עיבוד נתונים בסטטיסטיקה מודרנית

פרוייקט סיום - קליסטור לתתי מרחבים

אלה פאליק 208882134

1 המודל

$B_1, ..., B_K \subset \mathbb{R}^p$ תתי מרחבים לינאריים, $\dim(B_i) = d$ לכל i, כך שממוצע הזוויות בין כל זוג הוא θ, כלומר $\frac{1}{\binom{K}{2}} \sum_{i<j} \theta_{ij}$.
נתייחס לתת מרחב B_i כמטריצה $M_{p \times d}$ עם עמודות אורתונורמליות שיוצרות בסיס ל B_i.
נדגום נקודות $x_1, ..., x_n \in \mathbb{R}^p$ מתתי המרחבים באופן הבא:

$$\begin{aligned} z_i &\sim U(\{1, ..., K\}) \\ w_i &\sim N(0, I_d) \\ x_i | z_i, w_i &\sim N\left(B_{z_i} \cdot w_i, \sigma^2 \cdot I_p\right) \end{aligned}$$

וכל השלישיות (z_i, w_i, x_i) הן בלתי תלויות. המטרה היא לשחזר את תתי המרחבים $B_1, ..., B_K$ ואת החלוקה לקלסטרים $z_1, ..., z_n$.

2 איכות השחזור

נסמן את תתי המרחבים המקוריים $B_1, ..., B_K$, את החלוקה המקורית לקלסטרים $z_1, ..., z_n$, את תתי המרחבים המשוחזרים $\hat{B}_1, ..., \hat{B}_K$ ואת שחזור החלוקה לקלסטרים $\hat{z}_1, ..., \hat{z}_n$. נשתמש בשני מדדים כדי לבדוק את איכות השחזור:

1. $C_{subspace}$: מדידת הזוויות בין המרחבים המקוריים והמשוחזרים

$$C_{subspace} = \max_{\pi \in S_K} \sum_{k=1}^{K} \cos\left(\angle\left(\hat{B}_{\pi(k)}, B_k\right)\right)^2$$

ככל שהערך יותר גדול, הזווית בין המרחבים קטנה יותר, כלומר המרחבים המשוחזרים יותר דומים למקוריים. נקבל מספרים בטווח $[0, K]$.

2. $C_{cluster}$: מדידת חלקיות החלוקה הנכונה לקלסטרים (נקבל מספרים בטווח $[0, 1]$)

$$C_{cluster} = \max_{\pi \in S_K} \frac{1}{n} \sum_{i=1}^{n} 1_{\{\pi(\hat{z}_i) = z_i\}}$$

חלק I

סימולציה ללא רעש

נתבונן בפרמטרים הבאים עבור המודל:

$$
\begin{aligned}
&n = 2^3, 2^4, ..., 2^{10} \\
&p = 2^4, 2^5, 2^6, 2^7 \\
&d = 2^{-1} \cdot p, 2^{-2} \cdot p, 2^{-3} \cdot p, 2^{-4} \cdot p \text{ for each } p \\
&K = 4 \text{ clusters} \\
&\theta = 10^{-2} \cdot \theta_{max}, 10^{-1} \cdot \theta_{max}, \theta_{max} \\
&\sigma = 0
\end{aligned}
$$

כאשר θ_{max} = הזווית הממוצעת בין מרחבים שנדגמים באופן אחיד.

1 סימולציה לכל (p, d)

נריץ שני אלגוריתמים עבור הפרמטרים האלו:

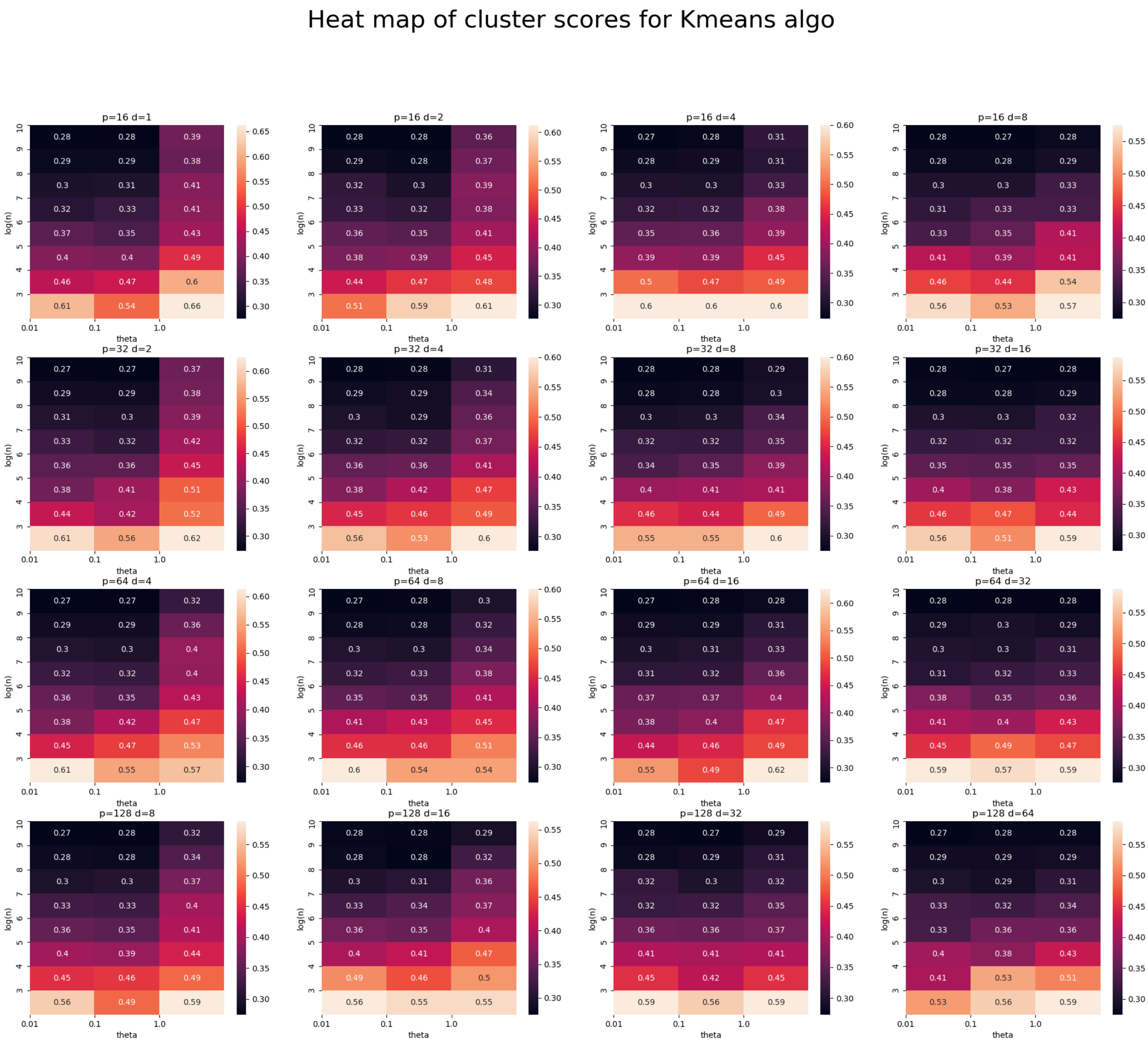
1. $Kmeans$ למציאת קלאסטרים ו PCA למציאת תתי המרחבים.
2. $EnSC$ למציאת קלאסטרים ו PCA למציאת תתי המרחבים.

לכל זוג ערכים (p, d), ניצור מפת חום שמראה את ביצועי האלגוריתמים הנ"ל כפונקציה של הזווית θ ומספר הדוגמאות n (מיצוע על 10 הרצות).

