

# Multiscale SVD

Estimation of the intrinsic  
dimensionality

Presented by Yael Barak and Charles Sutton  
September 28, 2016  
236327 – Signal and Image Processing by Computer

# Outline

- › Introduction to the problem
- › Multiscale SVD
- › Analysis on a new dataset
- › Conclusion and questions

# Introduction to the problem

# Estimation of intrinsic dimension

- › Occurs in many scientific problems
  - Number of variables in statistical models
  - Number of degree of freedom in dynamical systems
  - Estimation of probability distribution highly concentrated around low dimensional manifold
- › Input of algorithms in many scientific fields
  - Signal processing, economics, genomics ...

# Defining our datasets and notations

- › Let  $M$  be a smooth  $k$ -dimensional non-linear manifold, then:
  - Let  $X = \{x_i\}_{i=1}^n$  be a set of uniformly distributed random sample points of  $M$
  - Let  $\tilde{X} = \{x_i + \sigma\eta_i\}_{i=1}^n$  be the noisy samples, where  $\eta_i$  is a centered white noise with  $\sigma$  as its standard deviation
- › Given a set  $\tilde{X}$  of  $n$  sample points embedded in  $\mathbb{R}^D$  - they will be represented by a  $n \times D$  matrix

# Notions of dimensionality

- › Ambient dimension : the dimensions where the manifold is embedded (in the matrix representation : the number of columns)
- › Extrinsic dimension : minimum number of dimensions in which the shape of the manifold can be embedded (in the matrix representation : the rank of the matrix)
- › Intrinsic dimension : the number of parameters needed to generate the manifold

