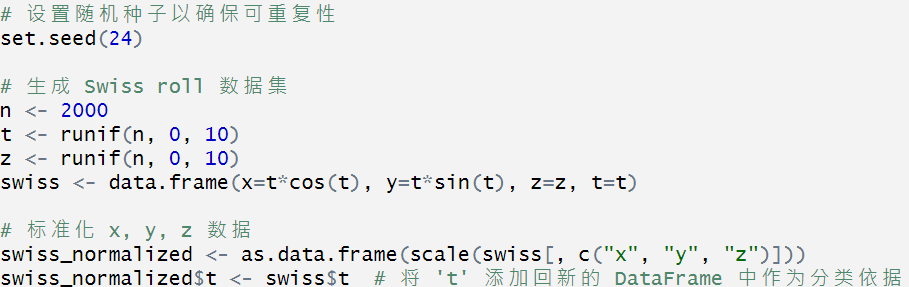
* 資料介紹

1. 生成方式

* t：從0到10的均勻分佈生成。在Swiss roll結構中，𝑡 代表從捲起原點出發沿螺旋結構的距離或者 "長度"。t 的值決定了點在螺旋上的位置，並通過cos(t)和sin(t)影響每個數據點的x和y座標。
* x 和 y 的計算：x=t⋅cos(t), y=t⋅sin(t)

這種計算方式使得數據點按照螺旋的方式排列，隨著 t 增加，每個點圍繞中心向外擴展，形成一個漸進捲起的結構。

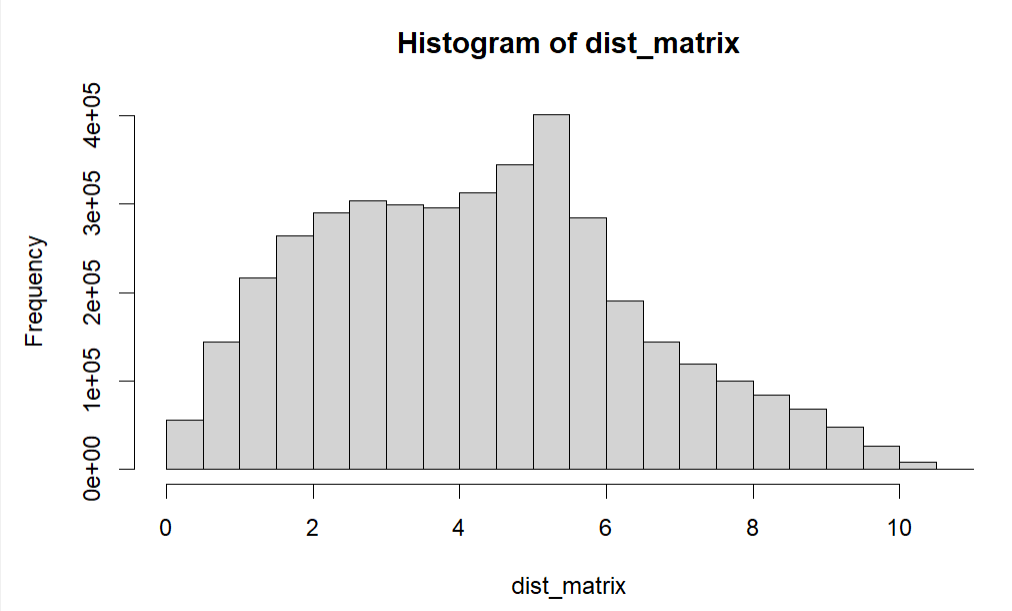
* 參數 z：z也是從0到10的均勻分佈生成的，它代表點在垂直於捲曲平面的方向上的位置。這使得 Swiss roll 有了第三維，增加了數據集的複雜性。
* 將以上基礎建構出的資料集標準化，便是此次分析所使用的資料



1. 距離矩陣

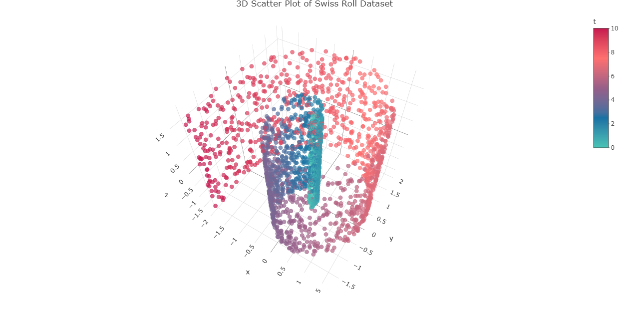
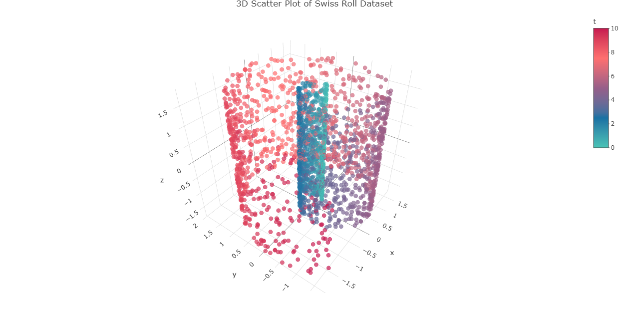
觀察標準化後的資料之距離矩陣，可以看出絕大多數的點距離介於1~6中間，在使用 ”enn” 方法轉換資料時，或許介於此中間的是較為妥當的。

