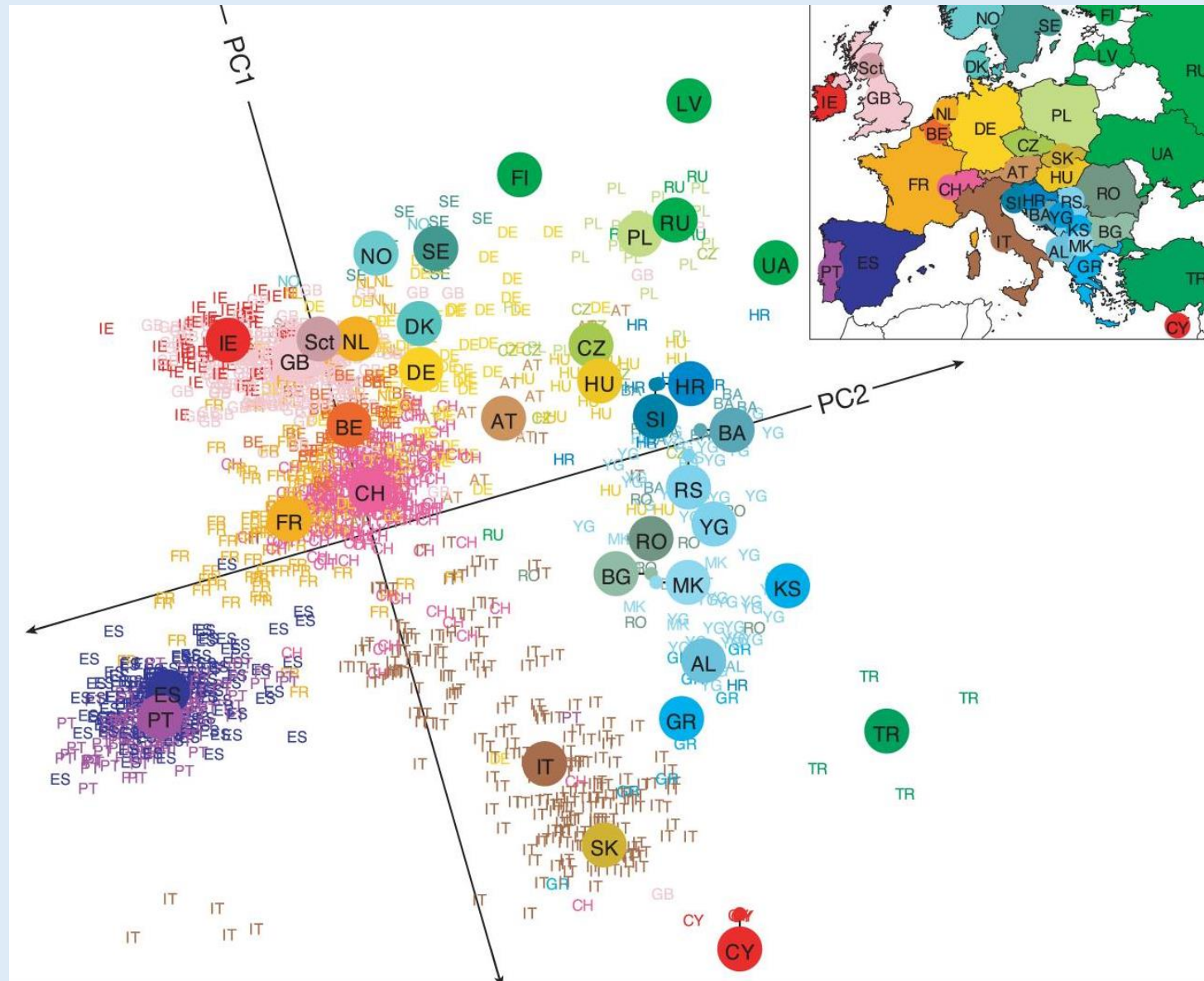


תרגול 13 ואחרון

PCA

PCA - מוטיבציה



PCA - מוטיבציה

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

נניח ויש לנו 6 עכברים ואנו רוצים לדעת אילו עכברים הכי דומים זה לזה.

**PCA עוזר לנו לצמצם את מספר המימדים
ולתפוס את השונות הגבוהה ביותר בנתונים**

מה זה PCA?

PCA הוא אלגוריתם שעוזר לנו להוריד את מספר המימדים של הנתונים שלנו תוך שמירה על רוב המידע המקורי.

נניח ויש לנו 10 מדידות עבור 100 מטופלים בחדר מיון:

- יש לנו 10 מימדים (ואי אפשר לייצג את כולם בגרף אחד)

- כל דגימה היא נקודה במרחב 10 מימדי (כמו שבדו-מימד, לכל נקודה יש 2 מימדים)

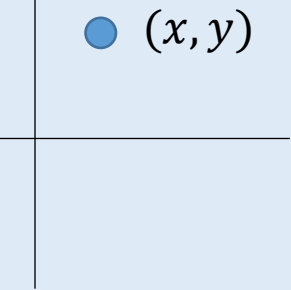
- האלגוריתם של PCA (Principle Component Analysis) ייצור מערכת צירים חדשה שלא דווקא דומה

למערכת הצירים המקורית של הנתונים. למשל, אם נתונים לנו גובה (h), משקל (w), BMI (BMI), מספר

