סטטיסטיקה לעידן הביג דאטה

מבחן סיום

מגיש: נמרוד דביר 021991294

שאלה 1:

סעיף (a)

נתון לנו כי $\mathcal{M}_1$ הוא differentially private $\frac{\epsilon}{2}$, כלומר ל-x,y כך ש-$\|x-y\|_1 \leq 1$ מתקיים:

$e^{-\frac{\epsilon}{2}} \leq \frac{P(\mathcal{M}_1(x)=s)}{P(\mathcal{M}_1(y)=s)} \leq e^{\frac{\epsilon}{2}}$ בש-$S \subseteq Range(\mathcal{M}_1)$.

כמו כן נתון גם כי $\mathcal{M}_2$ הוא differentially private $\frac{\epsilon}{2}$ בהינתן $\mathcal{M}_1$, כלומר עבור x,y כך ש-$\|x-y\|_1 \leq 1$ מתקיים: $e^{-\frac{\epsilon}{2}} \leq \frac{P(\mathcal{M}_2(x)=r \mid \mathcal{M}_1(x)=s)}{P(\mathcal{M}_2(y)=r \mid \mathcal{M}_1(y)=s)} \leq e^{\frac{\epsilon}{2}}$.

נשתמש בחוק בייס להתפלגות מותנה: $P(A \cap B) = P(A|B) \cdot P(B)$ ובמונוטוניות של e ונקבל:

$$e^{-\frac{\epsilon}{2}} \cdot e^{-\frac{\epsilon}{2}} \leq \frac{P(\mathcal{M}_2(x) = r \mid \mathcal{M}_1(x) = s)}{P(\mathcal{M}_2(y) = r \mid \mathcal{M}_1(y) = s)} \cdot \frac{P(\mathcal{M}_1(x) = s)}{P(\mathcal{M}_1(y) = s)} \leq e^{\frac{\epsilon}{2}} \cdot e^{\frac{\epsilon}{2}}$$

ומכאן התוצאה המבוקשת שהאלגוריתם המשותף משחרר differentially private ϵ:

$$e^{-\epsilon} \leq \frac{P(\mathcal{M}_2(x) = r\,,\ \mathcal{M}_1(x) = s)}{P(\mathcal{M}_2(y) = r\,,\ \mathcal{M}_1(y) = s)} \leq e^{\epsilon}$$

סעיף (b)

I. הרגישות, בנוסף לפרמטר הפרטיות ϵ, היא חלק מהחישוב של ההתפלגות שמייצרת את הרעש שאנחנו מוסיפים לתוצאות האמיתיות ונדרש לחשבה כדי לדעת כמה רעש נדרש להוסיף למידע המשוחרר על מנת לשמור על הפרטיות ברמה שהתבקשנו (באמצעות ϵ).

II. האלגוריתמים נדרשים לבחור SNPs M מתוך M' קיימים ולשחרר סטטיסטים שלהם עם רעש. תיאור האלגוריתמים:

- אלגוריתם 1 תחילה מוסיף רעש לפלס (התלוי ב-M, ברגישות ובפרמטר בפרטיות), אז בוחר את M ה-SNPS עם הסטטיסטיים המורעשים הגבוהים ביותר ומשחרר אותם עם רעש חדש (שהוא חצי מהרעש שהשתמש כדי לבחור בהם).

- אלגוריתם 2 משתמש בשיטה מתוחכמת יותר לבחירת ה-M אותם מפרסם (שלבים 2-5): הוא נותן לכל SNP משקולת שהיא פונק' של הציון שלו, M, הרגישות ופרמטר הפרטיות. לאחר מכן מנרמל את כל המשקולות להסתברויות כך שככל של-SNP ציון גבוה יותר יש לו הסתברות גבוהה יותר להיבחר. אז האלג' בוחר SNP אחד, מאפס לו את ההסתברות, מנרמל את שאר ההסתברויות וכך חוזר עד ש-M נבחרו. לבסוף מוסיף רעש (באותה צורה שאלגוריתם 1 הוסיף) ומשחרר את הסטטיסטים המורעשים של M הנבחרים.
מכיוון שההסתברות לבחור SNP היא פרופורציונאלית לפונק' אקספוננט של ציונו, הרגישות ופרמטר הפרטיות זהו מנגנון אקספוננציאלי. בנספח C במאמר יש הוכחה מתמטית לכך שמנגנון זה הוא differentially private $\frac{\epsilon}{2}$.

III. במאמר הוכח כי תהליך הבחירה של שני האלגוריתמים מבטיח differentially private $\frac{\epsilon}{2}$. נסתכל על ההרעשה טרם השחרור – זו הרעשה לפלסית. הוכחנו בביתה כי רעש לפלס עם פרמטר $\frac{s}{\epsilon}$ שומר על differentially private ϵ ולכן במקרה שלנו האלג' שומר על differentially private $\frac{\epsilon}{2}$ בהינתן תהליך בחירה כלשהו. בסעיף א' הוכחנו כי במקרה כזה סה"כ נשמרת differentially private ϵ.

IV. הסבר הגרף:
- השורות הן 4 אפשרויות שונות של כמה SNP נרצה לבחור (כלומר ערך M).
- הציר האופקי הוא פרמטר הפרטיות ϵ (או כפי שמוכנה בגרף – "תקציב" הפרטיות).
- הציר האנכי הוא פונק' התועלת מבוטאת באחוזים, כלומר כמה מה-M SNPs המשמעותיים באמת הצליח כל אלגוריתם לבחור (כתלות ב-ϵ בהינתן M).

V. כשקובעים את פרמטר הפרטיות $\epsilon = 50$ מקבלים עבור x, y כך ש-$\|x - y\|_1 \leq 1$:

$$5 \cdot 10^{-21} \approx e^{-50} \leq \frac{P(\mathcal{M}_1(x) = s)}{P(\mathcal{M}_1(y) = s)} \leq e^{50} \approx 5 \cdot 10^{21}$$

היחס בין ההסתברויות שישוחרר אותו מידע על מדגמים עם הבדל של אובייקט בודד הוא בין אפס לטריליארד... כלומר חסמים שאינם רלוונטיים כלל ובעצם אין כמעט כלל הבטחה לאיזושהי פרטיות.
אינני מבין את ההיגיון מאחורי בחירת הטווח של ציר ה-ϵ, מראים תועלת יפה אבל ללא הבטחת פרטיות כלל באופן פרקטי. הייתי אומר שהטווח המעניין הוא כמעט שני סדרי גודל מתחת לזה ($\epsilon = 1$ ייתן יחס של $e \approx 2.7$ בין ההסתברויות שזה גם יחסית כבר קל למצוא את האובייקט שהשתנה).