由於kernal-weight方法也有牽涉到距離矩陣，這個直方圖也可以給予其參數的相關資訊。

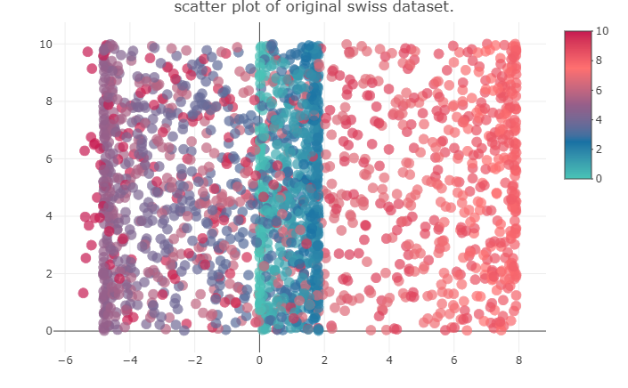
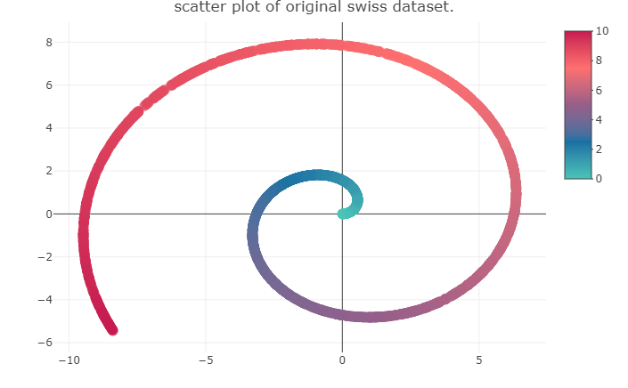


1. 立體圖與二維圖

* 3d：資料呈現捲狀從中心向外旋轉，由於資料建構基礎是以參數t為基準，所以我們將根據參數t的大小作顏色漸變，方便觀察後續的分析結果。

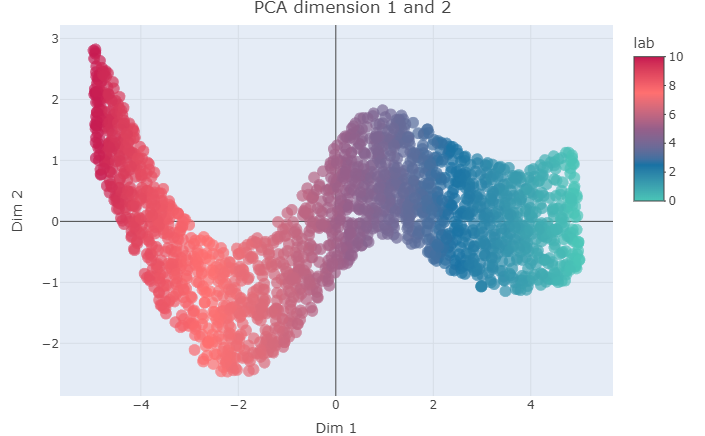


* 2d：左方圖為固定z軸，右方圖為固定y軸，可以看出顏色變化的位置以x,y軸上可觀測到的為主，故降維觀測後的顏色變化應該要與下左圖類似。



* 分析

1. PCA



由上圖我們可以看出：

1. 形狀和結構：

從三維空間中的“瑞士捲”形狀降維到二維空間中，我們可以看到數據呈現出一種獨特的卷曲結構。這個結構由t變量決定，它在這裡作為色彩編碼的依據，所以我們可以看到x和y如何隨著t（從紅色到藍色）的變化而變化。

2. 顏色和t變量的關係：

色彩從深紅色平滑過渡到深藍色，這表明t是一個連續變量。在圖中，t的低值對應於紅色，而高值對應於藍色，顏色的連續變化描繪出t值的變化是平滑的。

3. 主成分解釋的變異：

a. Dim 1和Dim 2：在圖表的軸標題中，Dim 1和Dim 2分別對應於PCA分析中提取的第一和第二主成分。這些軸代表數據變異最大的方向。對比原瑞士捲的2d圖像，Dim 1軸應該是捕捉到將瑞士捲長度變異，而Dim 2則捕捉瑞士捲厚度變異。