לויקוציטים (WBC) וכו', כל ציר ב-PCA יהיה קומבינציה לינארית של המימדים הנ"ל. לדוגמא:

$$PC1 = [0.1 \cdot h] + [0.2 \cdot w] + [0 \cdot BMI] + [0.6 \cdot WBC] + \dots$$

$$PC2 = [0.4 \cdot h] + [0 \cdot w] + [0 \cdot BMI] + [0.2 \cdot WBC] + \dots$$

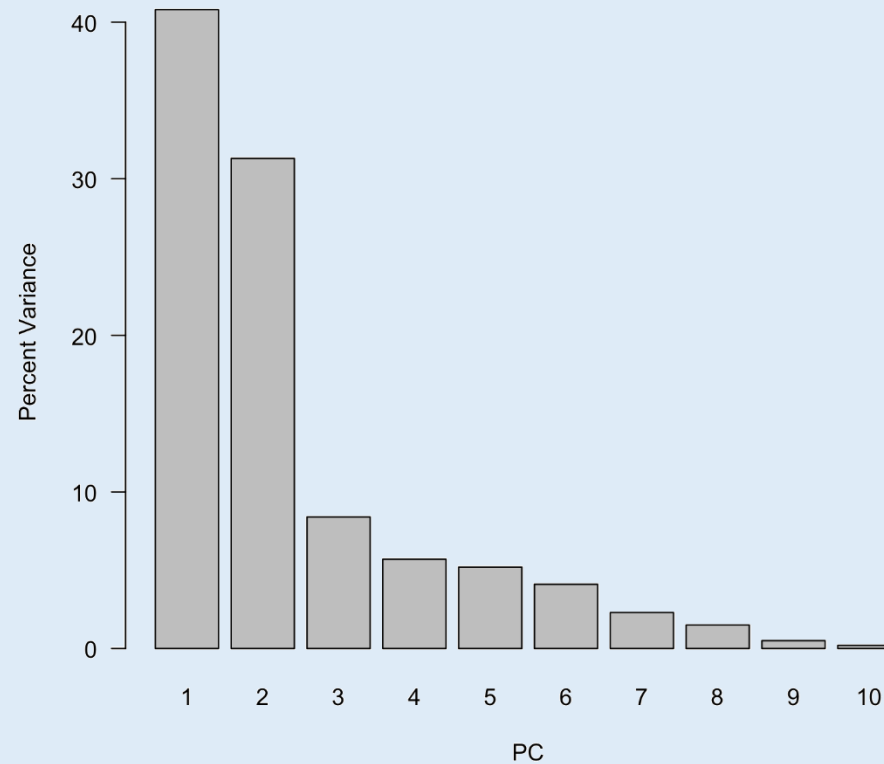


• (x, y)

מה זה PCA?

כל ציר PC מתאר אחוז מסויים מהשונות של הנתונים, והאלגוריתם מסדר את הצירים כך ש-PC1 מתאר את אחוז השונות הגבוה ביותר, PC2 השני הגבוה ביותר וכו'.

PCA Scree Plot



מה זה PCA?

Patient	H	W	H/W	BMI	WBC
001	1.65	70	42.4	25.7	5
002	1.85	85	45.9	24.8	12
...					

$$PC1 = [0.2 \cdot H] + [0.3 \cdot W] + [0 \cdot H/W] + [0.7 \cdot BMI] + [0.6 \cdot WBC] + \dots$$

$$PC2 = [0.4 \cdot H] + [0.9 \cdot W] + [0 \cdot H/W] + [0.1 \cdot BMI] + [0.2 \cdot WBC] + \dots$$

	H	W	H/W	BMI	WBC
PC1	0.2	-0.3	0	0.7	0.6
PC2	0.4	0.9	0	0.1	-0.2
...					

loadings

המשקל של כל משתנה בכל PC

נמשיך את הדוגמא של המיון:

אם נתונים לנו 2 מטופלים והנתונים שלהם:

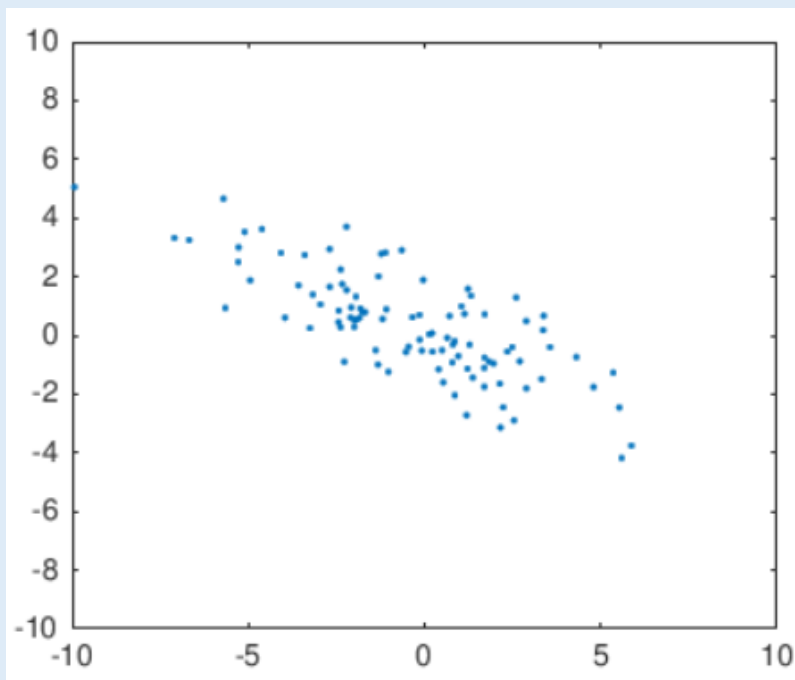
Patient	PC1	PC2	...
001	42.32	82.04	
002	50.43	82.12	
...			