VI. לפי הגרפים נראה שהמנגנון האקספוננציאלי (אלגוריתם 2) הוא העדיף מבין השניים שתוארו במאמר, מכיוון שלאותם ערכי "תקציב פרטיות" התועלת שלו היא תמיד גבוהה יותר. ההבדל בין האלגוריתמים היא אופן הבחירה אלו SNPs M ישוחררו ולכן אני חושב שההסבר להבדל בתועלות נובע משם.
אלגוריתם 2 משתמש במנגנון חכם שנותן עדיפות בהסתברות לבחירת SNP עם ציון גובה בניגוד לאלגוריתם 1 שבוחר לאחר הוספת רעש. התוצאה המוצגת בגרפים היא הוכחה לכך שהמנגנון של אלגוריתם 2 מוצלח (ואכן חכם).

VII. מנגנון *LocSig*, או JS, כפי שהוא מכונה במאמר, הוא מנגנון הדומה לאלג' האקספוננציאלי, אך עם פונק' ציון (שבו תלוי הסתב' הבחירה של SNP מסוים) שונה המתבססת על "מרחק" (כמה מדגם רחוק מקצה קבוצת המדגמים שעבורם פלט מסוים הוא האמיתי).
למנגנון הזה יש חסרונות של סיבוכיות ומחלות שמגיעות ב-ϵ גבוה ו/או ב-M נמוך, אבל באופן כללי התועלת שלו גבוהה משל שני האלגוריתמים שהוצגו במאמר לרוב ערכי M וערכי ϵ. ספציפית לערכי $\epsilon \leq 1$ אותם הגדרנו כרלוונטיים, מנגנון זה עדיף על האחרים לכל M.

שאלה 2:

סעיף (a)

I. נדרש לכתוב פורמולצית אות+רעש.

בשלב ראשון - מטריצת השונות/שונות משותפת A שתיבנה כך שהקורלציה בין משתנים באותה קבוצה (בגודל $\frac{p}{r} \times \frac{p}{r}$) היא 0.9 ובין משתנים בקבוצות שונות היא 0, הקורלציה של משתנה עם עצמו היא 1 (כך נקבל את האלכסון) וקיבלנו מטריצה כזו (לדוגמא כש- $p = 9\ ,\ \ r = 3$):

$$A = \begin{pmatrix} 1 & 0.9 & 0.9 & & & & & & \\ 0.9 & 1 & 0.9 & & 0 & & & 0 & \\ 0.9 & 0.9 & 1 & & & & & & \\ & & & 1 & 0.9 & 0.9 & & & \\ & 0 & & 0.9 & 1 & 0.9 & & 0 & \\ & & & 0.9 & 0.9 & 1 & & & \\ & & & & & & 1 & 0.9 & 0.9 \\ & 0 & & & 0 & & 0.9 & 1 & 0.9 \\ & & & & & & 0.9 & 0.9 & 1 \end{pmatrix}$$

כמתבקש, נציג את A בדרך הזו:

$$\mathrm{A} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \lambda_k v_k v_k^t + \sigma^2 I_p$$

כבר ניתן לראות ש- $\sigma^2 = 0.1$ ונקבל את המטריצה הבאה:

$$\sum_{\mathrm{k}=1}^{\mathrm{K}} \lambda_k v_k v_k^t = \begin{pmatrix} 0.9 & 0.9 & 0.9 & & & & & & \\ 0.9 & 0.9 & 0.9 & & 0 & & & 0 & \\ 0.9 & 0.9 & 0.9 & & & & & & \\ & & & 0.9 & 0.9 & 0.9 & & & \\ & 0 & & 0.9 & 0.9 & 0.9 & & 0 & \\ & & & 0.9 & 0.9 & 0.9 & & & \\ & & & & & & 0.9 & 0.9 & 0.9 \\ & 0 & & & 0 & & 0.9 & 0.9 & 0.9 \\ & & & & & & 0.9 & 0.9 & 0.9 \end{pmatrix}$$

כעת נמצא את וקטורי היחידה האורתוגונלים $v_k \in \mathbb{R}^p$. כדי שהוקטורים יהיו אורתוגונלים כל הרכיבים בהם יהיו 0 פרט ל- $\frac{p}{r}$ רכיבים חיוביים (שווים) ואז יתקיים עבור $j \neq k$: $v_j v_k^t = 0$. כדי שאלו יהיו וקטורי יחידה $\frac{p}{r}$ הרכיבים x צריכים לקיים: $\|v_k\| = \sqrt{\frac{p}{r}\mathrm{x}^2} = 1$ ומכאן ש- $x = \sqrt{\frac{r}{p}}$.

קיבלנו כי $K = r$ והוקטורים העצמיים הם מהצורה: $v_k^t = \left(0, \dots, 0, \sqrt{\frac{r}{p}}, \dots, \sqrt{\frac{r}{p}}, 0, \dots, 0\right)$ כש- $\frac{p}{r}$ האיברים החיוביים הם בהתאמה לקבוצה k.

נותר למצוא את הערכים העצמיים λ_k. המכפלה $v_k v_k^t$ נותנת מטריצה שבמטריצה הפנימית של קבוצה k כל הערכים הם $\sqrt{\frac{r}{p}} \cdot \sqrt{\frac{r}{p}} = \frac{r}{p}$ ומסביב אפסים. נרצה שהערכים האלו יהיו 0.9 ולכן נבחר בהתאם: $\lambda_k = 0.9\frac{p}{r}$.

