Manifold-Adaptive Dimension Estimation

or some organic thoughts regarding the nature of data and the way we may deal with it!

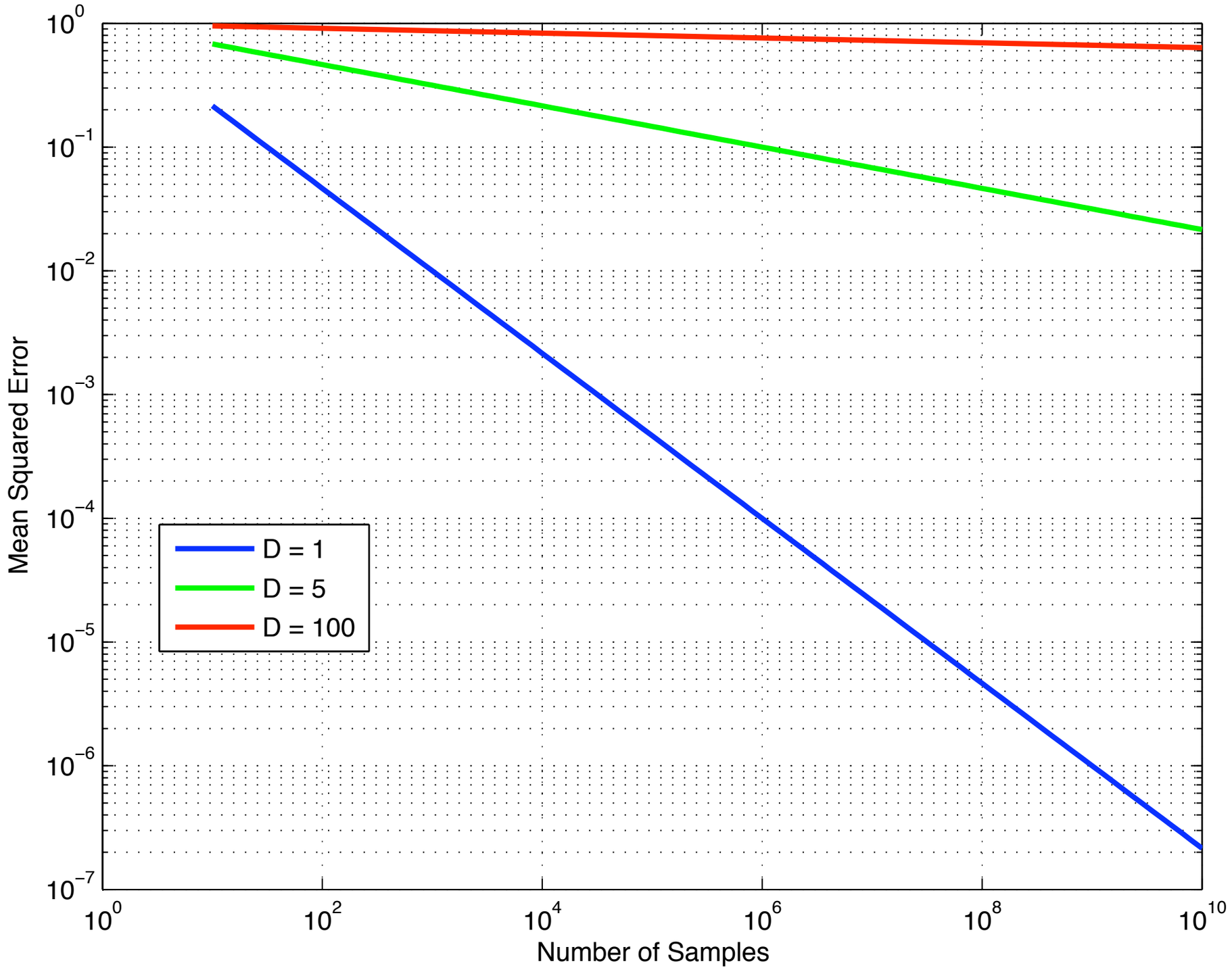
Amir massoud Farahmand[1], Csaba Szepesvári[1], Jean-Yves Audibert[2]

(1) Department of Computing Science, University of Alberta, Canada
(2) CERTIS, Ecole Nationale des Ponts, France

High-Dimensional Data Everywhere

- Vision
- Sensor Fusion
- Feature Expansion
- Kernel
- ...

Curse of Dimensionality

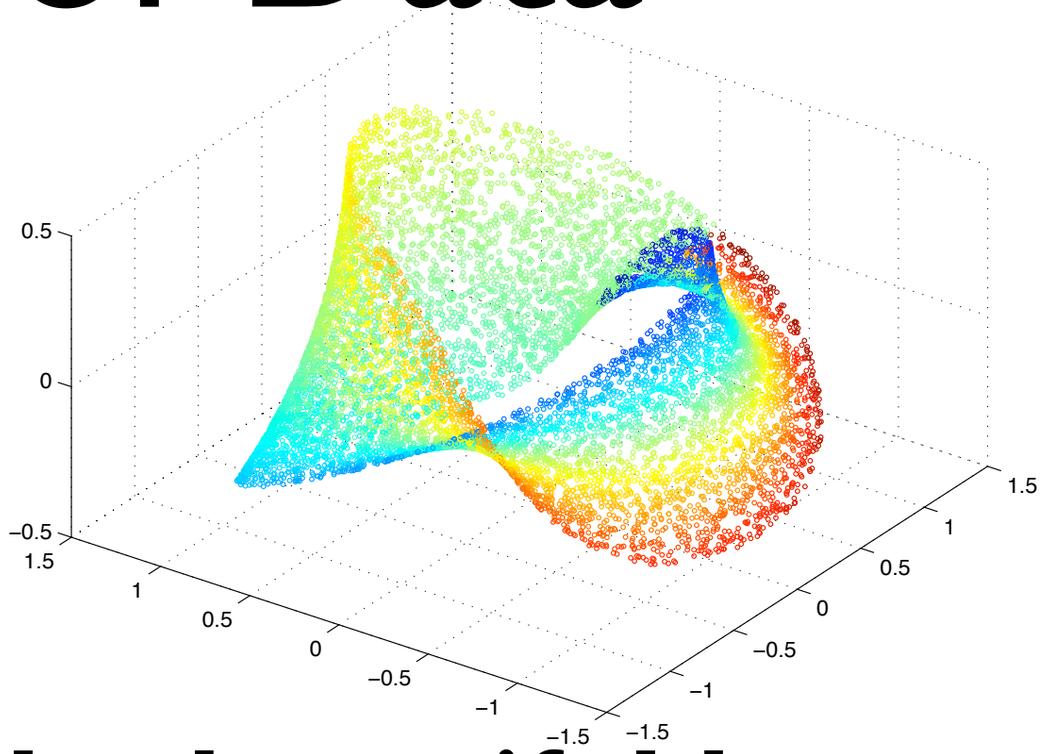


Practical Implications

- Thou shall reduce the dimension of the data before working with it!
- Thou shall not add features unnecessarily!
- Thou shall not accept projects with high-dimensional data!
- ... !

Regularities of Data

- Smoothness
- Sparsity
- Low Noise at Boundary



✓ **Lower dimensional submanifold**

- LLE, IsoMap, Graph Laplacian, HessianMap, Semi-supervised Learning, RL

Goal

- Manifold-adaptive machine learning methods
- Convergence rate independent of the dimension of the input space

Many open questions!

Here:
dimension estimation
(:

Why?

- Needed in various learning methods
- Not known a priori

New?