scores

הקואורדינטות החדשות של כל נקודה

PCA על נתונים דו-מימדיים

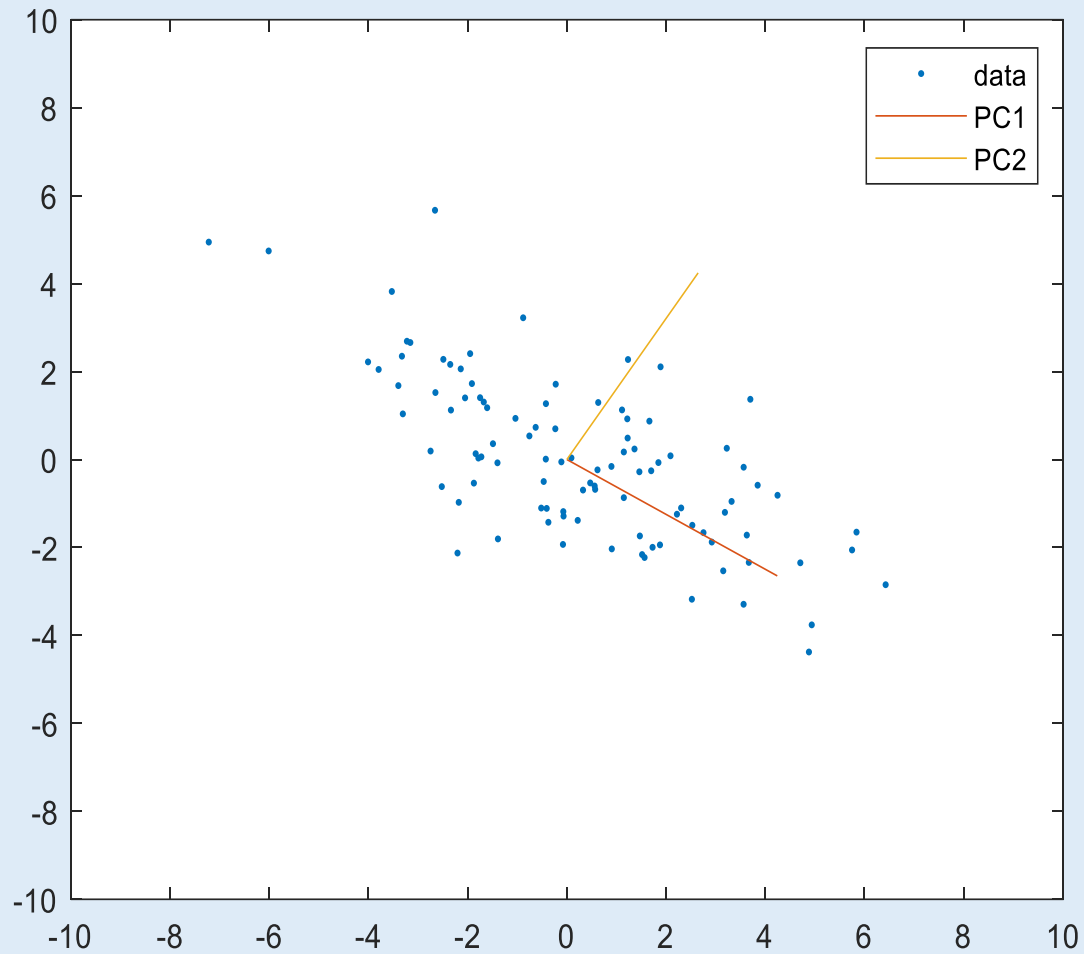
נתחיל משני מימדים ונסתכל איך זה נראה ברב מימד ברב מימד



(לא גרף PCA)

איזה ציר (מצוייר או לא) תופס את רוב השונות של הנתונים?

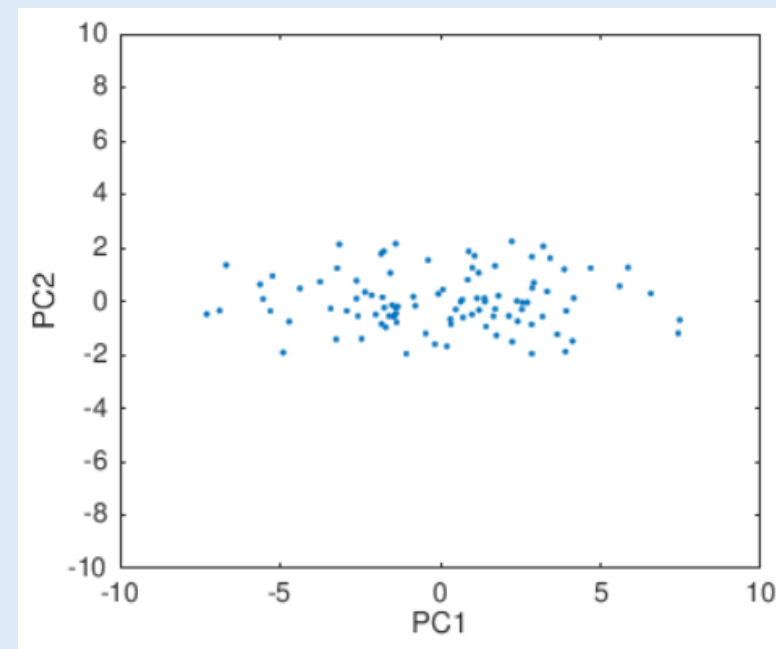
PCA על נתונים דו-מימדיים



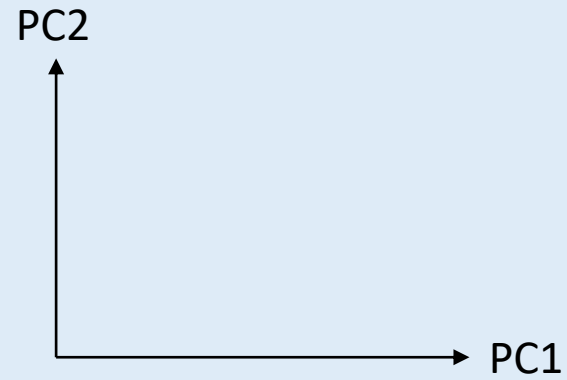
אחוז השונות המוסברת:

$PC1 \rightarrow 87.2\%$

$PC2 \rightarrow 12.8\%$

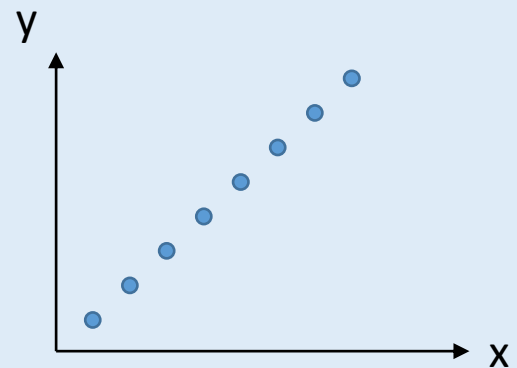


שאלה



- איך נצפה שיראה גרף PCA שבו ציר PC1 תופס 100% מהשונות?