שלב שני - נרצה להגיע לצורה של שורה $x_i \in \mathbb{R}^p$:

$$x_i = \sum_k s_{ik}\sqrt{\lambda_k} v_k + \sigma \xi_i$$

לאחר הצבת הערכים שכבר מצאנו:

$$x_i = \sum_k s_{ik}\sqrt{0.9\frac{p}{r}} v_k + \sqrt{0.1}\xi_i = \sum_k s_{ik} \begin{pmatrix} 0 \\ \vdots \\ \sqrt{0.9} \\ \vdots \\ 0 \end{pmatrix} + \sqrt{0.1}\xi_i$$

נותר להגדיר את $s_{ik} \in \mathbb{R}$ ואת $\xi_i \in \mathbb{R}^p$.

ידוע כי כל שורה x_i מתפלגת נורמלי סטנדרטי $(Z \sim N(0,1))$ באופן בלתי תלוי בשאר השורות, לכן נסיק (בעזרת הזהות $X = \sigma Z + \mu \Rightarrow X \sim N(\mu, \sigma^2)$) כי ξ_i הוא וקטור שכל רכיב שלו מתפלג נורמלי סטנדרטי וכך גם s_{ik} היא סדרה של סקלרים שכל אחד מתפלג (באופן ב"ת) נורמלי סטנדרטי (כך בתוך הקבוצה יש קורלציה ובין הקבוצות אין).

לסיכום קיבלנו כי כל שורה מורכבת מ- $r = K$ קבוצות שכל אחת מתפלגת באופן ב"ת:

$x_{ik} = 0.9N(0,1) + 0.1N(0,1)$, כשהחלק הראשון הוא אות והשני הוא רעש.

II. כפי שלמדנו מבועז (הרחבה של התיאוריה של M&P) התנאי סף ליכולת לשחזר את הוקטורים העצמיים הוא: $\lambda > \sigma^2\sqrt{\frac{p}{n}}$.

נציב את הערכים במקרה שלנו: $0.9\frac{p}{r} > 0.1\sqrt{\frac{p}{n}}$

ונקבל את היחס הנדרש בין p, n, r : $\frac{r}{\sqrt{np}} < 9$.

סעיף (b)

I. להבנתי שתי השיטות שיתפקדו באופן דומה הן Lasso ו-L_0 variable selection מכיוון ששתיהן בוחרות משתנים שביחד מסבירים הכי טוב את משתנה המטרה Y לעומת שיטת ה- Marginal regression שבוחרת משתנים לפי כמה טוב הם מסבירים בעצמם את Y ולכן גם תתפקד משמעותית פחות טוב.

- במשימה הראשונה – זיהוי הקבוצות המשתתפות בפיתרון:
MR לא תזהה נכון את כל הקבוצות מכיוון שתבחר משתנים רבים מהקבוצה בה $\beta_{j1} = 10$, זאת מפני שלכל משתנה בנפרד קורלציה גדולה עם Y, דבר הנובע מהקורלציה הגבוהה ביניהם ולא מהקורלציה האמיתית ל-Y (פרט למשתנה אחד בקבוצה לו קשר אמיתי עם Y).
לעומתה, שתי השיטות האחרות יצליחו לזהות את הקבוצות שבאמת חלק מהפיתרון, זאת בזכות השיטה שציינתי בפסקה הראשונה (בחירת משתנים שמסבירים ביחד).
- במשימה השניה – זיהוי המשתנים הספיציפים שמשתתפים בפיתרון:
MR כמובן לא תצלח במשימה זו מכיוון שבחרה את רוב המשתנים מקבוצה אחת.
לגבי שתי השיטות האחרות – קשה להגיד כמה יצליחו. בשל הקורלציה הגבוהה בין המשתנים בתוך כל קבוצה והרעש שהוספנו למטריצה ייתכן וייבחר משתנה שבגלל סיבות אלו נראה מתאים יותר מהמשתנה האמיתי באותה קבוצה.

II. במשימת הערכת ה-β-ות כמובן ש-MR שוב לא תצליח (טעות נגררת...) כי היא תיתן β חיובי למשתנים שה-β האמיתי שלהם הוא אפס ולהיפך.
ההצלחה של שתי השיטות האחרות במשימה זו תלויה בהצלחתן במשימה בחירת המשתנים המדויקים. בהנחה שבמשימה ההיא השיטות הללו לא כל כך הצליחו, אני משער ששיטת Lasso תצליח הכי טוב, או הכי פחות גרוע, במשימה הנוכחית. שיטה זו משלבת "עונש" ובכך מקטינה את ערכי ה-β ואם היו טעויות רבות בבחירת המשתנים, כך שניתן β חיובי למשתנה שה-β שלו הוא אפס ולהיפך, הקטנת ה-β-ות תהיה דבר חיובי ותקטין את סך הטעות בהערכה. שיטת L_0 בעצם תיענש על כך שלא נותנת עונשים ותהיה עם סך טעות הערכה גדולה יותר, ייתכן שאפילו בסדר גודל הטעות של MR (מכיוון שבחירת הקבוצה המתאימה לא תורמת, נדרש לבחור את המשתנים הנכונים והיא לא בהכרח הצליחה טוב יותר בכך). במידה ו-L_0 ו-Lasso הצליחו לבחור טוב את המשתנים, הקטנת β-ות תפעל הפוך ותגדיל את סך טעות ההערכה ואז L_0 תצליח הכי טוב במשימה הזו.

שאלה 3:

סעיף (a)

I. הקוד:

```
con = url("http:/[redacted].rdat")
load(con)

library(statnet)
library(lmtest)

# Full Disclosure: The code is based on the example in the course's website

# Fit a model with indegree, outdegree & mutuality:
p1.w_in<-pstar(fr.n,effects=c("outdegree","indegree","mutuality"))

# Fit a model with outdegree & mutuality:
p1.wo_in<-pstar(fr.n,effects=c("outdegree","mutuality"))

lrtest(p1.w_in,p1.wo_in)
```

(הסוף של) הפלט:

```
  #Df   LogLik  Df  Chisq Pr(>Chisq)
1 142  -946.54
2  72 -1093.81 -70 294.53  < 2.2e-16 ***
```

התקבל p-value שואף לאפס, כלומר בכל רמת מובהקות רלוונטית אנחנו דוחים את השערת האפס כי פופולריות לא משנה את המודל ומסיקים מכך שזהו משתנה מובהק.

