Estimating Manifolds

Rates, Methods, and Surrogates

Christopher R. Genovese

Department of Statistics
Carnegie Mellon University

19 Jun 2012
WNAR

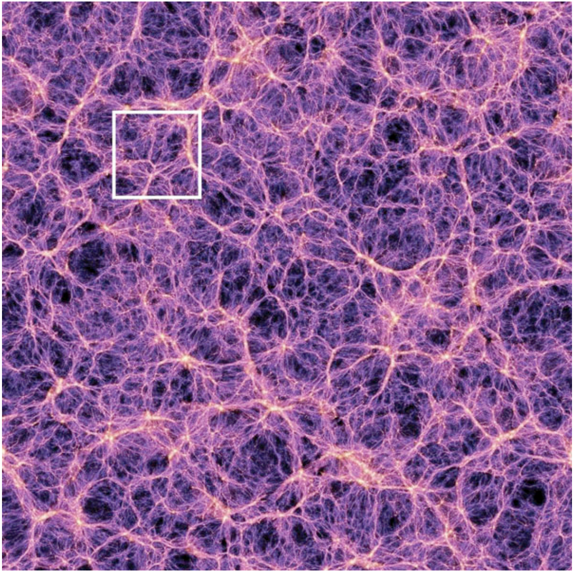
Collaborators

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:

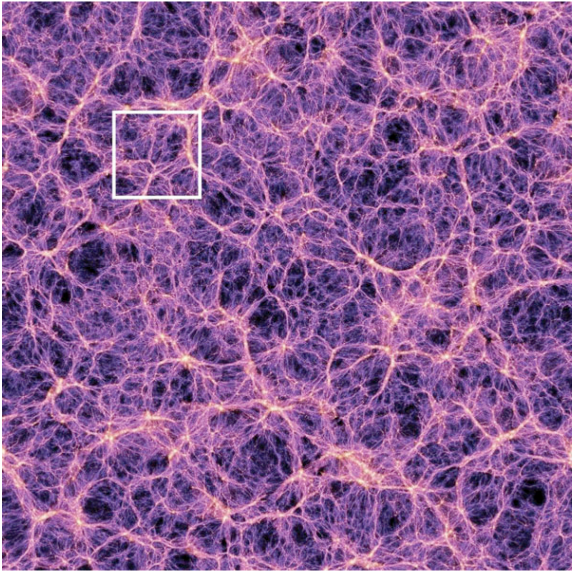
1. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2009. [*Ann. Stat.*, **37**]
2. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2010a. [arXiv:1003.5536 JASA]
3. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2010b. [arXiv:1007.0549 Annals]
4. Genovese, Perone-Pacifico, Verdinelli, Wasserman 2011.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

Motivating Example: the "Cosmic Web"



Gnedin (2005)

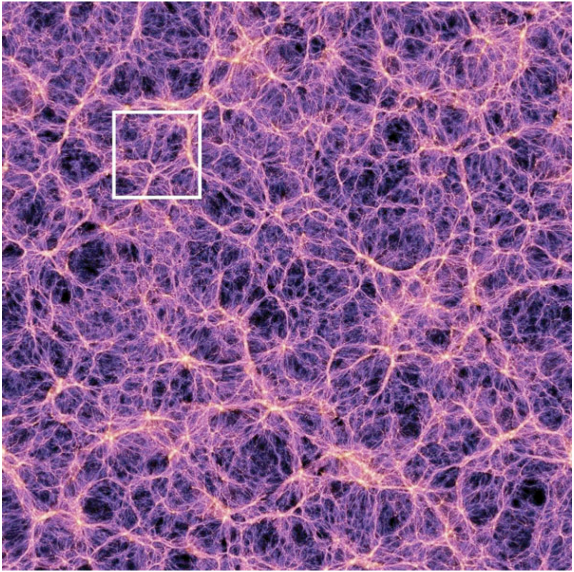
Motivating Example: the "Cosmic Web"



Matter is concentrated around lower dimensional features:

Gnedin (2005)

Motivating Example: the "Cosmic Web"

