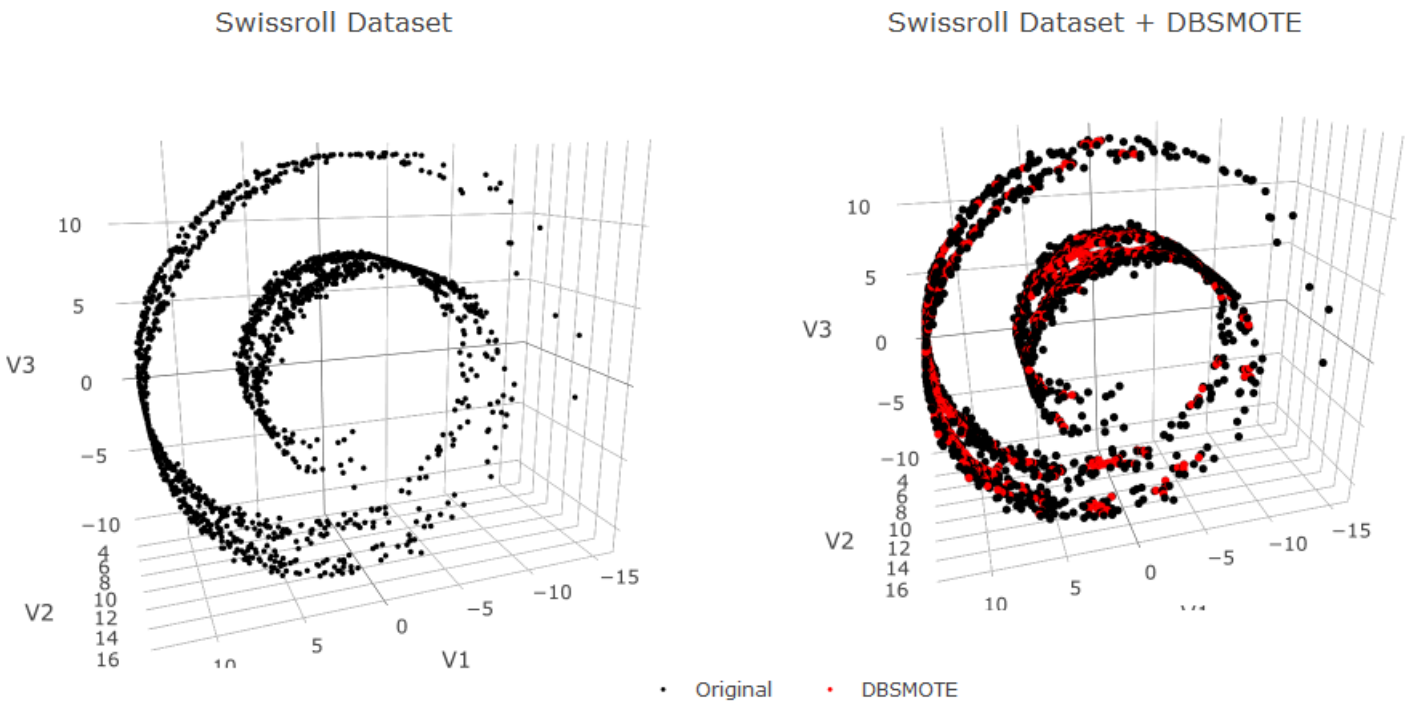
Oct 10, 2018

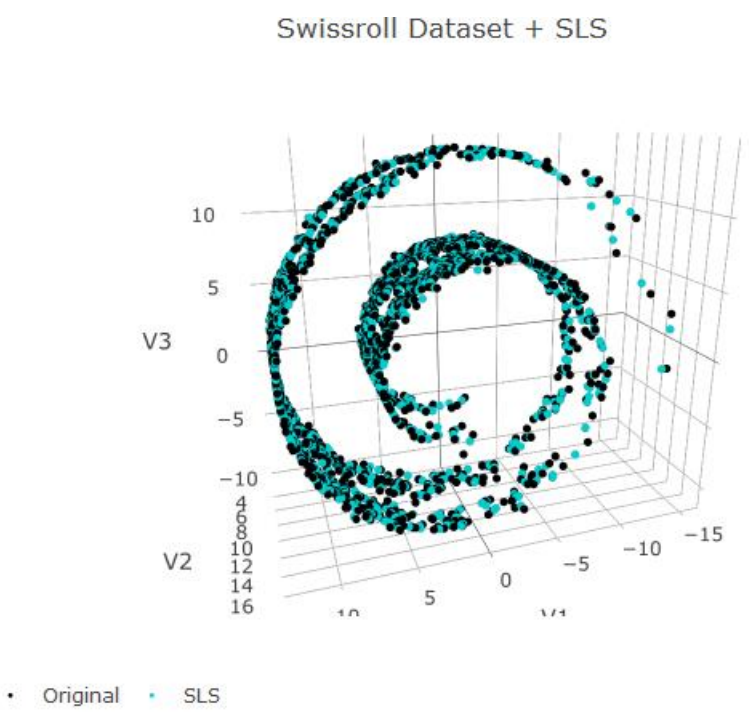
In this experiment,

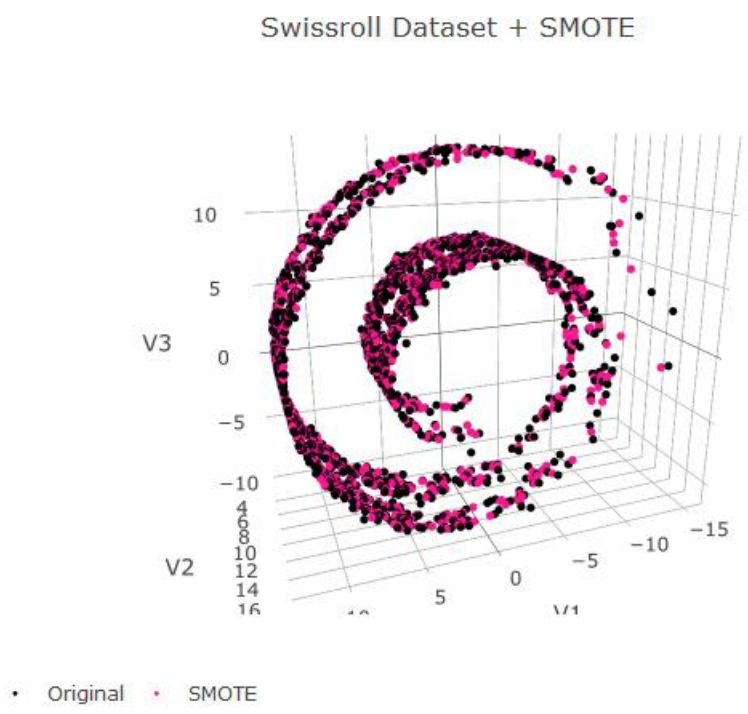
- We have applied oversampling techniques to different well-defined manifold datasets.
- Different oversampling techniques used include Smote, Density Based Smote & Safe Level Smote.
- For all the experiments, the oversampling percentage is kept at 100%, KNN parameter k at 5 and C (KNN in SLS) at 5 (default values)
- Data for swiss-roll manifold can be found at - (http://people.cs.uchicago.edu/~dinoj/manifold/swissroll.dat)
- Other manifolds used are generated using well defined equations. These include sphere, S-curve, toroidal helix etc.

Below are the Visualizations exported in 2D format –

1.

