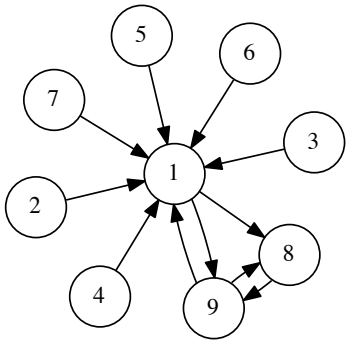
Internet Analytics (COM-308)

Problem Set 3

Problem 1



Consider the graph shown on the left. Estimate the PageRank score of every node, and order the nodes by decreasing score (ties are allowed), for the following two scenarios:

- (a) $\theta = 0.001$.
- (b) $\theta = 0.999$.

Explain what is happening in a few words (exact computations should not be necessary.)

Problem 2

In this problem, we explore the connection between PCA and the SVD. Let X be an $n \times m$ matrix, where each of the n rows is a datum represented by an m-dimensional vector. Furthermore, assume that X is zero-mean, i.e. each column of X sums up to zero ($1_n X = 0_m$). Consider the singular value decomposition (SVD) of X:

$$X = U\Sigma V^\intercal.$$

(a) Express the SVD of the matrix $X^\intercal$ in terms of the SVD of X.

(b) We define the $m \times m$ covariance matrix $\mathrm{Cov}_{m\times m}[X]$ as

$$\mathrm{Cov}_{m\times m}[X] = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \dots \\ \mathrm{Cov}(X_2, X_1) & \ddots & \\ \vdots & & \mathrm{Cov}(X_m, X_m) \end{pmatrix}$$

where $\mathrm{Cov}(X_i, X_j) = \frac{1}{n}\sum_{k=1}^{n} X_{ki} X_{kj}$. Express $\mathrm{Cov}_{m\times m}[X]$ in terms of multiplication of two matrices.

(c) Principal Component analysis (PCA) can be seen as the eigendecomposition of the covariance matrix:

$$\mathrm{Cov}_{m\times m}[X] = Q\Lambda Q^\intercal.$$

How does this interpretation of PCA relate to the SVD of X? Express Q and Λ in terms of U, Σ and V.

Problem 3

We stated in class without proof that the projection onto the principal directions maximizes the variance of the projected data $Y = XV$. Let us now formally establish this fact.

Let X be an $n \times m$ matrix, where each of the n rows is a datum represented by an m-dimensional vector. Furthermore, assume that X is zero-mean, i.e. each column of X sums up to zero ($1_n X = 0_m$).

(a) What are the averages of the columns of Y?

(b) Let us find the first principal component by maximizing the projected variance onto a vector v_1, subject to $v_1^T v_1 = 1$. For this, write the projected variance $\mathrm{Var}[Y]$ as a matrix expression.

(c) Now solve the constrained optimization problem

$$\begin{aligned} v_1 &= \arg\max \mathrm{Var}[Y] && (1) \\ &\quad \text{subject to } v_1^T v_1 = 1, && (2) \end{aligned}$$

using a Lagrange multiplier for the constraint, to show that v_1 is indeed the dominant eigenvector of the covariance matrix.

Problem 4

We generate a set of points $X_1, \ldots, X_n$ i.i.d. according to a Gaussian distribution $N(0, \Sigma)$, where Σ is the covariance matrix, and we then perform PCA on this set of points in order to visualize them in two dimensions.

More precisely, we project every point X_i onto the first and second principal component, i.e., the eigenvectors of the empirical covariance matrix associated with the largest and second-largest eigenvalue (as seen in class).

You are given the following six 2D plots as possible results.

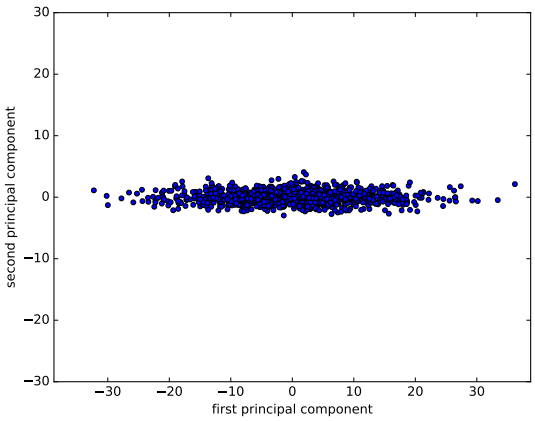
Which of these plots are guaranteed to never result from PCA, and which are plausible? Explain.

