CS 236327 : SIGNAL AND IMAGE PROCESSING BY COMPUTER

MULTISCALE SVD : ESTIMATION OF INTRINSIC DIMENSIONALITY

September 28, 2016

Students : Yael Barak and Charles Sutton
Supervisors : Yehuda Dar and Alfred Bruckstein

Technion - Israel Institute of Technology
Department of Computer Science

Multiscale SVD

Yael Barak and Charles Sutton

September 27, 2016

1 Introduction

Every manifold has both extrinsic and intrinsic dimensions. The extrinsic dimension is the minimum number of dimensions in which the shape of the manifold can be embedded while the intrinsic dimension is the number of parameters needed in order to generate it. For example, the unit circle embedded in $\mathbb{R}^2$ has an extrinsic dimension of 2 while its intrinsic dimension is 1 since it can be formulated by only one parameter $\theta : sin^2\theta + cos^2\theta = 1$.

In many problems, the most simple approach is to assume that the dimensionality of a dataset is its extrinsic dimension. However, it appears that in many fields of science (machine learning, statistical analysis and image processing ...) this is the intrinsic dimension that is of interest. In these fields of research, it is very common to assume that a dataset is drawn from a low dimensional manifold, that's why estimating the intrinsic dimension of a point cloud dataset is studied by many.

In this report we present a method (described in the paper [1]) of estimating the intrinsic dimensionality of datasets. We first describe the case study of the article : estimation of the intrinsic dimensionality of a 9 dimensional unit sphere. Then we present our implementation of the methods' algorithm and how we used it in order to run a dimensionality estimation on a k dimensional unit sphere using our own code and the results we received. Lastly, we describe our investigation of the method using Gaussian pulse-like signals as a case study from which it appears that the technique generalizes well to this case.

In this report, all the datasets we describe are non linear manifolds of intrinsic dimension k embedded in $\mathbb{R}^D$, where $k \ll D$, and with added white noise of dimension D. Datasets are represented by a matrix in $\mathbb{R}^{n \times D}$, where n is the number of samples.

Contents

2 Presentation and reimplementation of MSVD

2.1 Estimating dimensionality with SVD

Every matrix X admits a singular value decomposition (SVD) to the form $U\Sigma V$ where the diagonal of Σ is composed of $n-r$ zeros and the r non-zero singular values of X which correspond to the rank of X. This decomposition process is known as global SVD and it is a well-established method of determining the dimension of a dataset. In the case of linear manifolds the intrinsic dimension equals to the extrinsic dimension of the dataset. Hence, global SVD is an accurate estimator of the intrinsic dimension of the dataset. As long as the manifold is linear, global SVD is also robust to the noise : the estimation of the dimension $\hat{k}$ is done by detecting the gap in the singular values $\sigma_1 > ... > \sigma_{\hat{k}} \gg \sigma_{\hat{k}+1} > ... > \sigma_D$.

In the case of non-linear manifolds, when trying to estimate intrinsic dimensionality, SVD fails by over estimating the intrinsic dimension. The curvature of the dataset is the cause of the methods' failure. As a simple example : let us consider the 2D circle embedded in $\mathbb{R}^3$, since we need only one parameter to generate such a circle, its intrinsic dimensionality is 1 (fig. 1).

