Fast Algorithms for Big Data: Homework 4, Problem 3 Write-Up

Derek Driggs

April 26, 2016

Problem 3

We will use PCA to estimate the covariance and the mean of the distributions A, B, and C. The analysis of each distribution will utilize the same code, and this code is attached.

Part A

The distribution A lives in two dimensions. After running PCA, we see that the estimated mean is $(-0.01767, -0.09507)$. This singular values of X are 3.03492 and 0.98450. Because both of these values are on the same order of magnitude, we should include both dimensions in our analysis. The two principal directions are

$$u_1 = \begin{pmatrix} -0.88486 \\ 0.46586 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0.46586 \\ 0.88486 \end{pmatrix}.$$

A scatter-plot of the distribution as well as the two principal directions are shown below. If the principal-direction vectors seem to not be perpendicular, it is due to the scaling of the axis.

Finally, We find that the estimated covariance matrix is

$$\left(\frac{1}{n-1}\right) XX^* = \begin{pmatrix} 2.75788 & 3.39732 \\ 3.39732 & 7.42211 \end{pmatrix}.$$

This indicated that there is a direct correlation between the two variables.

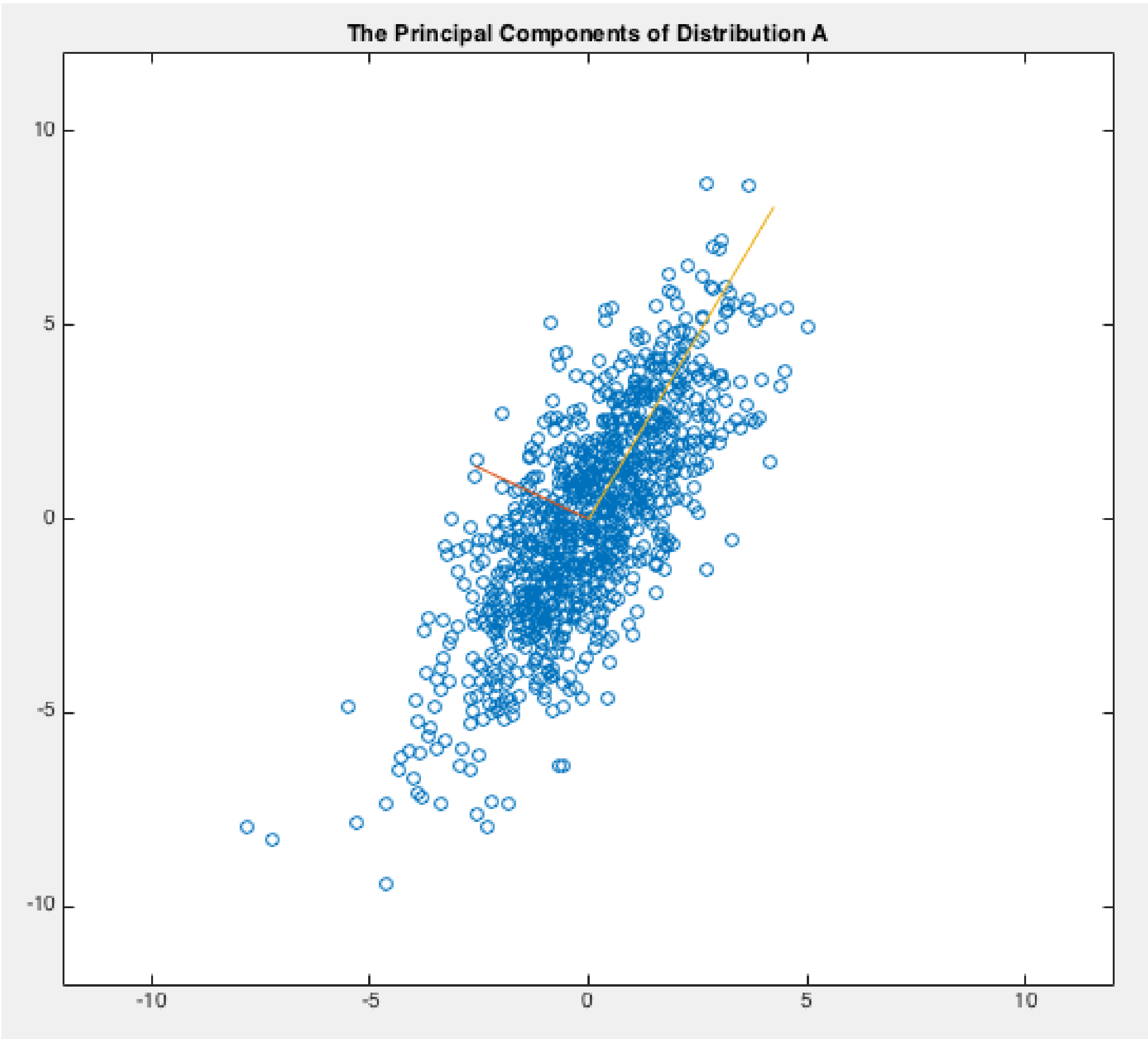
Part B

The distribution B is a bit more interesting. This distribution lives in three-dimensions, but it can be represented well using a two-dimensional subspace. To see this, observe the two plots on the following page. Each plot shows the distribution B as it exists in three-dimensions, but the plot on the right visualizes the data along the two-dimensional "active" subspace.

PCA estimates the mean to be $(1.10423, 0.98240, 1.08742)$. The estimated singular values of the covariance matrix are 2.99391, 1.00397, and 0.00970. As predicted, one of the singular values is several orders of magnitude smaller that the others, so we do not include this dimension in our analysis. The principal vectors are then

$$u_1 = \begin{pmatrix} 0.41701 \\ 0.39926 \\ -0.81651 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0.70204 \\ -0.71206 \\ -0.01036 \end{pmatrix}.$$

A plot of the distribution along the two-dimensional subspace spanned by these vectors is shown on the following pages.


The Principal Components of Distribution A
10
5
0
-5
-10
-10
-5
0
5
10

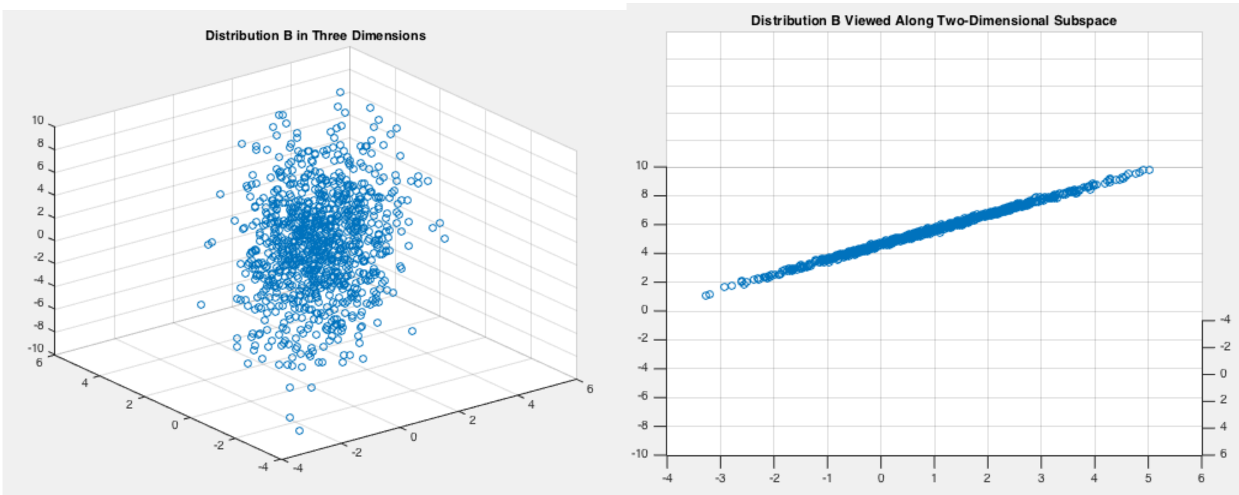
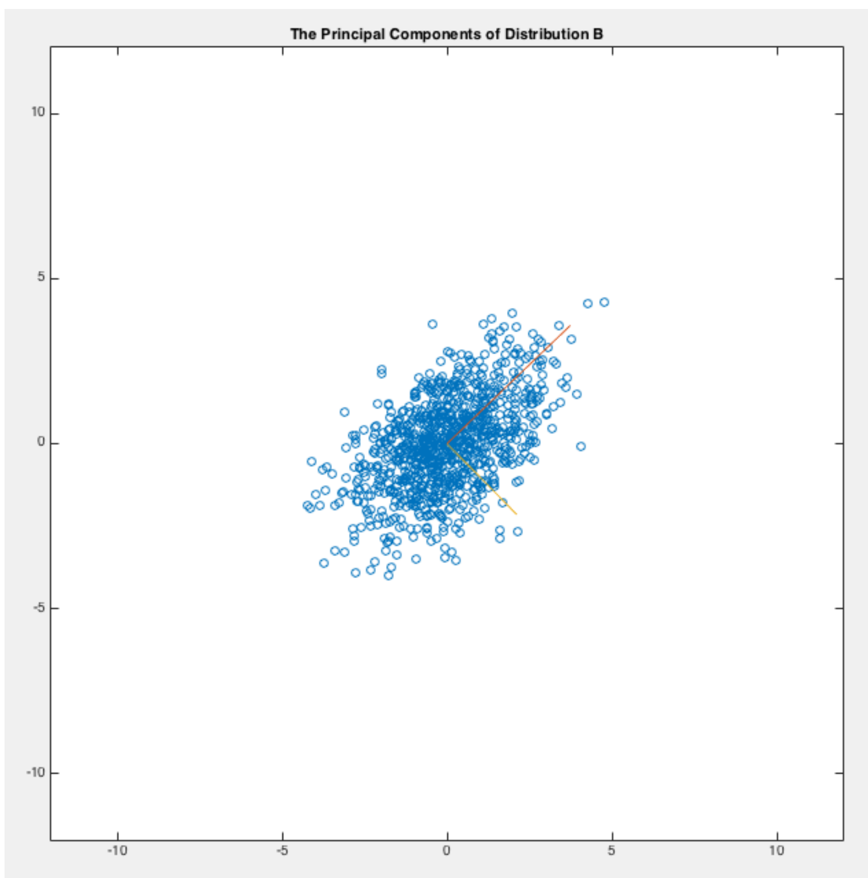


