פרק 12: הפחתת מימדיות מידע

12.1 הקדמה

אחת הבעיות המרכזיות בניתוח מידע היא בחירת מאפיינים. כאשר המידע רב מימדי יש חשיבות רבה לבחירת המאפיינים בהם נשתמש לאלגוריתמי הלמידה השונים. בחירת מאפיינים מאפשרת:

- להפחית את קללת המימדיות
- לשפר את ההכללה
- להאיץ את תהליך הלימוד
- לשפר את הבנת המידע ולתת אינטרפרטציה למודל הנלמד.

בעיית בחירת המאפיינים היא בעיה קשה וניתן להראות שבאופן כללי לא ניתן לפתור אותה בדרך שאינה שקולה לחיפוש ממצה (exhaustive search). חיפוש מקיף אינו מעשי בבעיות עם עשרות מאפיינים מהם מחפשים יותר ממספר מאפיינים בודדים.

באפן כללי יש שתי גישות להפחתת מימדיות:

1. דירוג מאפיינים: דירוג "איכות" המאפיינים והשמטת מאפיינים שאינם מועילים מספיק.
2. בחירת תת קבוצה: חיפוש אחר תת קבוצה מיטבית של מאפיינים.

לעיתים חייבים לבחור תת קבוצה: למשל כאשר הפלט הוא פונקציית XOR של שני מאפיינים שכל אחד מהם Bernoulli (.5).

יש שלל אלגוריתמי חיפוש יוריסטיים לבחירת התת קבוצה המיטבית: אלגוריתמים חמדנים, simulated annealing, אלגוריתמים גנטיים, Branch and bound וכ"ו. אלגוריתמים אלה מנסים למקסם קריטריונים שונים כגון: אינפורמציה הדדית, קורלציה, וכו'.

אלגוריתמים אינקרמנטליים מוסיפים מאפיין אחר מאפיין לפי גרסאות שונות של הכלל: מקסימום רלונטיות (קורלציה עם התגית הנכונה) מינימום יתירות (חוסר קורלציה עם המאפיינים האחרים). כעת נראה אלגוריתם סטנדרטי המשתמש בשיטה זו.

12.2 PCA (Principal Component Analysis)

בפרק זה ניגע בקצרה בנושא של הורדת מימדיות מידע. נתבונן בקבוצה של וקטורים רב-מימדיים $\{x_k \in \mathbb{R}^d\}_{k=1}^n$, כאשר המימד d גדול. האם ניתן לייצג וקטורים אלה (או את ה"מידע" הגלום בהם) על ידי וקטורים במימד נמוך יותר?

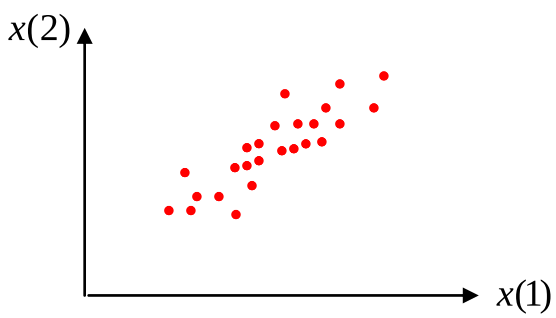
שאלה זו רלוונטית למספר רב של נושאים, וביניהם:

1. דחיסה: ייצוג קומפקטי של מידע (קובץ, תמונה).

2. סיווג: הורדת מאפיינים בעלי רלוונטיות נמוכה, לייעול תהליך הלמידה.

אנו נתמקד פה בשיטת PCA, המבוססת על הורדת מימדיות באמצעות התמרה (או הטלה) לינארית של וקטור המידע.

נתבונן ראשית באוסף נקודות במרחב הדו מימדי. נניח כי ברצוננו לתאר כל נקודה $x_k = (x_k(1), x_k(2))$ על ידי קואורדינטה אחת בלבד. כיצד נבחר אותה?