For the following three covariance matrices, state which plot would result (among the three plausible plots you identified above). Try to answer without resorting to any calculations.

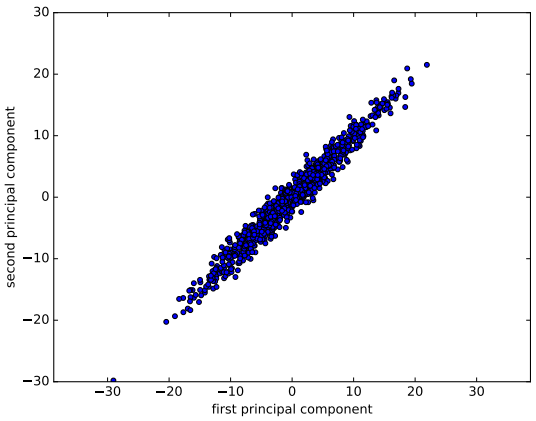
$$\Sigma_1 = \begin{bmatrix} 100 & 0 \\ 0 & 30 \end{bmatrix} \quad (3)$$

$$\Sigma_2 = \begin{bmatrix} 50 & 49 \\ 49 & 50 \end{bmatrix} \quad (4)$$

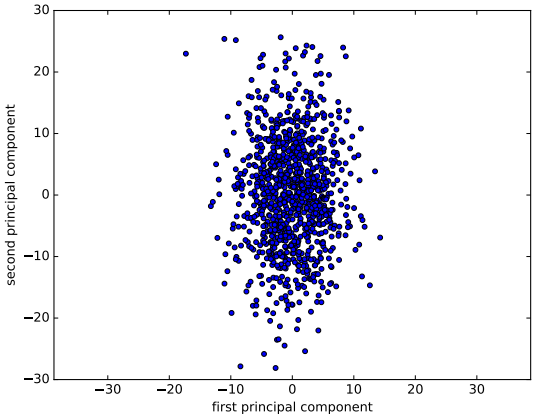
$$\Sigma_3 = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 100 \end{bmatrix} \quad (5)$$


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

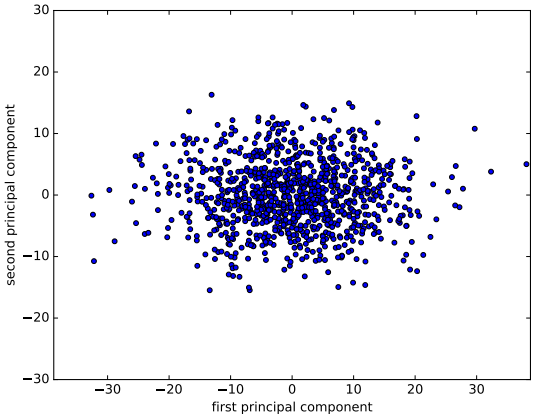
(a)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

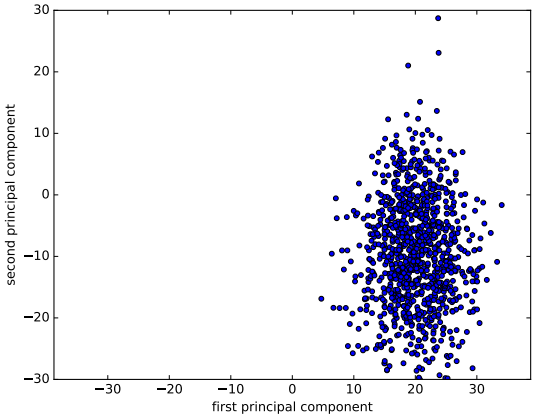
(b)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

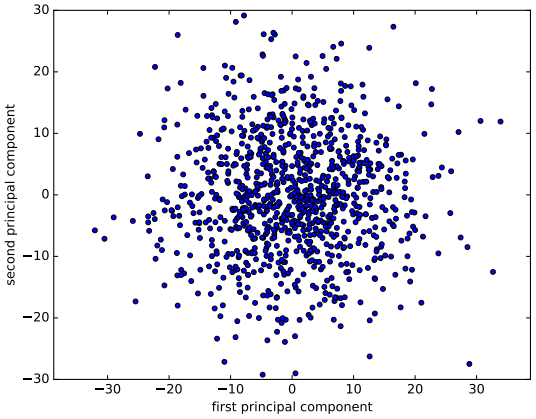
(c)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

(d)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

(e)


second principal component
first principal component
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

(f)