מבוא להסקה סיבתית – פרויקט סיום

Detecting Confounding in Multivariate Linear Models via Spectral Analysis

by Dominik Janzing and Bernhard Schölkopf

מגישים:

אורן פלוזניק – 203933551

שי כליפה - 318801859

תקציר:

אחת המטרות החשובות ביותר בניתוח סטטיסטי של נתונים היא הערכת הקשר הסיבתי שבין המשתנים המסבירים למשתנה המוסבר. למען מטרה זו, אחת ההנחות המרכזיות שיש להניח במודלים סיבתיים היא הנחת ה Ignorability אשר לפיה לא קיים ערפלן מוסתר. אמנם, במקרים רבים מניחים את הנחה זו מבלי להתבסס על מדדים כמותיים למרות שזו אינה הנחה טריוויאלית. למעשה, ללא הנחות נוספות על המודל והנתונים לא ניתן לקבוע בוודאות האם אכן הנחה זו מתקיימת.

בעבודה זו נסקור ונסכם את המאמר Detecting Confounding in Multivariate Linear Models via Spectral Analysis מאת Dominik Janzing וBernhard Schölkopf אשר עוסק בזיהוי קיומם של ערפלנים מוסתרים ובקיום הנחת ה Ignorability תחת תנאים מסוימים. השיטה שהמאמר מציע עובדת על מודל לינארי ומבוססת על עקרון ריכוז המידה ואנליזה ספקטרלית דרך בחינת האוריינטציה שבין מקדמי המודל הלינארי לבין הוקטורים העצמאיים של מטריצת השונות המשותפת של הנתונים המסבירים. בנוסף, ביצענו שחזור של חלק מהניסויים אותם ביצעו החוקרים ובחנו את השפעותיהם של מצבים שונים ושינויים ברכיבים מסוימים במודל על יעילות השיטה, בעזרת ניסויים נוספים.

כמו כן, נדון ברעיונות לכיווני מחקר עתידיים לרעיון המוצג במאמר.

הצגת הבעיה והגדרת המודל:

אחת המטרות החשובות ביותר בניתוח סטטיסטי של נתונים היא הערכת ההשפעה הסיבתית של משתנים מסבירים $X_1, \dots, X_d$ על המשתנה המוסבר Y. עם זאת, השגת מטרה זו ללא התערבות הינה משימה קשה אשר דורשת הנחות רבות. לפי עקרון הסיבתיות המשותפת של Reichenbach לא תמיד הקשר הנצפה בין X_i ל Y הוא ההשפעה הסיבתית של X_i על Y. על פי עקרון זה, ייתכן והקשר הנצפה נובע מהשפעת Y על X_i או ממשתנה שלישי Z, הנקרא ערפלן, אשר משפיע על שני המשתנים. בעבודה זו נתמקד בזיהוי המקרה השני בו קיים המשתנה Z. במאמר לא התייחסו לקשר בין המשתנים המסבירים ועל כן איגדו אותם כ$X = X_1, \dots, X_d$. כמו כן, המאמר מתייחס למקרה בו קיים ערפלן ממשי, יחיד ורציף בלבד ומניח קשרים לינאריים כדלהלן:

$$\mathbf{X} = \mathbf{b}Z + \mathbf{E} \tag{1}$$

$$Y = \langle \mathbf{a}, \mathbf{X} \rangle + cZ + F \tag{2}$$

הנחות וסימונים:

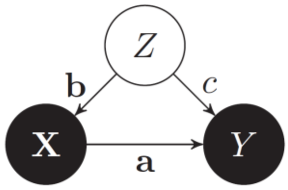
1) E וקטור אקראי בR^d.
2) Z, F משתנים אקראיים בR.
3) E , Z, F בלתי תלויים.
4) נסמן בa את ההשפעה הסיבתית של X על Y.

5) b פרמטר בR^d אשר מייצג את ההשפעה של Z על X.

6) c פרמטר בR אשר קובע את ההשפעה של Z על Y.

7) $Var(Z) = 1$.

8) כל המשתנים מתפלגים התפלגות נורמלית סביב 0.



התרחישים המתוארים במאמר מתייחסים לDAG זה ומשתנים רק בפרמטרים $a, b, c, \sigma_F, \Sigma_{EE}$

במקרים רבים, כאשר רוצים לאמוד את השפעת X על Y, אומדים את המשתנה a בעזרת רגרסיה לינארית באופן הבא:

$$\hat{\mathbf{a}} := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y}$$

אמנם, $\hat{a}$ למעשה מתאר את הקשר המתאמי ולא מתחשב באופן ישיר במבנה הסיבתי המכיל את X ו Y (המוצג בנוסחאות (1) ו(2))

כדי לבחון את הקשר שבין $\hat{a}$ ו a כותבי המאמר מתבססים על הפיתוח הבא (אשר מתקבל מפיתוח המשוואה לעיל על בסיס הנחת אי התלות בין Z, F , E ועל כך שZ הינו מנורמל):

3) $$\hat{\mathbf{a}} = \underbrace{\mathbf{a}} + \underbrace{(\Sigma_{\mathbf{EE}} + \mathbf{bb}^T)^{-1} c\, \mathbf{b}}$$

הטיה הנגרמת כתוצאה מנוכחות הערפלן.

