Principal Component Analysis

Vishal S, IIT Madras

1 Introduction

In this assignment we will try to implement Principal Component Analysis method to a data and try to analyze the variance of the data with respect to its principal components

2 Principal Component Analysis

Principal Component Analysis is a procedure by which the data is reduced to a set of variables which are linearly uncorrelated called as **Principal Components**. These principal components are orthogonal to each other. These components are so chosen such that the variance is maximum in that direction.

3 Code

A code was written in MATLAB to calculate the principal components of a given data.

```
mean = zeros(len(2),1);
sum_col = zeros(len(2),1);
% To make the mean zero
for j = 1:len(2)
    sum_col(j) = sum(fatiguedatademo(:,j));
    mean(j) = sum_col(j)/len(1);
end
for j = 1:len(2)
    for i = 1:len(1)
       fatiguedatademo(i,j) = fatiguedatademo(i,j) - mean(j);
    end
end

%% Calculating the principal components
[U,S,V] = svd(fatiguedatademo','econ') ;
error = zeros(26,1);
%% Calculating the error while reconstructing data
for k = 1:26
num_comp  = k ;
 c1 = U(:,1:num_comp)'*fatiguedatademo';
 temp = c1';
 recons = temp(:,1:num_comp)*U(:,1:num_comp)';
 for j = 1:len(2)
    for i = 1:len(1)
       recons(i,j) = recons(i,j) + mean(j);
    end
end
 err = immse(recons,fatiguedatademo);
 error(k) = max(err);
end


%% To plot the maximum error as function of components considered
comp = 1:26;
plot(comp,error,'-ro','linewidth',2.0);
xlabel('Number of Components');
ylabel('Max error in Data')
ax = gca ;
set(ax, 'linewidth',1.5);
xlim([2 10])
axis('square');
grid on;

%% To make scatter plots
scatter(c1(1,:),c1(5,:));
```

```
xlabel('PC1');
ylabel('PC5')
ax = gca ;
set(ax, 'linewidth',1.5);
axis('square');
grid on;


%% To make a Pareto plot
s = svd(fatiguedatademo','econ') ;
pareto(s)
xlabel('Components');
ylabel('Frequency')
ax = gca ;
set(ax, 'linewidth',2.0);
grid on;
legend('Frequency','Cumulative Frequency');
```

The above code was used to create *Pareto plots* and *Scatter Plot* of the principal components calculated for the data.

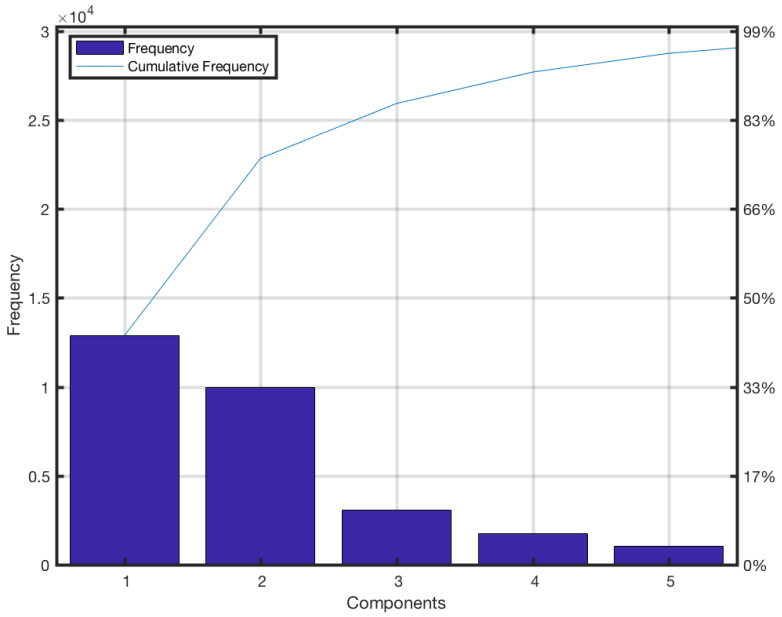
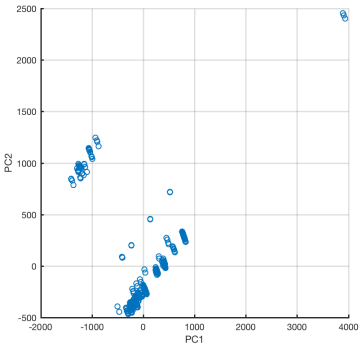
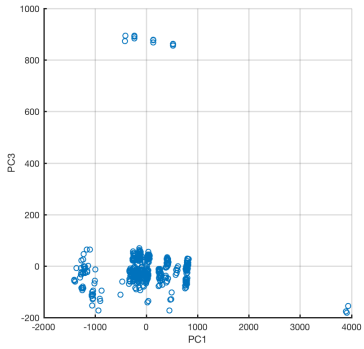


Figure 1: Pareto plot of frequency of Principal Components

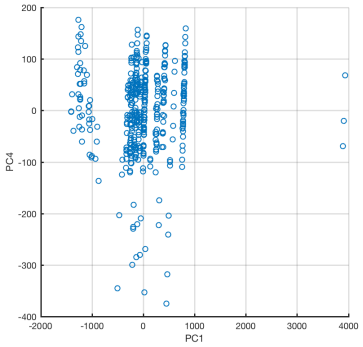
Then we reconstructed the data using the principal components and the maximum error from the original data was calculated as a function of number of components used in Fig. 3



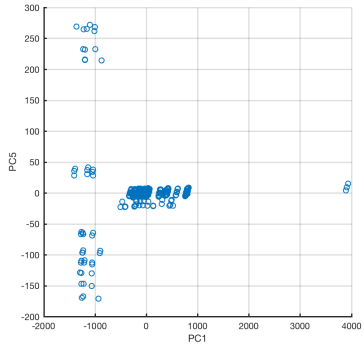
(a) Principal Components 1 and 2



(b) Principal Components 1 and 3



(c) Principal Components 1 and 4



(d) Principal Components 1 and 4

Figure 2: Scatter plots of first 5 principal components

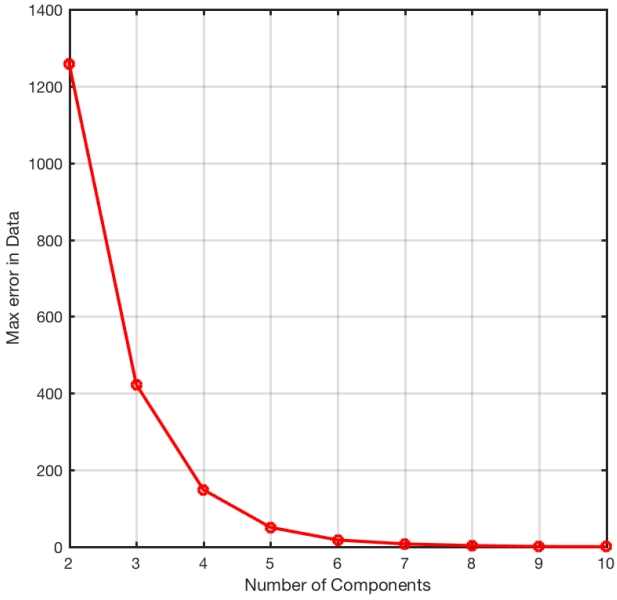


Figure 3: Max. error in reconstructed data

4 Results

We performed **Principal Component Analysis** on the data set provided to us. We made *Pareto plot* and *Scatter plots* on the principal components. We also reconstructed the data using the principal components and compared with the original data to find that 8 components are sufficient to reconstruct the data to reasonable accuracy.