Estimating Manifolds

Rates, Methods, and Surrogates

Christopher R. Genovese

Department of Statistics
Carnegie Mellon University

30 Apr 2012
Yale University

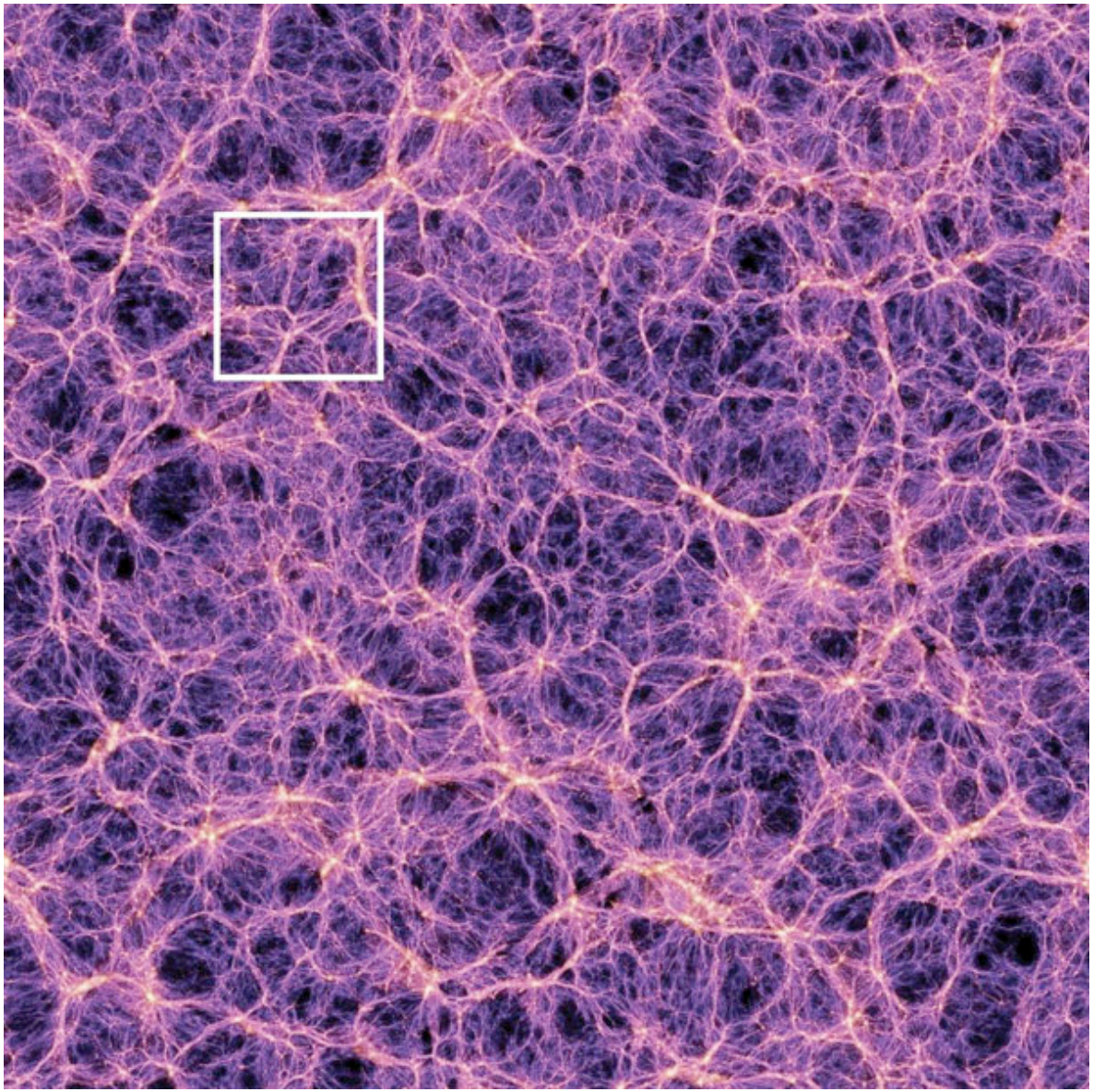
Collaborators

Larry Wasserman	Carnegie Mellon University
Isabella Verdinelli	Carnegie Mellon University and University of Rome
Marco Perone-Pacifico	University of Rome

Recent papers on this problem:

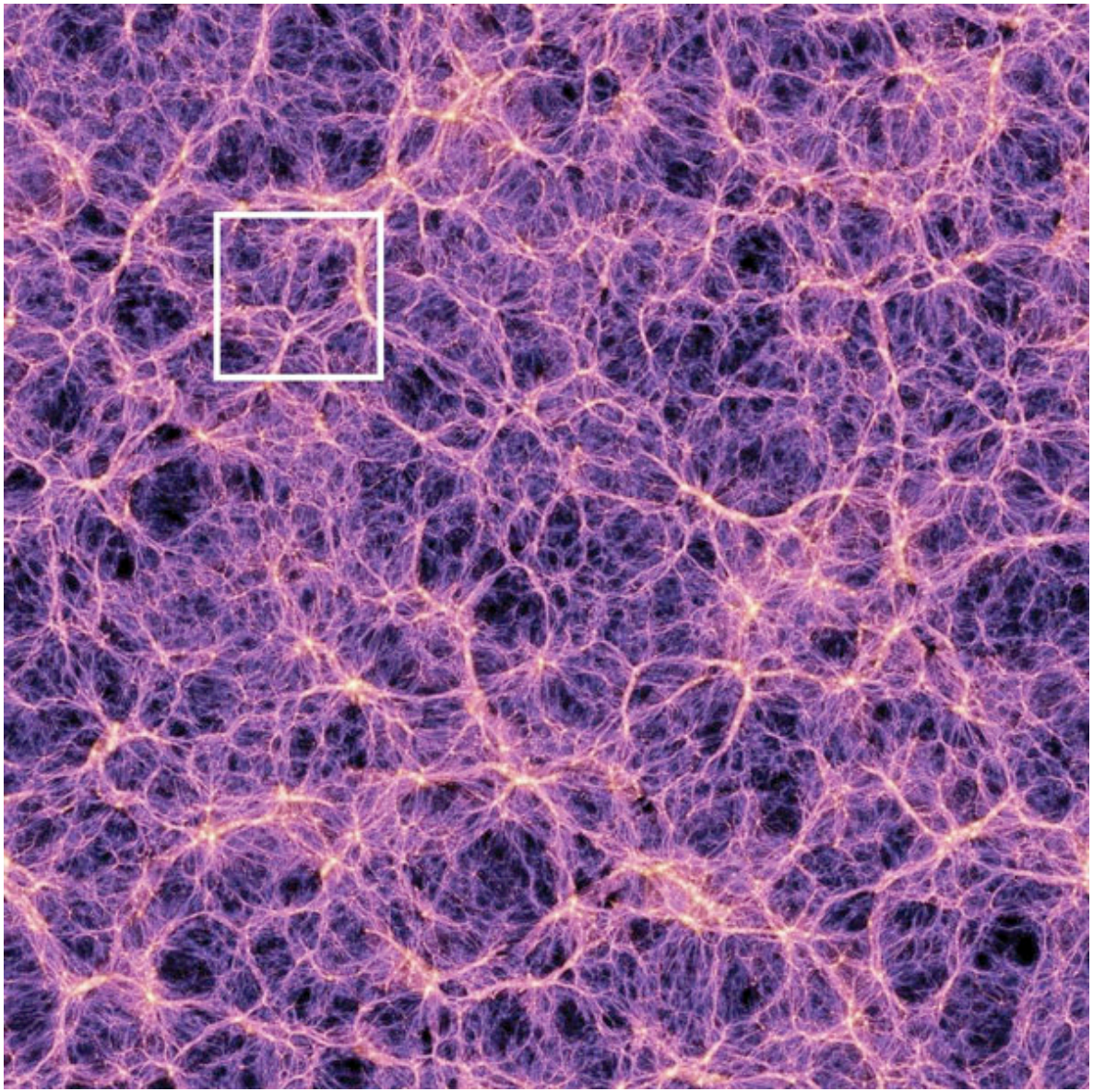
1. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2009. [*Ann. Stat.*, **37**]
2. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2010a. [arXiv:1003.5536 JASA]
3. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2010b. [arXiv:1007.0549 Annals]
4. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2011. [arXiv:1109.4540 JMLR]

Motivating Example: the "Cosmic Web"



Gnedin (2005)

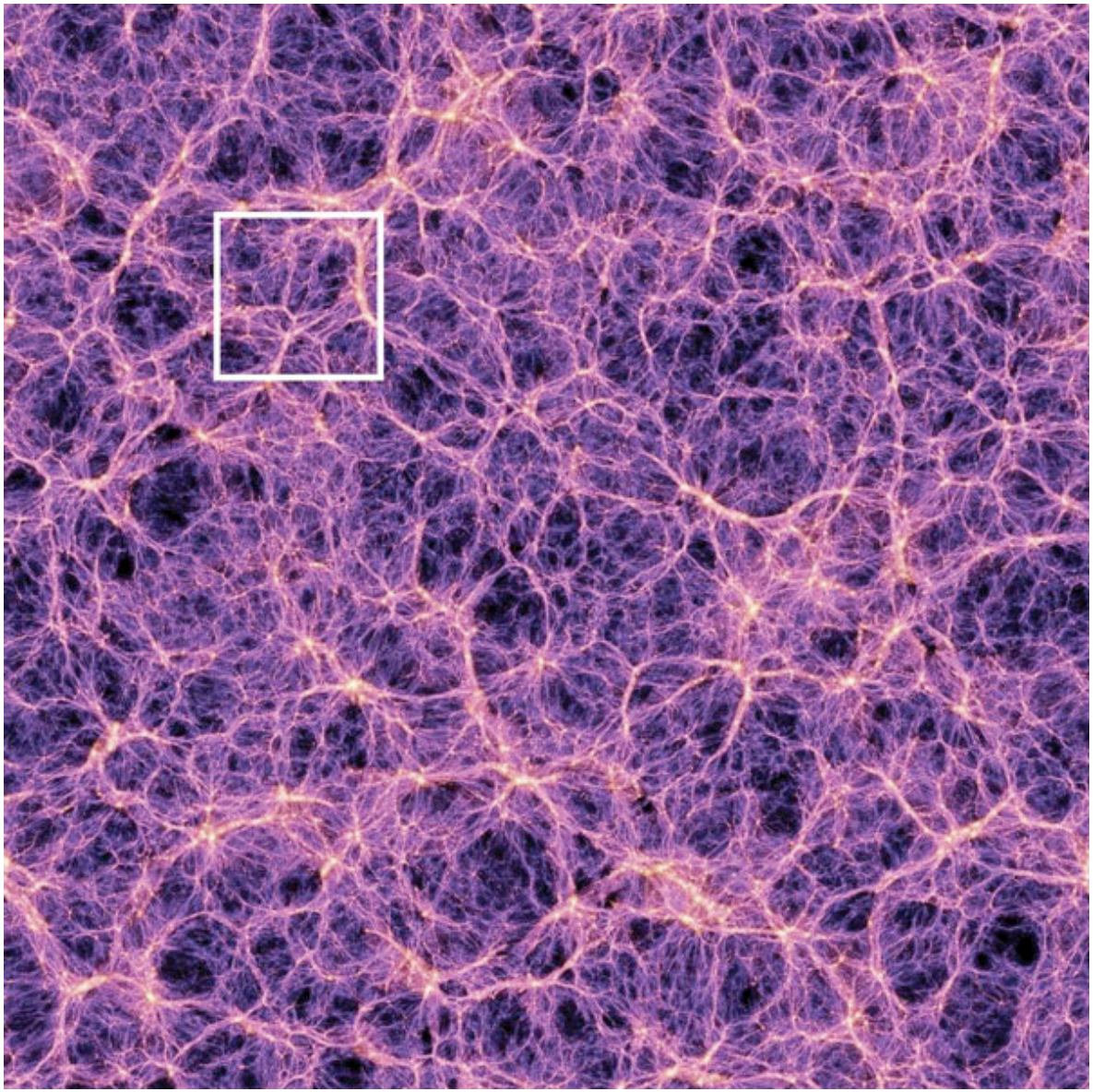
Motivating Example: the "Cosmic Web"



Matter is concentrated around lower dimensional features:

Gnedin (2005)

Motivating Example: the "Cosmic Web"