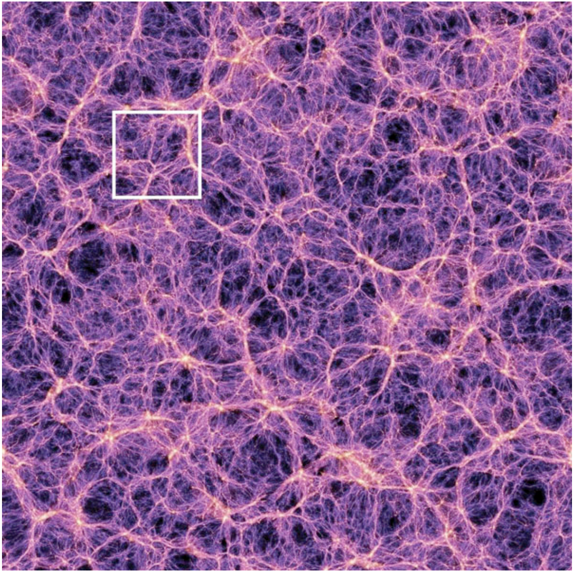


Matter is concentrated around lower dimensional features:

0-dimensional clusters

Gnedin (2005)

Motivating Example: the “Cosmic Web”

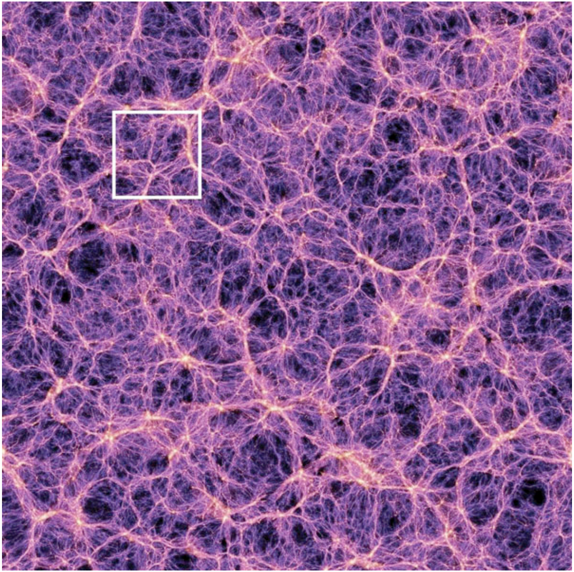


Matter is concentrated around lower dimensional features:

0-dimensional clusters
1-dimensional filaments

Gnedin (2005)

Motivating Example: the "Cosmic Web"

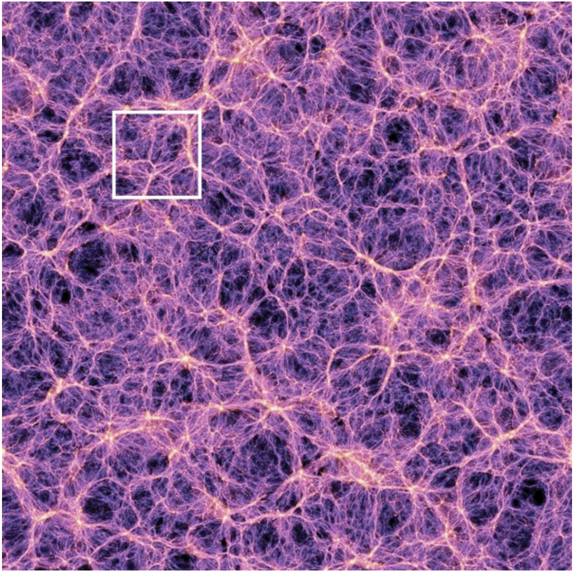


Gnedin (2005)

Matter is concentrated around lower dimensional features:

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

Motivating Example: the "Cosmic Web"



Matter is concentrated around lower dimensional features:

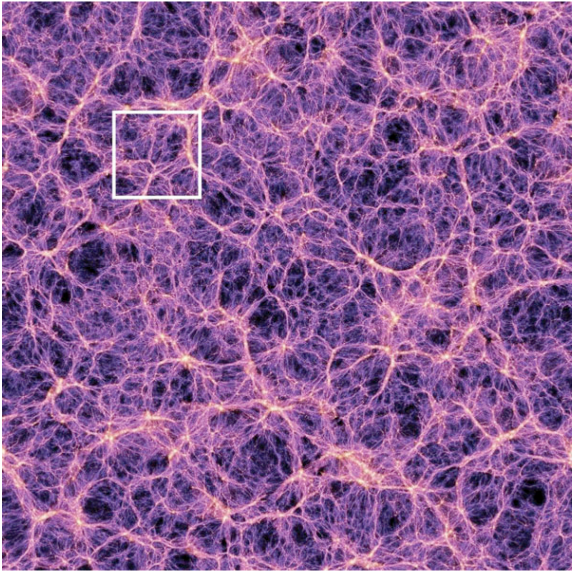
0-dimensional clusters
1-dimensional filaments
2-dimensional sheets

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.

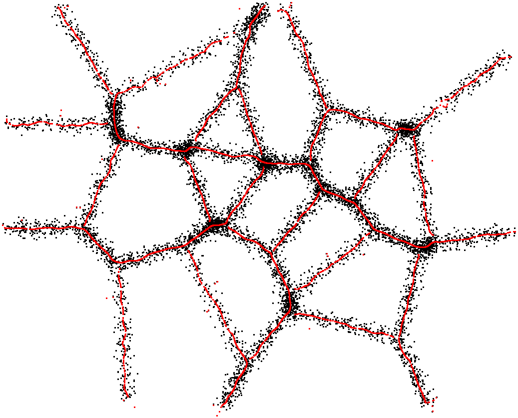
A manifold is a space that looks locally like a Euclidean space of some dimension (called the dimension of the manifold).

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

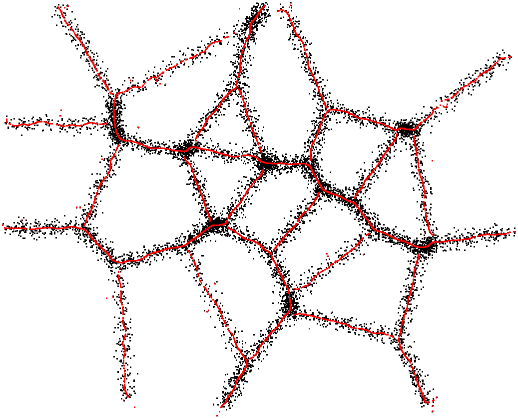
To allow for intersections and other complexities, consider a *union* of manifolds embedded in $\mathbb{R}^D$ with maximal dimensions $d < D$.

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

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 $Z_i \sim \Phi$ (e.g., spherical Normal).

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:

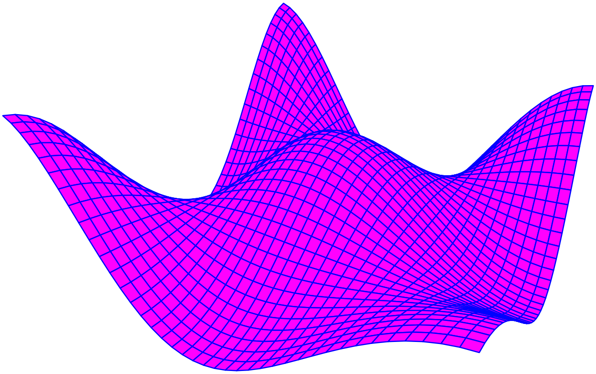
1. noiseless: $Y_i = X_i$.
2. clutter: $Y_i = X_i$ with probability π, otherwise $Y_i \sim$ Uniform.
3. perpendicular: $Y_i = X_i + Z_i$ where Z_i is normal to M.
   (See also Niyogi, Smale, Weinberger 2008.)
4. additive: $Y_i = X_i + Z_i$ and $Z_i \sim \Phi$ (e.g., spherical Normal).

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

The noise model strongly affects the difficulty of this problem.

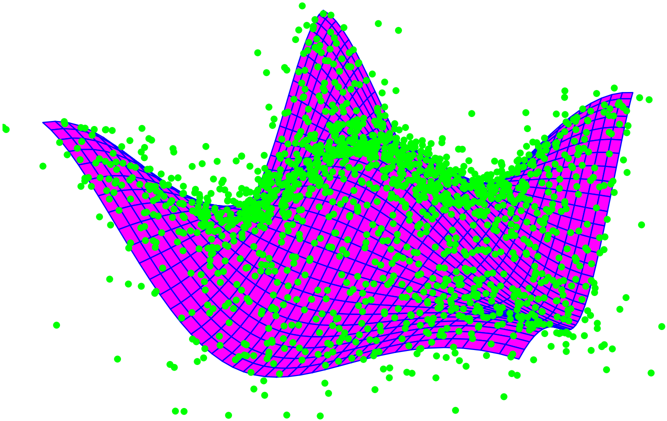
A Synthetic Example

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



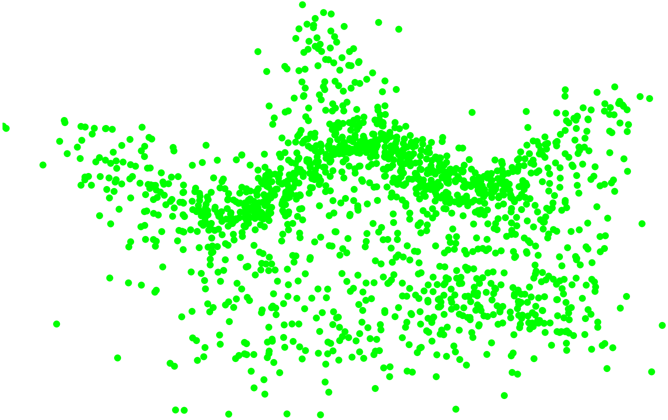
A Synthetic Example

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 (sup) r such that $d(x, M) \leq r$ implies that x has a unique projection onto M.*

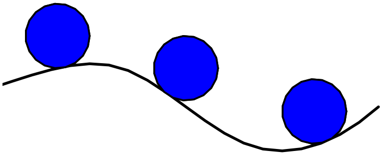
This is also called the thickness or condition number of the manifold; see Niyoki, Smale, and Weinberger (2009).

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

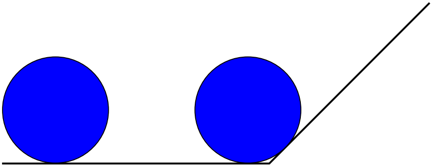
1. *Curvature.* A ball or radius $r \leq \kappa$ can roll freely and smoothly over M, but a ball or radius $r > \kappa$ cannot.
2. *Separation.* M is at least 2κ from self-intersecting.

Reach in One Dimension

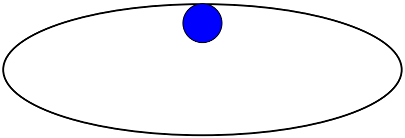
circles have radius r



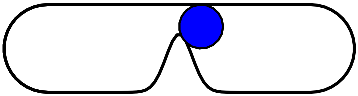
$\kappa > r$



$\kappa < r$



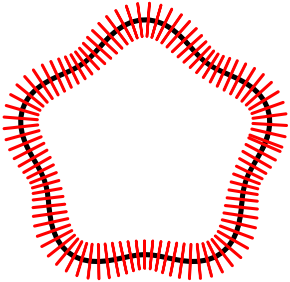
$\kappa > 2r$

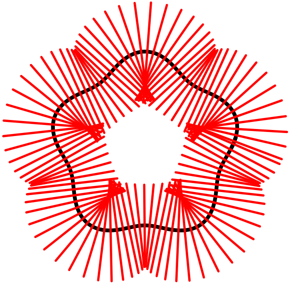


$\kappa < 2r$

Reach Visualized

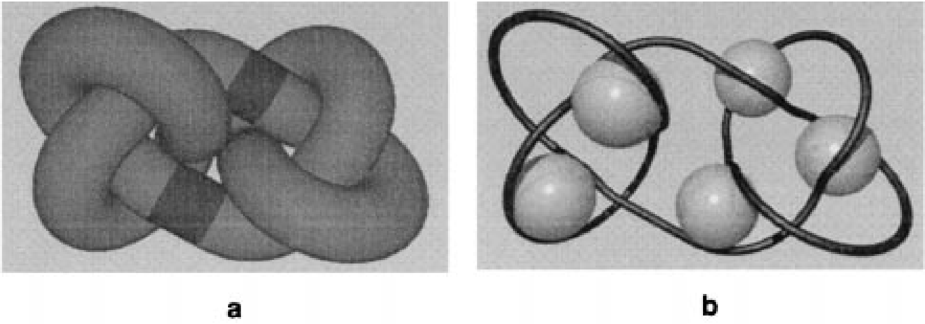
Normals of size $<$ reach(M) do not cross.





Reach Visualized

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



from Gonzalez and Maddocks (1999)

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

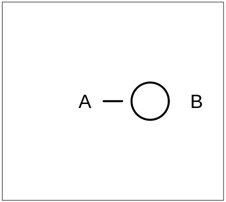
Hausdorff Distance

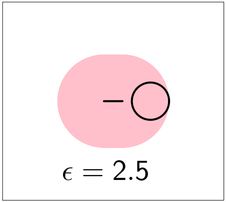
Given two subsets of $\mathbb{R}^D$, A and B:

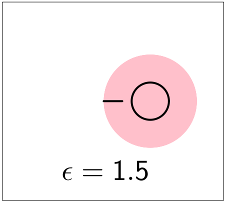
$$\text{Haus}(A, B) = \inf\{\epsilon : \ A \subset B \oplus \epsilon \text{ and } B \subset A \oplus \epsilon\}$$

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

Example:







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

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

Methods and Surrogates

Minimax Rates under Various Noise Models

$$\inf_{\widehat{M}_n} \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q \, \mathrm{Haus}(\widehat{M}_n, M) \asymp C\psi_n \quad \text{(up to log terms)}$$

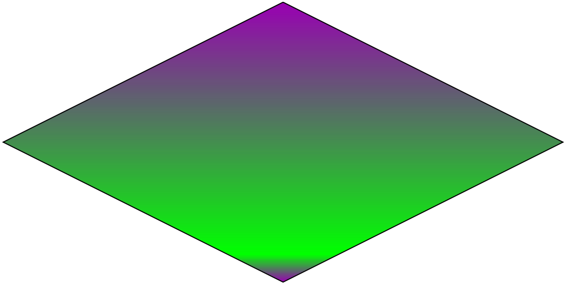
Noise Model	ψ_n
Clutter/Noiseless	$(\pi n)^{-\frac{2}{d}}$
Perpendicular Compact	$n^{-\frac{2}{2+d}}$
Additive Compact/Polynomial	*in progress*
Additive sub-Gaussian	$(\log n)^{-1}$

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.

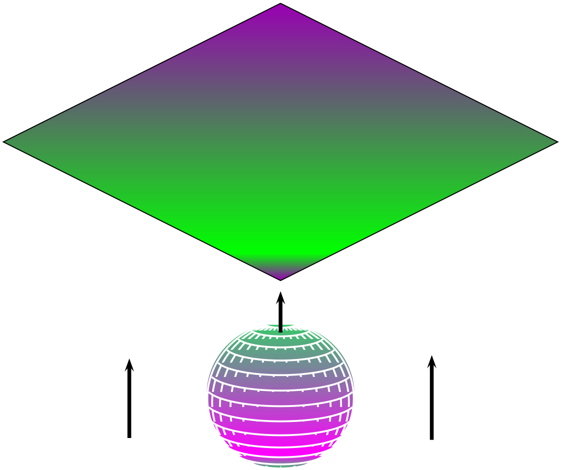
Proof Sketch: Lower Bound, Perpendicular Noise

Start with M_0 . . .



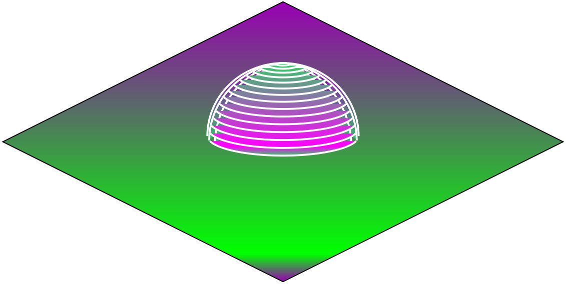
Proof Sketch: Lower Bound, Perpendicular Noise

Push up κ-ball,



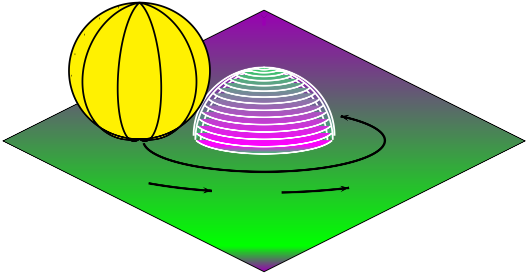
Proof Sketch: Lower Bound, Perpendicular Noise

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



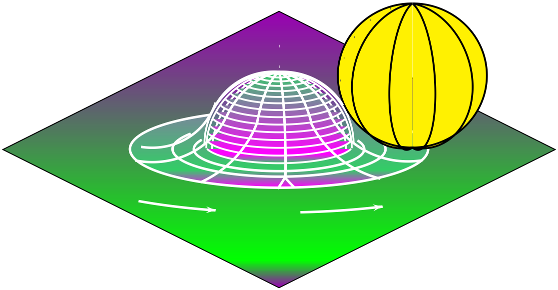
Proof Sketch: Lower Bound, Perpendicular Noise

But reach still 0, so smooth the corners.



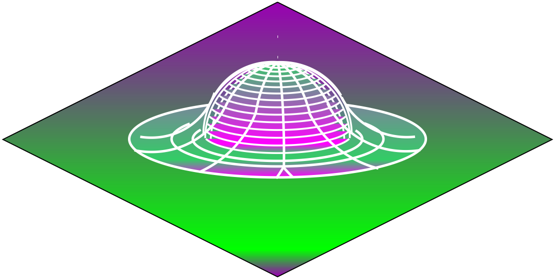
Proof Sketch: Lower Bound, Perpendicular Noise

Smooth the corners . . .



Proof Sketch: Lower Bound, Perpendicular Noise

Flying Saucer M_1



Road Map

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

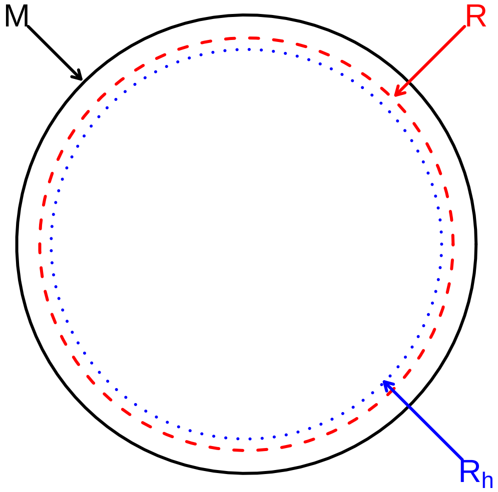
- q, g, and h be the density of Q and its gradient and Hessian,
- $\lambda_j(x)$ be the jth eigenvalue of $h(x)$ in *increasing* order,
- $V(x)$ to matrix whose columns are the eigenvectors of $h(x)$ for $\lambda_1(x), \ldots, \lambda_{D-d}(x)$.

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

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

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).

A Hyper-Ridge Set Estimator

Steps:

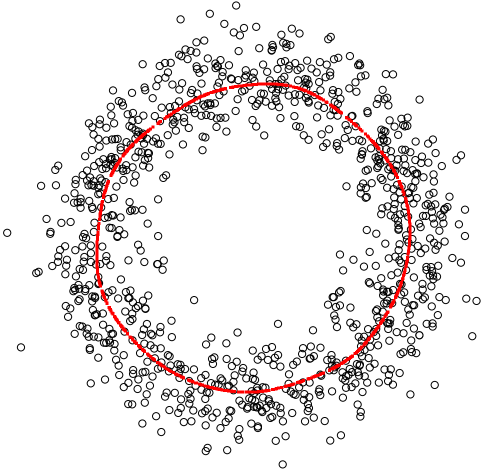
1. Estimation: estimate the density q, its gradient g, and its Hessian h.
2. Denoising: remove background clutter and low-probability regions, restricting attention to a set where q is not too small;
3. Mean-Shift: apply the SCMS algorithm within the restriction set.

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

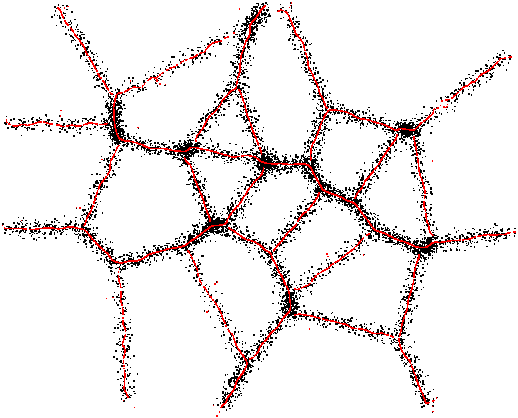
However, if we can live with bias, then we can set $h = O(\sigma)$ and then $H(R_h, \widehat{R}_h) = O_P\left(n^{-\frac{1}{2}}\right)$.

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

Example 1

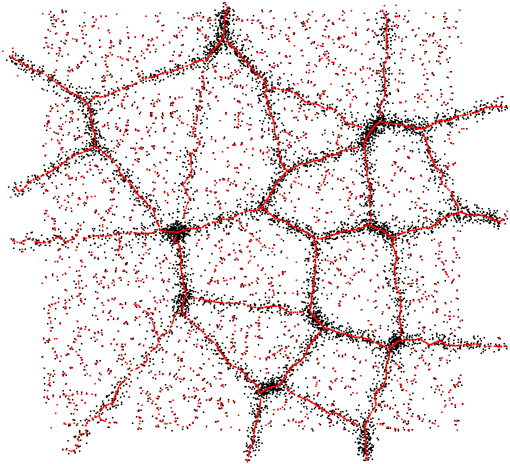


Example 2



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.