# EXAMPLE

Ambient dimension : 3

> The circle is embedded in  $R^3$

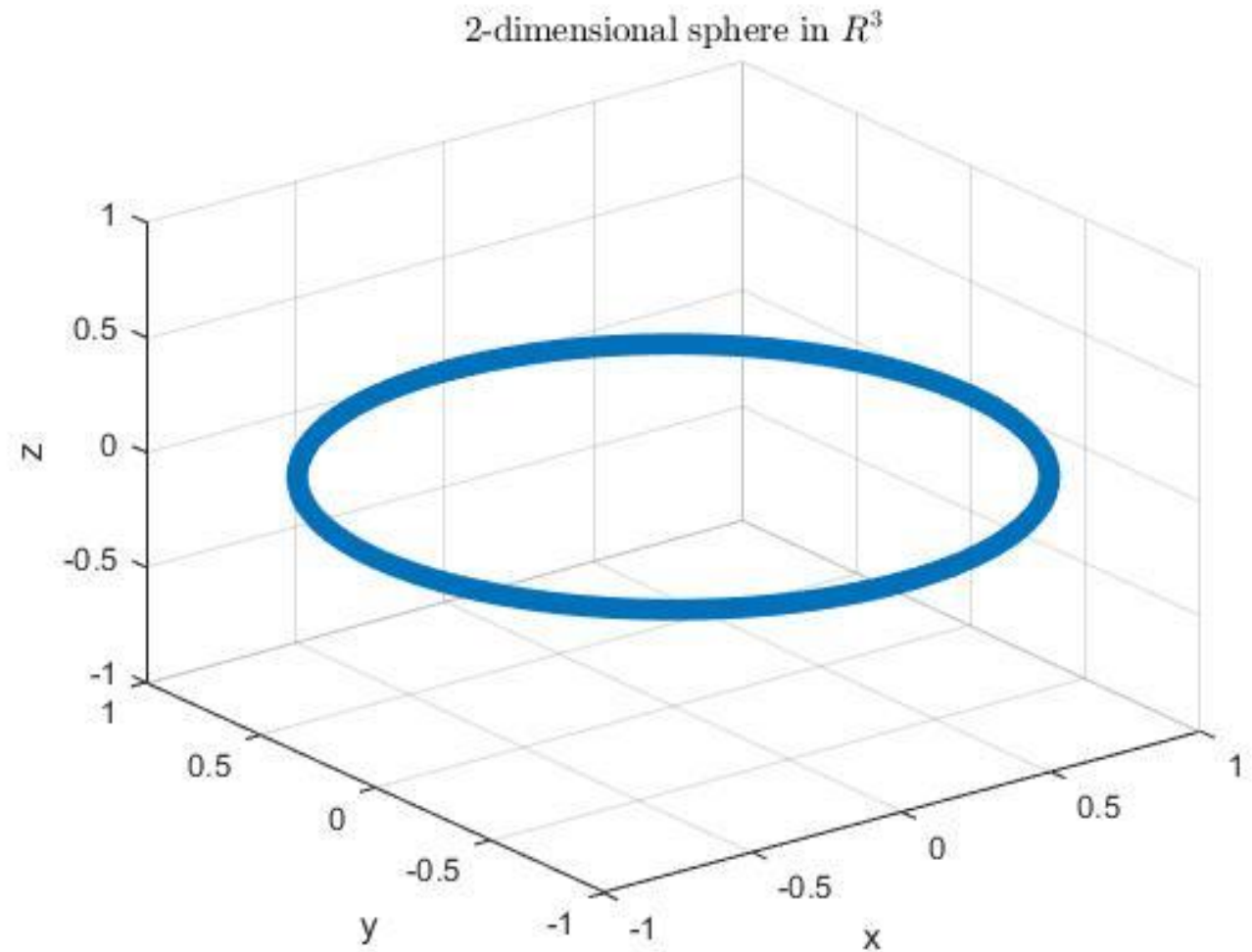
Extrinsic dimension : 2

> This is a 2 dimensional circle

Intrinsic dimension : 1

> The circle is generated with only one parameter

$(\cos(\alpha), \sin(\alpha), 0)$



# Multiscale SVD



# Estimation with SVD

The SVD is a good estimator of the extrinsic dimension, and is most often used to perform dimensionality estimation

## Linear manifolds

- › e.g. Hyperplane
- › Intrinsic = Extrinsic
- › SVD is a good estimator of the intrinsic dimensionality
- › Robust to noise

## Non linear manifolds

- › e.g. Sphere
- › Intrinsic  $<$  Extrinsic
- › SVD over estimates the intrinsic dimension

# LOCALLY VS GLOBALLY

Curvature causes  
over-estimation of the intrinsic  
dimension

SVD is performed globally

The manifold is locally linear  
and can be approximated by a  
tangent plane.

Local SVD should approximate  
the local extrinsic dimension  
(that is also the local intrinsic  
dimension)