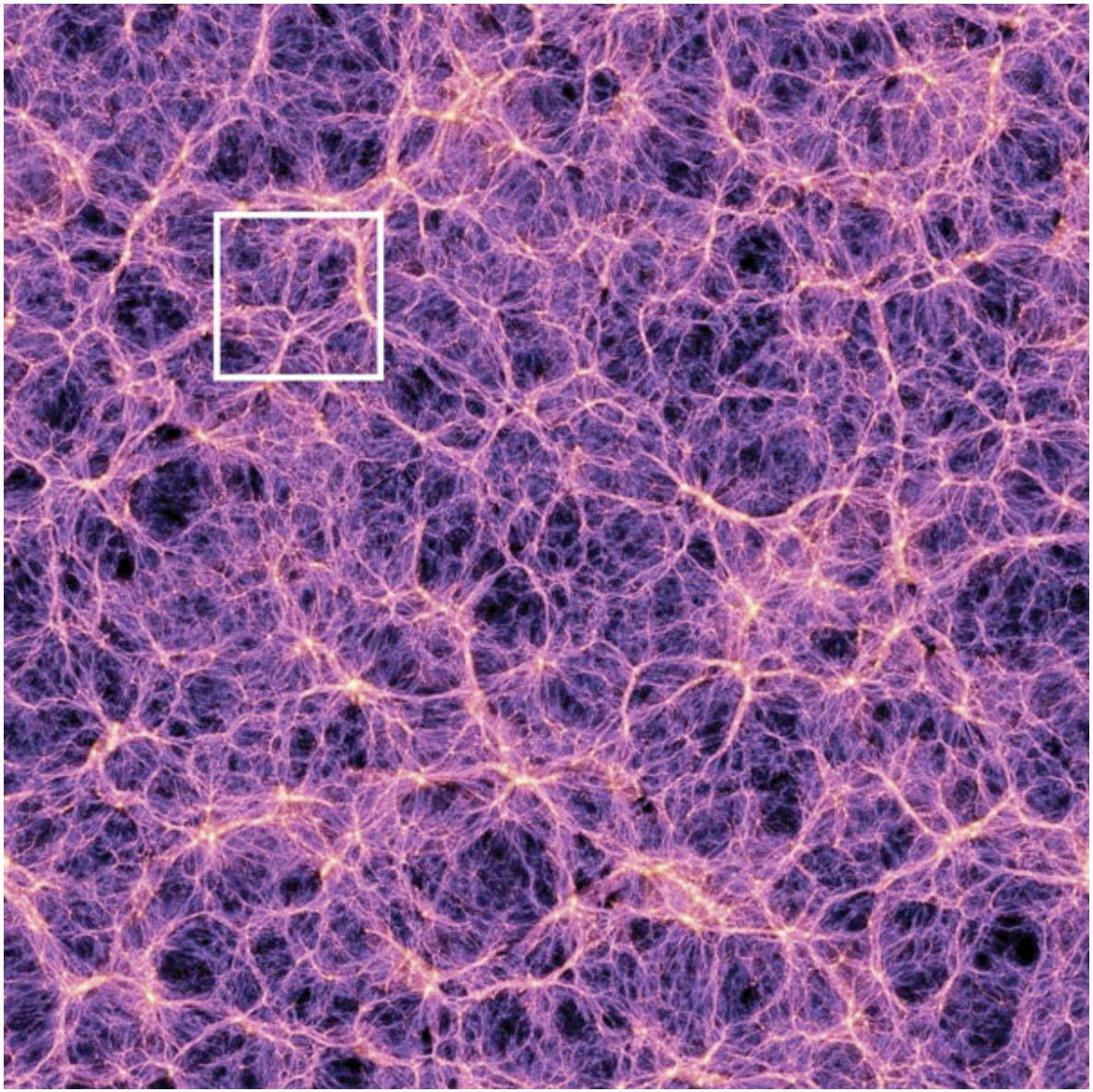


Matter is concentrated around lower dimensional features:

- 0-dimensional clusters

Gnedin (2005)

Motivating Example: the "Cosmic Web"

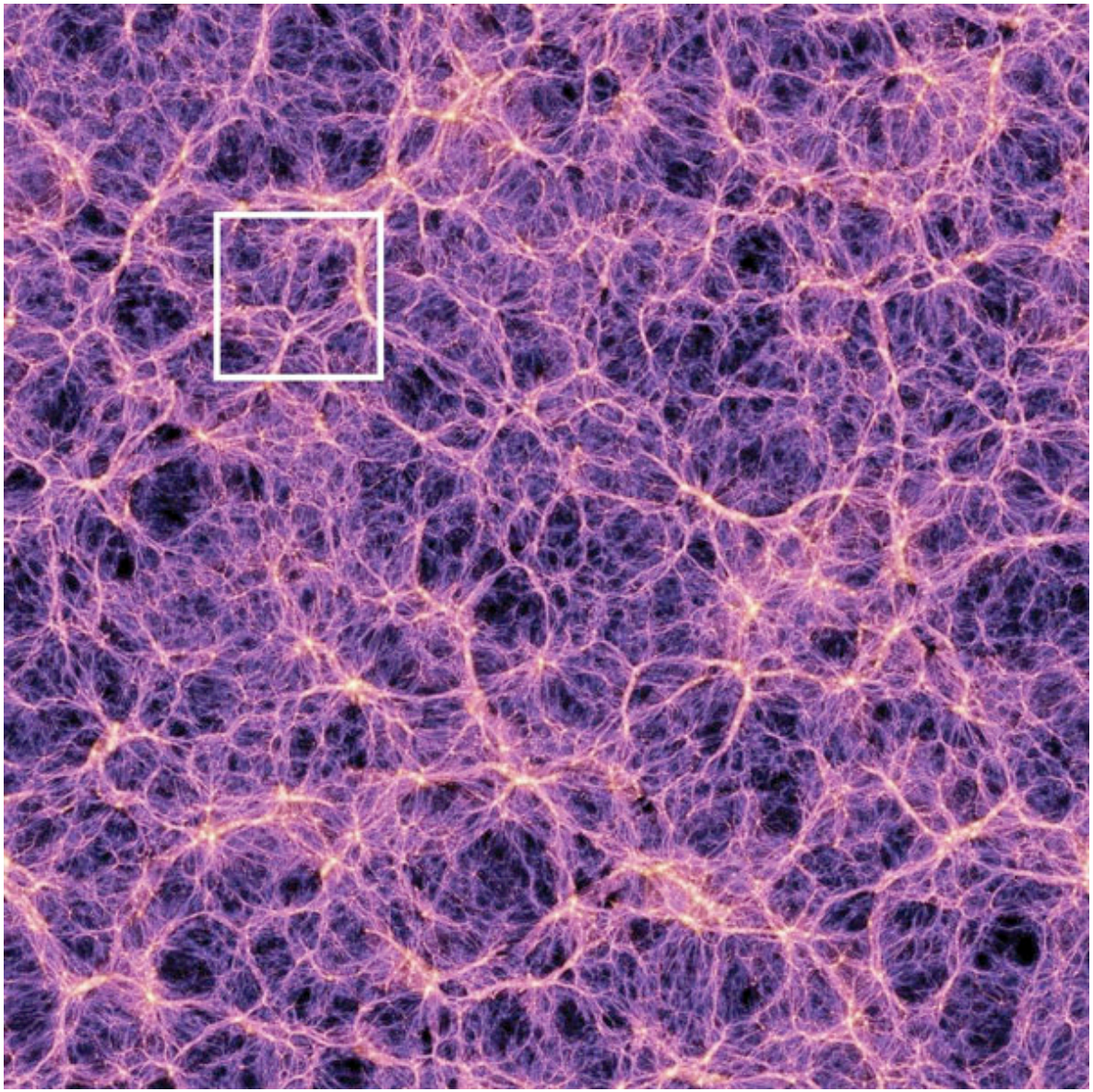


Matter is concentrated around lower dimensional features:

- 0-dimensional clusters
- 1-dimensional filaments

Gnedin (2005)

Motivating Example: the "Cosmic Web"

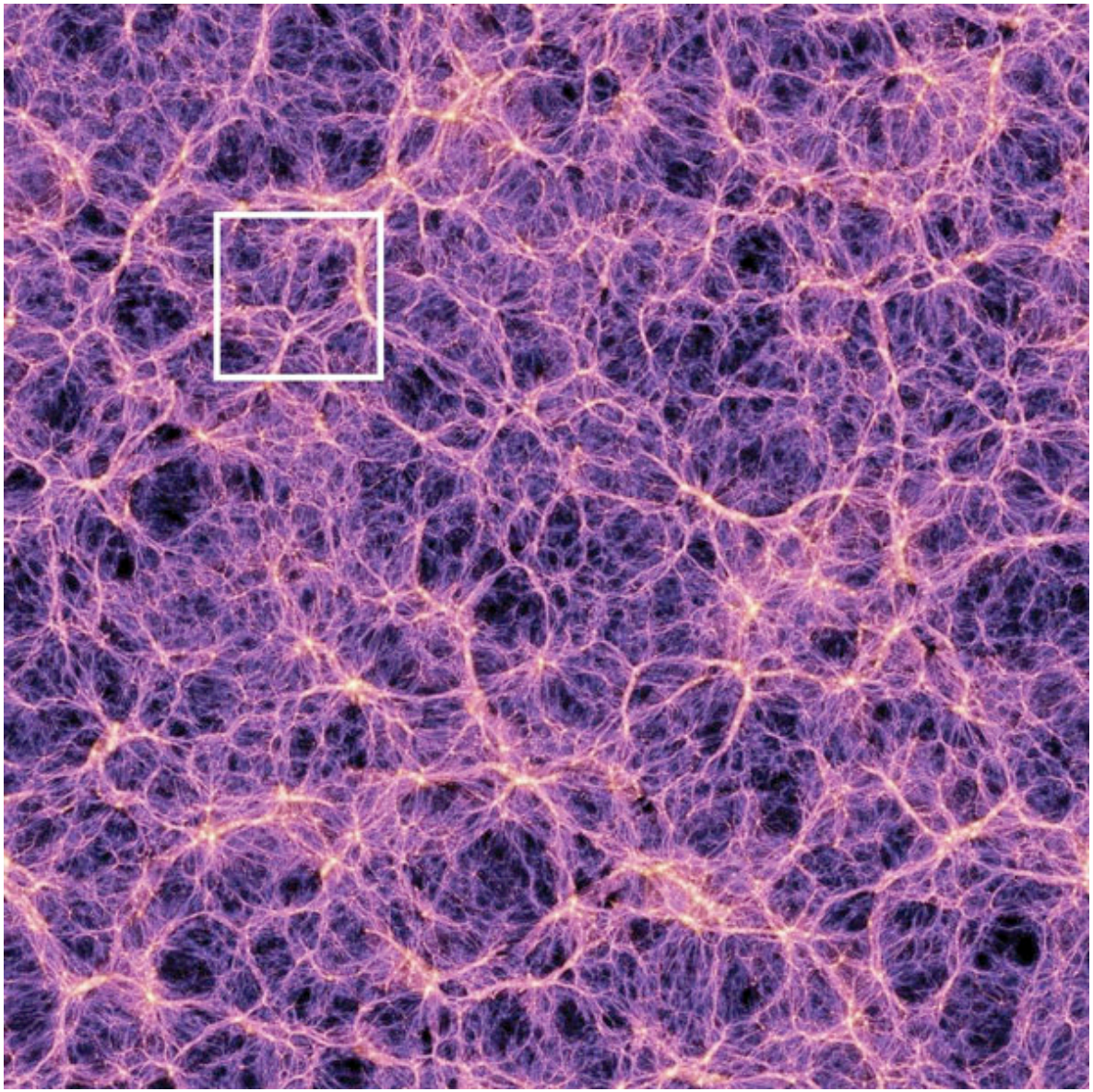


Matter is concentrated around lower dimensional features:

- 0-dimensional clusters
- 1-dimensional filaments
- 2-dimensional sheets

Gnedin (2005)

Motivating Example: the "Cosmic Web"



Matter is concentrated around lower dimensional features:

