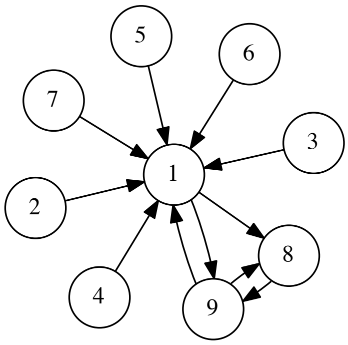
Internet Analytics (COM-308)
Problem Set 3 - Solutions

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.)

Consider a random walk with random restarts. In case (a), when θ is very small, we are mostly jumping around randomly (teleportation), so π is close to uniform. The differences between the PageRank scores of different nodes are therefore small deltas from $1/n$. We need to understand the influence of the rare occasions when we actually do follow a link. This is enough, as it is exceedingly unlikely that a random walker follows more than one link in succession, so no need to consider these (second-order effect). We can therefore approximate the PageRank of a node fairly well by considering that all nodes initially have 1 unit of PageRank, and that this value is evenly split and "pushed" to each node's neighbors:

- Node 1 receives 6.5 units;
- Nodes 2 to 7 receive 0 units;
- Node 8 receives 1 unit;
- Node 9 receives 1.5 units.

This suggests the ranking $1 \succ 9 \succ 8 \succ 2 = \ldots = 7$.

In case (b), there are almost no random restarts (teleportation). One can then consider nodes $2, \ldots, 7$ to be essentially transient, and therefore all the action happens in the subgraph $\{1, 8, 9\}$. Node 9 dominates, because it "receives everything" from 8 and only "gives half" to 1. Node 8 is second, because it "receives" from both 9 and 1, whereas 1 "receives" only from 9. This suggests the ranking $9 \succ 8 \succ 1 \succ 2 = \ldots = 7$.

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.

We have

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

(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.

We have

$$\mathrm{Cov}_{m\times m}[X] = \tfrac{1}{n}[X^\intercal X].$$

(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.

$$\begin{aligned}\mathrm{Cov}_{m\times m}[X] = \tfrac{1}{n}[X^\intercal X] &= \tfrac{1}{n}(U\Sigma V^\intercal)^\intercal U\Sigma V^\intercal \\ &= \tfrac{1}{n}V\Sigma^\intercal U^\intercal U\Sigma V^\intercal \\ &= \tfrac{1}{n}V\Sigma^\intercal I_m \Sigma V^\intercal \\ &= V\tfrac{1}{n}\Sigma^\intercal\Sigma V^\intercal\end{aligned}$$

Hence $Q = V$ and $\Lambda = \frac{1}{n}\Sigma^\intercal\Sigma$. In particular, this means that the principal components of PCA are equal to the right singular vectors of the SVD – a handy way to obtain the PCs from the data X directly via SVD.

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?

The projected means are, as expected, also zero:

$$\frac{1}{n}1_n^T XV = 0$$

.

(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.

$$\mathrm{Var}[Y] = \frac{1}{n}\sum_{i=1}^{n}[(Xv_1)_i]^2 = \frac{1}{n}(Xv_1)^T(Xv_1) = v_1^T\frac{X^TX}{n}v_1.$$

(c) Now solve the constrained optimization problem

$$\begin{aligned} v_1 &= \arg\max \mathrm{Var}[Y] && (1)\\ & \quad \text{subject to } v_1^Tv_1 = 1, && (2) \end{aligned}$$

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

$$L(v_1, \lambda) = v_1^T\frac{X^TX}{n}v_1 - \lambda(v_1^Tv_1 - 1)$$

.

$$\frac{\partial L}{\partial v_1} = 2\frac{X^TX}{n}v_1 - 2\lambda v_1 = 0.$$

$$\frac{\partial L}{\partial \lambda} = v_1^Tv_1 - 1 = 0.$$

Therefore each local maximum (v_1, λ) is an eigenvector-eigenvalue pair of the covariance matrix $\frac{X^TX}{n}$. Note that at the local maximum, $L(v_1, \lambda) = \lambda$ (the second term is zero). Therefore, to maximize the objective, it must be the dominant pair ($\lambda = \lambda_1$, i.e., largest eigenvalue).

A similar approach with the additional constraint that $v_2^Tv_1 = 0$ can be used to show that v_2 is the second-largest eigenvector, and so on.

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).