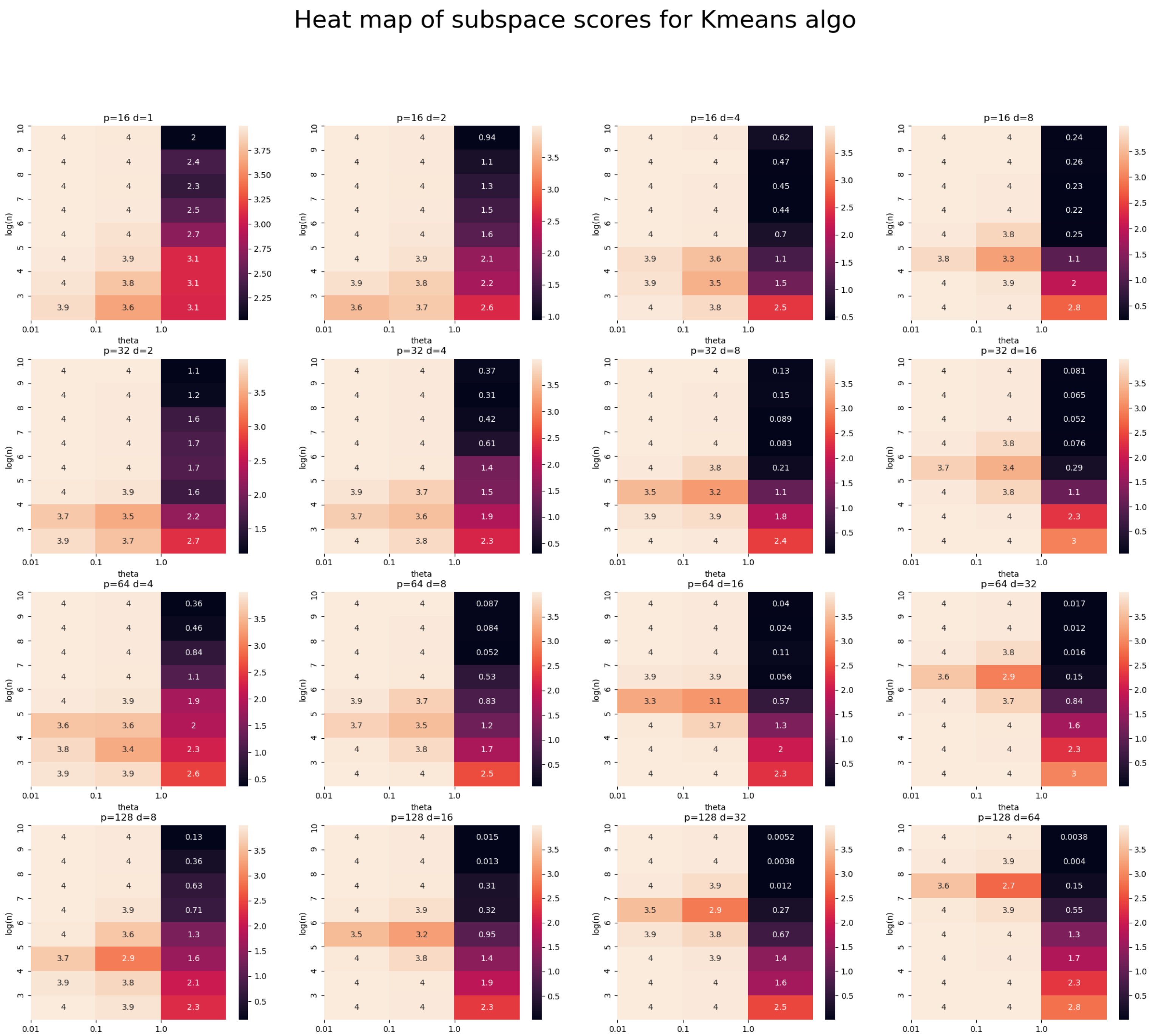
תוצאות

1.1 שחזור הקלאסטרים ע"י $Kmeans$

$Kmeans$ מחלק את הדאטה ל 4 קבוצות לפי מרחק ממרכזים בנורמה L_2, כאשר המרכזים מחושבים מחדש בכל איטרציה לפי המרכז של הקבוצה שהותאמה, עד שמקבלים התכנסות. על כן, לא נצפה שהחלוקה של $Kmeans$ תתאים למודל של תתי מרחבים, ואכן מדד $C_{cluster}$ נותן תוצאות מאוד נמוכות, לפחות עבור ערכי n גדולים (איור 1).

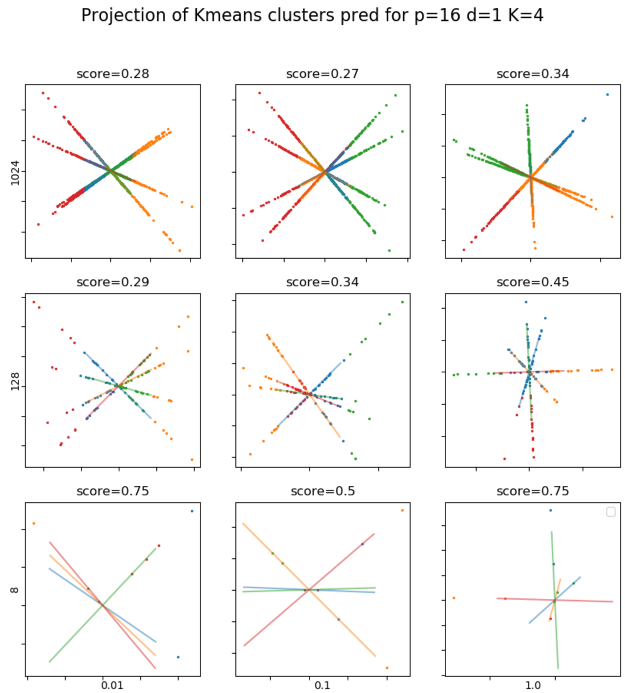


איור 1: מפת חום של ממוצע $C_{cluster}$ על 10 הרצות עבור אלגוריתם $Kmeans + PCA$.



איור 2: מפת חום של ממוצע $C_{subspace}$ על 10 הרצות עבור אלגוריתם $Kmeans + PCA$.

נרצה לחקור מעט את התוצאות: נתבונן בכמה דוגמאות מתוך הרצה של האלגוריתם עם פרמטרים $p = 16, d = 1, K = 4$ וערכי n ו θ שונים, לאחר הטלה באמצעות PCA ל $\mathbb{R}^2$ (איור 3).



איור 3: הטלה (בעזרת PCA עם שני רכיבים) של הדאטה, צבוע לפי חלוקה ל 4 קבוצות שהתקבלה מ $Kmeans$, עבור הרצה עם $p = 16, d = 1, K = 4$ וערכי $n = 8, 128, 1024$, $\theta = 0.01, 0.1, 1$. הישרים מייצגים את תתי המרחבים המקוריים.

נשים לב למספר תובנות:

1. עבור ערכי n קטנים, אין משמעות לזווית והחלוקה היא פחות או יותר אקראית. מכיוון שאנחנו לוקחים את הציון המקסימלי על כל הפרמוטציות נקבל ציון גבוה יחסית. נתבונן מעתה בערכי n גדולים יחסית.

2. הקבוצות של $Kmeans$ בד"כ מחלקות את המרחב לרבעים. על כן, אם היינו דוגמים את הדאטה באופן אחיד מקוביית היחידה, היינו מצפים שהאלגוריתם יקבל ציון דומה לחלוקה רנדומית לקבוצות. אך אנחנו דוגמים נקודות מתתי מרחבים B_{z_i} ע"י דגימה המיקום שלהם על הישר $w_i \sim N(0, I_d)$, כאשר החלוקה לתתי מרחבים z_i נדגמת באופן אחיד מ $\{1, ..., K\}$. על כן נצפה לראות שהנקודות מתרכזות סביב הראשית.

3. עבור זוויות קטנות, תתי המרחבים מאוד קרובים אחד לשני בקוביה לפי מרחק אוקלידי, וניתן לראות שבהטלה החלוקה היא בד"כ לפי ה"צד" של הראשית ואז לפי "קרוב" או "רחוק" מהראשית. נקבל בד"כ שני מרכזים מאוד קרובים לראשית (אחד מכל צד) ושניים רחוקים (אחד מכל צד), כי זו תהיה החלוקה הכי יציבה (הזזה קטנה של המרכז תשנה מעט מאוד את הקבוצה). ככל ש n גדל המרכזים מתקרבים יותר לראשית (כדי שהחלוקה תהיה יציבה, כי יש הרבה יותר נקודות קרוב לראשית), לכן החלוקה תהיה יותר אחידה ונתקרב לציון של חלוקה אחידה לקבוצות.

4. עבור זוויות גדולות, יכולת ההפרדה של $Kmeans$ גדלה, מכיוון שיש סיכוי גבוה יותר שמרכז מסויים "יתפוס" סביבו את כל הנקודות של (לפחות) תת מרחב אחד מצד אחד של הראשית, והנקודות על זרוע אחת לא יתחלקו לשני קבוצות כמו בזוויות קטנות יותר. לכן יקבל ציונים מעט יותר טובים. הציונים עדיין יהיו נמוכים מכיוון ש $Kmeans$ עדיין מחלק את הנקודות לפי ה"צד" של הראשית בו הן נמצאות ולא מתאים את עצמו למבנה של תת מרחב.

לסיכום, ניתן לומר שהאלגוריתם אינו מתאים למבנה של חלוקה לפי תתי מרחבים, ואינו משקף היטב את החלוקה של הדאטה. יש להעיר שההטלה באמצעות PCA ל $\mathbb{R}^2$ אינה משמרת זוויות, מכיוון שאנחנו לא בהכרח מטילים על המשטח שנפרש על ידי הוקטורים שנותנים את הזווית המקסימלית בכל זוג תתי מרחבים (בין היתר יתכן שזהו משטח אחר עבור כל זוג). אך מכיוון שלקחנו תתי מרחבים ממימד 1 (ישרים), בהינתן שהישרים מספיק רחוקים מלהיות מאונכים נקבל ייצוג יחסית טוב של הפיזור של הנקודות ביחס לישרים, אך שוב לא בהכרח של הזוויות בין הישרים.

1.2 שחזור תתי המרחבים ע"י $Kmeans + PCA$

ניתן לראות (איור 2) הבדל משמעותי בין התוצאות לפי הזווית בין תתי המרחבים. כאשר הזווית קטנה נקבל התאמה טובה (עד כדי מושלמת) בין תתי המרחבים המתקבלים מהפעלת PCA על הקבוצות שהתקבלו מ $Kmeans$, לבין תתי המרחבים המקוריים. את ההתאמה הטובה הזו ניתן לזקוף לזכות העובדה שבין תתי המרחבים זוויות מאוד קטנות ולכן הם מאוד קרובים אחד לשני בקוביה לפי מרחק אוקלידי. רוב הנקודות קרובות מאוד לראשית לכן גם הן מאוד קרובות אחת לשנייה לפי מרחק אוקלידי. סה"כ נקבל שרוב הנקודות מרוכזות בשתי גזרות ולכן נצפה לקבל מה PCA מרחבים שנמצאים בגזרות האלו. כלומר תתי המרחבים המשוחזרים יהיו קרובים לפי מרחק אוקלידי ולכן נצפה שגם בזווית.

1.3 שחזור הקלאסטרים ע"י $EnSC$

$EnSC$ עם הפרמטרים הדיפולטיים מחלק את הדאטה ל 4 קבוצות בשני שלבים:

1. בונים מטריצה A שמתארת מרחק בין נקודות הדאטה (Affinity matrix).
2. מפעילים על A אלגוריתם Spectral Clustering שמבצע הורדת מימדים ל $\mathbb{R}^K$ ואז מקלאסטר את הנקודות (בד"כ באמצעות $Kmeans$).

