exping fall 2012 hw4

4 PCA [Avi 25 pts]

In this question we will try to understand PCA by showing two cool ways of interpreting the first principal component. One is the direction of maximum variance after projection and the second is the direction that minimizes reconstruction error. Note that the first principal component is the first eigenvector of the sample covariance matrix.

Consider n points $X_1, ..., X_n$ in p-dimensional space, and let X be the $n \times p$ matrix representing these points. Assume that the data points are centered, ie, $\vec{1}^\top X = \vec{0}$. Consider a unit vector $v \in \mathbb{R}^p$ and project all the points onto this vector (hence every point becomes a one-dimensional point on the direction of unit vector v).

1 [1 pt] Argue that the projection is given by Xv.

Soln:- Let us decompose the vector representing X_i into two orthogonal vectors X_{iv} and $X_{iv'}$ where X_{iv} is parallel to v. Using notation that $X = [X_1^T, X_2^T, \ldots, X_n^T] = [X_i^T]$ we get

$$Xv = [(X_{iv} + X_{iv'})]v = [X_{iv}^T]v + [X_{iv'}^T]v = [X_{iv}^T]v = X_v v$$

Given that v is a unit vector, $X_v v$ given the component of X in direction v. Since $Xv = X_v v$, Xv represents projection of X onto v.

2 [2 pt] What is the sample mean of all the points after the projection?

Soln:- Sample mean after projection is given by

$$\frac{1}{n}[\vec{1}^T(Xv)] = \frac{1}{n}[(\vec{1}^T X)v] = \frac{1}{n}[\vec{0}] = 0$$

3 [2 pt] What is the sample variance of all the points after the projection?

Soln:- Given that the sample mean is zero we can write the variance as

$$\frac{1}{n}[(Xv)^T(Xv)] = \frac{1}{n}[v^T X^T X v] = v^T \left[\frac{X^T X}{n}\right] v = v^T \Sigma v$$

where $\Sigma = X^T X$ is the sample variance of original p-dimensional points (X).

4 [2 pt] Setup the problem of maximizing the sample variance of the projection onto v subject to a constraint on the L2-norm of v.

Soln:-

$$\max_{v} v^{\top} \Sigma v$$
$$st. \|v\|^2 = 1 \quad (1)$$

5 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

Soln:- By stationarity, at optimality we have

$$2\Sigma v^* + \lambda^* v^* = 0$$

Thus the optimal value is $v^{*T}\Sigma v^* = \lambda$ and so the vector that maximizes variance after projection, is the eigenvector associated with the largest eigenvalue λ of the covariance matrix Σ.

So we have now proved that the direction of maximum covariance is the first PC. Now we show that the direction that minimizes reconstruction error is also the first PC.

6 [1 pt] Argue that the reconstruction of X_i using v is $(X_i^{\top} v)v$.

Soln:- The reconstruction error of X_i using v can be written as the following optimization problem (with α being scalar): $\min_{\alpha} \|X_i - \alpha v\|^2$. Taking derivative wrt α and setting it to zero gives us the following:

$$2(X_i - \alpha v)^T v = 0 \Leftrightarrow X_i^T v = \alpha v^T v \Leftrightarrow \alpha = X_i^T v$$

Since v is a unit vector $v^T v = 1$. So, $\alpha v = (X_i^T v)v$ is the reconstruction of X_i using v.

7 [2 pt] You projected X_i to $X_i^{\top} v$ and then reconstructed it using $(X_i^{\top} v)v$. What is the reconstruction error of X_i, when measured in L2-norm?

Soln:-

$$\|(X_i^T v)v - X_i\|_2$$

8 [2 pt] What is the total squared reconstruction error over all points? **Soln:-**

$$\|(X^T v)v - X\|_F^2$$

9 [2 pt] Show that minimizing total squared reconstruction error is equivalent to minimizing $-\|Xv\|_2^2$.