b. 變異量：數據點在Dim 1和Dim 2上的廣泛分佈表明這兩個主成分解釋了數據的一大部分變異(也可以由圖二佐證，兩維度的主成分已經足夠解釋此筆資料)。

4. 分布密度：

圖中數據點的密集區域可能表明t值在這些範圍內變化不大，或者這些t值的樣本數較多。反之，數據點較少的區域可能對應於t值的快速變化，或者這些t值的樣本數較少。

1. Laplacian Eigenmap

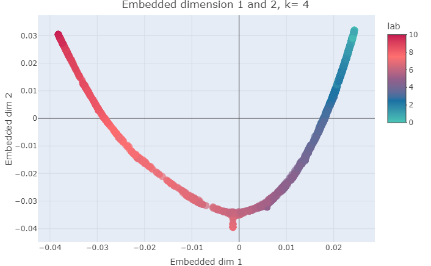
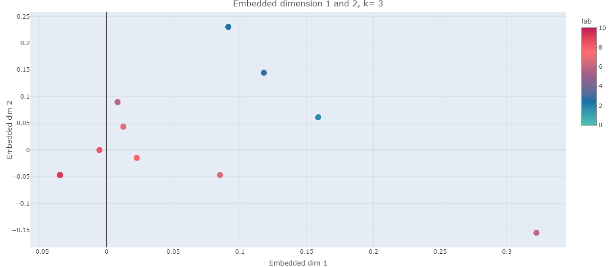
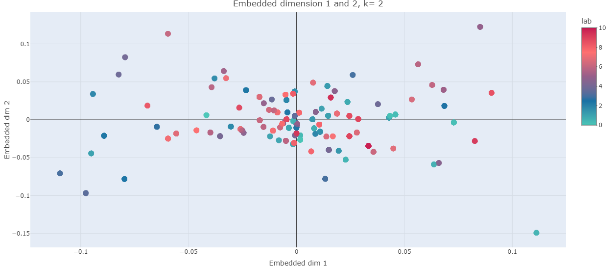
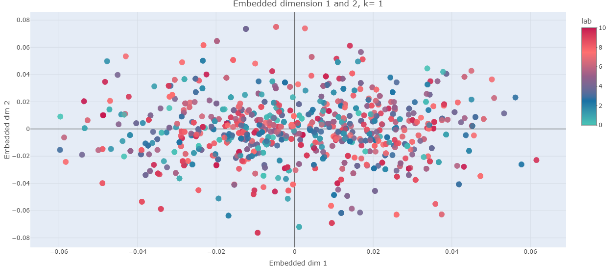
制定實驗計劃如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 實驗次數 | 是否採用熱核 | 建圖方法 | 觀察的維度 |
| 1 | 否 | knn | Dim1, Dim2 |
| 2 | 否 | knn | 比較維度 |
| 3 | 是 | knn | Dim1, Dim2 |
| 4 | 是 | knn | 比較維度 |
| 5 | 否 | enn | Dim1, Dim2 |
| 6 | 否 | enn | 比較維度 |
| 7 | 是 | enn | Dim1, Dim2 |
| 8 | 是 | enn | 比較維度 |

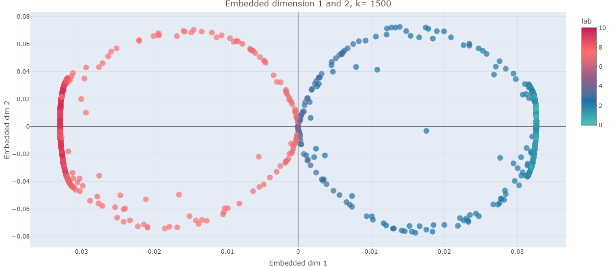
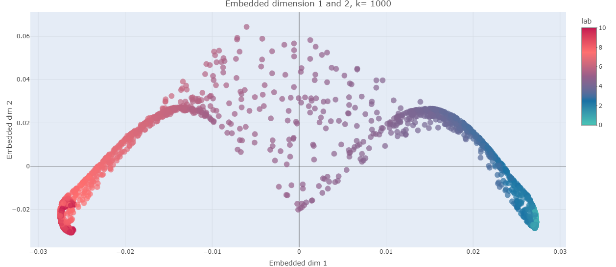
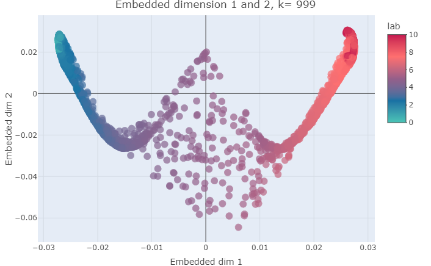
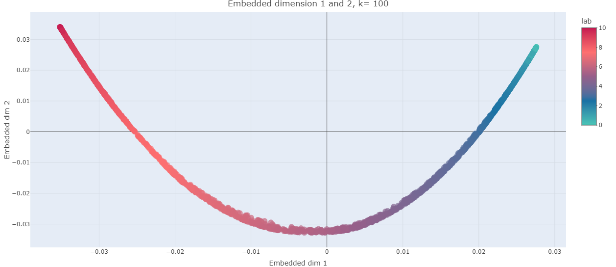
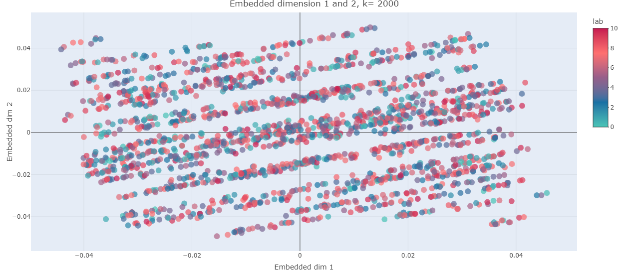
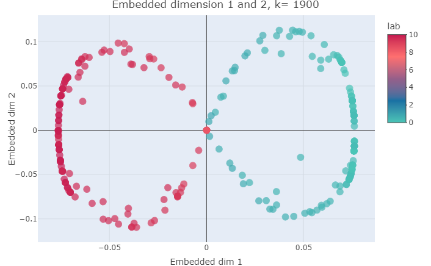
1. Knn without weighted (dim1&2)

嘗試 k = 1, 2, 3, 5, 10, 100, 1000, 1500, 2000：

* k=1 ~3時，很容易出現點重合在一起的情況。因為當2個或3個點是彼此的k近鄰，把它們映射到同一個點就可以完美地保持局部平滑，但當k大於等於4後此情況消失，代表當具有四個比較點時，轉換更能夠保留點與點的相對距離關係且同時保留局部平滑性質。



* 由於 Swiss data為全連通圖型，且資料點共2000個，當 k 具有一定大小內，都還是仍能夠在轉換過後維持原先的顏色變化 (即相近點保持原有的相近性質) ，由下右圖可以看出當近鄰點小於半數時，轉換大致上仍能夠保留資料內部與外圍原有的位置。
* 但當k值接近半數時，每個點的局部結構不再局限於真正的地理鄰近點，而是幾乎包括了半數的點，可能會破壞螺旋的連續性，尤其是對於處於螺旋中心的點，可以明顯看出當 k=999時，中心的點已經無法呈現自身真正的鄰近位置。