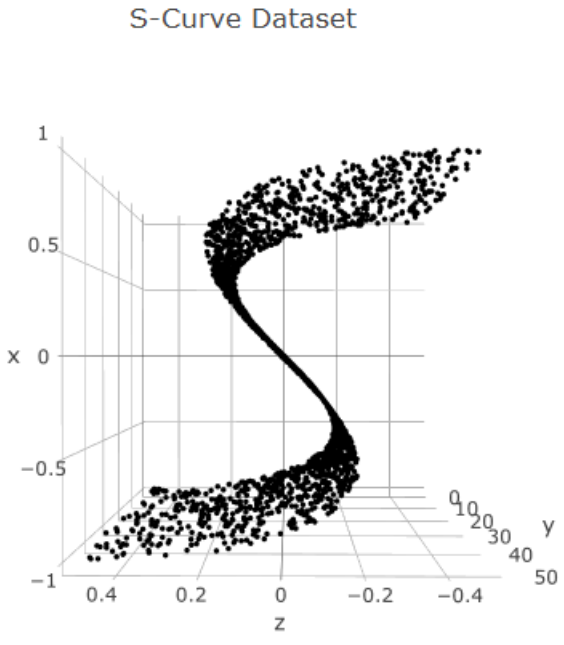


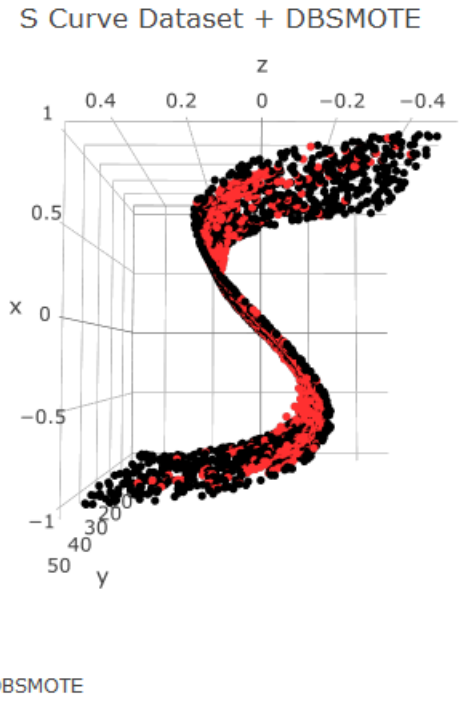


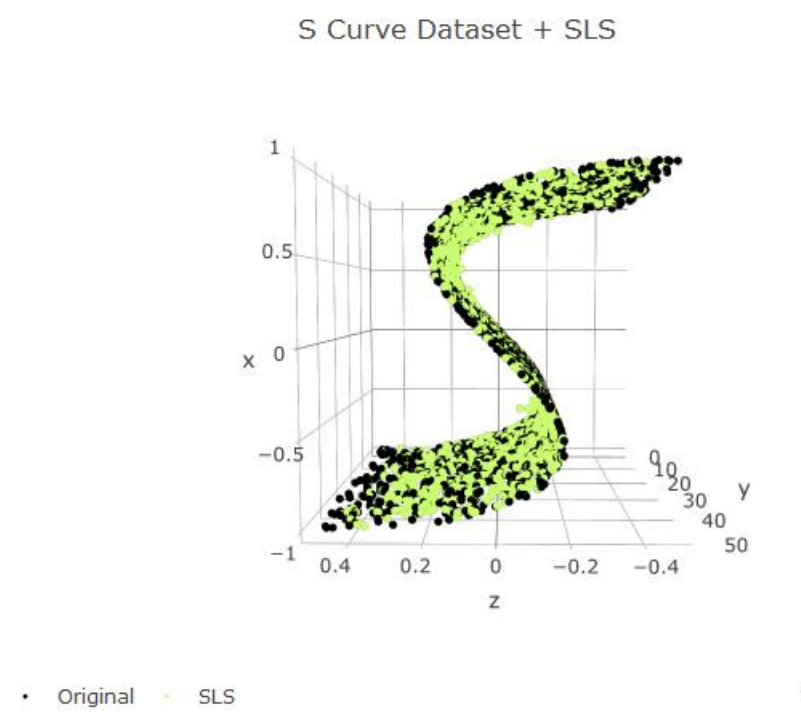


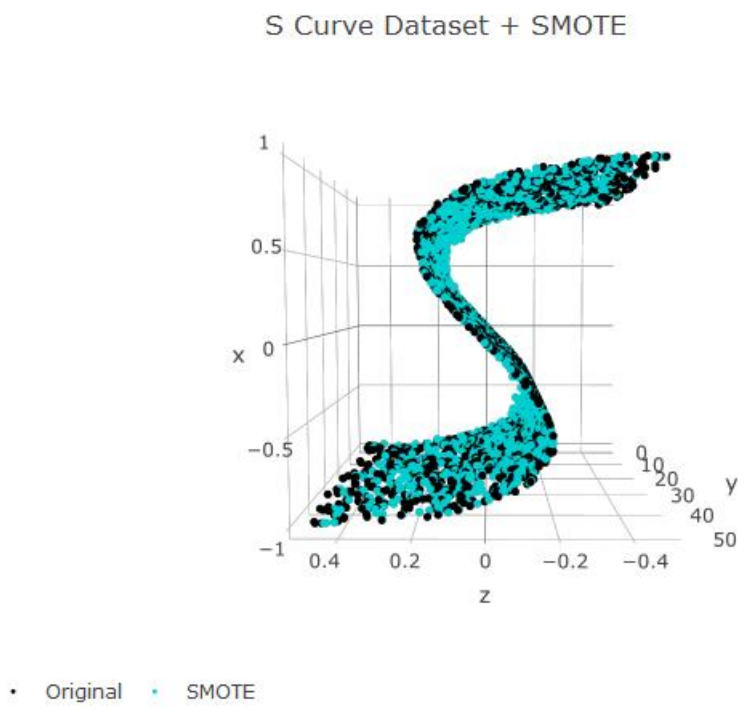
- For the swiss roll dataset, it was observed that all the oversampling techniques preserve the overall shape of the manifold
- No distortions are introduced by synthetic points as all the synthetic points lie well near the original surface
- No synthetic point lies in the empty space between the manifold surfaces