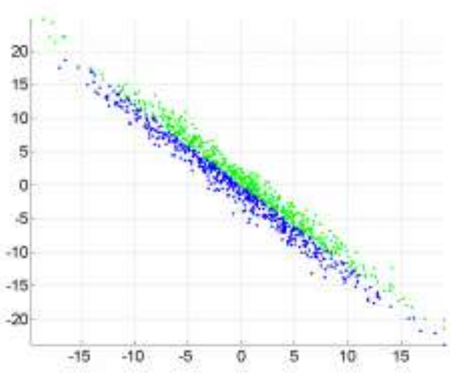
להלן קריטריונים אפשריים לבחירה כזו.

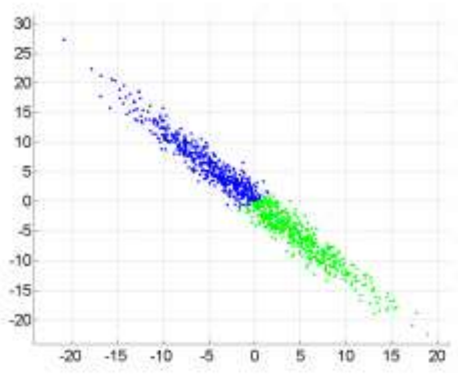
1. ווריאנס מכסימלי: נבחר כיוון במרחב X אשר לאורכו השונות (ווריאנס) של נקודות המידע היא מכסימלית.

2. שגיאת שחזור מינימלית: נבחר ייצוג (במימד מופחת) אשר מאפשר לשחזר את המידע המקורי בשגיאה (ריבועית) מינימלית.

נראה כי שני קריטריונים אלה מובילים לפיתרון זהה.

הערה חשובה: PCA אינו לוקח בחשבון את התיוג של הנקודות, אם ישנו,\ ולכן יתכן ויבחר כיוון שאינו אינפורמטיבי עבור התיוג.





האיור הימני מראה קבוצה שהכוון עם שונות מקסימלית הינו בעל קורלציה חזקה עם התיוג, בעוד באיור השמאלי המצב הפוך.

12.3 PCA עבור וקטור אקראי

נניח כי $\mathbf{x} \in \mathbb{R}^d$ הינו משתנה מקרי וקטורי בעל פילוג נתון, מטריצת קווריאנס Σ, וממוצע אפס. (אם הממוצע שונה מאפס נתבונן במשתנה הממורכז $\mathbf{x} - E(\mathbf{x})$).

תזכורת מאלגברה לינארית:

יהיו $\{\lambda_1,\ldots,\lambda_d\}$ הערכים העצמיים (ע"ע) של Σ, עם וקטורים עצמיים (ו"ע) מתאימים $\{v_1,\ldots,v_d\}$, כלומר $\Sigma v_j = \lambda_j v_j$. מכיוון ש-$\Sigma$ מטריצה סימטרית, ידוע כי:

1. כל הע"ע שלה <u>ממשיים</u> (ולמעשה אי-שליליים עקב $\Sigma \geq 0$).
2. קיימים d ו"ע בלתי תלויים. ו"ע המתאימים לע"ע שונים הם בהכרח אורטוגונליים ($\lambda_i \neq \lambda_j \;\Rightarrow\; v_i^T v_j = 0$). לפיכך ניתן לבחור d ו"ע אוטוגונליים $\{v_1,\ldots,v_d\}$.

נניח מעתה כי כל הערכים העצמיים מסודרים בסדר יורד: $\lambda_1 \geq \lambda_2 \geq \ldots$, וכי נבחרו ו"ע $\{v_1,\ldots,v_d\}$ אורטונורמליים (אוטוגונליים ומנורמלים, $\| v_j \| = 1$).

נציין גם כי ניתן לפרק את Σ באופן הבא (הייצוג הספקטרלי של מטריצה סימטרית):

$$\Sigma = \sum_{j=1}^{d} \lambda_j v_j v_j^T$$

<u>הגדרה</u>: **הכיוון העיקרי הראשון** של $\mathbf{x}$ הינו וקטור היחידה $w_1 \in \mathbb{R}^d$ אשר מביא למקסימום את הווריאנס של היטל $\mathbf{x}$ בכיוון w_1:

$$\max_{w:\|w\|=1} E(w^T \mathbf{x})^2$$

<u>טענה 1</u>: $w_1 = v_1$. כלומר, הכיוון העיקרי הראשון הינו הווקטור העצמי של Σ המתאים לערך העצמי הגדול ביותר.