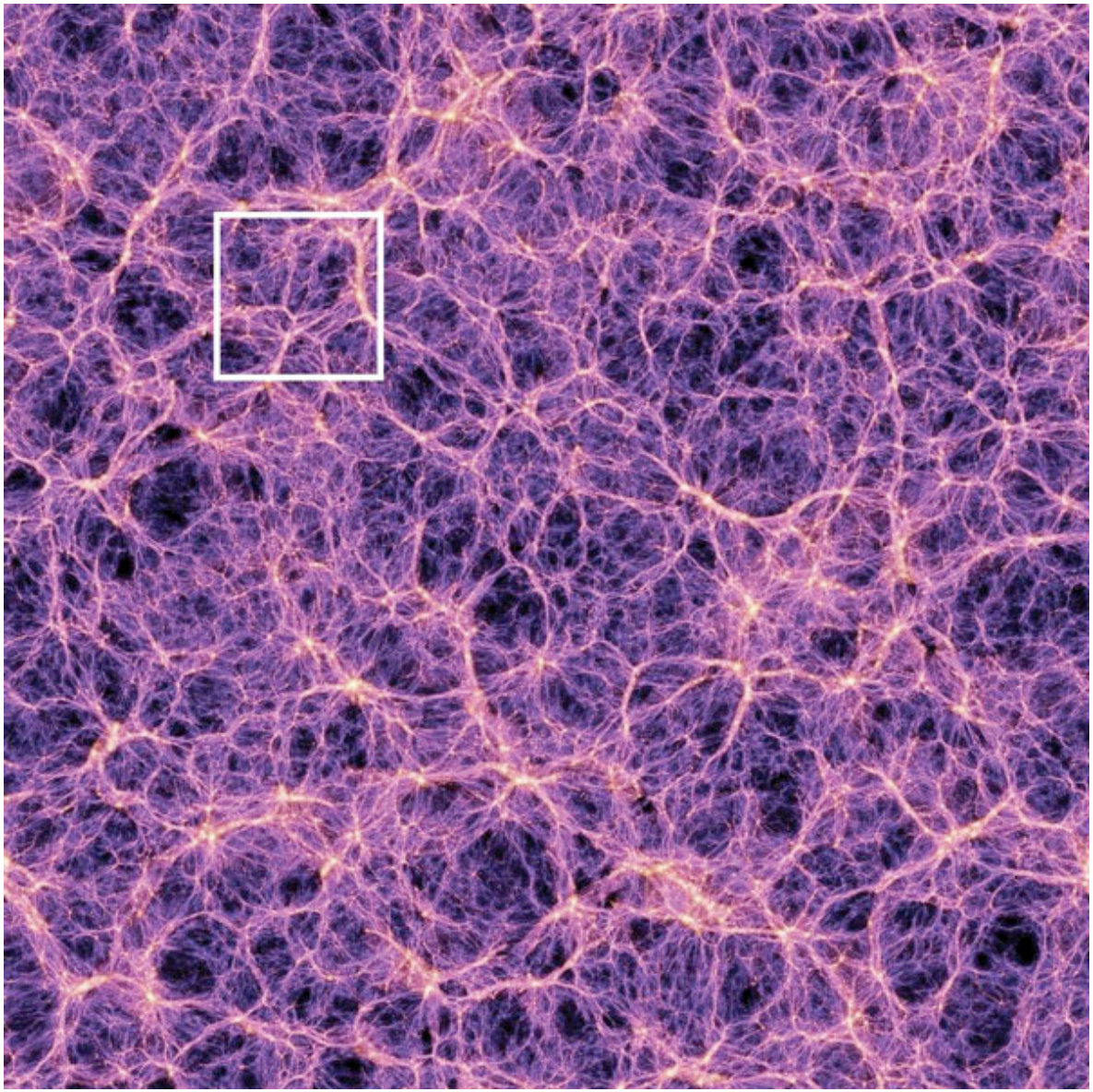
- 0-dimensional clusters
- 1-dimensional filaments
- 2-dimensional sheets

with intervening

- 3-dimensional voids.

Gnedin (2005)

Motivating Example: the "Cosmic Web"



Matter is concentrated around lower dimensional features:

- 0-dimensional clusters
- 1-dimensional filaments
- 2-dimensional sheets

with intervening

- 3-dimensional voids.

The distribution of these features has cosmological significance.

Gnedin (2005)

Low-Dimensional Structure in Point Cloud Data

Many datasets exhibit complex, low-dimensional structure.

Low-Dimensional Structure in Point Cloud Data

Many datasets exhibit complex, low-dimensional structure.

More Examples:

- Networks of blood vessels in medical imaging.
- River and road systems in remote sensing.
- Fault lines in seismology.
- Landmark paths for moving objects in computer vision.

In addition, high-dimensional datasets often have hidden structure that we would like to identify.

Low-Dimensional Structure in Point Cloud Data

Many datasets exhibit complex, low-dimensional structure.

More Examples:

- Networks of blood vessels in medical imaging.
- River and road systems in remote sensing.
- Fault lines in seismology.
- Landmark paths for moving objects in computer vision.

In addition, high-dimensional datasets often have hidden structure that we would like to identify.

Several distinct problems here, including:
Dimension Reduction, Clustering, and Estimation.

Low-Dimensional Structure in Point Cloud Data

Many datasets exhibit complex, low-dimensional structure.

More Examples:

- Networks of blood vessels in medical imaging.
- River and road systems in remote sensing.
- Fault lines in seismology.
- Landmark paths for moving objects in computer vision.

In addition, high-dimensional datasets often have hidden structure that we would like to identify.

Several distinct problems here, including:
Dimension Reduction, Clustering, and **Estimation**.

Manifolds and Manifold Complexes

Manifolds give a useful representation of low dimensional structure.

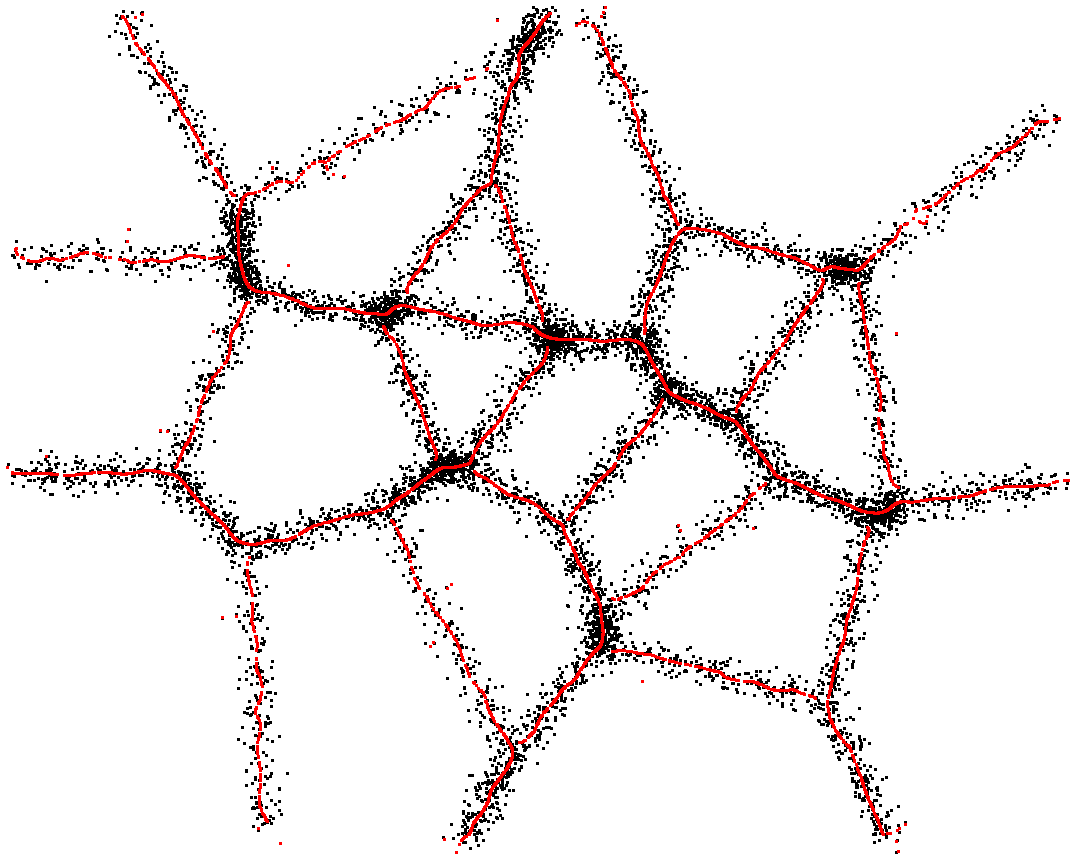
A manifold is a space that looks locally like a Euclidean space of some dimension (called the dimension of the manifold).

Examples: point (0-dim), filaments (1-dim), surface of the sphere or torus (2-dim), three-dimensional sphere, space-time (4-dim).

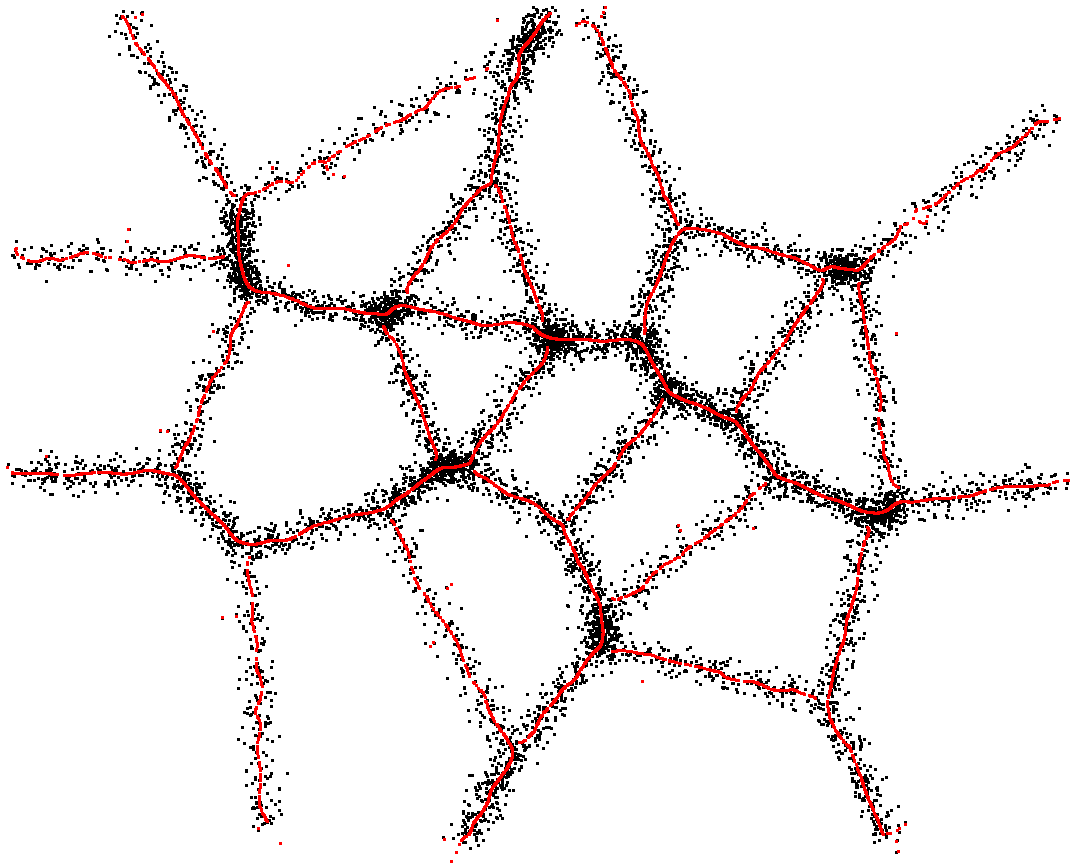
To allow for intersections and other complexities, consider a *union* of manifolds embedded in $\mathbb{R}^D$ with maximal dimensions $d < D$.

I will call this a d-dimensional manifold complex.

Example



Example



Challenge: Given a point cloud sampled from a manifold complex and then perturbed by noise, **accurately estimate the manifold complex.**

Road Map

Motivation

Manifold Estimation

Minimax Rates under Various Noise Models

Methods and Surrogates

Road Map

Motivation

Manifold Estimation

- Data Models and Objective
- Synthetic Example
- Reach and Distance
- Problem and Literature

Minimax Rates under Various Noise Models

Methods and Surrogates

Models for Manifold Estimation

Suppose M belongs to a class $\mathcal{M}$ (to be defined shortly) of d-dimensional "smooth" manifolds embedded in $\mathbb{R}^D$ for $D > d$.

G is a distribution on M, with density bounded away from 0 and ∞.

Draw $X_1, \ldots, X_n$ from G and then draw $Y_1, \ldots, Y_n$ according to one of four noise models:

1. noiseless: $Y_i = X_i$.
2. clutter: $Y_i = X_i$ with probability π, otherwise $Y_i \sim$ Uniform.
3. perpendicular: $Y_i = X_i + Z_i$ where Z_i is normal to M.
   (See also Niyogi, Smale, Weinberger 2008.)
4. additive: $Y_i = X_i + Z_i$ and $\epsilon_i \sim \Phi$.

Want to estimate M from $Y_1, \ldots, Y_n$.

Models for Manifold Estimation

Suppose M belongs to a class $\mathcal{M}$ (to be defined shortly) of d-dimensional "smooth" manifolds embedded in $\mathbb{R}^D$ for $D > d$.

G is a distribution on M, with density bounded away from 0 and ∞.

