Chapter: Dimensionality reduction

Proposed exercises

October 23, 2019

Contents

1 Notations

- If $A \in \mathbb{R}^{n \times m}$, then A_i (or a_i) is the i$^{\text{th}}$ column and $(A^\top)_i$ is the i$^{\text{th}}$ row as a column vector.
- If we have a data matrix X, we dispose the observations as columns, each row being an attribute. MUST REWRITE THE QUESTIONS IN EXERCISES IF IT'S OTHERWISE

2 Prerequisites

2.1 Linear Algebra in $\mathbb{R}^n$ Prerequisites

Solution: check the specific chapter in DP book or google the answer

2.1.1 Basics, Linear ..., Orthogonal ...

1. Compute the **determinant** and the **trace** of the following matrices: $M_1 = \begin{bmatrix} 1 & 1 \\ 3 & 4 \end{bmatrix}$, $M_2 = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 1 & 6 \\ 7 & 1 & 1 \end{bmatrix}$.

2. Given $u = \begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}$ and $v = \begin{bmatrix} 1 \\ 5 \\ 7 \end{bmatrix}$ compute the **dot product** (i.e., the Euclidian inner product; $\cdot$), the norm induced by this inner product (i.e., 2-norm or **Euclidian norm**), the **distance** induced by this norm (i.e., 2-norm distance or **Euclidian distance**), the **cosine of the angle**, the **outer product** ($\otimes$) between u and v.

3. (LA4ML. Review Packet 2. ex. 11b) Determine the **unit vector** that points in the same direction as the following vector: $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

4. (after DP) Determine whether the given matrix is in **row echelon form** (REF). If it is, state whether it is also in **reduced row echelon form** (RREF):

   $M_1 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 3 \\ 0 & 1 & 0 \end{bmatrix}$, $M_2 = \begin{bmatrix} 7 & 10 & 1 & 0 \\ 0 & 1 & -1 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $M_3 = \begin{bmatrix} 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$, $M_4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $M_5 = \begin{bmatrix} 1 & 0 & 3 & -4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 10 & 5 & 0 & 1 \end{bmatrix}$, $M_6 = \begin{bmatrix} 10 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, $M_7 = \begin{bmatrix} 1 & 20 & 30 \\ 1 & 0 & 0 \\ 0 & 1 & 10 \\ 0 & 0 & 1 \end{bmatrix}$, $M_8 = \begin{bmatrix} 2 & 10 & 30 & 50 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 30 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

5. (after DP) Use elementary row operations to reduce the given matrix to **row echelon form** and, then, to **reduced row echelon form**.

$$M_1 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},\ M_2 = \begin{bmatrix} 5 & 4 \\ 2 & 1 \end{bmatrix},\ M_3 = \begin{bmatrix} 3 & 5 \\ 5 & 2 \\ 2 & 4 \end{bmatrix},\ M_4 = \begin{bmatrix} 2 & 4 & -2 & 6 \\ 3 & 1 & 6 & 6 \end{bmatrix},$$
$$M_5 = \begin{bmatrix} -3 & 2 & -1 \\ 2 & -1 & -1 \\ 4 & -3 & -1 \end{bmatrix},\ M_6 = \begin{bmatrix} 2 & 4 & 7 \\ -3 & -6 & 10 \\ 1 & 2 & -3 \end{bmatrix}.$$

6. (DP) Solve the given system of equations using either **Gaussian or Gauss-Jordan elimination**, write the linear system in **parametric form** and specify the **free and pivot variables**:

$$(S_1): \begin{cases} -x_1 + 3x_2 - 2x_3 + 4x_4 = 0 \\ 2x_1 - 6x_2 + x_3 - 2x_4 = -3 \\ x_1 - 3x_2 + 4x_3 - 8x_4 = 2 \end{cases}, (S_2): \begin{cases} \frac{1}{2}x_1 + x_2 - x_3 - 6x_4 = 2 \\ \frac{1}{6}x_1 + \frac{1}{2}x_2 - 3x_4 + x_5 = -1 \\ \frac{1}{3}x_1 - 2x_3 - 4x_5 = 8 \end{cases},$$

$$(S_3): \begin{cases} \sqrt{2}x + y + 2z = 1 \\ \sqrt{2}y - 3z = -\sqrt{2} \\ -y + \sqrt{2}z = 1 \end{cases}, (S_4): \begin{cases} w + x + 2y + z = 1 \\ w - x - y + z = 0 \\ x + y = -1 \\ w + x + z = 2 \end{cases},$$

$$(S_5): \begin{cases} a + b + c + d = 4 \\ a + 2b + 3c + 4d = 10 \\ a + 3b + 6c + 10d = 20 \\ a + 4b + 10c + 20d = 35 \end{cases}$$

7. (DP)

   (a) Compute an **LU decomposition** (in two forms: **reduced (also called economy sized or thin)** and **full**) with $l_{ii} = 1$ of the following matrices and mention if it is **unique**: $M_1 = \begin{bmatrix} 2 & -4 \\ 3 & 1 \end{bmatrix}$, $M_2 = \begin{bmatrix} 2 & 2 & -1 \\ 4 & 0 & 4 \\ 3 & 4 & 4 \end{bmatrix}$, $M_3 = \begin{bmatrix} 1 & 2 & 0 & -1 & 1 \\ -2 & -7 & 3 & 8 & -2 \\ 1 & 1 & 3 & 5 & 2 \\ 0 & 3 & -3 & -6 & 0 \end{bmatrix}$, $M_4 = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 2 & -7 & 1 & 3 \\ 0 & 3 & 3 & -3 \\ -1 & 8 & 5 & -6 \\ 1 & -2 & 2 & 0 \end{bmatrix} = M_3^\top$, $M_5 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$

   (b) Can you compute an LU decomposition of the following matrix? If

not, compute a $P^\top LU$ **decomposition** of it: $M = \begin{bmatrix} 0 & 0 & 1 & 2 \\ -1 & 1 & 3 & 2 \\ 0 & 2 & 1 & 1 \\ 1 & 1 & -1 & 0 \end{bmatrix}$.

8. (DP)

   (a) Determine if the vector v is a **linear combination** of the remaining vectors: $v = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $u_1 = \begin{bmatrix} 4 \\ -2 \end{bmatrix}$, $u_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$

   (b) Determine whether v is in the **span** of the remaining vectors: $v = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $u_1 = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$, $u_2 = \begin{bmatrix} -1 \\ 1 \\ -3 \end{bmatrix}$.

9. (DP) Determine whether the following sets of vectors are **linearly independent**. If, for any of these, the answer can be determined by inspection (i.e., without calculation), state why. For any sets that are **linearly dependent**, find a dependence relationship among the vectors.

   (a) $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 3 \end{bmatrix}$

   (b) $\begin{bmatrix} 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} -8 \\ 1 \\ 10 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ -4 \\ -12 \\ 0 \end{bmatrix}$

   (c) $\begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix}$

   (d) $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$,

   (e) $\begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -5 \\ 2 \end{bmatrix}$

(f) $\begin{bmatrix}1\\1\\1\end{bmatrix}, \begin{bmatrix}4\\-1\\1\end{bmatrix}, \begin{bmatrix}2\\1\\0\end{bmatrix}, \begin{bmatrix}4\\0\\10\end{bmatrix}$

10. (DP) Show that the given vectors form an **orthogonal set**. Determine whether the set is orthonormal. If it is not, normalize the vectors to form an **orthonormal set**.

(a) $\begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\\\frac{1}{2}\end{bmatrix}, \begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\\-\frac{1}{2}\end{bmatrix}$

(b) $\begin{bmatrix}\frac{1}{2}\\\frac{1}{2}\\-\frac{1}{2}\\\frac{1}{2}\end{bmatrix}, \begin{bmatrix}0\\\frac{1}{3}\\\frac{2}{3}\\\frac{1}{3}\end{bmatrix}, \begin{bmatrix}\frac{1}{2}\\-\frac{1}{6}\\\frac{1}{6}\\\frac{1}{6}\end{bmatrix}$

11. (DP) Determine whether the given **matrix is orthogonal**. If it is, find its inverse.

(a) $\begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$

(b) $\begin{bmatrix}\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2}\\\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2}\\-\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\\\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2}\end{bmatrix}$

12. (DP) Determine whether the following vectors form a **basis** of a (sub)space. If the answer is positive, mention the corresponding **(sub)space**.

(a) $\begin{bmatrix}1\\-1\\3\end{bmatrix}, \begin{bmatrix}-1\\5\\1\end{bmatrix}, \begin{bmatrix}1\\-3\\1\end{bmatrix}$

(b) $\begin{bmatrix}1\\-1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\-1\end{bmatrix}, \begin{bmatrix}0\\0\\-1\\1\end{bmatrix}, \begin{bmatrix}-1\\0\\1\\0\end{bmatrix}$

(c) $\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\0\\1\end{bmatrix}$

(d) $\begin{bmatrix}1\\1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\1\\0\\1\end{bmatrix}, \begin{bmatrix}1\\0\\1\\1\end{bmatrix}$.

13. (DP) Find the **coordinate vector** of $v = \begin{bmatrix}1\\3\\4\end{bmatrix}$ with respect to the basis $\mathcal{B} = \left\{ \begin{bmatrix}3\\1\\4\end{bmatrix}, \begin{bmatrix}5\\1\\6\end{bmatrix} \right\}$.

14. (Gareth Williams) Find the **transition matrix** P from the given basis $\mathcal{B}$ to the basis $\mathcal{B}'$ of $\mathbb{R}^2$ (**change of basis**). Use the matrix to find the coordinate vector of u relative to $\mathcal{B}'$.

    (a) $\mathcal{B} = \left\{ \begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}3\\-1\end{bmatrix} \right\}, \mathcal{B}' = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \right\}, [u]_{\mathcal{B}} = \begin{bmatrix}3\\0\end{bmatrix}$

    (b) $\mathcal{B} = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \right\}, \mathcal{B}' = \left\{ \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}-1\\-1\end{bmatrix} \right\}, [u]_{\mathcal{B}} = \begin{bmatrix}8\\3\end{bmatrix}$

    (c) $\mathcal{B} = \left\{ \begin{bmatrix}3\\-1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix} \right\}, \mathcal{B}' = \left\{ \begin{bmatrix}2\\5\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix} \right\}, [u]_{\mathcal{B}} = \begin{bmatrix}7\\-2\end{bmatrix}$

15. (DP) Find a **basis for the span** of the given vectors: $\begin{bmatrix}1\\-1\\1\end{bmatrix}, \begin{bmatrix}1\\2\\0\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}2\\1\\2\end{bmatrix}$.

16. (DP) Find a **basis for $\mathbb{R}^3$ that contains** the vector $\begin{bmatrix}3\\1\\5\end{bmatrix}$.

17. (DP)

    (a) Show that $\mathbb{R}^2 = \text{span}\left(\begin{bmatrix}3\\-2\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \right)$.

    (b) Show that $\mathbb{R}^3 = \text{span}\left(\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}2\\1\\-1\end{bmatrix} \right)$.

18. (DP) Find the **orthogonal complement** $W^\top$ of W and give a basis for $W^\top$.

(a) $W = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : 2x - y + 3z = 0 \right\}$

(b) $W = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : x = \frac{1}{2}t, y = -\frac{1}{2}t, z = 2t \right\}$

(c) $W = \text{span}\left(\begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}, \begin{bmatrix} 4 \\ 0 \\ 1 \end{bmatrix} \right)$

(d) $W = \left\{ \begin{bmatrix} -y \\ x + 2y \\ 3x - 4y \end{bmatrix} : \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^2 \right\}$

19. (DP) Find the **matrix of the orthogonal projection** onto the subspace W. Then use this matrix to find the **orthogonal projection** of v onto W ($\text{proj}_W(v)$) and the **component of v orthogonal to** W ($\text{perp}_W(v)$).

    (a) $W = \text{span}\left(\begin{bmatrix} 1 \\ -2 \end{bmatrix} \right), v = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

    (b) $W = \text{span}\left(\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \right), v = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

20. (DP) Find the **orthogonal decomposition** of v with respect to W: $v = \begin{bmatrix} 3 \\ 2 \\ -3 \end{bmatrix}, W = \text{span}\left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \right)$.

21. (DP) Apply the **Gram-Schmidt Process** to the following basis to obtain an orthogonal (or even orthonormal) basis: $\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\}$.

22. (DP) Compute a **QR decomposition** (in two forms: **reduced (also called economy sized or thin)** and **full**) of the following matrices and mention if it is **unique**:

$$M_1 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \; M_2 = \begin{bmatrix} 1 & 2 & 2 \\ -1 & 1 & 2 \\ -1 & 0 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \; M_3 = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 2 \end{bmatrix} = M_2^\top,$$

$$M_4 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

23. (DP) Fill in the missing entries of Q to make Q an orthogonal matrix:

$$Q = \begin{bmatrix} \frac{1}{2} & \frac{2}{\sqrt{14}} & * & * \\ \frac{1}{2} & \frac{1}{\sqrt{14}} & * & * \\ \frac{1}{2} & 0 & * & * \\ \frac{1}{2} & -\frac{3}{\sqrt{14}} & * & * \end{bmatrix}$$

24. (DP) The given vectors form a basis (a subspace of $\mathbb{R}^n$). Obtain an **orthogonal basis** and an **orthonormal** basis.

    (a) $\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 4 \end{bmatrix}$

    (b) $\begin{bmatrix} 2 \\ -1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ -1 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

25. (DP + Gareth Will) Let $A = \begin{bmatrix} 2 & -4 & 5 & 8 & 5 \\ 1 & -2 & 2 & 3 & 1 \\ 4 & -8 & 3 & 2 & 6 \end{bmatrix}$. The **four fundamental subspaces of A are column space of A (col(A)), row space of A (row(A)), null space of A (null(A)), null space of $A^\top$ (null($A^\top$))**.

    (a) i. Find k such that null(A) is a subspace of $\mathbb{R}^k$.
    ii. Find k such that col(A) is a subspace of $\mathbb{R}^k$.

    (b) Find bases for the four fundamental subspaces of A.

    (c) i. Verify that every vector in row(A) is orthogonal to every vector in null(A).
    ii. Verify that every vector in col(A) is orthogonal to every vector in null($A^\top$).

    (d) Determine the **rank** and **nullity** of A.

2.1.2 Eigen...

26. (DP)

(a) Show that $x = \begin{bmatrix}4\\2\end{bmatrix}$ is an **eigenvector** of $A = \begin{bmatrix}4 & -2\\5 & -7\end{bmatrix}$ and find the corresponding **eigenvalue**.

(b) Show that 3 is an eigenvalue of $A = \begin{bmatrix}2 & 2\\2 & -1\end{bmatrix}$ and determine all eigenvectors corresponding to this eigenvalue. Find a basis for its **eigenspace**.

27. (after many sources; the most important ex. for PCA) For any of the following matrices answer to the questions:

(a) Compute the **eigenvalues** (Write down the **characteristic polynomial** and solve the **characteristic equation**). Are the eigenvalues in $\mathbb{R}$?

(b) Compute the **eigenspaces** (E_λ).

(c) Take one vector from each eigenspace and form a set. Verify that the set is **linearly independent**.

(d) Only for matrices that have only real eigenvalues and eigenvectors, take one vector from each eigenspace and form a set. Verify that the set is **orthogonal**.

(e) Compute the **algebraic and geometric multiplicities** of the eigenvalues.

(f) Is the matrix **diagonalizable**? If yes, compute an **eigendecomposition (or spectral decomposition)** of the matrix, **compute the matrix raised to the power of 2019 (M_i^{2019})**, and write the **spectral decomposition in the outer product form**. Is the **decomposition unique**?
(You do not have to calculate inverses, just write $\ldots^{-1}$.)

Complex eigenvalues:

$M_1 = \begin{bmatrix}-1 & -4\\1 & -1\end{bmatrix}$; Hint: $\lambda_i \in \{-1+2i, -1-2i\}$. Source:`http://www.math.utk.edu/~freire/complex-eig2005.pdf`

$M_2 = \begin{bmatrix}1 & -2 & -1\\1 & 3 & 1\\0 & 0 & 2\end{bmatrix}$; Hint: $\lambda_i \in \{2+i, 2-i, 2\}$. Source: `http://www.wright.edu/~chaocheng.huang/lecture/mth255/mth255lect13.pdf`

Algebraic multiplicity $\neq$ Geometric multiplicity:

$M_3 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$; Hint: $\lambda_i \in \{1\}$. Source: https://people.math.carleton.ca/~kcheung/math/notes/MATH1107/wk10/

$M_4 = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 2 & 1 \\ -1 & 2 & 2 \end{bmatrix}$; Hint: $\lambda_i \in \{1, 2\}$. Source: https://archive.uea.ac.uk/jtm/9/Lec9p7.pdf

Symmetric and non-symmetric matrices with 1, 2, 3 different eigenvalues:

$M_5 = \begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$; Hint: $\lambda_i \in \{3, 8\}$. Source: https://staff.csie.ncu.edu.tw/chia/Course/LinearAlgebra/sec8-1.pdf

$M_6 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{bmatrix}$; Hint: $\lambda_i \in \{-1, 1, 2\}$. Source: DP

$M_7 = \begin{bmatrix} 5 & -2 \\ 7 & -4 \end{bmatrix}$; Hint: $\lambda_i \in \{-2, 3\}$. Source: https://archive.uea.ac.uk/jtm/9/Lec9p7.pdf

$M_8 = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$; Hint: $\lambda_i \in \{3\}$.

$M_9 = \begin{bmatrix} 7 & 0 & 4 \\ 7 & 0 & 4 \\ 0 & 0 & 11 \end{bmatrix}$; Hint: $\lambda_i \in \{0, 7, 11\}$. Source: https://archive.uea.ac.uk/jtm/9/Lec9p7.pdf

$M_{10} = \begin{bmatrix} 4 & 0 & -2 \\ 2 & 5 & 4 \\ 0 & 0 & 5 \end{bmatrix}$; Hint: $\lambda_i \in \{4, 5\}$. Source: https://people.math.carleton.ca/~kcheung/math/notes/MATH1107/wk10

$M_{11} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & -1 \\ 0 & 1 & 0 \end{bmatrix}$; Hint: $\lambda_i \in \{1\}$. Source: https://math.stackexchange.com/questions/2244911/finding-eigenvectors-of-a-3x3-matrix-with-a-root-of-mult

$M_{12} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$

$M_{13} = \begin{bmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{bmatrix}$; Hint: $\lambda_i \in \{-1, 8\}$. Source: https://archive.uea.ac.uk/jtm/9/Lec9p7.pdf

$$M_{14} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

0 (zero) eigenvalue:

$$M_{15} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

28. (after DP)

    (a) Use the **power method** to approximate the dominant eigenvalue and dominant eigenvector of A. Use the given initial vector x_0, the specified number of iterations k, and three-decimal-place accuracy:

        i. $A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{bmatrix}$, $x_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $k = 6$, with 2-norm scaling.

        ii. $A = \begin{bmatrix} 0.2 & 0.4 \\ 0.8 & 0.6 \end{bmatrix}$, $x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $k = 6$, without scaling.

