tags: Data Science

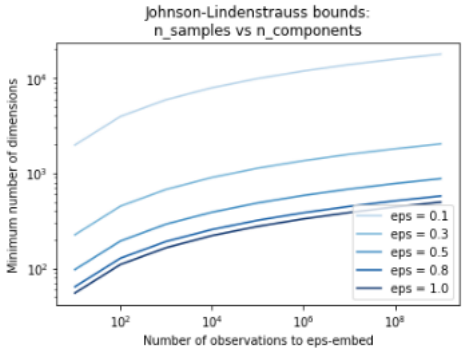
Project 2

408410042 林靖紳

Experimental Results

Experiment 1

```
Out[8]: <Figure size 432x288 with 0 Axes>
```



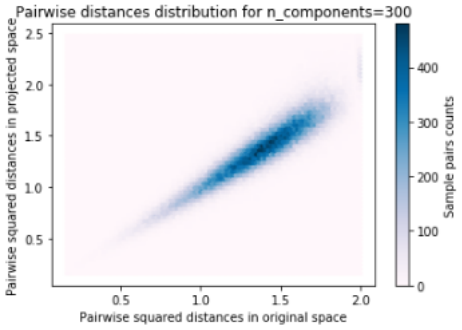

```
<Figure size 432x288 with 0 Axes>
```

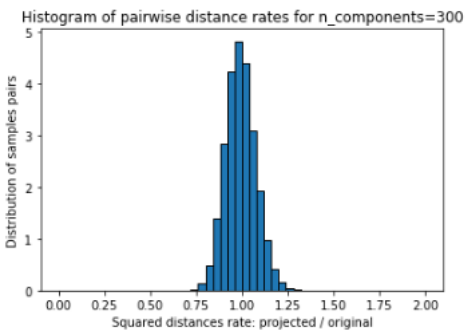
Experiment 2

- 再最後一個 跳出 memory error，應該是資料太大，分配的記憶體不夠

```
Embedding 500 samples with dim 130107 using various random projections
Projected 500 samples from 130107 to 300 in 1.747s
Mean distances rate: 0.99 (0.08)
Projected 500 samples from 130107 to 1000 in 7.546s
Mean distances rate: 1.01 (0.05)

---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
<ipython-input-13-bbb194d6fa64> in <module>
```

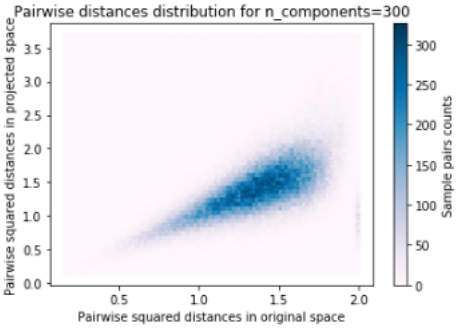


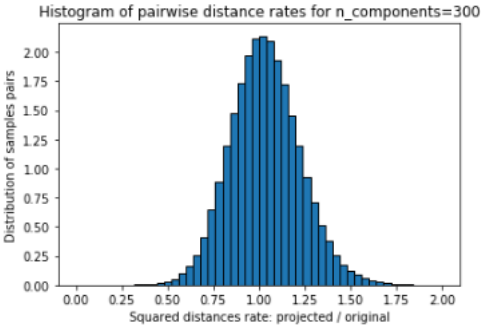


Experiment 3

- 明顯的比 2 快上許多

```
Embedding 500 samples with dim 130107 using various random projections
Projected 500 samples from 130107 to 300 in 0.246s
Mean distances rate: 1.03 (0.19)
Projected 500 samples from 130107 to 1000 in 0.712s
Mean distances rate: 0.96 (0.09)
Projected 500 samples from 130107 to 10000 in 6.526s
Mean distances rate: 1.00 (0.03)
```





Report

1. Explain what you observed from the figure generated from Experiment 1 with distortion bound ϵ = 0.1, 0.3, 0.5, 0.8, 1.0 , respectively. (eps =ϵ) (10%)

- 從實驗1生成的圖中觀察到，當 eps 增加時，原始數據和經過扭曲後的數據之間的平均距離增加，這意味著數據點之間的差異變得更大。
- 隨著 eps 的增加，實現所需的扭曲程度的 cluster 的數量也增加。這是因為當 eps 增加時，被允許離其分配的 cluster 中心更遠的點更多，因此需要形成更多的 cluster 以保持所需的扭曲水平。
- 對於給定的數據集，似乎選擇 ϵ = 0.5或0.8是一個不錯的選擇，它提供了所需 cluster 的數量和扭曲水平之間合理的折衷。選擇過小的ϵ值將需要大量的 cluster ，而選擇過大的ϵ值將導致高扭曲水平。

2. Explain what you observed from Experiment 2.

- 使用高斯隨機投影進行降維的方法可以有效地降低資料的維度，從而減少distortion的計算成本。

- 在實驗中，隨著投影後的維度減少，distortion的計算時間也會隨之減少，但歸納效果可能會有所下降。
- 在投影後的維度超過資料點的實際維度時，歸納效果會明顯下降，這是因為在這種情況下，資料丟失了部分有效資訊。

3. Compare the results of Experiment 2 and Experiment 3 with two properties given in lec12_v1 pp.17, and explain the difference

1. $\|E[\Phi(\mathbf{u}_i)]\|_2^2 = \|\mathbf{u}_i\|_2^2$ with high probability (whp)

2. Var $(\|[\Phi(\mathbf{u}_i)]\|_2^2 - \|\mathbf{u}_i\|_2^2)$ is small enough to hold the inequality (whp)

- 實驗2 和 實驗3 的結果均顯示出，對於所有的k，1. 皆成立。這表明了這兩種投影方法在保留向量的範數上是一致的，都具有良好的保留性。
- 然而， 實驗3 的結果顯示，2. 比 實驗2 的要小得多。這表明稀疏隨機投影比高斯隨機投影更能夠在保留向量之間的距離上保持較低的變異性。因此，在某些應用中，稀疏隨機投影可能是更好的選擇，因為它可以更好地保留向量之間的相對距離。