<u>הוכחה</u>: $E(w^T \mathbf{x})^2 = w^T E(\mathbf{x}\mathbf{x}^T) w = w^T \Sigma w$. נפתור את בעיית האופטימיזציה תחת האילוץ $\| w \| = 1$ (אקויוולנטית, $w^T w = 1$) בעזרת כופל לגרנז' (λ). פונקציית הלגרנז'יאן היא:

$$L(w,\lambda) = w^T \Sigma w + \lambda(1 - w^T w)$$

על ידי גזירה:

$$2(\Sigma w - \lambda w) = 0 \quad \Rightarrow \quad \Sigma w = \lambda w$$

מכאן כי w הינו ו"ע של Σ עם ע"ע λ, דהיינו $\lambda \in \{\lambda_1, \ldots, \lambda_d\}$. כמו כן

$$w^T \Sigma w = w^T (\lambda w) = \lambda$$

מכאן שהערך המכסימלי של $w^T \Sigma w$ הינו $\lambda = \max\{\lambda_1, \ldots, \lambda_d\} = \lambda_1$, וערך זה מתקבל עבור הו"ע המתאים $w = v_1$. □

<u>הגדרה</u>: **הכיוון העיקרי ה-** m של $\mathbf{x}$ הינו וקטור היחידה $w_m \in \mathbb{R}^d$ אשר מביא למקסימום את הווריאנס, $E(w_m^T \mathbf{x})^2$ מבין כל הוקטורים המאונכים ל- $\{v_1, \ldots, v_{m-1}\}$.

<u>טענה 2</u>: $w_m = v_m$. כלומר, הכיוון העיקרי ה-m הינו הווקטור העצמי של Σ המתאים לערך העצמי λ_m (כאשר $\lambda_1 \geq \lambda_2 \geq \ldots$).

<u>הוכחה</u>: נתחיל עם $m = 2$. כזכור עלינו להביא למכסימום את $E(w_2^T \mathbf{x})^2$. כיוון ש-w_2 מאונך ל- v_1 לפי הגדרה, הרי

$$w_2^T \mathbf{x} = w_2^T \tilde{\mathbf{x}}, \quad \text{where} \quad \tilde{\mathbf{x}} = \mathbf{x} - v_1\left(v_1 \cdot \mathbf{x}\right)$$

ולכן ניתן אקוויולנטית להביא למכסימום את $E(w_2^T \tilde{\mathbf{x}})^2$.

נשים לב כי $E(\tilde{\mathbf{x}}) = E(\mathbf{x}) = 0$ וכן כי

$$\begin{aligned}
&E(\tilde{\mathbf{x}}\tilde{\mathbf{x}}^T) \\
&= E\left(\left[\mathbf{x} - v_1\left(v_1 \cdot \mathbf{x}\right)\right]\left[\mathbf{x} - v_1\left(v_1 \cdot \mathbf{x}\right)\right]^T\right) \\
&= E\left(\mathbf{x}\mathbf{x}^T - \mathbf{x}\mathbf{x}^T v_1 v_1^T - v_1 v_1^T \mathbf{x}\mathbf{x}^T + v_1 v_1^T \mathbf{x}\mathbf{x}^T v_1^T v_1^T\right) \\
&= \Sigma - \Sigma v_1 v_1^T - v_1 v_1^T \Sigma + v_1 v_1^T \Sigma v_1^T v_1^T \\
&= \Sigma - \lambda_1 v_1 v_1^T - \lambda_1 v_1 v_1^T + \lambda_1 v_1 v_1^T \\
&= \Sigma - \lambda_1 v_1 v_1^T
\end{aligned}$$

ולכן מטריצת הקווריאנס $\tilde{\Sigma}$ של $\tilde{\mathbf{x}}$ היא בעלת ו"ע $\{v_1,\ldots,v_d\}$ (ללא שינוי) ועם ע"ע מתאימים $\{0,\lambda_2,\ldots,\lambda_d\}$. עתה λ_2 הוא הערך העצמי המכסימלי, ולפי הטענה הקודמת $w_2 = v_2$ הוא הוקטור המביא למכסימום את $E(w_2^T\tilde{\mathbf{x}})^2$. וקטור זה מאונך ל- v_1 כנדרש.