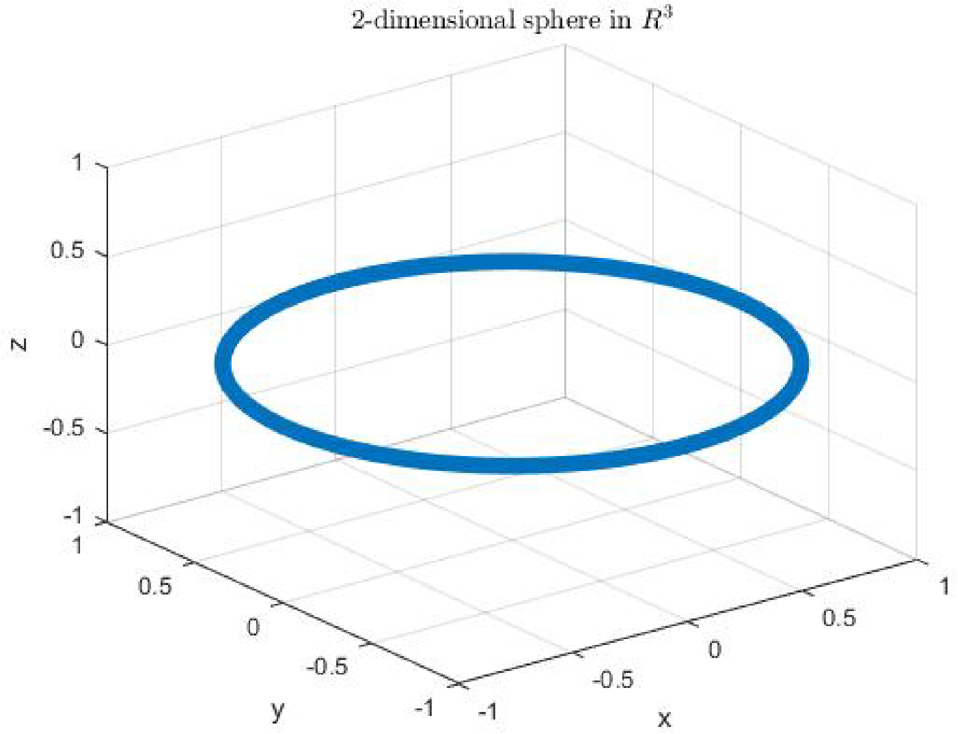


Figure 1: 2 dimensional circle embedded in $\mathbb{R}^3$

Preforming SVD on this dataset results in estimating the dimension to be 2 – that is an over estimation that was cause due to the curvature of the circle.

2.2 A multiscale SVD approach

As we saw, global SVD fails on non-linear datasets due to the curvature that causes over estimation of the intrinsic dimension. The SVD process is performed globally – over all of the dataset. The multiscale SVD method suggests thinking locally vs. globally : even due the whole dataset is non-linear, if we split the manifold into small enough areas, then each one is locally linear and can be approximated by a tangent plane.

Before explaining the technique, let us define what is the problem we are trying to solve and the notations used throughout the report. These notations are similar to the notations of the paper [1].

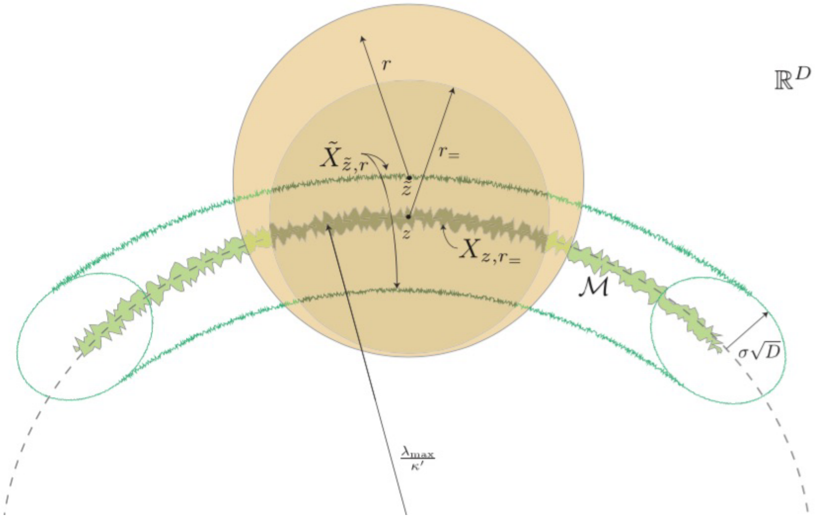


Figure 2: Pictorial representation of SVD at a lower scale (figure from [2])

2.3 Notations

Let $\mathcal{M}$ be a smooth k-dimensional non-linear manifold, let $X = \{x_i\}_{i=1..n}$ be a set of uniformly distributed random sample points of $\mathcal{M}$ and let $\tilde{X} = \{x_i + \sigma\eta_i\}_{i=1...n}$ be the noisy samples, where η_i is a reduced white noise and $\sigma > 0$ is its standard deviation. Finally, let us define $\tilde{X}^{(z,r)}$ as the intersection of X with the ball of center z and radius r.