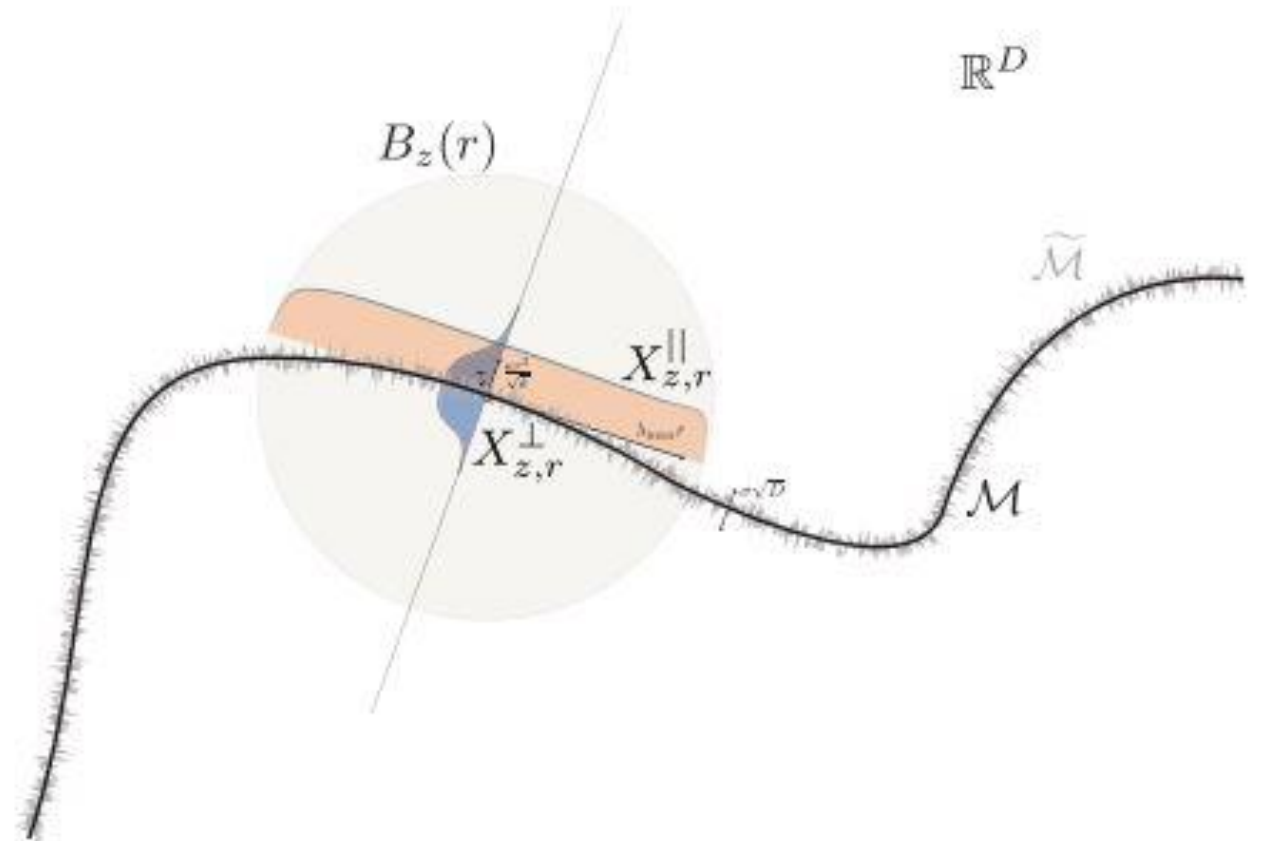


Figure taken from : Little, A. V., Maggioni, M., & Rosasco, L. (2011). *Multiscale geometric methods for data sets I: Intrinsic dimension*.

# MSVD

Compute the singular values for  $X$  intersected with each ball centered around point  $z$  with radius  $r$ .

This process is performed over a wide range of radii.

