Problem 1

1. 當每個資料點之間的大小差異(或數量級差)愈大的話，經過covariance matrix的計算後會無法凸顯彼此之間的差異。例如題中Joe的兩筆資料數量級差10倍而boss的差10000倍，所以boss的GDP資料可能就會比較集中在一小段範圍(數字上)，而過度的集中就可能無法藉由每筆資料間的variance獲得正確的相對關係。
2. 所以正確評價每個pc要避免以上情況，例如某些資料點較其他點離散時可以用log normalization的方式重新分布，反之可以用exponential之類的方式展開之。

Problem 2

。

1. eigenvalue和其eigenvector分別為

可以看出只會影響eigenvalue的值，不會影響eigenvector(主值方向)。

Problem 3

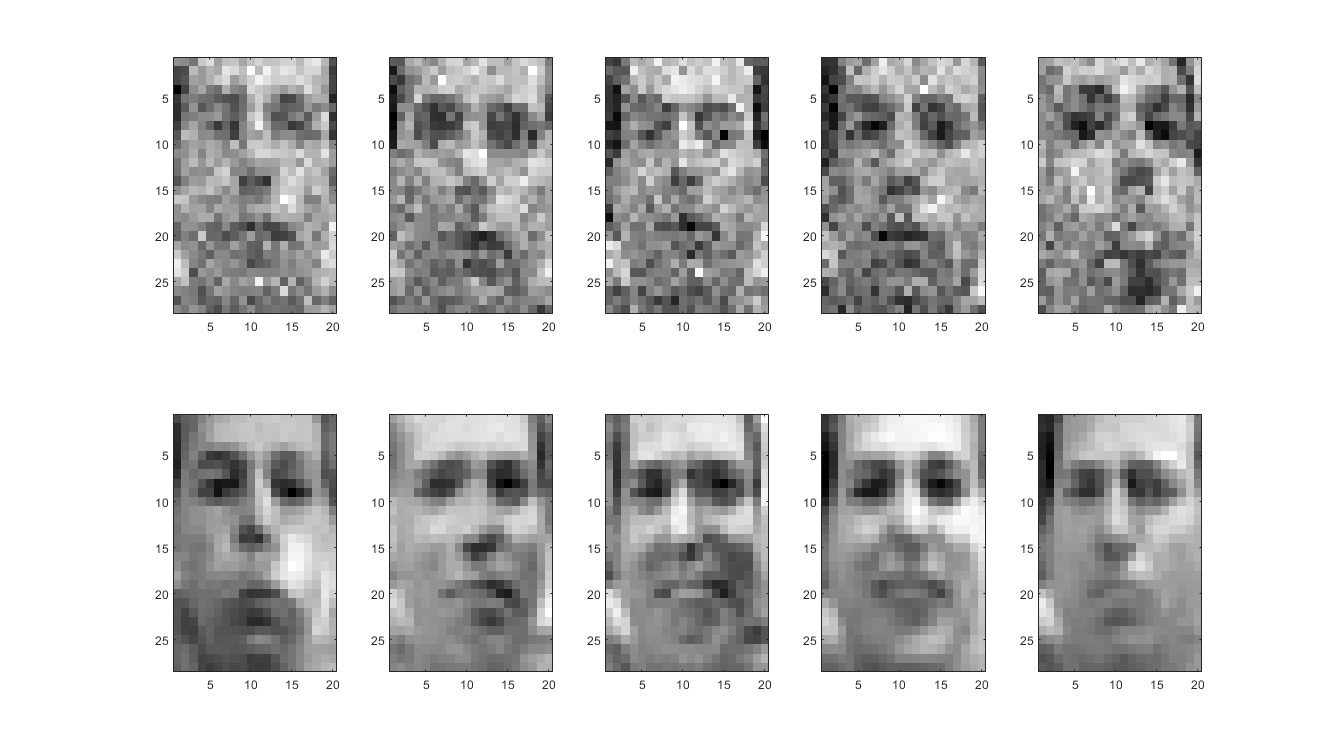


Figure 1 Result of taking first 10 eigenvalues

結果看似不錯，變得比較清晰且和原圖有神似之處。

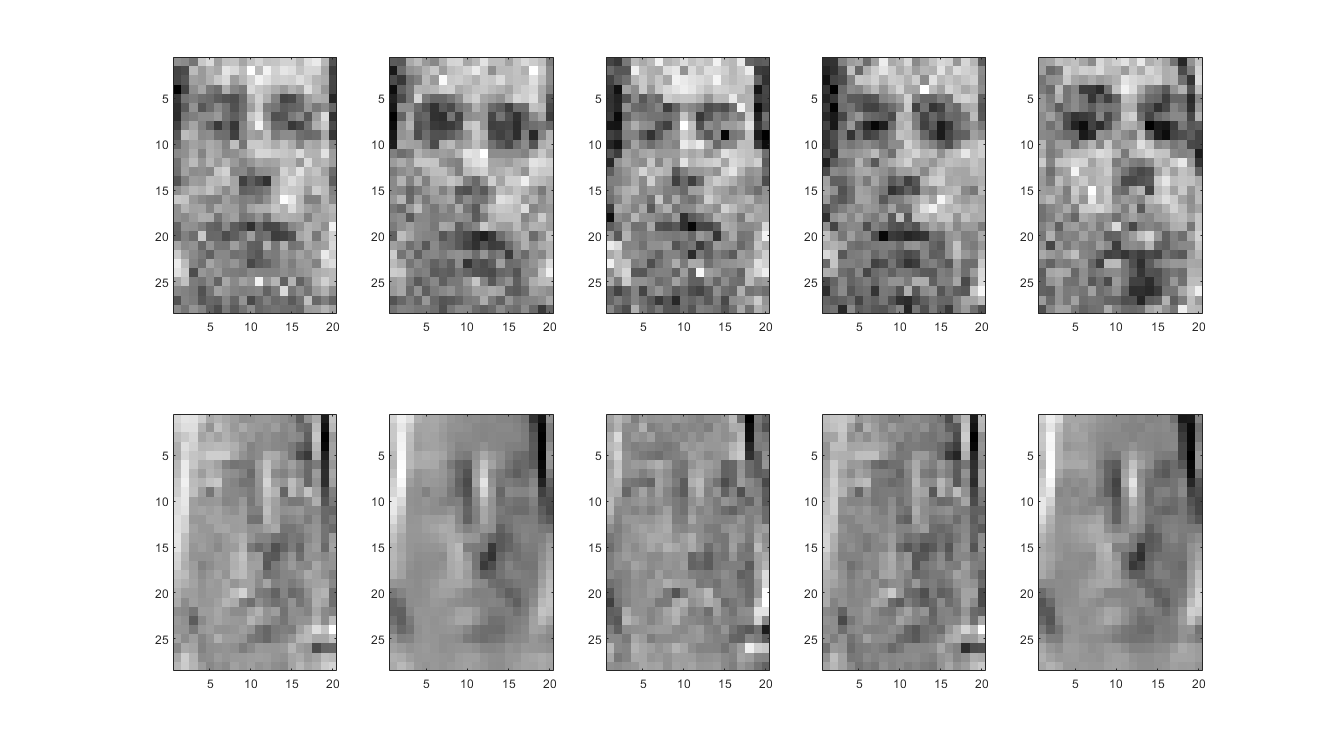


Figure 2 Result of taking 2nd and 30th eigenvalue

結果明顯比較差，雜訊不僅沒有去除，反而更簡辨識度。

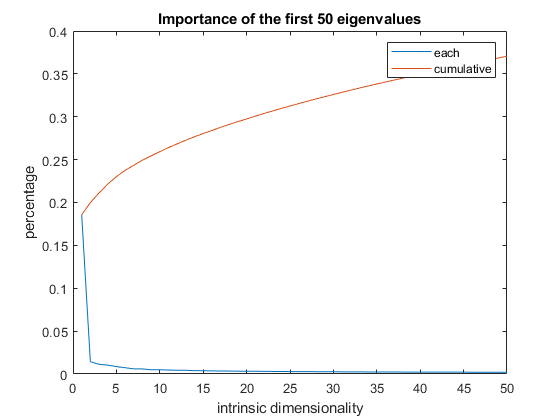


Figure 3 Importance of each eigenvalue

將每個eigenvalue除以總值作圖，藍線為每個eigenvalue所佔的比例，橘線為累積值，由於超過第50筆以後的值都呈現相同趨勢，故予以省略。

由上圖可以得知在intrinsic dimensionality超過5以後幾乎對整體不太佔影響，所以我認為取前5個或附近，前4或前6或前7，就足以消除大部分雜訊了。

以下為程式碼

load('04HW2\_noisy.mat');

[U,S,V] = svd(X);

newS = zeros(size(S));

for i=1:5

newS(i,i) = S(i,i);

end

recordS = zeros(560,1);