- חוקר ביצע PCA על הנתונים. מה יהיה אחוז השונות המוסבר ע"י כל אחד מהצירים?

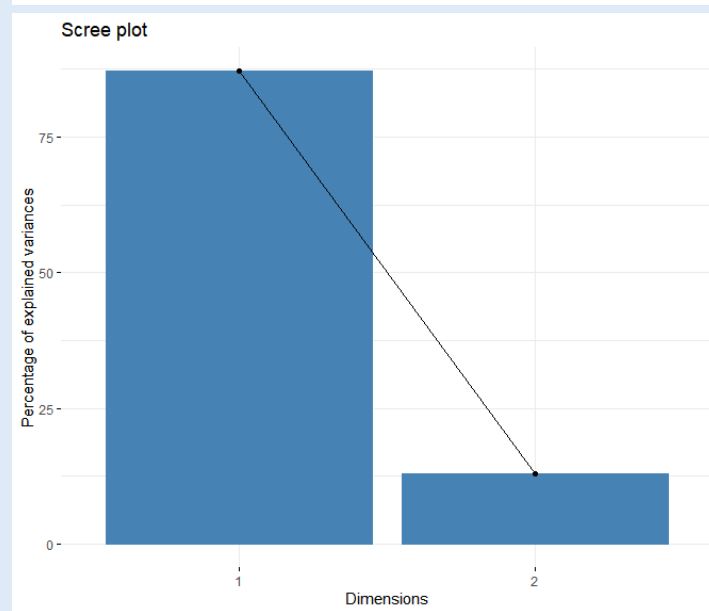


PCA על נתונים דו-מימדיים

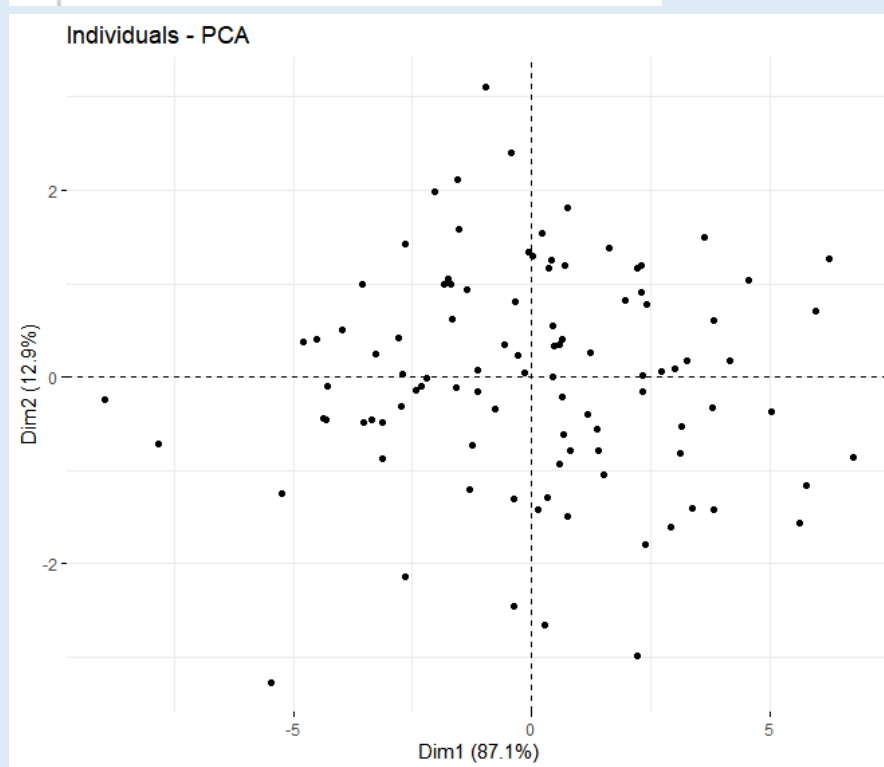


```
> summary(p)
Importance of components:
                PC1    PC2
Standard deviation 3.0166 1.1585
Proportion of Variance 0.8715 0.1285
Cumulative Proportion 0.8715 1.0000
```

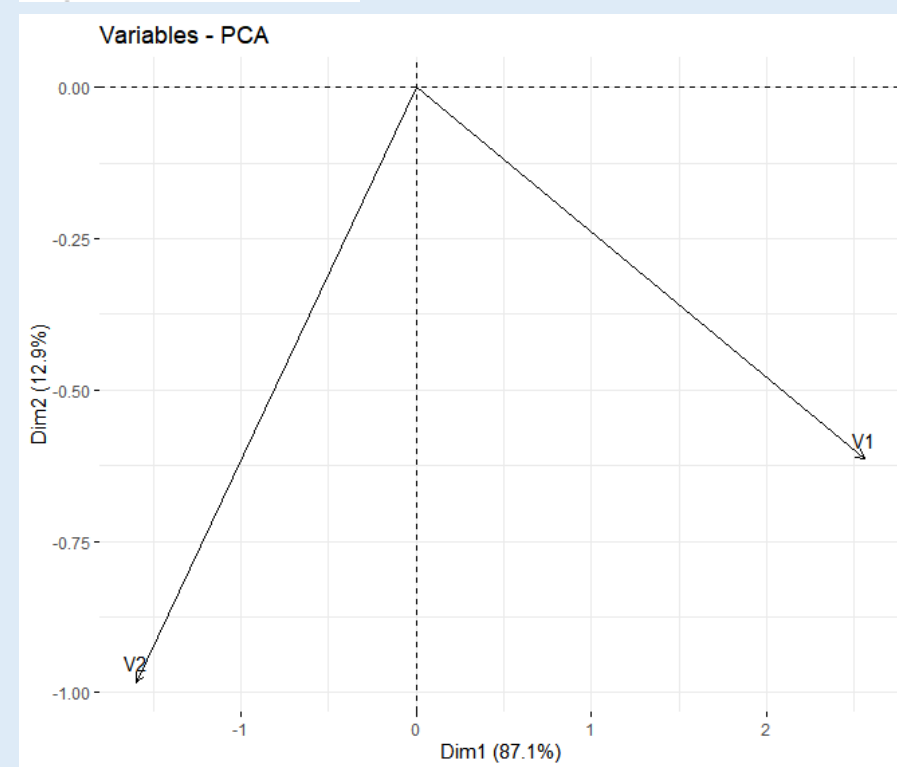
```
> fviz_screplot(p)
```



```
> fviz_pca_ind(p, geom = "point")
```