Average the results for each  $r$  over the dataset

Identifying a range of scale three groups of singular values

(intrinsic, extrinsic, noise) can be distinguished

Estimation of the dimensionality

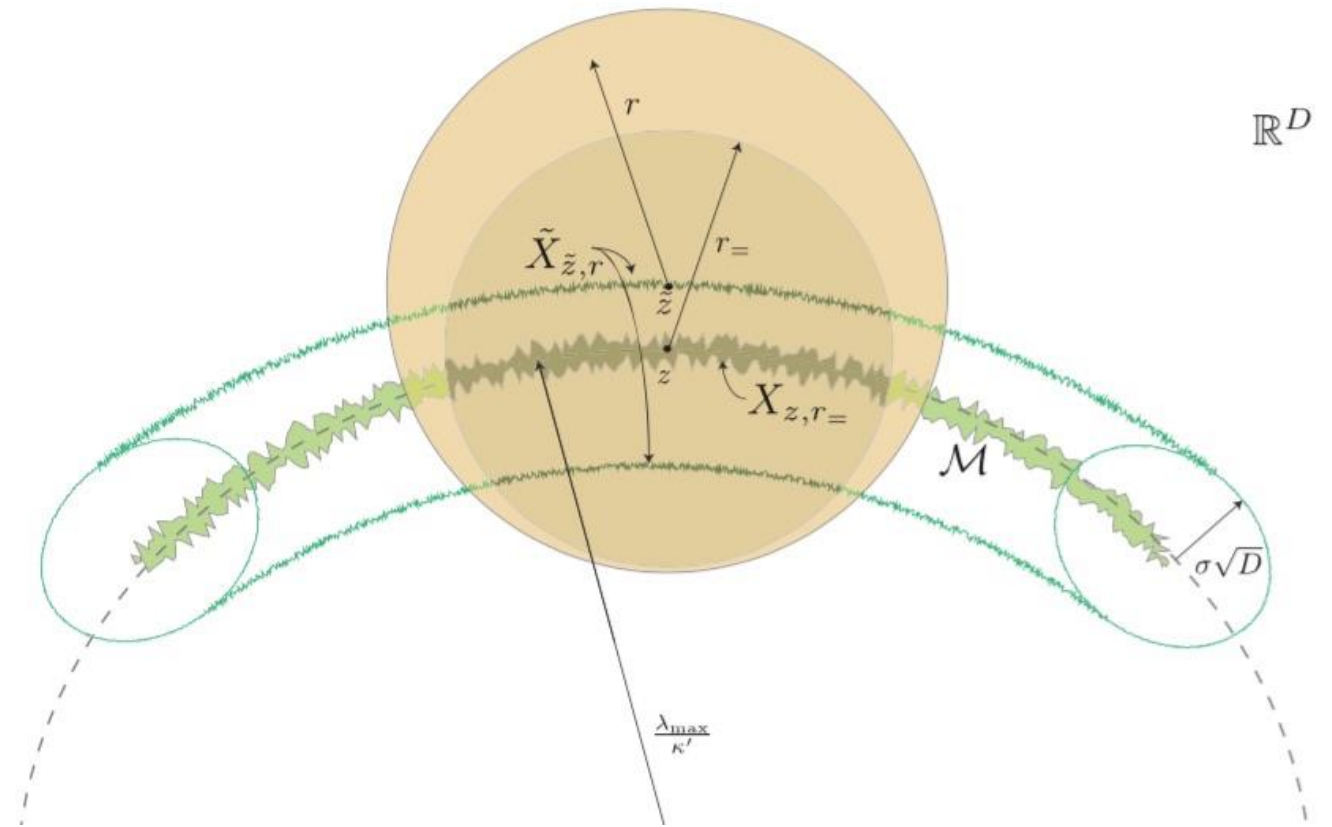


Figure taken from : Little, A. V., Maggioni, M., & Rosasco, L. (2011). *Multiscale geometric methods for data sets I: Intrinsic dimension*.

# The paper's case study

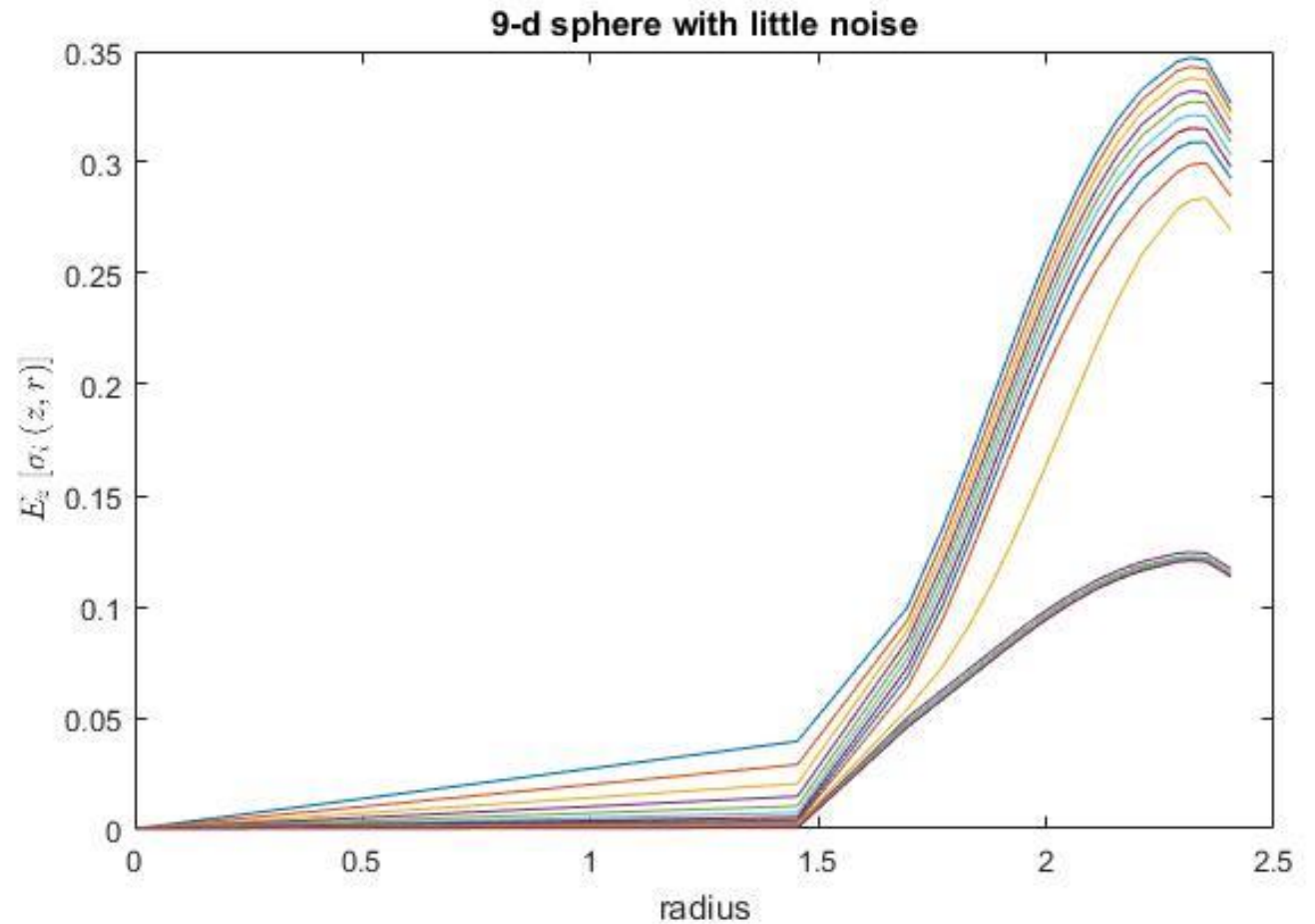
- › Sphere is defined as  $S^k = \{x \in \mathbb{R}^{k+1} \mid \|x\|_2 = 1\}$
- › extrinsic dimension :  $k+1$
- › intrinsic dimension :  $k$ 
  - since the equation which describes it is of  $k+1$  parameters with one constraint, hence it has  $k$  degrees of freedom which are the intrinsic dimension.
- › The sphere is embedded in  $\mathbb{R}^D$