אלגוריתמים רבים משתמשים בגישה הזו, השוני ביניהם הוא בדרך בה הם מגדירים את המטריצה A. שיטה אחת היא לחשב את המטריצה על ידי self-represention של הדאטה, כלומר להציג כל נקודה באמצעות נקודות אחרות, כאשר באלגוריתם הספציפי הזה נרצה שהייצוג יהיה הכי דליל שאפשר. כלומר, נרצה לפתור את בעיית האופטימיזציה הבאה:

$$
\begin{aligned}
(1)\ \min_{c_j}\ & \|c_j\|_0 \\
s.t.\ & Xc_j = x_j \\
& c_{jj} = 0
\end{aligned}
$$

לכל נקודת דאטה j. רלקסציה של הבעיה הזו (שהיא NP קשה) תהיה

$$
\begin{aligned}
(2)\ \min_{c_j}\ & \|c_j\|_1 + \frac{\gamma}{2}\|x_j - Xc_j\|_2 \\
s.t.\ & c_{jj} = 0
\end{aligned}
$$

הרעיון של ייצוג עצמי דליל, כלומר לייצג כל נקודה באמצעות מספר מינימלי של נקודות, נובע מהעובדה שאם נקודה נמצאת בתת מרחב מסויים ממימד d, ניתן לייצג אותה באמצעות d נקודות במרחב, וסביר להניח שזו כמות מינימלית. המטרה היא ליצור A כך שלכל נקודה x_i (שורה i) רק d כניסות לא מתאפסות, שמתאימות כולן לנקודות שנמצאות באותו תת מרחב כמו x_i. אם נחשוב על A כמתארת גרף (אם כניסה אינה 0 קיימת צלע בין קודקודים, אחרת לא) הבעיה שעולה מהרעיון הזה היא שהגרף יהיה מאוד לא קשיר - over segmentation. כלומר, אם נפתור את הבעיה היטב, קודקודים יהיו מקושרים בצלע רק אם הנקודות המתאימות להם נמצאות באותו תת מרחב, אך יתכן שיהיו רכיבי קשירות רבים בתת גרף המתאר כל קלאסטר, וכאשר נעביר את המטריצה הזו לאלגוריתם Spectral Clustering ונבצע $Kmeans$, לא יהיה מבחינתו הבדל בין רכיבי קשירות שונים שמקורם מאותו קלאסטר וכאלו שמקורם בקלאסטרים שונים. כאשר הגרף יהיה over segmentated, החלוקה שמתקבלת יכולה להיות כמעט אקראית.
כדי לטפל בבעיה הזו מוסיפים במאמר רגולריזציה נוספת - מזעור של נורמה L_2 ($\frac{1}{2}\left\|\cdot\right\|_2^2$) של וקטור המקדמים, שידוע שמתעדף grouping, כלומר קשירות. סה"כ מקבלים את הבעיה (רגרסיית Elastic Net)

$$
(3)\ \min_{c_j} \lambda \left\|c_j\right\|_1 + \frac{(1-\lambda)}{2}\left\|c_j\right\|_2^2 + \frac{\gamma}{2}\left\|x_j - Xc_j\right\|_2
$$
$$
s.t.\ c_{jj} = 0
$$

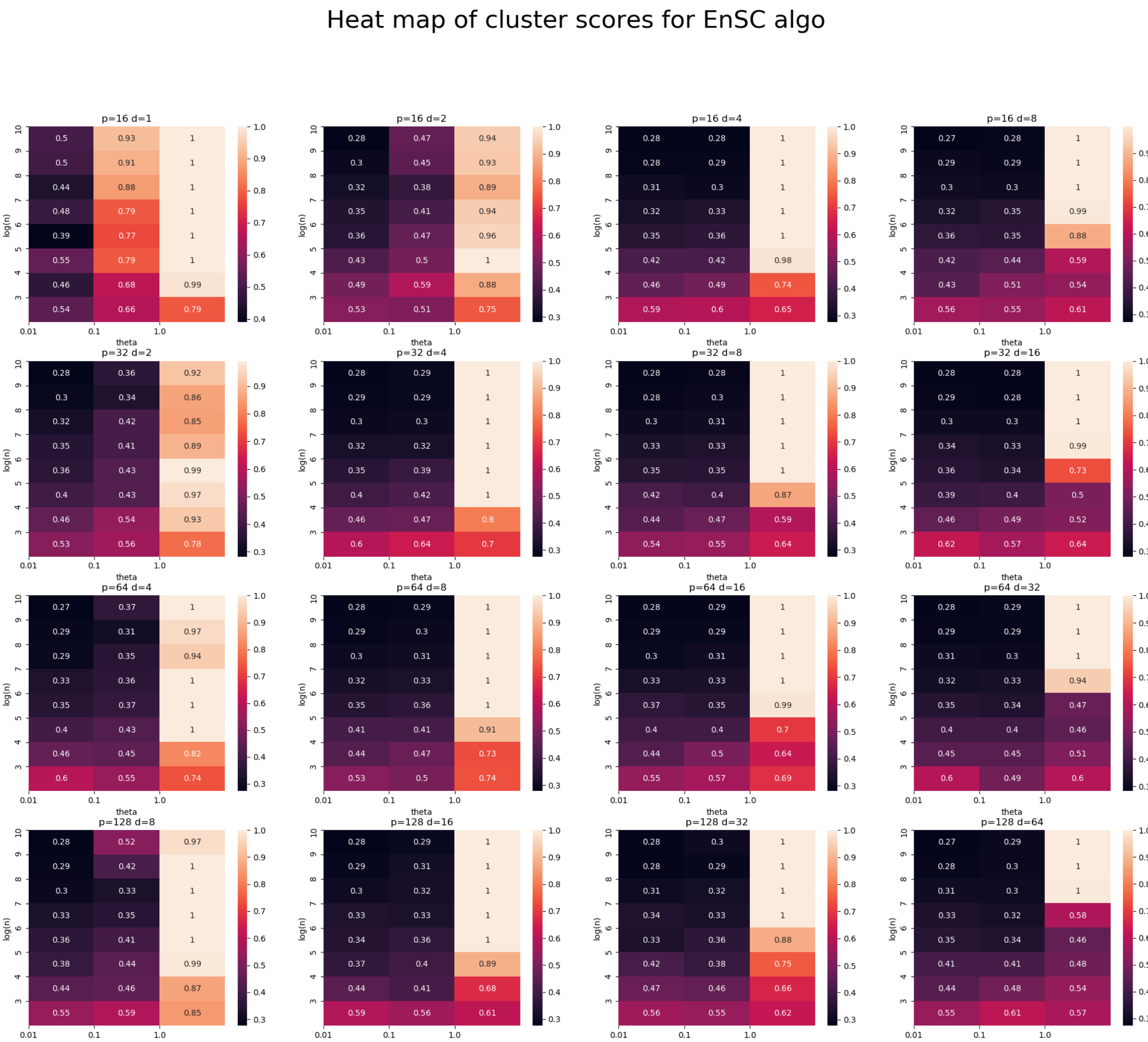
כאשר λ שולט בטריידאוף בין קשירות ודלילות.
בספריה בה השתמשתי יש מימוש של אלגוריתם Active set שפותר את הבעיה (3), אך לא השתמשתי בו מכיוון שהאלגוריתם שפותר את זה לא ניתן להתקנה על המחשב שלי, אלא באלגוריתם שפותר את (2) (רגרסיית Lasso). על כן חלק גדול מהבעיות בתוצאות ניתן לשייך לחוסר קשירות.
בבעיה זו ניתן לראות תוצאות שונות מאוד עבור זוויות שונות (איור 4):

1. אם n קטן מקבלים ציונים די דומים בלי קשר לזווית או לפרמטרים אחרים, מכיוון שאנחנו מסתכלים על כל הפרמוטציות האפשריות (באותו אופן כמו $Kmeans$). נתבונן מעתה בערכי n גדולים יחסית.