    (b) Regarding the above two matrices, compute the **percentage of the dominant eigenvalue from the sum of eigenvalues without computing the other eigenvalues**.

2.1.3 Singular...

29. (DP) Find the **singular values** of $A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \end{bmatrix}$.

30. (DP)

    (a) Find a **singular value decomposition (SVD)** (in two forms: **reduced (also called economy sized or thin)** and **full**) for the following matrices and mention if it is **unique**.

        i. $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

        ii. $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

        iii. $A = \begin{bmatrix} 0 & 0 \\ 0 & 3 \\ -2 & 0 \end{bmatrix}$

    (b) For each the matrices A above, mention:

- the **singular values**
- the **left singular vectors**
- the **right singular vectors**
- the **outer product form of the SVD**
- rank(A), nullity(A)
- an orthonormal basis for col(A)
- an orthonormal basis for row(A)
- an orthonormal basis for null(A)
- an orthonormal basis for null($A^\top$).

31. Compute the **eigendecomposition and the SVD** of the following matrix: $\begin{bmatrix} 4 & 2 \\ 2 & 7 \end{bmatrix}$. Are they **the same**?

32. (DP) Let $u_1 = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}$, $u_2 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$, and $v = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$. Find the **best approximation** to v in the plane $W = \text{span}(u_1, u_2)$ and find the Euclidian distance from v to W.

33. (DP)

    (a) Compute the **pseudoinverse (or Moore-Penrose inverse)** of the matrix A (i.e., A^+) via the **formula available only when A has linearly independent columns**.
    $A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \end{bmatrix}$.

    (b) Find a **least squares solution** to the inconsistent system $Ax = b$ by constructing and solving the **normal equations**, where $A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

    (c) A **QR factorization** of A is given. Use it to find a **least squares solution** of $Ax = b$:
    $A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ -1 & 1 \end{bmatrix}$, $Q = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{bmatrix}$, $R = \begin{bmatrix} \sqrt{6} & -\frac{\sqrt{6}}{2} \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

(d) Show that the **least squares solution of $Ax = b$ is not unique** and solve the normal equations to find all the least squares solutions, where $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ and $b = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$.

(e) Compute the **pseudoinverse (or Moore-Penrose inverse)** of the matrix A (i.e., A^+) via the **SVD**, where $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

(f) Compute the **minimal length least squares solution** to $Ax = b$, where $A = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ and $b = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$.

2.1.4 Inner product, norm, distance

34. (DP) Let $u = \begin{bmatrix} -1 \\ 1 \\ -5 \end{bmatrix}$ and $v = \begin{bmatrix} 2 \\ -2 \\ 0 \end{bmatrix}$.

(a) Compute the **1-norm** (or the sum norm or the taxicab norm or the Manhattan norm; $\|\cdot\|_1$) of u and the 1-norm of v. Compute the induced **distance** from the 1-norm between u and v ($d_1(u, v)$).

(b) Compute the **2-norm** (or the Euclidian norm; $\|\cdot\|_2$) of u and the 2-norm of v. Compute the induced **distance** from the 2-norm between u and v ($d_2(u, v)$).

(c) Compute the **∞-norm** (or the max norm or the uniform norm; $\|\cdot\|_\infty$) of u and the ∞-norm of v. Compute the induced **distance** from the ∞-norm between u and v ($d_\infty(u, v)$).

35. (DP) Let $A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Compute the following **matrix norms** and **condition numbers**:

(a) $\|A\|_{\text{Fro}}$ in two modes and $\text{cond}_{\text{Fro}}(A)$, knowing that the singular values of A are $\sqrt{2}$ and 1.

(b) $\|A\|_1$ and $\text{cond}_1(A)$, knowing that its pseudoinverse is $A^+ = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix}$.

(c) $\|A\|_2$ and $\text{cond}_2(A)$, knowing that the singular values of A are $\sqrt{2}$ and 1.

(d) $\|A\|_\infty$ and $\text{cond}_\infty(A)$, knowing that its pseudoinverse is $A^+ = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{bmatrix}$.

2.1.5 Hyperplanes

36. (after MIT, fall 2011 video "Problem Solving: Projection onto Subspaces")

    (a) The following is an equation of a **(linear) hyperplane** which represents a subspace of $\mathbb{R}^3$: $x + y + z = 0$.

        i. Specify a **normal vector to this hyperplane**.

        ii. Find the **orthogonal complement** of the given subspace **in two ways**.

        iii. Compute the Euclidian **distance from the point** $v = \begin{bmatrix} 8 \\ 0 \\ 9 \end{bmatrix}$ **to the hyperplane**.

    (b) The following is an equation of a **affine hyperplane** which represents an affine subspace of $\mathbb{R}^3$: $x + y + z = 2$.

        Having the new hyperplane, answer to the three questions of the exercise above.

2.1.6 Revision

37. (LA4ML - (from course - Jeopardy Round) or source is specified) Circle the correct answer.

    (a) How could I measure how far apart (i.e., how different) two observations, y_1 and y_2, are from each other?

        i. Compute $\|y_1 - y_2\|$

        ii. Compute $\|y_2 - y_1\|$

        iii. Compute $\|y_1\| - \|y_2\|$

        iv. Compute covariance(y_1,y_2)

        v. Either 37(a)i or 37(a)ii

        Solution: E

    (b) What is the span of one vector in $\mathbb{R}^3$?

        i. A plane

        ii. A line

        iii. The whole 3-dimensional space

        iv. A point

        v. A vector

        Solution: B

(c) What is the span of two linearly independent vectors in $\mathbb{R}^3$?

i. A plane
ii. A line
iii. The whole 3-dimensional space
iv. A point
v. A vector

Solution: A

(d) For 3 vectors, x, y and z, suppose that $2x + 3y + 5z = 0$

i. Then x, y and z are linearly independent.
ii. Then x, y and z are linearly dependent.
iii. Then x, y and z are orthogonal.
iv. None of the above.

Solution: B

(e) If a collection of vectors is mutually orthogonal, then those vectors are linearly independent.

i. True
ii. False

Solution: A

(f) If U is an orthogonal matrix, then:

i. $U^\top U = UU^\top = I$
ii. $U^\top$ is the inverse of U
iii. U is a covariance matrix
iv. $U^\top U = 0$
v. Both 37(f)i and 37(f)ii.

Solution: E

(g) If the span of 3 vectors x, y, and z is a 2-dimensional subspace (a plane) then ...

i. x, y, and z are linearly dependent
ii. x, y, and z are linearly independent
iii. x, y, and z are orthogonal
iv. x, y, and z are all multiples of the same vector

Solution: A

(h) In order for a matrix to have eigenvalues and eigenvectors, what must be true?

i. All matrices have eigenvalues and eigenvectors
ii. The matrix must be square
iii. The matrix must be orthogonal
iv. The matrix must be a covariance matrix

Solution: B

(i) If I multiply a matrix A by its eigenvector x, what can I say about the result, Ax?

i. The result is a unit vector
ii. The result is a scalar, which is called the eigenvalue
iii. The result is a scalar multiple of x
iv. The result is orthogonal

Solution: C

(j) (CMU, Scalable ML, BPoczos, 2018f, HW1, ex. 1.28) If A is a symmetric n by n real matrix, then the eigenvalues are all real.

i. True
ii. False

Solution: A

(k) (CMU, Scalable ML, BPoczos, 2018f, HW1, ex. 1.29) If A is a symmetric n by n real matrix, then A can be written in the form of $A = U^\top \Lambda U$ where the rows of U are an orthonormal set of eigenvectors for A and Λ is a diagonal matrix of eigenvalues for A.

i. True
ii. False

Solution: A

(l) (Final CS 189 Spring 2013 Introduction to Machine Learning, ex. Q1.(v)) The eigenvectors of $AA^\top$ and $A^\top A$ are the same.

i. True
ii. False

Solution: False

(m) (Final CS 189 Spring 2013 Introduction to Machine Learning, ex. Q1.(w)) The non-zero eigenvalues of $AA^\top$ and $A^\top A$ are the same.

i. True

ii. False

Solution: True

(n) (Final CS 189 Spring 2013 Introduction to Machine Learning, ex. Q2.(d)) The left singular vectors of a matrix A can be found in ...

i. Eigenvectors of $A^\top A$
ii. Eigenvectors of A^2
iii. Eigenvectors of $A^\top A$
iv. Eigenvectors of $AA^\top$

Solution: A

(o) (Final CS 189 Spring 2013 Introduction to Machine Learning, ex. Q2.(f)) Let A be a symmetric matrix and S be the matrix containing its eigenvectors as column vectors, and D a diagonal matrix containing the corresponding eigenvalues on the diagonal. Which of the following are true:

i. $AS = SD$
ii. $SA = DS$
iii. $AS = DS$
iv. $AS = DS^\top$

Solution: A

(p) (Final CS 189 Spring 2014 Introduction to Machine Learning, ex. Q1.1) The singular value decomposition of a real matrix is unique.

i. True
ii. False

Solution: False

(q) (Final CS 189 Spring 2015 Introduction to Machine Learning, ex. Q1.(k)) Given any matrix X, its singular values are the eigenvalues of $XX^\top$ and $X^\top X$.

i. True
ii. False

Solution: False

(r) (Final CS 189 Spring 2015 Introduction to Machine Learning, ex. Q1.(l)) Given any matrix X, $(XX^\top + \lambda I)^{-1}$ for $\lambda \neq 0$ always exists.

i. True
ii. False

Solution: False

(s) (CS 189 Spring 2017 Introduction to Machine Learning Final, ex. Q1.(4)) With the SVD, we write $X = UDV^\top$. For which of the following matrices are the eigenvectors the columns of U?

i. $X^\top X$
ii. $XX^\top$
iii. $X^\top XX^\top X$
iv. $XX^\top XX^\top$

Solution: B

(t) (CS 189 Spring 2017 Introduction to Machine Learning Final, ex. Q1.(6)) Consider the matrix $X = \sum_{i=1}^{r} \alpha_i u_u v_i^\top$ where each α_i is a scalar and each u_i and v_i is a vector. It is possible that the rank of X might be

i. r+1
ii. r-1
iii. r
iv. 0

Solution: B,C,D

(u) (CS 189 Spring 2019 Introduction to Machine Learning Midterm, ex. Q1.(a)) Let A be a real, symmetric $n \times n$ matrix. Which of the following are true about A's eigenvectors and eigenvalues?

i. A can have no more than n distinct eigenvalues
ii. The vector 0 is an eigenvector, because $A0 = \lambda 0$
iii. A can have no more than $2n$ distinct unit-length eigenvectors
iv. We can find n mutually orthogonal eigenvectors of A

Solution: A,D

There can be infinitely many unit-length eigenvectors if the multiplicity of any eigenvector is greater than 1 (so the eiganspace is a plane, and you can pick any vector on the unit circle on that plane). The 0 vector is not an eigenvector by definition.

(v) (CS 189 Spring 2019 Introduction to Machine Learning Midterm, ex. Q1.(b)) The matrix that has eigenvector $[1, 2]^\top$ with eigenvalue 2 and eigenvector $[-2, 1]^\top$ with eigenvalue 1 (note that these are not unit eigenvectors!) is

i. $\begin{bmatrix} 9 & -2 \\ -2 & 6 \end{bmatrix}$

ii. $\begin{bmatrix} 6 & 2 \\ 2 & 9 \end{bmatrix}$

iii. $\begin{bmatrix} 9/5 & -2/5 \\ -2/5 & 6/5 \end{bmatrix}$

iv. $\begin{bmatrix} 6/5 & 2/5 \\ 2/5 & 9/5 \end{bmatrix}$

Solution: D

(w) (CS189 Spring 2018 midterm, ex. 8.a) Let $X \in \mathbb{R}^{n \times d}$ with $n \geq d$. Suppose $X = U\Sigma V^\top$ is the singular value decomposition of X where $\sigma_i = \Sigma_{i,i}$ are the diagonal entries of Σ and satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ while u_i and v_i are the i-th columns of U and V respectively. Which of the following is the rank k approximation to X that is best in the Frobenius norm. That is, which low rank approximation, X_k, for X yields the lowest value for $\|\cdot\|_{\text{Fro}}^2$?

i. $\sum_{i=1}^{k} \sigma_i u_i v_{n-i}^\top$

ii. $\sum_{i=1}^{k} \sigma_i u_i v_i^\top$

iii. $\sum_{i=d-k+1}^{d} \sigma_i u_i v_i^\top$

iv. $\sum_{i=1}^{k} \sigma_i u_{n-i} v_i^\top$

Solution: The solution comes from the Eckhart-Young theorem which states we should use the singular vector expansion from 1 to k, $\sum_{i=1}^{k} \sigma_i u_i v_i^\top$

(x) (CS246 Final Exam March 16, 2016, ex. 20.5) If M is sparse, the SVD of M is guaranteed to be sparse as well.

i. True

ii. False

Solution: False

(y) (CS246 Final Exam March 16, 2016 ex. 20.4) The singular values of a matrix are unique (except for 0's).

i. True

ii. False

Solution: False

(z) (EPFL final exam 2018, ex. 12) Which of the following statements about the SVD of an $N \times D$ matrix X are correct?

i. We can compute the singular values of X by computing the eigenvalues of $XX^\top$. This has complexity $O(N^3)$.

ii. We can compute the singular values of X by computing the eigenvalues of $XX^\top$. This has complexity $O(D^3)$.

iii. We can compute the singular values of X by computing the eigenvalues of $X^\top X$. This has complexity $O(N^3)$.

iv. We can compute the singular values of X by computing the eigenvalues of $X^\top X$. This has complexity $O(D^3)$.

v. We can compute the singular values of X by computing the eigenvalues of $XX^\top$ if and only if X is a square matrix. This has complexity $O(D^3) = O(N^3)$.

Solution: We can compute the singular values of X by computing the eigenvalues of $X^\top X$. This has complexity $O(D^3)$. We can compute the singular values of X by computing the eigenvalues of $X^\top X$. This has complexity $O(N^3)$.

2.2 Probability and Statistics Prerequisites

38. Let $X = \begin{bmatrix} 1 & 2 & 1.3 \\ 1 & 3 & 3.2 \end{bmatrix}$ and $Y = \begin{bmatrix} 0 & 1 & 1 \\ 0 & -1 & 3 \\ 1 & 1 & 2 \end{bmatrix}$ be two data matrices (row = attribute, column = observation).

(a) Compute the **sample mean** for each data matrix.

(b) Compute the **sample covariance matrix** for each data matrix.

(c) Compute the **sample cross-covariance matrix** between X and Y.

39. (a) Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a **random vector**, with:

$$E[X] = E\left[\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\right] = \begin{bmatrix} E[X_1] \\ E[X_2] \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$E[XX^\top] = E\left[\begin{bmatrix} X_1^2 & X_1X_2 \\ X_2X_1 & X_2^2 \end{bmatrix}\right] = \begin{bmatrix} E[X_1^2] & E[X_1X_2] \\ E[X_2X_1] & E[X_2^2] \end{bmatrix} = \begin{bmatrix} 20 & 1 \\ 1 & 10 \end{bmatrix}$$

i. Compute the **covariance matrix** of X: $\mathrm{Cov}(X)$.

ii. Compute $\mathrm{E}[AX]$, where $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$.

(b) Let $X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ and $Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ be random vectors, with:

$$E[X] = E\left[\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}\right] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ E[X_3] \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$E[Y] = E\left[\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}\right] = \begin{bmatrix} E[Y_1] \\ E[Y_2] \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$E[XY^\top] = E\left[\begin{bmatrix} X_1Y_1 & X_1Y_2 \\ X_2Y_1 & X_2Y_2 \\ X_3Y_1 & X_3Y_2 \end{bmatrix}\right] = \begin{bmatrix} E[X_1Y_1] & E[X_1Y_2] \\ E[X_2Y_1] & E[X_2Y_2] \\ E[X_3Y_1] & E[X_3Y_2] \end{bmatrix} = \begin{bmatrix} 7 & 0.2 \\ 0.3 & 6 \\ 1 & 2 \end{bmatrix}$$

i. Compute the **cross-covariance matrix** for X and Y: $\mathrm{Cov}(X, Y)$.

ii. Compute $\mathrm{Cov}(AX, BY)$, where $A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$.

40. (CS189 spring 2013 midterm, ex.Q2.c) Which of the following transformations when applied on $X \sim \mathcal{N}(\mu, \Sigma)$ transforms it into an axis aligned Gaussian? ($\Sigma = UDU^\top$ is the spectral decomposition of Σ)

(a) $U^{-1}(X - \mu)$
(b) $(UD)^{-1}(X - \mu)$
(c) $UD(X - \mu)$
(d) $(UD^{1/2})^{-1}(X - \mu)$
(e) $U(X - \mu)$

Solution: A,B,D

41. (after LA4ML - Book(course) chapter 2, ex.1.e) Let $u = \begin{bmatrix} 1 \\ 2 \\ -4 \\ -2 \end{bmatrix}$ and $v = \begin{bmatrix} 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}$. Suppose these vectors are observations on four independent variables, which have the following covariance matrix:

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Determine the **Mahalanobis distance** between u and v.

2.3 Matrix factorizations: the Machine Learning context

2.3.1 Generalities

42. (after LA4ML. Introduction to Vector Space Models - Worksheet. Part Three, ex.1 AND LA4ML - slides, Chapter 7) Interpret the following Non-negative Factor Output for a small collection of text documents, answering the following questions:

    (a) What meaning (theme/topic) would you give to each of the two factors?

    (b) What is the dominant factor (theme/topic) for each document?

    (c) What is the loading of the word *dog* on Factor 2?

    (d) What is the coordinate/score of document 4 along Factor 2?

    (e) Interpret the factorization as a **soft clustering** technique applied to attributes, but also to instances.

$$\text{TermDocMatrix} \approx \begin{array}{c} \\ \text{"cat"} \\ \text{"dog"} \\ \text{"tired"} \\ \text{"injured"} \\ \text{"ankle"} \\ \text{"sprained"} \end{array} \begin{array}{c} \begin{array}{cc} \text{Factor1} & \text{Factor2} \end{array} \\ \begin{bmatrix} 1.0 & 0 \\ 1.6 & 0 \\ 0.4 & 0.4 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} \text{doc1} & \text{doc2} & \text{doc3} & \text{doc4} \end{array} \\ \begin{bmatrix} 1.0 & 1.7 & 0 & 0 \\ 0 & 0.1 & 0.9 & 1.1 \end{bmatrix} \end{array}$$

Observation: Change the text of the exercise after researching the fact that indeed the left matrix can be generally called the loading matrix and the right matrix can be called the score matrix. I think that this terminology is only for Factor Analysis, not for any matrix factorization.

