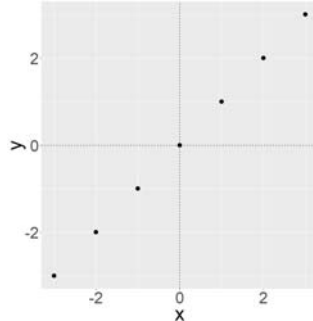


פתרון לשאלה פתוחה 3 :

נתונות הנקודות הבאות במרחב דו-מימדי :



א. (1 נק) מהו אחוז השונות בנתונים המוסבר ע"י כל ציר.

כל ציר מסביר 50% מהשונות הכוללת בנתונים.

ב. (1 נק) חוקר ביצע PCA על הנתונים. מה יהיה אחוז השונות המוסבר ע"י כל אחד מצירי ה-PCA?

הציר הראשון יסביר 100% מהשונות בעוד שהציר השני יסביר 0% מהשונות (הנתונים לא ישתנו לאורכו).

בסעיפים הבאים מפורטים שינויים שנוספו לנקודות המתוארות לעיל. בכל סעיף, לאחר ביצוע השינויים החוקר ביצע PCA על הנקודות. שערכו והסבירו עבור כל סעיף מה יהיה אחוז השונות המוסבר ע"י כל אחד מצירי ה-PCA.

ג. (2 נק) החוקר הוסיף לנקודות הנ"ל רעש קבוע: $\varepsilon(x) = 0.01 \cdot x$.

כמו (ב), כיוון שהרעש קבוע והתלות הליניארית נשמרת.

ד. (2 נק) החוקר הוסיף לנקודות הנ"ל רעש קבוע: $\varepsilon(x) = 0.01 \cdot x^2$.

הציר הראשון יסביר קרוב ל-100% מהשונות בעוד שהציר השני יסביר קרוב ל-0% מהשונות.

ה. (2 נק) החוקר הוסיף לנקודות הנ"ל רעש שנדגם אקראית מהתפלגות נורמלית בעלת ממוצע 0 וסטיות תקן 100.

כיוון שהתלות הליניארית כמעט ולא קיימת, כל ציר יסביר קרוב כ-50% מהשונות בנתונים.

פתרון לשאלה פתוחה 4:

1. (2 נק) Batch effect הינו מצב שבו יש השפעה של נסיבות טכניות בניסוי על הסיגנל המתקבל. למשל, שימוש בריכוז מעט יותר גבוה של ראגנט מסויים ביום א של הניסוי, בו נמדדו 10 דוגמאות, לעומת יום ב של הניסוי, בו נמדדו 10 דוגמאות אחרות של הניסוי. הסבירו כיצד PCA יכול לעזור בקביעה האם יש batch effect עבור הניסוי הבא. הסבירו מה התוצאות הצפויות של ה-PCA במידה שיש batch effect ובמידה שאין batch effect. בניסוי נלקחו דוגמאות דם מחולים ובריאים ומדדו ביטוי גנים ב- microarrays ב-3 ימים שונים (בכל יום נמדדו 10 דוגמאות בבת אחת):

	חולים	בריאים
יום א	5 דוגמאות	5 דוגמאות
יום ב	5 דוגמאות	5 דוגמאות
יום ג	5 דוגמאות	5 דוגמאות

פתרון: PCA מראה מה משפיע על השונות המירבית בנתונים. ניתן לשרטט PCA של כל הנתונים, לצבוע את נקודות המידע לפי חולים/ בריאים ולפי ימי ניסוי ולראות האם השונות בנתונים מגיעה מאחד מהם. אם השונות המירבית מגיעה מימי הניסוי השונים סימן שיש batch effect שהוא גדול יותר מהסיגנל האמיתי בנתונים.

2. האם תמיד PCA יכול לענות על השאלה האם קיים batch effect בניסוי כלשהו? עבור כל אחד מהמקרים הבאים קבעו האם PCA יכול או לא יכול לענות על השאלה והסבירו למה:
 (א) (2 נק) כל דוגמאות החולים נמדדו בבית חולים אחד וכל דוגמאות הבריאים נמדדו בבית חולים אחר, ורוצים לבדוק האם היה batch effect לכל מוסד (בתהליך לקיחת הדוגמאות והפקת ה-RNA שנעשה במרכז בכל בית חולים).
 (ב) (2 נק) בית חולים אחד סיפק 7 דוגמאות של בריאים ו-7 דוגמאות של חולים ובית חולים אחר סיפק 8 דוגמאות של בריאים ו-8 דוגמאות של חולים. רוצים לבדוק האם היה batch effect לכל מוסד (בתהליך לקיחת הדוגמאות והפקת ה-RNA שנעשה במרכז בכל בית חולים).

פתרון: בראשון לא ניתן, ובשני ניתן. זאת מכיוון שבראשון לא ניתן להפריד בין סיגנל אמיתי שאנחנו מחפשים לבין מקור הרעש אותו אנחנו רוצים לבדוק.

3. (2 נק) Outliers הם נקודות מידע עם ערכים קיצוניים מאוד לעומת אחרים. איך outliers ישפיעו על PCA? הסבירו 2 דרכים להתמודד עם השכלות אלה.
 פתרון: Outliers ישפיעו חזק על ה-PCA מכיוון שהם יגדירו את ציר השונות הגדול ביותר. ניתן להתמודד עם הבעיה על ידי הורדת הנקודות האלה מה-data או ע"י scaling.