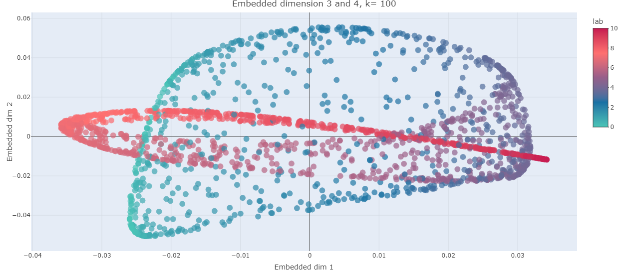
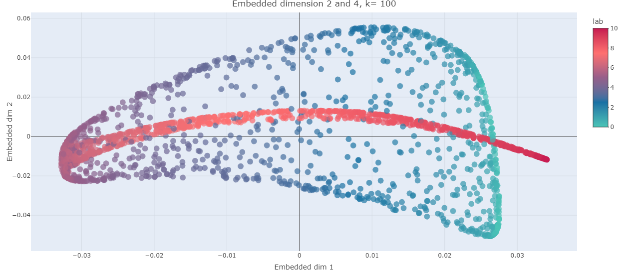
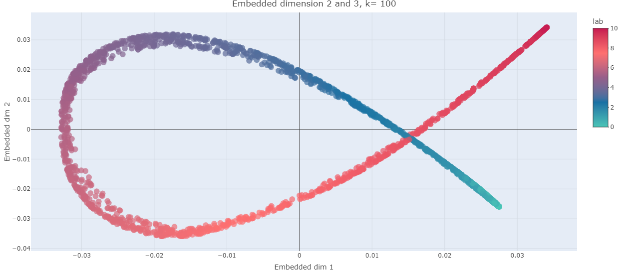
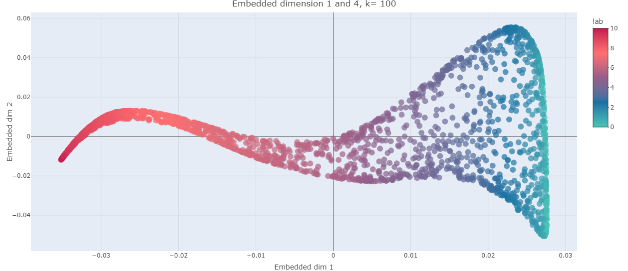
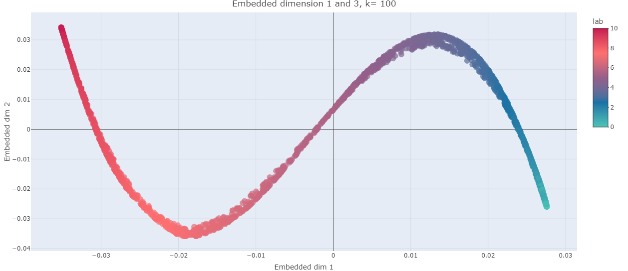
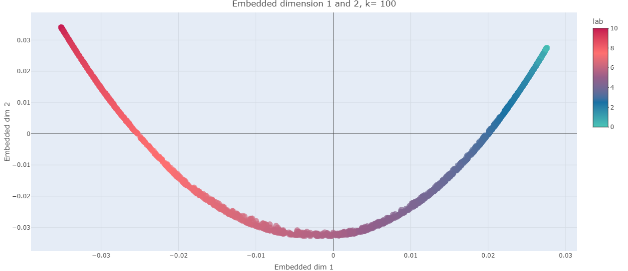
 

* 由上圖可以看出，當 k達到 1500時，螺旋中心的紫色點徹底消失，大k值導致降維後的數據點過於集中，不同類別或群組之間的區分度變差，僅有距離較遠的紅橘色群與藍青色群被分別，此情況一直維持到k<2000，當k=2000時徹底失去點之間的相鄰關係。