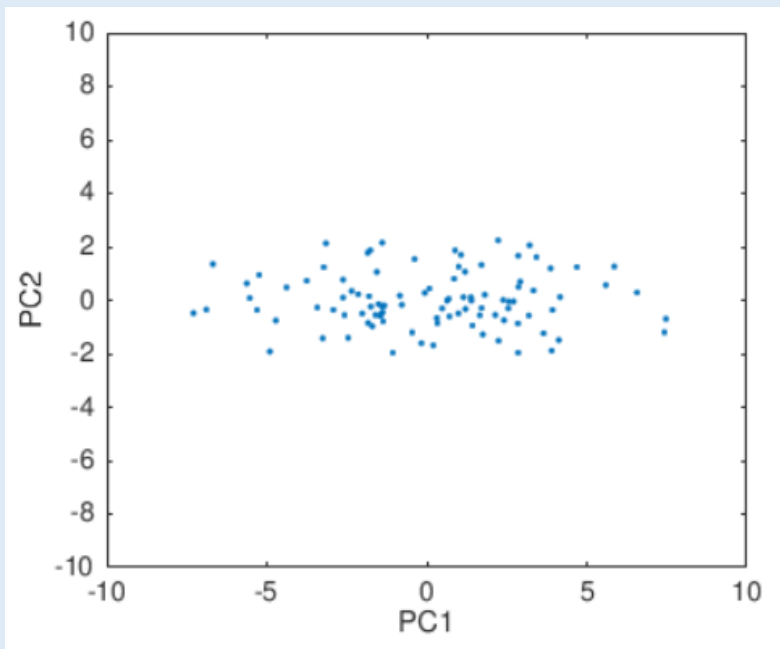


```
> fviz_pca_var(p)
```



Dim1 = PC1
Dim2 = PC2

מה זה PCA?



נשים לב:

1. סכום % השונות המוסברת של ה-PCs צריך להיות 100
2. מספר המימדים (PC=) האפשריים שווה למספר המימדים של הנתונים
3. במקרה הזה עשינו רק סיבוב בדו-מימד של הנתונים – ברב מימד, הסיבוב יהיה במרחב הרב-מימד.
4. ברב מימד נוכל לבחור את ה-PCs בהם ראינו את השונות המירבית של הנתונים, ואין צורך להסתכל על כל ה-PCs האפשריים.

חישוב PCA

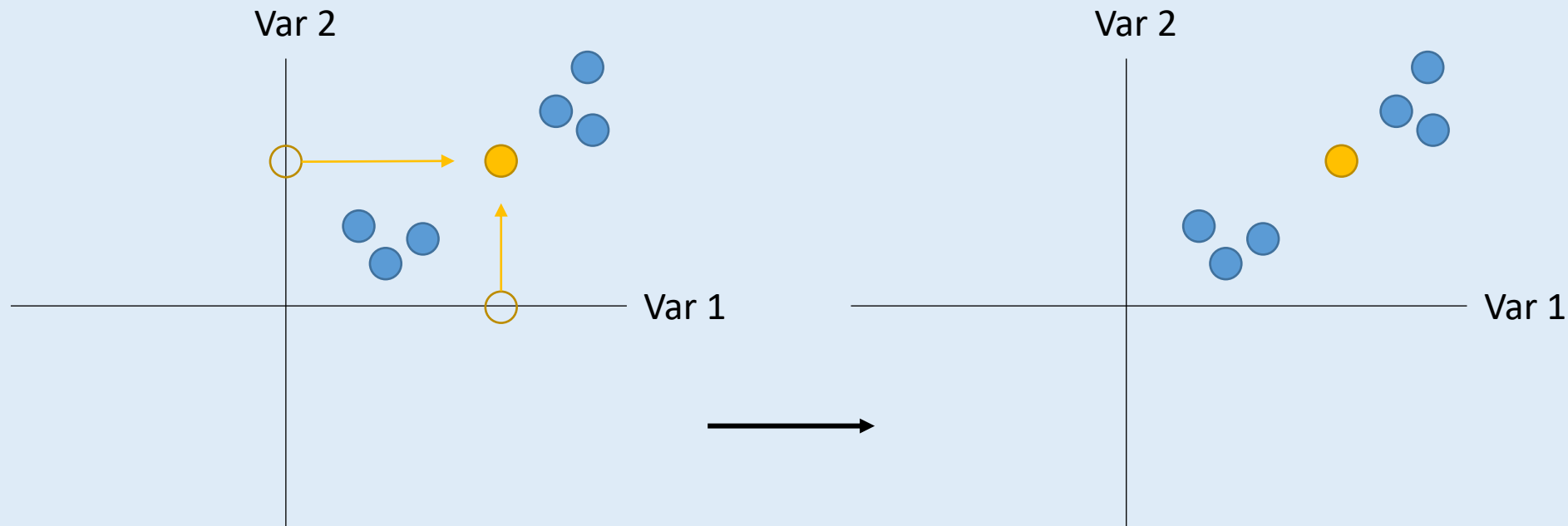
משתנים (מימדים של הנתונים)

Patient	H	W	BMI	WBC
001	1.65	70	25.7	5
002	1.85	85	24.8	12
...				