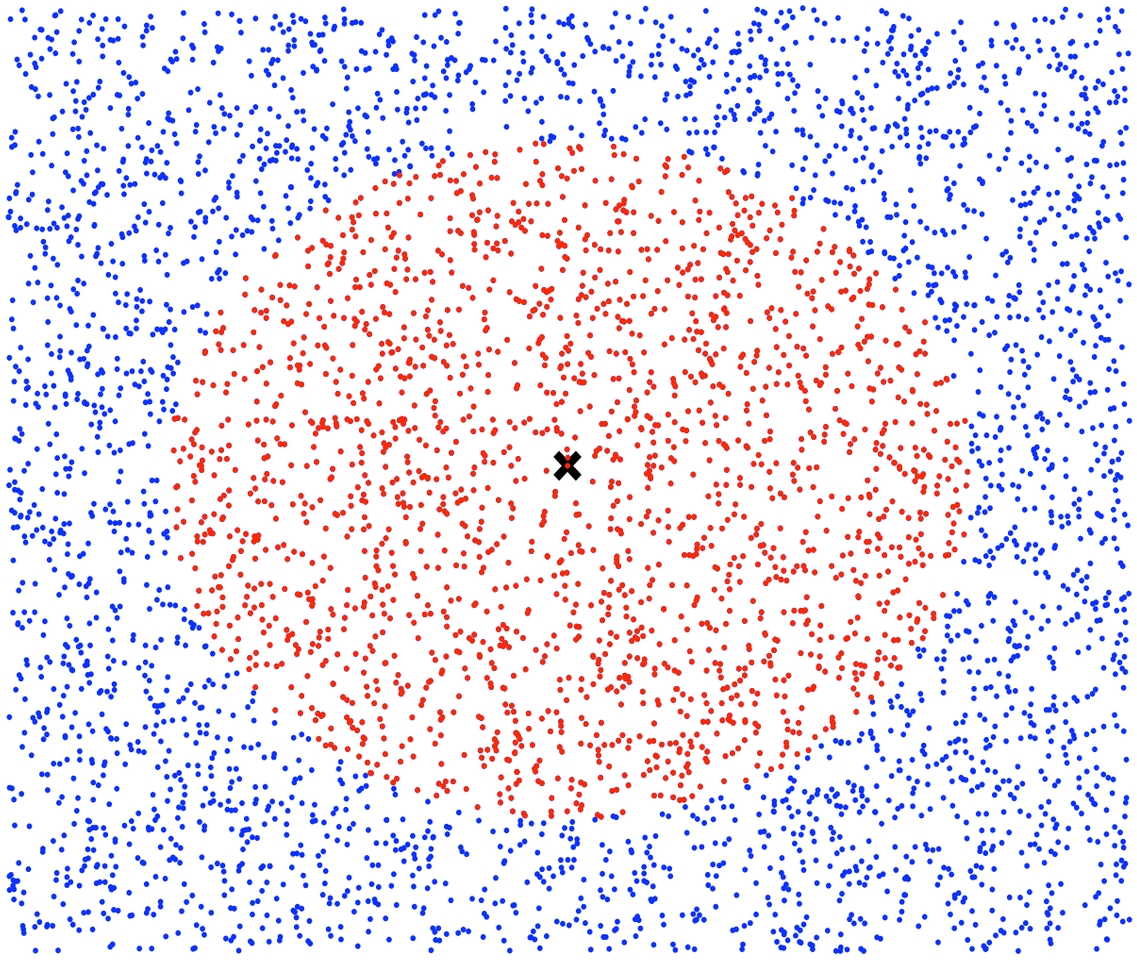
- Many existing methods [Pettis *et al.* (1979), Kegl (2002), Levina & Bickel (2005), Hein & Audibert (2005)]
- No rigorous analysis
  - Asymptotic result [Levina & Bickel (2005)]

Our Contribution

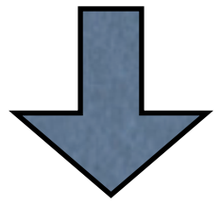
- New algorithm
  - K-NN
- Manifold-adaptive convergence rate

General Idea

$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$



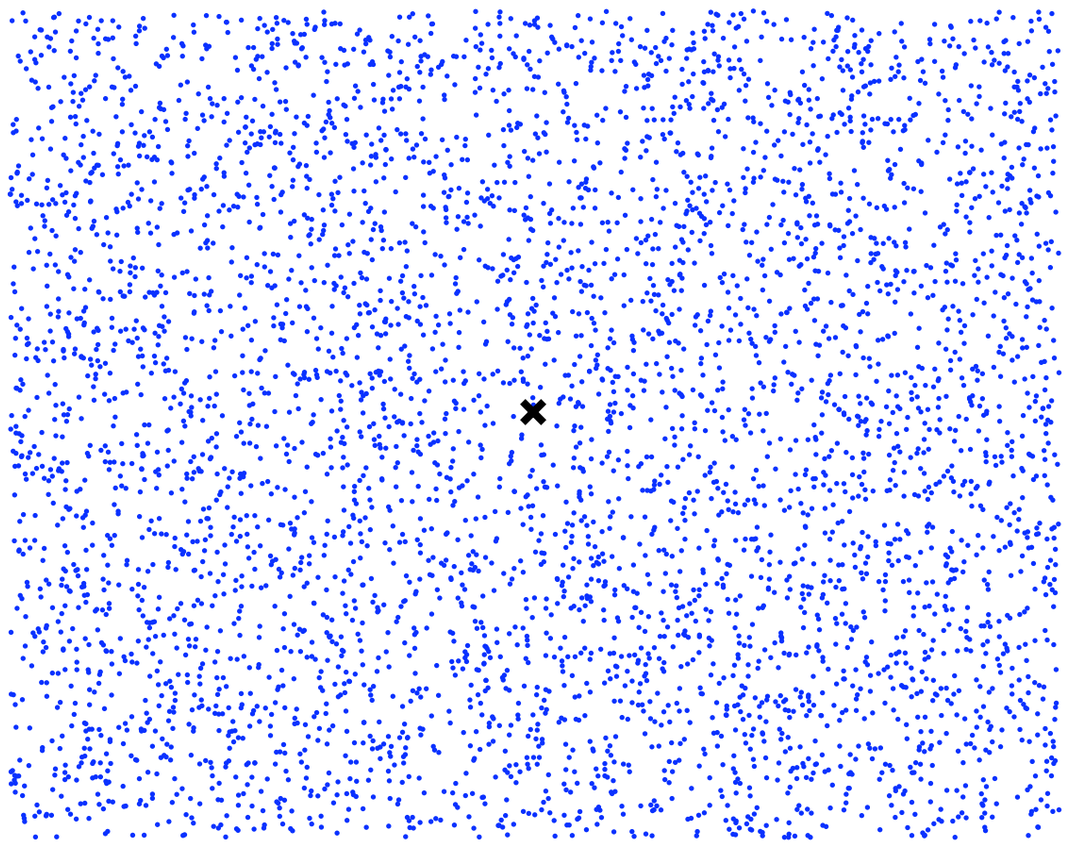
$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$