עבור $m > 2$ ניתן להפעיל טיעון זהה

לגבי $\tilde{\mathbf{x}} = \mathbf{x} - v_1(v_1 \cdot \mathbf{x}) - \ldots - v_m(v_m \cdot \mathbf{x})$. □

<u>הערה 1</u>: במקום ההגדרה האיטרטיבית שבה השתמשנו עבור הכיוונים העיקריים, ניתן לאפיין כיוונים אלה באמצעות הטלות על תת-מרחב. יהיו $\{w_1,\ldots,w_m\}$ וקטורים אורטונורמליים כלשהם, אשר המהווים בסיס לתת-מרחב $S_m \subset \mathbb{R}^d$. הטלה של $\mathbf{x}$ על תת-מרחב זה נותנת:

$$\tilde{\mathbf{x}} \equiv \langle \mathbf{x}, w_1 \rangle w_1 + \ldots + \langle \mathbf{x}, w_m \rangle w_m$$

כאשר $\langle \mathbf{x}, w \rangle = \mathbf{x}^T w$ הינה המכפלה הפנימית. ניתן להראות (באופן דומה להוכחות דלעיל) כי $\tilde{\mathbf{x}}$ הינו בעל וריאנס מכסימלי כאשר $\{w_1,\ldots,w_m\} = \{v_1,\ldots,v_m\}$.

<u>הערה 2</u>: לאחר זיהוי הכיוונים העיקריים $\{v_1,\ldots,v_m\}$, הוקטור $\mathbf{x}$ יותמר לוקטור $\mathbf{z} = (\langle \mathbf{x}, v_1 \rangle, \ldots, \langle \mathbf{x}, v_m \rangle)^T$. ניתן לראות התמרה זו כתהליך דו-שלבי: א. מעבר לבסיס אלטרנטיבי $\{v_1,\ldots,v_d\}$ עם קואורדינטות $\langle \mathbf{x}, v_1 \rangle, \ldots, \langle \mathbf{x}, v_d \rangle$, ב. לקיחת m הקואורדינטות הראשונות.

<u>הערה 3</u>: מאורטוגונליות הוקטורים העצמיים $\{v_1,\ldots,v_d\}$ נובע חוסר-קורלציה של הקואורדינטות $\mathbf{z}_j = \langle \mathbf{x}, v_j \rangle$:

$$(i \neq j) \quad E(\mathbf{z}_i \mathbf{z}_j) = E(\langle \mathbf{x}, v_i \rangle \langle \mathbf{x}, v_j \rangle) = E(v_i^T \mathbf{x}\mathbf{x}^T v_j) = v_i^T \Sigma v_j = v_i^T \lambda_j v_j = 0 .$$

מינימיזציה של שגיאת השחזור: נראה עתה כי ההטלה על הכיוונים העיקריים מביאה למינימום את "שגיאת השחזור". נגדיר ייצוג של $\mathbf{x}$ במימד מופחת על ידי טרנספורמציה לינארית $\mathbf{z} = A\mathbf{x}$, כאשר $\mathbf{z}$ וקטור במימד $m<d$, ואילו A מטריצה במימד $m\times d$. נגדיר גם טרנספורמציית שחזור $\hat{\mathbf{x}} = B\mathbf{z}$, כאשר B מטריצה $d\times m$. מטרתנו לבחור את A,B אשר מביאים למינימום את שגיאת השחזור הריבועית:

$$J_m = E(\|\mathbf{x}-\hat{\mathbf{x}}\|^2) \quad \to \quad \min, \quad where\ \ \hat{\mathbf{x}} = BA\mathbf{x}$$

סימון מפורש יותר: נסמן ב- a_j את שורות A, כלומר $A^T = [a_1,\ldots,a_m]$. אזי

$$\mathbf{z} = A\mathbf{x} = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix}\mathbf{x} = \begin{bmatrix} a_1^T\mathbf{x} \\ \vdots \\ a_m^T\mathbf{x} \end{bmatrix} \equiv \begin{bmatrix} \langle a_1,\mathbf{x}\rangle \\ \vdots \\ \langle a_m,\mathbf{x}\rangle \end{bmatrix}$$