ההשפעה הסיבתית האמיתית של X על Y.

המאמר מציין עבודות נוספות העוסקות בזיהוי ערפלנים תחת הנחות שונות([1] [3] [2]), אך המאמר לעומתן מציע שיטה לזיהוי נוכחות ערפלן מוסתר, תוך כדי שהוא מסתמך רק על סטטיים מסדר שני ולא מסתמך על רעש שאינו גאוסייני ואי תלות מסדר גבוה.

<u>מדדים לאפיון הערפול:</u>

<u>מדידת הערפול המבני:</u>

המדד העיקרי בו כותבי המאמר משתמשים עבור מדידת הסטייה של $\hat{a}$ מ a הינו מדד הערפול המבני הנתון באופן הבא:

$$\beta := \frac{\|c \cdot \Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\|^2}{\|\mathbf{a}\|^2 + \|c \cdot \Sigma_{\mathbf{XX}}^{-1}\mathbf{b}\|^2} \in [0, 1]$$

כאשר המונה הינו בעצם נורמת ההטיה הנגרמת מנוכחות הערפלן בריבוע והמכנה הינו סכום נורמת ההסבר הסיבתי בריבוע ונורמת ההטיה בריבוע.

<u>מדידת השפעת הערפלן על X:</u>

מאחר וניתן לקבל את אותו הערך לβ בשני מצבים שונים מאד:

1) b קטן וc גדול – במצב זה, ידיעת Z משפיעה במידה מועטה על חוסר הודאות בנוגע לערכי X.
2) b גדול וc קטן – במצב זה, ידיעת Z מקטינה את חוסר הודאות בנוגע לערכי X ומאפשר לשערך אותו בקלות יחסית

כדי להבדיל בין שני המצבים משתמשים במדד הבא: $\eta := \mathrm{tr}(\Sigma_{\mathbf{XX}}) - \mathrm{tr}(\Sigma_{\mathbf{XX}|Z}) = \mathrm{tr}(\Sigma_{\mathbf{XX}}) - \mathrm{tr}(\Sigma_{\mathbf{EE}}) = \|\mathbf{b}\|^2 \le \|\Sigma_{\mathbf{XX}}\|$

<u>אינטואיציה לפתרון הבעיה המוצע במאמר</u>

נניח שיש לנו גרף DAG המכיל את המשתנים $Z_1, \ldots, Z_n$ *ונסמן את קבוצת הוריו של כל משתנה כ*PA_i . לפי עקרון ה *Independent Conditionals)* [4] [5]*(תחת ההנחה שגרף DAG מייצג מודל סיבתי, על המשתנים בו להיות בלתי תלויים בהינתן הוריהם בגרף.*

כך שנניח ואנחנו במצב בו ההתפלגות של $Z_i|PA_i$ *הינה התפלגות פרמטרית המאופיינת על ידי* θ_i *ונניח כי קיימת תלות כלשהי בין* $\theta_1, \ldots, \theta_n$ *אז סביר כי קיימת גם תלות בין ההתפלגויות של* $Z_i|PA_i$ *ועל כן על בסיס העיקרון סביר שהגרף אינו מייצג את המבנה הסיבתי נכונה.*

כעת נבחן את המודל הסיבתי $X \to Y$ *, כאשר במודל זה* $Y|X$ *ו* X *אינם תלויים. במקרה זה, הפרמטר המאפיין של ההתפלגות* $P(X)$ *הינו* Σ_{XX} *והפרמטר המאפיין של ההתפלגות* $P(Y|X)$ *הינו* a.

כך שעל בסיס מה שכתבנו למעלה, אם קיימת תלות בין Σ_{XX} *ל*a *אז סביר שהמודל הסיבתי של* $X \to Y$ *לא מתקיים. לכן אם אכן* $X \to Y$ *נצפה למצוא בסבירות גבוה כי*a *נמצא באוריינטציה גנרית (יוסבר בהמשך) ביחס ל*Σ_{XX}.