$$\ln\left(P(X_i \in B(x,r))\right) = \ln(\eta(x,r)) + d\ln(r)$$

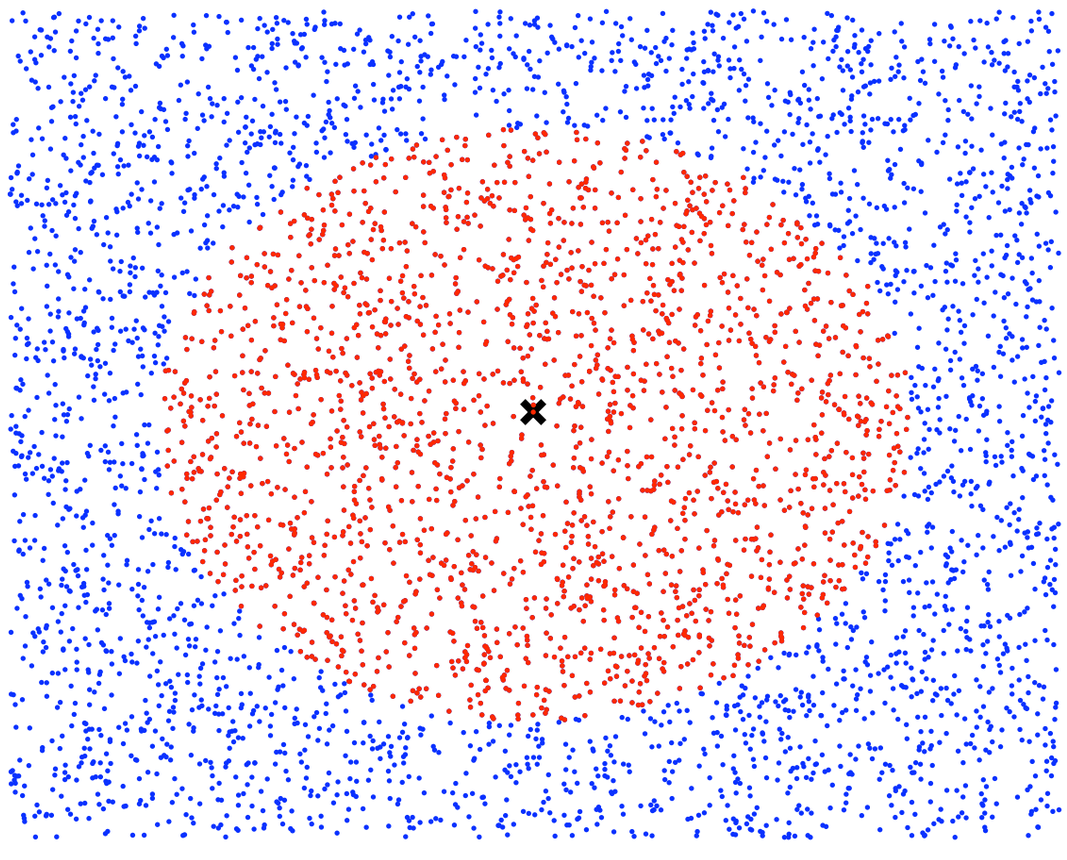
$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$

$$\ln\left(P(X_i \in B(x,r))\right) = \ln(\eta(x,r)) + d\ln(r)$$



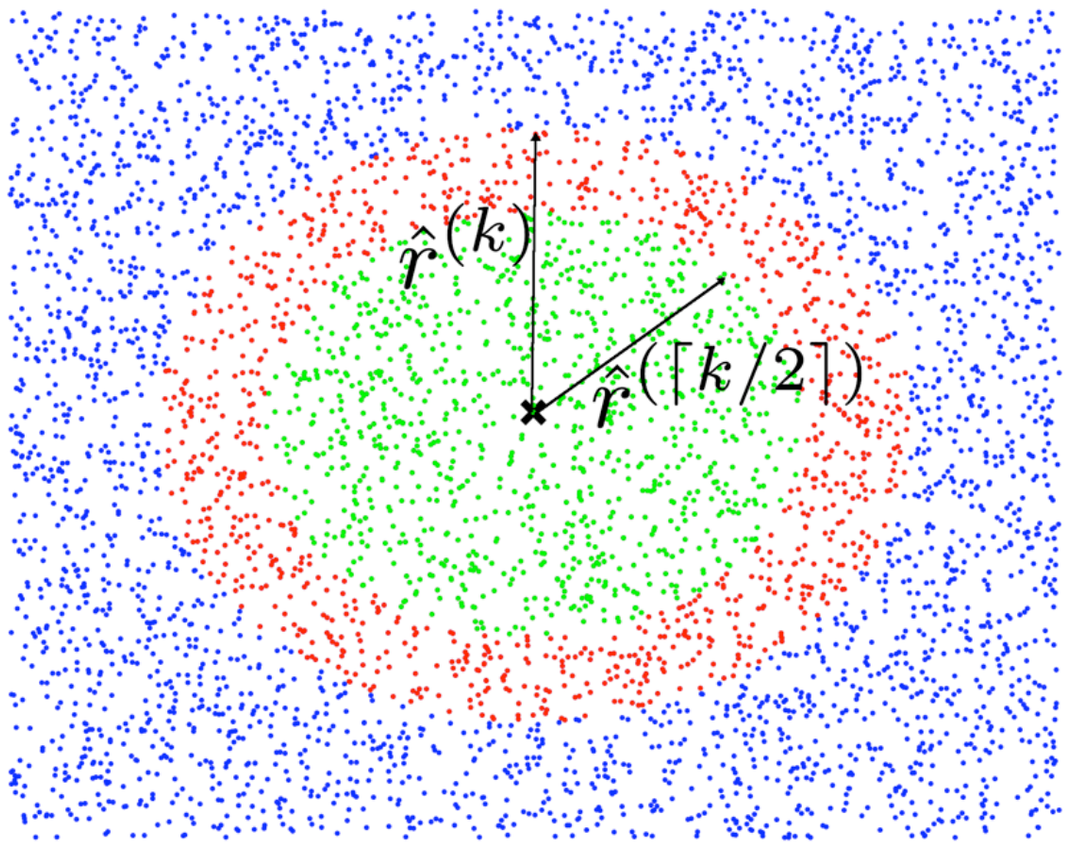
$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$