43. (after `https://www.fbbva.es/wp-content/uploads/2017/05/dat/greenacre_c01_2010.pdf`) *True or False*: The *bi* in biplot refers to the fact that the display is usually two-dimensional, not to the fact that two sets of points (i.e., the rows and columns of the target matrix) are visualized by scalar products.

    Solution: False

44. (EPFL final-exam-2017, ex.20) Is it true that K-means can be equivalently written as the following matrix factorization problem? Here X denotes

the $N \times D$ data matrix. The μ_k denote columns of M, rows of Z, and $L(z, \mu) = \|X^\top - MZ^\top\|_{\text{Fro}}$.

$$\min_{z,\mu} L(z, \mu)$$

$$\text{s.t.} \mu_k \in \mathbb{R}^D,$$

$$z_{nk} \in \{0, 1\}, \sum_{k=1}^{K} z_{nk} = 1$$

Solution: Yes

45. (EPFL 2018 final, ex.15) Consider optimizing a matrix factorization $WZ^\top$ in the matrix completion setting, for $W \in \mathbb{R}^{D \times K}$, for $W \in \mathbb{R}^{D \times K}$. We write Σ for the set of observed matrix entries. Which of the following statements are correct?

    (a) Given any Σ, for $K := \min\{N, D\}$, there is an exact solution to the problem

    (b) In general, a step of SGD will change all entries of the W and Z matrices

    (c) Adding a Frob-norm regularizer for W and Z to the matrix factorization objective function makes the objective convex.

    (d) A step of alternating least squares is more costly than an SGD step.

    (e) For complete observations $\Sigma = [1 \dots D] \times [1 \dots N]$, the problem can be solved by the singular value decomposition.

    (f) The cost of an SGD step depends on the number of observed entries.

    Solution: A step of alternating least squares is more costly than an SGD step. Given any Ω, for $K := \min\{N, D\}$, there is an exact solution to the problem. For complete observations $\Omega = [1 \dots D] \times [1 \dots N]$, the problem can be solved by the singular value decomposition.

2.3.2 Singular Value Decomposition - SVD

46. (CS168 Final Exam - spring 2017, ex.6.a) Consider the following grayscale image of the Moon. Draw a sketch of your best guess for what the best rank-1 approximation of the image would look like in the space indicated below. Assume that the representation of the image stores black as zero.

Figure 1: The Moon

Figure 2: Sketch the best rank-1 approx. here!

47. (CS168 final exam - spring 2017, ex.6.b-e)

    (a) Netflix has an enormous dataset of ratings of movies given by its users, and it needs your help to analyze it. Assume there are 10000 users and 1000 movies, and each user has rated every movie. This data can be represented as a 10000 $\times$ 1000 dimensional matrix M, where each

entry $M(i; j)$ denotes the rating given by user i to movie j. Suppose you perform an SVD $M = UDV^\top$ of the matrix M. What concepts might be captured by the top few right singular vectors (the top singular vectors are the ones corresponding to the largest singular values)? What concepts might be captured by the top few left singular vectors?

(b) Continuing the previous part, you carry out a SVD of M, and observe that the top right singular vectors do seem to correspond to interesting concepts. You want to perform a low-dimensional projection of the 10,000 users, to visualize them in the 2-dimensional space spanned by the top 2 right singular vectors. How can you use the SVD to directly find the projection of each user (i.e. each row of the matrix M) onto the top 2 right singular vectors, without actually computing the inner product?

(c) Assume now that we are in the more realistic setting where all the 10000×1000 ratings corresponding to every (user,movie) pair are not available. Assume that 10% of the entries are missing. Low rank approximation can be used in this case to complete the matrix M to infer the missing entries. You use the following algorithm to infer the missing entries: 1) You first set them to be the average overall rating to obtain a new matrix $\hat{M}$. 2) You find a rank-k approximation $\tilde{M}$ of the matrix $\hat{M}$. 3) You set the missing entries to be their value in $\tilde{M}$. Explain: i) How and ii) Why can you use the top k singular vectors to find a rank-k approximation.

(d) Using the approach outlined in the previous part, how do you expect the error in estimating the missing entries to vary as a function of k? Explain.

48. (CMU, 15-826 - Multimedia databases and data mining, Spring 2008, Homework 3, ex.Q6) PRACTICAL EXERICISE (**Known as PureSVD in recc. systems literature**) When there are some entries in the matrix missed, we can use the reconstructed value of SVD to estimate them. Consider the following 2-d data set `http://www.cs.cmu.edu/~htong/15826-S08/hw3/missing.data` (the missing values are denoted by NA), where each line is a data point in 2-d space, you are asked to run SVD to recover the missing values.

    What to Hand in:

    (a) Treat the missing values (i.e., 'NA') as zeros and give the scatter-plot the whole dataset. Use dots ('.') for the existing points, and stars ('*') for the points with missing values.

(b) Now, run the SVD on the whole dataset (by treating the 'NA' as zeros). We can use the reconstructed values (i.e., the rank-1 approximation) as the prediction for the missing values. Give the scatter-plot again, with dots '.' for the existing points, and circles 'o' for the reconstructed ones. What is your prediction for the missing values?

(c) A possible improvement for the above technique is to do it in a recursive way, - to (1) use the reconstructed values as the initial prediction for missing value, (2) do SVD one more time and (3) use the new reconstructed values as a 'better' prediction and so on.

(d) This problem is related with the ratio rules we saw in the class, where we only have one missing value;

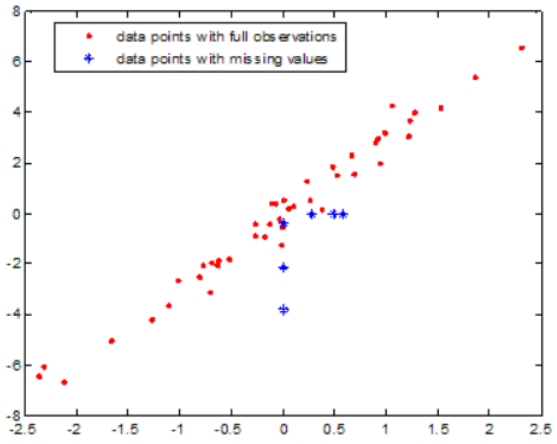
(e) A (much more elaborate) version of this technique is leading the Netflix competition. See the paper by Yehuda Koren `http://www.research.att.com/~yehuda/pubs/cf.pdf` for details.

Observations for PureSVD (first two subpoints):

- values are highly biased towards 0
- not good for rating prediction
- its not a big problem for ranking task

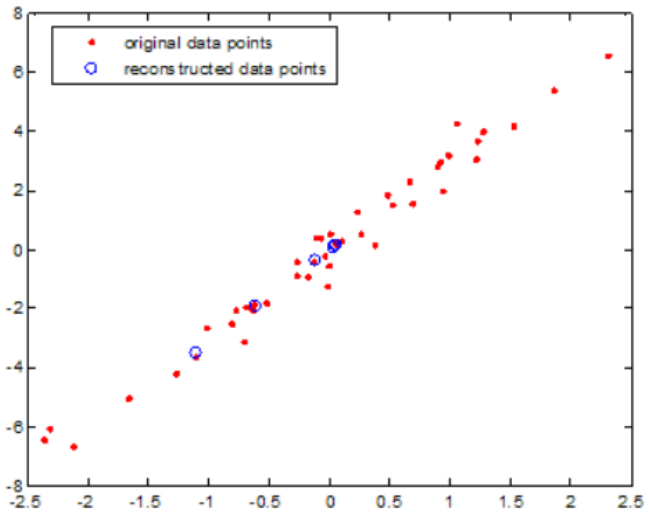
(from `https://www.coursera.org/lecture/machine-learning-applications-big-data/recsys-svd-ii-3cR7d`)

Solution:



(a)

(b)



Index	Reconstructed values
$\text{Data}_{3,x}$	-1.12
$\text{Data}_{10,x}$	-0.61
$\text{Data}_{19,y}$	0.17
$\text{Data}_{20,x}$	-0.12
$\text{Data}_{31,y}$	0.14
$\text{Data}_{50,y}$	0.08

49. (CMU, 15-826 - Multimedia databases and data mining, Spring 2008, Homework 3, ex.Q5) PRACTICAL EXERCISE SVD for latent semantic index

    Consider the following Document by Term matrix `http://www.cs.cmu.edu/~htong/15826-S08/hw3/docterm.data` with some noise. In this file each row denotes a document, and the columns denote terms. An entry (i,j) denotes the number of occurrences of the jth term in the ith document. Based on their occurrence in the documents, the terms can be clustered as belonging to a particular topic. Run SVD to determine the number of topics present. You may use any standard package (e.g., Matlab, R, mathematica, python/Numeric/numpy, e.t.c.).

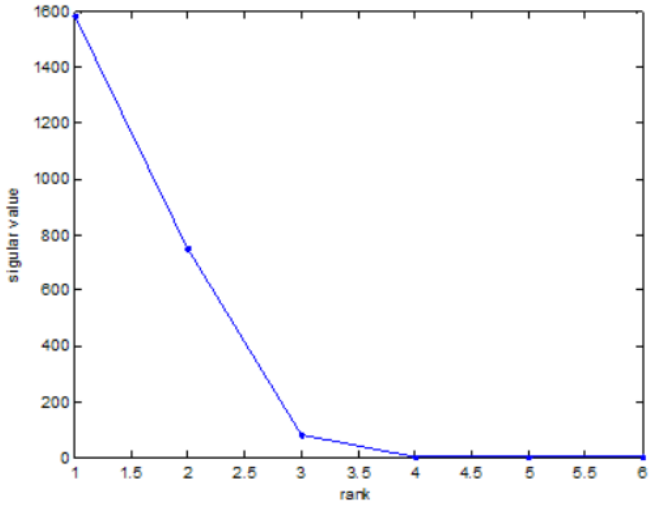
    What to Hand in:

    (a) Give your estimation of the number of topics present in the dataset, together with your justification.

    (b) Give the top 5 terms for each topic.

    Solution:

(a) Based on the singular-value vs. rank plot, we conclude there are 3 topics:



(b)

Topic 1	Topic 2	Topic 3
Term 10	Term 1	Term 2
Term 6	Term 7	Term 2
Term 4	Term 5	Term 15
Term 3	Term 9	Term 13
Term 1	Term 14	Term 11

2.3.3 Latent Semantic Indexing/Analysis - LSI/LSA

50. (CS168, Spring 2016 Final Exam, ex.7) Your aunt runs a small local news website, with articles falling into four different categories: Politics, Sports, Crime, and Weather. Fresh from CS168, you decide to help her out and spend a few weeks in the summer to add a 'Related Articles' feature to the website which would suggest articles similar to the article being currently read, so that people spend more time on the website. You build a large document-word matrix M, where the rows index articles and the columns index words. Each entry in the matrix $M[i; j]$ represents how many times word j appears in article i.

    We now consider applying the SVD to understand the structure of the document-word matrix M. Consider computing the SVD $M = USV^\top$. We will try and understand what the columns of U and V represent.

As a very simple example, suppose the document-word matrix M is the following:

$$M = \begin{bmatrix} 5 & 5 & 0 & 0 & 0 \\ 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 5 & 5 \\ 0 & 0 & 5 & 5 & 5 \\ 0 & 0 & 5 & 5 & 5 \end{bmatrix}$$

(a) What is the rank of M?

(b) Compute the SVD of the 4×5 matrix M given above. Specifically, if r is the rank of M, then write matrices U, S and V such that $M = USV^\top$, where U is a $4 \times r$ matrix, D is a $r \times r$ matrix and $V^\top$ is a $r \times 5$ matrix.

(c) Now imagine an actual (large) real-world document-word matrix. Suppose that, in general, articles on the same topic will use similar vocabularies, and that the different topics (Sports, Politics, Crime, Weather) will each use significantly different vocabularies. Of course, common words like "the" and "and" will occur often in all documents. What would you expect the columns of U corresponding to the largest singular values to represent for your news website articles? What you would expect the columns of V corresponding to the largest singular values to represent? (Write at most 2 sentences for each part. Do not just say that "they are the singular vectors" — you should give concrete and meaningful potential interpretations.)

51. (Example exam Multimedia Retrieval 2016-2017, Utrecht University, ex.5)

(a) What is the relationship between LSA (Latent Semantic Analysis) and SVD (Singular Value Decomposition)?

(b) Assume we have a corpus (i.e., a set) of 3 documents (i.e., d_1, d_2 and d_3) containing a total of 4 words (i.e., w_1, w_2, w_3 and w_4). For this corpus, the following word-by-document (or term-by-document) table has been computed, where the cells contain the word frequencies:

$$\begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \end{array} \begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\ \begin{bmatrix} 0 & 2 & 1 \\ 4 & 0 & 2 \\ 0 & 3 & 0 \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

SVD is supplied to the original word-by-document table above, yielding a decomposition $USV^\top$ with U, S and V rounded off for the purpose of this exercise as:

$$U = \begin{bmatrix} 0.2 & 0.3 & 0.6 \\ 0.4 & -0.7 & 0.3 \\ 0.2 & 0.6 & 0.7 \\ 0.9 & 0 & -0.2 \end{bmatrix}, S = \begin{bmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}, V = \begin{bmatrix} 0.8 & 0.5 & -0.3 \\ 0.6 & -0.7 & 0.2 \\ 0.1 & 0.3 & 1.0 \end{bmatrix}$$

Assume the corpus deals almost exclusively with two topics. Is it then reasonable to assume the the three dimensional document space can be mapped into a two dimensional semantic space?

(c) Suppose a user searches for information in the corpus using the words $w_1 w_2$ as query. Show where the concept expressed by these words would lie in an (appropriately chosen) semantic space, by calculating the coordinates of the query w1w2 in that space. (If you completely simplify your outcome, which is optional, round it off to no more than one decimal). Can the coordinates in the semantic space have negative values?

3 Principal Component Analysis - PCA

3.1 PCA

52. (CMU, 2012f, EXing, ASingh, HW4, pr.4) 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 $X_1, \ldots, X_n$ in p-dimensional space, and let X be the $n \times p$ matrix representing these points. Assume that the data points are centered, ie, $1^\top X = 0$. 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).

    (a) Argue that the projection is given by Xv.

    (b) What is the sample mean of all the points after the projection?

    (c) What is the sample variance of all the points after the projection?

    (d) Setup the problem of maximizing the sample variance of the projection onto v subject to a constraint on the L2-norm of v.

    (e) Solve the minimization problem to show that the solution is the rst PC. (Hint: take the Lagrangian of the above problem, differentiate and substitute to zero, to get to the optimum solution)

    (f) 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.

    Argue that the reconstruction of X_i using v is $(X_i^\top v)v$.

    (g) 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?

    (h) What is the total squared reconstruction error over all points?

    (i) Show that minimizing total squared reconstruction error is equivalent to minimizing $-\|Xv\|_2^2$.

    (j) 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)

    Solution:

(a) 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 $X = [X_1^\top, X_2^\top, \ldots, X_n^\top] = [X_i^\top]$ we get

$$Xv = [(X_{iv} + X_{iv'})]v = [X_{iv}^\top]v + [X_{iv'}^\top]v = [X_{iv}^\top]v = X_v v$$

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.

(b) Sample mean after projection is given by

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

(c) Given that the sample mean is zero we can write the variance as

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

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

(d)

$$\max_v v^\top \Sigma v$$
$$st. \|v\|^2 = 1$$

(e) By stationarity, at optimality we have

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

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

(f) 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:

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

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

(g)

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

(h)
$$\|(X^\top v)v - X\|_{\text{Fro}}^2$$

(i)
$$\begin{aligned}\|(X^\top v)v - X\|_{\text{Fro}}^2 &= \text{tr}(((X^\top v)v - X)^\top((X^\top v)v - X)) \\ &= \text{tr}(vv^\top X^\top Xvv^\top) - 2\text{tr}(vv^\top X^\top X) + \text{tr}(X^\top X) \\ &= \text{tr}(v^\top X^\top Xvv^\top v) - 2\text{tr}(v^\top V^\top Xv) + \text{tr}(X^\top X) \\ &= \text{tr}(v^\top X^\top Xv) - 2\text{tr}(v^\top X^\top Xv) + \text{tr}(X^\top X) \\ &= -\text{tr}(v^\top X^\top Xv) + \text{tr}(X^\top X) \\ &= -\|Xv\|_2^2 + \|X\|_2^2\end{aligned}$$

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

(j) The optimization problem is the same as in the 5-th part.

53. Suppose that our dataset consists of two points: $\begin{bmatrix}0\\0\end{bmatrix}$ and $\begin{bmatrix}2\\2\end{bmatrix}$. Draw the points on a grid and find the new coordinates of points **(principal components scores) using mainly the plot** you just drew.

54. (CS 189 Spring 2016 Introduction to Machine Learning Final, ex.Q5) You are given a design matrix $X = \begin{bmatrix}6 & -4\\ -3 & 5\\ -2 & 6\\ 7 & -3\end{bmatrix}$. Let's use PCA to reduce the dimension from 2 to 1.

    (a) Compute the covariance matrix for the sample points. (Warning: Observe that X is not centered.) Then compute the unit eigenvectors, and the corresponding eigenvalues, of the covariance matrix. Hint: If you graph the points, you can probably guess the eigenvectors (then verify that they really are eigenvectors).

    (b) Suppose we use PCA to project the sample points onto a one-dimensional space. What one-dimensional subspace are we projecting onto? For each of the four sample points in X (not the centered version of X!), write the coordinate (in principal coordinate space, not in $\mathbb{R}^2$) that the point is projected to.

    (c) Given a design matrix X that is taller than it is wide, prove that every right singular vector of X with singular value σ is an eigenvector of the covariance matrix with eigenvalue σ^2.

Solution:

(a) The covariance matrix $X^\top X = \begin{bmatrix} 82 & -80 \\ -80 & 82 \end{bmatrix}$.

Its unit eigenvectors are $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ with eigenvalue 2 and $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ with eigenvalue 162. (Note: either eigenvector can be replaced with its negation.)

(b) We are projecting onto the subspace spanned by $\begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$. (Equivalently, onto the space spanned by $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$). Equivalently, onto the line $x + y = 0$.) The projections are $(-6, 4) \to \frac{10}{\sqrt{2}}$, $(-3, 5) \to -\frac{8}{\sqrt{2}}$, $(-2, 6) \to -\frac{8}{\sqrt{2}}$, $(7, -3) \to \frac{10}{\sqrt{2}}$.