2. S-Curve Data



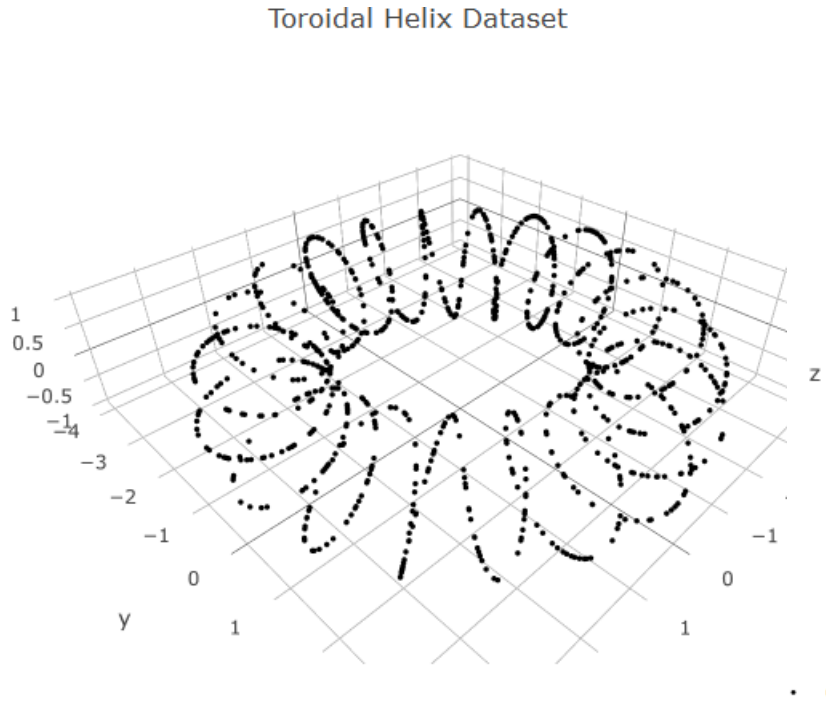


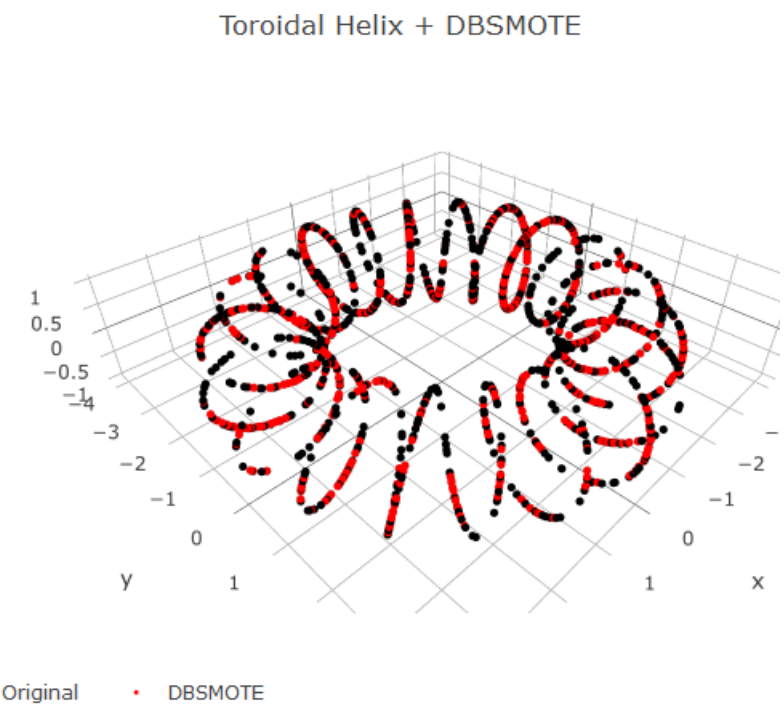


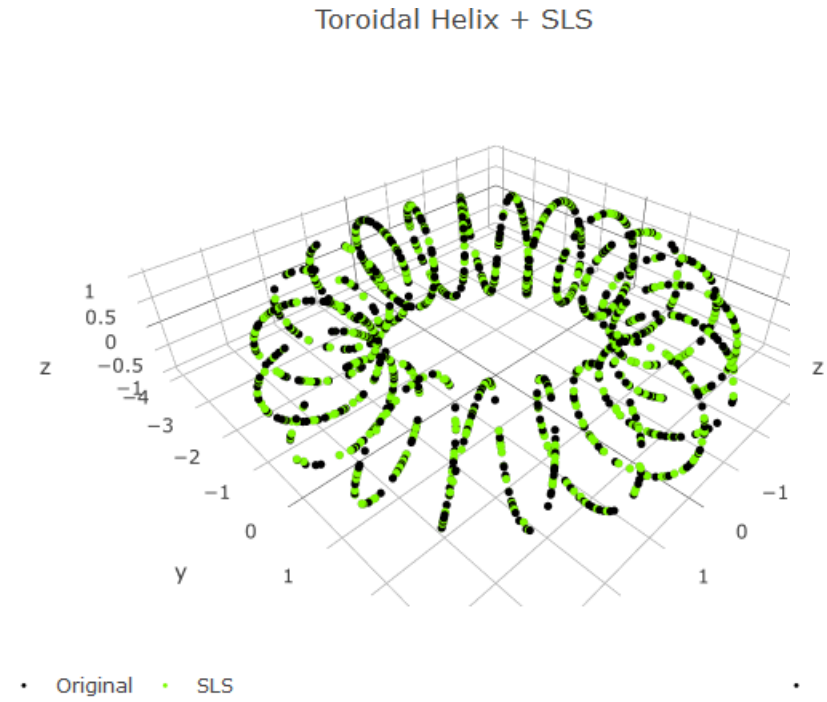


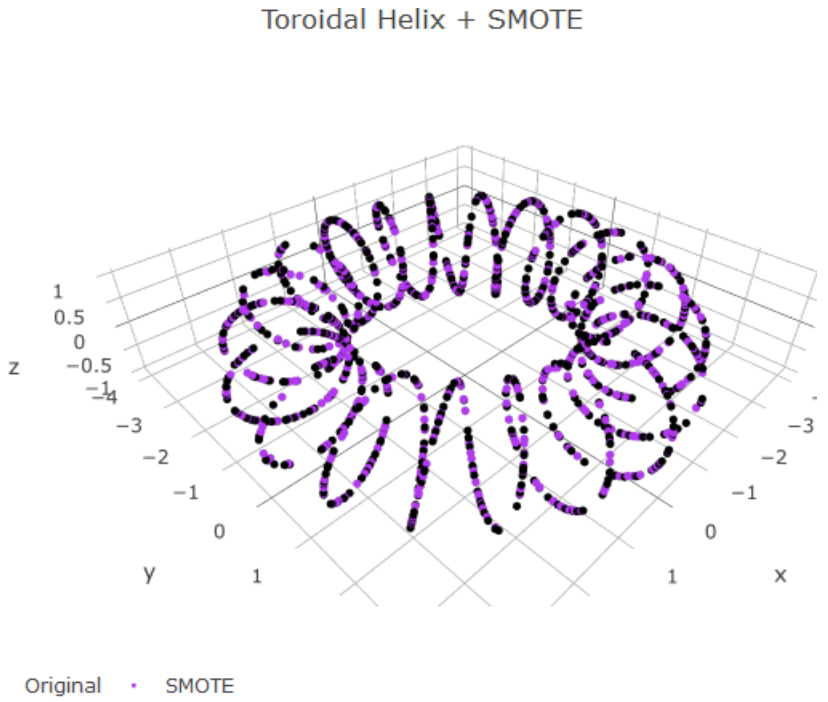
- For S-Curve Manifold it was observed that all oversampling techniques preserve the overall shape of the manifold and no distortions were introduced to the manifold