$$\ln\left(P(X_i \in B(x,r))\right) = \ln(\eta(x,r)) + d \ln(r)$$



$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$

$$\ln\left(P(X_i \in B(x,r))\right) = \ln(\eta(x,r)) + d\ln(r)$$

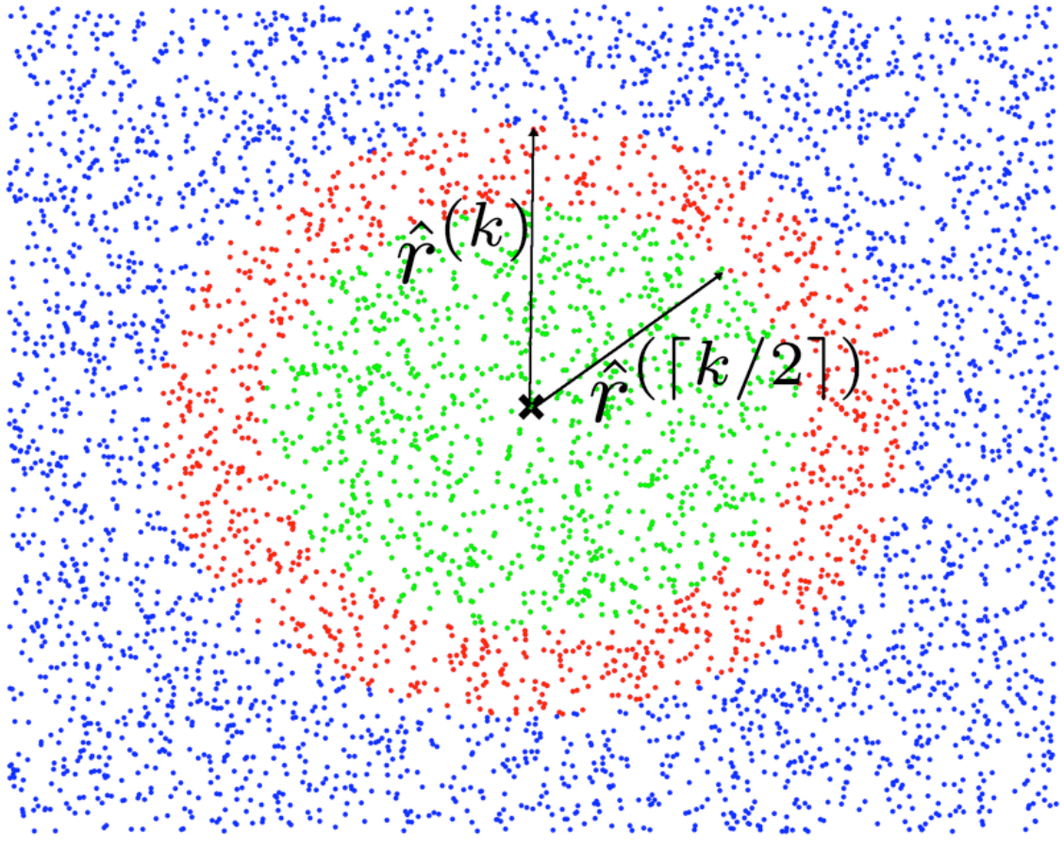


$$P(X_i \in B(x,r)) = \eta(x,r) r^d$$

$$\ln\left(P(X_i \in B(x,r))\right) = \ln(\eta(x,r)) + d\ln(r)$$

$$\ln(k/n) \approx \ln(\eta_0) + d\ln(\hat{r}^k(x))$$

$$\ln(k/(2n)) \approx \ln(\eta_0) + d\ln(\hat{r}^{\lceil k/2 \rceil}(x))$$



$$\hat{d}(x) = \frac{\ln 2}{\ln(\hat{r}^k(x)/\hat{r}^{\lceil k/2 \rceil}(x))}$$

Finite Sample Convergence Rate

$$\hat{d}(X_1) = \frac{\ln 2}{\ln(\hat{r}^k(X_1)/\hat{r}^{\lceil k/2 \rceil}(X_1))}$$

Theorem: Under some regularity assumptions on η, provided that $\frac{n}{k} > \Omega(2^d)$, with probability at least $1 - \delta$

$$|\hat{d}(X_1) - d| \leq O\left(d\left[B\left(\frac{k}{n}\right)^{\frac{1}{d}} + \sqrt{\frac{\ln(4/\delta)}{k}}\right]\right)$$

Issues

$$\hat{d}(X_1) = \frac{\ln 2}{\ln(\hat{r}^k(X_1)/\hat{r}^{\lceil k/2 \rceil}(X_1))}$$

High variance of $d(X_1)$

Inefficient use of data

Aggregation

- Averaging
- Voting

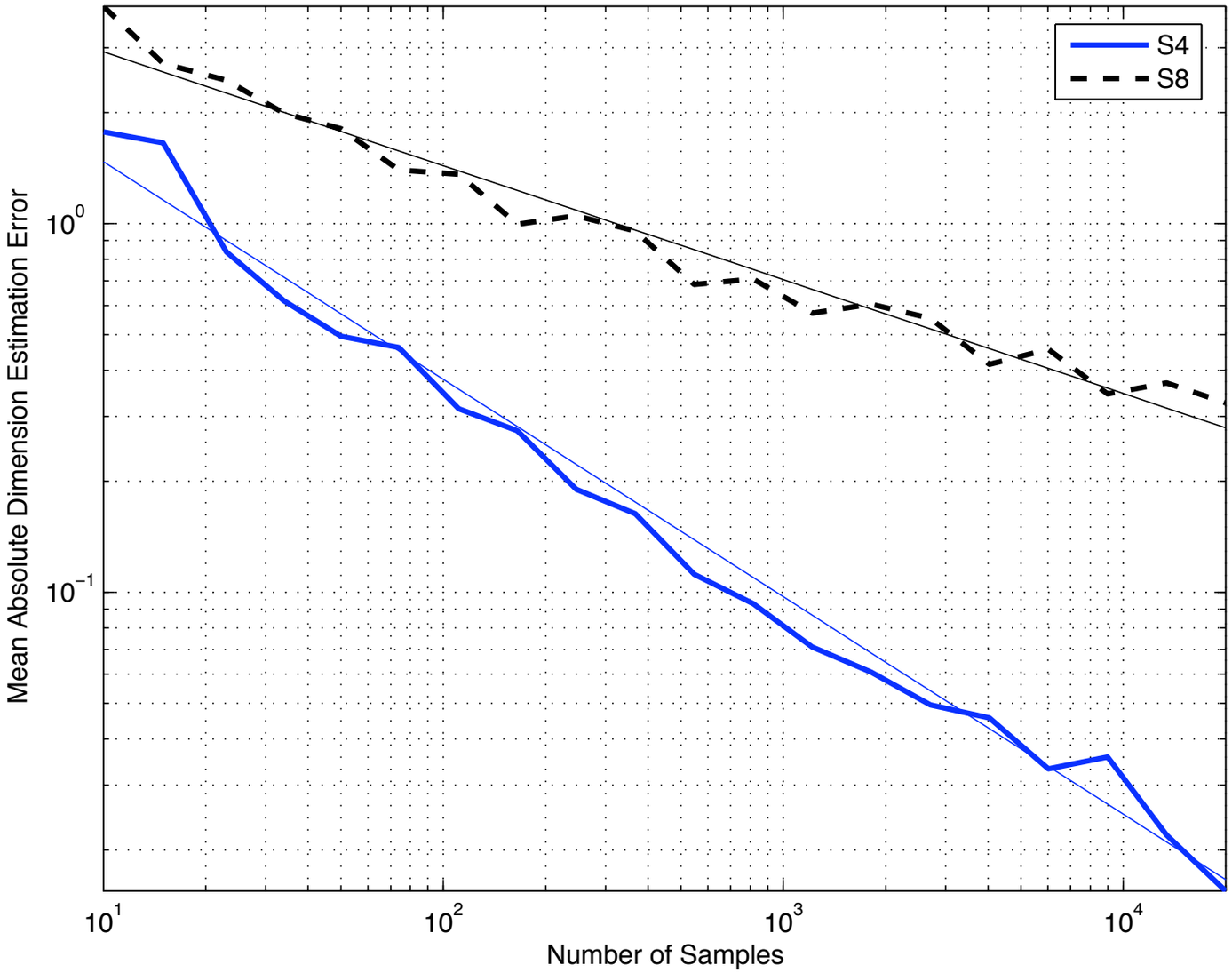
$$\hat{d}_{avg} = \frac{1}{n} \sum_{i=1}^{n} \min(\hat{d}(X_i), D)$$

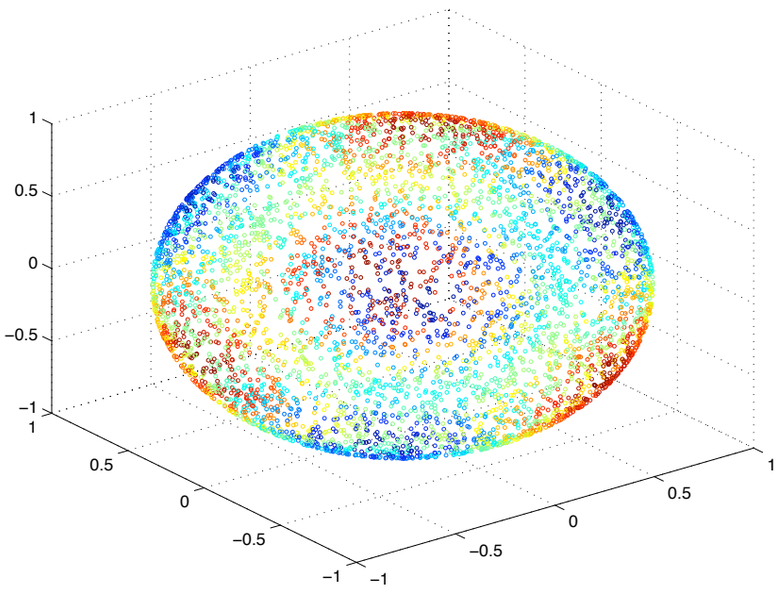
Theorem:

$$\mathbb{P}\left(\hat{d}_{\text{vote}} \neq d\right) \leq e^{-\frac{c' n}{(c^d k)^2}},$$

$$\mathbb{P}\left(\hat{d}_{\text{avg}} \neq d\right) \leq e^{-\frac{c'' n}{(D c^d k)^2}}$$