Given a set $\tilde{X}$ of n sample points embedded in $\mathbb{R}^D$ and represented by a $n \times D$ matrix, our goal is to accurately estimate the intrinsic dimension of $\mathcal{M}$.

2.4 Description of the algorithm

The key idea of MSVD is that for r small enough, $\tilde{X}^{(z,r)}$ is almost linear, thus it can be approximated by a tangent plane. Consequently, we expect $\tilde{X}^{(z,r)}$ to have k singular values corresponding to the intrinsic dimension, then $rank\left(\tilde{X}^{(z,r)}\right) - k$ smaller singular values corresponding to the curvatures. The $D - rank\left(\tilde{X}^{(z,r)}\right)$ singular values almost equal to zero correspond to the noise.

We need to scale with the following consideration in mind : r must not be too small, because we want $\tilde{X}^{(z,r)}$ to contain enough points to perform SVD on it successfully. On the other hand, r must not be too large, because we want the approximation of $\tilde{X}^{(z,r)}$ by a tangent plane to be as accurate as possible. Another consideration is the noise factor : for small r the noise is more prominent in $\tilde{X}^{(z,r)}$ and thus effects the results more.

With these constraints in mind, let $\left\{\sigma_i^{(z,r)}\right\}_{i=1...D}$ be the singular values of $\tilde{X}^{(z,r)}$. The MSVD algorithm computes $\left\{\sigma_i^{(z,r)}\right\}_{i=1...D}$ for each z and for an wide range of radii. For every r, i , the algorithm computes the average of σ_i^r for every point z meaning : $\sigma_i^r = \mathbb{E}_z\left[\sigma_i^{(z,r)}\right]$. The final step of the algorithm is to identify a range of scales where the intrinsic dimension singular values can be distinguished from the curvature and noise singular values.

Therefore, we can retrieve the intrinsic dimensionality by performing SVD on multiple $\tilde{X}^{(z,r)}$ datasets at different scales.

The process described is a computationally expensive one. As the size of the dataset grows, the complexity of the process increases quadratically and depending on the r scaling, this process can be even more complex.

2.5 Reimplementation of the paper

In the paper [1], the MSVD method of estimating intrinsic dimension is demonstrated over the the k dimensional unit sphere $\mathcal{S}^k = \left\{x \in \mathbb{R}^{k+1} : \|x\|_2 = 1\right\}$ naturally embedded in $\mathbb{R}^D$ where $k \ll D$.

The intrinsic dimension of $\mathcal{S}^k$ is k since the equation which describes it is of $k+1$ parameter with one constrain, hence it has k degrees of freedom which are the intrinsic dimension.

The dataset in the paper was constructed using the parameters $k = 9$ and $D = 100$. The dataset was of size $n = 1000$ and the samples were corrupted by a D dimensional centered gaussian noise, and with $\sigma = 0.1$.

The results presented in the paper show that the MSVD algorithm yielded that the intrinsic dimension of this dataset is 9, it has one S.V. corresponding to the sphere curvature and another 90 S.V. which correspond to the noise. As the results in the paper show, the process works well on this dataset.

On this dataset global SVD doesn't work since it return 10 S.V. that are significantly larger than the rest of the S.V. (90), that is an over estimation of the intrinsic dimension by 1. A demonstration of that is shown in the script `global_doesnt_work_for_non_linear_manifold_with_gui.m` in our code repository.

In order to test the MSVD algorithm we implemented it ourselves and tested it on the $\mathcal{S}^9$ sphere dataset. We have generated the dataset by using n points drawn uniformly on the S^9 sphere, and by adding a D dimensional gaussian centered noise with $\sigma = 0.1$. Results are shown in figure 3.