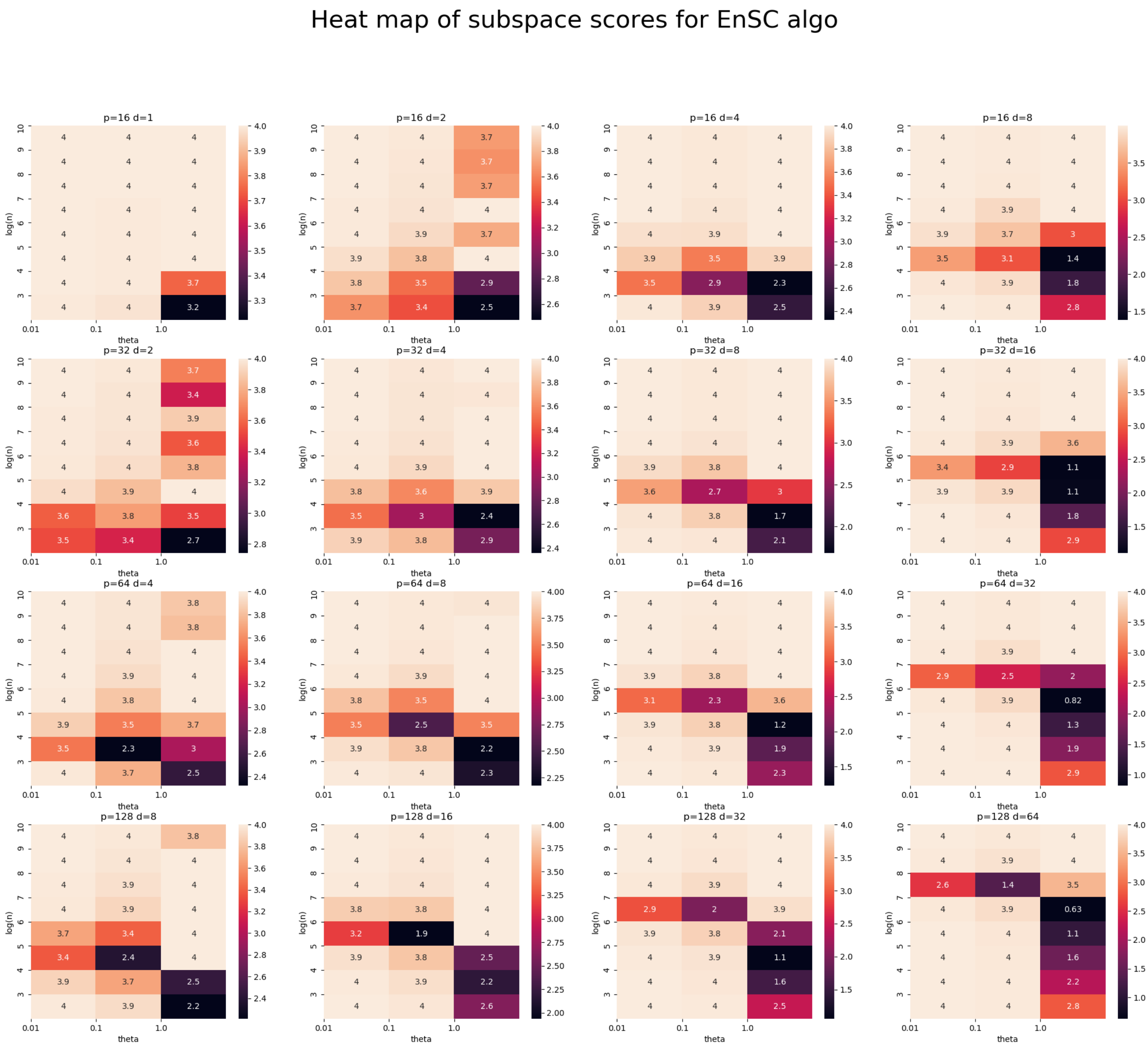
2. עבור זוויות קטנות ו d גדול, נקבל ציון קרוב מאוד לציון עבור חלוקה רנדומית ל 4 קבוצות. כפי שהסברתי למעלה, זה ככל הנראה נובע מחוסר קשירות של הגרף ש A יוצרת. כאשר מבצעים על המטריצה הזו הורדת מימדים ו $Kmeans$, אין הבדל בין רכיבי קשירות שונים שמקורם מאותו קלאסטר וכאלו שמקורם בקלאסטרים שונים ולכן החלוקה שמתקבלת היא כמעט אקראית. בנוסף, בגלל הזוויות הקטנות בין תתי המרחבים והעובדה שרוב הנקודות נדגמות סביב הראשית, יש סיכוי טוב שנקודות יתארו נקודות ממרחבים שונים, כך שאפילו רכיבי הקשירות שכן קיימים בגרף לא מתארים טוב את החלוקה לקלאסטרים. הציון מתקרב לציון של חלוקה רנדומית ככל ש n גדל, כי מאבדים את היתרון של להסתכל על כל הפרמוטציות האפשריות.

3. עבור זוויות קטנות ו d קטן (יחסית ל p), נקבל ציונים יותר טובים, מכיוון שבמקרים כאלו נקבל בד"כ שתיאור של נקודה יכלול מעט נקודות, לכן יותר סביר שהנקודות יגיעו מתוך התת מרחב שלה. למשל, במקרה ש $d=1$, נקבל שהייצוג של כל הנקודות על ישר יהיה פשוט כפל בסקלר (משקולת) של אחת הנקודות. ייצוג כזה הוא הכי פשוט ואכן מתקבל בד"כ מהאלגוריתם.

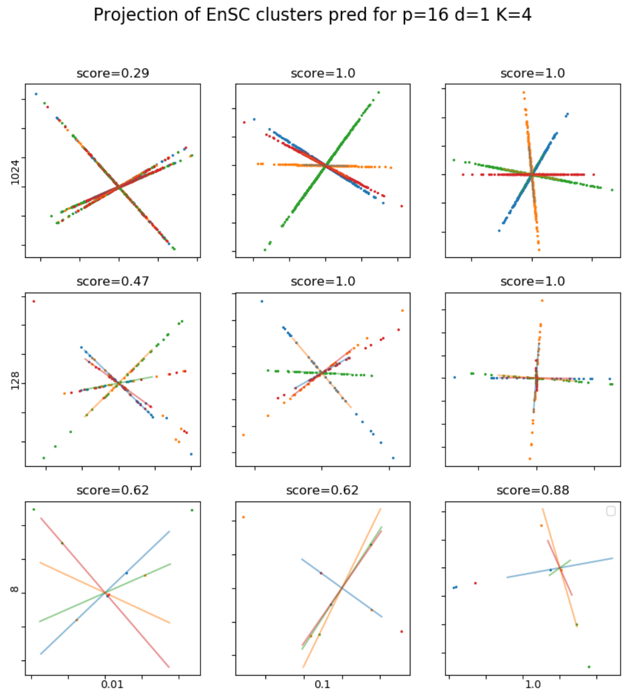
4. עבור זוויות גדולות ו d גדול, נקבל ציונים טובים מאוד, עד כדי 100% נכונות עבור ערכי n גדולים. כאן אנחנו לא רואים בהכרח קשירות בגרף, אבל מכיוון שהזוויות גדולות, הנקודות בד"כ רחוקות אחת מהשנייה, כך שרכיבי הקשירות שמתקבלים מכילים בסיכוי טוב רק נקודות מאותו תת מרחב. מכיוון ש d גדול אנחנו מקבלים יחסית הרבה נקודות שמשתתפות בייצוג, ולכן יש הרבה צלעות בגרף יחסית, לכן נצפה שלא יהיו המון רכיבי קשירות. כל זה נותן לנו הפרדה די טובה לרכיבי קשירות, ולכן ציון טוב בהפרדה.



איור 4: מפת חום של ממוצע $C_{cluster}$ על 10 הרצות עבור אלגוריתם $EnSC + PCA$.



איור 5: מפת חום של ממוצע $C_{subspace}$ על 10 הרצות עבור אלגוריתם $EnSC + PCA$.



איור 6: הטלה (בעזרת PCA עם שני רכיבים) של הדאטה, צבוע לפי החלוקה ל 4 קבוצות שהתקבלה מ $EnSC$, עבור הרצה עם $p=16, d=1, K=4$ וערכי $n=8,128,1024$, $\theta=0.01,0.1,1$. הישרים מייצגים את תתי המרחבים המקוריים.

1.4 שחזור תתי המרחבים ע"י $EnSC+PCA$

כאן (איור 5) אנחנו רואים את ההשפעה של קליסטור מוצלח. הנימוקים עבור $Kmeans$ תקפים גם כאן (עבור זוויות קטנות, ההתאמה טובה מכיוון שרוב הנקודות מרוכזות בשתי גזרות ולכן נצפה לקבל מה PCA מרחבים שנמצאים בגזרות האלו), רק שמכיוון שהקליסטור כל כך מדוייק עבור זוויות גדולות (בניגוד ל $Kmeans$), נקבל תוצאות מעולות גם שם (הנקודות מתחלקות בצורה מדוייקת לקלאסטרים ולכן ה PCA יכול לשחזר את תת המרחב, בהינתן שיש מספיק נקודות בקלאסטר).

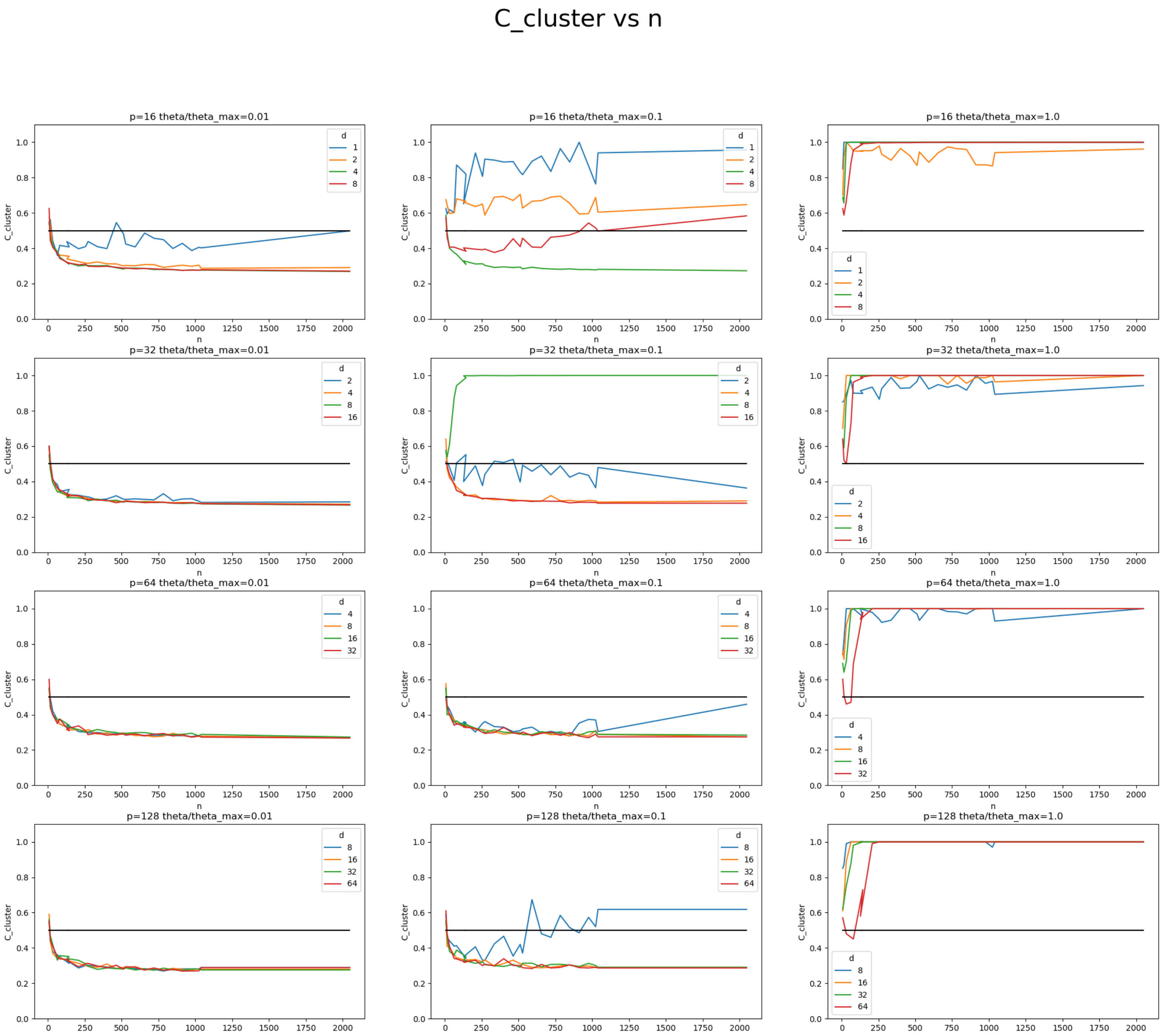
2 מציאת n_q ו $b\left(p,\frac{\theta}{\theta_{max}}\right)$