Soln:-

$$\|(X^T v)v - X\|_F^2 \quad = \quad tr(((X^T v)v - X)^T((X^T v)v - X))$$

$$
\begin{aligned}
&= tr(vv^T X^T X vv^T) - 2tr(vv^T X^T X) + tr(X^T X) \\
&= tr(v^T X^T X vv^T v) - 2tr(v^T X^T X v) + tr(X^T X) \\
&= tr(v^T X^T X v) - 2tr(v^T X^T X v) + tr(X^T X) \\
&= -tr(v^T X^T X v) + tr(X^T X) \\
&= -\|Xv\|_2^2 + \|X\|_2^2
\end{aligned}
$$

since the minimization is wrt to v, $\|X\|_2^2$ is constant.

10 [4 pt] Solve the minimization problem to show that the solution is the first PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

Soln:- The optimization problem is the same as in part 5.

4.1 [3 pts] SVD and PCA

Let us define a new variable Y as

$$Y = X^T \tag{2}$$

where X is a $n \times p$ matrix containing the data points as defined before. If the SVD of Y is given by $Y = U\Sigma V^T$ then show that the columns of V are the PCA of X.

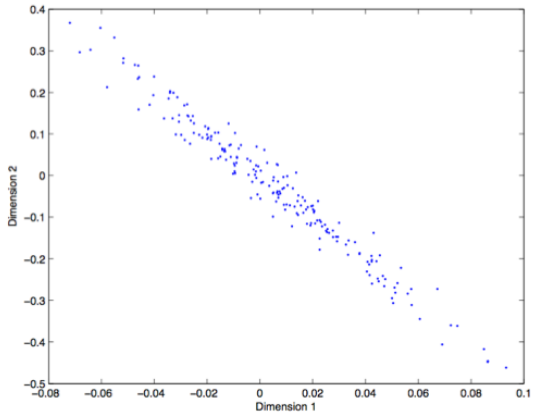
Soln:

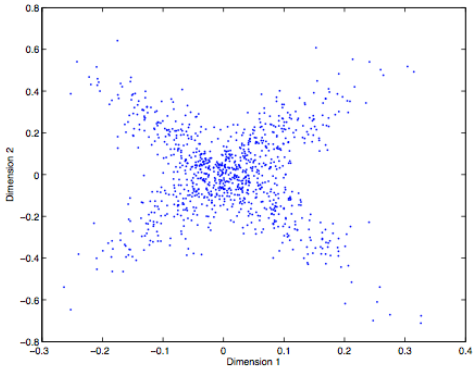
$$XX^T = Y^T Y = V\Sigma U U^T \Sigma V^T = V\Sigma^2 V^T \tag{3}$$

Thus $(XX^T)V_i = \Sigma_{ii}^2 V_i$ which implies that the columns of V are the PCA of X.

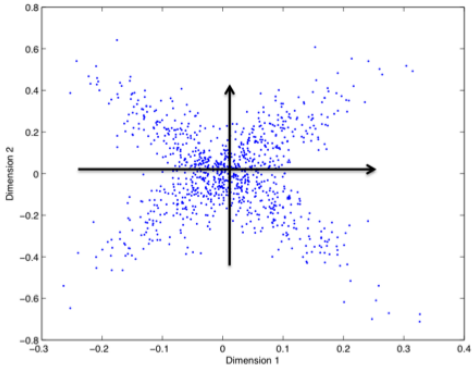
fsfb

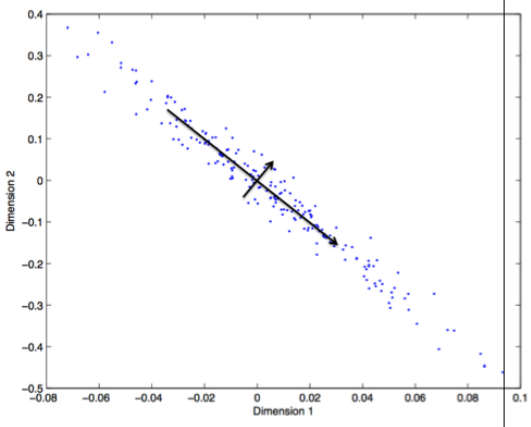
(e) [2 points] Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principle components on each plot.