<u>הגדרת אוריינטציה גנרית דרך מידה ספקטרלית מושרת:</u>

<u>אזהרה:</u> *המטרה תהיה להגיע להנחות מסוימות על התנהגות המידות (שיוגדרו בהמשך) שיגדירו את האוריינטציה. אנו עוברים על החלק המתמטי במאמר בסקירה שטחית יחסית, ללא התעמקות בהוכחות ומעברים. נסקור בעיקר הגדרות, טענות ולמות שיובילו לבניית מושג האוריינטציה הגנרית באופן מתמטי.*

הגדרת האוריינטציה מבוססת על אנליזה ספקטרלית של מטריצות השונות המשותפת ועל עקרון ריכוז המידה.

לשם כך, נתבסס על המשפטים וההגדרות הבאים:

1) <u>הגדרה:</u> העקבה המנורמלת עבור $A \in R^{dxd}$ הינה: $\tau(A) := \frac{1}{d}\mathrm{tr}(A)$

- במאמר מתרכזים במקרה בו הערכים העצמיים של כל המטריצות הם שונים.

2) <u>הגדרה:</u> מידת העקבה הספקטרלית: תהי $A \in R$ מטריצה סימטרית עם ערכים עצמיים $\lambda_1 > \cdots > \lambda_d$ אז מידת העקבה הספקטרלית μ_A^τ של A היא המידה הדיסקרטית על R הניתנת על ידי ההתפלגות האחידה על הערכים העצמיים של A. כך ש $\mu_A^\tau := \frac{1}{d}\sum_{j=1}^{d}\delta_{\lambda_j}$ כאשר δ_s הינה מידת דיראק עבור $s \in R$.

3) <u>למה</u> ([6]) : התוחלת עבור מידת העקבה הספקטרלית:
התוחלת של כל פונקציה מן ערכים עצמיים לערך ממשי, ביחס למידת העקבה הספקטרלית, ניתן לחישוב על ידי: $\int f(s)d\mu_A^\tau(s) = \tau(f(A))$

4) <u>הגדרה:</u> מידה ספקטרלית המושרית על ידי וקטור:
תהי A מטריצה סימטרית עם ערכים עצמיים $\lambda_1 > \cdots > \lambda_d$ ווקטורים עצמיים $\phi_1, \ldots, \phi_d$ עבור $\psi \in R^d$ המידה הספקטרלית המושרית על ידו מוגדרת באופן הבא: $\mu_{A,\psi}(S) = \sum_{j \text{ with } \lambda_j \in S} \langle\psi, \phi_j\rangle^2$
לכל קבוצה בת-מניה $S \subset R$

5) <u>למה:</u> תוחלת עבור מידה ספקטרלית המושרית על ידי וקטור:
התוחלת עבור פונקציה f מעל הספקטורם של A ביחס ל $\mu_{A,\psi}$ ניתנת לכתיבה באופן הבא: $\int f(s)d\mu_{A,\psi}(s) = \langle\psi, f(A)\psi\rangle$

- אחד מהרעיונות העיקריים העומדים במרכז השיטה המוצגת במאמר לזיהוי הערפלן הוא ש $\mu_{A,\psi}$ קרוב ל $||\psi||^2\mu_A^\tau$ לכל בחירה טיפוסית של ψ .

6) <u>למה:</u> תהי $(A_d)_{d\in N}$ כאשר, $||A_d|| \le a$, סדרה של מטריצות סימטריות שהמידה הספקטרלית שלהן מתכנסת בצורה חלשה למידת הסתברות μ^∞, כלומר –

$$\mu_{A_d}^\tau \to \mu^\infty$$

ותהי $(c_d)_{d\in N}$ עם $c_d \in R^d$ סדרה המוגרלת באופן אקראי מכדור ברדיוס r. אז $\mu_{A_d,c_d} \to r^2\mu^\infty$ בהסתברות באופן חלש.

- אחד הרעיונות שהמאמר מציג הוא שנוכחותו של ערפלן משבשת את $\mu_{\Sigma_{XX},\hat{a}}$ בצורה הניתנת לאפיון בהינתן שהפרמטרים של המודל הם טיפוסיים. כדי לבחון את מאפיינים אלה מגדירים את המודלים הבאים:
  - $(\Sigma_{EE}^d)_{d \in N}$ סדרה חסומה על ידי חסם משותף של מטריצות בגודל dxd כך שמידת העקבה הספקטרלית שלהם מתכנסת באופן חלש למידת הסתברות μ^∞
  - $(a_d)_{d\in N}$ סדרה של וקטורים בR^d המוגרלים באופן אחיד מכדור בעל רדיוס קבוע r_a
  - $(b_d)_{d\in N}$ סדרה של וקטורים בR^d המוגרלים באופן אחיד מכדור בעל רדיוס קבוע r_b באופן בלתי תלוי ב a_d אז,

$$\Sigma^d_{\mathbf{XX}} = \Sigma^d_{\mathbf{EE}} + \mathbf{b}_d\mathbf{b}^T_d \text{ and } \hat{\mathbf{a}}_d = \mathbf{a}_d + c(\Sigma^d_{\mathbf{XX}})^{-1}\mathbf{b}_d$$

כאשר c קבוע לכל d.