נגדיר

$$n_q = \text{sample size for which } C_{cluster}\left(p,d,\theta,n_q\right)=q$$

לכל קונפיגורציה (p,d,θ), נשערך את $n_{0.5}$ עבור האלגוריתם $EnSC+PCA$. לכל זוג ערכים $\left(p,\frac{\theta}{\theta_{max}}\right)$ ניצור עקום של $n_{0.5}$ כפונקציה של $\frac{d}{p}$, וננסה למצוא קבוע $b\left(p,\frac{\theta}{\theta_{max}}\right)$ כך שאם ננרמל את העקום של $n_{0.5}$ שקיבלנו עבור הערכים האלו, בקבוע הזה, נקבל שכל העקומים דומים.

תוצאות



איור 7: הערכה של $C_{cluster}$ (מיצוע על 10 הרצות) עבור הפעלת $EnSC$ על ערכים שונים של $16 \le n \le 2048$.

ראשית, נתחיל את החישובים מ $n=16$, מכיוון שמתחת לזה הדיוק של הקליסטור הוא תוצאה יחסית אקראית שלא משקפת את איכות האלגוריתם. נתבונן בהערכה של $C_{cluster}$ עבור ערכים שונים של $16 \leq n \leq 2048$ (איור 7). נזכור שאנחנו רוצים ליצור עקום $n_{0.5}$ עבור כל $\left(p, \frac{\theta}{\theta_{max}}\right)$ כפונקציה של $\frac{d}{p}$, ולשם כך נצטרך שלכל ערך $\frac{d}{p}$ עבור זוג מסויים $\left(p, \frac{\theta}{\theta_{max}}\right)$ נוכל למצוא את $n_{0.5}$.

1. עבור $\frac{\theta}{\theta_{max}}=1$, רואים שלכל ערכי p ו d, $C_{cluster}$ לא מגיע ל 0.5 עבור $n \geq 16$, ומתכנס מהר מאוד לדיוק 1 (כפי שהוסבר בסעיף א), לכן לא נוכל ליצור עקום של $n_{0.5}$.
2. עבור $\frac{\theta}{\theta_{max}}=0.01$, רואים שעבור $p=32,64,128$, לכל d מקבלים ש $C_{cluster}$ לא מגיע ל 0.5 עבור $n \geq 16$, ומתכנס מהר מאוד לדיוק 0.25 (כפי שהוסבר בסעיף א), לכן לא נוכל ליצור עקום של $n_{0.5}$. עבור $p=16$, הערכים $d=2,4,8$ נותנים תוצאה זהה, אך נראה שעבור $d=1$ ניתן יהיה למצוא את $n_{0.5}$ באיזור $n=500$, אך ערך d יחיד לא מספיק כדי ליצור עקום, כמו שהסברתי למעלה.
3. עבור $\frac{\theta}{\theta_{max}}=0.1$, נראה שיש עליה שיתכן שתגיע ל 0.5 בחלק מהמקרים. עם זאת, שוב, רואים ערכים בודדים של d עבורם יתכן שנוכל למצוא את $n_{0.5}$, אך זה לא מספיק ליצירת עקום.

סה"כ, לא הצלחתי למצוא זוג $\left(p, \frac{\theta}{\theta_{max}}\right)$ כך שאוכל ליצור עקום של $n_{0.5}$ כפונקציה של $\frac{d}{p}$. מהתוצאות שקיבלתי נראה שהאלגוריתם הנ"ל עובד בצורה מאוד דיכוטומית: מצד אחד תוצאות נהדרות עבור פרמטרים מסויימים (כל $\frac{\theta}{\theta_{max}}=1$ וחלק מ $\frac{\theta}{\theta_{max}}=0.1$), ומצד שני תוצאות גרועות על פרמטרים אחרים (רוב $\frac{\theta}{\theta_{max}}=0.01$ וחלק מ $\frac{\theta}{\theta_{max}}=0.1$), כאשר התוצאות באמצע הן מעטות ובד"כ קשורות ליחס $\frac{d}{p}$, לכן לא ניתן ליצור באמצעותן עקום.
עם זאת, חשוב לשים לב שהתוצאות שהבאתי כאן הן מיצוע על 10 ריצות. אם נריץ את האלגוריתם מספיק פעמים (עם כל בחירה של פרמטרים) נקבל מתישהו ערך עבור $n_{0.5}$. למרות זאת, מכיוון שבאופן ממוצע האלגוריתם לא מתנהג כך, הרגשתי שזה לא נכון להוציא דוגמאות לא מייצגות לריצות ולהסיק מהן מסקנות.
בכל מקרה, מימשתי את הפונקציות הדרושות כדי למצוא את $n_{0.5}$ ואת $\left(p, \frac{\theta}{\theta_{max}}\right)$ b. האסטרטגיה למציאת $n_{0.5}$ כוללת חיפוש בינארי, בעוד שמציאת b מורכבת מהשוואה של כל העקומים לעקום הראשון (או אחד אקראי) ו line_search כדי למצוא את ה b שנותן את הקרבה המקסימלית על פי נורמה L_1. בחרתי בנורמה L_1 מכיוון שכדי שעקומים יראו "דומים", חשבתי שהכי נכון לקרב כמה שיותר נקודות.

חלק II

אנליזה של דאטה אמיתי - fashion MNIST

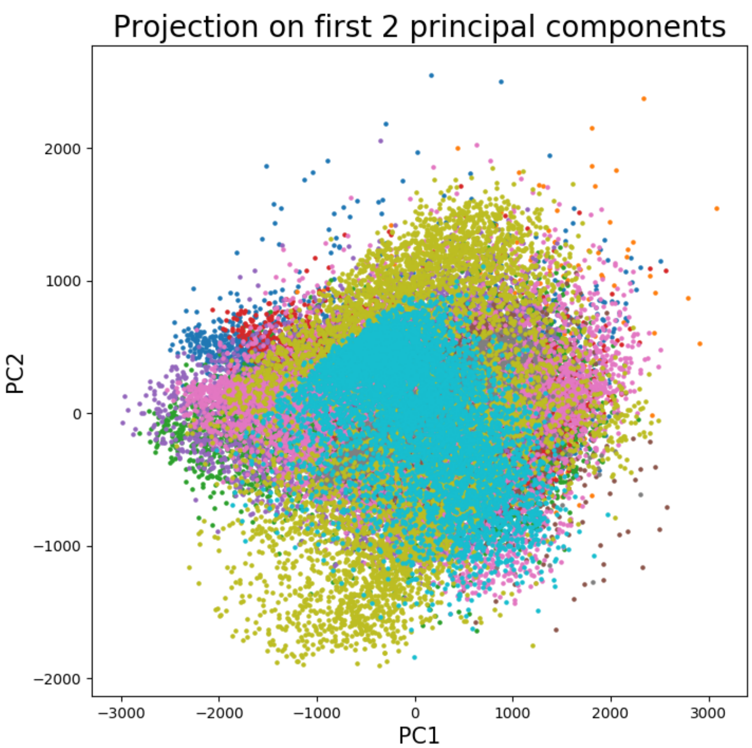
1 נרמול הדאטה

נוריד את הדאטה סט fashion MNIST, עם קבוצת אימון של $60,000$ תמונות בגודל 26×26 פיקסלים שמראות 10 סוגי בגדים. המטרה היא לקלסטר את התמונות באופן לא מפוקח. כדי לגרום להקשות על המשימה ננרמל את הנקודות: נחסר את הממוצע של כל מחלקה, כך שכל 10 מרכזי הקלאסטרים נמצאים בנקודת 0.

נריץ PCA על הדאטה סט ונוציא גרף של ההטלה של שני ה PC הראשונים, ונסמן את הנקודות של כל מחלקה בצבע אחר. האם המחלקות מופרדות היטב?

תוצאות

ניתן לראות (איור 8) שהמחלקות אינן מופרדות כאשר מנרמלים את המרכזים שלהם ל 0.