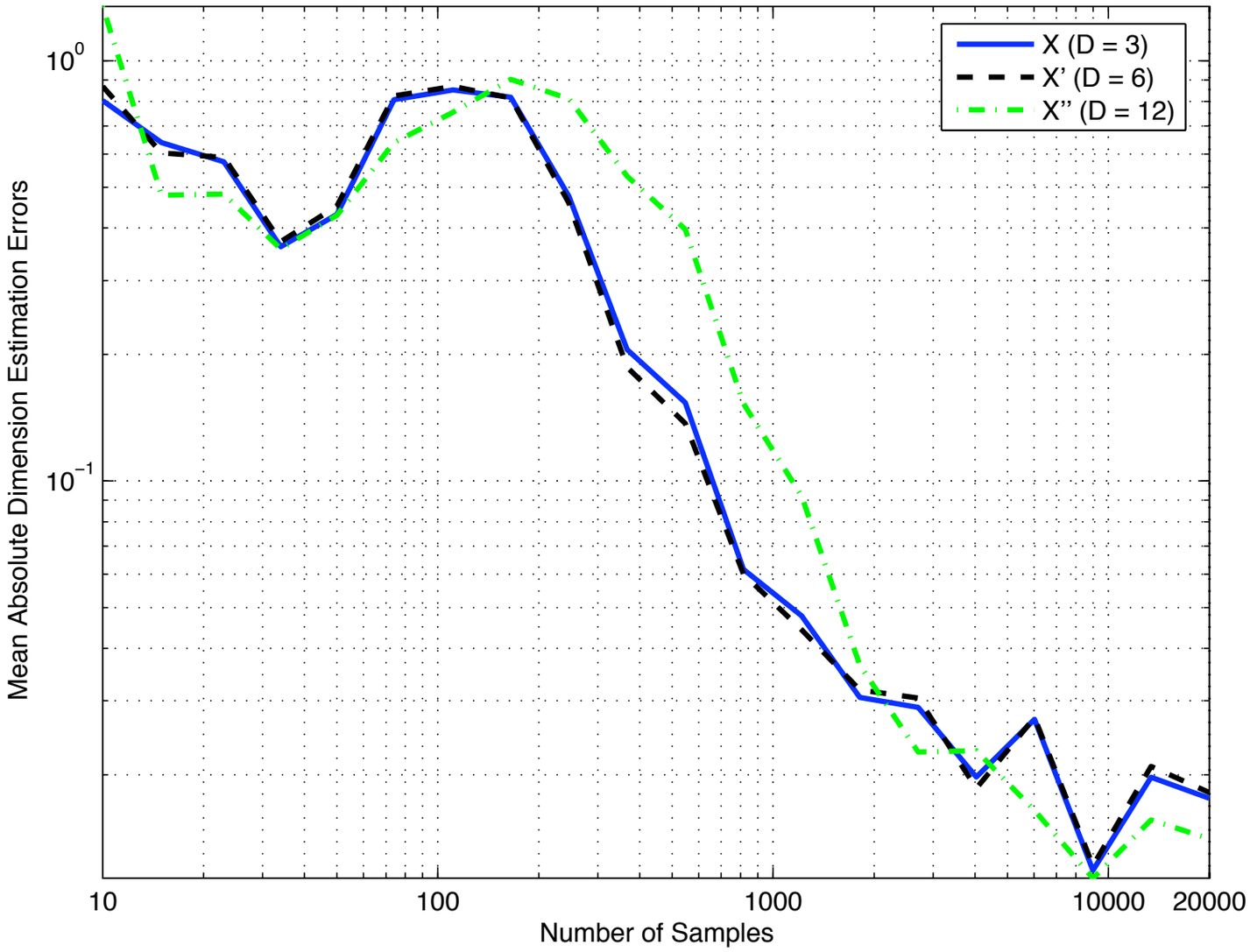
Experiments
Different Manifold Dimension

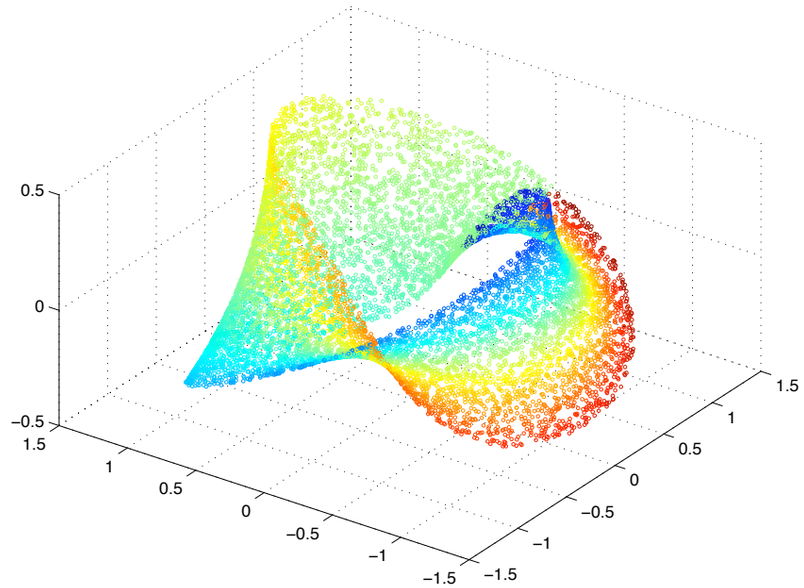




Experiments

Different Embedding Space Dimension





Experiments

Data set	n=50	n=100	n=500	n=1000	n=5000
S^1	98 (99)	100 (100)	100 (100)	100 (100)	100 (100)
S^3	75 (19)	95 (20)	100 (15)	100 (19)	100 (62)
S^5	33 (5)	50 (10)	100 (9)	98 (2)	100 (0)
S^7	18 (2)	17 (3)	57 (1)	54 (1)	100 (0)
Sinusoid	92 (98)	100 (100)	100 (100)	100 (100)	100 (100)
10-Möbius	69 (47)	13 (74)	100 (98)	100 (99)	100 (100)
Swiss roll	62 (71)	49 (91)	88 (96)	100 (100	100 (100)

Conclusions and Future Work

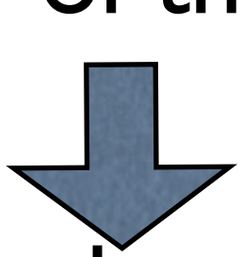
- Manifold-adaptive convergence rate
- Other ML methods?
- K-NN regression can!
- Penalized least squares in the works

TnX!

Curse of Dimensionality

High-Dimensional Data

Increase the complexity of the function space



Higher variance with the same number of samples

More samples for the same precision

Lower Bound

Assume that m_n is a regression function that estimate random variable Y based on X and $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$, and $m(X) = E[Y|X]$. What is the best possible performance of m_n in L_2 sense, i.e. $E\{\|m_n(X) - m(X)\|^2\}$?