Draw $X_1, \ldots, X_n$ from G and then draw $Y_1, \ldots, Y_n$ according to one of four noise models:

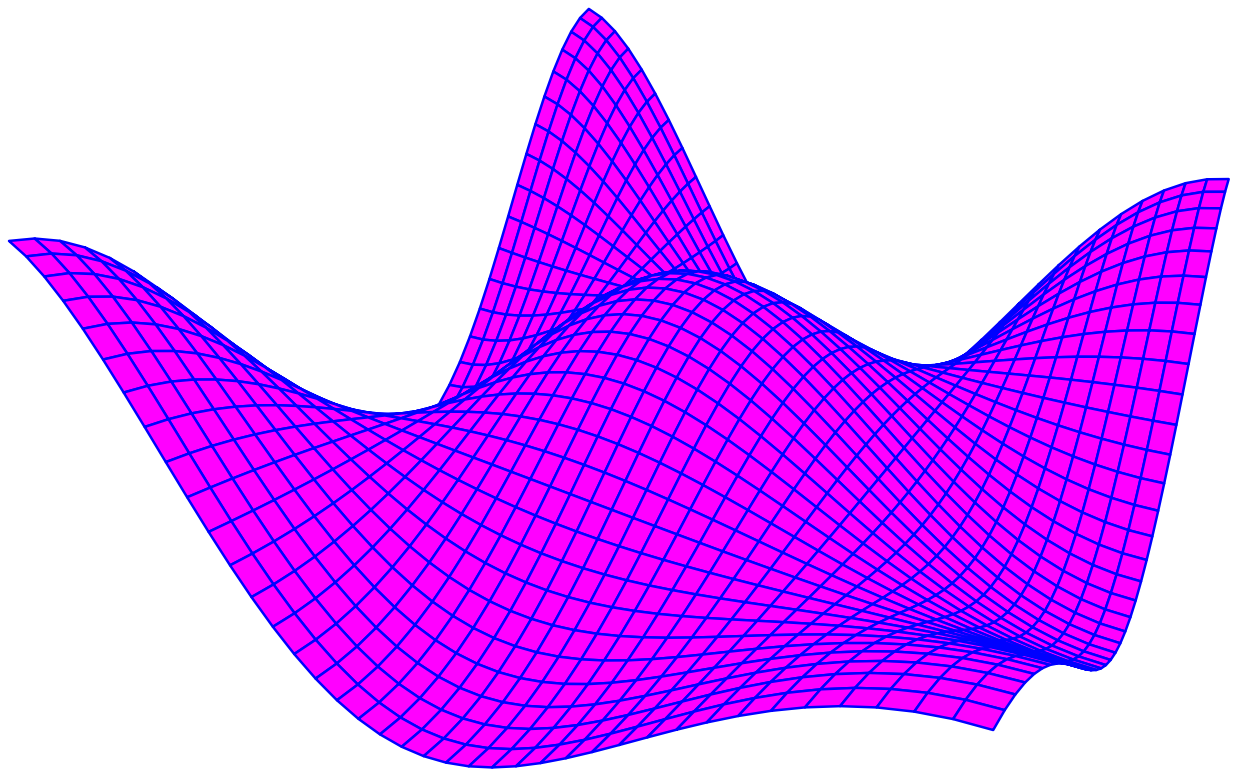
1. noiseless: $Y_i = X_i$.
2. clutter: $Y_i = X_i$ with probability π, otherwise $Y_i \sim$ Uniform.
3. perpendicular: $Y_i = X_i + Z_i$ where Z_i is normal to M.
   (See also Niyogi, Smale, Weinberger 2008.)
4. additive: $Y_i = X_i + Z_i$ and $\epsilon_i \sim \Phi$.

Want to estimate M from $Y_1, \ldots, Y_n$.

The noise model strongly affects the difficulty of this problem.

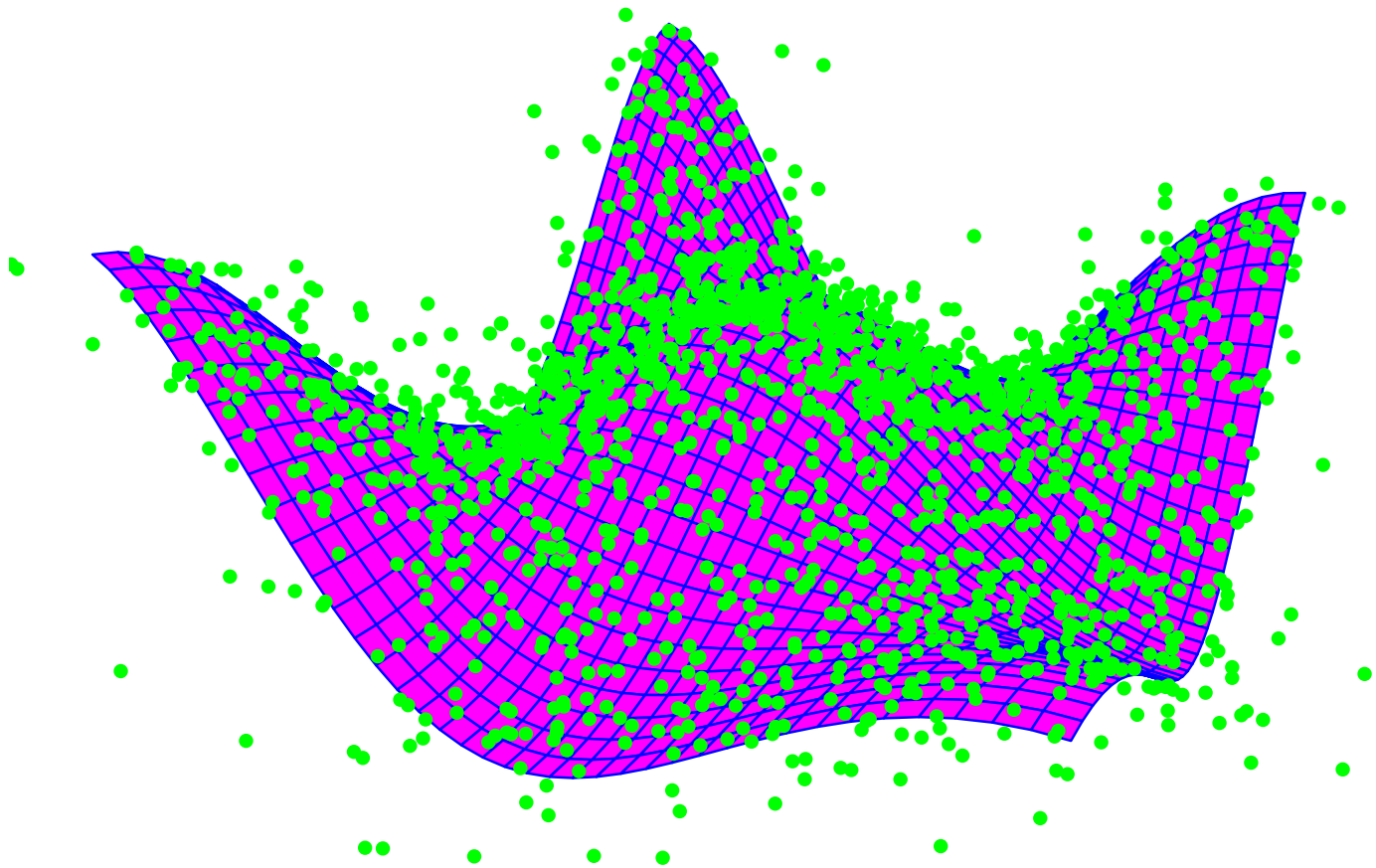
A Synthetic Example

An smooth manifold with $d = 2, D = 3$



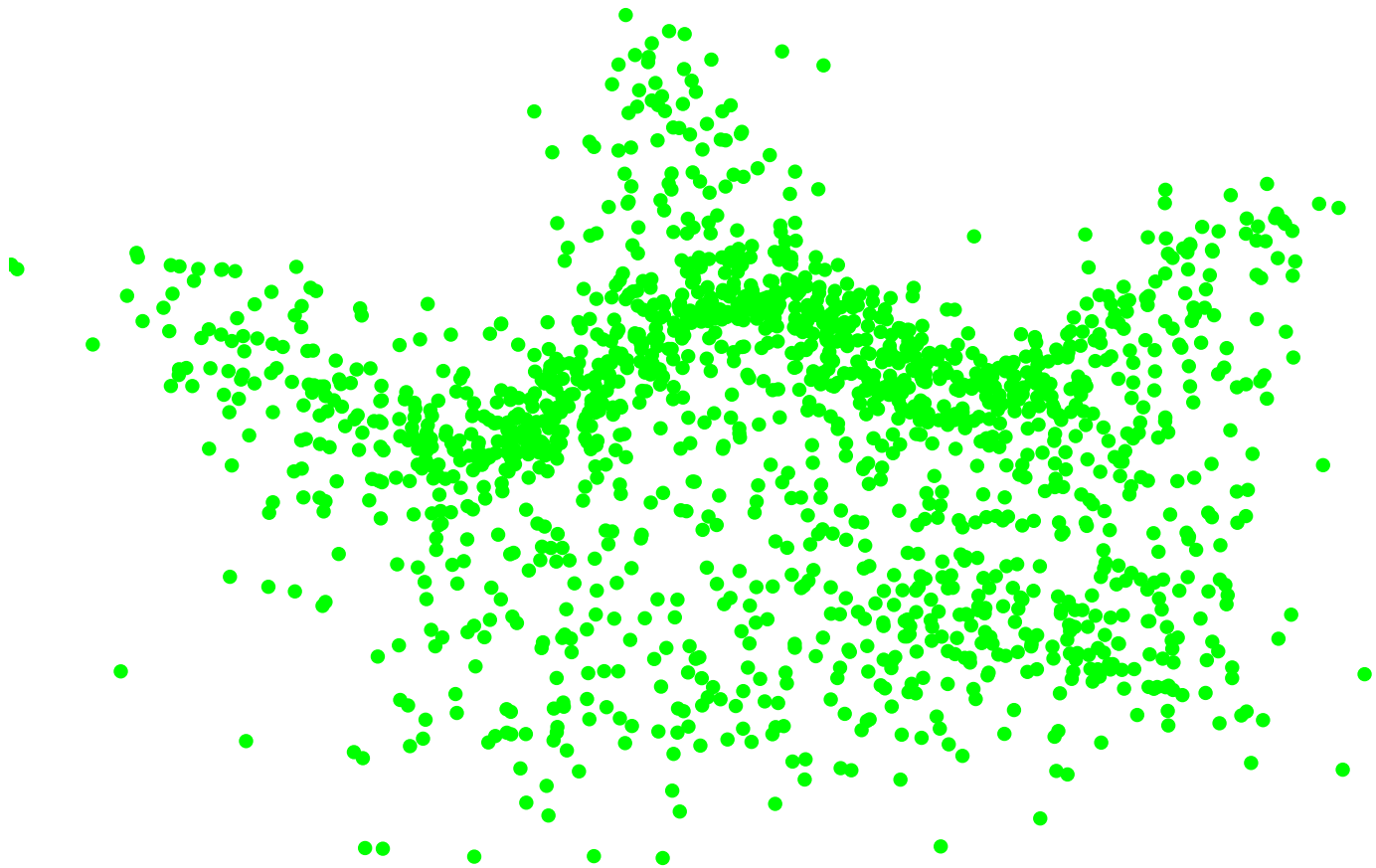
A Synthetic Example

An smooth manifold with $d = 2, D = 3$ plus data drawn from the additive model



A Synthetic Example

The data drawn from the additive model



Minimax Manifold Estimation

Define $\mathcal{M} \equiv \mathcal{M}_\kappa = \{M : \text{ reach}(M) \geq \kappa\}$ and $\mathcal{Q} = \{Q_M : \ M \in \mathcal{M}\}$, where

$$Q_M(A) = \int_M \Phi(Y \in A \mid X = x)\, \mathrm{d}G(x)$$

is the induced distribution on Y.

Draw $Y_1, Y_2, \ldots, Y_n$ IID from Q_M and estimate $\widehat{M} \equiv \widehat{M}_n$.

Goal: determine the minimax risk

$$R_n = \inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \, \text{Haus}(\widehat{M}_n, M),$$

at least up to rates, with Hausdorff loss.

The Reach of a Manifold

Define the reach of a manifold M as follows:

> reach(M) *is the largest* r *such that* $d(x, M) \leq r$ *implies that* x *has a unique projection onto* M.

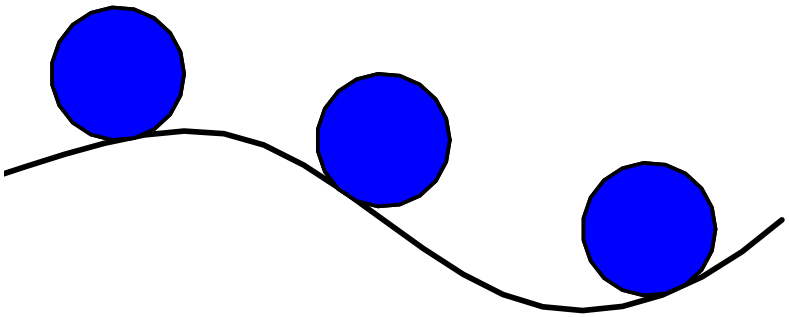
This is also called the thickness or condition number of the manifold; see Niyoki, Smale, and Weinberger (2009).

Intuitively, a manifold M with reach(M) $= \kappa$ has two constraints:

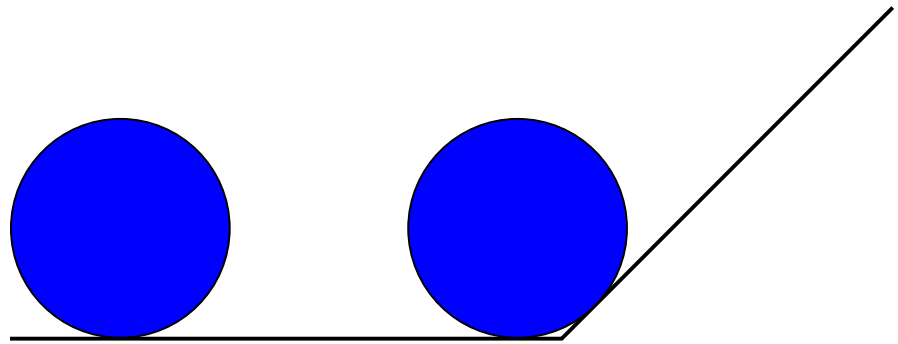
1. *Curvature.* A ball or radius $r \leq \kappa$ can roll freely and smoothly over M, but a ball or radius $r > \kappa$ cannot.
2. *Separation.* M is at least 2κ from self-intersecting.