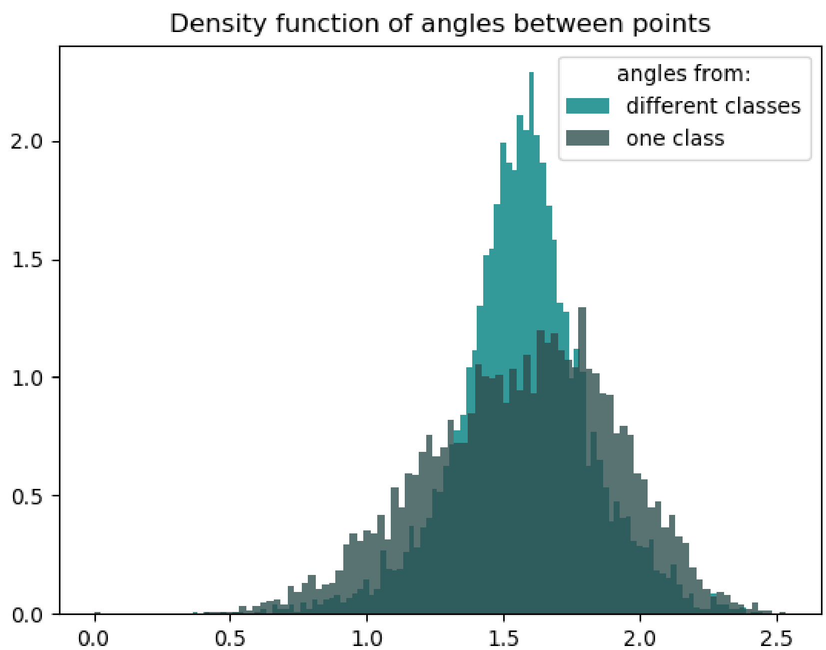


איור 8: הטלה של הדאטה סט fashion MNIST על שני הרכיבים העיקריים הראשונים, כאשר כל מחלקה צבועה בצבע אחר.

2 התפלגות הזוויות בין ובתוך קלאסטרים

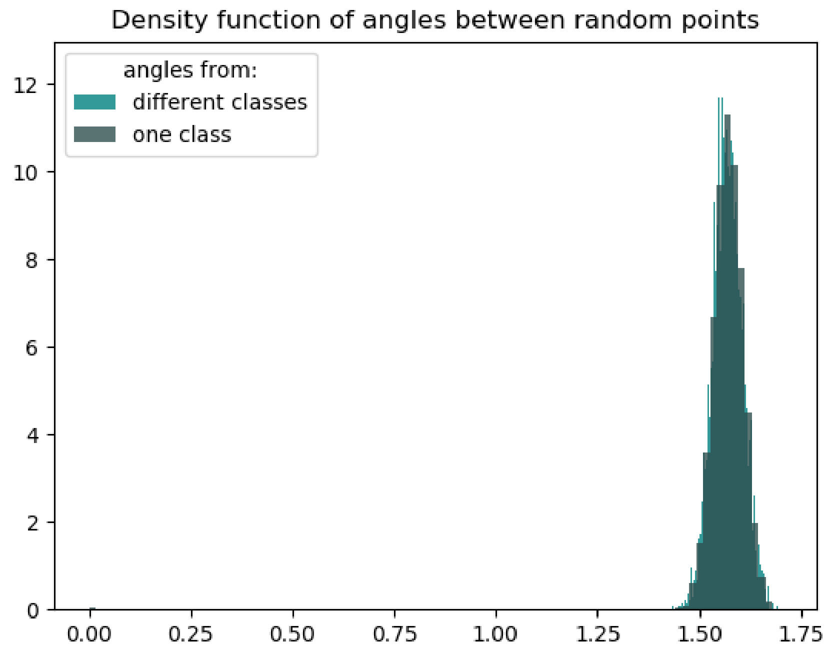
נדגום 5000 זוגות של נקודות ממחלקה אחת, ו 5000 זוגות של נקודות ממחלקות שונות.
נחשב את הזוויות בין זוגות של נקודות מאותה מחלקה ואת הזוויות בין זוגות של נקודות ממחלקות שונות.
נוציא גרפים של התפלגות הזווית בתוך קלאסטר ובין קלאסטרים.
האם אפשר לראות הבדל בין ההתפלגויות?

תוצאות



איור 9 : פונקציית צפיפות של הזוויות בין שתי נקודות מאותו קלאסטר ומקלאסטרים שונים, על פי שיערוך מ 5000 זוגות.

ניתן לראות (איור 9) שפונקציות הצפיפות של ההתפלגויות שתיהן נורמליות סביב $\frac{\pi}{2}$, כאשר הזוויות בין שתי נקודות מאותה מחלקה עם שונות גדולה יותר. כלומר הדוגמאות מקלאסטרים שונים הן בסבירות יחסית גבוהה קרובות ללהיות מאונכות, לפחות בסבירות יותר גבוהה מאשר שדוגמאות מאותה מחלקה יהיו קרובות להיות מאונכות. אפשר לראות את זה כסימן לכך שיתכן שקליסטור לפי תתי מרחבים יעשה עבודה טובה יחסית על הדאטה הזה, מכיוון שאנחנו יכולים לצפות שקיימים "צירים" שבהם תתי המרחבים שמשכנים כל קלאסטר נבדלים. לצורך המחשה, אם נדגום נקודות באופן אקראי מתוך $\mathbb{R}^{784}$ (המרחב שבו שוכנות הדוגמאות, ספציפית נדגום מתוך $[-255, ..., 255]^{784}$ כי שם הדוגמאות שלנו שוכנות לאחר נרמול), נחלק אותן באופן אחיד ל 10 מחלקות, ננרמל את המרכזים ונבדוק את התפלגות הזוויות בין ובתוך מחלקות, נקבל באופן צפוי שלא קיים הבדל בין שתי ההתפלגויות (איור 10).



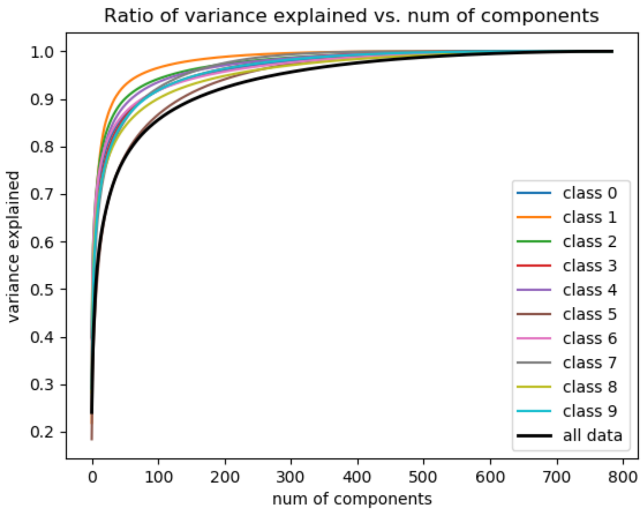
איור 10 : פונקציית צפיפות של זוויות בין שתי נקודות מאותו קלאסטר ומקלאסטרים שונים, על פי שיערוך מ 5000 זוגות של נקודות שנדגמו באופן אקראי עם תיוג אקראי.

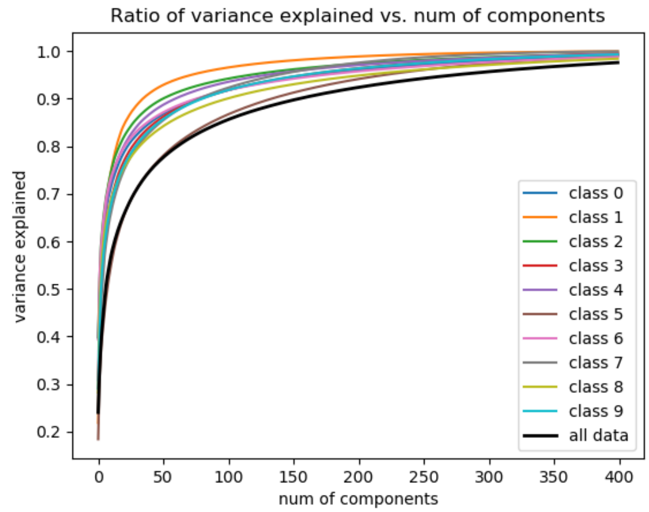
3 השונות המוסברת ע"י ה PC השונים

לכל מחלקה בנפרד, נריץ PCA ונוציא גרף של השונות שהוסברה כפונקציה של מספר ה PC, מסודר החל מהרכיב העיקרי הראשון על האחרון.
נחזור על התהליך על כל הדאטה סט.
איזה מספר רכיבים ניקח לאנליזה עתידית?

תוצאות

ניתן לראות מהגרפים (איור 11) שאחרי 300 רכיבים כבר אין כמעט עלייה מבחינת השונות המוסברת כאשר מוסיפים רכיבים. כבר ב 100 רכיבים ניתן לראות את הקפיצה הגדולה מבחינת רוב המחלקות, אבל ההתכנסות מתקבלת מאוחר יותר. מבחינת אנליזה עתידית, נרצה לקחת מספר רכיבים שנותן לנו את רוב האינפורמציה על הדאטה, אבל שאינו גדול מידי כדי לייתר את החלוקה לתתי מרחבים (הזוויות יהיו קטנות מאוד ולא יהיה ניתן להבדיל בין תתי המרחבים). אני בחרתי ב **250** רכיבים לאנליזה עתידית, תוך מחשבה שזו פשרה טובה בין מספר לא גדול מידי של רכיבים (יש סה"כ 784) לבין אחוז גבוה של שונות מוסברת (ניתן לראות שהגרף לא משתנה הרבה אחרי 250).





איור 11: אחוז השונות המוסברת לפי מספר הרכיבים העיקריים עליהם מטילים, עבור כל מחלקה בנפרד ועבור כל הדאטה יחד. בצד ימין נבחן יותר מקרוב את הגרף, עד 400 רכיבים.

4 הרצת והשוואת אלגוריתמים

נריץ על הדאטה סט את האלגוריתמים הבאים:

1. $Kmeans$ עם $K = 10$.
2. PCA עם מספר הרכיבים שבחרנו בסעיף הקודם (250), ואז $Kmeans$ עם $K = 10$ על ההטלה של הדאטה על ה PC.
3. $EnSC$ עם מספר הרכיבים שבחרנו בסעיף הקודם (250) כמימד תתי המרחבים ו $K = 10$.