For the class of $D^{(p,C)}$ of (X, Y) distributions, when $X \in \mathbb{R}^D$, we have the the following behavior:

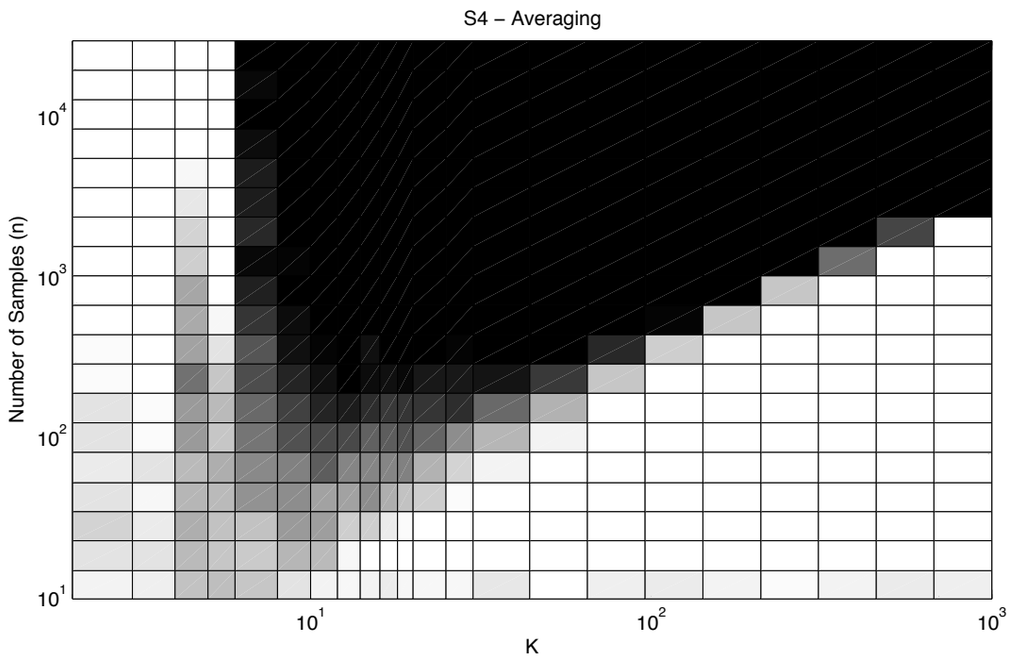
$$E\{\|m_n(X) - m(X)\|^2\} > O\left(n^{-\frac{2p}{2p+D}}\right)$$

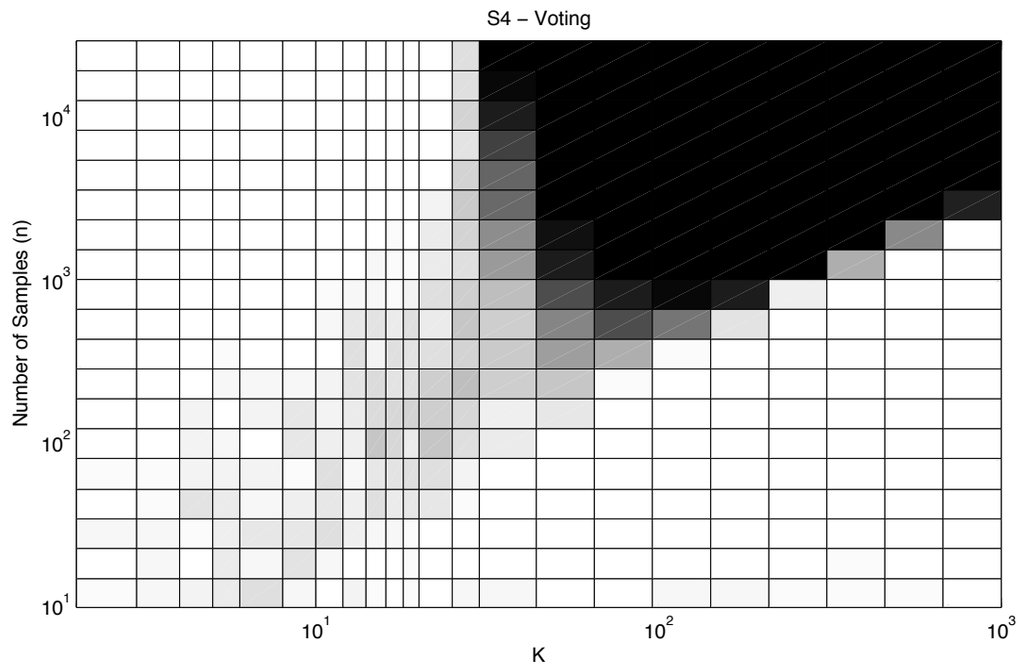
Two sources of error:

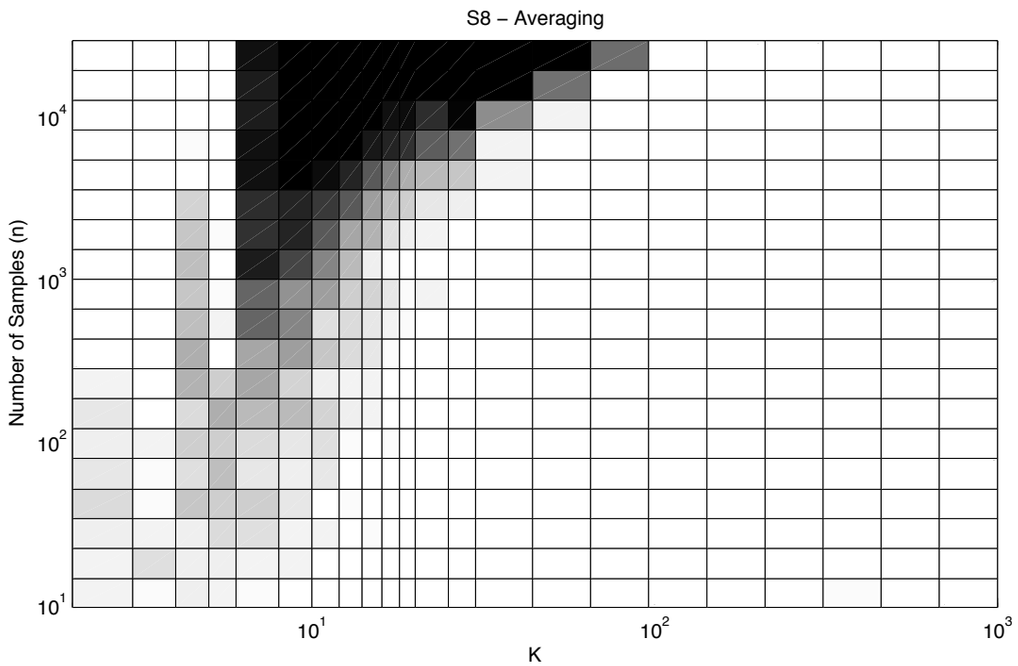
- *Approximation Error*: assuming fixed $\eta(x, r)$
- *Estimation Error*: estimating $P(X \in B(x, r))$ with the empirical estimate k/n.

