Denoising on high-dimension datasets

Jiayuan Wang

May 2, 2016

1 Introduction

Experiments have already shown that the parameter-free denoising algorithm in [1] works well on low-dimension datasets. The motivation of this project is to see the performance of the denoising algorithm on high dimension datasets. After denoising, We compare the persistence diagrams to evaluate the performance of the algorithm. Three datasets are used in this project, they are the primal circle[2], the survivin protein[3] and the MNIST[4][5].
The structure of this paper is described as following: Section 2 gives brief introduction to the denoising algorithm, Section 3 shows the experiment results, Section 4 is observations and explainations for the results and also discusses future works.

2 The Denoising Algorithm

[1] proposes a denoising algorithm that is free of parameter. It first gives a Declutter algorithm with one parameter k and also argues the parameter is unavoidable taking the scale of the data into consideration. Then under a slight stronger sampling condition a parameter-free denoising algorithm is given.
There are two constants $c_{declutter}$ and $c_{resample}$ used in the algorithm. Increasing $c_{declutter}$ and decreasing $c_{resample}$ will move more noise. To get the theoretical guarantees, $c_{declutter}$ is taken as 2 and $c_{resample}$ is taken as $10+2\sqrt{2}$. In practice, $c_{declutter}=2$ and $c_{resample}=4$ works well. But if the dataset doesn't satisfy the assumptions of the noise model, we need to choose the constants carefully.

3 Experiment

The high-dimension datasets used in the denoising are primal circle (50000×25), survivin protein (252996×150) and digit 7s from the MNIST (1279×784). The number means $\#\,of\ points\times dimensions$.
For all datasets except the MNIST, first we apply PCA on them and reduce the dimension to 3 because it takes a long time to run the code for datasets with both high dimensions and large sizes. This might be a problem for those with large distance distortions after PCA. So the

figures of the high-dimension datasets are actually plotted after the dimension reduction. We can still handle the digit 7s in 784 dimensions from the MNIST because the size isn't so large. When calculating the persistence we reduce the dimension to 100 for the MNIST. The persistence is calculated by using code from [6].

Primal Circle

Table 3.1 shows the preserved distance information after dimension reduction for the two datasets of primal circle.

Dataset	Reduce to dimension	Preserved information(%)
50k original primal circle	3	51.28
15k clean primal circle	3	75.67

Table 3.1: Distance information after dimension reduction for the two datasets of primal circle.

Figure 3.1 and Figure 3.2 shows the original 50k primal circle dataset and the clean 15k primal circle dataset and their persistence barcodes respectively. The cleaning set is obtained by taking a threshold on the density of the points. We can see that the clean 15k primal circle has a clearer H_1 feature compared with the original 50k primal circle.

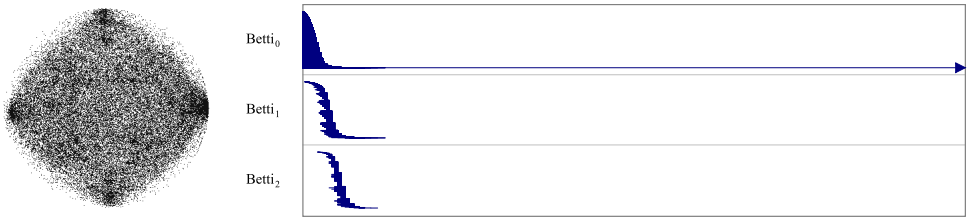

Figure 3.1: The original 50k primal circle dataset and its persistence bar code.

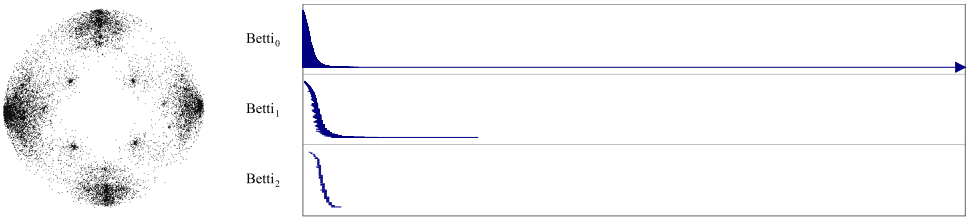

Figure 3.2: The clean 15k primal circle dataset and its persistence bar code.