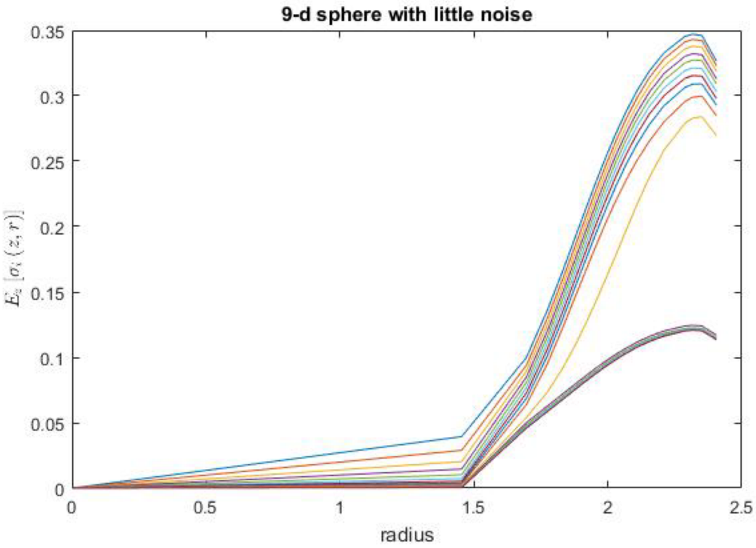


Figure 3: Our implementation of MSVD on $\mathcal{S}^9$

There is a visible gap between the 9 upper S.V. and the singular values below them. The S.V. corresponding to the noise can be distinguished for the rest of the S.V. as well. Meaning that on this dataset our code performs well.

2.6 Computational enhancements

Efficient generation of the sphere in high dimension

Sampling n points uniformly on a k dimensional sphere is not trivial. One may have (the wrong) idea to sample a vector belonging to the k dimensional hypercube and then normalizing the vector.

The problem with this method is that it doesn't produce a uniformly distributed dataset, since the projection is not equally weighted on the surface.

In fact, the previous method can work if we add a restriction : omit all the points outside of the unit sphere. By doing so, the projection is equally weighted over k dimensional sphere surface.

However, if this method works it is far from efficient. Let us recall that the volume of a k dimensional sphere over the volume of an hypercube is $\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}$. Therefore the ratio decreases more than exponentialy and tends to 0 as the dimensionality grows. Since the volume corresponds to the probability of accepting of rejecting a point, we reject most of the candidates in high dimensions.

An efficient solution to this problem is to use the property of symmetry of the multidimensional gaussian distribution. Since the gaussian distribution is symmetric, it selects a direction uniformly. Consequently, normalizing the n samples drawn from the $k+1$ multidimensional gaussian distribution is an efficient solution to get n samples uniformly distributed on $\mathcal{S}^k$.

Finding points in $\tilde{X}^{(z,r)}$

When performing multiscale SVD, using the naive way to find the points in $\tilde{X}^{(z,r)}$ is very inefficient. To overcome this problem, we compute a symmetric $n \times n$ distance matrix where entry (i, j) in the matrix represents the Euclidian distances between points i and j. From this matrix we compute two others $n \times n$ matrix , where in each row i, the j cell is the distance or the index of the j^{th} farthest neighbor of i. By doing so, the process of finding all the neighbors of z becomes much more efficient.

Smart radius scaling

This enhancement is directly inspired from §4 in [1]. Instead of selecting a uniform range of radii in $[0, r]$, we select a range of radii $\{r_1, r_2, ...\}$ such that $r_i = min\left\{r : \mathbb{E}_z\left[\left|\tilde{X}^{(z,r)}\right|\right] \geq s_i\right\}$ where $\{s_1, s_2, ...\}$ is a range of of predefined steps.

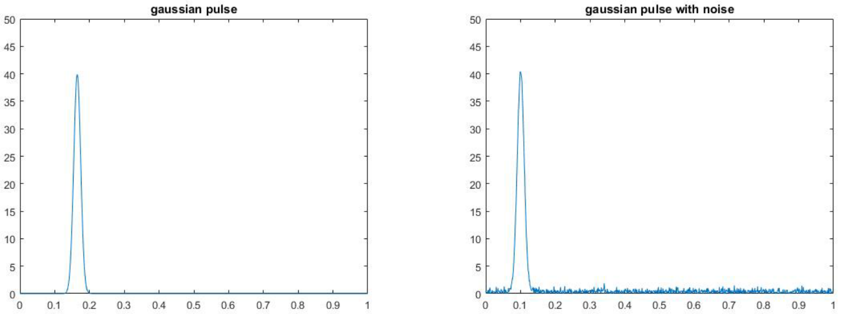
3 Extension to pulses

In the previous section, we have successfully reimplemented the results of the main article [1]. Since it appears the technique works very well on the theoretic framework given by the authors, a major issue of our project was to test if this technique can be extended to a wider collection of signals. Hence we have conducted the analysis to a case very related to signal processing : pulses.