Reach in One Dimension

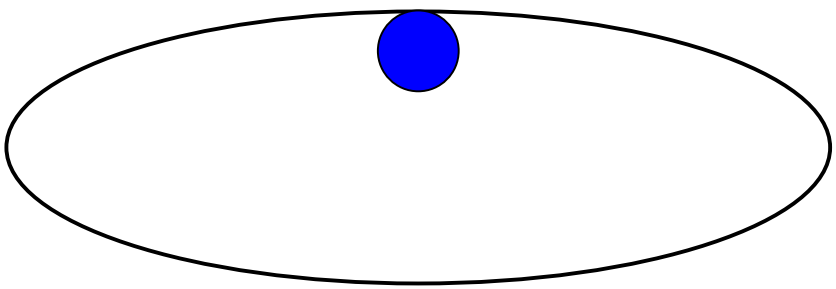
circles have radius r



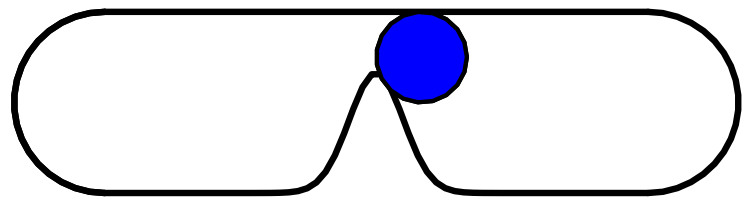
$\kappa > r$



$\kappa < r$



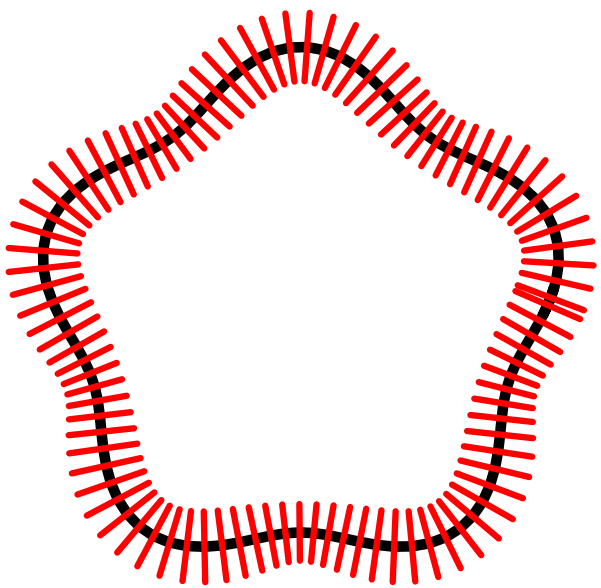
$\kappa > 2r$

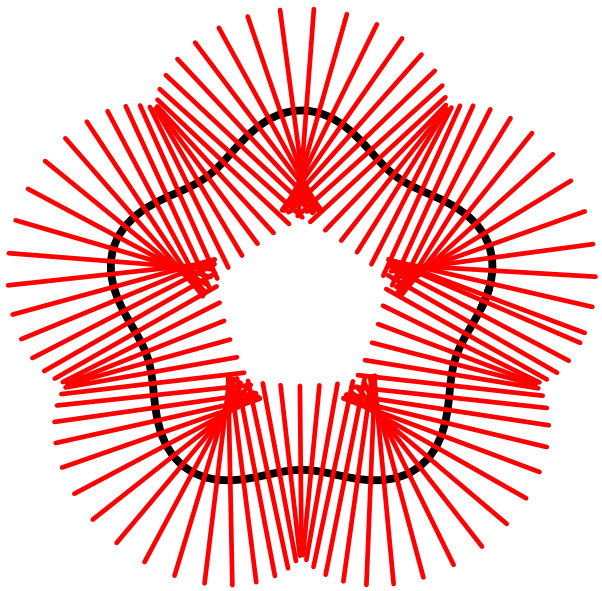


$\kappa < 2r$

Reach Visualized

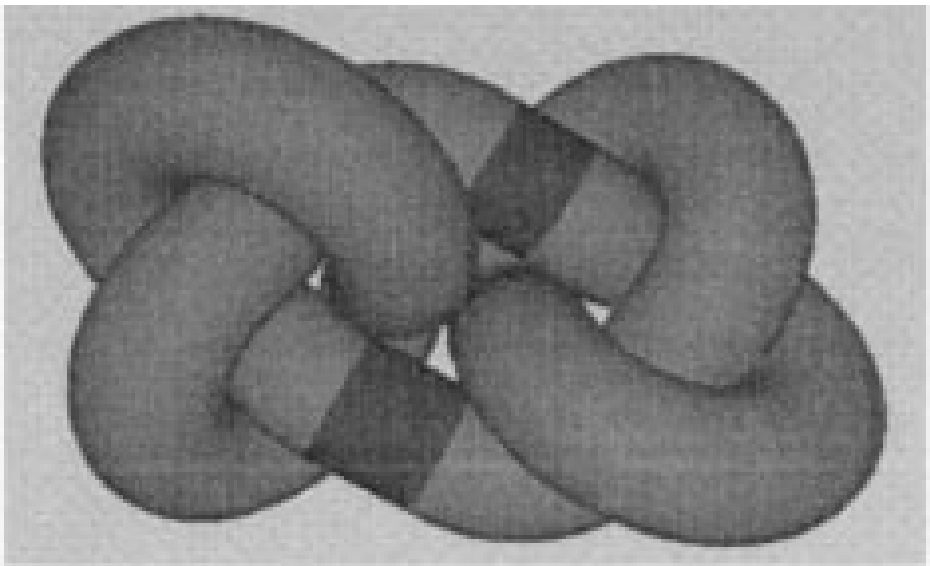
Normals of size $<$ reach(M) do not cross.



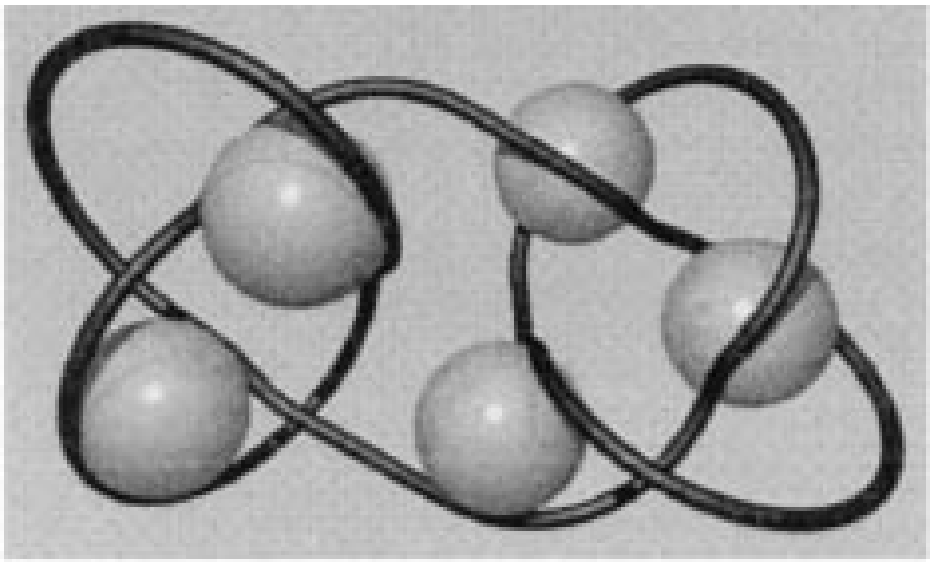


Reach Visualized

A large value of reach(M) implies that the manifold M is smooth and not too tightly looped around itself



a



b

from Gonzalez and Maddocks (1999)

Reach of case (a) $\ll$ Reach of case (b)

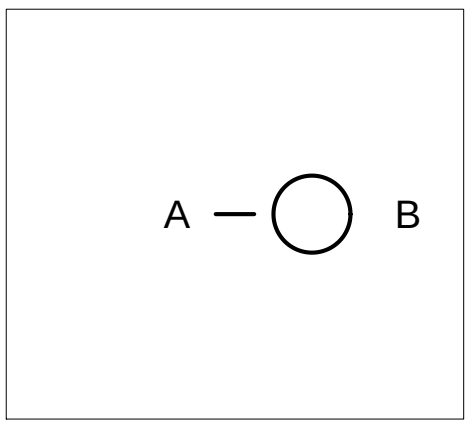
Hausdorff Distance

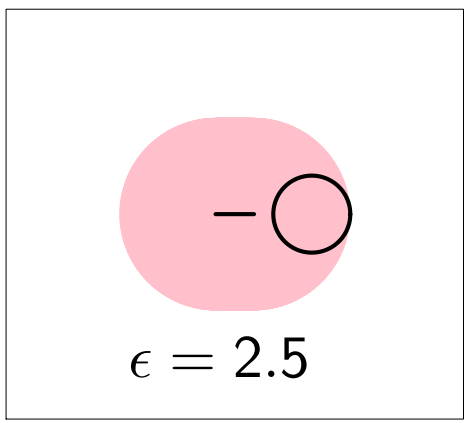
Given two subsets of $\mathbb{R}^D$, A and B:

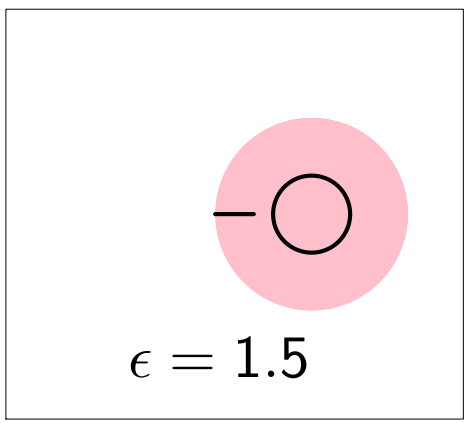
$$\text{Haus}(A, B) = \inf \{\epsilon : \ A \subset B \oplus \epsilon \text{ and } B \subset A \oplus \epsilon\}$$

where $A \oplus \epsilon = \bigcup_{x \in A} B(x, \epsilon)$ and $B(x, \epsilon) = \{y : \ \|x - y\| \leq \epsilon\}$.

Example:







$$\text{Haus}(A, B) = \max \{2.5, 1.5\} = 2.5$$

Minimax Manifold Estimation

Define $\mathcal{M} \equiv \mathcal{M}_\kappa = \{M : \text{reach}(M) \geq \kappa\}$ and $\mathcal{Q} = \{Q_M : M \in \mathcal{M}\}$, where

$$Q_M(A) = \int_M \Phi(Y \in A \mid X = x)\, \mathrm{d}G(x)$$

is the induced distribution on Y.

Draw $Y_1, Y_2, \ldots, Y_n$ IID from Q_M and estimate $\widehat{M} \equiv \widehat{M}_n$.

Goal: determine the minimax risk

$$R_n = \inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \, \text{Haus}(\widehat{M}_n, M),$$

at least up to rates, with Hausdorff loss.

Existing Literature

Computational geometry (e.g., Cheng et al. 2005, Dey 2006)

Here, "noise" does *not* have the statistical meaning of points drawn randomly from a distribution; instead, data must be close to M but not too close to each other. (There are a few notable exceptions.)

Manifold learning (e.g., Ozertem and Erdogmus 2011)

The primary focus here is on dimension reduction

Homology estimation (e.g., Niyoki, Smale, and Weinberer 2009)

Focus on topological rather than geometric information

Filaments, principle curves, support estimation, ...

e.g., Hastie and Stuetzle (1989), Tibshirani (1992), Arias-Castro et al. (2006)

Road Map

Motivation

Manifold Estimation

Minimax Rates under Various Noise Models

- Results
- Lower Bound under Perpendicular Noise
- Upper Bound under Perpendicular Noise
- Clutter
- Additive Model

Methods and Surrogates

Minimax Rates under Various Noise Models

$$\inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \, \mathrm{Haus}(\widehat{M}_n, M) \asymp C\psi_n$$

Noise Model	ψ_n
Clutter/Noiseless	$(\pi n)^{-\frac{2}{d}}$
Perpendicular Compact	$n^{-\frac{2}{2+d}}$
Additive Compact/Polynomial	*in progress*
Additive sub-Gaussian	$(\log n)^{-1}$

Note that these rates do not depend on the ambient dimension D.

There are strong connections between the additive noise model and errors-in-variables regression but also some notable differences.

Proof Sketch

The lower bound is established with Le Cam's Lemma.

Suppose $Y_1, \ldots, Y_n$ drawn IID from $\mathcal{Q}$, an estimator $\widehat{\theta} \equiv \widehat{\theta}(Y_1, \ldots, Y_n)$, and a (weak, semi-) metric ρ.

Then for any pair $Q_0, Q_1 \in \mathcal{Q}$

$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \rho(\widehat{\theta}, \theta(Q)) \geq C \rho(\theta(Q_0), \theta(Q_1))(1 - \mathrm{TV}(Q_0, Q_1))^{2n},$$

where

$$TV(Q_0(A), Q_1(A)) = \sup_A |Q_0(A) - Q_1(A)| = \frac{1}{2} \int |q_0 - q_1|.$$

Hence, for each given Hausdorff distance, we want to choose a least favorable pair of manifolds whose distributions are as hard to distinguish as possible.