1. Knn without weighted (dim)

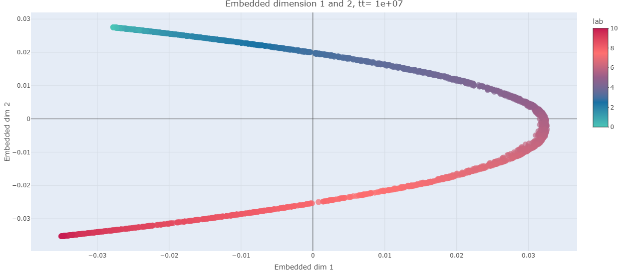
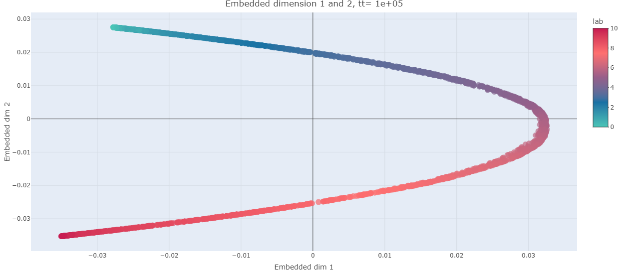
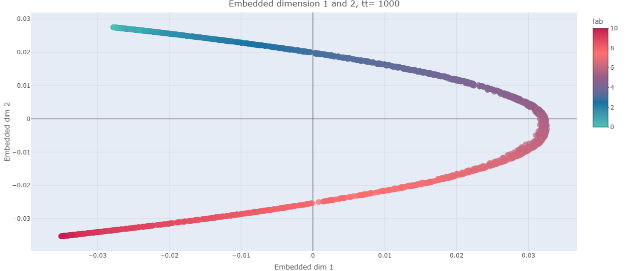
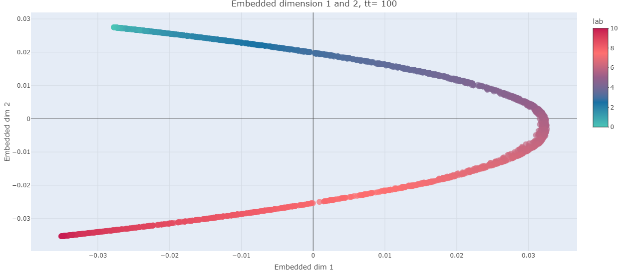
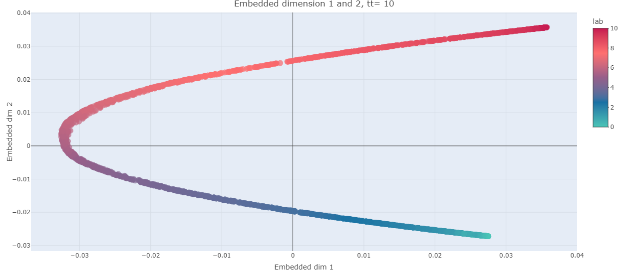
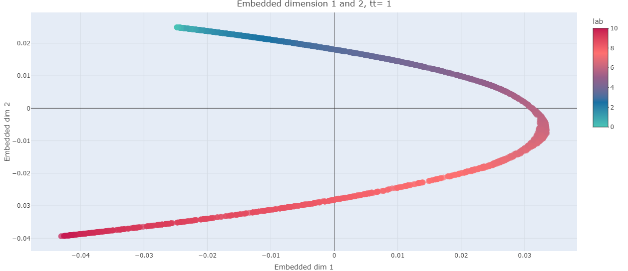
根據上點，我們選擇k=100作為合適的參數大小，並比較不同

可以看出第一維度(顏色變化???)、第二維度(



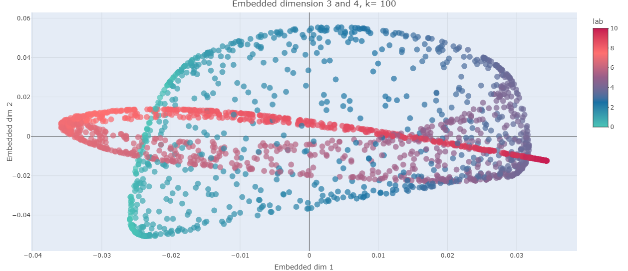
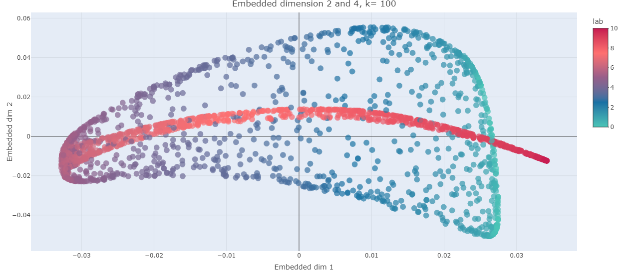
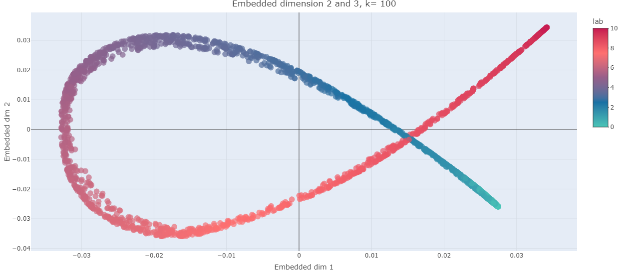
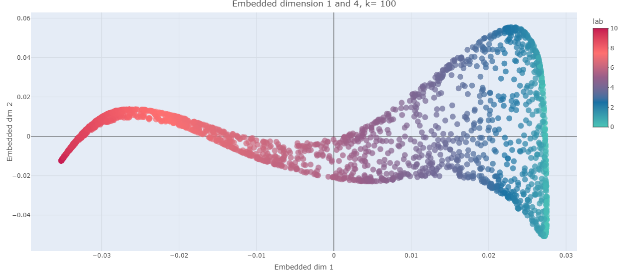
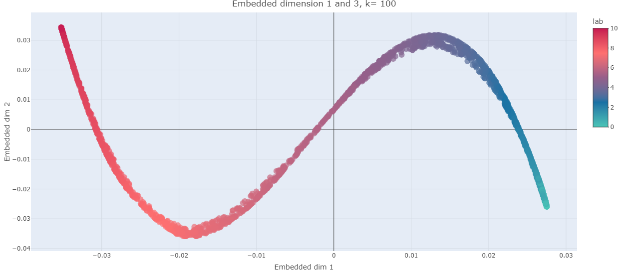
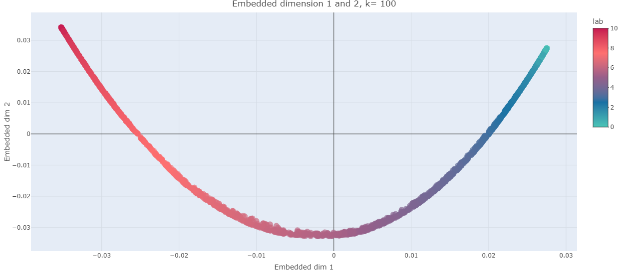
1. Knn with weighted(dim1&2)