3. Toroidal Helix



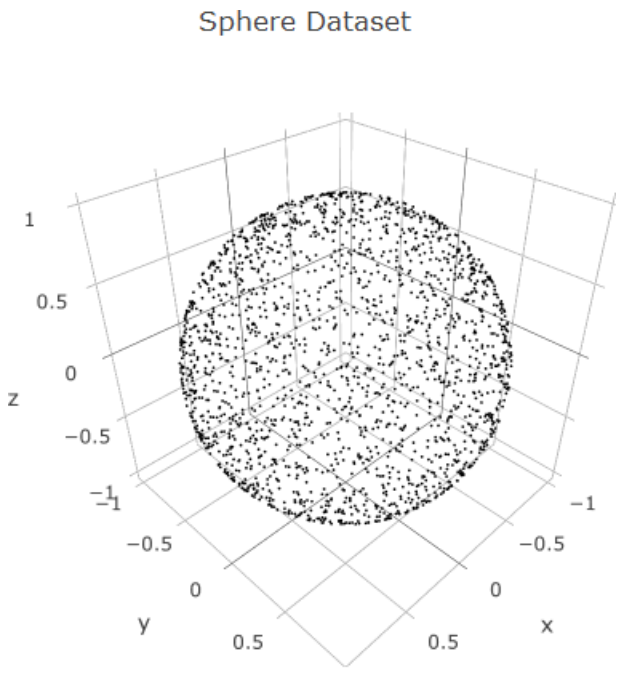


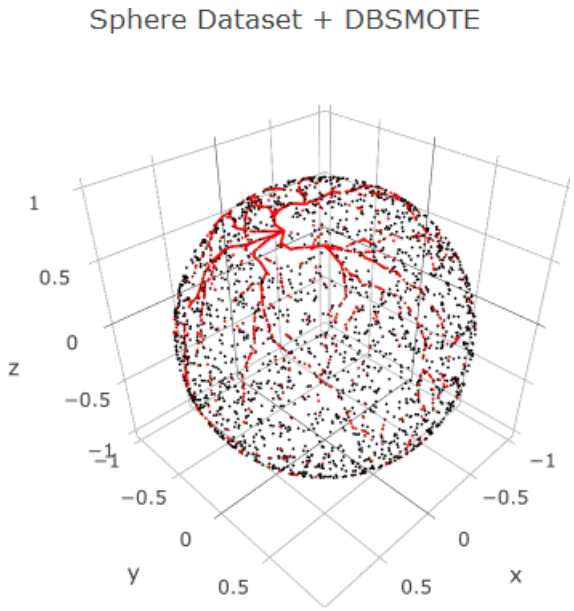


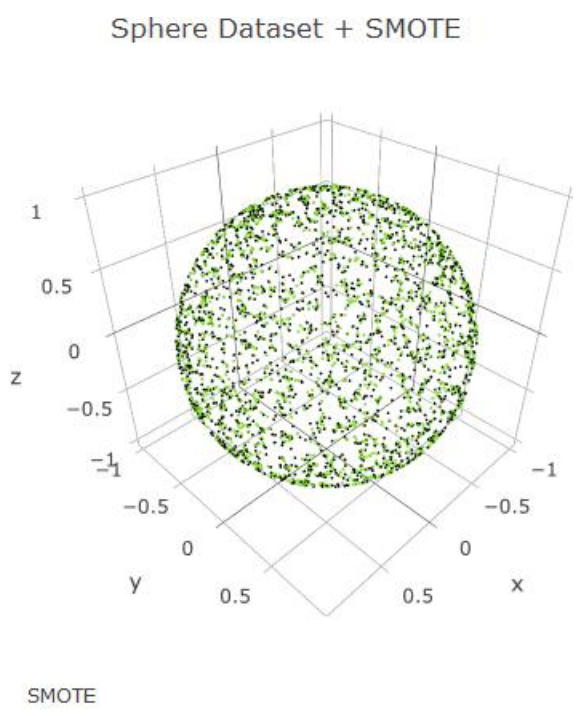


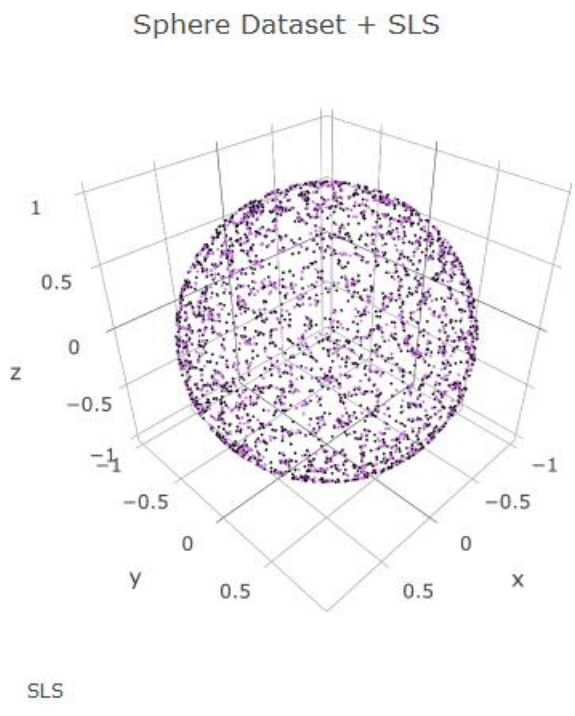
- For toroidal helix dataset it was observed that no distortions were introduced and synthetic points maintain the overall shape of the manifold.