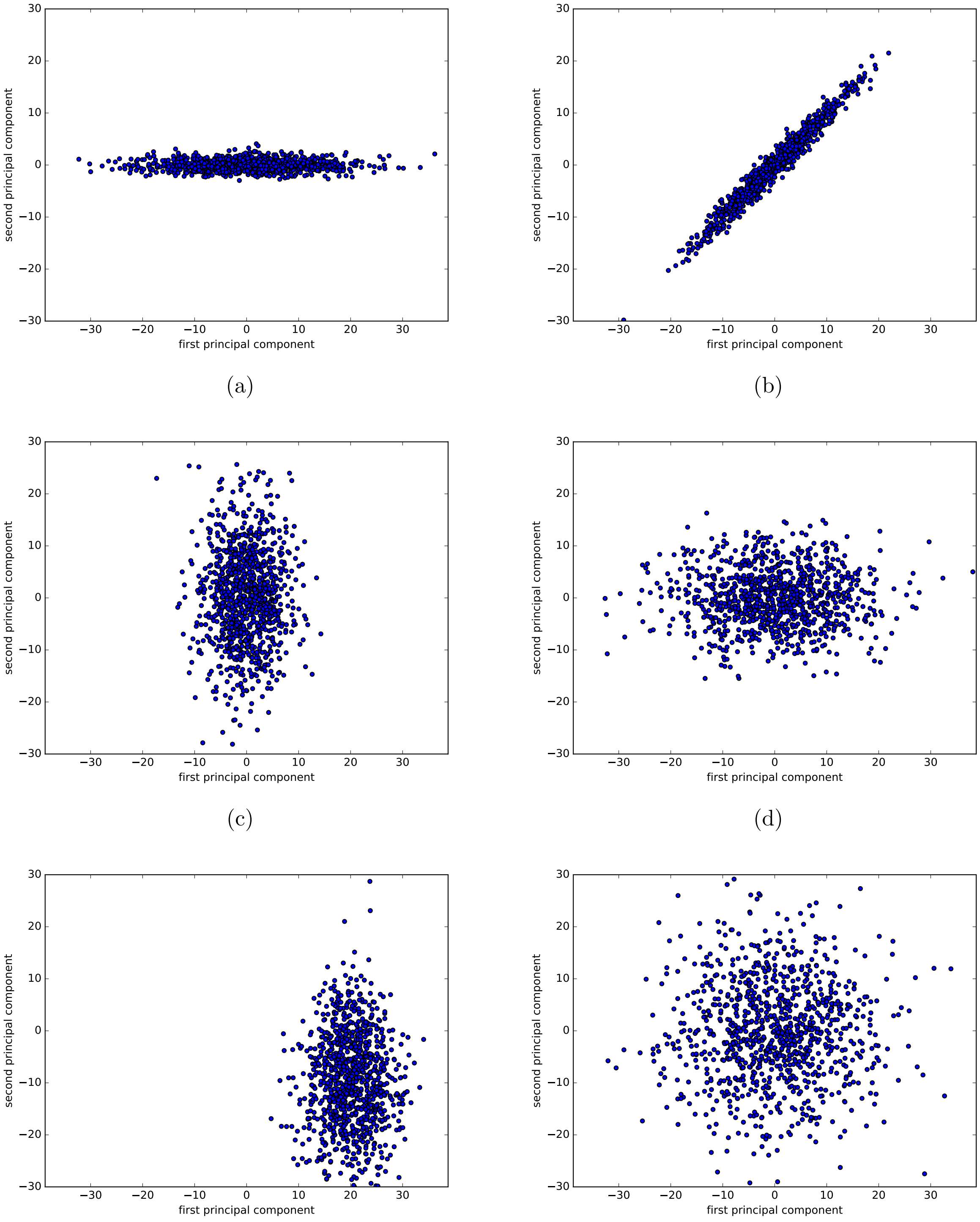
You are given the following six 2D plots as possible results.

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

The plots (b), (c), and (e) are not plausible.

- (b) the two principal components are not independent.
- (c) the second PC has larger variance than the first PC.
- (e) not centered.

The plots (a), (d) and (f) may arise for some choice of Σ – which is the topic of the next question.


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

(a) (b) (c) (d) (e) (f)

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.

$$\Sigma_1 = \begin{bmatrix} 100 & 0 \\ 0 & 30 \end{bmatrix} \tag{3}$$

$$\Sigma_2 = \begin{bmatrix} 50 & 49 \\ 49 & 50 \end{bmatrix} \tag{4}$$

$$\Sigma_3 = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 100 \end{bmatrix} \tag{5}$$

- (d) $= \Sigma_1$: this covariance matrix is diagonal, so we should see an axis-aligned ellipse with standard deviations on PC1 of 10, and on PC2 between 5 and 6. That's clearly plot (d).
- (f) $= \Sigma_3$: this covariance matrix is diagonal, and the first and third coordinate clearly dominate the variance. PCA will therefore simply project these two dimensions (with variance 100, so standard deviation of 10) onto PC1 and PC2. This results in the "circular" distribution in (f).
- (a) $= \Sigma_2$: this covariance matrix is almost singular (its determinant is small $(50^2 - 49^2)$), so its second eigenvalue would be much smaller than its largest EV. In terms of the distribution of samples, this should look a bit like a "cigar", ie, a narrow ellipse – as in (a).