固定 k = 100，嘗試tt= 1, 10, 100, 1000, 1e5, 1e7，無明顯差異，可知全連通圖型若是已選擇有效的k值，weighted本身影響不大。



1. Knn with weighted(dim)

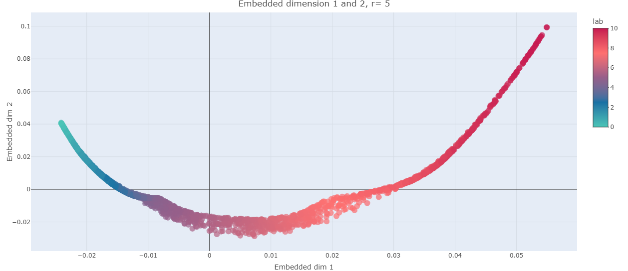
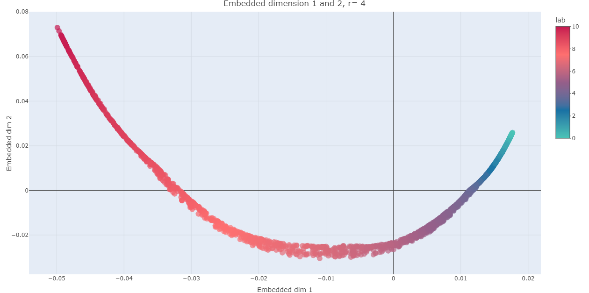
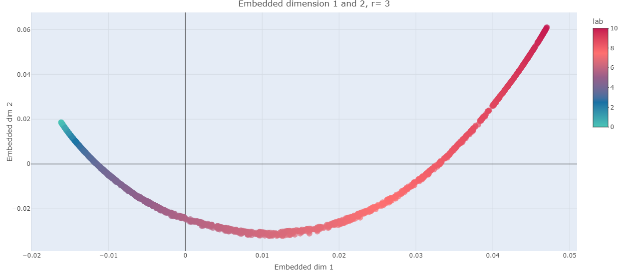
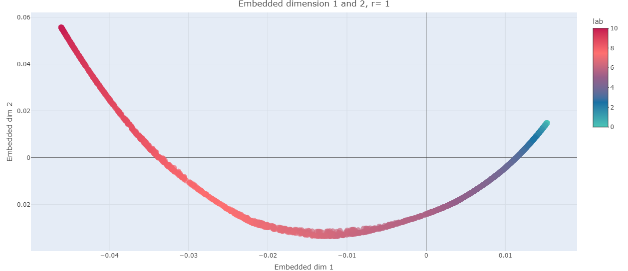
固定 k = 100，tt=100，嘗試觀察不同維度之結果(與無weighted版本相同)

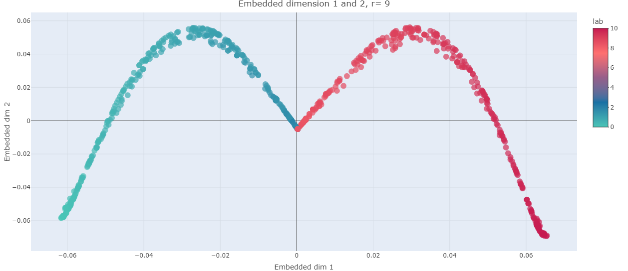
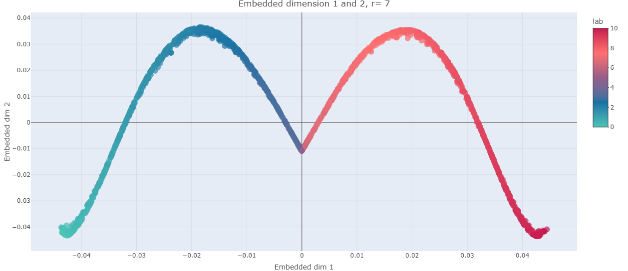


1. enn without weighted(dim12)

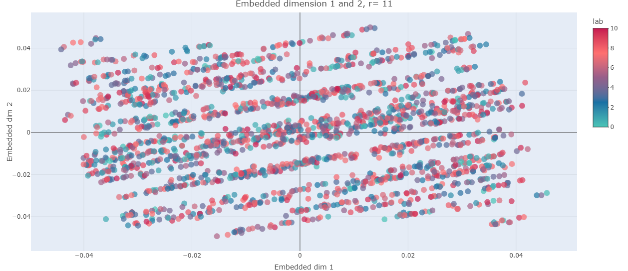
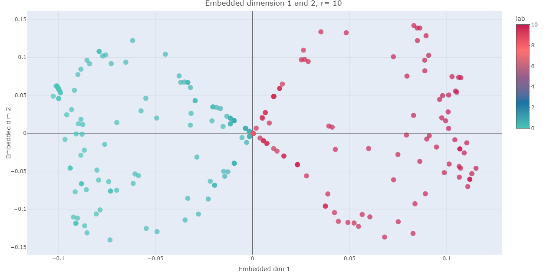
嘗試 r= 1, 3, 4, 5, 7, 9, 10, 11

* 可以看出由於資料本身距離大多數小於7，所以參數小於此的資料降維結果較能夠保持原本不同顏色間的點之鄰近關係
* 但當參數大於4時，已經開始無法完美捕捉不同顏色之間的相對關係