7) <u>משפט:</u> מידה ספקטרלית אסימפטוטית עבור מודל המייצר שמורת-סיבוב:

עבור המודל המוגדר לעיל, הקירובים הבאים מוצדקים באופן אסימפטוטי.

(1
$$\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d} \to r^2_{\mathbf{a}}\mu^\infty$$

(2
$$\mu_{\Sigma^d_{\mathbf{EE}},\mathbf{b}_d} \to r^2_{\mathbf{b}}\mu^\infty$$

(3
$$\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d+c(\Sigma^d_{\mathbf{XX}})^{-1}\mathbf{b}_d} - (\mu_{\Sigma^d_{\mathbf{XX}},\mathbf{a}_d} + \mu_{\Sigma^d_{\mathbf{XX}},c(\Sigma^d_{\mathbf{XX}})^{-1}\mathbf{b}_d}) \to 0$$

בהסתברות, באופן חלש.

- *אינטואיטיבית המשוואה הראשונה אומרת שהמידה שמשרה a על Σ_{XX} קרובה למידת העקבה הספקטרלית עד כדי נורמליזציה. מאחר ולא יש אוריינטציה "גנרית" לΣ_{XX} מאחר והוא נבחר באופן בלתי תלוי בΣ_{EE} ובb. מצד שני b נבחר באופן בלתי תלוי בΣ_{EE}, ולכן המידה הספקטרלית המושרה על ידי b וΣ_{EE} קרובה למידת העקבה המתאימה עד כדי נורמליזציה, כפי שרואים במשוואה 2.*

במשוואה 3 רואים כי $\mu_{\Sigma_{XX},a+\Sigma^{-1}_{XX}b}$ כמעט מתפרק ל$\mu_{\Sigma_{XX},a}$ ועוד $\mu_{\Sigma_{XX},\Sigma^{-1}_{XX}b}$, זו היא המידה הספקטרלית המושרה על ידי a וההפרעה שלו $\Sigma^{-1}_{XX}b$. זה שימושי כי משוואה 1 מתארת את ההתנהגות האסימפטוטית של $\mu_{\Sigma_{XX},a}$. לעומת זאת כדי להבין את ההתנהגות האסימפטוטית של $\mu_{\Sigma_{XX},\Sigma^{-1}_{XX}b}$ נצטרך שהשתמש במשוואה 2.

<u>הנחות האוריינטציה גנרית</u> – אם משוואות המודל הלינארי הן המשוואות המבניות המייצגות את הDAG המייצג את ההתפלגות וd גדול מספיק מתקיים:

i. לוקטור a יש אוריינטציה גנרית לΣ_{XX} במובן ש:

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} \approx \mu^\tau_{\Sigma_{\mathbf{XX}}}\|\mathbf{a}\|^2$$

ii. לוקטור b יש אוריינטציה גנרית לΣ_{EE} במובן ש:

$$\mu_{\Sigma_{\mathbf{EE}},\mathbf{b}} \approx \mu^\tau_{\Sigma_{\mathbf{EE}}}\|\mathbf{b}\|^2$$

iii. הוקטור a גנרי ביחס לb, Σ_{EE} במובן ש:

$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}+c\,\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}} \approx \mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} + \mu_{\Sigma_{\mathbf{XX}},c\,\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}}$$

השיטה:

בניית מידה ספקטרלית טיפוסית עבור ערכי פרמטרים נתונים:

נסקור, בקצרה ככל הניתן מעברים, הגדרות, וסימונים חשובים כדי לבנות את האלגוריתם ולהסביר כיצד פותח.

בסוף חלק זה נראה כי,

$$\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}} \approx \|\hat{\mathbf{a}}\|^2 \nu_{\beta,\eta}$$

כאשר $\nu_{\beta,\eta}$ מורכב משני חלקים אותן נפרט כעת:

<u>החלק הסיבתי</u>- חלק זה מתאר את המידה הספקטרלית בהינתן קשר סיבתי ללא ערפלן כאשר מידה זאת מושרת על ידי a וΣ_{XX}. לפי הנחת האוריינטציה הגנרית של הוקטור a ביחס לΣ_{XX} בהינתן המודל הסיבתי שהגדרנו בתחילת העבודה וd גדול מידה זו ניתנת לקירוב על ידי ההתפלגות האחידה מעל הערכים העצמיים של Σ_{XX}. לכן, כותבי המאמר הגדירו:

$$\nu^{\text{causal}} := \mu^{\tau}_{\Sigma_{\mathbf{XX}}}$$

<u>החלק המערפל</u>- מקרבים את המידה המושרת על ידי Σ_{XX} ו$\Sigma_{XX}^{-1}b$ *נגדיר מטריצה אלכסונית* $M_X = diag(\nu_1^X, \dots, \nu_d^X)$ *כאשר* ν_i^X הם הערכים העצמיים של Σ_{XX} בסדר יורד. כעת מגדירים פרטובציה מדרגה 1 של M_X על ידי:

$$T := M_X + \eta \mathbf{g}\mathbf{g}^T$$

כאשר $g := (1, \dots, 1)^T/\sqrt{d}$.