II. בחרתי כסטטיסטי את סכום המרחקים בין ערך נצפה לערך צפוי ($\sum_{i=1}^{71}|O_i - E_i|$), כלומר, בין מספר הקשתות הנכנסות של כל אובייקט למספר הקשתות שהיו לו אם הייתה התפלגות אחידה ($E_i = \mathrm{E}$). בחרתי אותו כי הוא פשוט והרבה non-uniformity tests מבוססים עליו.

הקוד:

```
O = colSums(fr)                          # Observed original indegree
E = sum(fr)/nrow(fr)                     # Expected (uniformed) indegree
stat.O = sum(abs(O-E))                   # Calculating the statistic for the original

permutaions = 10^4
count = 0

for (p in 1:permutaions){
  P = fr                                 # Just for simple initial
  for (i in 1:nrow(P)){
    P[i,] = sample(fr[i,])               # Mixing the row for the permutaion
    while (P[i,i]==1){                   # Making sure no self edges
      P[i,] = sample(fr[i,])             # Trying again
    }
    if (sum(fr[i,])!=sum(P[i,])){        # Making sure the sum of the row remained
      print('Error')
    }
  }

  O = colSums(P)                         # Permuted indegree
  stat.P = sum(abs(O-E))                 # Calculating the statistic for the original

  if (stat.P > stat.O){                  # Counting how many permuted statistic passed the original
    count=+1
  }
}

pv = count/permutaions                   # Calculating the p-value
cat("The p-value is:", pv)
```

הפלט:

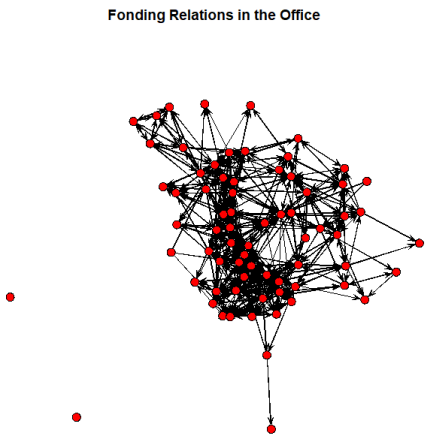
```
> cat("The p-value is:", pv)
The p-value is: 0
```

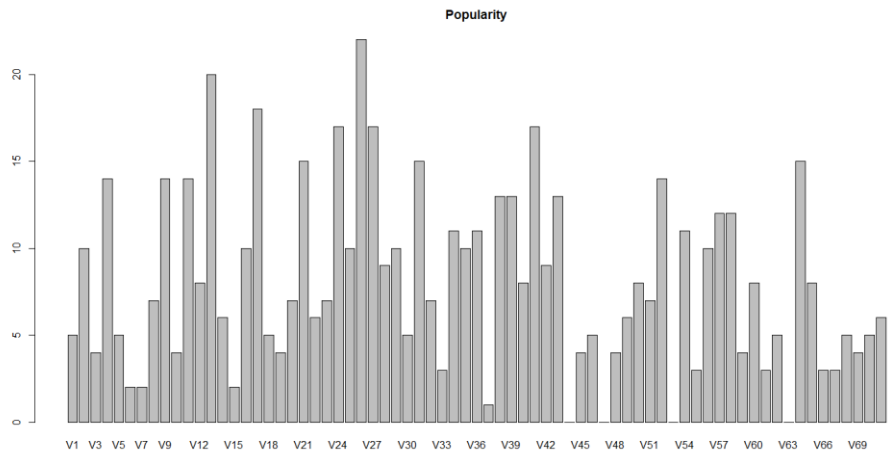
התקבל כי ה-p-value, קרי אחוז המקרים בהם הסטטיסטי של מטריצה "מוחלפת" הוא גבוה יותר מהסטטיסטי של המטריצה המקורית, הוא 0.
כלומר, כל המטריצות שערבבנו באופן אקראי קרובות יותר להתפלגות אחידה מהמטריצה המקורית ולכן מגיעים לאותה מסקנה והיא שהפופולריות לא מתפלגת באופן אחיד בין האובייקטים, כלומר היא משתנה מובהק (כל עו"ד הרוויח (או "הרוויח") את הפופולריות שלו).

להלן שני עזרים חזותיים:

הקוד:

```
# Graphs:
barplot(colSums(fr), main = "Popularity")
plot (fr.n, main = "Fonding Relations in the Office")
```





הגרפים הללו מחזקים את המסקנה שהצגתי בסעיף הקודם:
בגרף העמודות רואים כי התפלגות הפופולריות מאוד לא אחידה בין האובייקטים.
בגרף הקודקודים והקשתות ניתן לראות כי עבור קודקודים מסוימים יש הרבה קשתות נכנסות ועבור אחרים מעט עד כדי בכלל לא, קרי, חלק מעורכי הדין מחובבים ע"י הרבה מעמיתיהם וחלק ע"י מעט או אף אחד מהם.

סעיף (b)

I. אציג את המודל בארבעת השילובים של 2-3 מימדים ו-2-3 אשכולות.

הקוד:

```
# Full Disclosure: The code is based on the example in the course's website

erg2.2 = ergmm(fr.n~euclidean(d=2,G=2))
plot (erg2.2,labels=T, main = "2 Clusters in 2D")
summary(erg2.2)

erg2.3 = ergmm(fr.n~euclidean(d=2,G=3))
plot (erg2.3,labels=T, main = "3 Clusters in 2D")

erg3.2 = ergmm(fr.n~euclidean(d=3,G=2))
plot (erg3.2,labels=T,use.rgl = TRUE, main = "2 Clusters in 3D")
plot (erg3.2,labels=T, main = "2 Clusters in 3D projected on 2 PC")

erg3.3 = ergmm(fr.n~euclidean(d=3,G=3))
plot (erg3.3,labels=T,use.rgl = TRUE, main = "3 Clusters in 3D")
plot (erg3.3,labels=T, main = "3 Clusters in 3D projected on 2 PC")

bic.ergmm(erg2.2)$Z
bic.ergmm(erg2.3)$Z
bic.ergmm(erg3.2)$Z
bic.ergmm(erg3.3)$Z
```