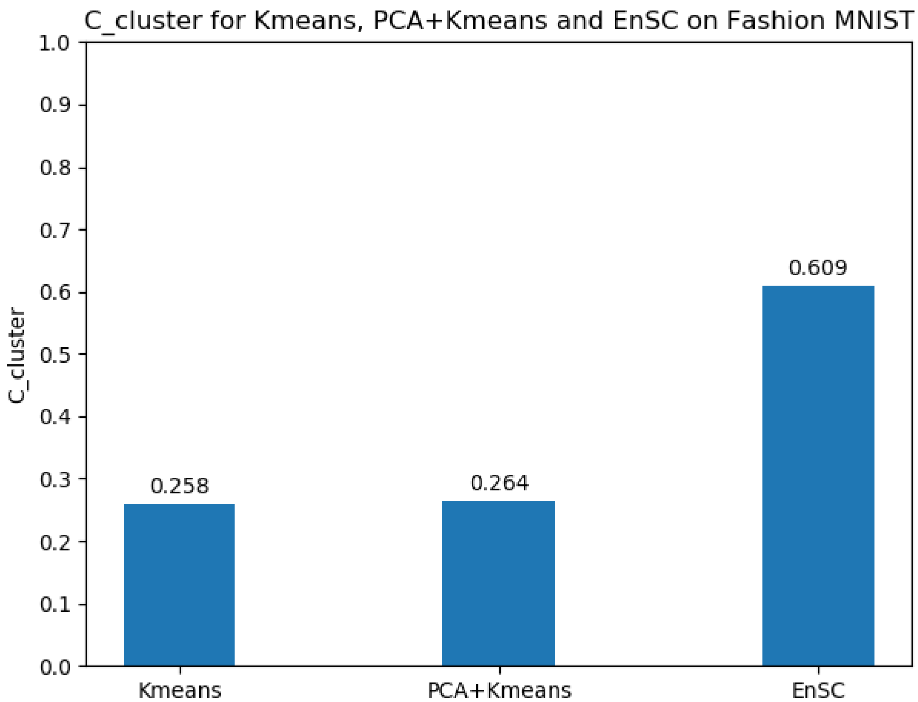
לכל אלגוריתם, נחשב את המדד $C_{cluster}$. איך ניתן להסביר את התוצאות?

תוצאות

נתבונן בתוצאות של כל האלגוריתמים (איור 12). ניתן לראות שהאלגוריתמים Kmeans, PCA+Kmeans נותנים תוצאות זהות באיזור 0.25, בעוד ש EnSC מקבל תוצאה גבוהה באופן משמעותי, 0.6. נרצה לחקור את התוצאות האלו:

1. יש הבדל מזערי בין Kmeans ו Kmeans+PCA: ניתן לראות שלהטלה על 250 הרכיבים העיקריים יש השפעה מעטה על הדיוק. מכך אפשר להסיק שהמימד של הדאטה הוא בעצם קרוב יותר ל 200 מאשר ל 784, וההורדה במימד לא מאבדת הרבה מידע. זה מאשר לנו שהבחירה ב 250 כמימד בסעיף ג' הייתה בחירה סבירה.
2. Kmeans נותן דיוק נמוך: קליסטור לפי Kmeans מתאים למקרים בהם כל קלאסטר מרוכז באיזור מסויים במרחב. לעומת זאת, כפי שראינו במקרה ללא רעש, Kmeans עובד רע מאוד (כמו חלוקה רנדומית) עבור דאטה שנדגם מתת מרחב לכל דלאסטר. מכיוון שאנחנו לא מקבלים כאן תוצאה דומה לחלוקה רנדומית ל 10 מחלקות (0.1), אלא קצת יותר טוב, נוכל לומר שהדאטה אינו שוכן באיחוד תתי מרחבים בדיוק ויש רעש במידול הזה, אך הוא גם לא מרוכז סביב נקודת מרכז באופן שיתן תוצאות טובות מ Kmeans.

3. EnSC נותן תוצאות משמעותית יותר טובות: בניגוד ל PCA, EnSC לא מניח שהדאטה משוכן בתת מרחב יחיד ממימד 250 אלא באיחוד של 10 תתי מרחבים ממימדים כלשהם. העלייה הגדולה בדיוק והמסקנות מההתפלגויות השונות של זוויות שקיבלנו (שקיימים "צירים" שמבדילים את תתי המרחבים בהם משכנים את הקלאסטרים) מספרים לנו שיש אפשרות סבירה שהדאטה מפוזר באופן כזה, כאשר העובדה שאנחנו עדיין מקבלים שגיאה גבוהה מצביעה על כך שהייצוג הזה לא מושלם ויש הרבה רעש בהצגה כזו של הדאטה.



איור 12: ציוני $C_{cluster}$ על כל דאטה סט Fashion MNIST עבור האלגוריתמים Kmeans, PCA+Kmeans, EnSC.

חלק III

הרחבה של האנליזה -

מציאת נושאים בפוסטים של חברי כנסת בפייסבוק

האנליזה הבאה תהיה אינטגרציה של הכלי בו התעסקנו בפרוייקט הזה - אלגוריתמים לקליסטור לפי תתי מרחבים, בפרוייקט שהכנתי בקורס אחר (סדנא ב NLP בעברית, סיכום של הפרוייקט מצורף). הפרוייקט כלל אנליזה של פוסטים (בעברית) של חברי כנסת בפייסבוק, כאשר שאלת המחקר הייתה ״מהם הנושאים שאפשר למצוא בפוסטים של חברי כנסת ומה העמדות שהם מבטאים לגביהם?״. השאלה הזו כוללת שתי משימות: Topic Analysis, כלומר אנליזה של הנושאים שניתן למצוא בפוסטים, ו Sentiment Analysis, כלומר עבור פוסטים מנושא מסויים, אילו עמדות מבטאים בהם.
המשימה הראשונה של זיהוי נושאים בפוסטים היא משימה של למידה לא מפוקחת, ולכן חשבתי שיהיה מעניין לראות את הביצועים של האלגוריתם בו השתמשתי בפרוייקט זה לעומת אלגוריתמים קלאסיים לינאריים של Topic Analysis, ומה אפשר ללמוד מכך על הדאטה.

1 דאטה

הדאטה מגיע מדאטה בייס של כיכר המדינה, אתר של הסדנא לידע ציבורי. האתר מכיל פוסטים של חברי כנסת בצירוף מידע כמו שם הכותב, שם מפלגתו, תאריך ושעת הפרסום ועוד, ויש אפשרות לחפש פוסטים על פי מילות חיפוש. על מנת לנתח את התוצאות של זיהוי נושאים ולקבל הערכה מספרית, נרצה לנתח דאטה סט שמכיל פוסטים שהנושאים בהם ידועים, בהנחה שכלי שישיג תוצאות טובות על דאטה סט כזה יתן חלוקה טובה גם בדאטה סט הכללי. התמקדתי בשלושה נושאים: ״בנימין נתניהו״, ״חוק הלאום״ ו ״דונאלד טראמפ״. אספתי ~ 100 פוסטים מכל אחד מהנושאים האלו באמצעות שימוש באפשרות באתר לחפש פוסטים לפי מילות חיפוש. מילות החיפוש היו ״נתניהו״, ״חוק הלאום״ ו ״טארמפ״ בהתאמה.
משהו שצריך לשים לב אליו הוא העובדה שפוסט יחיד כולל לעיתים קרובות יותר מנושא אחד, לכן חילקתי את הפוסטים למשפטים, בהנחה שמשפט מכיל לא יותר מנושא יחיד. זה גורם לכך שבדאטה סט יהיו משפטים משלושת הנושאים שהגדרנו, אבל גם הרבה משפטים מנושאים אחרים. תייגתי את המשפטים לפי שייכות לכל אחד מהנושאים או ״אחר״.

יאיר לפיד, כחול לבן, 7 בנובמבר 2018, 12:54 ['ה נצחון ה דמוקרטי ב ה מירוץ ל ה קונגרס ב ארהב יוצר לא מעט אתגרים ל ה ממשלה ה נוכחית', 'כבר יותר מ שנתיים ש כולם אומרים לנתניהו ש זה יקרה', 'אבל הוא חשב ש הוא יודע יותר טוב', 'קודם כל מפני ש זה יקשה על טראמפ להמשיך לעבוד', 'לא היה ל אנחנו ו אולי גם לא יהיה נשיא אמריקאי ידידותי יותר ל ישראל', 'ב ה שנתיים ה אחרונות הוא עשה סדרה של צעדים מבורכים', 'העברת ה שגרירות', 'ביטול הסכם ה גרעין ו החזרת ה סנקציות', 'ביטול ה תקציב ל ה אונרא', 'ה מלחמה ה גלויה ב ה אום', ...]

איור 13: דוגמא לתוצאת עיבוד חלק מפוסט שפורסם ע״י יאיר לפיד.

עיבוד הדאטה כולל ניקיון של המשפטים וחלוקה למורפמות (היחידות הלשוניות הקטנות ביותר הנושאות משמעות, מילה או חלק ממילה) באמצעות ספריית YAP. נקבל את הדאטה סט הבא:

	סה"כ	טראמפ	נתניהו	חוק הלאום	"אחר"
מספר פוסטים (לפי מילות חיפוש)	303	100	103	100	-
מספר משפטים (לפי מילות חיפוש)	2696	871	862	964	-
מספר משפטים (לפי תיוג נושא)	2696	314	442	348	1592

טבלה 1: הרכב הדאטה סט לאחר תיוג.

נשווה בין מספר אלגוריתמים קלאסיים של זיהוי נושאים לבין אלגוריתם קליסטור לפי תתי מרחבים, EnSC, פעם אחת בדאטה סט של 3 נושאים + נושא "אחר" ופעם אחת בדאטה סט של 3 נושאים בלבד.