כעת, מחשבים את המידה הספקטרלית המושרת על ידי הוקטור $T^{-1}g$ וT ומגדירים:

$$\nu_\eta^{\text{confounded}} := \frac{1}{\|T^{-1}\mathbf{g}\|^2}\mu_{T,T^{-1}\mathbf{g}}$$

<u>שילוב החלקים</u>- כעת נגדיר את $\nu_{\beta,\eta}$ כאינטרפולציה לינארית של שני החלקים, כאשר, המשקולות יוגדרו על בסיס עוצמת הערפול β:

$$\nu_{\beta,\eta} := (1-\beta)\nu^{\text{causal}} + \beta\nu_\eta^{\text{confounded}}$$

במשפט הבא נראה שכאשר הנחות האוריינטציה הגנרית מתקיימות עם דיוק מספיק ועבור ממד d גדול מספיק אז $\nu_{\beta,\eta}$ הוא קירוב טוב למידה הספקטרלית המושרית.

<u>משפט:</u> תהה (Σ_{XX}^d) סדרה של מטריצות שונות משותפת שההתפלגות של הערכים העצמיים שלהן מתכנסת באופן חלש למידת הסתברות μ^∞ שהתומך שלה הוא מקטע קומפקטי ב R_0^+. נניח כי נתונות לנו סדרות של וקטורים (a_d) ו(b_d) עם נורמות $\|a_d\| = r_a$ ו $\|b_d\| = r_b$ וc קבוע ככה שהמשוואות ממשפט מספר 1 מתקיימות אז ה$\nu_{\beta,\eta}$ מקרב את $\mu_{\Sigma_{xx,\widehat{a_d}}}$ עד כדי נורמליזציה במובן ש: $\frac{1}{\|\hat{\mathbf{a}}_d\|^2}\mu_{\Sigma^d_{\mathbf{XX}},\hat{\mathbf{a}}_d} - \nu^d_{\beta,\eta} \to 0$ בהסתברות במובן החלש. כאשר $\eta := r_b^2$ וβ ניתן לחישוב בעזרת μ^∞, r_b וc.

האלגוריתם משערך את β כאשר הוא מתבסס על משפט 2 בכך שהוא מוצא את הערך של $\hat{\beta}$ ו $\hat{\eta}$ שמביאים למינימום את המרחק בין $\nu_{\hat{\beta},\hat{\eta}}$ ל $\mu_{\Sigma_{XX,\hat{a}}}$. כותבי המאמר מציינים כי מאחר וההתכנסות המוצגת במשפט 2 היא חלשה מרחקים מסוג l_1 וl_2 לא יתאימו ולכן הם משתמשים בהחלקה עם גרעין גאוסייני ואז משתמשים במרחק l_1 כפי שניתן לראות בנוסחאות הבאות:

$$D(w, w') := \|K(w - w')\|_1$$

$$K(i, j) := e^{-\frac{(\nu_i^X - \nu_j^X)^2}{2\sigma^2}}$$

האלגוריתם שכותבי המאמר מציעים הוא:

Algorithm 1 Estimating the strength of confounding

1: **Input:** I.i.d. samples from $P(\mathbf{X}, Y)$.
2: Compute the empirical covariance matrices $\Sigma_{\mathbf{XX}}$ and $\Sigma_{\mathbf{X}Y}$
3: Compute the regression vector $\hat{\mathbf{a}} := \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}Y}$
4: PHASE 1: Compute the spectral measure $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$
5: Compute eigenvalues $\nu_1^X > \cdots > \nu_d^X$ and the corresponding eigenvectors $\phi_1, \ldots, \phi_d$ of $\Sigma_{\mathbf{XX}}$
6: Compute the weights $w'_j = \langle \hat{\mathbf{a}}, \phi_j \rangle^2$ and then the normalized weights $w_j := w'_j / \sum_j w'_j$.
7: PHASE 2: find the parameter values $\hat{\beta}, \hat{\eta}$ that minimize the distance $D(w, w^{\beta,\eta})$ with D defined by eq. (26), where $w^{\beta,\eta}$ denotes the weight vector of the measure $\nu_{\beta,\eta}$.
8: **Output:** Estimated confounding strength $\hat{\beta}$

