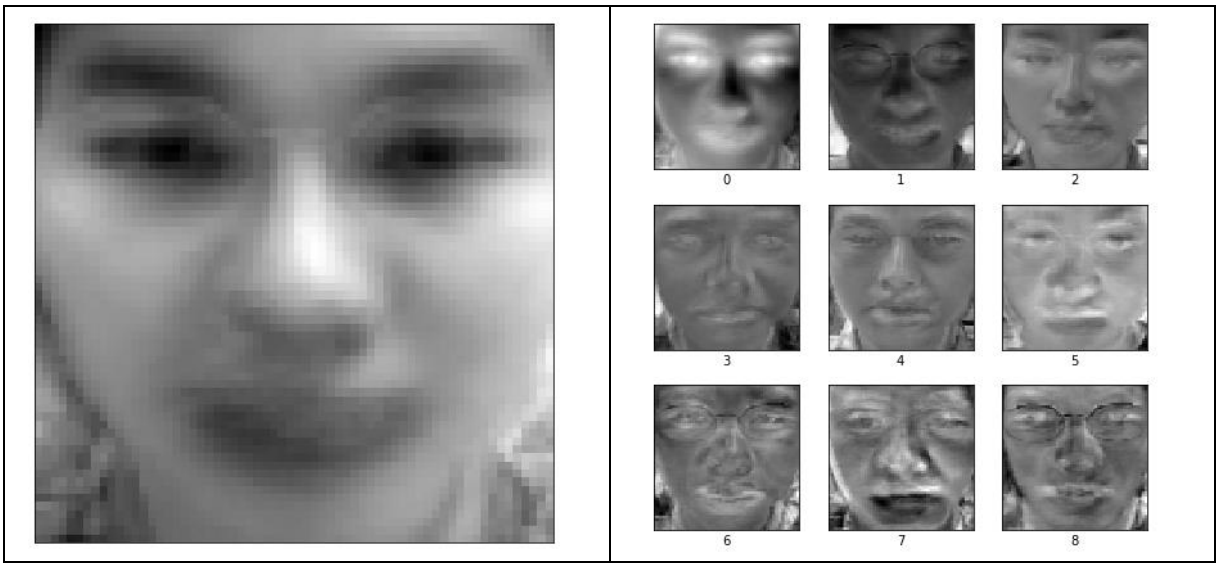
學號：R05942056 系級： 電信所碩一 姓名：時丕澔

1.1. Dataset 中前 10 個人的前 10 張照片的平均臉和 PCA 得到的前 9 個 eigenfaces:

答： (左圖平均臉，右圖為 3x3 格狀 eigenfaces, 順序為 左到右再上到下)

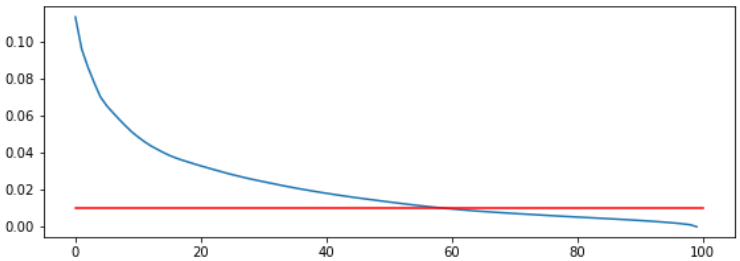


1.2. Dataset 中前 10 個人的前 10 張照片的原始圖片和 reconstruct 圖 (用前 5 個 eigenfaces):

答：(左右各為 10x10 格狀的圖, 順序一樣是左到右再上到下)

1.3. Dataset 中前 10 個人的前 10 張照片投影到 top k eigenfaces 時就可以達到 < 1% 的 reconstruction error.

答： k = 59



2.1. 使用 word2vec toolkit 的各個參數的值與其意義:

答：

train, output: 輸入跟輸出的檔案路徑

size: 800，蒐集的 feature 的維度大小

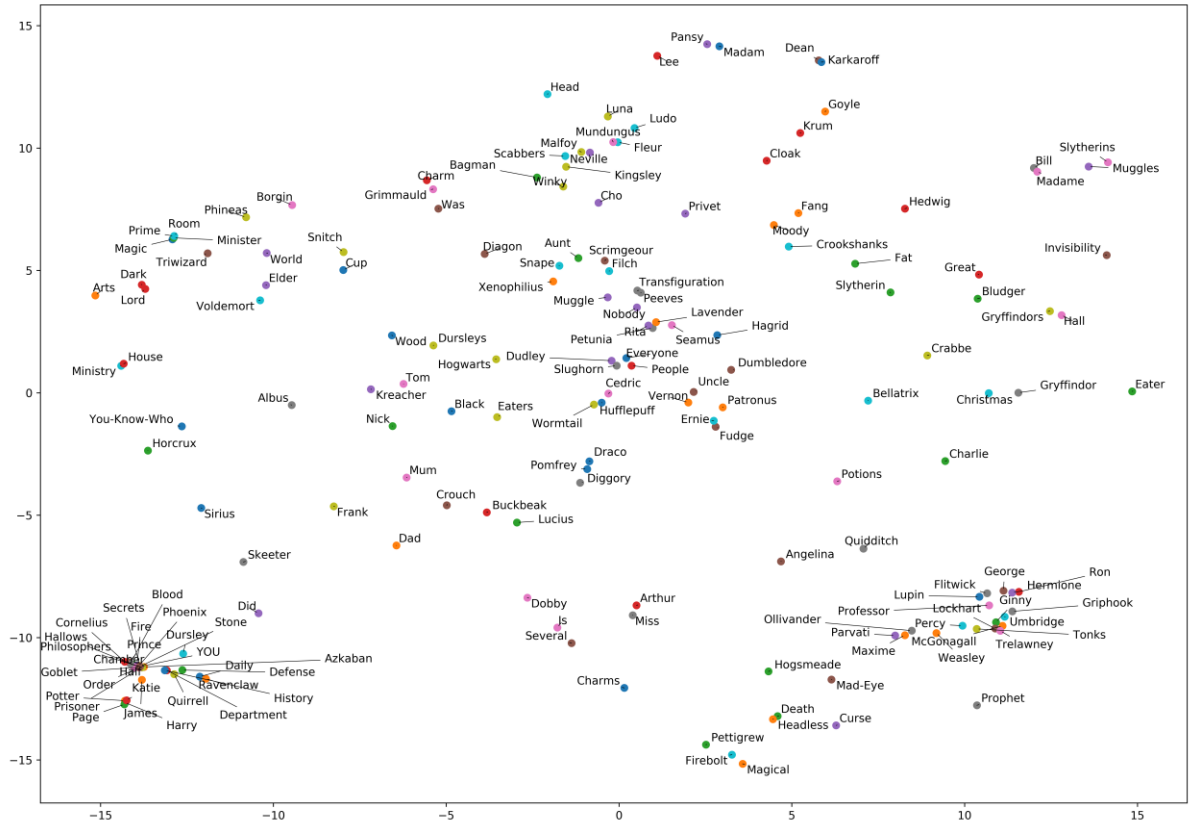
min_count: 50，一個字至少要出現幾次才被計入

iter_ = 20，train 幾個 epoch

negative = 1，是否使用 negative sampling

2.2. 將 word2vec 的結果投影到 2 維的圖:

答：



2.3. 從上題視覺化的圖中觀察到了什麼？

答：

有關聯的單字普遍會靠近，尤其是詞性相同、句構位置相同的通常會比較集中。例如左下角很多都是通篇常常出現的單字集合，右下角集中的則是文中出現的特定家族名稱。

3.1. 請詳加解釋你估計原始維度的原理、合理性，這方法的通用性如何？
答：

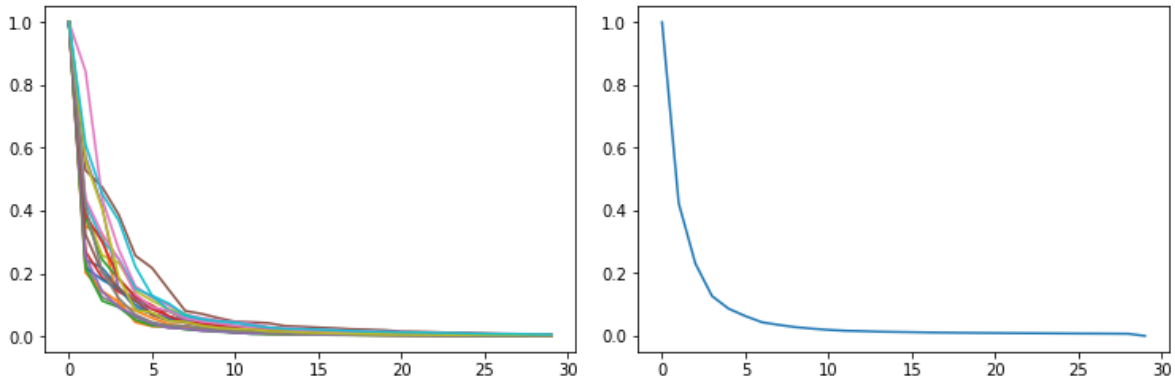
我的作法是在資料中取出數個資料點，並用 tree-structure 找它們附近的點集，對這些點集合做 PCA，最後觀察 PCA 出來的特徵值分布來找出最有可能的維度值。

這個方法的前提假設是此資料代表的流形沒有非常明顯的皺褶或是誇張的折疊，在數學上這種資料有一些局部的平滑性可以利用；也就是說，如果取一個點附近的點集合出來看它會非常像一個高維度中的低維平面，而 PCA 正是解釋有這樣行為的資料集合很有用的工具。

這方法的核心挑戰在於如何使用 PCA 出來的特徵值分布找到最有可能的維度值，在本題中直接利用了作業中有給資料集合產生方法的情況，先將一些 pretrain 的資料集合用 SVM 做好一個特徵值分布對結果維度值映射的模型，最後將測試資料計算出的特徵分布直接套用模型得出維度值。

3.2. 將你的方法做在 hand rotation sequence datatset 上得到什麼結果？合理嗎？請討論之。
答：

雖然 handrotate 的資料和本題的產生方法不同，但還是可以套用之前 train 出來的模型來預測，但因為此圖片集的維度大小很大，故我先把它們做二維的 sample，再拉直成 feature vector 後套用上述做法。下圖是 handrotate dataset 取 20 個取樣點、每個點分別取 30 個鄰域點，計算出來的 PCA 分布，還有它們的平均。



套用原題的模型得出的維度值為 4。雖然肉眼看來合理的猜測應該為 1，但可以想像此圖集因為在很高維(24W)的空間中散布，故樣本數相對非常小，做 sampling 後 fit 出來的流形也可能長的較彎，會因此對 PCA 造成一些誤差，但考慮到本題的樣本數這麼小，計算維度會是相對 ill-posed 的問題，故此結果實際上並不差。