באופן דומה, אם נסמן $B=[b_1,\ldots,b_m]$, אזי

$$\hat{\mathbf{x}} = BA\mathbf{x} = [b_1,\ldots,b_m]A\mathbf{x} = \sum_{j=1}^{m}\left\langle a_j,\mathbf{x}\right\rangle b_j$$

<u>**טענה 3**</u>: שגיאת השחזור הריבועית הינה מינימאלית עבור $B = A^T = \left[v_1,\ldots,v_m\right]$.

<u>הוכחה</u>: נניח ללא הגבלת הכלליות כי עמודות B הינן אורטונורמליות [אחרת "נלבין" את B, כלומר נגדיר $\tilde{B} = BC$ כאשר עמודות $\tilde{B}$ מהוות בסיס אורטונורמלי לעמודות B (תהליך גרהם-שמידט), ובהתאם להגדיר $\tilde{A} = C^{-1}A$ כך ש- $BA = \tilde{B}\tilde{A}$]. נשים לב כי

$$\hat{\mathbf{x}} = BA\mathbf{x} = \sum_{j=1}^{m}\left\langle a_j,\mathbf{x}\right\rangle b_j$$

נבטא עתה גם את $\mathbf{x}$ באמצעות עמודות B. לשם כך נשלים ראשית את $[b_1,\ldots,b_m]$ לבסיס אורטונורמלי שלם של $\mathbb{R}^d$ ע"י וקטורים מתאימים $[b_{m+1},\ldots,b_d]$. אזי

$$\mathbf{x} = \sum_{j=1}^{d}\left\langle b_j,\mathbf{x}\right\rangle b_j$$

$$\hat{\mathbf{x}}-\mathbf{x}=\sum_{j=1}^{m}\left\langle a_j-b_j,\mathbf{x}\right\rangle b_j-\sum_{j=m+1}^{d}\left\langle b_j,\mathbf{x}\right\rangle b_j$$

ועקב אורטונורמליות הוקטורים $\{b_j\}$ נקבל

$$\|\hat{\mathbf{x}}-\mathbf{x}\|^2=\sum_{j=1}^{m}\left\langle a_j-b_j,\mathbf{x}\right\rangle^2+\sum_{j=m+1}^{d}\left\langle b_j,\mathbf{x}\right\rangle^2$$

כיוון שכל האיברים חיוביים, ברור כי המינימום של $E(\|\mathbf{x}-\hat{\mathbf{x}}\|^2)$ מתקבל עבור $a_j=b_j$, $j=1,\ldots,m$.

עדיין עלינו לבחור את הוקטורים $\{b_j\}$. עבור הבחירה (האופטימאלית) הנ"ל של $\{a_j\}$ נקבל

$$\|\hat{\mathbf{x}}-\mathbf{x}\|^2=\sum_{j=m+1}^{d}\left\langle b_j,\mathbf{x}\right\rangle^2=\cdots=\|\mathbf{x}\|^2-\sum_{j=1}^{m}\left\langle b_j,\mathbf{x}\right\rangle^2$$

מכאן כי מינימיזציה של $E(\|\mathbf{x}-\hat{\mathbf{x}}\|^2)$ שקולה למכסימיזציה של איבר הסכום האחרון, אולם לפי הערה 1 לעיל מכסימום זה מתקבל עבור $b_j=v_j$, $j=1,\ldots,m$. □

12.4 PCA עבור אוסף נקודות

נחזור לבעייה המקורית שבה נתון אוסף נקודות $\{x_k\in\mathbb{R}^d\}_{k=1}^n$.

פעולה מקדימה: מרכוז הדגימות, כלומר $x_k\leftarrow x_k-\bar{\mu}_n$, כאשר $\bar{\mu}_n=\frac{1}{n}\sum_{k=1}^{n}x_k$. לאחר המירכוז נקבל $\hat{x}_n=0$.