3.1 1D pulses dataset

The case of 1D pulses is an interesting theoretical framework related to hot applications in signal processing. Many cases rely on pulses as the elementary signal : music tempo, heartbeats, breathing etc.

Here, we define our dataset as a collection of pulses generated by a parameter θ of dimension one in the euclidian sense. Each pulse is embedded in $\mathbb{R}^D$, where D is called the “ambient dimension”.



(a) without noise

(b) with noise

Figure 4: Gaussian pulses

In our experiments we have handled two kinds of pulses defined as follows :

- The *gaussian pulse* from the density function of a normal distribution :

$$p_{gaussian}\left(x, \theta = \mu\right) = \frac{1}{\sqrt{2\pi\sigma_p^2}} exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma_p^2}\right) \tag{1}$$

- *The stair pulse :*

$$p_{stair}\left(x, \theta = \mu\right) = \mathbf{1}_{\left[\mu - \frac{\sigma_p}{2}, \mu + \frac{\sigma_p}{2}\right]}\left(x\right) \tag{2}$$

Where $\mathbf{1}$ is the indicator function of the interval $\left[\mu - \frac{\sigma_p}{2}, \mu + \frac{\sigma_p}{2}\right]$.

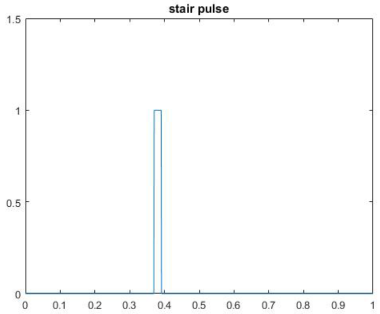
These pulses are illustrated in fig. 4 and fig. 5.

In our experiments, we discretize the function in D points uniformly on the $[0, 1]$ interval, σ_p is fixed and only μ is changing. μ is uniformly selected over the $[0, 1]$ interval. σ_p controls the width of the signal in both cases.

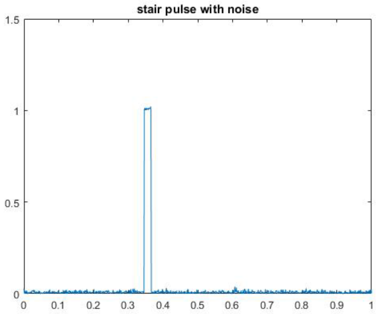
The dataset is represented by a matrix of $\mathbb{R}^{n \times D}$, where n is the number of sample and D the ambient dimension.

3.2 Estimation of the intrinsic dimension

Our first experiment is to check if the multiscale SVD estimates the intrinsic dimension k of the 1D pulse dataset. Here, even if the intrinsic dimension is trivially 1, the manifold representing the collection has a complex shape, that's



(a) without noise



(b) with noise

Figure 5: Stair pulses

why the global SVD technique estimates a dimension $\hat{k} \gg 1$. We perform the experiment on a dataset with the following parameters :

- number of samples : $n = 1000$
- Ambient dimension : $D = 1000$
- Noise : $\sigma_n = 0.1$
- Pulse width : $\sigma_p = 0.1$

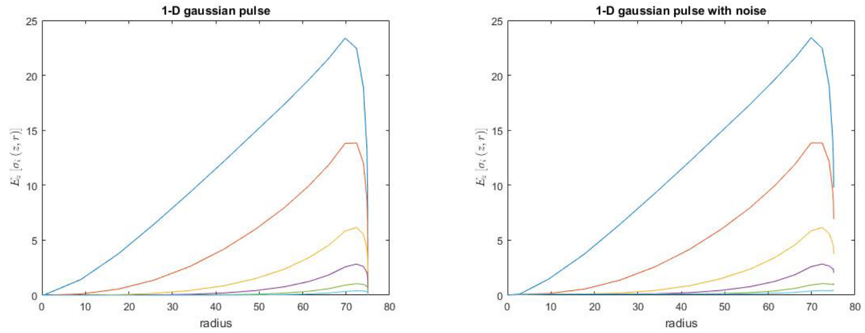
We add a gaussian centered noise of variance σ_n on all the matrix representing the dataset.