2 שיטות

2.1 Embedding

ראשית נרצה ליצור ייצוג לדאטה. כל דוגמא מורכבת מאוסף המורפמות המופיעות במשפט, ונרצה לייצג אותה בעזרת וקטור. ישנם שני ייצוגים קלאסיים ב NLP:

Term Frequency: נייצג כל דוגמא כוקטור ב $\mathbb{R}^N$, כאשר N = מספר המורפמות במילון (כל המורפמות בכל המשפטים). נמספר את המורפמות במילון באופן אקראי. עבור דוגמא x_i עם אוסף המורפמות המופיעות במשפט ה i, נסמן

$$[x_i]_j = \begin{cases} 1 & \text{morpheme j appear in sentence i} \\ 0 & \text{else} \end{cases}$$

Tf-idf: שוב נייצג כל דוגמא כוקטור ב $\mathbb{R}^N$, אך בייצוג זה נשכלל את המשקל של המילה במשפט כך שיקח בחשבון כמה המילה נפוצה בקורפוס, כלומר כמה מידע היא מוסיפה על המשפט (למשל, אם מילה מופיעה בכל המשפטים נרצה שיהיה לה משקל נמוך, כי היא לא נותנת שום מידע נוסף על משפט ספציפי). יש דרכים רבות להגדיר את השכלול הזה.

2.2 אלגוריתמים קלאסיים לינאריים לזיהוי נושאים

נשתמש ב - 5 סוגי אלגוריתמים קלאסיים עם סוגי Embedding שונים (סה"כ 6 אלגוריתמים):
- SVD, Random SVD, PCA, Non-Negative Matrix Factorization (NMF) with TF vectorization,
- NMF with TF-IDF vectorization

כל האלגוריתמים הנ"ל הם אלגוריתמים שמבצעים matrix factorization. שם כללי לטכניקה הוא **(LSA) Latent Semantic Analysis**:

בהינתן מטריצה A המתארת קשר בין המשפטים בקורפוס למילים באוצר מילים (ייצוג מאלו שתיארנו בסעיף הקודם), נשים לב שהיא בדרך כלל מאוד דלילה, רועשת, ומפוזרת בצורה מיותרת על פני מימד גבוה מאוד. לכן נרצה לבצע הורדת מימדים כדי למצוא מספר נושאים שמייצגים היטב את הקשר בין מילים למסמכים.

הטכניקה הבסיסית היא Truncated SVD: נמצא פירוק לערכים סינגולריים (SVD) $A = USV$ ונוריד מימדים על ידי בחירה של K הערכים העצמיים הגדולים ביותר, ושמירה על K הוקטורים הראשונים מכל מטריצה (כאשר K = מספר הנושאים

שנקבל). כך נקבל את K המימדים הכי משמעותיים של המרחב ש A פורשת. לאחר הורדת מימדים, U תייצג את הקשר בין משפטים לנושאים, ו V תייצג את הקשר בין נושאים למילים באוצר מילים.
האלגוריתמים נבדלים בדרך שבה מפרקים את המטריצה A.

2.3 מודל לזיהוי 3 נושאים + נושא "אחר"

נזכור שאנחנו רוצים לבחון את הביצועים של EnSC מול האלגוריתמים שתיארנו למעלה, פעם אחת בדאטה סט של שלושת הנושאים בלבד ופעם אחת בדאטה סט של 3 נושאים + נושא "אחר". בפרוייקט (בקורס השני) מתוארים מספר מודלים לקליסטור ל 3 נושאים + נושא "אחר", אך כאן אתרכז במודל אחד שנתן את התוצאות האופטימליות: הרצה בשני שלבים.

שלב 1 - זיהוי נושא "אחר":

1. נריץ את האלגוריתמים עבור $K = 3$. לאחר מכן נשנה סיווג של חלק ממשפטים לפי רמת הביטחון בציון התאמה בין משפט לנושא:

<u>Certainty levels threshold:</u>

T = אחוז המסמכים עבורם נאפשר בדיקה של "נושא אחר".
$W_{d,i}$ = ציון התאמה בין משפט לנושא שהתקבל מהאלגוריתם.
עבור משפט d בקורפוס ונושא $i \in [K]$, נגדיר:

$$\text{certainty level}\,(d) = \sum_{i=1}^{K} W_{d,i}$$

$$\begin{aligned}
\text{certainty threshold}\,(d) &= \lambda \\
&\quad s.t. \text{for } T\% \text{ of docs, certainty level}\,(d) < \lambda \\
W_{d,K+1} &= \text{certainty score for doc } d \text{ to be in topic "other"} \\
&= \begin{cases} 1 - \text{certainty level}(d) & if \text{ certainty level}\,(d) < \lambda \\ 0 & else \end{cases}
\end{aligned}$$

אז נקבל מסווג:

$$\hat{f}\,(d) = \arg\max_{i \in [K+1]} W_{d,i}$$

נקבל סיווג של המשפטים ל 4 נושאים. T נקבע ל 30%.

2. ניקח את הנושא עם הכי הרבה סיווגים להיות נושא "אחר". עבור משפטים שסווגו כך נקבע את התווית להיות 0.

שלב 2 - זיהוי 3 הנושאים במשפטים שנותרו:

1. נריץ את האלגוריתמים עבור $K = 3$ על כל המשפטים שקיבלו סיווג אחר בשלב הקודם, נקבל סיווג שלהם ל 3 נושאים.

בסה"כ קיבלנו סיווג של כל המשפטים ל 4 נושאים.

האלגוריתם שהראה תוצאות אופטימליות בשלב 1 היה PCA. הציון מחושב על פי הדיוק (Accuracy) עבור משפטים שתוייגו "אחר", כאשר נורמל את הציון הזה במספר האפסים שתוייגו סה"כ כדי למנוע מצב שתיוג של כל המשפטים כנושא "אחר" יקבל ציון גבוה.
לכן נשווה בין האלגוריתמים על פי ביצועיהם בשלב 2, כאשר את שלב 1 מריצים עם PCA.

2.4 הערכת התוצאות

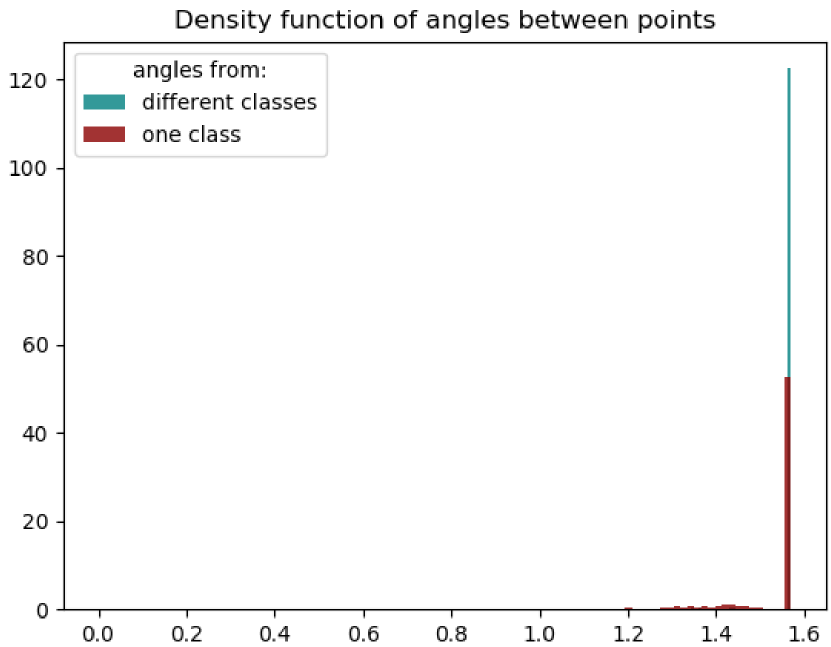
עבור ההשוואה בדאטה סט של שלושת הנושאים בלבד, נשווה את ממוצע ה Accuracy של האלגוריתמים על 10 הרצות. עבור הדאטה סט של 3 נושאים + נושא "אחר", נשווה מספר מדדים (ממוצע על 10 הרצות): Accuracy ו Accuracy עבור כל המשפטים שלא תוייגו "אחר" ידנית (נקרא לזה Accuracy משוכלל). המדד העיקרי הוא Accuracy כללי, אך מכיוון שיש הרבה יותר משפטים שמתוייגים "אחר" מאשר משפטים שמתוייגים עבור כל אחד מהנושאים, המדד השני יכול לתת תובנות לגבי כמה מהנושאים באמת תוייגו נכון. בנוסף נחשב ממוצע של שני המדדים.

3 תוצאות ומסקנות

SVD	Random SVD	PCA	NMF	NMF with TF-IDF	EnSC	מדד	דאטה סט
0.564	0.792	0.599	**0.811**	**0.817**	0.393	Accuracy	3 נושאים
0.270	0.603	0.594	0.642	0.663	0.174	Accuracy עבור שלב 2	3 נושאים +
0.555	0.624	0.530	0.643	0.650	0.419	Accuracy משוכלל עבור שלב 2	נושא "אחר"
0.412	0.614	0.562	**0.643**	**0.656**	0.296	ממוצע המדדים עבור שלב 2	עם אלגו 1 = PCA

טבלה 2: תוצאות עבור משימת Topic Analysis אחרי מיצוע על 10 הרצות עבור האלגוריתמים שבחרנו, על שני הדאטה סטס.

בטבלה 2 ניתן לראות ש NMF ו NMF with TF-IDF vectorization נותנים את התוצאות הכי טובות בשתי ההרצות, כאשר PCA מעט מאחוריהם אך כל השאר בפער משמעותי. ספציפית, EnSC מקבל תוצאות מאוד גרועות על הדאטה סטס האלו (חלוקה רנדומית ל 3 נושאים תיתן $\frac{1}{3}$ דיוק). אם נסתכל על פונק' הצפיפות של הזוויות בין ובתוך מחלקות (איור 14), נקבל שרוב הנקודות מאונכות זו לזו, אך בתוך הקלאסטרים כן יש פיזור יותר רחב.
בסך הכל, מכל האמור לעיל נוכל להסיק שהדאטה סטס האלו אינם מתאימים כלל למידול לפי איחוד תתי מרחבים.



איור 14 : פונקציית צפיפות של הזוויות בין שתי נקודות מאותו קלאסטר ומקלאסטרים שונים, על פי שיערוך מ 5000 זוגות.