We want the original 50k dataset to be similar to the clean 15k one shown on the left of 3.2 after denoising. But even after trying many different combinations of $c_{declutter}$ and $c_{resample}$, we didn't get good results. Figure 3.3 shows the denoising output and the persistence barcode of the original 50k primal circle dataset with $c_{declutter} = 2$ and $c_{resample} = 2$. We can't find a clear H_1 feature in the barcode.

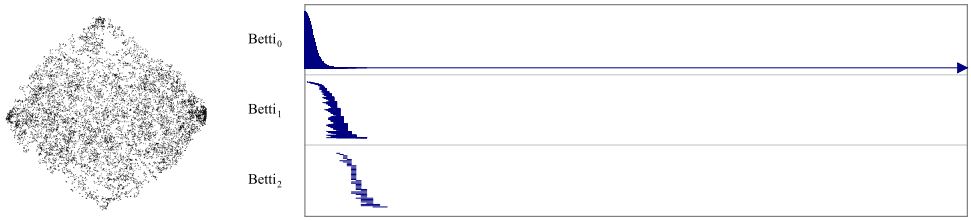

Figure 3.3: Denoising output and the persistence barcode of the original 50k primal circle dataset with $c_{declutter} = 2$ and $c_{resample} = 2$.

I also run the denoising algorithm on the clean 15k dataset. The points along the circle with relatively low density are removed and four clusters remain. The appearance of the H_1 feature is suspended because there are no points between the clusters as it is in Figure 3.2 .

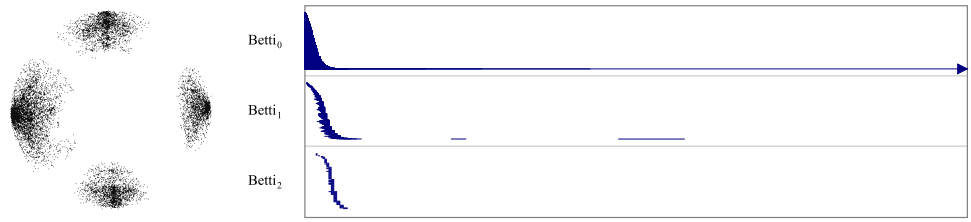

Figure 3.4: Denoising output and the persistence barcode of the clean 15k primal circle dataset with $c_{declutter} = 2$ and $c_{resample} = 4$.

Survivin protein

Table 3.2 shows the preserved distance information after dimension reduction for the survivin protein.

Dataset	Reduce to dimension	Preserved information(%)
Survivin protein	3	99.77

Table 3.2: Distance information after dimension reduction for the two datasets of survivin protein.

Figure 3.5 shows the 250k survivin protein dataset. Every point has a corresponding energy value. We are more interested in the clustered points with low energy but there is a dominated H_0 feature so we can hardly see the details of them. So we simply take the 20% points with lowest energy out, also to avoid the effect from dominated feature we apply PCA only on that 20% points. Figure 3.6 shows the whole picked points on the left and the details of the cluster on the left. Note the left figure in Figure 3.6 is the whole 20%points, the right figure is zooming in of the cluster on the left, so it's just part of the 20% points.

Figure 3.5: 250k survivin protein dataset.

Figure 3.6: Left is 20% points with lowest energy from the 250k survivin protein dataset. Right is the details of the cluster on the left, it's just part of the dataset.

If we compute the persistence of the 20% points with low energy directly, we can hardly see the details of the cluster on the left because of the scale (Shown in Figure 3.7). So our denoising goal here is to remove the noises on the right of the upper figure in Figure 3.6 and keep the cluster on the left.

Figure 3.7: Barcode of survivine protein dataset with 20% lowest energy. The maximum scale of the filtration is 67341.5. Up is drawn with a maximum x axis = 65662. Down is drawn with a maximum x axis = 750.

After denoising, we successfully remove the noises. The output of denoising and the persistence barcode are shown in Figure 3.8. Now we are able to see the details of the features in the cluster.