לדוגמא, כך נראה פלט סיכום למודל (2 מימדים, 2 אשכולות):

```
==========================
Summary of model fit
==========================

Formula:   fr.n ~ euclidean(d = 2, G = 2)
Attribute: edges
Model:     Bernoulli
MCMC sample of size 4000, draws are 10 iterations apart, after burnin of 10000 iterations.
Covariate coefficients posterior means:
            Estimate   2.5%  97.5% 2*min(Pr(>0),Pr(<0))
(Intercept)   1.7782 1.5468 2.0346            < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Overall BIC:        2793.818
Likelihood BIC:     2082.174
Latent space/clustering BIC:     711.6434

Covariate coefficients MKL:
            Estimate
(Intercept) 1.255632
```

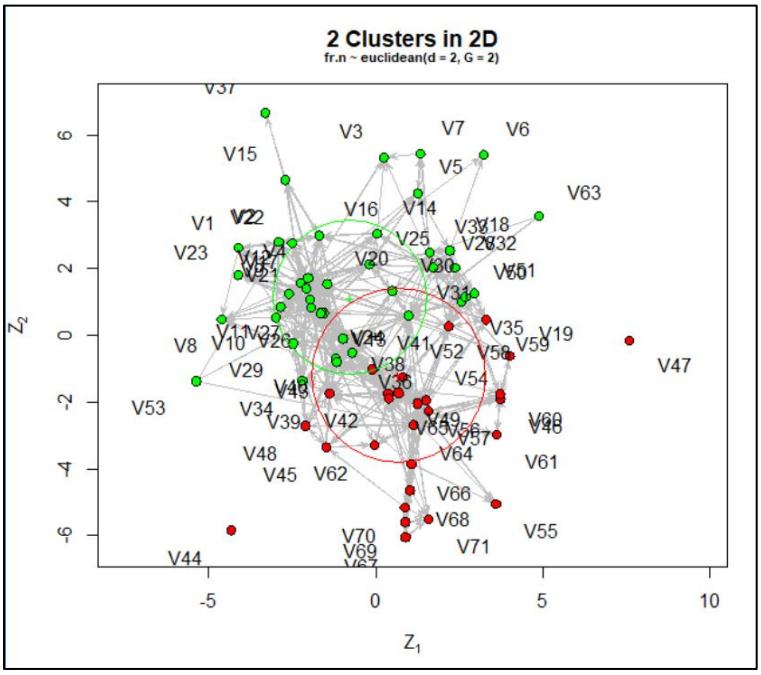
בקו הכתום המקווקו מסומן הנתון הרלוונטי ביותר להשוואת המודלים.
BIC נמוך יותר משמעותו נראות גבוהה יותר להיות המודל האמיתי.

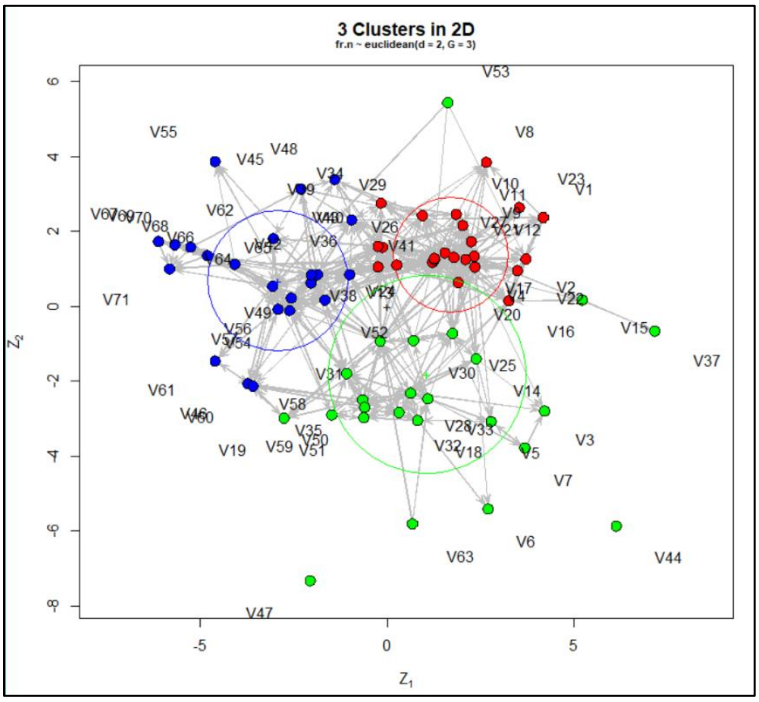
להלן **<u>התוצאות הסטטיסטיות:</u>**

```
> bic.ergmm(erg2.2)$Z
[1] 711.6434
> bic.ergmm(erg2.3)$Z
[1] 707.2822
> bic.ergmm(erg3.2)$Z
[1] 1065.848
> bic.ergmm(erg3.3)$Z
[1] 1072.499
```

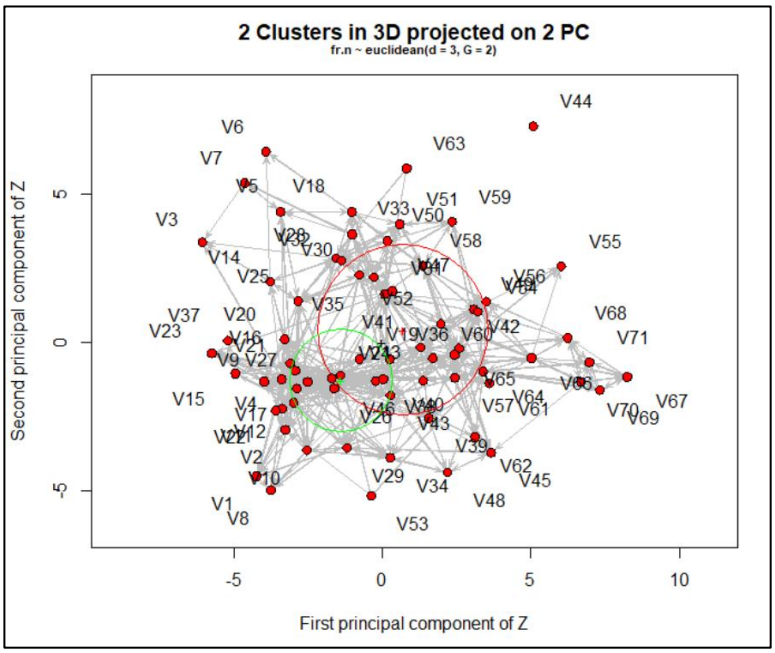
אני מבין מזה שלמודלים עם 2 או 3 אשכולות התאמה דומה למציאות עם יתרון קל לשלוש אשכולות. לצערי לא הצלחתי להריץ את המודל ל-4 אשכולות, מעניין לראות אם היה שיפור נוסף, להערכתי לא, מכיוון שבד"כ בגרפים של בחירת אשכולות אין "פלאטו" באזור השיא, אם שני הערכים של 2,3 קרובים אני צופה שב-4 כבר תהיה ירידה (זה כמובן מבוסס אינטואיציה בלבד).
בנוסף, מודלים ב-3 מימדים משמעותית פחות תואמים למציאות.