for j=1:560

recordS(j) = S(j,j);

end

newX = U\*newS\*V;

colormap gray

subplot(2,5,1)

imagesc(reshape(X(:, 10), 20, 28)')

subplot(2,5,2)

imagesc(reshape(X(:, 121), 20, 28)')

subplot(2,5,3)

imagesc(reshape(X(:, 225), 20, 28)')

subplot(2,5,4)

imagesc(reshape(X(:, 318), 20, 28)')

subplot(2,5,5)

imagesc(reshape(X(:, 426), 20, 28)')

subplot(2,5,6)

imagesc(reshape(newX(:, 10), 20, 28)')

subplot(2,5,7)

imagesc(reshape(newX(:, 121), 20, 28)')

subplot(2,5,8)

imagesc(reshape(newX(:, 225), 20, 28)')

subplot(2,5,9)

imagesc(reshape(newX(:, 318)\*-1, 20, 28)')

subplot(2,5,10)

imagesc(reshape(newX(:, 426)\*-1, 20, 28)')

%%

sum\_S = sum(recordS);

recordS(:,2) = recordS(:,1)./sum\_S;

recordS(1,3) = recordS(1,2);

for i=2:560

recordS(i,3) = recordS(i,2)+recordS(i-1,3);

end

%%

plot(1:15,recordS(1:15,2),1:15,recordS(1:15,3))

xlabel('intrinsic dimensionality')

ylabel('percentage')

title('Importance of the first 50 eigenvalues')

legend('each','cumulative')