(c) If v is a right singular vector of X, then there is a singular value decomposition $X = UDV^\top$ such that v is a column of V. Here each of U and V has orthonormal column, V is square, and D is square and diagonal. The covariance matrix is $X^\top X = VDU^\top UDV^\top = VD^2V^\top$. This is an eigendecomposition of $X^\top X$, so each singular vector in V with singular value σ is an eigenvector of $X^\top X$ with eigenvalue σ^2.

55. (CMU, 2008f, Exing, final, pr.3.1) Given 3 data points in 20d space: (1,1), (2,2) and (3,3).

(a) what is the first principal component?

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

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

Solution:

(a) pc = $(1/\sqrt{2}, 1/\sqrt{2})^\top = (0.707, 0.707)^\top$ (the negation is also correct)

(b) $4/3 = 1.33$

(c) 0

56. (CMU, 2014s, SKim, HW3, pr. 1.3.4-5) Assume you are given 4 data points in $\mathbb{R}^3$, $(4,4,0)$, $(5,5,0)$, $(0,0,1)$, $(0,0,2)$, represented in the following matrix where each row corresponds to a data point:

$$\begin{bmatrix} 4 & 4 & 0 \\ 5 & 5 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \end{bmatrix}$$

(a) What is the first principal component for the original data points?

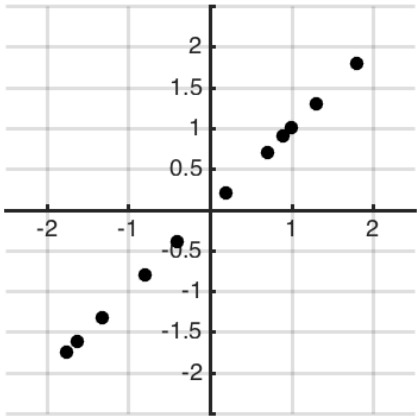
i. $[0,0,1]$
ii. $[0,0,2]$
iii. $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0]$
iv. $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$

(b) What is the second principal component for the original data points?

i. $[0,0,1]$
ii. $[0,0,2]$
iii. $[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0]$
iv. $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0]$

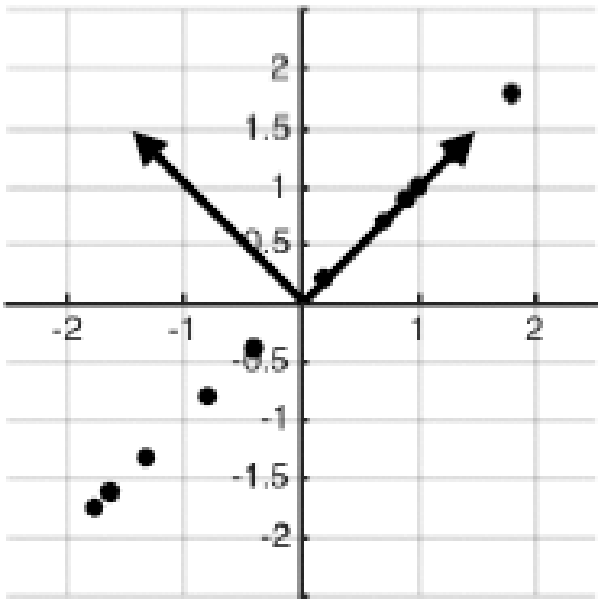
57. (CS189 final 2016 final, ex.Q3.(c-d))

(a) Draw the principal components of this data set on the plot.



(b) What is the ratio of the smallest eigenvalue to the largest eigenvalue of the covariance matrix of this data?

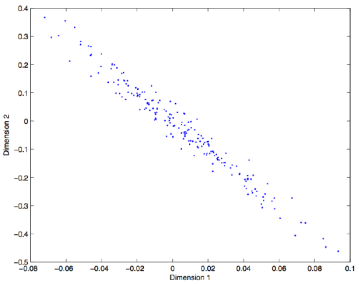
Solution:

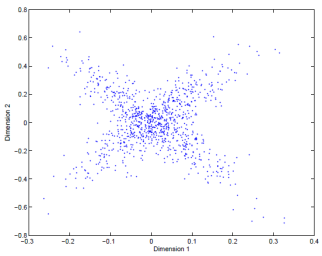


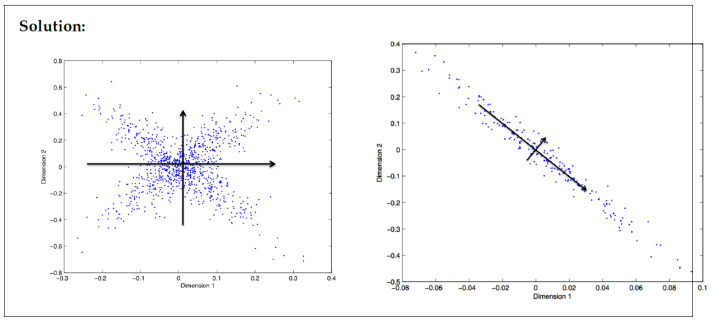
(a)

(b) 0

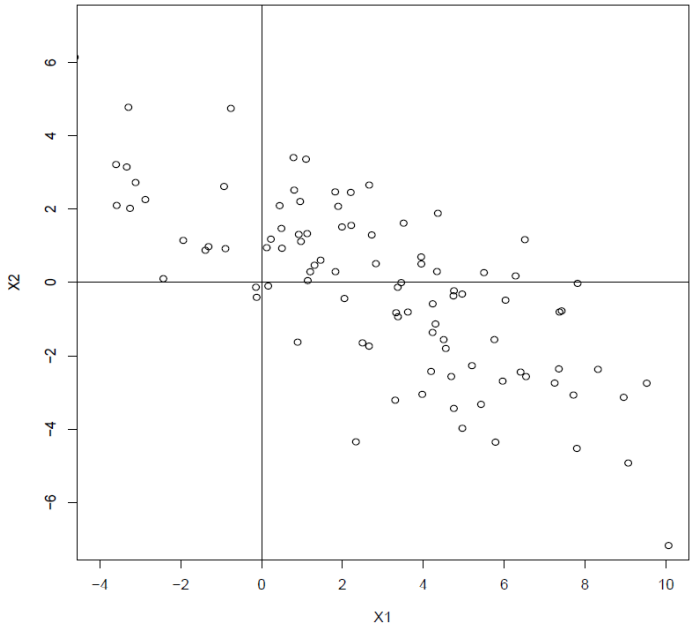
58. (CMU, 2012f, Tom Mitchell, Ziv Bar Joseph, final exam, ex.Q1.e) Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principal components on each plot.







59. (Radford, 2008f, midterm, ex.4) Here is a scatter plot of a sample of $n = 100$ values for two variables:



Judge by eye what the eigenvectors and eigenvalues of the sample covariance matrix for these variables is, choosing one of the following:

(a) $e_1 = [0.65, 0.75]^\top$, $\lambda_1 = 41.2$, $e_2 = [0.75, -0.65]^\top$, $\lambda_2 = 0.9$

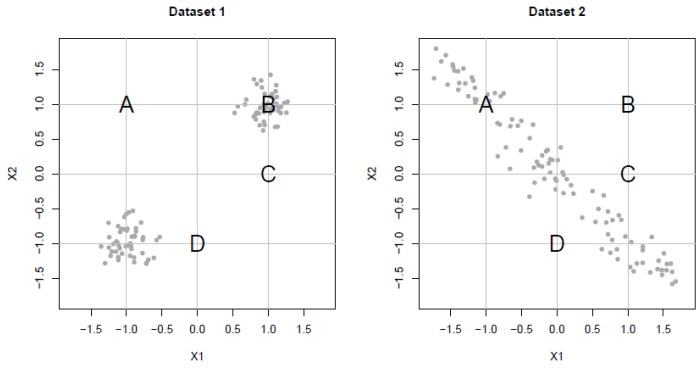
(b) $e_1 = [0.65, 0.75]^\top$, $\lambda_1 = 17.2$, $e_2 = [0.75, -0.65]^\top$, $\lambda_2 = 1.9$

(c) $e_1 = [0.87, -0.50]^\top$, $\lambda_1 = 3.7$, $e_2 = [0.50, 0.87]^\top$, $\lambda_2 = 0.9$

(d) $e_1 = [0.87, -0.50]^\top$, $\lambda_1 = 18.3$, $e_2 = [0.50, 0.87]^\top$, $\lambda_2 = 1.9$

60. (Radford, Spring 2006, pr.2)

Below are scatterplots of the input variables, X_1 and X_2, in two datasets with 100 cases. For both datasets, the variables have been standardized to have mean zero and standard deviation one. (Note that the response variable is not shown here.)



Rather than use these two inputs for a classification or regression method, we would like to reduce them to only one variable by projecting the points onto the first principal component direction. If $x = (x_1, x_2)$ is a vector representing the inputs for some case, and v is a vector of length one pointing in the direction of the first principal component, the projection on the first principal component will be $x^\top v$. Note that $-v$ would also be a valid indicator of the principal component direction; you may use either, but you must be consistent. Draw the vectors you use on the plots above.

The questions below ask you to find (approximately) the projections on the first principle component of the four point marked A, B, C, and D. You should determine the principle component directions by eye, not by trying to work numerically through some formula. (Note that you should figure out the answers for dataset 1 and for dataset 2 separately - these datasets have nothing to do with each other.)

(a) For dataset 1, and for each point marked in the plot, circle the number below that is approximately the projection of that point on the first principal component direction:

Point A: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

Point B: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

Point C: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

Point D: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

(b) For dataset 2, and for each point marked in the plot, circle the number below that is approximately the projection of that point on the first principal component direction:

Point A: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

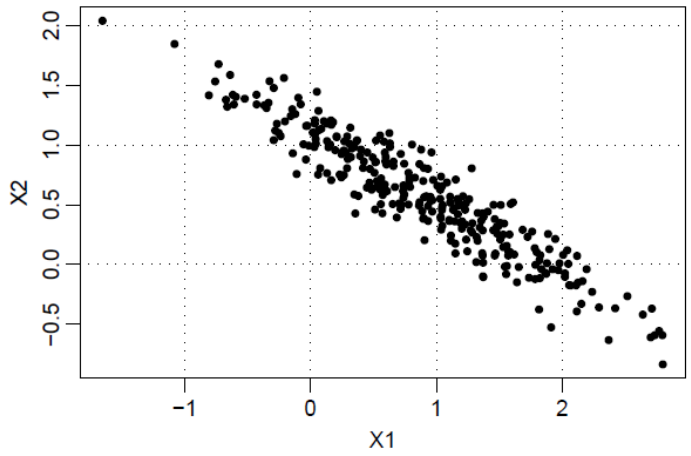
Point B: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

Point C: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

Point D: -2.0 -1.4 -1.2 -1.0 -0.7 -0.3 0.0 0.3 0.7 1.0 1.2 1.4 2.0

61. (Radford, 2010f, midterm, pr.5)

Here is the scatterplot of 300 observations on two variables:



The sample mean vector for this data is $[0.880.57]^\top$.

(a) Which of the following is the sample covariance matrix for this data?

i. $\begin{bmatrix} 0.58 & 0.45 \\ 0.45 & 0.24 \end{bmatrix}$

ii. $\begin{bmatrix} 4.89 & -1.37 \\ -1.37 & 2.20 \end{bmatrix}$

iii. $\begin{bmatrix} 0.63 & -0.37 \\ -0.37 & 0.25 \end{bmatrix}$

iv. $\begin{bmatrix} 0.68 & -0.32 \\ -0.32 & 0.71 \end{bmatrix}$

(b) Which of the following is a vector pointing in the direction of the first principal component?

i. $\begin{bmatrix}0.55\\0.84\end{bmatrix}$

ii. $\begin{bmatrix}0.85\\-0.52\end{bmatrix}$

iii. $\begin{bmatrix}0.52\\-0.85\end{bmatrix}$

iv. $\begin{bmatrix}-0.55\\-0.84\end{bmatrix}$

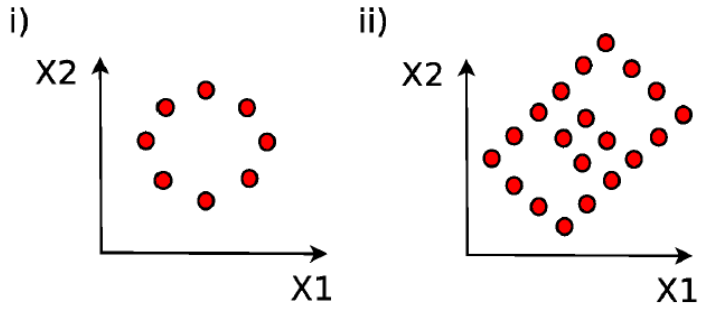
v. $\begin{bmatrix}-0.80\\-0.60\end{bmatrix}$

(c) What is the projection of the data point $[-0.12, 1.07]^\top$ on the first principal component?

(d) Write down a vector that points in the direction of the second principal component.

62. (CMU, 2012s, ZBarJoseph, final, pr.7.1)

Which of the following unit vectors expressed in coordinates $(X_1; X_2)$ correspond to theoretically correct directions of the 1st (p) and 2nd (q) principal components (via linear PCA) respectively for the data shown in the following figure? Choose all correct options if there are multiple ones.



(a) i) p(1,0), q(0,1); ii) p($1/\sqrt{2}$,$1/\sqrt{2}$), q($1/\sqrt{2}$,-$1/\sqrt{2}$)

(b) i) p(1,0), q(0,1); ii) p($1/\sqrt{2}$,-$1/\sqrt{2}$), q($1/\sqrt{2}$,$1/\sqrt{2}$)

(c) i) p(0,1), q(1,0); ii) p($1/\sqrt{2}$,$1/\sqrt{2}$), q(-$1/\sqrt{2}$,$1/\sqrt{2}$)

(d) i) p(0,1), q(1,0); ii) p($1/\sqrt{2}$,$1/\sqrt{2}$), q(-$1/\sqrt{2}$,-$1/\sqrt{2}$)

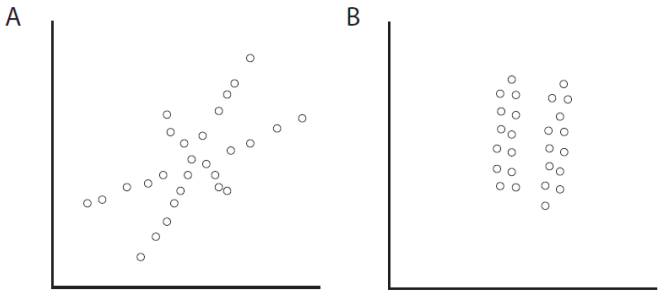
(e) All of the above are correct.

Solution:

first and third answers. Three points are tested: PCs are ordered according to variability; PCs are orthogonal; PC axes are potentially unidentifiable.

63. (CMU, 2016f, NBalcan, MGormley, HW5, pr.3)

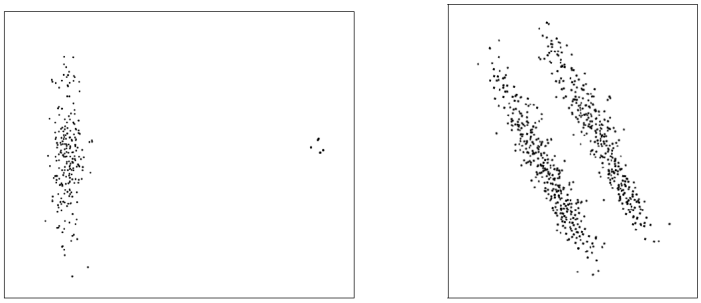
    (a) For each of the two data sets shown below, draw vectors that correspond to the first and second principal components that would be produced by Principal Components Analysis (PCA). For each diagram, label which vector is the first component, and which is the second.



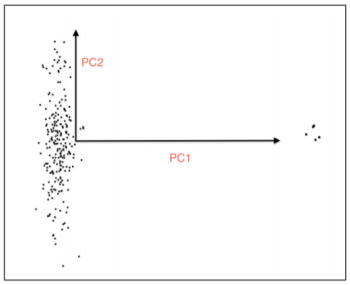
    (b) Is it possible for you to draw a third principal component for the first of these datasets (in A)? Please do, or explain in one sentence why it is impossible.

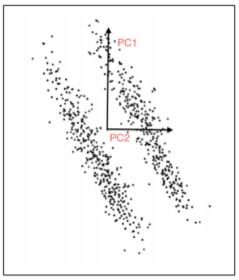
    (c) PCA is typically performed before clustering high-dimensional data. For this procedure, what assumption do we make about the data's covariance structure in relation to the distances between clusters? Explain why this assumption holds or does not hold for the second dataset (in B). Be brief (2-3 sentences).

64. (10-601 Machine Learning, Fall 2011: Homework 5 Machine Learning Department Carnegie Mellon University, ex.3.2) Draw the first two principal components for the following two datasets. Based on this, comment on a limitation of PCA.



Solution:





One of the limitations of PCA is that is can be highly affected by just a few outliers. Another limitation is that it may fail to capture meaningful components that do not pass through the origin.

65. (CMU, 2012s, ZBarJoseph, final, pr.7.2)

In linear PCA, the covariance matrix of the data $C = C^\top X$ is decomposed into weighted sums of its eigenvalues (λ) and eigenvectors p:

$$C = \sum_i \lambda_i p_i p_i^\top$$

Prove mathematically that the first eigenvalue λ_1 is identical to the variance obtained by projecting data into the first principal component p_1 (hint: PCA maximizes variance by projecting data onto its principal components).

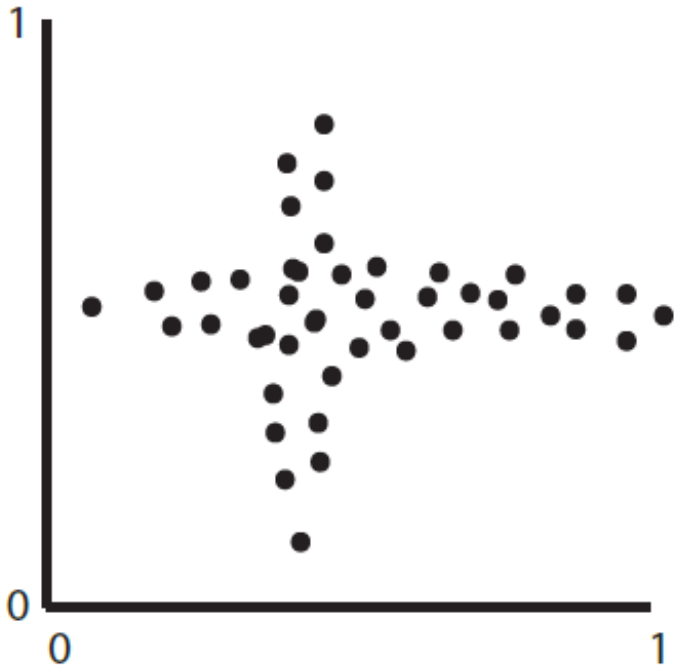
Solution:

the variance in the first PC is $v = p_1^\top C p_1$. Since λ_1 is the eigenvector, $\lambda_1 p_1 = C p_1 \Rightarrow \lambda_1 p_1^\top p_1 = p_1^\top C p_1 \Rightarrow \lambda_1 \cdot 1 = v \Rightarrow p_1^\top p_1 = 1$ (Q.E.D.).