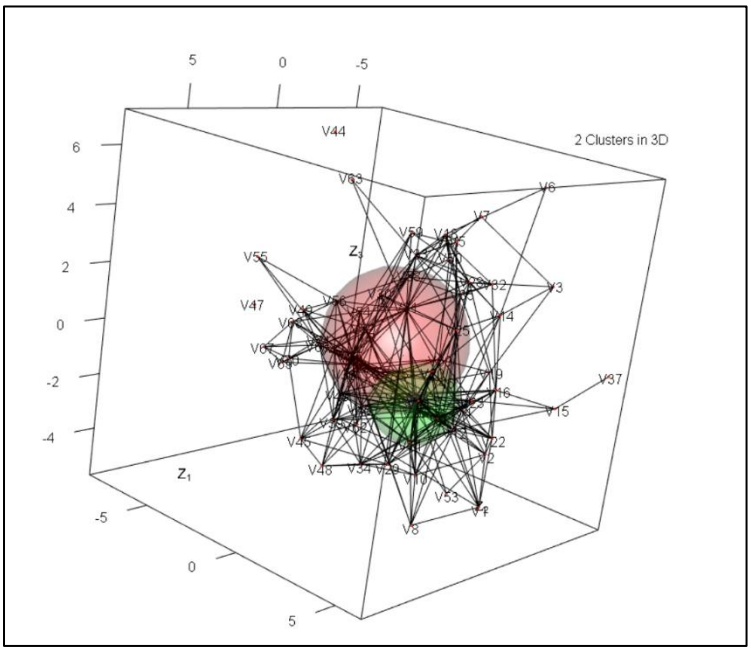
כעת, **<u>גרפים</u>**...

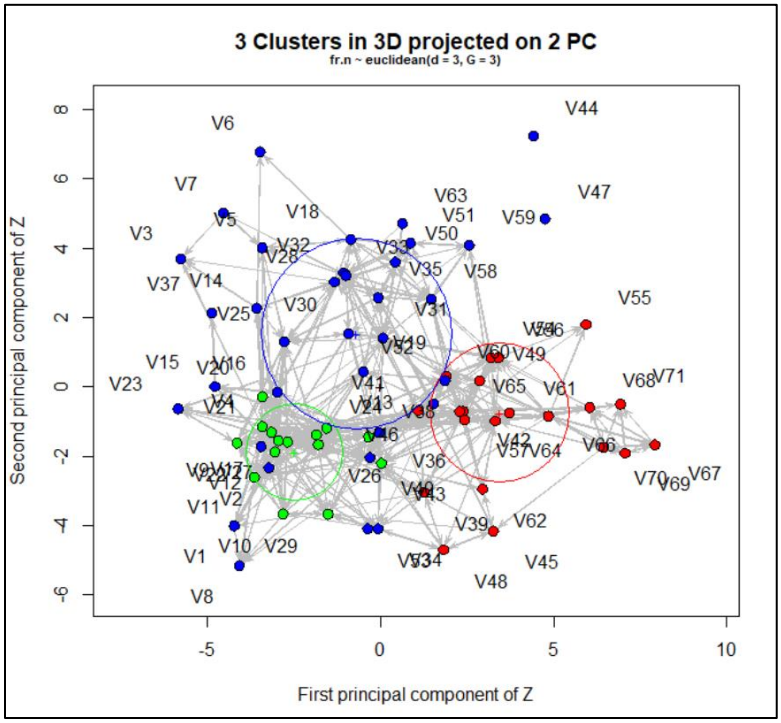


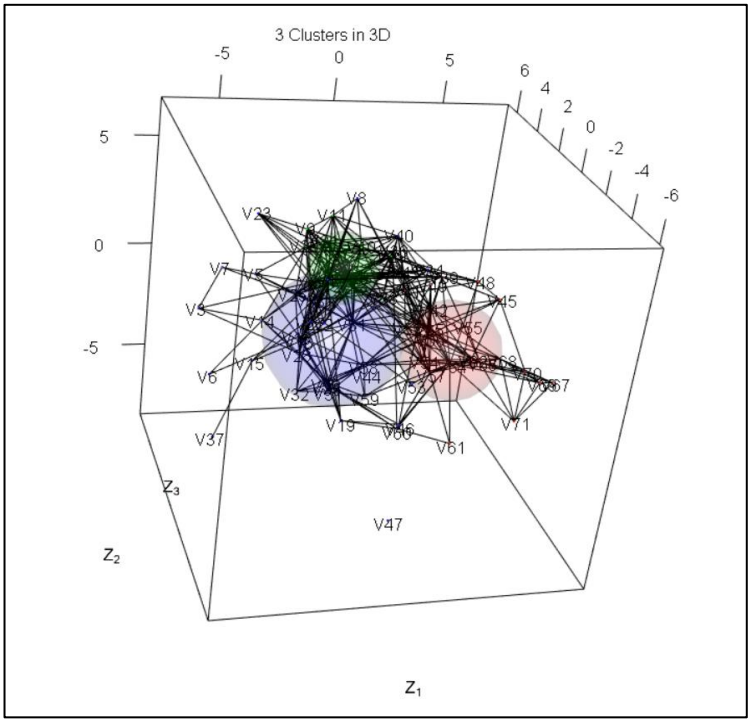


<u>מסקנות:</u> ב-2 מימדים אכן נראית חלוקה די טובה גם ל-2 אשכולות וגם ל-3 (פרט ל-3 אלמנטים עם מעט מאוד קשתות שנראה יותר הגיוני שישוייכו לאשכול האדום מאשר הירוק, כנראה מיעוט הקשתות איכשהו משפיע על החישוב).