כאשר את $w^{\beta,\eta}$ מחשבים על ידי חישוב המטריצה T לפי הנוסחה שהגדרנו לעיל וחישוב הוקטורים העצמיים שלה. לאחר מכן, נחשב את $v = T^{-1}g/\|T^{-1}g\|$. ריבוע השונות המשותפת של v והוקטורים העצמיים של T מתאר את המשקולות של $\nu_\eta^{confounded}$. כדי לחשב את $\nu_{\beta,\eta}$ נוסיף את התרומה של ν^{causal} ונקבל את המשקולות: $w_j^{\beta,\eta} := \frac{1}{d}(1-\beta) + \beta\langle v, \phi_j\rangle^2$

ניסויים:

לאחר שקראנו את המאמר והבנו את השיטה המוצגת בו בחרנו להתעמק בתוצאות השיטה ולנסות ולבחון מספר הערות שהעירו החוקרים לאורך המאמר וביצענו שחזור של ניסויים נבחרים. ביצענו את הניסויים בעזרת שפת התכנות python וניתן למצוא את הקוד ב https://github.com/Plozel/ci_project . חלק מהערות שנתקלנו בהן במהלך הקריאה של המאמר התייחסו לנכונות אמפירית ללא הוכחה או ניסוי מאושש.

הערות שבחרנו להתעמק בהן:

- במאמר מציינים כי נרמול מטריצת הנתונים מפר את הנחות האי תלות של המודל אך אינו פוגע אמפירית בתוצאות השיטה כפי שניתן לראות בתוצאות הניסויים שבוצעו בנתוני אמת מנורמלים (חשוב לציין שבנתונים אלו הערך האמיתי לא ידוע ולכן לא בוצעה בחינה של הביצועים באופן מדויק אלה רק בדיקת נראות).
- במאמר מציעים בחינת שימוש בגורם regularization בחישוב האומד ל$\hat{a}$ למרות שהדבר מפר את ההנחות התאורטיות שבבסיס חישוב $\hat{a}$.
- כותבי המאמר מעלים אפשרות להשתמש בשיטות kernel כדי להתמודד עם מודלים שאינם לינאריים.

<u>שחזור ניסויים:</u>

<u>שחזור ניסוי 6.1 לאמידת עוצמת הערפול:</u>

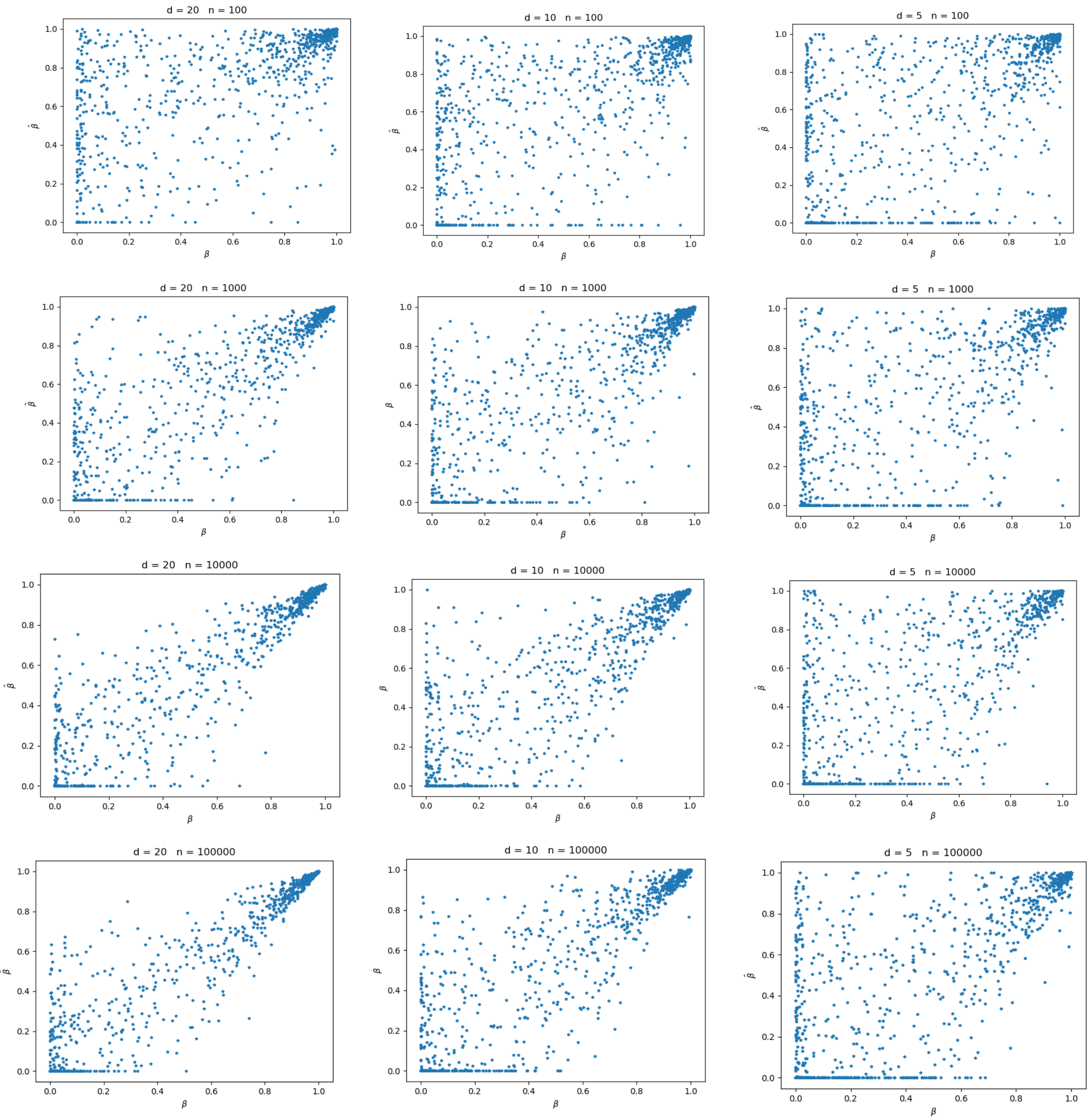
בניסוי זה כותבי המאמר ערכו סימולציה בה יצרו את המידע לפי הנחות המודל.