66. (after CMU, 2017f, NBalcan, class test 3, pr. 2) **PCA vs uncentered PCA** Consider the following plot of data.

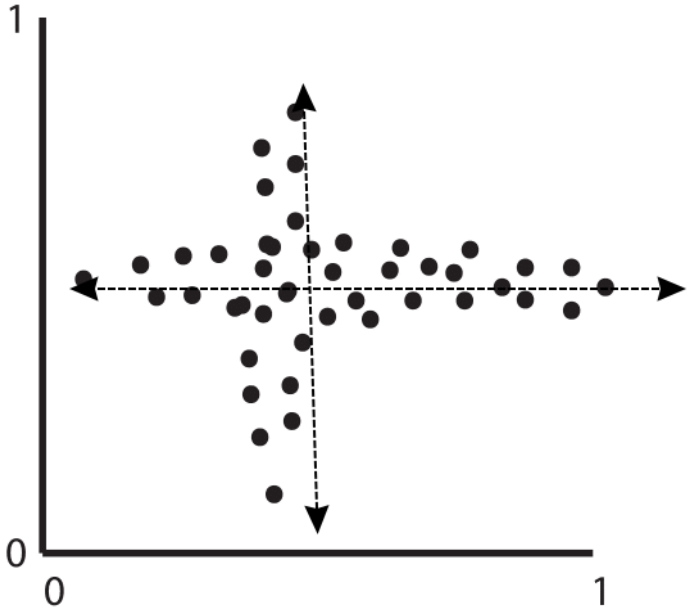
(a) Draw arrows from the mean of the data to denote the direction and relative magnitudes of the principal components. (PCA = best **affine** subspace)

(b) Draw arrows from the origin to denote the direction and relative magnitudes of the principal components when data was not centered. (uncentered PCA = best **linear** subspace)



Solution:

(a)

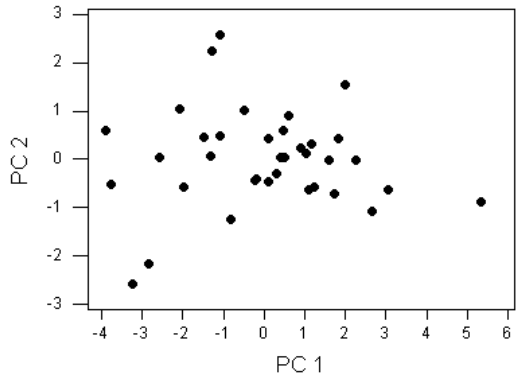


67. (Penn State Department of Statistics , Statistics 460, 2000s, final exam, ex. 2, exams/New Folder (5)) (PCA with given output) In a sample of 36 flea beetles, 6 different physical characteristics are measured for each beetle: head length, body length, length os one joint, length os a second joint, total weight, and body temperature. The first 4 are measured in micrometers, the 5th is measured in milligrams, and the 6th is measured in degrees Celsius. The outcome of a principal components analysis and a related plot are given below.

```
Eigenanalysis of the Correlation Matrix

Eigenvalue    4.0424    1.0419    0.8711    0.0372    0.0064    0.0010
Proportion     0.674     0.174     0.145     0.006     0.001     0.000
Cumulative     0.674     0.847     0.993     0.999     1.000     1.000

Variable         PC1       PC2       PC3
head           0.489    -0.001    -0.077
body           0.495     0.065     0.028
jnt1           0.437    -0.084     0.393
jnt2           0.372     0.083    -0.723
weight         0.496    -0.057     0.023
temp           0.035     0.824     0.562
```



(a) As you can see, the **correlation** matrix is used here instead of the covariance matrix. Explain why the correlation matrix is a more sensible choice than the covariance matrix for this analysis.

(b) How much of the variation in this dataset is explained by the first principal component? How much is explained by the first and the second principal components together?

(c) A plot of the principal component scores for the 36 beetles is shown. Imagine that the x and y axes are drawn onto the plot, dividing it into 4 quadrants: UR (upper right), UL (upper left), LR (lower right), and LL (lower left).

Suppose two new beetles, Bert and Ernie, are measured. Bert is much larger than any of the other beetles and is also slightly warmer. Ernie, on the other hand, is very small compared with the other beetles but like Bert, Ernie is warmer than the others.

For both Bert and Ernie, tell which quadrant o fthe above graph each would lie in. Explain briefly how you know.

Solution:

(a) The variables are on completely different scales, so it makes sense to standardize them all before analyzing their contributions in the variation. This is what the correlation matrix does, not the covariance matrix.

(b) The first PC explains 67.4%. The first two together explain 84.7%.

(c) Bert: UR

Ernie: UL

From the loadings, we can tell that the first PC is essentially nothing but the sum of the size measurements and the second PC is essentially nothing but the temperature. Thus, Bert will have a high PC 1 score and a high PC 2 score, whereas Ernie will have a low PC 1 score and high PC 2 score.

3.2 PCA and SVD

68. (CMU, 2014s, BPoczos, ASingh, midterm, pr. 1.19) **PCA and SVD**

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 .

Solution:

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

3.3 PCA and Least Squares

69. (Rob Tibshirani, Sta306b midterm test May 20, 2015, ex.1 - exams/New folder) (**PCA and linear regression**) In linear regression, principal components is sometimes used as an initial step to reduce the dimensionality of a set of features. The idea is to regress the outcome on the first few principal components of the data matrix. Comments on the pros and cons of this idea.

    Solution:

    The direction of variation of the outcome may or may not be in the direction of largest variation in X. So regression on the first few principal components can sometimes be effective, and sometimes is not. However one can argue that features are often chosen for an experiment because they are likely to be relevant for predicting y, and hence their direction of variation will often be relevant for predicting y.

70. (CMU, 2007s, final, pr.1.7) Both **PCA and linear regression** can be thought of as algorithms for minimizing a sum of squared errors. Explain which error is being minimized in each algorithm.

    Solution:

    PCA: $\arg\min_u (x - \sum_{i=1}^{k} (x \cdot u_i) u_i)^2$ - "reconstruction error"

    Linear regression: $\arg\min_\beta (y - x\beta)^2$ - "residual" error

71. (CS168, Spring 2016, Final Exam, ex.4) **(TLS, PCA and least squares)** For the first two questions, choose an answer from the following choices:

    - PCA
    - Least squares (i.e., linear regression)
    - Both
    - None of the above

    (a) Finds the best fit line that minimizes the average squared Euclidian distance between the line and the data points.

    (b) Finds the best fit line that minimizes the sum of squares of residuals $\sum_i (y_i - \hat{y}_i)^2$.

    (c) Given a set $(x_1, y_1), \ldots, (x_n, y_n)$ of two-dimensional points, suppose we run linear regression twice, first regressing x onto y and then regressing y onto x. Will we get the same best-fit line in both cases?

(d) Consider the set of points depicted in the figure below with with the horizontal axis corresponding to the x coordinate of each point, and the vertical axis corresponding to the y-coordinate. Draw two lines, one depicting the first principal component of the point set, and the second depicting the least squares regression line corresponding to fitting y as a linear function of x. Make sure you clearly label which line is which.

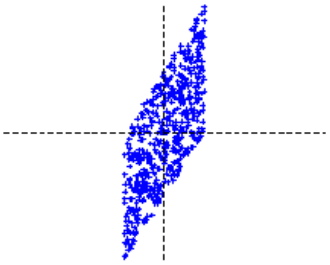


Figure 2: Draw two lines, and label one "PCA" and one "least squares".

3.4 PCA and Whitening

72. (Independent Component Analysis book, Aapo Hyvarinen, Chapter 6, Problems, ex.6.5 - d bookfinal_ICA.pdf: ICA, pca, fa + some th. ex...) The covariance matrix of vector x is

$$C_x = \begin{bmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{bmatrix}$$

Compute a **whitening** transformation for x.

3.5 Dual PCA and Kernel PCA

73. (CS 189 Spring 2014 Introduction to Machine Learning Final, ex.Q7) **Kernel PCA**

You are given d-dimensional real-valued data$\{x_i\}_{i=1}^N$ and a feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^m$. In the following questions, you will investigate how to do PCA in feature space on the feature vectors $\{\phi(x_i)\}_{i=1}^N$. Assume that the data is centered in feature space; that is, $\sum_{i=1}^N \phi(x_i) = 0$.

In the following, Ψ is a design matrix whose i$^{\text{th}}$ row is $\psi(x_i)$.

(a) Recall that as a part of PCA, we must solve the eigenvalue problem $Sv = \lambda v$, where S proportional to the sample covariance matrix. For PCA in feature space, we have $S = \sum_{i=1}^{N} \phi(x_i)\phi(x_i)^\top = \Phi^\top \Phi$. Why is this a problem if m is large?

(b) Now, you are given a kernel function $k(x, x') = \psi(x) \cdot \psi(x')$. Define the kernel matrix $K_{ij} = k(x_i, x_j)$. Show that if $\lambda \neq 0$, then λ is an eigenvalue of S if and only if λ is also an eigenvalue of K (in other words, finding feature-space principal components can be done by finding eigenvectors of K).

(c) Let v be an eigenvector of S with nonzero eigenvalue λ. Show that v can be written as $v = \Phi^\top \alpha_v$, where α_v is an eigenvector of K with eigenvalue λ.

(d) You are given a new data point $x \in \mathbb{R}^d$. Find the scalar projection of its feature representation $\phi(x)$ onto $v/\|v\|$ (with v defined as above). Write your answer in terms of α_v and λ. Use the kernel k, and do not explicitly use ϕ. You should use the notation $k_x = [k(x_1, x) \dots k(x_n, x)]^\top$.

Solution:

(a) Working in feature space directly is too expensive if m is large. The covariance matrix $\Phi^\top \Phi$ is $m \times m$, which is too large to compute and work with.

(b) Notice that $K = \Phi\Phi^\top$. The nonzero eigenvalues of $K = \Psi\Phi^\top$ and $S = \Phi^\top\Phi$ are the same.

(c) First, write $Sv = \Phi^\top \Phi v = \lambda v$. Multiplying this equation by Φ gives $K\alpha = \lambda\alpha$, where $\alpha = \Phi v$. Then, since $\Phi^\top \alpha = \Phi^\top \Phi v = Sv = \lambda v$, we have $v = \Phi^\top \alpha/\lambda$. Choosing $\alpha_v = \alpha/\lambda$ gives the desired result.

(d) First, let's calculate the squared norm of v:

$$\|v\|^2 = v^\top v = \alpha_v^\top \Phi\Phi^\top \alpha_v = \alpha_v^\top K \alpha_v = \lambda \|\alpha_v\|^2$$

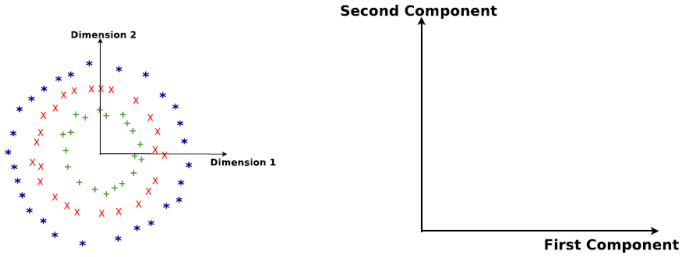
Then, the projection is:

$$\frac{v^\top \phi(x)}{\|v\|} = \frac{\alpha_v^\top \Phi \phi(x)}{\sqrt{\lambda}\|\alpha_v\|} = \frac{\alpha_v^\top k_x}{\sqrt{\lambda}\|\alpha_v\|}$$

74. (CMU, 2017f, NBalcan, mideterm, pr. 3.5) The left plot in the following figure shows toy data in $\mathbb{R}^2$. Suppose we perform **kernel PCA** on this data using the following kernel

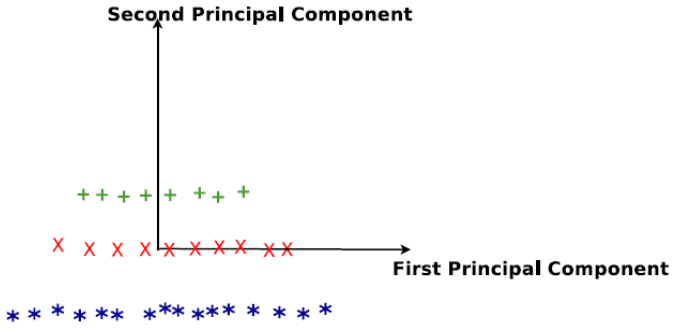
$$K(x_1, x_2) = (1 + < x_1, x_2 >)^2$$

In the right plot show how the data looks after projection onto the first two principal components. Recall that in kernel PCA we map each point x_i to a high dimensional space using the feature map, ϕ, associated with the kernel K. We then perform PCA in the ϕ space. Use a computer!



Solution:

The following figure shows the data after projection onto the first two principal components. Notice that the data becomes linearly separable in the kernel space.



3.6 Revision

75. (CS168 Final Exam, Spring 2017, ex.5) **(PCA vs. JL)**

    (a) Briefly explain how the PCA and Johnson-Lindenstrauss (JL) approaches to dimensionality reduction work.

(b) For each of the following parts, specify whether PCA or JL would be the more appropriate dimensionality reduction method, and give a 1–2 sentence explanation for your answer.

i. You want to visualize your data.

ii. You suspect that your data has low-dimensional linear structure.

iii. The method of data collection that you were using introduced a large amount of noise along some specific direction or low-dimensional subspace.

iv. You want to use a nearest neighbor subroutine to explore your data and classify new data.

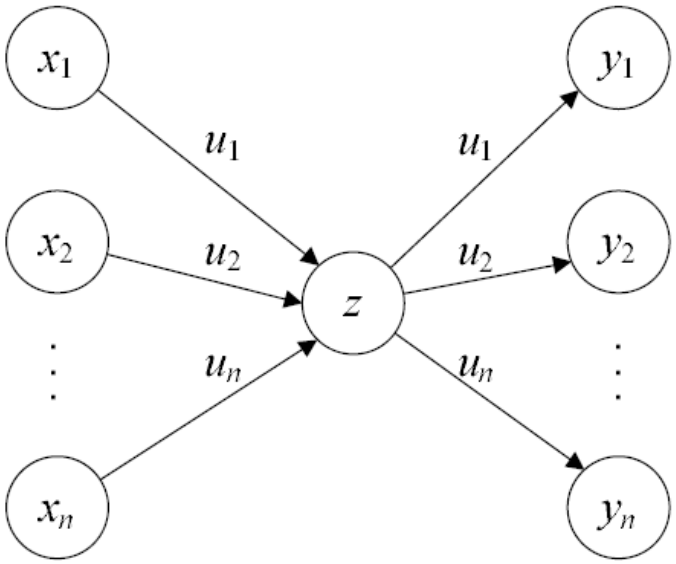
76. (CMU, 2006s, final, pr. 10)

PCA and neural networks

It turns out that neural networks can be used to perform dimensionality reduction. We are given a dataset $x_1, \ldots, x_m$, where each point is a n-dimensional vector: $x^i = (x_1^i, ..., x_n^i)$. Suppose that the dataset is centered:

$$\bar{x} = \sum_{i=1}^{m} x^i = 0$$

Consider the following network with 1 node in the hidden layer:



In this network, the hidden and the output layer share the weight parameters, i.e., the edge $x_j \to z$ uses the same weight u_j as the edge $z \to y_j$. The nodes have linear response functions:

$$z = \sum_{j=1}^{n} u_j x_j, y_j = u_j z$$

Typically, in neural networks, we have training examples of the form (x^i, t^i). Here, the network is trained as follows: for each point x^i of our dataset, we construct a training example (x^i, x^i) that assigns the point to both the inputs and the outputs, i.e., $t^i = x^i$.

(a) Suppose that we minimize a square loss function,

$$l(w; x) = \sum_i \sum_j (t_j^i - \mathrm{j}(x^i : w))^2$$

where out(x;w) is the prediction y of the network on input x, given weights w. Write down the loss function in terms of the network parameters u and the way we are training this neural network.

(b) Recall that, in principal component analysis with m components, we approximate a data point x_i, by projecting it onto a set of basis vectors $(u_1, \ldots, u_k)$:

$$\hat{x}^i = \hat{x} + \sum_{j=1}^{k} z_j^i u_j$$

where $z_j^i = x^i \cdot u_j$, in order to minimize the squared reconstruction error:

$$\sum_{i=1}^{m} \|x^i - \hat{x}^i\|_2^2$$

where $\|v\|_2^2 = \sum_{j=1}^{n} (v_j)^2$. Relate the optimization problem from part 1 to the problem optimized in PCA.

(c) Construct a neural network that will result in the same reconstruction of data as PCA with k first principal components. Write down both the structure of the network and the node response functions.

Solution:

training this neural network.

$$\ell(U;X) = \sum_i \sum_j \left(x_j^i - U_j \cdot \sum_{\ell=1}^{n} U_\ell \, x_\ell^i \right)^2$$

(a)

$$\sum_{i=1}^{m} ||\mathbf{x}^i - \hat{\mathbf{x}}^i||_2^2,$$

where $||v||_2^2 = \sum_{j=1}^{n} (v_j)^2$. Relate the optimization problem from part 1 to the problem optimized in PCA.

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (x_j^i - \hat{x}_j^i)^2 = \sum_i \sum_j \left(x_j^i - \bar{x} - \left[\sum_{d=1}^{k} (x^i \cdot u_d) u_d \right]_j \right)^2$$

$$= \sum_i \sum_j \left(x_j^i - \bar{x} - \sum_{d=1}^{k} u_{dj} \left(\sum_{\ell=1}^{n} x_\ell^i u_{d\ell} \right) \right)^2$$

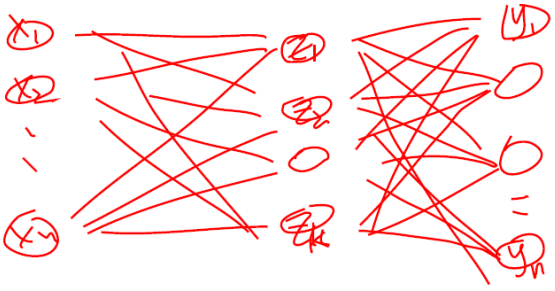
$x_j^i - \bar{x}$: t_j^i above

$\sum_{d=1}^{k} u_{dj} \left(\sum_{\ell=1}^{n} x_\ell^i u_{d\ell} \right)$: $\text{out}_j(x^i, u_d)$ of them

(b)

network and the node response functions.