<u>מסקנות</u>: אכן נראה שב-3 מימדים החלוקה לאשכולות פחות אחידה.
ל-3 קבוצות נראה שיש חפיפה רבה לעומת 2 מימדים.
ל-2 קבוצות יצא גרף מוזר, עם 2 אשכולות אך נראה שכולם שייכים לאותה קבוצה, מכיוון שאין הרבה איפה לטעות בקוד, העניין נשאר כתעלומה בעיני...

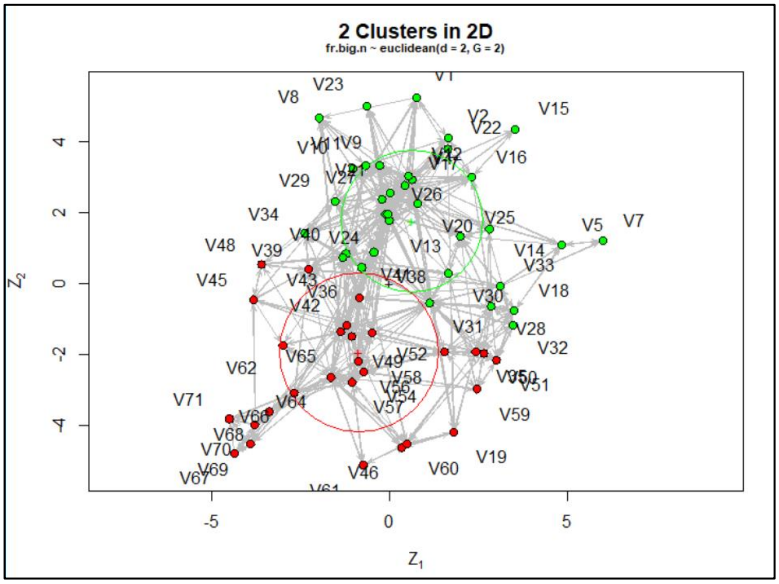
II. נחזור על אותם מודלים לאחר השמטת 8 פרקליטים לא סוציאליים (לאו דווקא אגב עם אפס קשתות נכנסות/יוצאות לפי גרף העמודות שהצגתי קודם, אבל עד כדי קשתות בודדות). הקוד כמובן דומה מאוד.

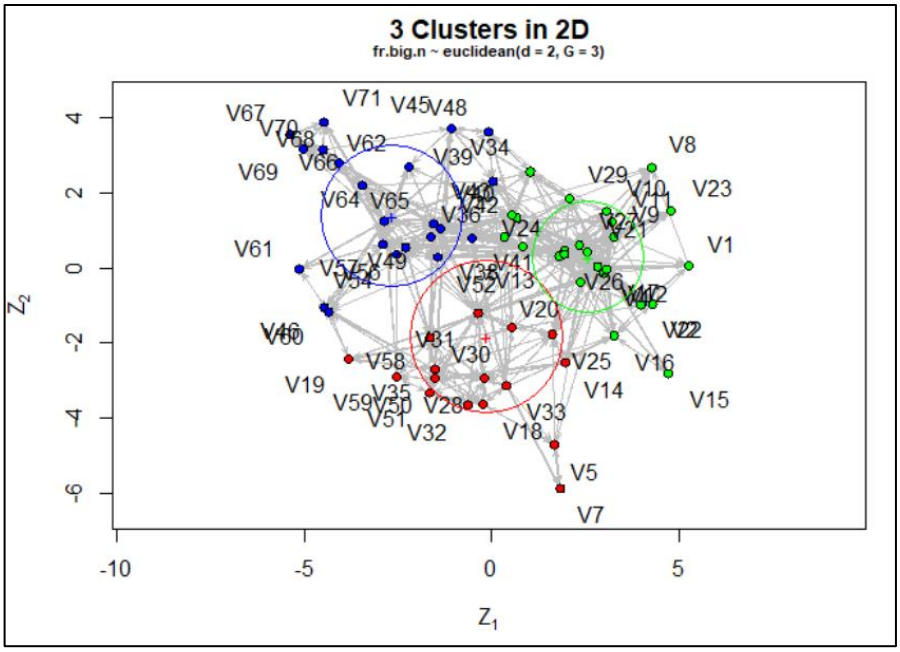
```
> bic.ergmm(erg2.2b)$Z
[1] 610.0752
> bic.ergmm(erg2.3b)$Z
[1] 615.2608
> bic.ergmm(erg3.2b)$Z
[1] 904.7188
> bic.ergmm(erg3.3b)$Z
[1] 912.3526
```

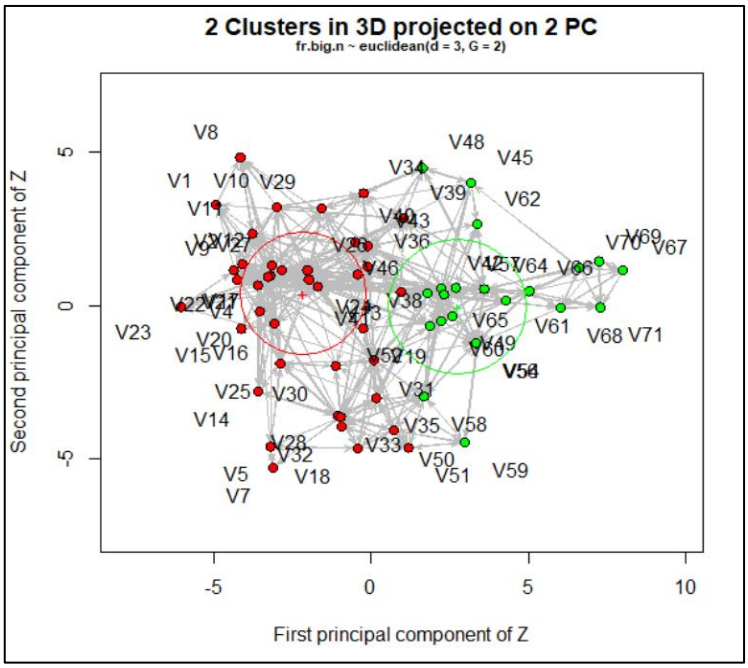
תחילה, **<u>התוצאות הסטטיסטיות:</u>**
ניתן לראות קודם כל כי כל המודלים השתפרו.
דבר שני, הסדר בגדול נשמר – ללא הבדל גדול בין 2 ל-3 אשכולות ועדיפות ניכרת ל-2 מימדים על 3. אך בקטן הסדר התהפך, הפעם יתרון קל ל-3 אשכולות לעומת 2.