שלבי הסימולציה:

- ייצור E: יוצרים n דוגמאות של וקטור d ממדי עם ערכים נורמליים רנדומליים $\tilde{E}$ עם תוחלת 0 ומטריצת I *covariance*. יוצרים מטריצת $d \times d$ רנדומלית G שערכיה הם רנדומליים נורמליים בלתי תלויים וקובעים $E = G\tilde{E}$.
- יוצרים סקלרים רנדומליים של משתני Z וF על ידי הגרלת n דוגמאות של כל אחד באופן בלתי תלוי אחד בשני מהתפלגות נורמלית סטנדרטית.
- יוצרים את פרמטרי המודל c, r_a, r_b על ידי הגרלות מהתפלגות אחידה על אינטרוול היחידה.
- מגרילים וקטורים a, b באופן בלתי תלוי מכדורים ברדיוסים r_a וr_b, בהתאמה.
- מחשבים את X וY על ידי המודל הלינארי שתואר לעיל.

נשים לב כי בתהליך ייצור הנתונים כל הפרמטרים ידועים ולכן נוכל לחשב את β במדויק.

תוצאות שחזור ניסוי 6.1:



כפי שניתן לראות, אכן התוצאות שלנו דומות לתוצאות המאמר. השיטה היינה שיטה סטטיסטית שעובדת על נתונים רנדומליים ולכן לא נצפה לקבל תוצאות זהות. תוצאות אלו מחזקות את אמונתנו בנכונות התרגום שלנו לאלגוריתם, כמו כן, הן ישמשו לבחינת סבירות תוצאות אחרות. בשחזור הניסוי הראנו שאכן התוצאות של הניסוי הינן ניתנות לשחזור וכי לא מדובר בתוצאות מקריות.

<u>שחזור ניסוי 8.1 על טעם היין:</u>

בניסוי זה לקחו כותבי המאמר נתונים מתחום בו יש להם אמונה מקדימה על הקשרים הסיבתיים בין המשתנים ועליה התבססו בבחינת התוצאות. לכן, לא ניתן לדעת בוודאות האם המודל עמד נכון את עוצמת הערפול ויש השענות רבה על אמונת החוקרים עם זאת אין לכך השפעה על תועלת השחזור לבדיקת השינוי בפיזור $\mu_{\Sigma_{\{XX\}},\hat{a}}$.

את הנתונים הורדנו מהמקור המצוין במאמר ונרמלנו כפי שכותבי המאמר נרמלו.