4. Sphere data





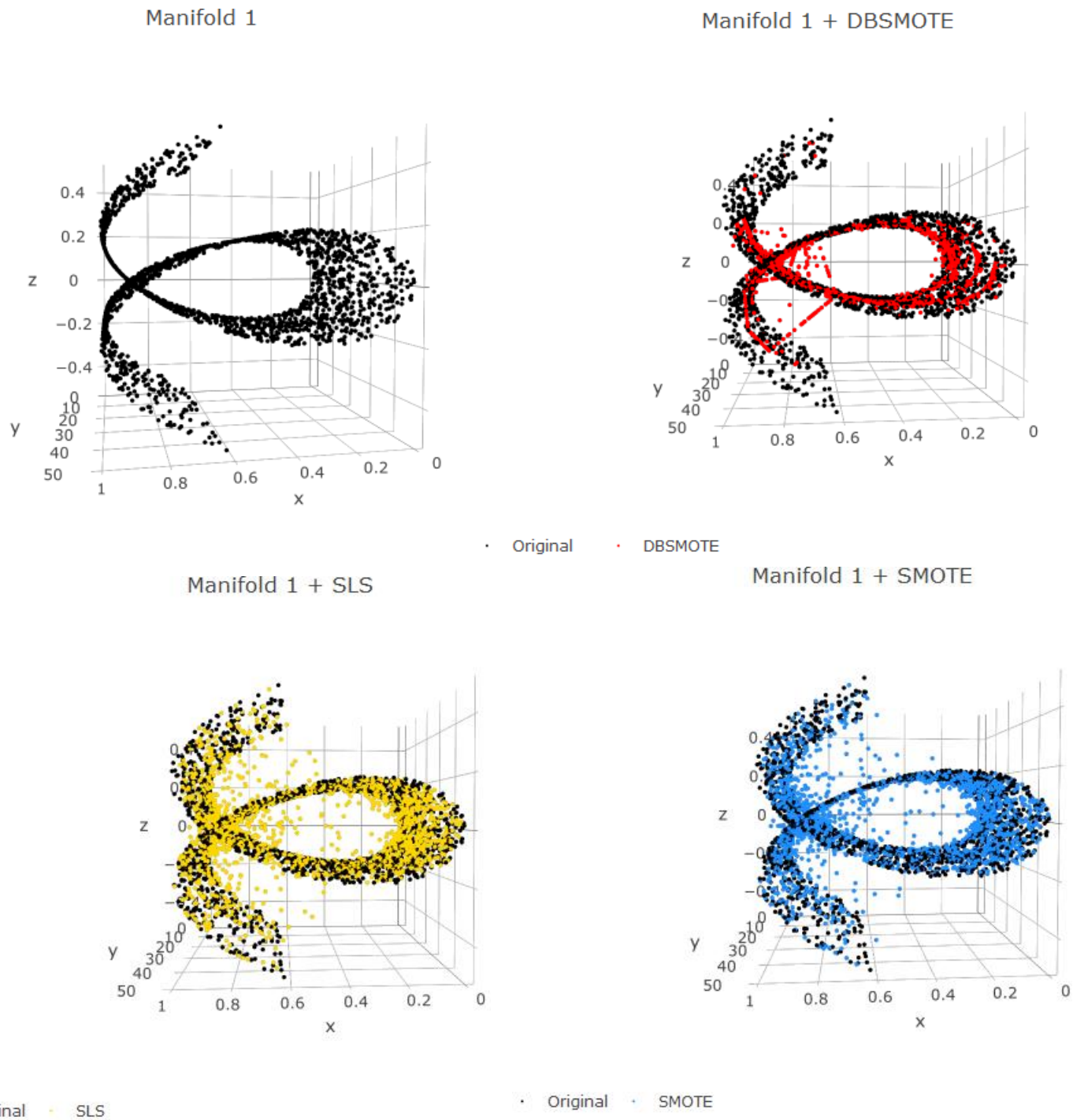




- For (unit) sphere data, the black points represent the original data and the colored points are the synthetic points generated using oversampling.
- As it is hard to examine if the synthetic points are generated inside or outside the sphere, we can numerically check if the euclidian distance of synthetic point is less than, greater than and equal to the radius of the sphere (=1).
- It was observed that for all oversampling techniques, no sythetic points have radius =1 and there is no sythetic point with radius more than 1 and less than 0.99.
- This implies that all synthetic points are generated below the surface of the manifold and are no deeper than 1% unit length from the surface (i.e. all points lie between .99 and 1 from the centre of the sphere)
- Diagonostic in R-

```
sum(sqrt(syn$x*syn$x + syn$y*syn$y + syn$z*syn$z)>1) = 0
sum(sqrt(syn$x*syn$x + syn$y*syn$y + syn$z*syn$z)==1) = 0
sum(sqrt(syn$x*syn$x + syn$y*syn$y + syn$z*syn$z)<0.99) = 0
```

5. **Other Intricate manifolds** –



- For this manifold, it was observed that the ovesampling techniques produces sythetic points that although follow the overall shape of the surface but also introduces a lot of noise.
- All types of oversamping produced many data points that lie in the voids of the manifold.

Thanks.

Falak