We can see the technique predicts well the intrinsic dimension of the 1D gaussian pulses dataset in both noisy and non-noisy cases (Fig. 6). In this case, we see that estimated dimension is one because of the quadratic shape of the second eigenvalue.

3.3 A case of higher dimension

The gaussian and stair pulses are a vanilla case, in the nature it is rare to find data reduced to the a single pulse as used in the previous experiment. In fact, it is more common to see sequences of pulses or combination of pulses.

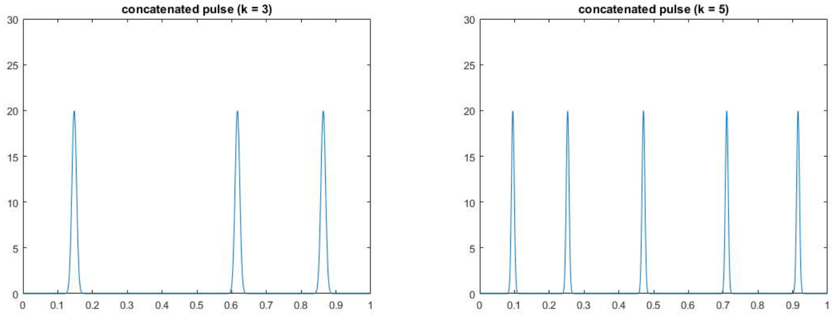
An interesting case to develop is the concatenation of k pulses (fig. 7). The intrinsic dimensionality of this dataset is k since it consists in choosing μ_i uniformly in $\left[\frac{i-1}{k}, \frac{i}{k}\right]$ for $i \in \{1, ..., k\}$. For $k \in \{2, ..., 10\}$, SVD gives an estimation of the dimension of the manifold that is over 50 and so hugely overestimate the true intrinsic dimension.



(a) without noise

(b) with noise

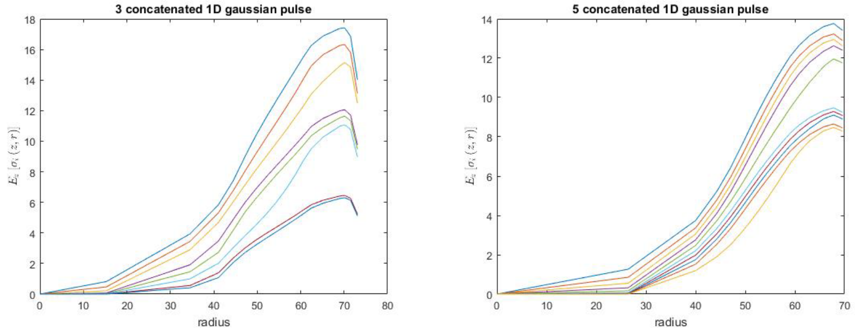
Figure 6: Singular value behavior on the gaussian pulses dataset



(a) $k = 3$

(b) $k = 5$

Figure 7: k concatenated pulses



(a) $k = 3$

(b) $k = 5$

Figure 8: MSVD on k concatenated pulses

We have chosen concatenation because it is not obvious to conduct an experiment on datasets representing the addition or multiplication of k independent pulses. In fact, to conduct an experiment we have to know the true intrinsic dimensionality *a priori* which is not in the case of addition and multiplication. Even if k is an upper bound of the true intrinsic dimension, but it is probably lower than k due to the "collision" between pulses. The average number of collision depends on the value of the parameters, and a closed formula or a well approximation of its value is not easy. Since these collisions cannot appear in the concatenation case, we decided to work on this case.