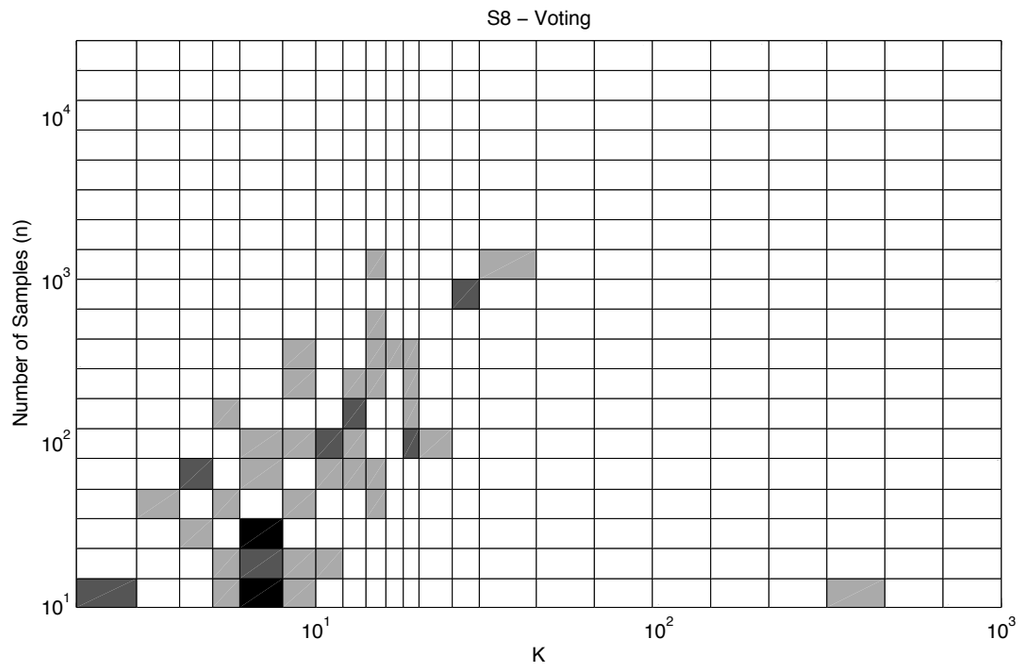
Both of them can be controlled by changing the size of neighborhood r (which is related to k/n).