Solution:





2.1 Principal components analysis vs. Fisher's linear discriminant

Principal components analysis (PCA) reduces the dimensionality of the data by finding projection direction(s) that *minimizes the squared errors in reconstructing the original data* or equivalently *maximizes the variance of the projected data.* On the other hand, Fisher's linear discriminant is a supervised dimension reduction method, which, given labels of the data, finds the projection direction that *maximizes the between-class variance relative to the within-class variance of the projected data.*

[10 points] In the following Figure 2, **draw** the first principal component direction in the left figure, and the first Fisher's linear discriminant direction in the right figure. Note: for PCA, ignore the fact that points are labeled (as round, diamond or square) since PCA does not use label information. For linear discriminant, consider round points as the positive class, and both diamond and square points as the negative class (since in the course lecture we only discuss the two-class case).

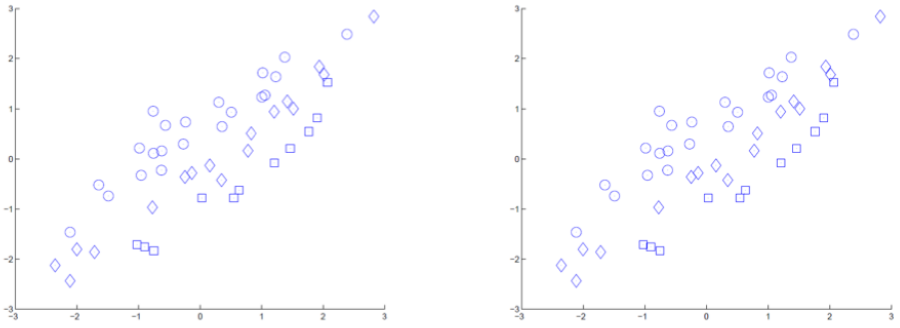
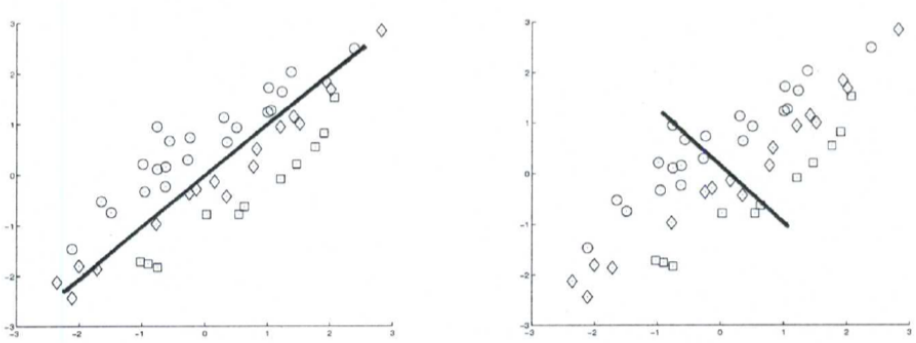


Figure 2: Draw the first principal component and linear discriminant component, respectively

★ **SOLUTION:** The PCA and LDA directions are shown in the following figure.



1(a) First PCA component

1(b) First LDA component

Figure 3: The first principal component and linear discriminant component, respectively

2.3 More Principal Components Analysis

Consider 3 data points in the 2-d space: (-1, -1), (0,0), (1,1).

[6 points] What is the first principal component (write down the actual vector)?

★ **SOLUTION:** The first principal component is $\mathbf{v} = [\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T$ (you shouldn't really need to solve any SVD or eigenproblem to see this). Note that the principal component should be normalized to have unit length. (The negation $\mathbf{v} = [-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}]^T$ is also correct.)

[7 points] If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the variance of the projected data?