# BEHAVIOR OF THE S.V.

The noise S.V. converge to the std of the white noise (here  $N(0,0.1)$ )

Intrinsic S.V. are the top 9 S.V.

Extrinsic S.V. is the 10<sup>th</sup> S.V.  
recognizable by the gap



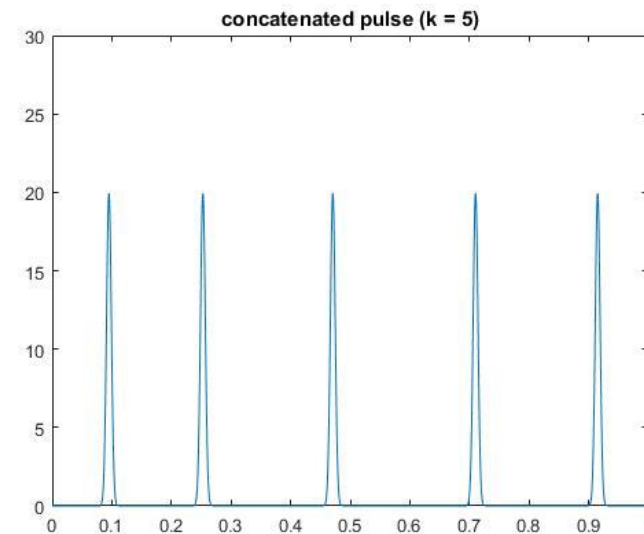
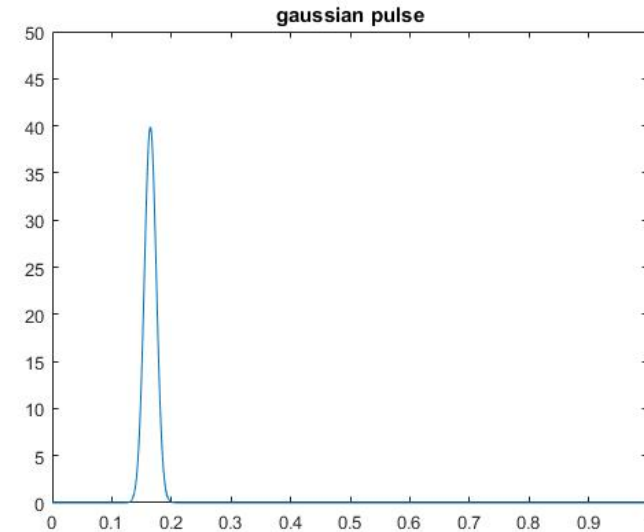
# Analysis on the pulse dataset

# THE PULSE DATASET

This dataset is more related to signal processing : heartbeats, musical tempo, breathing ...