Perpendicular Noise: Sketch of Lower Bound

Construct M_0 and M_1 such that:

- $M_i \in \mathcal{M}_\kappa$
- $\text{Haus}(M_1, M_0) = \gamma$
- $\text{TV} \equiv \int |q_1 - q_0| = O(\gamma^{(d+2)/2})$, which is minimum possible.

Apply Le Cam's Lemma: For any $\widehat{M}$:

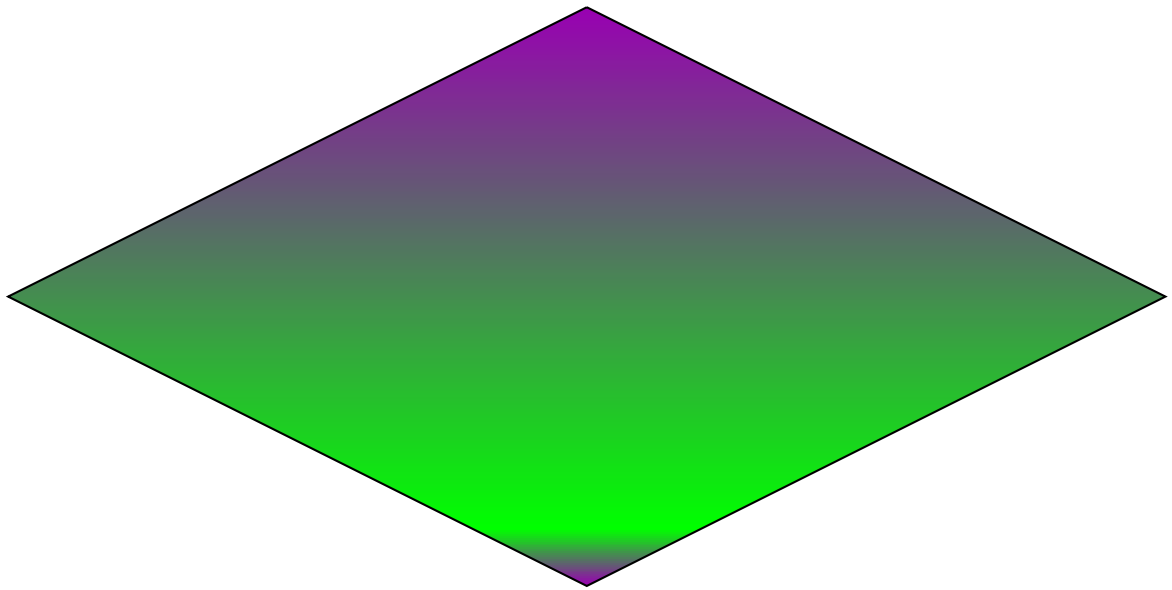
$$\sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \text{Haus}(M, \widehat{M}) \geq \text{Haus}(M_1, M_0) \times (1 - \text{TV})^{2n}$$
$$= \gamma(1 - c\gamma^{(d+2)/2})^{2n}.$$

Setting $\gamma = n^{-2/(d+2)}$ yields the result.

Least Favorable Pair M_0 and M_1: M_0 = plane and M_1 = "flying saucer".

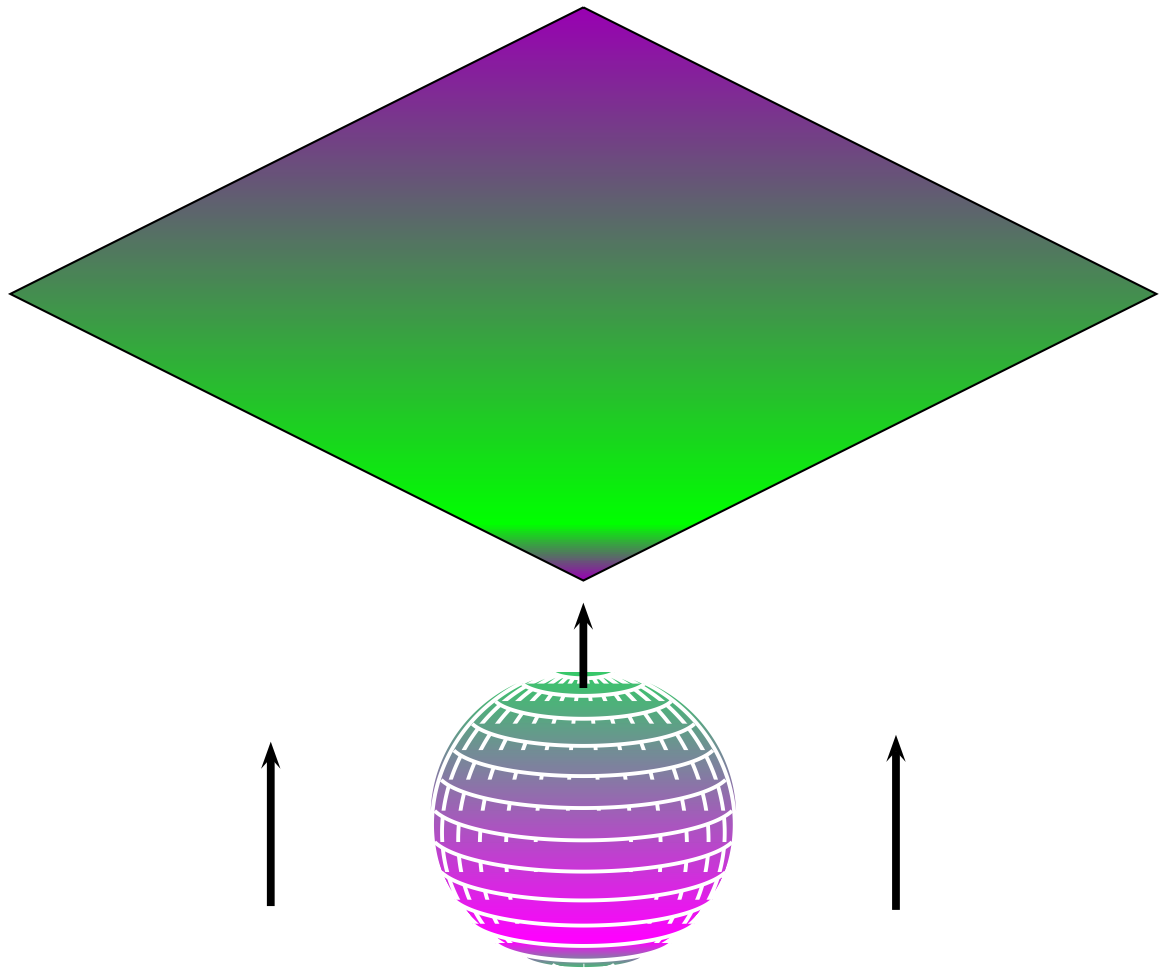
Constructing M_1

Start with M_0 . . .



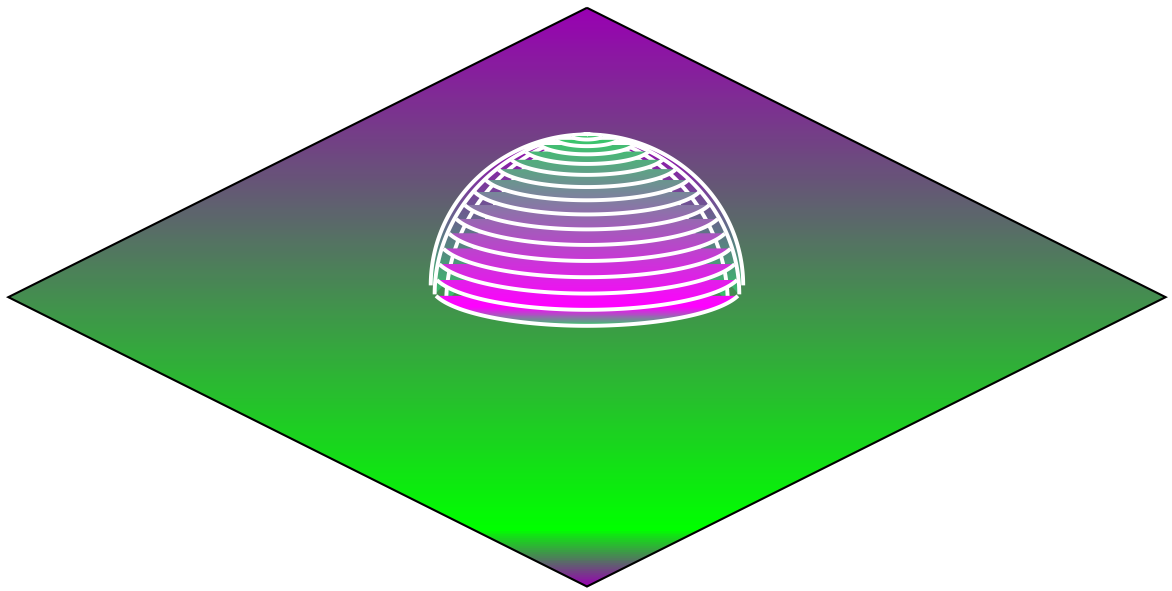
Constructing M_1

Push up κ-ball,



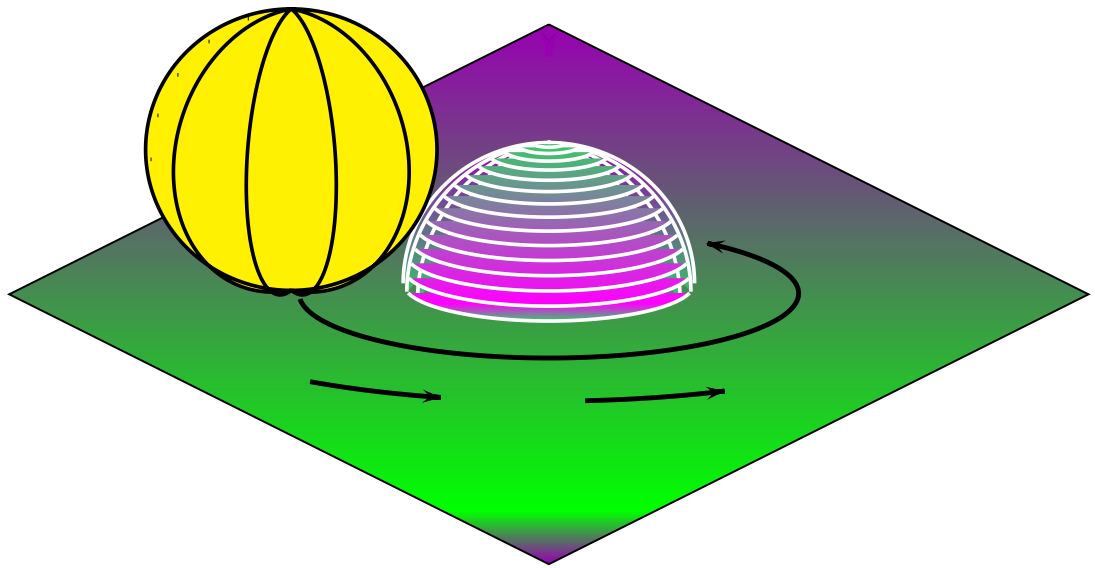
Constructing M_1

Push up κ-ball, through the plane to height γ. But reach still 0 ...



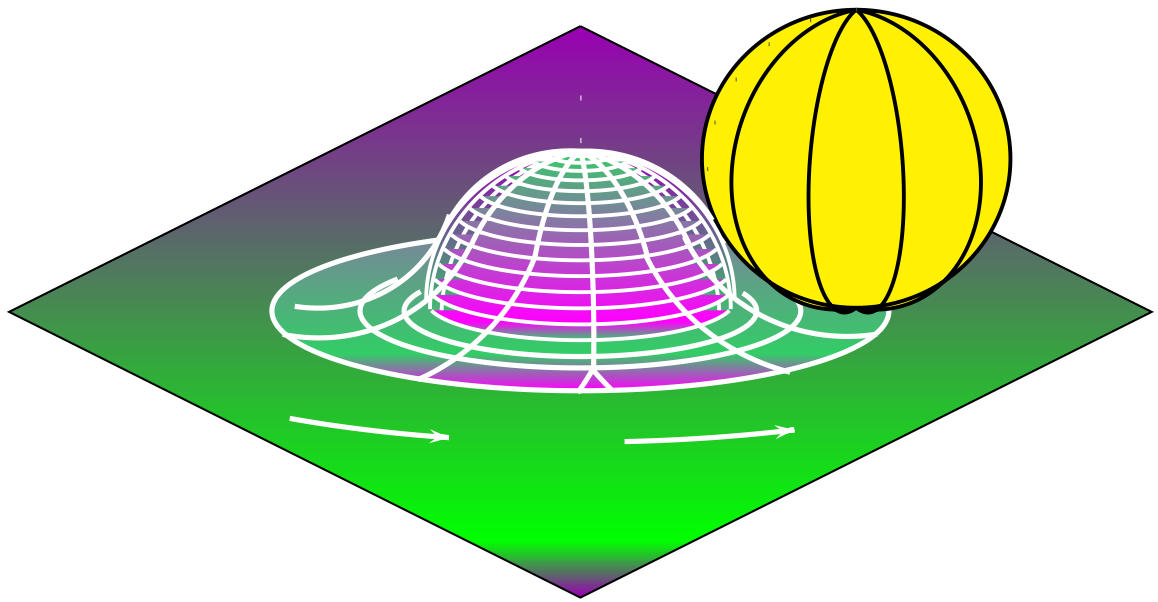
Constructing M_1

But reach still 0, so smooth the corners.