איך האלגוריתם עובד?

1. מרכז (centering) של הנתונים כך שהממוצע של כל משתנה בכל מימד הוא אפס.

שימו לב שהמרכז לא שינה את איך שהנתונים מפוזרים על הצירים



חישוב PCA

משתנים (מימדים של הנתונים)

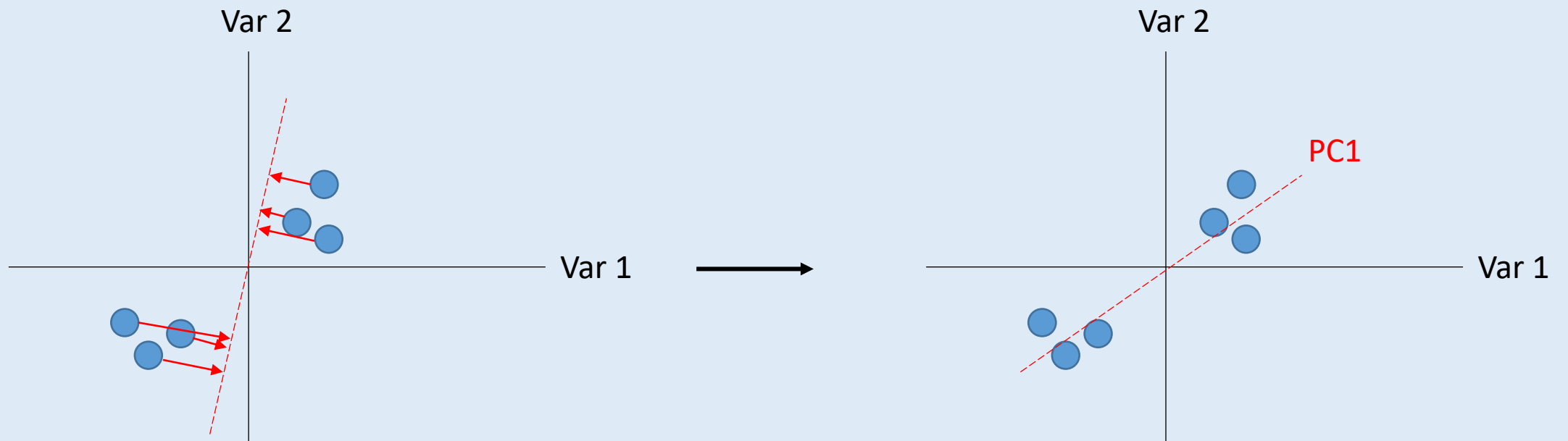
Patient	H	W	BMI	WBC
001	1.65	70	25.7	5
002	1.85	85	24.8	12
...				

2. נחפש ציר (שעובר דרך הראשית) וייתן לנו את השונות הגבוהה ביותר האפשרית.

לכל קו שעובר דרך הנקודות מחשבים את סכום ריבועי המרחקים מכל נקודה לקו.

הציר שמצמצם את הסכום הזה בצורה המשמעותית ביותר יהיה PC1.

$$\frac{SS_{\text{distances}}(\text{PC1})}{n - 1} = \text{Variance}(\text{PC1})$$



חישוב PCA

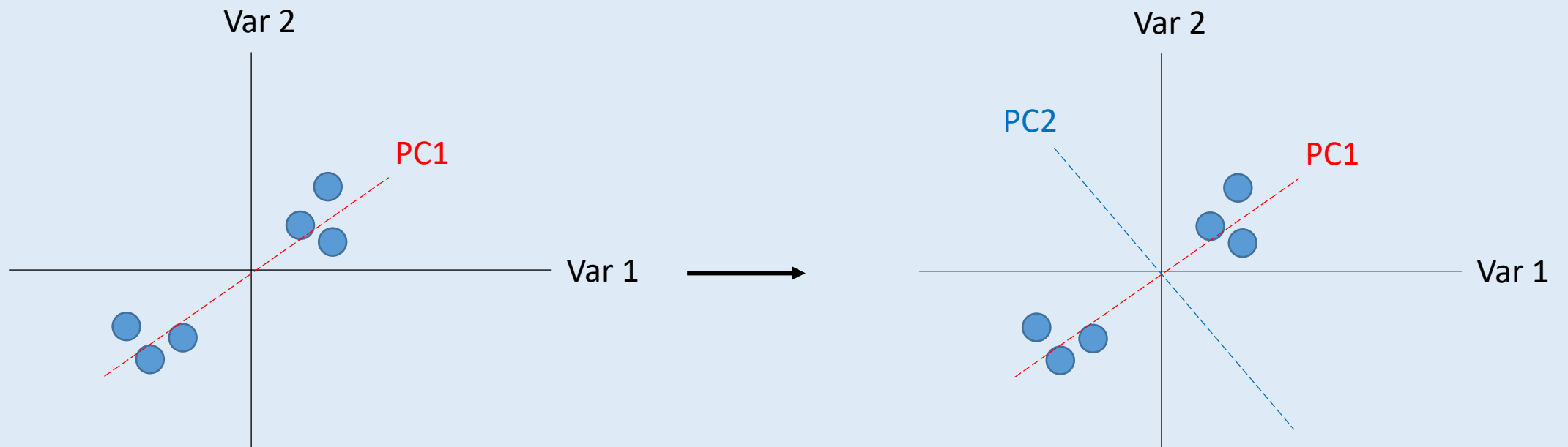
משתנים (מימדים של הנתונים)

Patient	H	W	BMI	WBC
001	1.65	70	25.7	5
002	1.85	85	24.8	12
...				