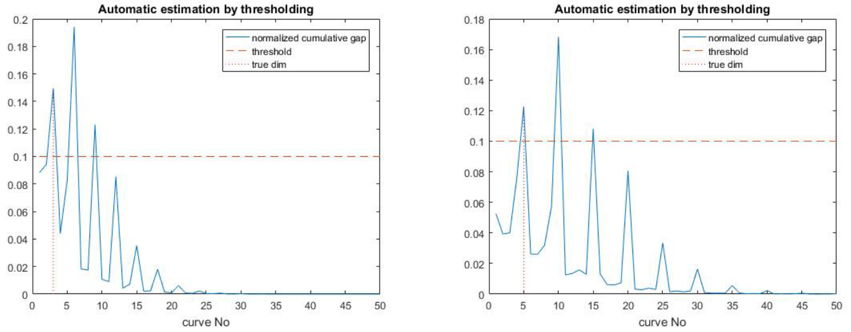
In our experiment, we performed the MSVD estimation on k concatenated pulses datasets for the cases $k = 3$ and $k = 5$ and with the following parameters :

- Number of samples : $n = 500$
- Ambient dimension : $D = 1000$
- Noise : $\sigma_n = 0.01$
- Pulse width : $\sigma_p = 0.1$

We clearly see in figure 8 both cases that there is a gap between the k first singular values curves and the others. Therefore, the estimation given by this technique is correct and much more accurate than global SVD.

4 Generalization

A major issue is of this report was to know if MSVD can be generalized to new type of datasets or if it fits only the specific technical frameworks introduced in the founding paper [1]. To assess the generalization to pulses, we perform the sensitivity analysis on the most important parameters of the main experiment.



(a) 3 concatenated pulse

(b) 5 concatenated pulse

Figure 9: Automatic estimation by thresholding

In this report, we test the robustness of the technique with regard to the noise σ, the sample size n and the ambient dimension D.

4.1 Automatic estimation

Given the curves of sorted singular values as a function of the radius (fig.6, fig.8), it is simple to visually detect the gap between the $\hat{k}$ first singular values and the others. Nevertheless, for obvious experimental and practical issues we had to program an automatic and objective estimation given the curves.

To do so, we compute the cumulative gap between each curve and then the estimation is performed when the first "gap peak" is detected. The gap peak is detected when the normalized cumulative gap exceeds a threshold. Empirically, we found that a threshold of 0.1 works well. In the case our automatic estimation doesn't find any peak, it returns 0 by default.

An illustration of this technique is given in figure 9.

4.2 Sensitivity analysis

Our sensitivity analysis try to recover most of the datasets and parameters mentioned. More precisely, we test the sensitivity of the parameters σ, n and D, and we focus on two families of datasets : the k dimensional spheres for $k \in \{5, 9, 12\}$, and the concatenated pulses for $k \in \{3, 5, 7\}$.
The complexity[1] of the multiscale SVD is $O\left(D\,n\,m\,log\left(n\right)\right)$ where $m = min(n, D)$ [1], therefore the computational cost is determined by parameters n and D. Here, we perform the sensitivity analysis parameter-wise over all the dataset, and when a parameter is tested the two others are fixed to their default values.

[1] in the worst case

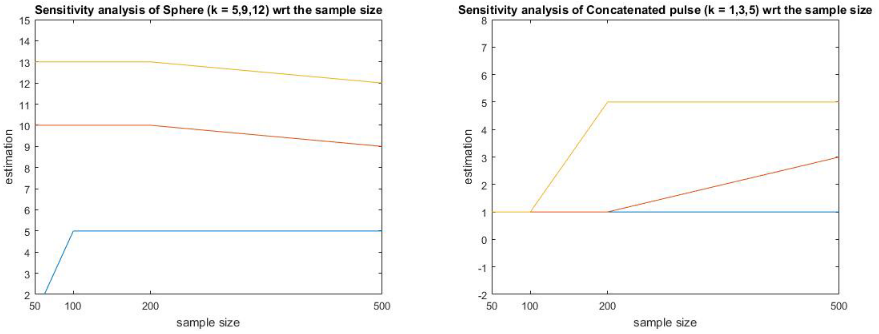
After preliminary results on n and D, we use the following default values for the parameters : $n = 500$, $D = 100$, $\sigma = 0.01$; while in the case of concatenated pulses we use $n = 500$, $D = 800$, $\sigma = 0.01$.

In the case of concatenated pulses, we let $D = 800$ since the resolution a k concatenated pulse is damaged for lower values of D.

In figure 10, we can see that a number of sample $n > 500$ is necessary to have a correct estimation in the results for all the datasets tested. For k concatenated pulses datasets, we can observe that the accuracy of the technique is not linearly correlated to k as it is claimed in [1] : here MSVD works with lower number of points when $k = 5$ than when $k = 3$.