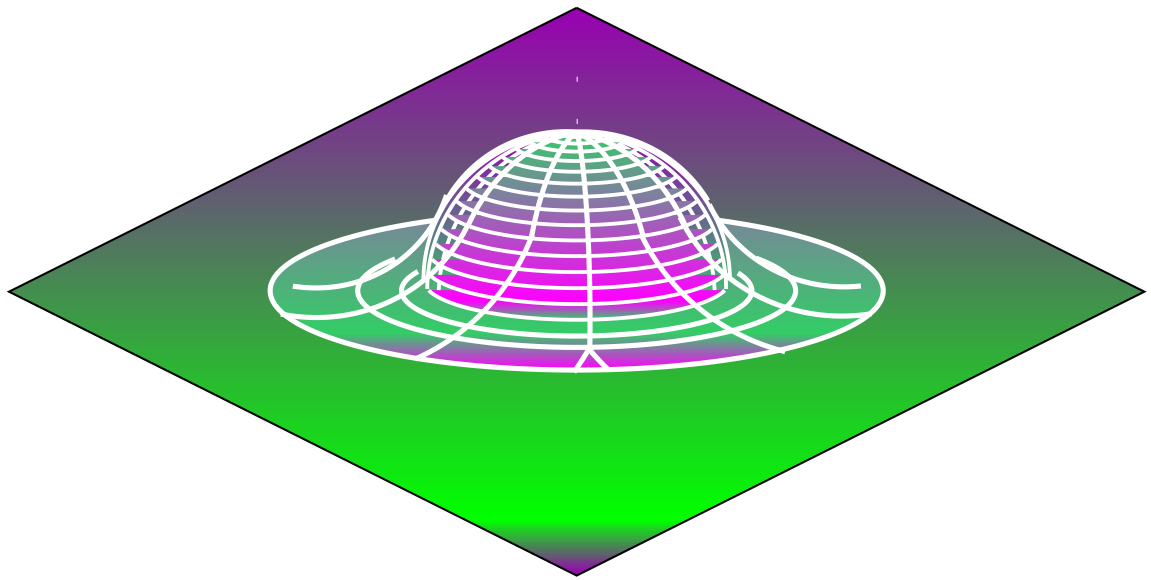
Constructing M_1

Smooth the corners ...



Constructing M_1

Flying Saucer M_1



Perpendicular Noise: Sketch of Upper Bound

Construct an "estimator" that achieves the bound:

1. Split the data into two halves.
2. Using the first half, construct a pilot esimator.
   This is a (sieve) maximum likelihood estimator.
3. Cover the pilot estimator with thin, long, slabs.
4. Using the second half of the data, fit local linear estimators $\widehat{M}_j$ in slab j
5. $\widehat{M} = \bigcup_j \widehat{M}_j$.

The details are messy and the estimator is not practical, but it suffices for establishing the bound.

Clutter Model

Suppose

$$Y_1, \ldots, Y_n \sim Q \equiv (1-\pi)U + \pi G$$

where $0 < \pi \leq 1$, U is uniform on the compact set $\mathcal{K} \subset \mathbb{R}^D$, and G supported on M as before.

Then,

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{Q^n} \text{Haus}(\widehat{M}, M) \asymp^* C \left(\frac{1}{n\pi}\right)^{\frac{2}{d}}.$$

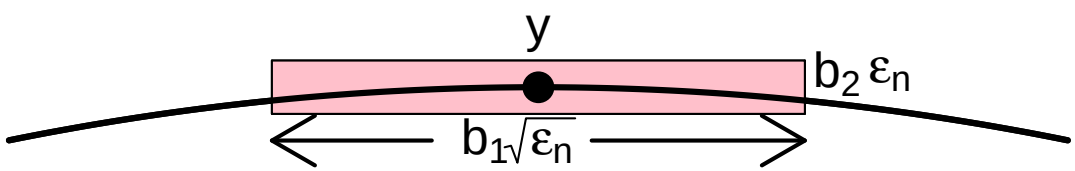
(The $\asymp^*$ means I am hiding log factors.)

Lower bound uses the same least favorable pair.

Clutter Model: Upper Bound

Let

- $\epsilon_n = (\log n/n)^{2/d}$.
- $\widehat{Q}_n$ be the empirical measure.
- $S_M(y)$ denotes a $\epsilon^{d/2} \times \epsilon^{D-d}$ slab:



Define

$$s(M) = \inf_{y \in M} \widehat{Q}_n[S_M(y)] \quad \text{and} \quad \widehat{M}_n = \operatorname*{argmax}_M s(M).$$

Additive Model

$X_1, X_2, \ldots, X_n \sim G$ where support$(G) = M$, and

$$Y_i = X_i + Z_i, \qquad i = 1, \ldots, n,$$

where $Z_i \sim \Phi =$ Gaussian.

This is analogous to an errors-in-variables problem, except:

1. We want to estimate the support of G not G itself.
2. G is singular.
3. The underlying object is a manifold not a function.

Additive Model

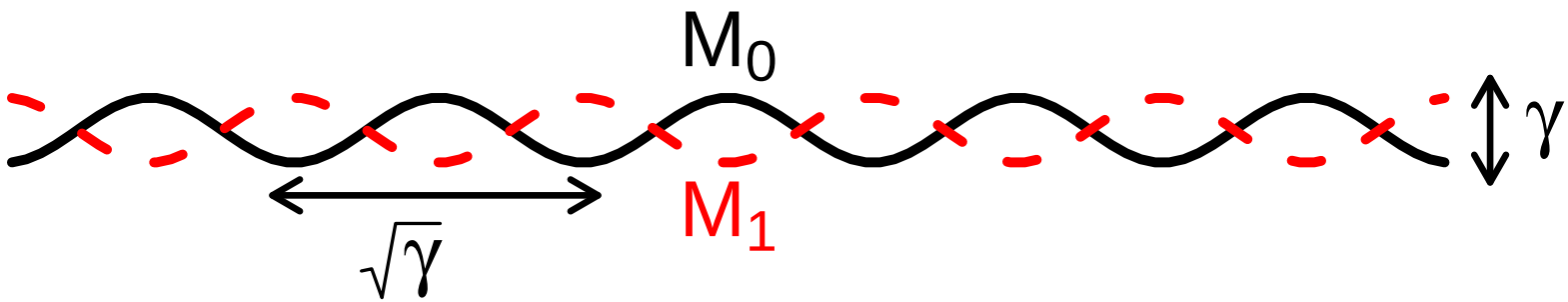
For technical reasons, we allow the manifolds to be noncompact. Define a truncated loss function,

$$L(M, \widehat{M}) = H(M \cap \mathcal{K}, \widehat{M} \cap \mathcal{K}).$$

Then,

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[L(M, \widehat{M})] \geq \frac{C}{\log n}.$$

Rate is similar to deconvolution but the proof is somewhat different (since Q_0 and Q_1 have different supports). Least favorable pair:



Additive Model: Upper Bound

Let $\widehat{g}$ be a deconvolution density estimator (though G has no density), and let $\widehat{M} = \{\widehat{g} > \lambda\}$.

Fix any $0 < \delta < 1/2$.

$$\inf_{\widehat{M}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[L(M, \widehat{M})] \leq C \left(\frac{1}{\log n} \right)^{\frac{1-\delta}{2}}.$$

In some special cases, we can achieve $\frac{1}{\log n}$ but, in general, not.

Road Map

Motivation

Manifold Estimation

Minimax Rates under Various Noise Models

Methods and Surrogates

- Rate-Hard Problems and Surrogates
- The Mean Shift Algorithm and Modifications
- A Hyper-Ridge Estimator

Surrogates for Rate-Hard Problems

Problems with rates like $1/\log n$ seem to offer little practical hope for good performance.

But it is sometimes possible to define a surrogate for the true object that

- captures essential features of the true object, and
- can be estimated with a good rate of convergence.

Example: Uniform confidence bands (Genovese and Wasserman 2008).

Strategy: Define a surrogate $\widetilde{M}$, called the **hyper-ridge set**, for the manifold complex M. Focus on estimating $\widetilde{M}$ accurately.

$\widetilde{M}$ is, roughly speaking, a smoother, slightly biased version of M.

Once we accept some bias, the curse of dimensionality becomes less daunting.

Hyper-Ridge Sets

$Y_1, \ldots, Y_n$ sampled IID from $Q = (1-\pi)U + \pi(G \star \Phi_\sigma)$, the additive model with clutter.

Let

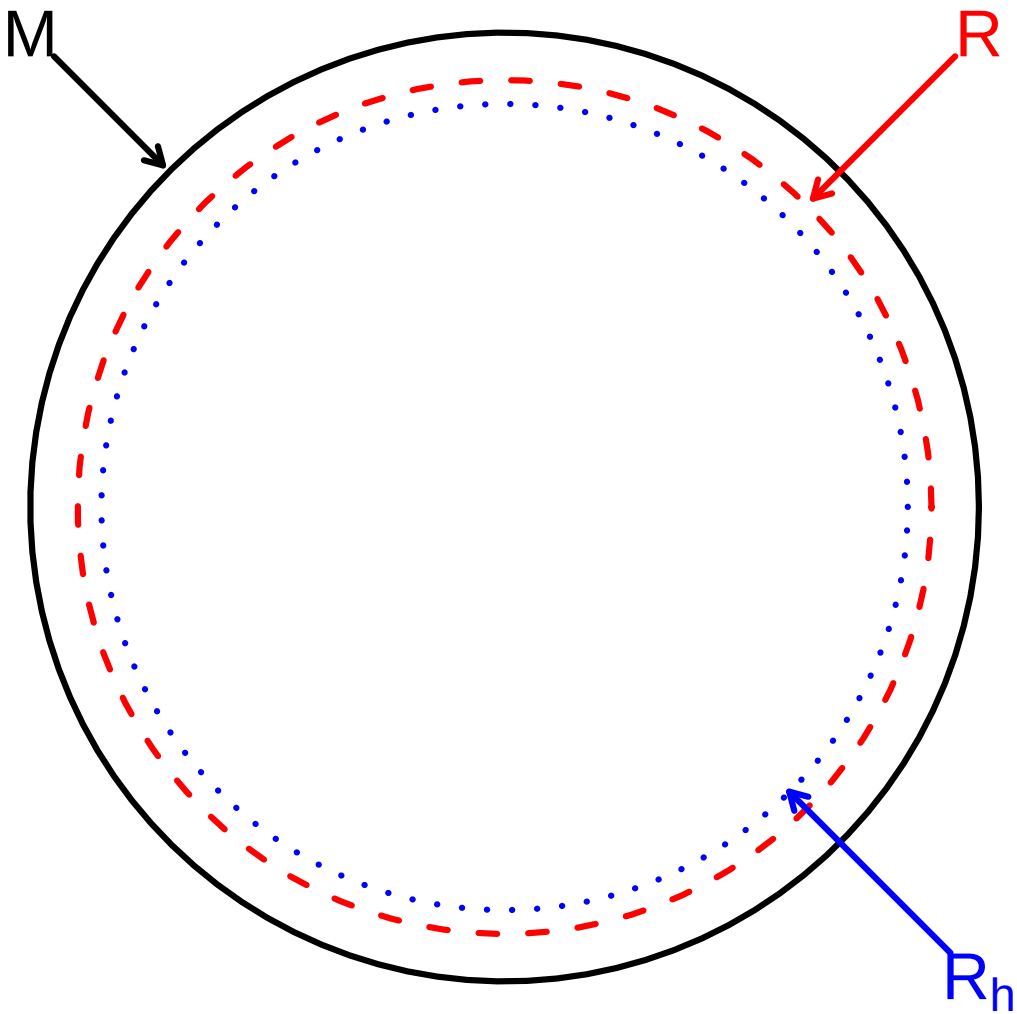
- q, g, and h be the density of Q and its gradient and Hessian,
- $\lambda_j(x)$ be the jth eigenvalue of $h(x)$ in *increasing* order,
- $V(x)$ to matrix whose columns are the eigenvectors of $h(x)$ for $\lambda_1(x), \ldots, \lambda_{D-d}(x)$.

Define the hyper-ridge set $R \equiv R(q)$ as follows:

$$x \in R(q) \text{ iff } \lambda_{D-d}(x) < 0 \text{ and } V(x)^T g(x) = 0.$$

If $\text{Haus}(M, R) = O(\sigma)$ and if R and M have a common topology, then R will be an effective surrogate.

Example Hyper-Ridge Set



Modified Mean-Shift Methods

Our hyper-ridge estimator uses a modification the mean-shift algorithm, which carries arbitrary points on trajectories towards (local) modes of a density.

Genovese, Perone-Pacifico, Verdinelli and Wasserman (2009) use the mean-shift trajectories to trace out ridges of the density and find filaments.