3. כעת, מתוך כל הצירים שניצבים (אנכיים) לציר הראשון, נחפש את הציר שנותן לנו

את השונות המקסימלית. זהו PC2.

$$\frac{SS_{\text{distances}}(\text{PC2})}{n - 1} = \text{Variance}(\text{PC2})$$



חישוב PCA

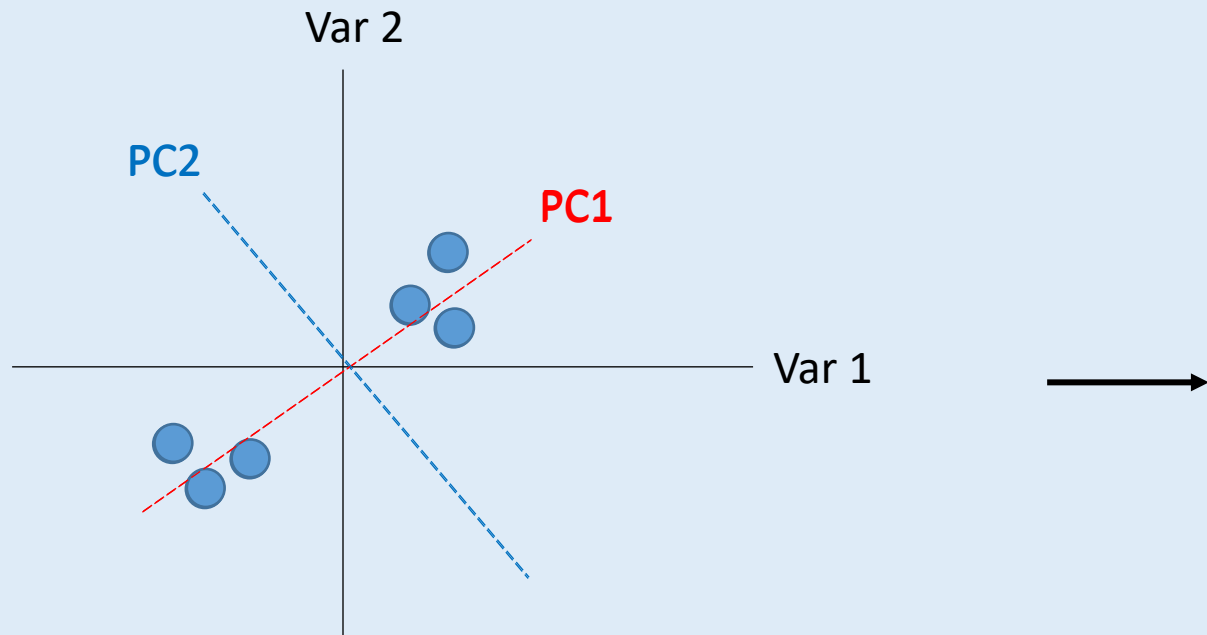
משתנים (מימדים של הנתונים)

Patient	H	W	BMI	WBC
001	1.65	70	25.7	5
002	1.85	85	24.8	12
...				

3. נמשיך בצורה הזו עד שנסיים.

4. לאחר שנמצא את כל ה-PCs, נעביר את הנתונים למערכת הצירים החדשה

5. ניתן להזניח חלק מהמימדים שהשונות המוסברת בהן נמוכה.

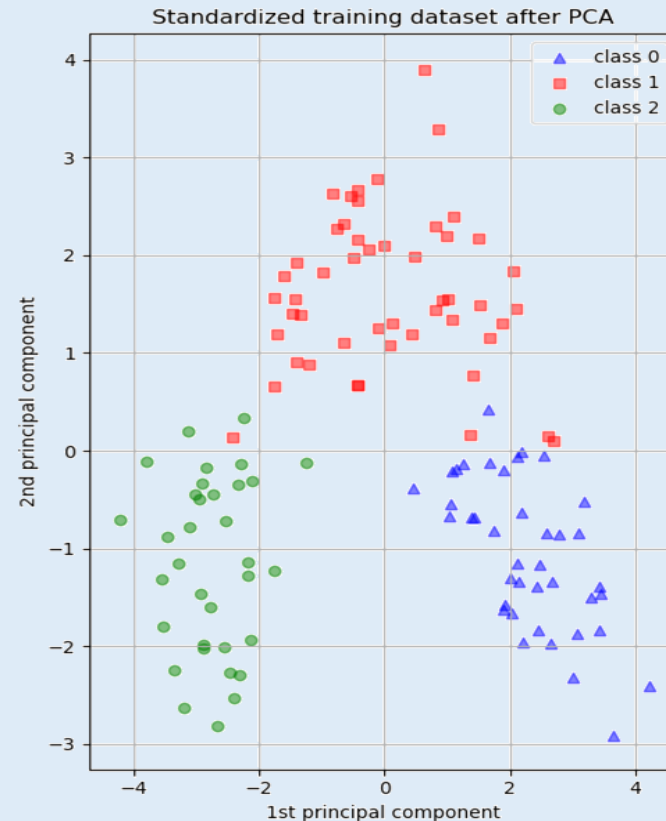
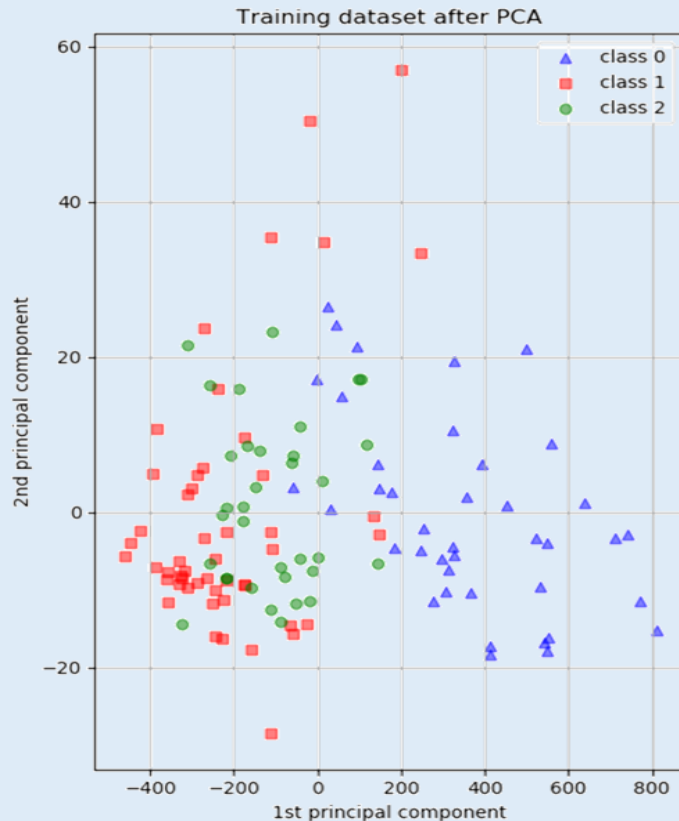