Figure 1: Distribution B in 3-space. The plot on the right shows that the distribution exists primarily along a two-dimensional subspace.



The estimated covariance matrix is

$$\left(\frac{1}{n-1}\right) XX^* = \begin{pmatrix} 2.05553 & 0.98855 & 3.04465 \\ 0.98855 & 1.93997 & 2.92952 \\ 3.04465 & 2.92952 & 5.97603 \end{pmatrix}.$$

Part C

The distribution C is five-dimensional, so we will exclude any visualizations. PCA estimates the mean to be $(0.31200, 0.32095, 0.78057, 0.50640, 0.28385)$. The estimated singular values of the covariance matrix are $(2.97833, 2.95970, 0.99572, 0.00348, 0.00328)$; we notice that the first three dimensions contain most of the information from the distribution. The principal directions are then

$$u_1 = \begin{pmatrix} 0.34583 \\ 0.03202 \\ 0.12527 \\ 0.05863 \\ 0.92750 \end{pmatrix}, \quad u_2 = \begin{pmatrix} -0.45546 \\ 0.24128 \\ -0.75875 \\ 0.31475 \\ 0.24408 \end{pmatrix}, \quad u_3 = \begin{pmatrix} -0.65706 \\ 0.26607 \\ -0.63579 \\ -0.27511 \\ -0.13254 \end{pmatrix}.$$

The estimated covariance matrix is

$$\left(\frac{1}{n-1}\right) XX^* = \begin{pmatrix} 3.30611 & -1.03776 & 2.99733 & -1.25513 & 1.78506 \\ -1.03776 & 0.58926 & -1.40037 & 0.75446 & 0.81427 \\ 2.99733 & -1.40037 & 5.58298 & -1.85338 & -0.50807 \\ -1.25513 & 0.75446 & -1.85338 & 0.97332 & 1.19145 \\ 1.78506 & 0.81427 & -0.50807 & 1.19145 & 8.17009 \end{pmatrix}.$$

```matlab
% An implementation of PCA
% NOTE: Professor Martinsson has posted
% code on the class website as well

% Load Test Distributions
testMats = load('testmatrices');

A = testMats.A;
B = testMats.B;
C = testMats.C;

% Choose matrix A, B, or C
% for the following 3 lines
n = size(C,2);
Mu = 1/n*C*ones(n,1); % Estimate Mean
X  = C-Mu*ones(1,n);

% Form Covariance Matrix
Cov = 1/(n-1)*(X*X.');

% Calculate Eigenvalues and Eigenvectors of Covariance Matrix
[U,D] = eig(Cov);
D     = diag(D.^(1/2));

% Order Principal Directions by Magnitude
[~,ind] = sort(D,'descend');
D       = D(ind);
U       = U(:,ind);

% k is Chosen After Analyzing the Decay of the Singular Values
k = 3;

% Compute Principal Directions and Magnitudes
U_principal = U(:,1:k);
D_principal = D(1:k);


% Plot the Distribution and Principal Directions
figure;
plot(X(1,:),X(2,:),'o')
axis([-12 12 -12 12])
hold on
plotv(3*D(1).*U_principal(:,1))
plotv(3*D(2).*U_principal(:,2))

% Choose matrix A, B, or C
title('The Principal Components of Distribution A')
```