Same as before but with $z_1, \ldots, z_k$



train ex: $(x^i, x^i - \bar{x})$

$$z_i = \sum_{j=1}^{n} u_{ij} x_j \quad , \quad y_j = \sum_i u_{ij} z_i$$

(c)

77. Circle the correct answer and justify your choice:

(a) (CS189 practice midterm spring 2018, ex.2.(a)) What is the primary purpose of PCA?

i. Dimension reduction

ii. Linear regression

iii. Outlier removal

iv. Optimization

(b) (Final CS 189 Spring 2015 Introduction to Machine Learning, ex.Q1.(e)) Given a design matrix $X \in \mathbb{R}^{n \times d}$, where $d << n$, if we project our data onto a k dimensional subspace using PCA where k equals the rank ok X, we recreate a perfect representation of our data with no loss.

i. True

ii. False

Solution: A

(c) (CS 189 Spring 2016 Introduction to Machine Learning Final, ex.Q1.(21)) Both PCA and Lasso can be used for feature selection. Which of the following statements are true?

i. Lasso selects a subset (not necessarily a strict subset) of the original features

ii. PCA and Lasso both allow you to specify how many features are chosen

iii. PCA produces features that are linear combinations of the original features

iv. PCA and Lasso are the same if you use the kernel trick

Solution: A,C

(d) (CS 189 Spring 2017 Introduction to Machine Learning Final, ex.Q1.(5)) Why is PCA sometimes used as a preprocessing step before regression?

i. To reduce overfitting by removing poorly predictive dimensions.

ii. To expose information missing from the input data.

iii. To make computation faster by reducing the dimensionality of the data.

iv. For inference and scientific discovery, we prefer features that are not axis-aligned.

Solution: A,C

(e) (CS189 Spring 2018 midterm, ex.8.(e)) Assume you have n input data points, each with d high quality features ($X \in \mathbb{R}^{n \times d}$) and associated labels ($y \in \mathbb{R}^n$). Suppose that $d >> n$ and you want to learn a linear predictor. Which of these approaches would help you to avoid overfitting?

i. Preprocess X using $k << n$ random projections.

ii. Preprocess X using PCA with $k << n$ components.

iii. Preprocess X using PCA with n components.

iv. Add polynomial features

v. Use a kernel approach

vi. Add a ridge penalty to OLS

vii. Do weighted least squares

Solution: The goal here is simple dimensionality reduction for overfitting avoidance. As the earlier problem in the exam showed, ridge regression can be viewed as a softer form of k-PCAOLS where the different dimensions are downweighted softly rather than as a strict truncation.

So both PCA and ridge are clearly valid approaches to reduce overfitting.

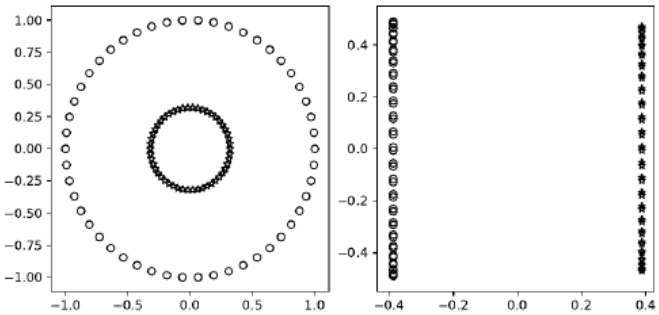
However, because the problem said that these are high-quality features, what you know about random projection also applies and so they too can be useful for dimensionality reduction.

The other answers are mostly wrong. If you use n components of PCA, there are still too many parameters relative to the data points. And so some overfitting will still occur.

Adding polynomial features makes the overfitting issue worse and not better, while weighing samples doesn't help us in any way.

We did not penalize for also marking kernel approaches since you know from lecture that the right kernel can also regularize because the kernel approach serves the same purpose as choosing features and priors together.

(f) (CS189 Spring 2018 midterm, ex.8.(f)) Which methods could yield a transformation to go from the two-dimensional data on the left to the two-dimensional data on the right?



i. Random projections
ii. PCA
iii. Use of a kernel
iv. Adding polynomial features

Solution:

The plot here was literally obtained by doing RBF kernel-PCA on the data and choosing the two dominant components. PCA is clearly being used, but also something that allows us to get nonlinear relationships. Either the right kernel or polynomial features would suffice since circles are involved.

Random projections would not help here since they are linear.

(g) (CMU, 2016s, Matt Gormley, Final Exam Review, Lecture 29, Sample Questions, ex.4.i - exams/lecture29-final: 1 ex T/F) The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

i. True
ii. False

(h) (CMU, 2016s, Matt Gormley, Final Exam Review, Lecture 29, Sample Questions, ex.4.ii - exams/lecture29-final: 1 ex T/F) The output pf PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

i. True
ii. False

(i) (CMU, 2016s, Matt Gormley, Final Exam Review, Lecture 29, Sample Questions, ex.4.iii - exams/lecture29-final: 1 ex T/F) Subsequent principal components are always orthogonal to each other.

i. True
ii. False

(j) (Radford, 2008f, pr.7.b) When principal component analysis is done, it makes no difference whether the eigenvectors of the covariance matrix or of the correlation matrix are found - the results are essentially the same.

i. True
ii. False

(k) (Radford, 2010f, midterm, pr.6b) Suppose we have data on the height in metres and the weight in kilograms of 100 people, for which the

sample correlation between height ans wight is 0.59. Will the sample correlation between height and weight change if we re-express heights in feet ans weights in pounds?

i. Yes

ii. No

(l) (Radford, 2010f, midterm, pr.6b) Suppose we have data on the lengths of arms, legs ans noses of 100 people. We intend to find the direction of the first principal component for this data, using the sample covariance matrix. Does it matter for this purpose whether we express all these lengths in inches, or instead express all these lengths in centimeters?

i. Yes

ii. No

78. (CMU, 2014s, SKim, HW3, pr.1.6)

For each of the tasks below, indicate which of the following methods would be best-suited?

(a) K-means clustering

(b) Hierarchical Clustering

(c) PCA

(d) Mixture Model

(e) Semi-supervised learning

(f) Co-training

1. You have been given a set of news articles from the New York Times and asked to use the contents in the documents to cluster the articles into topics. You are not told exactly how many topics are sufficient or what granularity of topics to use. For instance, the topic "Olympics" could be a subtopic of "Sports". You want to create as informative a clustering as possible given the limited specifications of your task.

2. You are interning at Facebook and they have told you that they want to use the social network to figure out how to target advertising campaigns. Specifically, they believe that certain social groups are more receptive to certain campaigns. It is your job to specify these groups of friends while realizing that each person has different social circles they belong to.

3. You're given one gene expression (microarray) data set that includes 1000 patients and you want to train a classifier to determine which patients

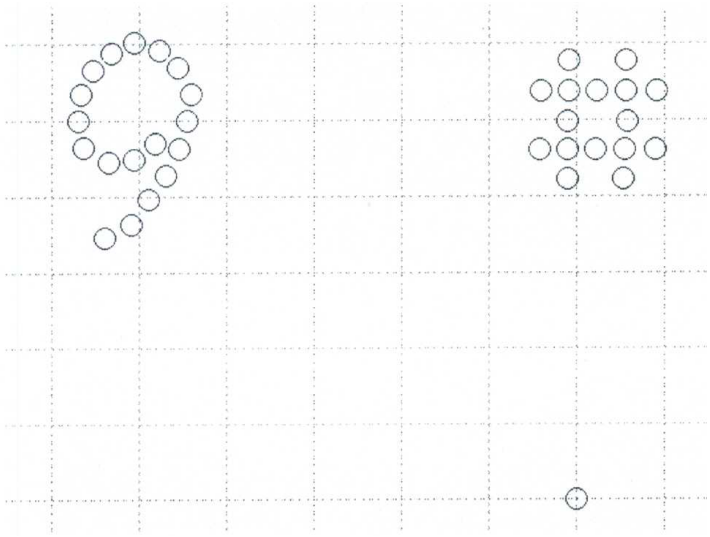
are "normal" and which patients have "cancer". The problem is the clinical group from whom you got the data isn't great at book keeping and seem to have lost half the labels.

4. You have been given an image data set and you want to classify the images into their object categories. The number of pixels is huge and the classifier is taking too long to run. You don't really care about recovering which pixels are most important to the classification task, you only care about correctly labeling images. What could make your life easier?

3.7 Spectral++

3.7.1 Spectral clustering - ML version

79. (CMU, 2006f, final, ex.8.B) We are in the setting of a 2-clustering problem. The data is given in the following grid. The grid unit is 1.



(a) If we use Euclidean distance to construct the affinity matrix W as follows:

$$W_{ij} = \begin{cases} 1 & \text{if } \|x_i - x_j\|_2^2 \leq \sigma^2 \\ 0 & \text{otherwise} \end{cases}$$

What σ^2 value would you choose? Briefly explain.

(b) The next step is to compute the $k = 2$ dominant eigenvectors of the affinity matrix W. For the value of σ^2 you chose in the previous

question, can you compute analytically eigenvalues corresponding to the first two eigenvectors? If yes, compute and report the eigenvalues. If not, briefly explain.

(c) Suppose the data is of very high dimension so that it is impossible to visualize them and pick a good value as we did at the first subpoint. Suggest a heuristic that could find an appropriate σ^2.

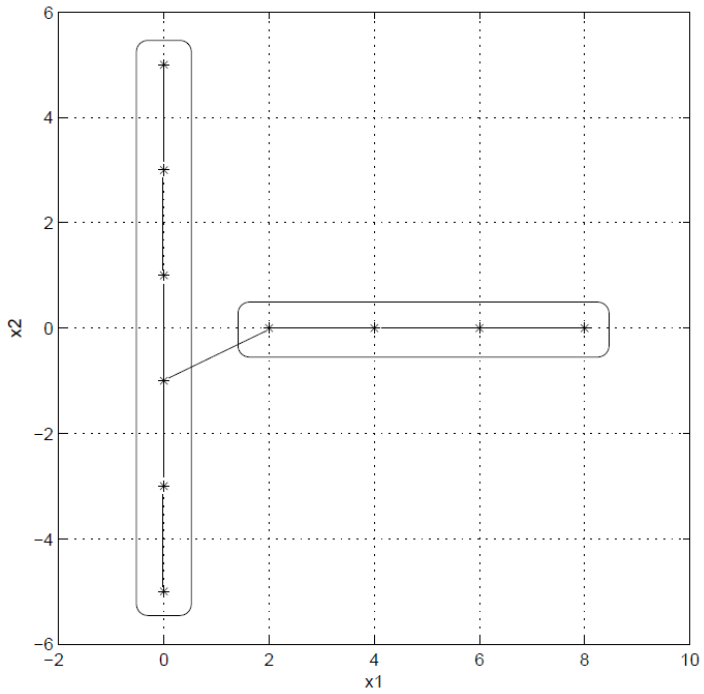
Solution:

(a) $\sigma^2 = 9 \sim 16$. W_{ij} should be 1 for every pair of points within "9", "#"; it should be 0 for other cases.

(b) $W = \begin{bmatrix} 1_{18\times 18} & 0 & 0 \\ 0 & 1_{16\times 16} & 0 \\ 0 & 0 & 1 \end{bmatrix}$

First two eigenvalues: 18, 16.

80. (MIT, 2004f, final exam, ex.3.2-3)



(a) Consider the data in the figure above. We will use spectral clustering to divide these points into two clusters. Our version of spectral clustering

uses a neighborhood graph obtained by connecting each point to its two nearest neighbors (breaking ties randomly), and by weighting the resulting edges between points x_i and x_j by $W_{ij} = \exp(-\|x_i - x_j\|)$. Indicate on the figure the clusters that we will obtain from spectral clustering. Provide a brief justification.

(b) Can the solution obtained in the previous part for the data in in the figure also be obtained by k-means clustering ($k = 2$)? Justify your answer.

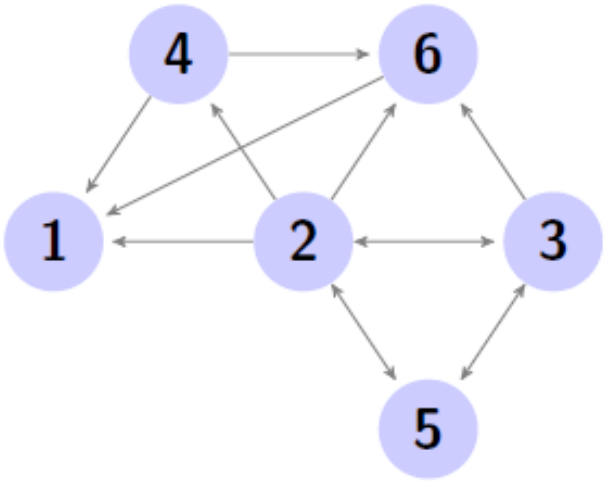
Solution:

(a) The random walk induced by the weights can switch between the clusters in the figure in only two places, (0,-1) and (2,0). Since the weights decay with distance, the weights corresponding to transitions within clusters are higher than those going across in both places. The random would would therefore tend to remain within the clusters indicated in the figure.

(b) No. In the k-means algorithm points are assigned to the closest mean (cluster centroid). The centroids of the left and right clusters in the figure are (0,0) and (5,0), respectively. Point (2,0), for example, is closer to the left cluster centroid (0,0) and wouldn't be assigned to the right cluster. The two clusters in the figure therefore cannot be fixed points of the k-means algorithm.

3.7.2 Ranking Webpages

81. (CS246: Mining Massive Data Sets Winter 2013 Final, ex.11) A directed graph G has the set of nodes $\{1, 2, 3, 4, 5, 6\}$ with the edges arranged as follows.



(a) Set up the PageRank equations, assuming $\beta = 0.8$ (jump probability $= 1 - \beta$). Denote the PageRank of node a by $r(a)$.

Solution:

$$r(1) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{2}r(4) + r(6) + \frac{1}{5}r(2)\right) + \frac{0.2}{6}$$

$$r(2) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{3}r(3) + \frac{1}{2}r(5)\right) + \frac{0.2}{6}$$

$$r(3) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{5}r(2) + \frac{1}{2}r(5)\right) + \frac{0.2}{6}$$

$$r(4) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{5}r(2)\right) + \frac{0.2}{6}$$

$$r(5) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{5}r(2) + \frac{1}{3}r(3)\right) + \frac{0.2}{6}$$

$$r(6) = 0.8\left(\frac{1}{6}r(1) + \frac{1}{5}r(2) + \frac{1}{3}r(3) + \frac{1}{2}r(4)\right) + \frac{0.2}{6}$$

(b) Order nodes by PageRank, from lowest to highest. (Note: No need to explicitly compute the scores. We are just asking for the ordering.)

Solution: In descending order: 1,6,2,3,5,4

82. (CMU, 15-826 Multimedia databases and data mining, Spring 2008, ex.Q7) Ranking on graphs PRACTICAL EXERCISE

Download the co-authorship network `http://www.cs.cmu.edu/~htong/15826-S08/hw3/pagerank.data` for the machine learning community. Each line of the dataset has three numbers: the index (author-id) for author i, the index for author j, and the count of paper they have been authored. You can also download the (author-id, name) list here `http://www.cs.cmu.edu/~htong/15826-S08/hw3/author_name.data`. You are asked to find the top 5 most influential authors by the PageRank algorithm (c=0.85, that is, fly-out probability = 1-c = 0.15, as Google suggests).

Hint: the PageRank vector v satisfies the equation

v = c A v + (1-c)/n 1

where v is an nx1 column vector, A is the nxn to-from column normalized transition matrix, and 1 (in bold) is an nx1 vector full of 'ones'.

What to Hand in:

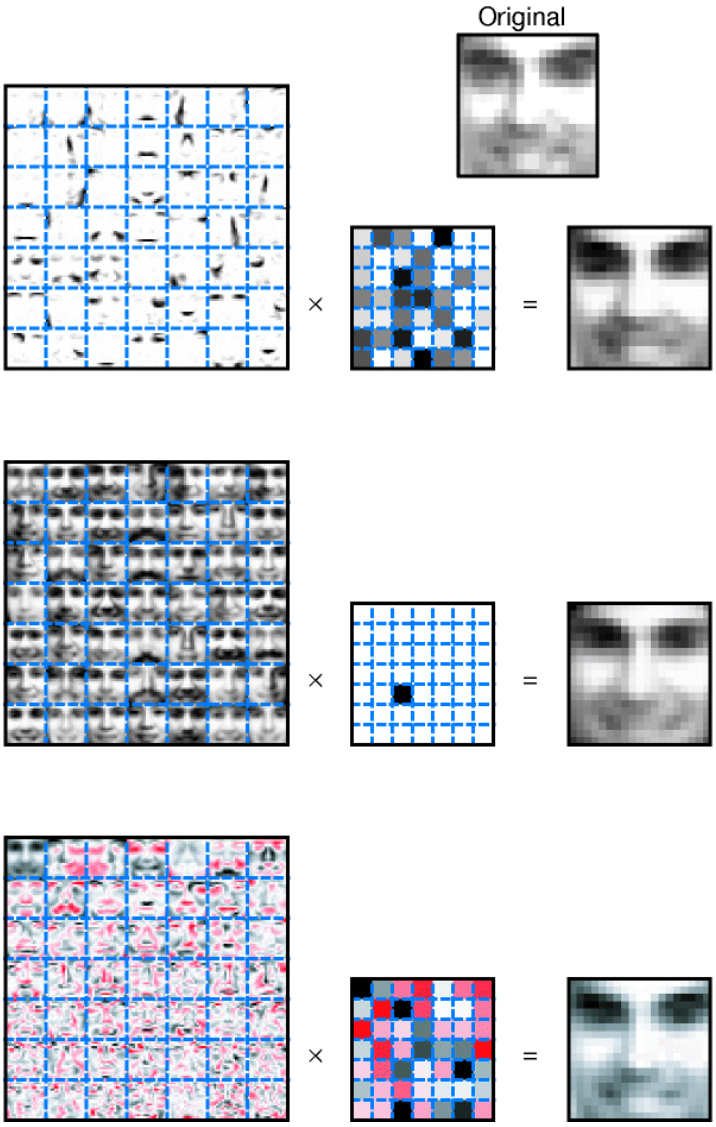
The names of top 5 authors in decreasing order, together with their ranking values.