Ozertem and Erdogmus (2011) take this further, projecting each mean-shift point onto the space spanned by the smallest (most-negative) $D - d$ eigenvectors of Hessian$(\widehat{q})$.

The latter is called the subspace-constrained mean-shift algorithm (SCMS).

The Mean-Shift Algorithm

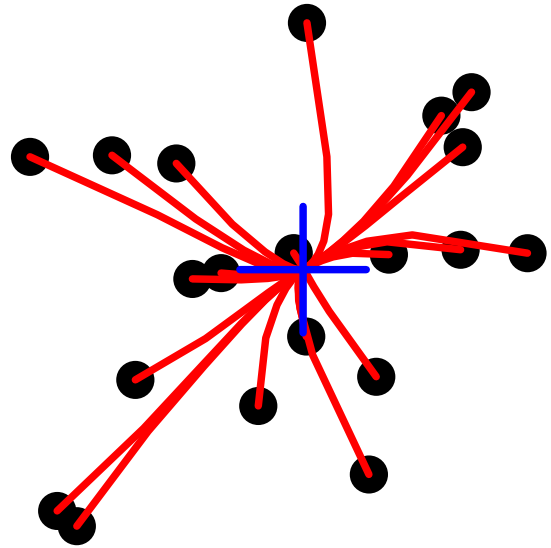
Finds the modes of a kernel density estimator $\widehat{q}$.

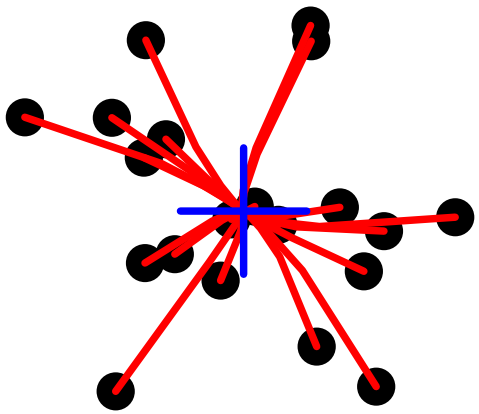
Input: Kernel density estimator $\widehat{q}_h$, tolerance $\tau \geq 0$

1. Choose initial mesh points $v_{1,0}, \ldots, v_{m,0}$
2. $t \leftarrow 0$
3. **repeat**
4. **for** $j = 1$ **to** m **do**
5. $$v_{j,t+1} \longleftarrow \frac{\sum_i Y_i \, K_h(\|v_{j,t} - Y_i\|)}{\sum_i K_h(\|v_{j,t} - Y_i\|)}$$
6. **end for**
7. $t \leftarrow t + 1$
8. **until** $\max_j |v_{j,t+1} - v_{j,t}| \leq \tau$
9. **return** $v_{1,t}, \ldots, v_{m,t}$

The $v_{j,t}$ converge to (local) modes as $t \to \infty$.

Mean Shift Paths





A Hyper-Ridge Set Estimator

Steps:

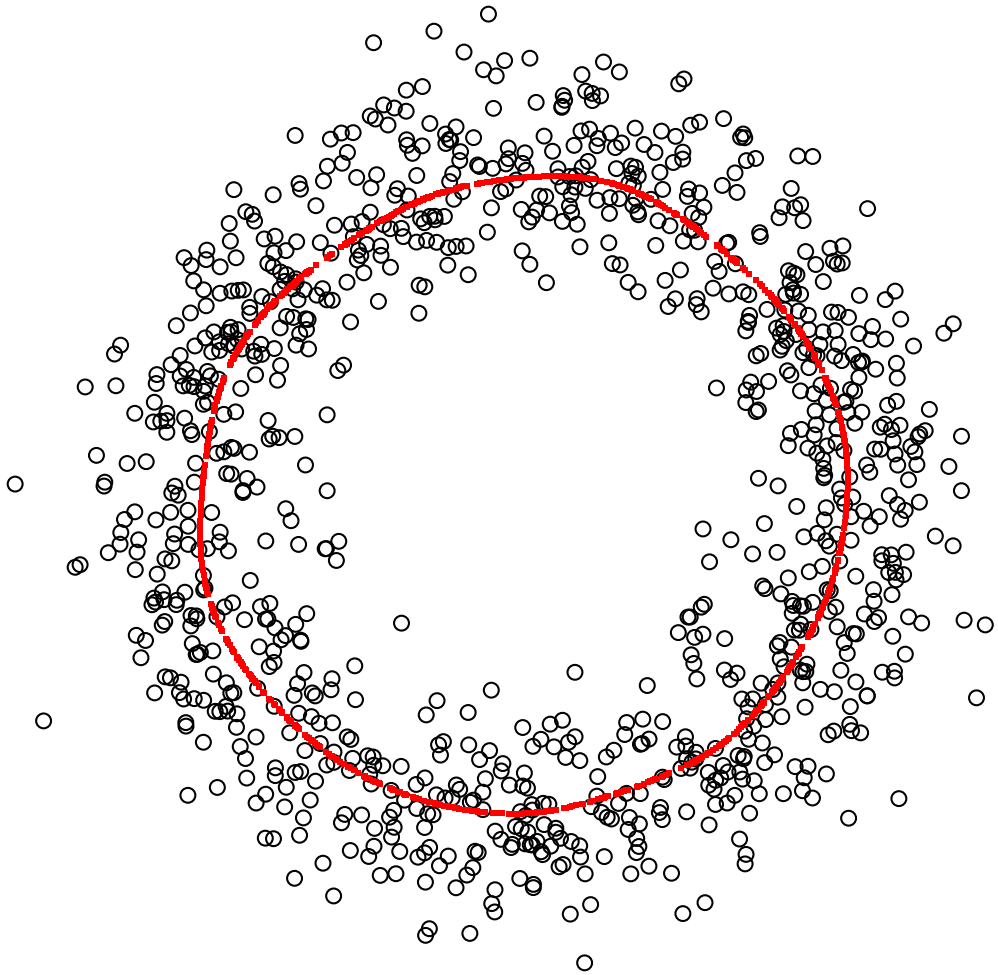
1. Estimation: estimate the density q, its gradient g, and its Hessian h.
2. Denoising: remove background clutter and low-probability regions, restricting attention to a set where q is not too small;
3. Mean-Shift: apply the SCMS algorithm within the restriction set.

We can show that: $H(R, \widehat{R}) = O_P\left(n^{-\frac{2}{4+D}}\right)$.

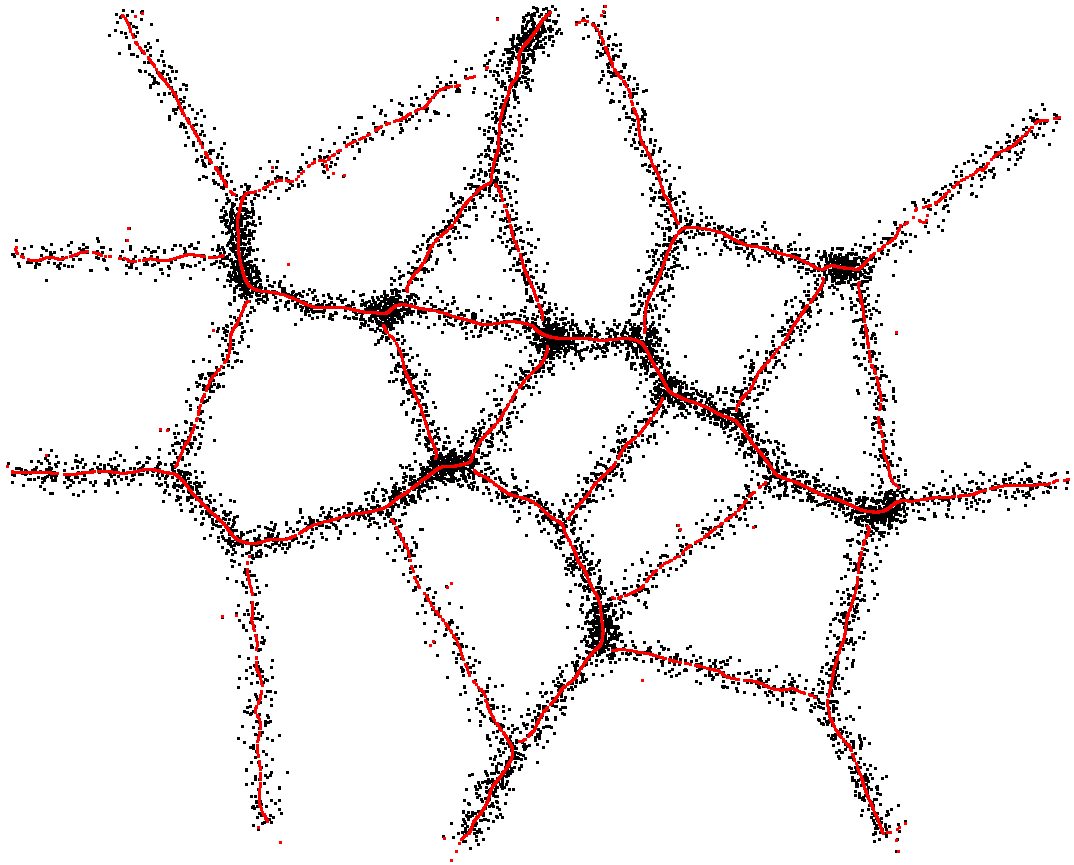
However, if we can live with bias, then we can set $h = O(\sigma)$ and then $H(R_h, \widehat{R}_h) = O_P\left(n^{-\frac{1}{2}}\right)$.

We are currently developing more of the theory. Here are two examples.

Example 1

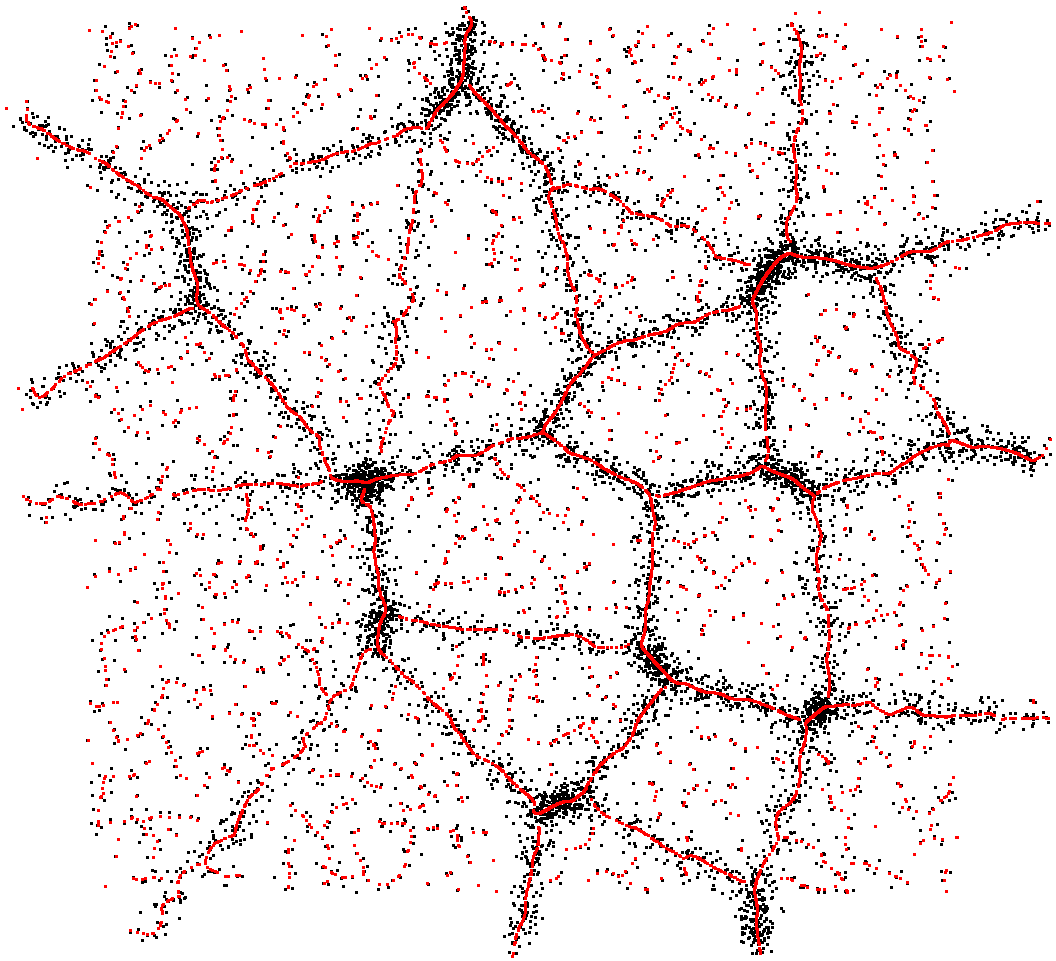


Example 2



Example 2

But we need to denoise first or else ...



Take-Home Points

1. Manifold complexes arise in many problems.

2. Manifold estimation is a special case; more generally, we want to find structure in data.

3. Minimax rates can be obtained for a variety of noise models.
   They do not depend on the dimension of the embedding space but are highly sensitive to the noise model.

4. Surrogates provide a useful (and computationally efficient) alternative even in very high dimensions.
   We accept some bias to capture some features accurately.