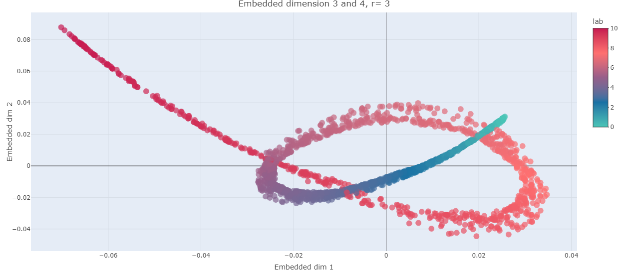
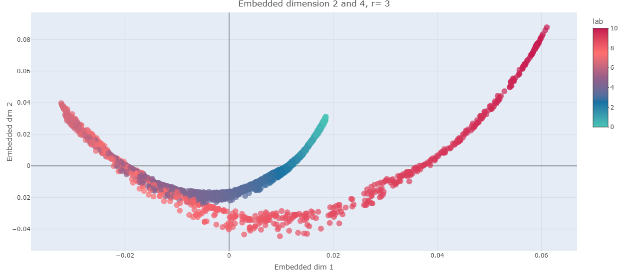
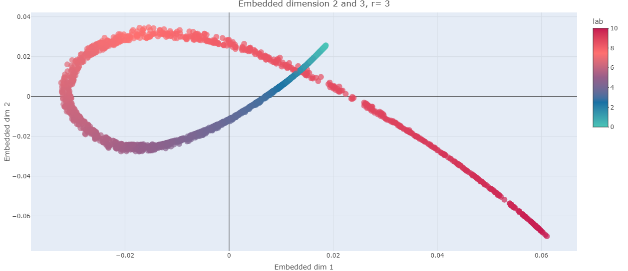
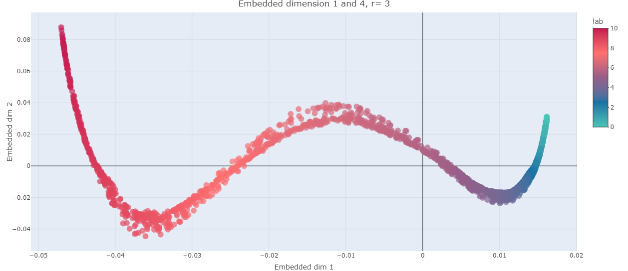
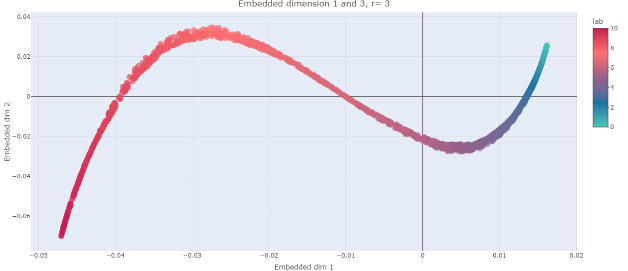
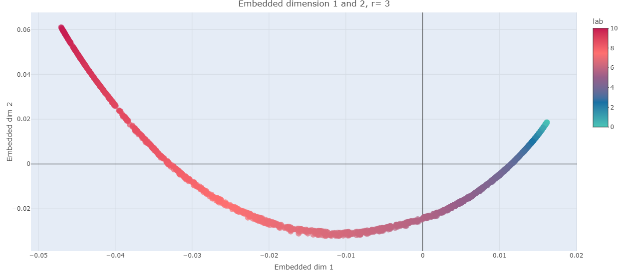


* 而當參數等於資料本身最遠距離時(10)，僅能夠捕捉相對距離最遠的兩種顏色的資料(深紅與淺藍)
* 超過最遠距離時，無法將不同色塊的資料點完美分隔



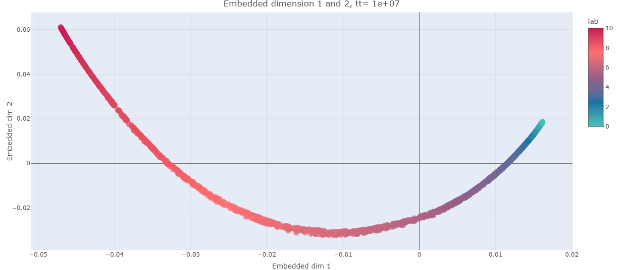
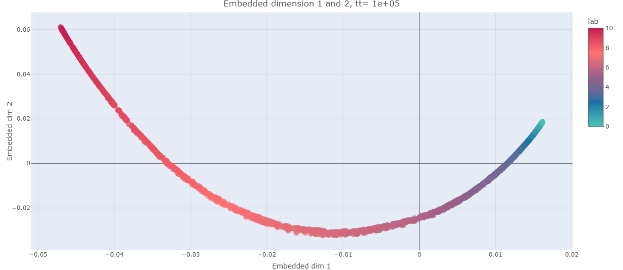
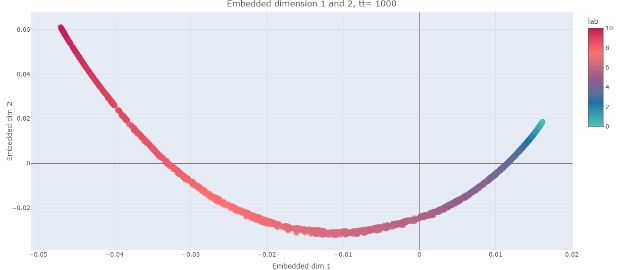
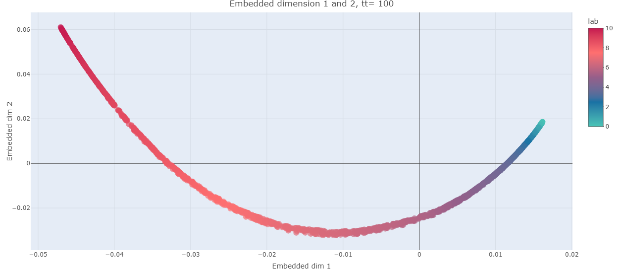
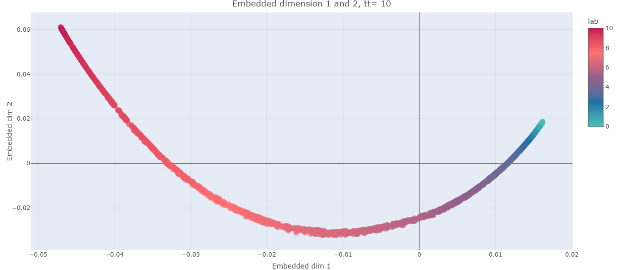
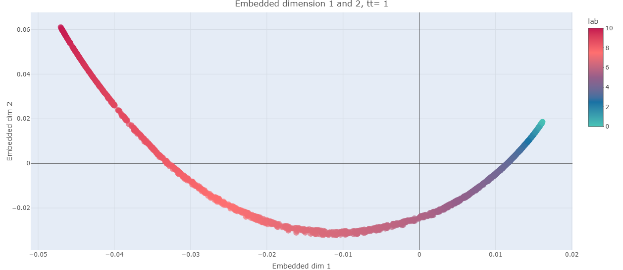
1. enn without weighted(dim)