Solution:

Author name	PageRank Score
Michael_I._Jordan	0.0036
Terrence_J._Sejnowski	0.0033
Bernhard_Schlkopf	0.0027
Satinder_P._Singh	0.0025
Andrew_W._Moore	0.0023

4 Non-negative Matrix Factorization - NMF

83. (CMU, 2006f, final, pr. 7.3; source of photo: `https://www.nature.com/articles/44565`) The goal of NMF is to reduce the dimensionality given non-negativity constraints. That is, we would like to find principal components $u_1, \ldots, u_r$, each of which is of dimension $d > r$, such that the d-dimensional data $x \approx \sum_{i=1}^{r} z_i u_i$, and all entries in x, z, $u_{1:r}$ are non-negative. NMF tends to find **sparse** (usually small L1 norm) basis vectors u_i's. Below is an example of applying PCA, NMF and k-means on a face image. Please point out the basis vectors in the equations and give them correct labels (PCA, NMF, k-means).



Solution:

The big squares = basis vectors

In order: NMF, k-means, PCA

5 LDA, GDA, QDA, FDA

5.1 Linear/Gaussian Discriminant Analysis - LDA=GDA

84. (Applied Multivariate Statistical Analysis by Johnson Wichern, ex.11.1) IS IT OK? Consider the two data sets

$$X_1 = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{bmatrix}$$

for which

$$\overline{x}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \overline{x}_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

and

$$S_{\text{pooled}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Calculate the linear discriminant function

$$\hat{y} = (\overline{x}_1 - \overline{x}_2)^\top S_{\text{pooled}}^{-1} x = \hat{a}^\top x.$$

(b) Classify the observations $x_0 = \begin{bmatrix} 2 \\ 7 \end{bmatrix}$ as population π_1 or population π_2, using the following rule with equal priors and equal costs:

$$\hat{m} = \frac{1}{2}(\overline{x}_1 - \overline{x}_2)^\top S_{\text{pooled}}^{-1}(\overline{x}_1 + \overline{x}_2) = \frac{1}{2}(\overline{y}_1 + \overline{y}_2)$$

$$\overline{y}_1 = \hat{a}^\top \overline{x}_1$$
$$\overline{y}_2 = \hat{a}^\top \overline{x}_2$$

85. Circle the correct answer and justify your choice:

(a) (CS 189 Spring 2019 Introduction to Machine Learning Midterm, ex. Q1.(h)) We are using linear discriminant analysis to classify points $x \in \mathbb{R}^d$ into three different classes. Let S be the set of points in $\mathbb{R}^d$ that our trained model classifies as belonging to the first class. Which of the following are true?

i. The decision boundary of S is always a hyperplane

ii. S can be the whole space $\mathbb{R}^d$

iii. The decision boundary of S is always a subset of a union of hyperplanes

iv. S is always connected (that is, every pair of points in S is connected by a path in S)

Solution: B,C,D

A: Given that we have three classes, S is defined by two linear inequalities, and therefore its boundary may not be a hyperplane.

C: Given that S is defined as the points satisfying a set of inequalities, its boundary is a subset of the hyperplanes defined by each of the linear inequalities.

B: If the prior for the first class is high enough, the probability of that class could be higher everywhere, and hence S would be the whole space. For example, take $\mu_1 = \mu_2 = \mu_3$ and $\pi_1 > \pi_2 = \pi_3$.

D: S is a convex polytope defined by the intersection of half-spaces (i.e. the points satisfying a set of linear inequalities). This is a convex set, and therefore it is connected.

(b) (Indian Institute of Technology Madras, Assignment 3, Introduction to Machine Learning, Prof. B. Ravindran, ex.7 - Assignment3.pdf)

With respect to Linear Discriminant Analysis, which of the following is/are true. (Consider a two class case)

i. When both the covariance matrices are spherical and equal, the decision boundary will be the perpendicular bisector of the line joining the means.

ii. When both the covariance matrices are spherical and equal, the decision boundary will be perpendicular to the line joining the means.

iii. When both the covariance matrices are spherical and equal and the priors $\pi_1 = \pi_2$ then the decision boundary will be perpendicular bisector of the line joining the means.

(c) (Indian Institute of Technology Madras, Assignment 3, Introduction to Machine Learning, Prof. B. Ravindran, ex.8 - Assignment3.pdf) For a two class classification problem, which among the following are true?

i. In case both the covariance matrices are spherical and equal, the within class variance term has an effect on the LDA derived direction.

ii. In case both the covariance matrices are spherical and equal, the within class variance term has no effect on the LDA derived direction.

iii. In case both the covariance matrices are spherical but unequal, the within class variance term has an effect on the LDA derived direction.

iv. In case both the covariance matrices are spherical but unequal, the within class variance term has no effect on the LDA derived direction.

5.2 Quadratic Discriminant Analysis - QDA

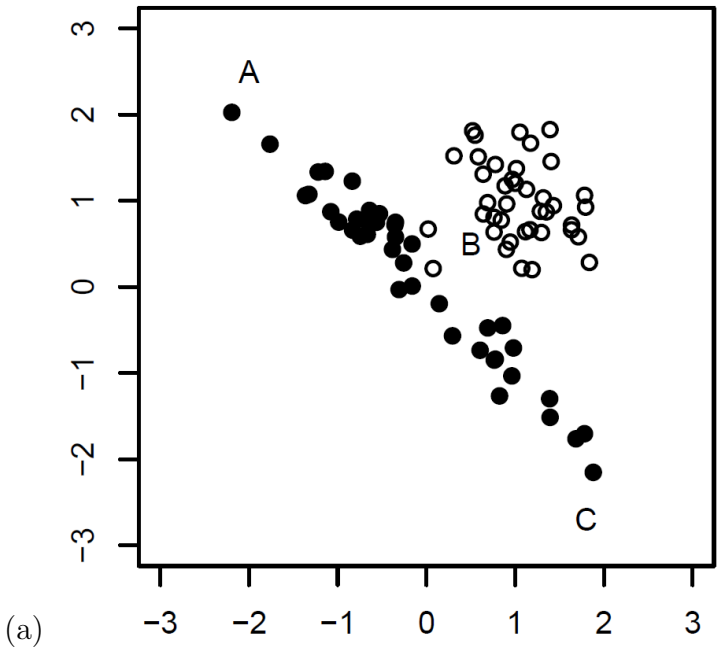
86. (Radford, 2007s, Second test, ex.1)

Recall that linear and quadratic discriminants can be derived from Gaussian models for the distribution of the inputs within each class. Using such models, we can find the probability of class k given values for the inputs, x, in a case by Bayes' Rule:

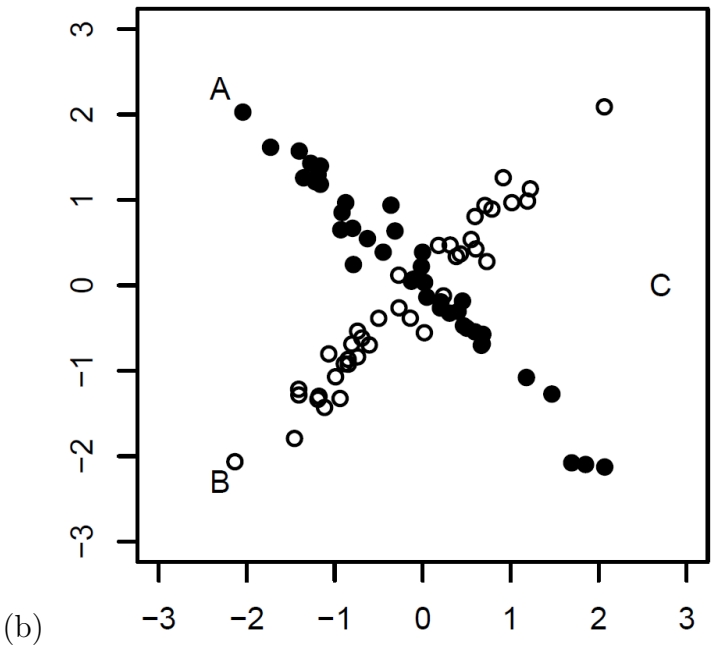
$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_j P(x|C_j)P(C_j)}$$

Linear discriminants, in which a case is classified according to the value of $w^\top x + w_0$, for some vector w and scalar w_0, arise when we assume that $P(x|C_k)$ is Gaussian, with the same covariance matrix for all classes k. Quadratic discriminants arise when we allowed different covariance matrices for different classes. In both cases, we find the means and covariance matrices by maximum likelihood.

Each of the two scatterplots below shows training cases for a classification problem with two inputs and two classes, with the class of the case indicated b the colour of dot (black is class 1). Three test cases with inputs x_A, x_B, and x_C, are also indicated, by the letters A, B, and C. For each scatterplot, write down the approximate probability of class 1 for each of the three test cases, if the probabilities are found assuming that the covariance matrix is the same for both classes (as with linear discriminants) and if the probabilities are found assuming that the covariance matrices for the two classes may be different (as with quadratic discriminants). You should give rough approximations for these probabilities - either "near 0", "near 1", or "near 1/2".



(a)



(b)

Solution:

(a) Linear discriminant:

$$P(C_1|x_A) = 1/2, P(C_1|x_D) = 1/2, P(C_1|x_C) = 1$$

With the covariance for the two classes constrained to be the same, the classification boundary will be a straight line going approximately through points A and B.

Quadratic discriminant:

$$P(C_1|x_A) = 1, P(C_1|x_B) = 1/2, P(C_1|x_C) = 1$$

(b) Linear discriminant:

$$P(C_1|x_A) = 1/2, P(C_1|x_B) = 1/2, P(C_1|x_C) = 1/2$$

Class means are nearly the same, covariances constrained to be the same, so class distributions are nearly identical, and all class probabilities are near $1/2$.

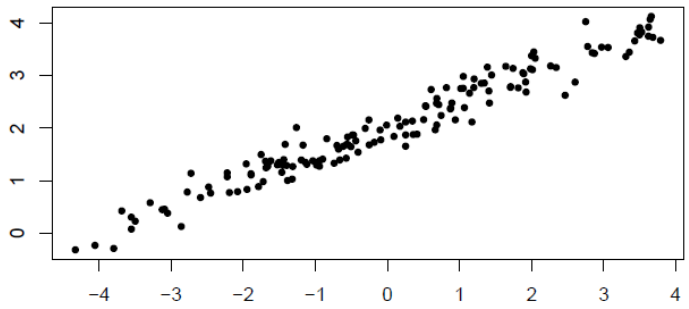
Quadratic discriminant:

$$P(C_1|x_A) = 1, P(C_1|x_B) = 0, P(C_1|x_C) = 1/2$$

5.3 Fisher Discriminant Analysis - FDA

5.3.1 PCA issue

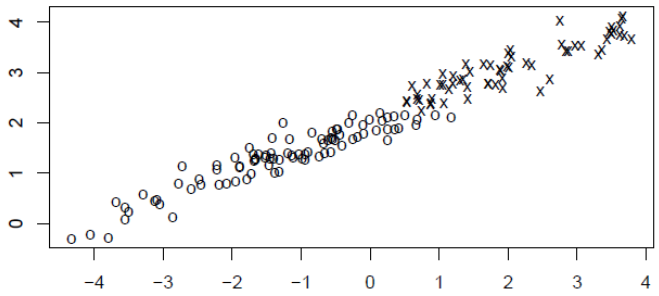
87. (Radford, 2014f, Practice Problem Set #3, ex.4) Below is a scatterplot of 150 observations of two variables:



(a) Write down a vector pointing in the direction of the first principal component for this data. An approximate answer found by eye is sufficient. The vector need not have length one. Also, draw the direction of the first principal component on the scatterplot above.

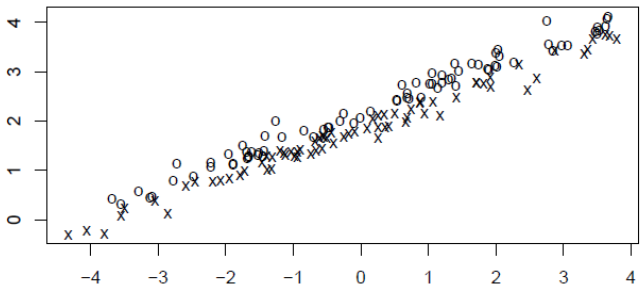
(b) What is the approximate standard deviation in the first principal component's direction?

(c) Suppose that each of these data points are associated with one of two classes, as shown below (with one class marked by "o" and the other by "x"):



If we reduce the data to just the projection on the first principal component, how well will we be able to classify the data points using this one number, compared to how well we would have benn able to classify using the two original numbers?

(d) Suppose instead that the two classes are as shown below:



In this case, how well will we be able to classify using just the projection on the first principal component, compared to using the two original numbers?

Solution:

(a) One answer is $[2, 1]$. I won't try to draw this on the plot.

(b) Somewhere around 2 or 3.

(c) We will be able to classify almost as well as with the original data.

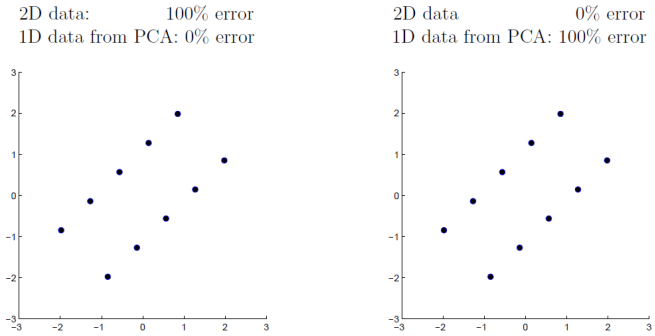
(d) The projection on the first principal component will give almost no information about the class. One could do much better using the

original data, since one can see in the plot that points in the circle class are usually above those in the x class. So there is a diagonal line that separates the classes fairly well.

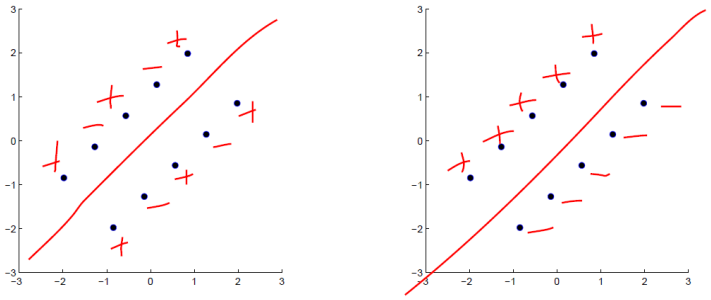
88. (CMU, 2006s, final, ex.9) Recall that PCA should be used with caution for classification problems, because it does not take information about classes into account. In this problem you will show that, depending on the dataset, the results may be very different. Suppose that the classification algorithm is 1-nearest-neighbor, the source data is 2- dimensional and PCA is used to reduce the dimensionality of data to 1 dimension. There are 2 classes (+ and -). The datapoints (without class labels) is pictured on plots below (the two plots are identical).

    (a) On one of the plots draw a line that PCA will project the datapoints to.

    (b) For each of the plots, label the source datapoints so that 1-NN will have the following leave-one-out cross-validation error:

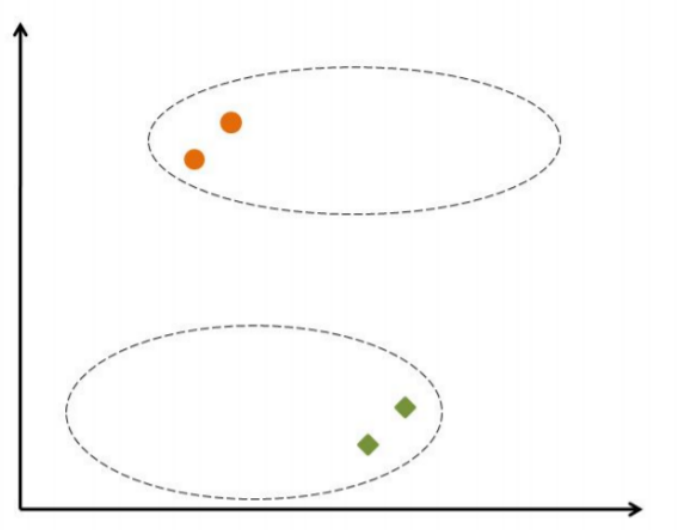


Solution:



5.3.2 PCA and FDA

89. (Balaraman Ravindran, Assignment 3, Introduction to Machine Learning, ex.9 - d Assignment3.pdf, pr.9) Suppose you are given a two dimensional two class data set as shown below with only two samples for each class. The dashed curves show the underlying (but unknown) distribution of each class. Which among the two methods for identifying a one dimensional representation of the given data would you suggest for building a classifier that will perform well on test data coming from the same underlying distributions

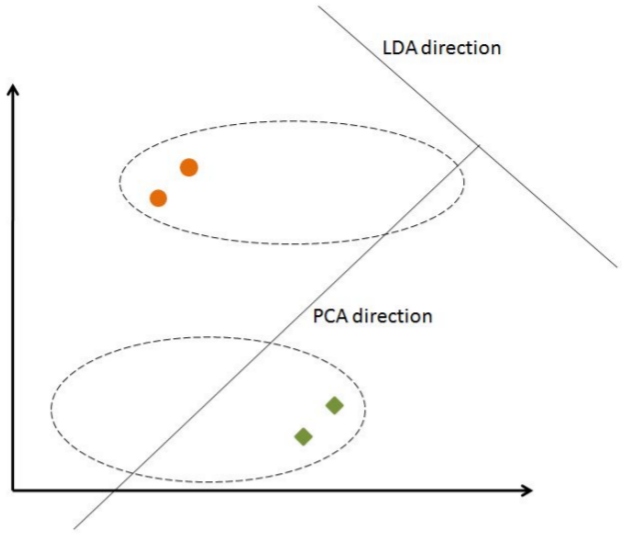


(a) Linear Discriminant Analysis (LDA)

(b) Principal Components Analysis (PCA)

Solution: (b)

As shown in the following figure, the single dimension identified by PCA is superior to the one identified by LDA for the purpose of classification if we take into consideration the underlying class distributions, since the overlap among the points of the two classes would be minimal in the case of the PCA dimension.



90. (Midterm exam CS 189/289, Fall 2015, True-False, ex.17) The Linear Discriminant Analysis (LDA) classifier computes the direction maximizing the ratio of between-class variance over within-class variance.

    (a) True

    (b) False

    Solution: True

6 Factor Analysis - FA

6.1 FA

91. (Radford, 2009f, final exam, ex.6) Consider a factor analysis model for six variables, $X = [X_1, X_2, X_3, X_4, X_5, X_6]^\top$, using two common factors, $F[F_1, F_2]^\top$. Suppose that the mean vector is

$$\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{bmatrix}$$

the factor loadings matrix is

$$L = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

and the diagonal matrix of specific variances (uniquenesses) is

$$\Psi = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

(a) What is the covariance matrix of X?