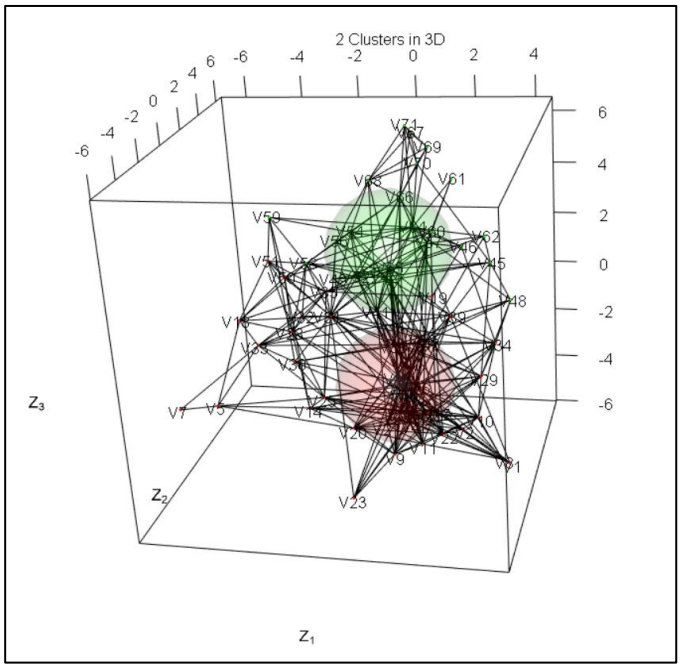
<u>גרפים</u>:

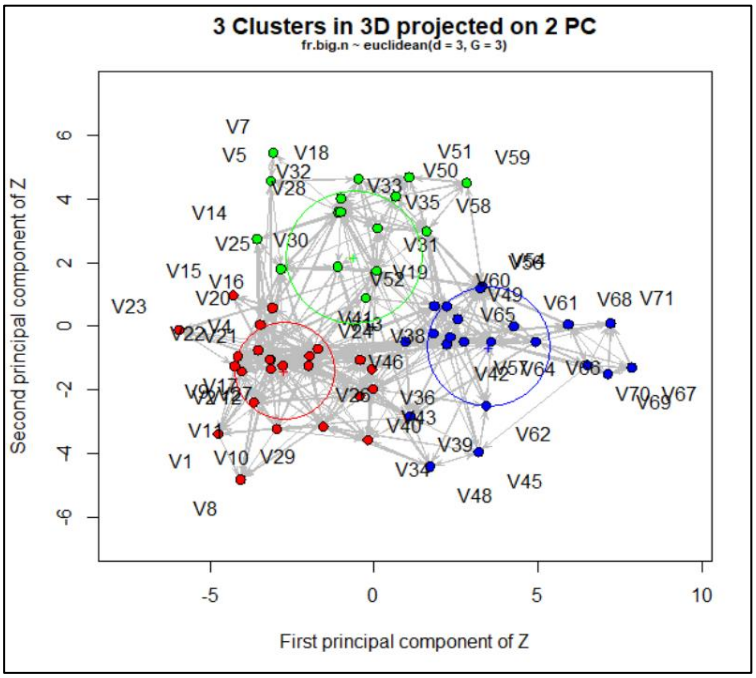
ב-2 מימדים החלוקה לאשכולות אכן נראית טובה עוד יותר גם ל-2 אשכולות וגם ל-3.

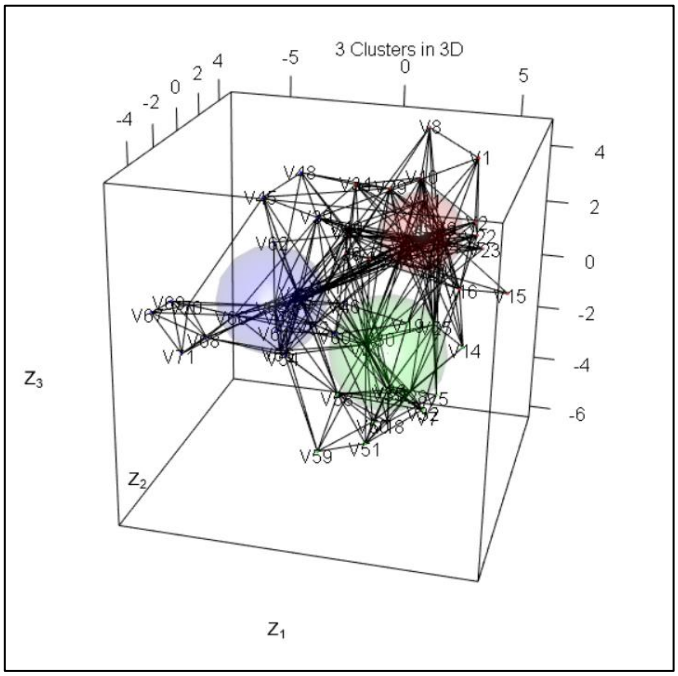












ניתן לראות שההפרדה בין האשכולות ב-3 מימדים טובה לאחר עדכון המודל.

לסיכום היה שיפור ניכר, אך לא גדול מאוד, בין המודלים לאחר הסרת האובייקטים הלא/במעט לא מקושרים. זה לא מפתיע אותי שזו הייתה ההשפעה להסרה שלהם, אני מניח שקשה למודל לקבוע לאיזו קבוצה הם שייכים וזה מה שפוגע באיכות, עם זאת הם מהווים חלק יחסית קטן מהכלל ולכן השיפור בהסרתם סה"כ שומר על אותו סדר הגודל. מעניין גם לראות שהעדיפות המינורית בין 2 ל-3 אשכולות התהפכה, כנראה מדגיש שההבדל בין טיב שני המודלים הללו באמת מזערי.