The MSVD technique is highly robust with regard to the parameter D. We can see in figure 11 the technique is totally accurate in its estimations for both families. For the datasets of k concatenated pulses, MSVD works surprisingly well since for lower values of D the signal is damaged because of the low resolution of the discretization. For higher values of D, the techniques is accurate, it is mainly due to the properties of SVD and the low default value of noise.

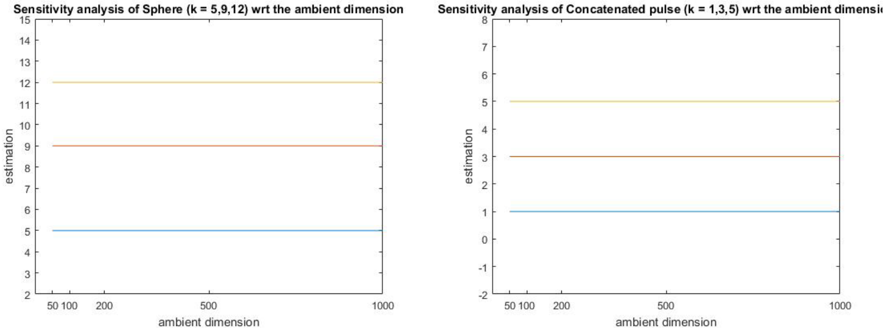
Finally, this technique appears robust to the noise (fig.12), up to the limit that the noise remains “small” with regard to the data. In the case of k concatenated pulses, noise level a relatively small and the technique works perfectly. For k dimensional spheres noise level become high from $\sigma > 0.05$. For small noises, our reimplementation slightly overestimates the k, but the results are analog to the ones found in the main papers. For higher levels of noise, it appears that our thresholding techniques fails to detect the gap peak that’s why the estimation vanishes.



(a) k dimensional sphere

(b) k concatenated pulses

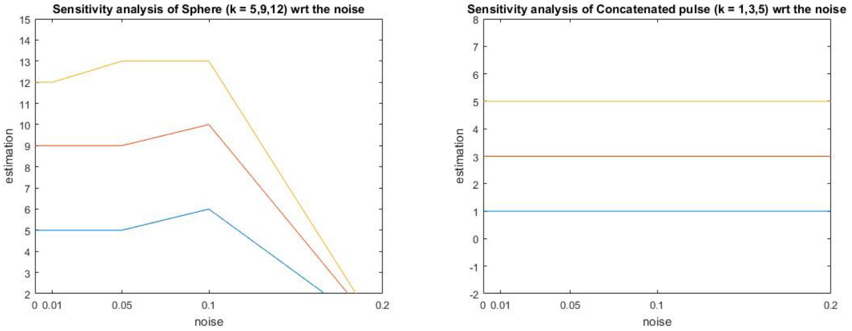
Figure 10: Sensitivity w.r.t the sample size



(a) k dimensional sphere

(b) k concatenated pulse

Figure 11: Sensitivity w.r.t the ambient dimension



(a) k dimensional sphere

(b) k concatenated pulse

Figure 12: Sensitivity w.r.t the noise

5 Conclusion

In this report we presented the multiscale SVD technique and its application to a dataset of concatenated pulses. It appears that on this dataset belonging to the signal processing field, this technique is reliable and much more accurate than global SVD.

The analysis of MSVD in this new dataset nuance some of the claims of the initial papers especially regarding the robustness of the technique for a low number of sample.

Nevertheless, our results show that multiscale SVD generalizes well to this dataset and is an accurate method to robustly estimate the intrinsic dimensionality.

Future improvements of this project consists in overcoming the computational issues by implementing this highly parallelizable method in a distributed way or by trying to perform the method only on subsamples.

Finally, a future research direction include to prove that this technique can generalize to the analysis of a wide range of high dimensional datasets.

References

[1] Little, A. V., Jung, Y. M., & Maggioni, M. (2009, November). Multiscale Estimation of Intrinsic Dimensionality of Data Sets. In AAAI fall symposium: manifold learning and its applications (Vol. 9, p. 04).

[2] Little, A. V., Maggioni, M., & Rosasco, L. (2011). Multiscale geometric methods for data sets I: Intrinsic dimension. Preprint.