(b) What is the conditional distribution of X_5 given $X_3 = 2$? Describe the distribution fully.

(c) Give values for the missing elements marked A, B, C, D, and E in the matrix L^* below so that the distribution for X produced by the factor analysis model using L^* as the loadings matrix will be the same as the distribution produced using L above (assume μ and Ψ stay the same).

Explain how you obtained your answer.

$$L^* = \begin{bmatrix} 0 & -1 \\ 0 & A \\ -1 & B \\ -1 & C \\ D & 0 \\ E & 0 \end{bmatrix}$$

92. (Radford, 2014s, Practice Problem Set #3, ex.2) Consider the factor analysis model, $x = \mu + Wz + \epsilon$, where x is an observed vector of p variables, μ is the mean vector for x, z is an unobserved vector of m common factor, W is the matrix of "factor loadings", and ϵ is a random residual. We assume that $z \sim \mathcal{N}(0, I)$ and independently $\epsilon \sim \mathcal{N}(0, \Sigma)$, where Σ is diagonal with diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$.

    Let the number of observed variables be $p = 4$ and the number of common factors be $m = 1$. Give en explicit example (specifying μ, W, and Σ) showing that it is possible for the correlation of x_1 and x_2 to be negative, the correlation of x_1 and x_3 to be positive, and the correlation of x_1 and x_4 to be zero. Compute the covariance and correlation matrices of x for your example.

    Solution:

    One possible example is $\mu = [0, 0, 0, 0]^\top$, $\Sigma = I$, and $W = [1, -1, 1, 0]^\top$. The covariance matrix of x will then be

$$E[(Wz+\epsilon)(Wz+\epsilon)^\top] = E[Wzz^\top W^\top + \epsilon\epsilon^\top] = WW^\top + \Sigma = \begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 2 & -1 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

    The correlation matrix will be

$$\begin{bmatrix} 1 & -1/2 & 1/2 & 0 \\ -1/2 & 1 & -1/2 & 0 \\ 1/2 & -1/2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

93. IS IT OK TO INCLUDE THIS? (University of Nevada, STAT755, 2018s, midterm 2, ex.5 - exams/New folder (3)) The correlation matrix

$$\rho = \begin{bmatrix} 1 & 0.05 & 0.09 \\ 0.05 & 1 & 0.45 \\ 0.09 & 0.45 & 1 \end{bmatrix}$$

for a three dimensional variable $Z = [Z_1, Z_2, Z_3]^\top$ is approximated by a factor model with $m = 1$:

$$Z_i = l_i F_1 + \epsilon_i, i = 1, 2, 3$$

and

$$L = [l_1, l_2, l_3]^\top = [0.1, 0.5, 0.9]^\top$$

(a) Find the communalities h_i^2, $i = 1, 2, 3$.

(b) Find the specific variances Ψ_i, $i = 1, 2, 3$.

(c) Find Corr(Z_i,F_1), $i = 1, 2, 3$. Which variable might carry the greatest weight in naming the common factor? Why?

94. (Bishop book, ex.12.19) Let W be a $D \times M$ matrix whose columns define a linear subspace of dimensionality M embedded within a data space of dimensionality D, and let μ be a D-dimensional vector. Given a data set $\{x_n\}$ where $n = 1, \ldots, N$, we can approximate the data points using a linear mapping from a set of M-dimensional vectors $\{z_n\}$, so that x_n is approximated by $Wz_n + \mu$. The associated sum-of-squares reconstruction cost is given by

$$J = \sum_{i=1}^{N} \|x_n - \mu - Wz_n\|^2$$

First show that minimizing J with respect to μ leads to an analogous expression with x_n and z_n replaced by zero-mean variables $x_n - \overline{x}$ and $z_n - \overline{z}$, respectively, where $\overline{x}$ and $\overline{z}$ denote sample means. Then show that minimizing j with respect to z_n, where W is kept fixed, gives rise to the PCA E step (12.58):

$$\Omega = (W_{\text{old}}^\top W_{\text{old}})^{-1} W_{\text{old}}^\top \tilde{X}$$

and that minimizing J with respect to W, where $\{z_n\}$ is kept fixed, gives rise to PCA M step (12.59):

$$W_{\text{new}} = \tilde{X}^\top \Omega^\top (\Omega\Omega^\top)^{-1}$$

Solution: NOTE: In PRML, there are errors in equation (12.58) and the preceding text. In (12.58), $\tilde{X}$ should be $\tilde{X}^\top$ and in the preceding text we define Ω to be a matrix of size $M \times M$ whose n-th *column* is given by the vector $E[z_n]$.

Setting the derivative of J with respect to μ to zero gives

$$0 = -\sum_{n=1}^{N} (x_n - \mu - Wz_n)$$

from which we obtain

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n - \frac{1}{N}\sum_{n=1}^{N} W z_n = \overline{x} - W\overline{z}$$

Back-substituting into J we obtain

$$J = \sum_{n=1}^{N} \|(x_n - \overline{x} - W(z_n - \overline{z}))\|^2$$

We now define X to be a matrix of size $N \times D$ whose n-th row is given by the vector $x_n - \overline{x}$ and similarly we define Z to be a matrix of size $D \times M$ whose n-th row is given by the vector $z_n - \overline{z}$. We can then write J in the form

$$J = \mathrm{Tr}((X - ZW^\top)(X - ZW^\top)^\top)$$

Differentiating with respect to Z keeping W fixed gives rise to the PCA E-step (12.58). Similarly setting the derivative of J with respect to W to zero with $\{z_n\}$ fixed gives rise to the PCA M-step (12.59).

6.2 PCA: an FA point of view

95. (LA4ML - (Principal Components Analysis - Worksheet. Part one, ex. 3)) The following output is produced after running PCA on the iris dataset:

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.91849782	2.00446735	0.7296	0.7296
2	0.91403047	0.76727360	0.2285	0.9581
3	0.14675688	0.12604204	0.0367	0.9948
4	0.02071484		0.0052	1.0000

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
Sepal_Length	0.521066	0.377418	-.719566	-.261286
Sepal_Width	-.269347	0.923296	0.244382	0.123510
Petal_Length	0.580413	0.024492	0.142126	0.801449
Petal_Width	0.564857	0.066942	0.634273	-.523597

PRACTICAL EXERCISE

Using a computer, reproduce the results in the tables and compute the **loadings and the standardized loadings**.

6.3 Revision

96. Circle the correct answer and justify your choice:

    (a) (Radford, 2009f, final exam, ex.1h) Suppose a factor analysis model with one common factor is fit by maximum likelihood (without rescaling variables). If the estimates of the specific variances (uniquenesses), ψ_i, are all equal, then the factor loadings found will be equal to the first principal component of the covariance matrix times some scalar.

        i. True
        ii. False

    (b) (Radford, 2009f, final exam, ex.1i) When factor analysis is performed on observations of p variables, if the variables are actually independent, the factor analysis model will need to have p common factors in order to fit the data well.

        i. True

ii. False

Solution: False

(c) (Radford, 2008f, final exam, ex.7.c) When factor analysis is done, it makes no difference whether the covariance matrix or the correlation matrix is modeled as having the form $LL^\top + \Psi$ - the results are essentially the same.

i. True

ii. False

Solution: True

7 Independent Component Analysis - ICA

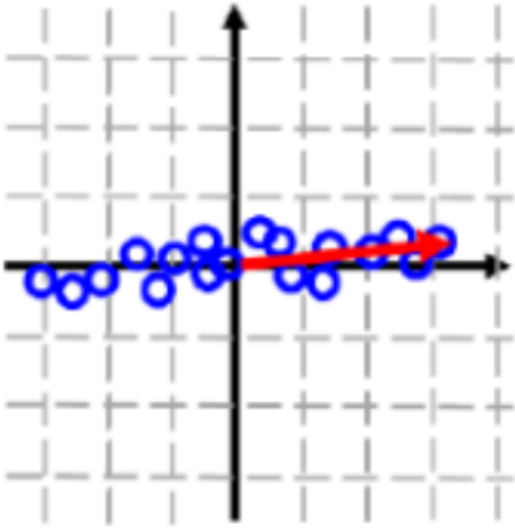
97. Is it OK to include it? (Fordham University, ML CISC 5800, HW3, ex.A.2) (`https://www.studocu.com/en-us/document/fordham-university/machine-learning/mandatory-assignments/homework-3answers/1257164/view`) Before you get started on the homework, please remember a few important points from lecture:

    - In PCA, components are orthogonal and have unit magnitude.
    - In ICA, components are not necessarily orthogonal, but they do have unit magnitude.
    - In NMF, components are not necessarily orthogonal, but they do have unit magnitude and they are all strictly non-negative.
    - You can project a vector x onto a unit vector u by taking the dot product: $u^\top x$. For PCA, this essentially can be considered u's coefficient (z) to represent the vector x.

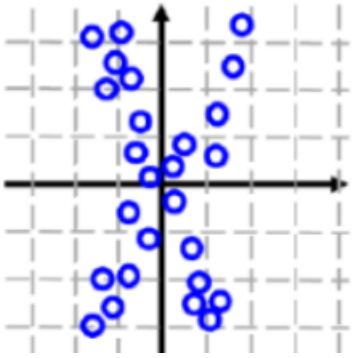
      For each o the following data sets (set A and set B), provide the top two principal components and the top one or two independent components (if there are two clear independent components, you must provide both). Express each component as a 2-element vector $\begin{bmatrix}\text{num}_1\\ \text{num}_2\end{bmatrix}$.

    You may print out this page and draw arrows for partial credit. The vector estimate of each component should be estimated to the nearest tenth. E.g., for the following data, we may have a direction as a roughly horizontal arrow.

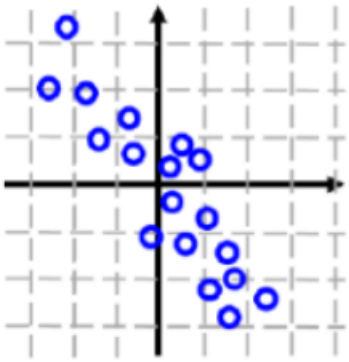
    Roughly estimated as: dir $\approx \begin{bmatrix}3.0\\ 0.5\end{bmatrix}$ or $u \approx \begin{bmatrix}1.0\\ 0.2\end{bmatrix}$



Set A:



Set B:



Solution:

Set A:

PC1: $\begin{bmatrix} 0.5 \\ 2 \end{bmatrix} \to \begin{bmatrix} 0.24 \\ 0.87 \end{bmatrix}$

PC2 (orthogonal to PC1): $\begin{bmatrix} 0.87 \\ -0.24 \end{bmatrix}$

IC1: $\begin{bmatrix} -1 \\ 2 \end{bmatrix} \to \begin{bmatrix} -0.45 \\ 0.89 \end{bmatrix}$

IC2: $\begin{bmatrix} 1 \\ 2 \end{bmatrix} \to \begin{bmatrix} 0.45 \\ 0.89 \end{bmatrix}$

Set B:

PC1: $\begin{bmatrix} 1 \\ -1 \end{bmatrix} \to \begin{bmatrix} 0.71 \\ -0.71 \end{bmatrix}$

PC2 (orthogonal to PC1): $\begin{bmatrix} 0.71 \\ 0.71 \end{bmatrix}$

IC1: $\begin{bmatrix} 1 \\ -1 \end{bmatrix} \to \begin{bmatrix} 0.71 \\ -0.71 \end{bmatrix}$

98. (from Unsupervised ML, Helsinki, Aapo Hyvarinen, Ex. set 5, ex.2) For zero-mean random variable, skewness of a distribution is defined to be its third moment, i.e.

$$\text{skew}(X) = E(X^3)$$

It measures the asymmetry of a distribution. If the independent variables s have highly asymmetric distribution, skewness can be used to perform ICA.

Suppose Z is $N \times N$ data matrix, and denote as z_k the columns of Z with each fixed $1 \leq k \leq K$. We would like to maximize

$$J(w) = \frac{1}{K} \sum_{k=1}^{K} (w \cdot z_k)^3$$

with respect to a vector $w \in \mathbb{R}^N$ under the constraint $\|w\| = 1$

(a) Find the gradient $\nabla J(w)$.

(b) What is the gradient-ascent optimization iteration, taking into account the constraint?

(c) Take the limit of large stepsizes, ie.e, $\mu \to \infty$. What is the optimization iteration now?

99. (from Unsupervised ML, Helsinki, Aapo Hyvarinen, Ex. set 5, ex.3) Assume the data $z_1, \ldots, z_K$ is iid sample generated by the model $Z = AS$, where Z and S are random vectors, A is orthogonal $N \times N$ matrix, and the components of S,S_n when $a \leq n \leq N$, are independent from each other.

    (a) Write down the log-likelihood $l(A|z_1, \ldots, z_K)$ of A in terms of the distributions $p_{S_n}(s_n)$ which may be arbitrary.

    (b) Show that the log-likelihood does not depend anymore on the matrix A is S_n are Gaussian.

100. (EPFL, Unsupervised and reinforcement learning in neural networks Week 4, ex.1) In class, it was argued that a mixture of statistically independent sources tends to be more Gaussian that the sources themselves. This argument served as the basis for ICA algorithms that rely on non-Gaussianity. In this exercise, we want you to show that the non-Gaussianity argument does not rely on the summation of a large number of statistically independent sources, but that it works already for two sources. Remember that the kurtosis is defined as $\kappa(x) = E[x^4] - 3E[x^2]^2$. Now let x_1 and x_2 be statistically independent and let both have zero mean.

    (a) Show that the kurtosis of $y = x_1 + x_2$ is given by $\kappa(x) = \kappa(x_1) + \kappa(x_2)$

    (b) Show that the kurtosis of $y = \alpha x$ with $\alpha \in \mathbb{R}$ is given by $\kappa(y) = \alpha^4 \kappa(x)$.

    (c) Use the first two subpoints to show that the kurtosis of $y = \sqrt{a}x_1 + \sqrt{1-a}x_2$, $a \in [0, 1]$, is given by

    $$\kappa(y) = a^2\kappa(x_1) + (1-a)^\kappa(x_2)$$

    (d) Let $\kappa(x_1) = c$ and $\kappa(x_2) = d$ be the kurtoses of x_1 and x_2. Assume that both signals are super-Gaussian and that $0 < c < d$. Show that the kurtosis of the mixture $y = \sqrt{a}x_1 + \sqrt{1-a}x_2$ has maxima for $a = 0$ and $a = 1$, and that $a = 0$ is the global maximum.

    (e) Which value(s) of a maximize the kurtosis if the signals x_1 and x_2 are sub-Gaussian: $c < d < 0$?

101. (EPFL, Unsupervised and reinforcement learning in neural networks Week 4, ex.2) Consider an ICA algorithm, that aims at maximizing $J(w) =< F(y) >$, where $y = w^\top x$ and $F(y) = \frac{1}{a}\log\cosh(ay)$. The maximization is done by gradient descent.

    (a) Show that: $\frac{dF}{dy} = \tanh(ay)$.

(b) Calculate $\frac{dF}{dw_j}$ for $y = \sum_k w_k x_k$.

(c) Show that a gradient ascent on $J(w) =< F(w^\top x) >$ leads to a Hebbian rule. (Hint: Make the transition from a bath rule to an online rule.)

102. (EPFL, Unsupervised and reinforcement learning in neural networks Week 4, ex.3) n the previous exercise, we discussed a simple ICA algorithm based in gradient ascent. Here, we will go one step further and maximize the non-Gaussianity of the mixture using the Newton method, that yields a faster convergence. The resulting learning algorithm is known as **fastICA**.

(a) We want to maximize the measure of non-Gaussianity F under the constraint of a normalized weight vector, i.e. $w^\top w = 1$. This corresponds to finding the maximum of the function $J(w) =< F(w^\top x) >$. Derive the Taylor expansion $J^*(w)$ of $J(w)$ around w_0 up to second order in w.

(b) A Newton step consists of setting the next value w_{new} to the vector that maximizes the second-order approximation J^* around the previous weight vector w_0. Show that this leads to the *fastICA* update rule:

$$w_{\text{new}} =< g(w_0^\top x)x > - < g'(w_0^\top x) > w_0$$

with $g := \frac{dF(y)}{dy}$ and $g' = \frac{dg(y)}{dy}$.

(Hint: Make the approximation that $< g'(w_0^\top x)xx^\top > \approx < g'(w_0^\top x) > \cdot < xx^\top >$ and exploit the fact that the data is pre-whitened, $< xx^\top >= E$ with identity matrix E. Finally, remember that the weight vector gets re-normalized to unity in the *fastICA* algorithm after the above update rule is applied.)

8 Canonical Correlation Analysis - CCA

8.1 CCA

103. IS IT OK? (THE UNIVERSITY OF CHICAGO Graduate School of Business Business 41912, Spring Quarter 2010, Mr. Ruey S. Tsay, final exam, ex.3 - exams/New Folder (7)) Consider the following four variables:

- X_1: 1973 nonprimary homicides
- X_2: 1973 primary homicides (homicides involving family or acquaintances)
- Y_1: 1970 severity of punishment (median months served)
- Y_2: 1970 certainty of punishment (number of admissions to prison divided by number of homicides)

The correlation matrix of $(X_1, X_2, Y_1, Y_2)^\top$ is

$$R = \begin{bmatrix} 1.0 & 0.615 & -0.111 & -0.266 \\ 0.615 & 1.0 & -0.195 & -0.085 \\ -0.111 & -0.195 & 1.0 & -0.269 \\ -0.266 & -0.085 & -0.269 & 1.0 \end{bmatrix}$$

(a) Find the sample canonical correlations between $X = (X_1, X_2)^\top$ and $Y = (Y_1, Y_2)^\top$.

(b) Find the first canonical pair $\hat{U}_1$ and $\hat{V}_1$.

(c) Find the second canonical pair $\hat{U}_2$ and $\hat{V}_2$.

Solution:

(a) The canonical correlations are 0.327 and 0.171.

(b) The vectors are $(1, -0.003)^\top$ and $(-0.524, -0.851)^\top$.

(c) The vectors are $(-0.523, 0.852)^\top$ and $(-0.923, 0.384)^\top$.

104. (Applied Multivariate Statistical Analysis by Johnson Wichern, ex.10.1) IS IT OK?Consider the covariance matrix:

$$\text{Cov}\left(\begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} 100 & 0 & 0 & 0 \\ 0 & 1 & 0.95 & 0 \\ 0 & 0.95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{bmatrix}$$

Verify that the first pair of canonical variates are $U_1 = X_2^{(1)}$, $V_1 = X_1^{(2)}$ with canonical correlation $\rho_1^* = 0.95$.