根據前述結果，我們固定 r=3，觀察不同維度的結果



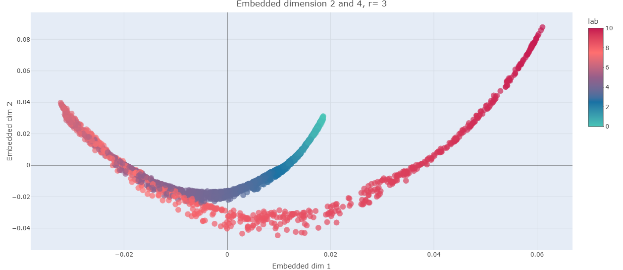
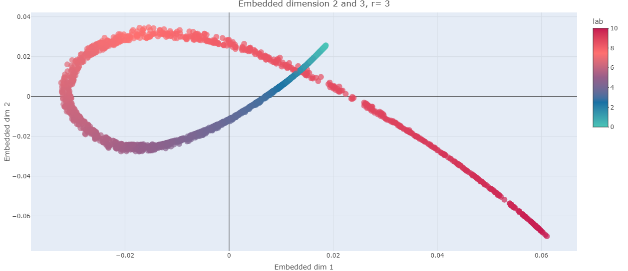
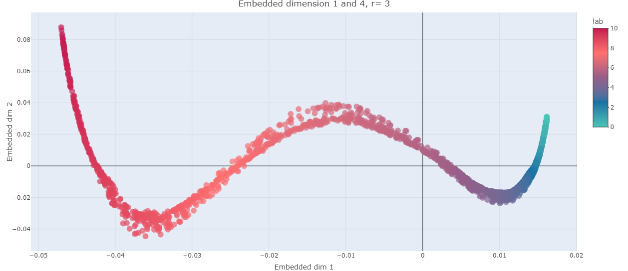
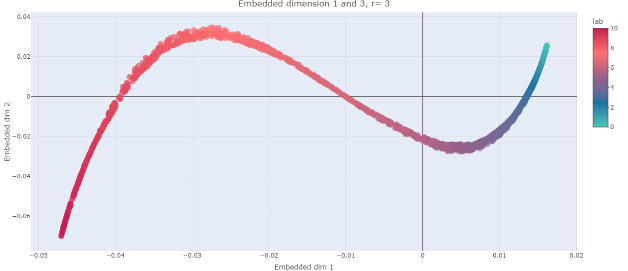
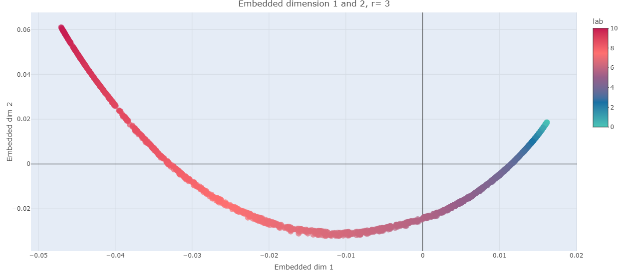
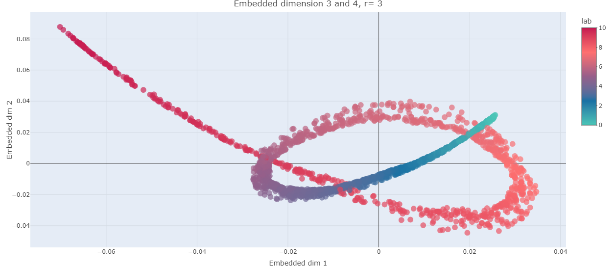
1. enn with weighted(dim12)

根據前述結果，我們固定 r=3，且嘗試不同 tt= 1, 10, 100, 1000, 1e5, 1e7下的kernal-weight是否會出不同結果，但由下圖可知，全連通圖若是已選擇有效的r值，並不會改變結果



1. enn with weighted(dim)

根據前述結果，我們固定 r=3，tt=100，觀察不同維度的結果

* 資料介紹