★ **SOLUTION:** The coordinates of three points after projection should be $z_1 = \mathbf{x}_1^T\mathbf{v} = [-1,-1][\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T = -\sqrt{2}$, $z_2 = \mathbf{x}_2^T\mathbf{v} = 0$, $z_3 = \mathbf{x}_3^T\mathbf{v} = \sqrt{2}$. Note that the sample mean is 0, and thus the variance is $\frac{1}{3}\sum_{i=1}^{3}(z_i - 0)^2 = \frac{4}{3}$ (or you can also choose to use the unbiased estimation $\frac{1}{3-1}\sum_{i=1}^{3}(z_i - 0)^2 = 2$).

[6 points] For the projected data you just obtained above, now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error?

★ **SOLUTION:** The reconstruction error is 0, since all three points are perfectly located on the direction of the first principal component. Or, you can actually calculate the reconstruction: $\hat{\mathbf{x}}_1 = z_1 \cdot \mathbf{v} = -\sqrt{2}\cdot[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]^T = [-1,-1]^T$, $\hat{\mathbf{x}}_2 = [0,0]^T$, $\hat{\mathbf{x}}_3 = [1,1]^T$, which are exactly $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$.

exping hw5

4. **[2 points]** PCA (Principal Component Analysis) can also be used to learn "topics" from a set of documents. Give at least two differences between PCA and topic models. You don't have to explain the differences, just list them.

Answer:

- The contributions of words to a topic can be negative in PCA while they are constrained to be non-negative and < 1 in topic models.
- The topics obtained from PCA are orthogonal to each other while there is no such restriction on topics in topic models.

midterm 10701 spring 2014

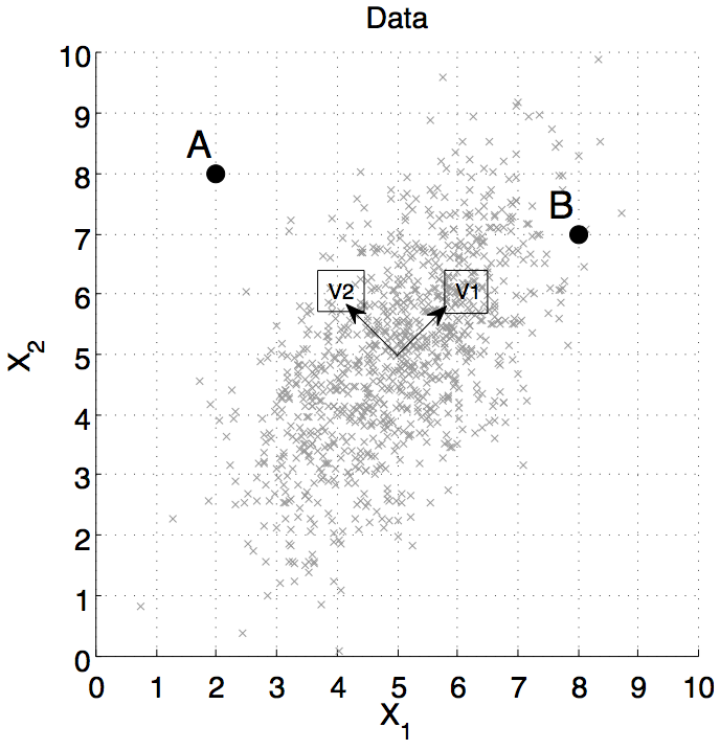
Let $X \in \mathbb{R}^{n \times m}$ be the data matrix of m n-dimensional data instances. Let $X = UDV$ denote the singular decomposition of X where D is a diagonal matrix containing the singular values ordered from largest to smallest. The 1st principal component of X is the first column vector of V.

False. The 1st principal component of X is the first column vector of U.