Only one parameter :  $\mu$  (fixed width)

Intrinsic dimension is scalable (concatenation)



# MSVD WORKS ON PULSES

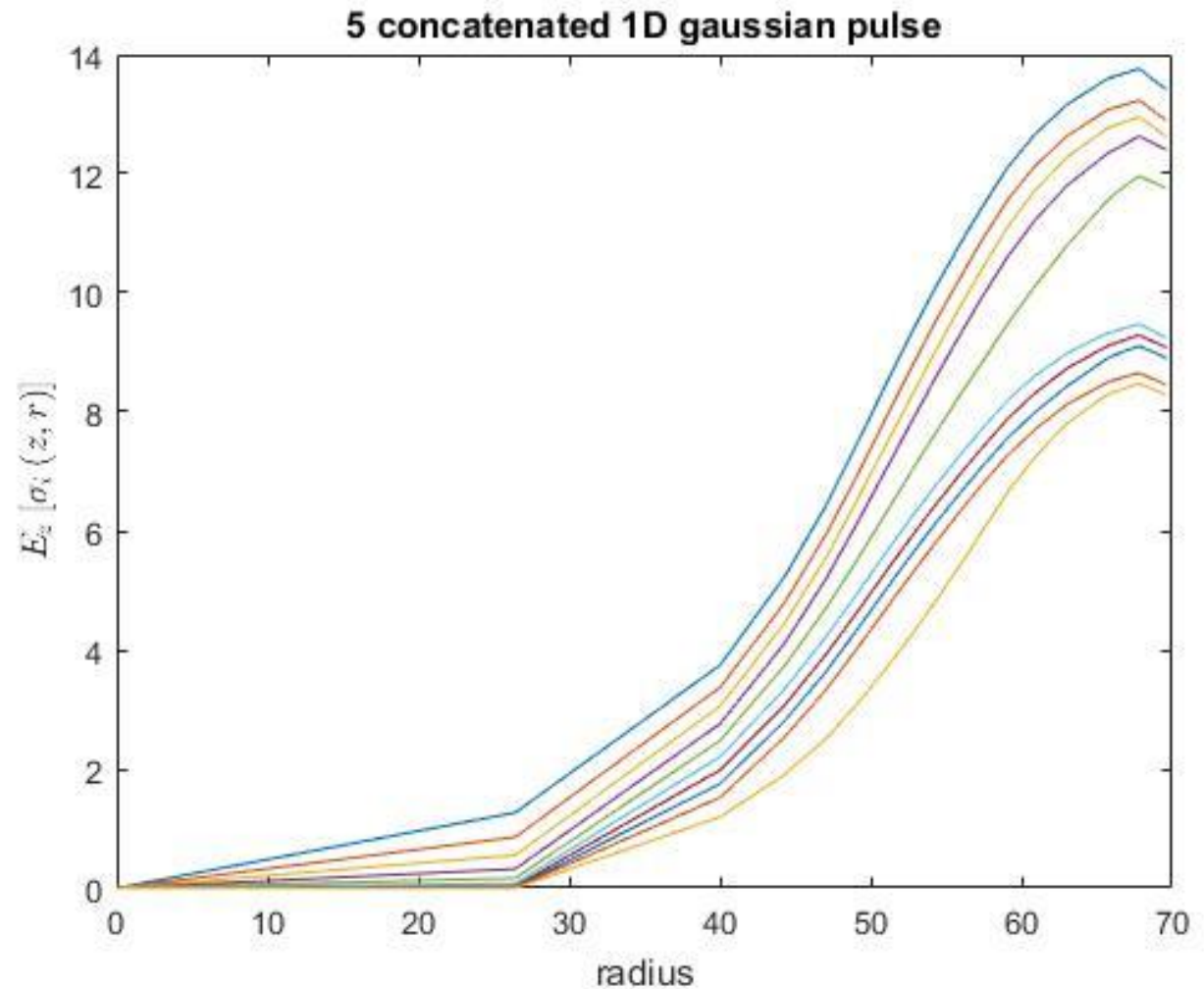
MSVD on the 5 concatenated pulses ( $k=5$ )