נגדיר עתה כיוונים עיקריים ביחס לאוסף זה. ההגדרות דומות למקרה הקודם, פרט לכך שהתוחלת $E(w^T\mathbf{x})^2$ תוחלף בסכום $S_n(w)=\sum_{k=1}^{n}(w^Tx_k)^2$. נשים לב להקבלה הבאה:

$$E(w^T\mathbf{x})^2=w^T\Sigma w$$

$$\frac{1}{n}S_n(w)=w^T\left(\frac{1}{n}\sum_{k=1}^{n}x_kx_k^T\right)w\triangleq w^T\hat{\Sigma}_nw$$

<u>הגדרה</u>: **הכיוון העיקרי הראשון** של $\{x_k\}_{k=1}^n$ הינו וקטור היחידה $w_1 \in \mathbb{R}^d$ אשר מביא למקסימום את הסכום $S_n(w_1)$. **הכיוון העיקרי ה-** m של $\{x_k\}_{k=1}^n$ הינו וקטור היחידה $w_m \in \mathbb{R}^d$ אשר מביא למקסימום את הסכום $S_n(w_m)$, מבין כל הוקטורים המאונכים ל- $\{v_1,\ldots,v_{m-1}\}$.

<u>טענה 4</u>: $w_m = v_m$, כאשר v_m הינו הווקטור העצמי של $\hat{\Sigma}_n$ המתאים לערך העצמי λ_m של $\hat{\Sigma}_n$ (כאשר $\lambda_1 \geq \lambda_2 \geq \ldots$).

הוכחת טענה זו נובעת מיידית מהטענה המקבילה עבור וקטור מקרי (טענה 2), לאור ההקבלה שצוינה לעיל.

תוצאה דומה תתקבל עבור מינימיזציה של "שגיאת השחזור". במקרה זה אנו מייצגים כל וקטור $x_k \in \mathbb{R}^d$ על ידי הוקטור $z_k = Ax_k \in \mathbb{R}^m$ ($m<d$), ומבצעים שחזור על ידי $\hat{x}_k = Bz_k$. שגיאת השחזור הריבועית הינה:

$$J_m = \sum_{k=1}^{n} \| x_k - \hat{x}_k \|^2$$

הבחירה האופטימאלית של A, B הינה ללא שינוי.

<u>הערות נוספות</u>:

1. שיטת ה-PCA רגישה לסקלה (scaling) של רכיבי הוקטורים המקוריים, לכן יש להקפיד על נרמול מתאים של הקואורדינטות לפני ביצוע החישוב.
2. קיימים אלגוריתמים איטרטיביים לחישוב הכיוונים העיקריים, אשר מתאימים גם לפעולה בזמן אמת. גישה מעניינת היא זיהוי הטרנספורמציה הלינארית $\mathbf{z} = A\mathbf{x}$ עם פרספטרון רב מימדי, ושימוש באלגוריתמי לימוד של רשתות נוירונים למציאת A.
3. שיטת ה-PCA שהוצגה היא גלובלית, כלומר הכיוונים העיקריים זהים בכל המרחב. במקרים רבים ניתן להשיג שיפור משמעותי על ידי גישה מקומית, כלומר שימוש בכיווני הטלה שונים באזורים שונים של המרחב. ניתן כמובן לקבוע מראש את אופן חלוקת מרחב לאזורים, אולם אלגוריתמים מתוחכמים יותר קובעים את החלוקה באופן אוטומטי מתוך המידע.