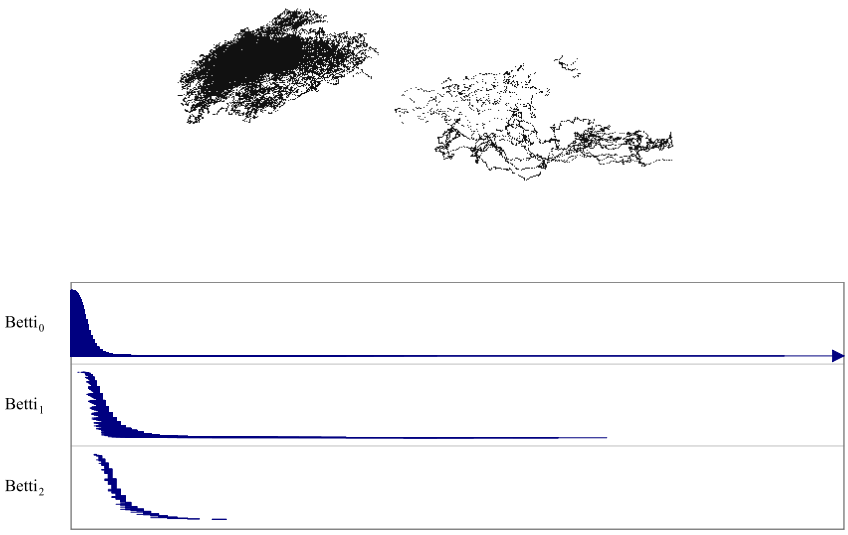

Figure 3.8: Denoising output of 20% points with lowest energy and its persistence barcode.

MNIST

We consider add digits with background image as noises. To simplify the experiment, we just use the dataset of digit 7. Figure 3.9 shows the images of some points in the original MNIST and MNIST-back-image. Figure 3.10 shows the persistence barcode of digit 7s without noises.

Table 3.3 shows the preserved distance information after dimension reduction for the MNIST. Note reducing the dimension is to calculate the persistences faster, the denoising is done in 784 dimension. The PCA is applied to the three datasets seperatelly.

Dataset	Reduce to dimension	Preserved information(%)
Original digit 7	100	94.37
Digit 7 with background noise	100	94.11
After denoising	100	94.46

Table 3.3: Distance information after dimension reduction for the two datasets of primal circle.

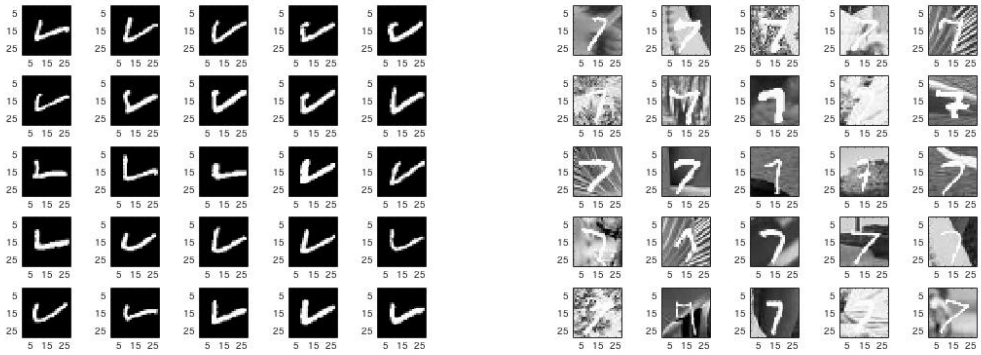

Figure 3.9: Left are digit 7s from the original MNIST. Right are digit 7s from MNIST-back-image.