There is a gap between the fifth and the sixth curve

-> The MSVD estimation is correct

MSVD is much more accurate than SVD

-> Global SVD estimation exceeds 50 in this case





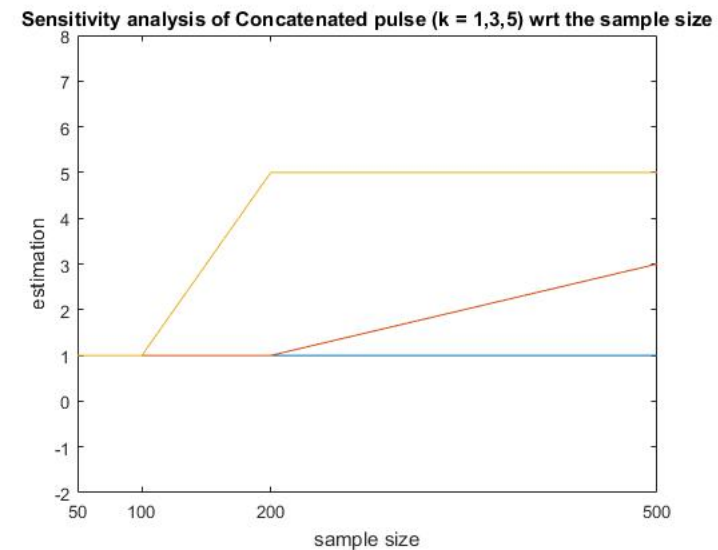
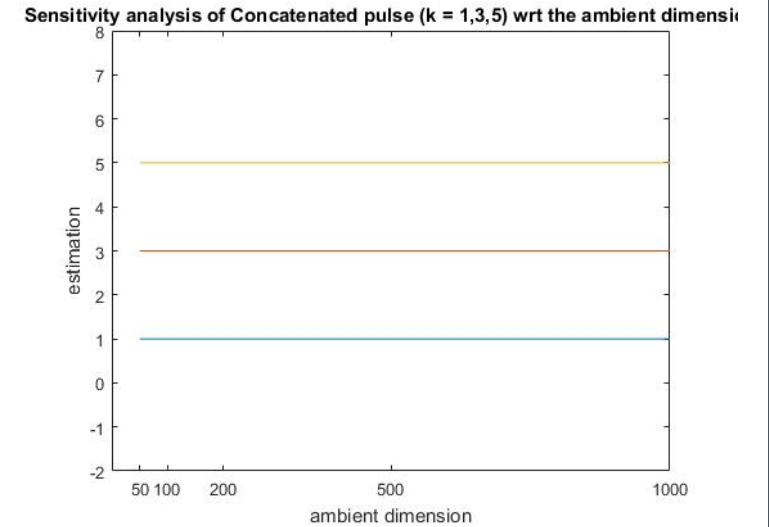
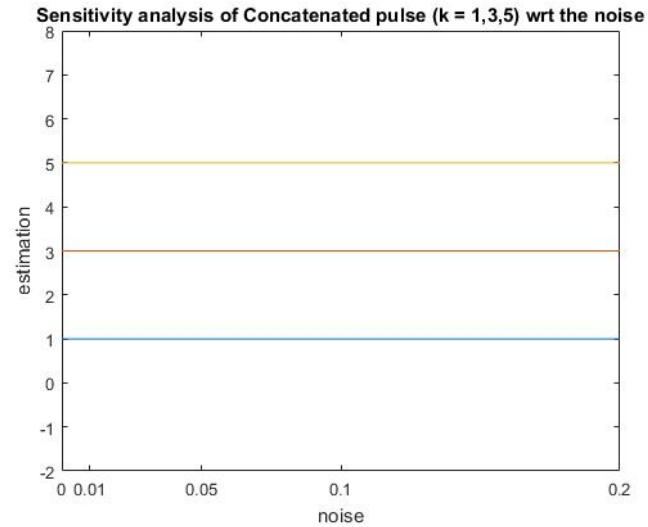
# SENSITIVITY ANALYSIS