Effect of K and n

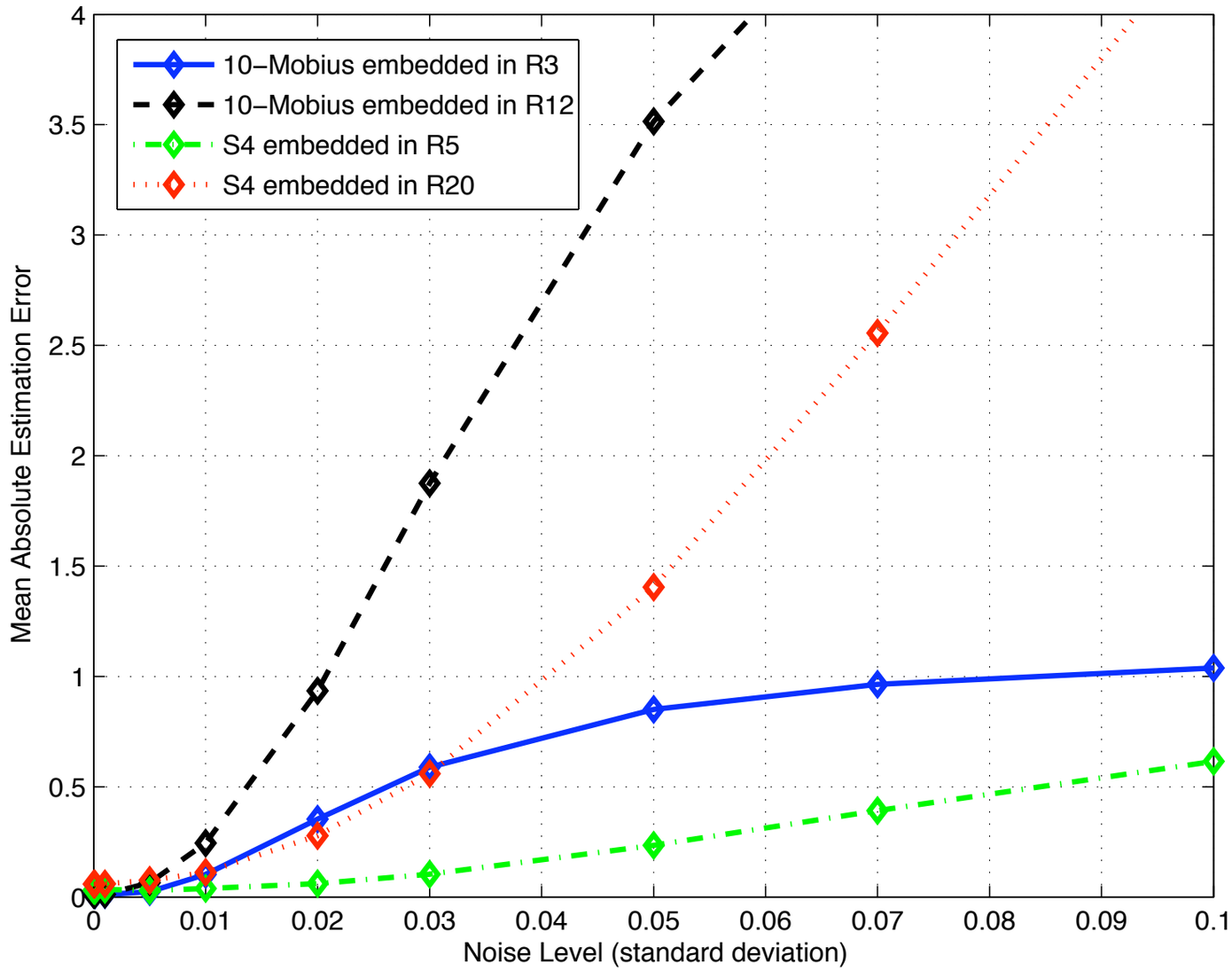




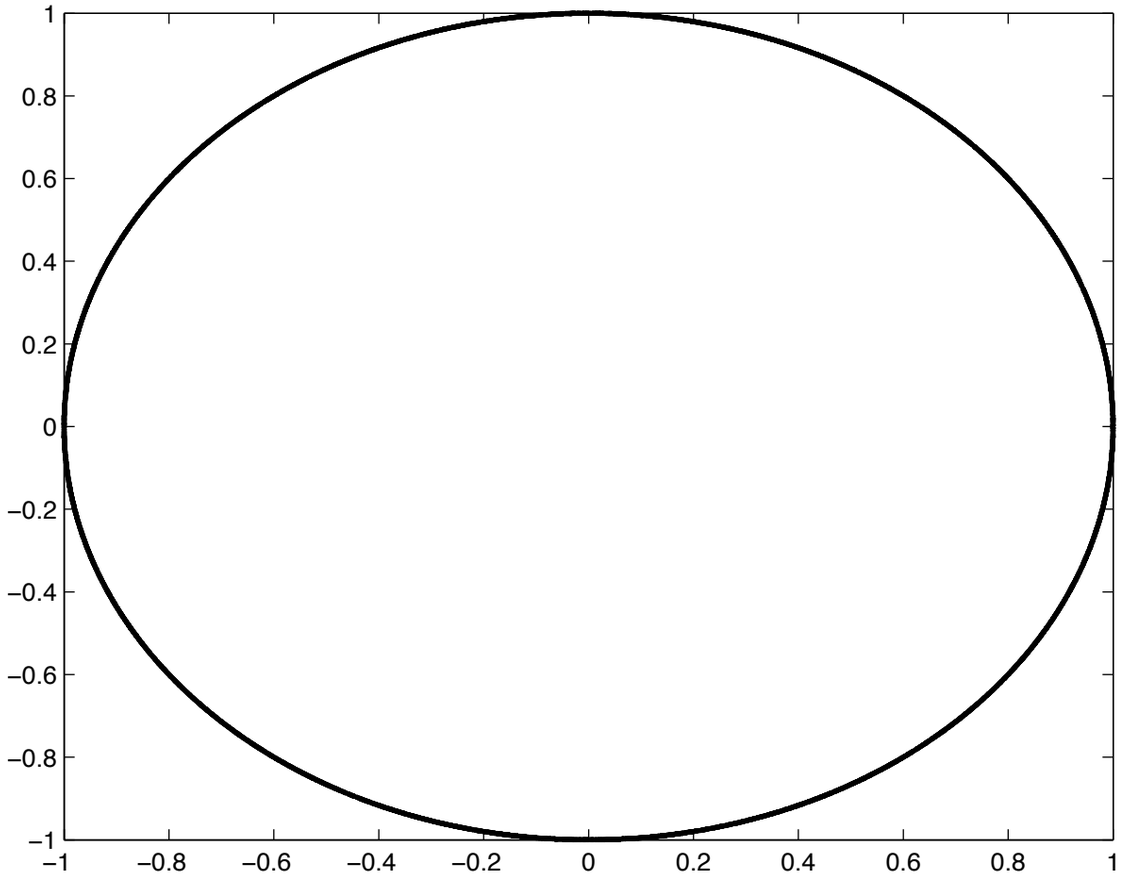




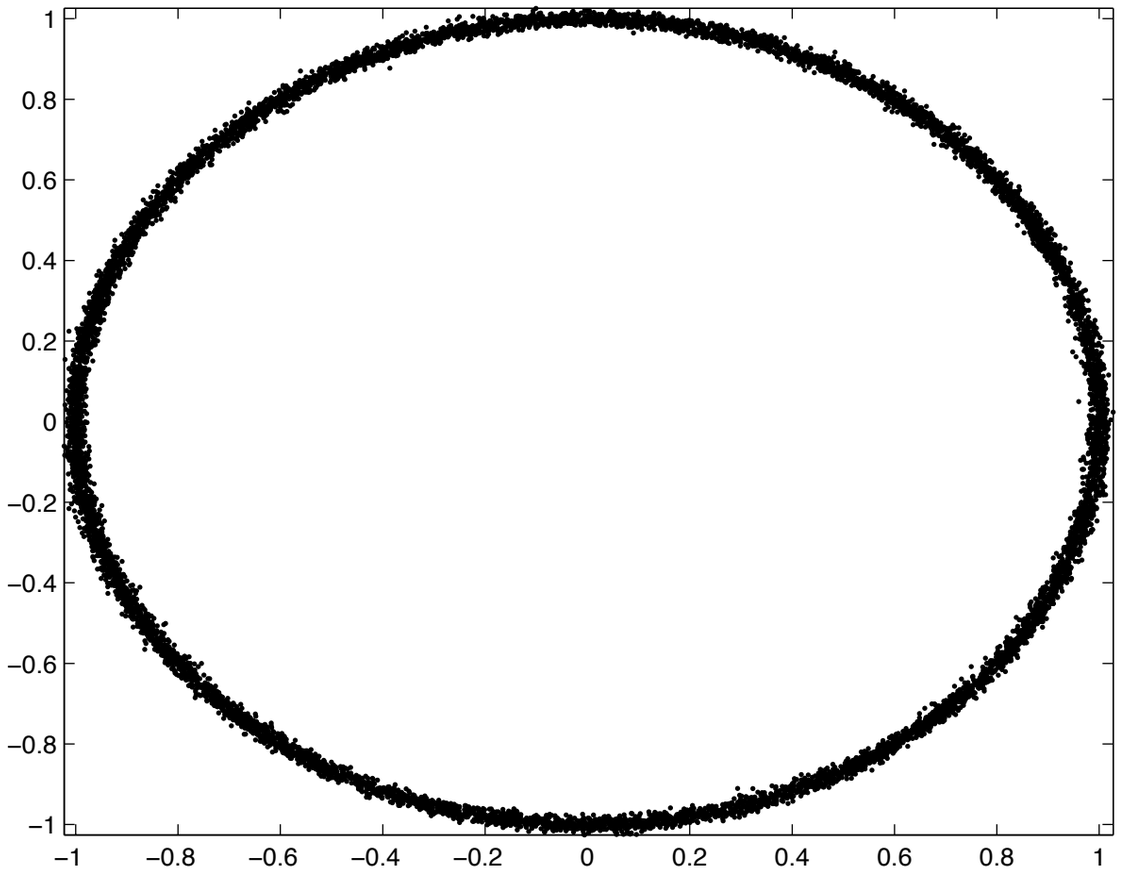
Experiments
Noise Effect



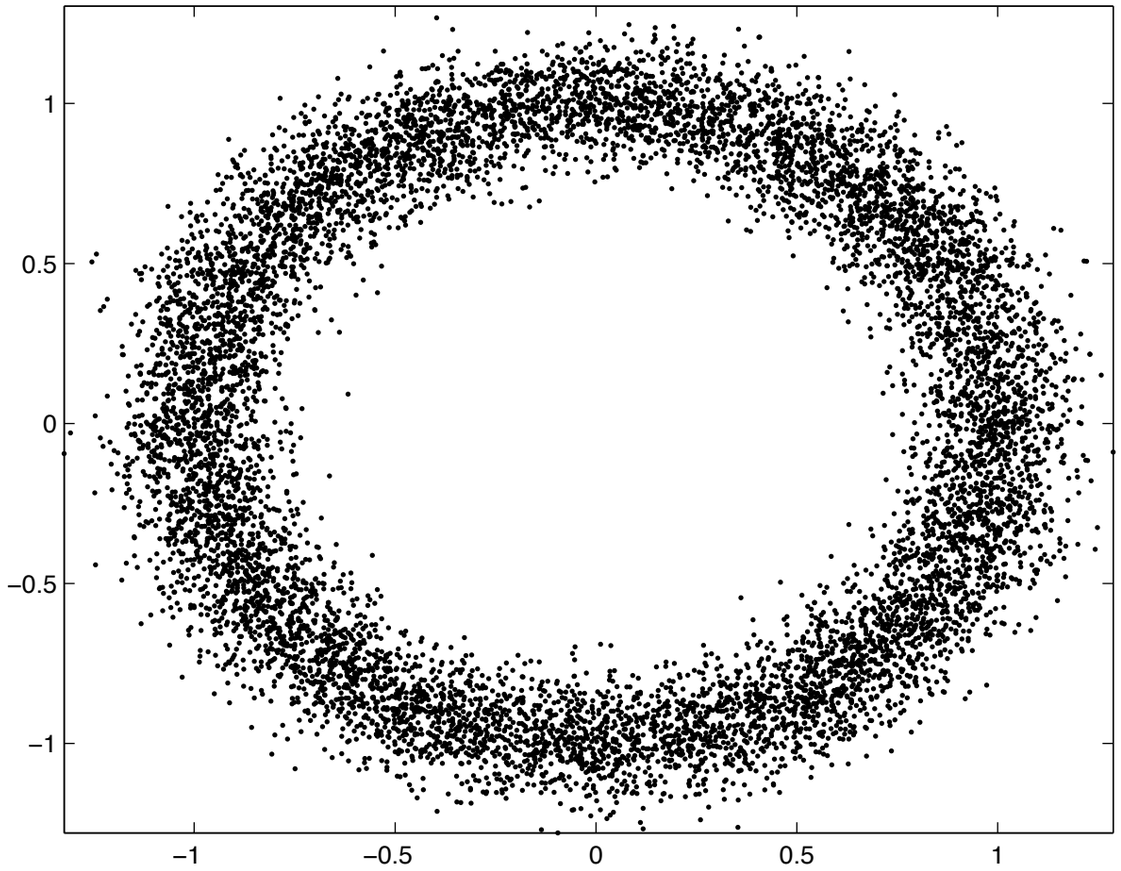
Effect of Noise



Effect of Noise



Effect of Noise



Effect of Noise