Define v_1 and v_2 as the directions of the first and second principal component, with $\|v_1\| = \|v_2\| = 1$. These directions define a change of basis

$$\begin{aligned} Z_1 &= (X - \mu) \cdot v_1 \\ Z_2 &= (X - \mu) \cdot v_2 \ . \end{aligned}$$

1. **[Points: 4 pts]** Sketch and label v_1 and v_2 on the following figure (a copy of Fig. 3). The arrows should originate from the mean of the distribution. You do not need to solve the SVD, instead visually estimate the directions.



★ **SOLUTION:** See above figure. Notice that both arrows are unit length.

2. **[Points: 2 pts]** The covariance $\mathrm{Cov}(Z_1, Z_2)$, is (circle):
   (a) negative
   (b) positive
   (c) approximately zero ★

3. **[Points: 2 pts]** Which point (A or B) would have the higher reconstruction error after projecting onto the first principal component direction v_1? Circle one:

   Point A ★ Point B

1.2 (7 pts) **PCA and SVD**

Given 6 data points in 5-d space, $(1,1,1,0,0)$, $(-3,-3,-3,0,0)$, $(2,2,2,0,0)$, $(0,0,0,-1,-1)$, $(0,0,0,2,2)$, $(0,0,0,-1,-1)$. We can represent these data points by a 6×5 matrix X, where each row corresponds to a data point:

$$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

(a) (1 pt) What is the sample mean of the data set?

Solutions: $[0,0,0,0,0]$

(b) (3 pts) What is SVD of the data matrix X you choose?

hints: The SVD for this matrix must take the following form, where $a, b, c, d, \sigma_1, \sigma_2$ are the parameters you need to decide.

$$X = \begin{bmatrix} a & 0 \\ -3a & 0 \\ 2a & 0 \\ 0 & b \\ 0 & -2b \\ 0 & b \end{bmatrix} \times \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \times \begin{bmatrix} c & c & c & 0 & 0 \\ 0 & 0 & 0 & d & d \end{bmatrix}$$

Solutions: $a = \pm 1/\sqrt{14} = \pm 0.267$, $b = \pm 1/\sqrt{6} = \pm 0.408$,
$\sigma_1 = 1/(a \cdot c) = \sqrt{42} = 6.48$, $\sigma_2 = 1/(b \cdot d) = \sqrt{12} = 3.46$,
$c = \pm 1/\sqrt{3} = \pm 0.577$,$d = \pm 1/\sqrt{2} = \pm 0.707$.

(c) (1 pt) What is first principle component for the original data points?

Solutions: pc=$\pm[c, c, c, 0, 0] = \pm[0.577, 0.577, 0.577, 0, 0]$ (Intuition: First, we want to notice that the first three data points are co-linear, and so do the last three data points. And also the first three data points are orthogonal to the rest three data points. Then, we want notice that the norm of the first three are much bigger than the last three, therefor, the first pc has the same direction as the first three data points)

(d) (1 pt) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

Solutions: var=$\sigma_1^2/6 = 7$ (Intuition: we just the keep the first three data points, and set the rest three data points as $[0, 0, 0, 0, 0]$ (since they are orthogonal to pc), and then compute the variance among them)

(e) (1 pt) For the projected data in (d), now if we represent them in the original 5-d space, what is the reconstruction error?

Solutions: var=$\sigma_2^2/6 = 2^1$(Intuition, since the first three data points are orthogonal with the rest three, here the rerr is the just the sum of the norm of the last three data points (2+8+2=12), and then divided by the total number (6) of data points, if we use average definition

1.1 (3 pts) **Basic PCA**

Given 3 data points in 2-d space, $(1, 1)$, $(2, 2)$ and $(3, 3)$,

(a) (1 pt) what is the first principle component?

Solutions: $pc = (1/\sqrt{2}, 1/\sqrt{2})' = (0.707, 0.707)'$, (the negation is also correct)

(b) (1 pt) If we want to project the original data points into 1-d space by principle component you choose, what is the variance of the projected data?

Solutions: $4/3 = 1.33$

(c) (1 pt) For the projected data in (b), now if we represent them in the original 2-d space, what is the reconstruction error?

Solutions: 0