MSVD is robust :

-> noise, up to a limit

-> ambient dimension

-> sample size when it isn't too low



# Conclusion

- › Accurate technique to estimate the intrinsic dimensionality of high dimensional datasets
- › Generalizes well on the pulse dataset
  - Robust to main parameters
- › Future improvements and research direction :
  - Measure improvements on real data
  - Computational : MSVD is natively a distributed technique

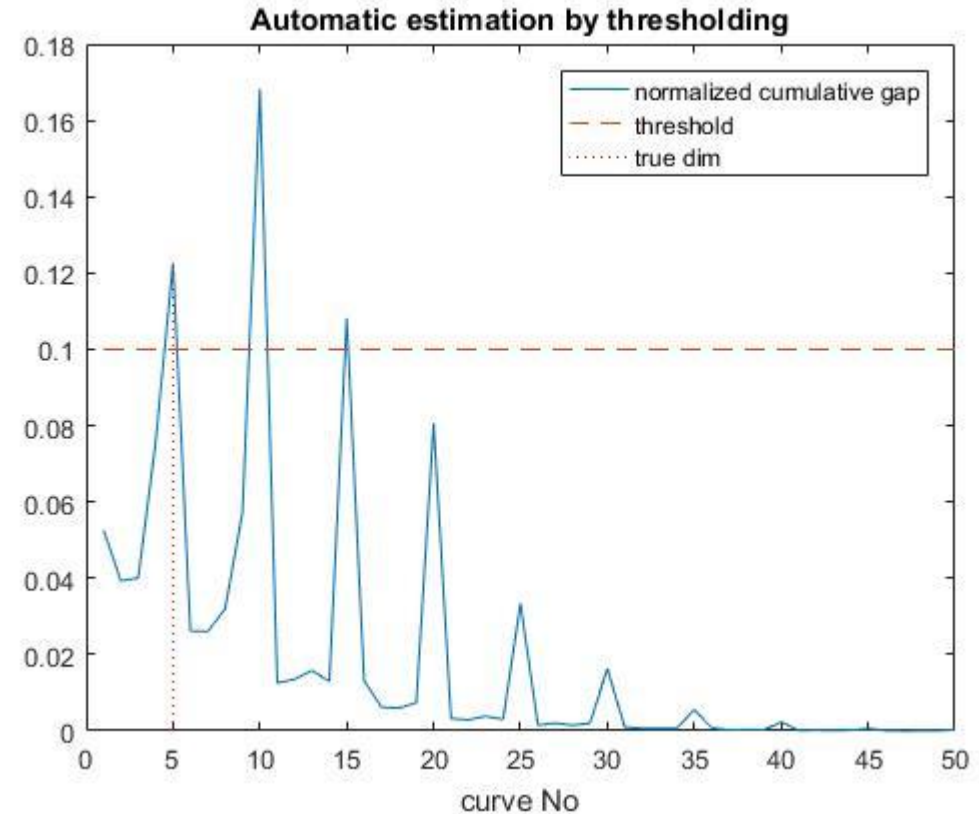
# Questions

# BONUS 1 : AUTOMATIC ESTIMATION

We compute the average gap between the curves given of the MSVD

We estimate the intrinsic dimensionality by thresholding on the normalized gap

Best threshold found : 0.1



# Bonus 2 : Uniform sampling on a sphere

## USING THE HYPERCUBE

- › Sampling points in the hypercube, keep only point that belong to the ball  $(z,r)$  and normalize
- › Problem : the volume of the sphere decreases too fast with regard to the dimensionality
- › Very inefficient in high dimension

## USING NORMAL DISTRIBUTION

- › Gaussian distribution is symmetric, therefore it is uniformly distributed over all the directions
- › Only normalize points drawn from the Gaussian distribution