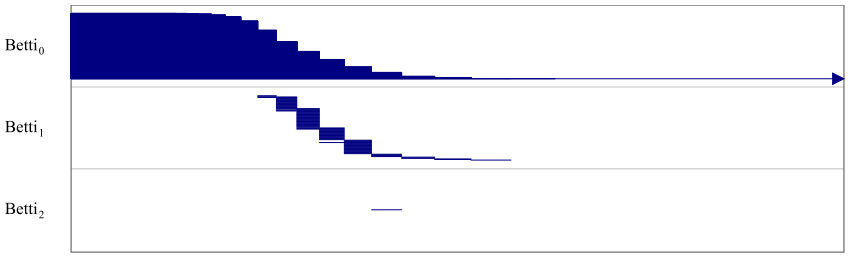

Figure 3.10: Persistence barcode of digit 7s without noises.

We add 250 MNIST-back-image digit 7s to 1279 digit 7s. After denoising, 277 points are removed of which 250 are from MNIST-back-image, which means we remove all noises. The bar in the red square in Figure 3.11 should correspond to the part of the background noises. They are far from each other compared to the original 7s so the H_1 feature appears when the scale is relatively large. Figure 3.12 shows the persistence barcode of digit 7s with background noises after denoising. After denoising, we remove the H_1 feature of the noise.

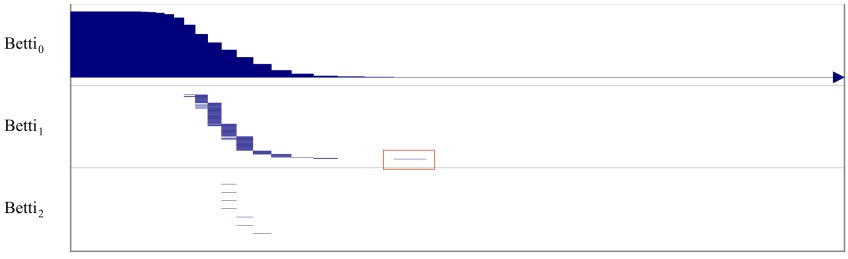

Figure 3.11: Persistence barcode of digit 7s with background noises.

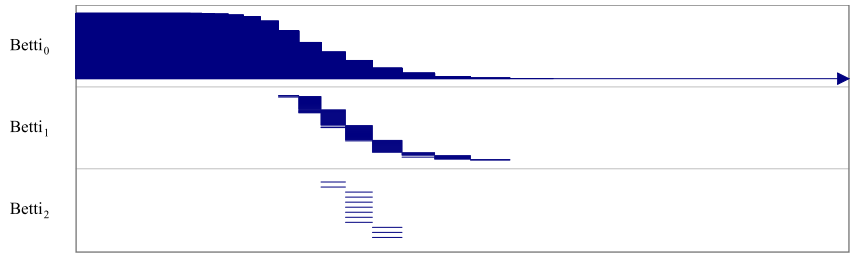

Figure 3.12: Persistence barcode of digit 7s with background noises after denoising

4 Analysis and future works

For the survivin protein, we manage to remove the noises thus reducing the scale of the data which allows us to see the details of the part we are interested in. For MNIST, we also remove all background noises and remove the feature of the noise in the persistence diagram. However, we didn't get good results for the primal circle. This could because PCA distorts the distance too much or the noise goes far away from out assumption of the noise model. So one of the future works is to improve the code and denoise directly in high dimension.

References

[1] Buchet, M., Dey, T. K., Wang, J., Wang, Y. (2015). *Declutter and Resample: Towards parameter free denoising*. arXiv preprint arXiv:1511.05479.

[2] Adams, H., & Carlsson, G. (2009). *On the nonlinear statistics of range image patches*. SIAM Journal on Imaging Sciences, 2(1), 110-117.

[3] Park, I. H., & Li, C. (2010). *Dynamic ligand-induced-fit simulation via enhanced conformational samplings and ensemble dockings: A survivin example.* The Journal of Physical Chemistry B, 114(15), 5144-5153.

[4] *http://yann.lecun.com/exdb/mnist/.*

[5] *http://www.iro.umontreal.ca/ lisa/twiki/bin/view.cgi/Public/MnistVariations*

[6] Tamal K. Dey, Dayu Shi, Yusu Wang. *SimBa: An Efficient Tool for Approximating Rips-filtration Persistence via Simplicial Batch-collapse*. Submitted.

[7] *Scikit-learn: Machine Learning in Python*, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.