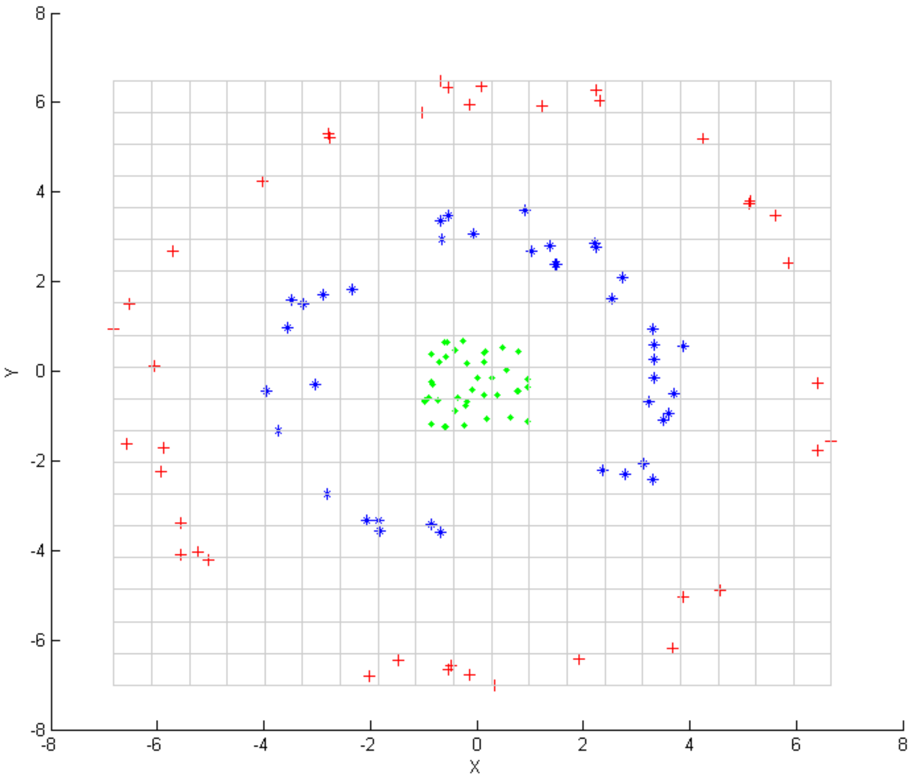
* הדגמות PCA ברשת:

http://diwww.epfl.ch/mantra/tutorial/english/pca/html/

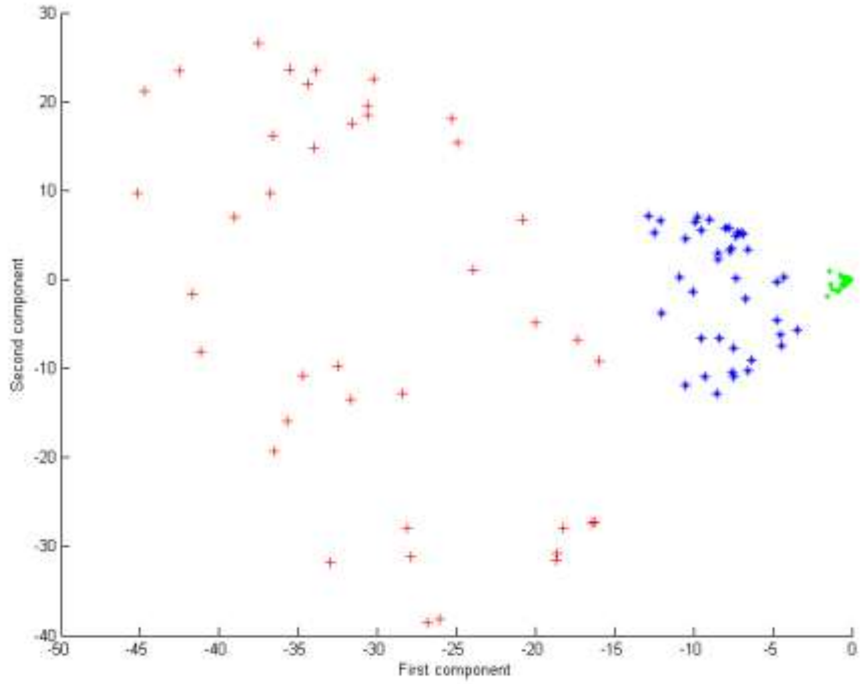
Kernel PCA 12.5

קל לראות שאלגוריתם ה- PCA מוגדר רק על ידי מכפלות פנימיות. לכן ניתן עבור kernel נתון לחשב את מטריצת הקווריאנס ולחפש את הוקטורים העצמיים במרחב הילברט המתאים.

למשל, עבור ה data הבא : (http://en.wikipedia.org/wiki/Kernel_PCA)



עבור גרעין פולינומיאלי : $k(x,y)=(1+x^T y)^2$ מתקבל :



ועבור kernel גאוסי מתקבל: