Estimating Manifolds

Supplementary Materials

Christopher R. Genovese

Department of Statistics
Carnegie Mellon University

19 Jun 2012
WNAR

Proof Sketch: Lower Bounds

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”.

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,

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

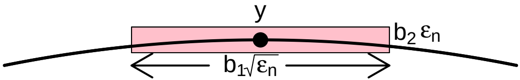
(The $\asymp^*$ means I am hiding log factors.)

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

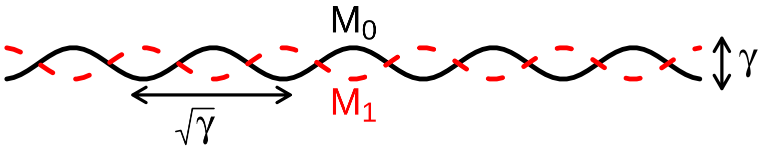
For technical reasons, we allow the manifolds to be noncompact. Define a truncated loss function,

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

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.