נרמול של הנתונים

כמו שלמדנו בשיעורים הקודמים, לפעמים יש הבדלים משמעותיים בין הסקאלות של המשתנים השונים (למשל גובה הוא

בין 0.5-2, אבל ערכי משקל גבוהים ב-2 סדרי גודל).

מסיבה זו, צריך לנרמל את הנתונים – על מנת למנוע "השתלטות" מלאכותית של מימד אחד על כל השונות בנתונים.



דוגמא בדו-מימד

הדוגמא הזו הייתה קצת טריוויאלית, כי מלכתחילה הנתונים שלנו היו ממימד נמוך ולא היה צורך להפחית מימדים.

PCA רק סובב את מערכת הצירים כך שהצירים יהיו לפי הגורמים הראשיים (PCs) במקרה של נתונים רב-מימדיים הסיבוב יהיה במרחב רב מימדי.

דוגמא ברב-מימד

1. טענו את מאגר הנתונים של חולי הלימפומה. זוהי מטריצה בגודל 130×29 כאשר העמודה הראשונה מתארת האם הטיפול בחולה הצליח או לא (TRUE/FALSE) וערכי 28 גנים. כלומר, עבור כל מטופל יש מידע 29 מימדי – כל נבדק מיוצג ע"י נקודה במרחב 29 מימדי).
2. בצעו PCA וציירו את הגרף המתאים. כמה שונות תופסים שני הצירים הראשונים?
3. הביטוי של איזה גן או גנים יורד ככל שנמצאים יותר ימינה על PC1?
4. מהם ה-scores של כל מטופל?
5. הציגו את כל הדגימות על גרף PCA ואת הנקודות הציגו בצבע שונה עבור מטופלים עבורם הטיפול הצליח ונכשל.
6. מה יקרה אם נשתמש ב-PC2 ו-PC3?
7. מה קורה כאשר לא מנרמלים את הנתונים לפני?

דוגמא ברב-מימד

1. טענו את מאגר הנתונים של חולי הלימפומה. זוהי מטריצה בגודל 130×29 כאשר העמודה הראשונה מתארת האם הטיפול בחולה הצליח או לא (TRUE/FALSE) וערכי 28 גנים. כלומר, עבור כל מטופל יש מידע 29 מימדי – כל נבדק מיוצג ע"י נקודה במרחב 29 מימדי).
2. בצעו PCA וציירו את הגרף המתאים. כמה שונות תופסים שני הצירים הראשונים?
3. הביטוי של איזה גן או גנים יורד ככל שנמצאים יותר ימינה על PC1?
4. מהם ה-scores של כל מטופל?
5. הציגו את כל הדגימות על גרף PCA ואת הנקודות הציגו בצבע שונה עבור מטופלים עבורם הטיפול הצליח ונכשל.
6. מה יקרה אם נשתמש ב-PC2 ו-PC3?
7. מה קורה כאשר לא מנרמלים את הנתונים לפני?

דוגמא ברב-מימד

1. טענו את מאגר הנתונים של חולי הלימפומה. זוהי מטריצה בגודל 130×29 כאשר העמודה הראשונה מתארת האם הטיפול בחולה הצליח או לא (TRUE/FALSE) וערכי 28 גנים. כלומר, עבור כל מטופל יש מידע 29 מימדי – כל נבדק מיוצג ע"י נקודה במרחב 29 מימדי).
2. בצעו PCA וציירו את הגרף המתאים. כמה שונות תופסים שני הצירים הראשונים?
3. הביטוי של איזה גן או גנים יורד ככל שנמצאים יותר ימינה על PC1?
4. מהם ה-scores של כל מטופל?
5. הציגו את כל הדגימות על גרף PCA ואת הנקודות הציגו בצבע שונה עבור מטופלים עבורם הטיפול הצליח ונכשל.
6. מה יקרה אם נשתמש ב-PC2 ו-PC3?
7. מה קורה כאשר לא מנרמלים את הנתונים לפני?

דוגמא ברב-מימד

1. טענו את מאגר הנתונים של חולי הלימפומה. זוהי מטריצה בגודל 130×29 כאשר העמודה הראשונה מתארת האם הטיפול בחולה הצליח או לא (TRUE/FALSE) וערכי 28 גנים. כלומר, עבור כל מטופל יש מידע 29 מימדי – כל נבדק מיוצג ע"י נקודה במרחב 29 מימדי).
2. בצעו PCA וציירו את הגרף המתאים. כמה שונות תופסים שני הצירים הראשונים?
3. הביטוי של איזה גן או גנים יורד ככל שנמצאים יותר ימינה על PC1?
4. מהם ה-scores של כל מטופל?
5. הציגו את כל הדגימות על גרף PCA ואת הנקודות הציגו בצבע שונה עבור מטופלים עבורם הטיפול הצליח ונכשל.
6. מה יקרה אם נשתמש ב-PC2 ו-PC3?
7. מה קורה כאשר לא מנרמלים את הנתונים לפני?