התוצאות מן המאמר:

$$\hat{\mathbf{a}} = (0.044, -0.194, -0.036, 0.023, -0.088, 0.046, -0.107, -0.034, -0.064, 0.155, 0.294)$$

ו$\hat{\beta} = 0.0$

תוצאות השחזור:

$$\hat{a} = (0.043, -0.194, -0.0355, 0.023, -0.088, 0.045, -0.107, -0.033, -0.063 \;\; 0.155 \;\; 0.294)$$

ו$\hat{\beta} = 0.0$

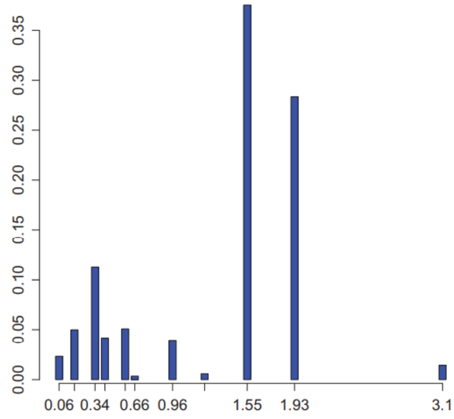
ניתן לראות כי כותבי המאמר ביצעו עיגול שאנו לא ביצענו ולכן קיבלו תוצאות מעט שונות (ערכי $\hat{a}$ שלנו נחתכו לשלוש ספרות אחרי הנקודה ולא עוגלו). אנו נראה כי חוסר עיגול זה מביא להבדל נוסף בתוצאות בהמשך.

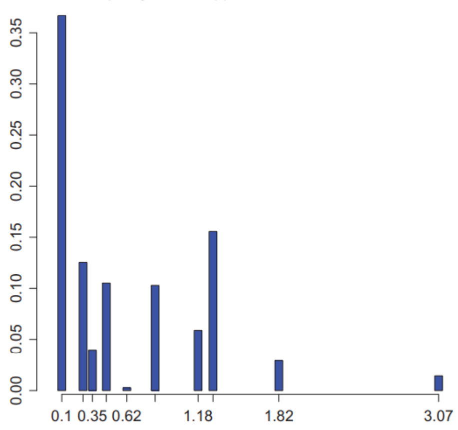
כדי לראות אם אכן ניתן למצוא נוכחות ערפלן בעזרת שיטה זו על נתונים מן המציאות הכותבים החליטו להשמיט משתנה יחיד בכל פעם ולחשב $\hat{\beta}$ בלעדיו. בעבור כל משתנה שאינו "רמת האלכוהול" (משתנה 11) התקבל כי $\hat{\beta} = 0.0$ (את הניסיונות הללו לא ניסינו לשחזר שכן מצענו אותם פחות מעניינים). בעבור "רמת האלכוהול" התקבל $\hat{\beta} = 0.55$ תוצאה זו הגיונית שכן משתנה זה קיבל גם את המשקל החזק ביותר בהשפעת X על Y וכן כותבי המאמר ציינו כי על בסיס אמונה רווחת סביר כי ישנו קשר סיבתי בין "רמת האלכוהול" לאיכות היין. כמו כן, למשתנה זה ערכי מתאם גבוהים עם שאר המשתנים. ועל כן, הסרת "רמת האלכוהול" מן הנתונים ייצרה באופן מלאכותי ערפלן נסתר שהשיטה זיהתה בהצלחה.

התוצאות שלנו הן: $\hat{\beta} = 0.52$ אנו מאמינים כי המקור להבדל הוא בעיגול שמבוצע במאמר.

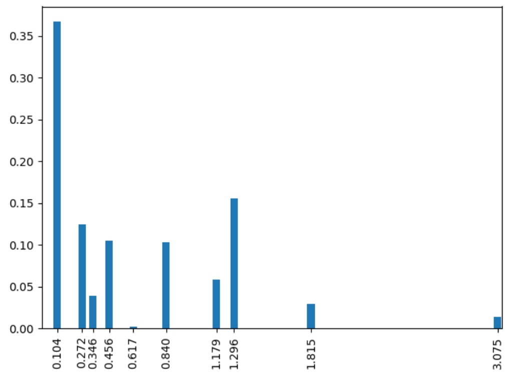
במקרה זה הוצג במאמר השינוי ב$\mu_{\Sigma_{XX},\hat{a}}$ לפי הערכים העצמיים השונים ואכן ניתן לראות כי כפי שמצוין במאמר נוכחות ערפלן מסיטה את המידה לערכים העצמיים הקיצוניים הקטנים.

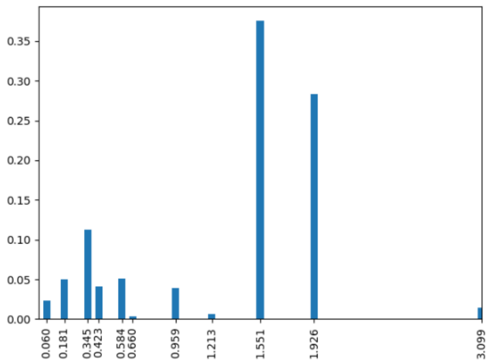
תוצאות המאמר:





מאחר וגרפים אלו מציגים עקרון בסיסי מאוד בדרך עבודת השיטה היה חשוב לנו לראות שנוכל לשחזר אותם.
התוצאות השחזור שלנו הן:





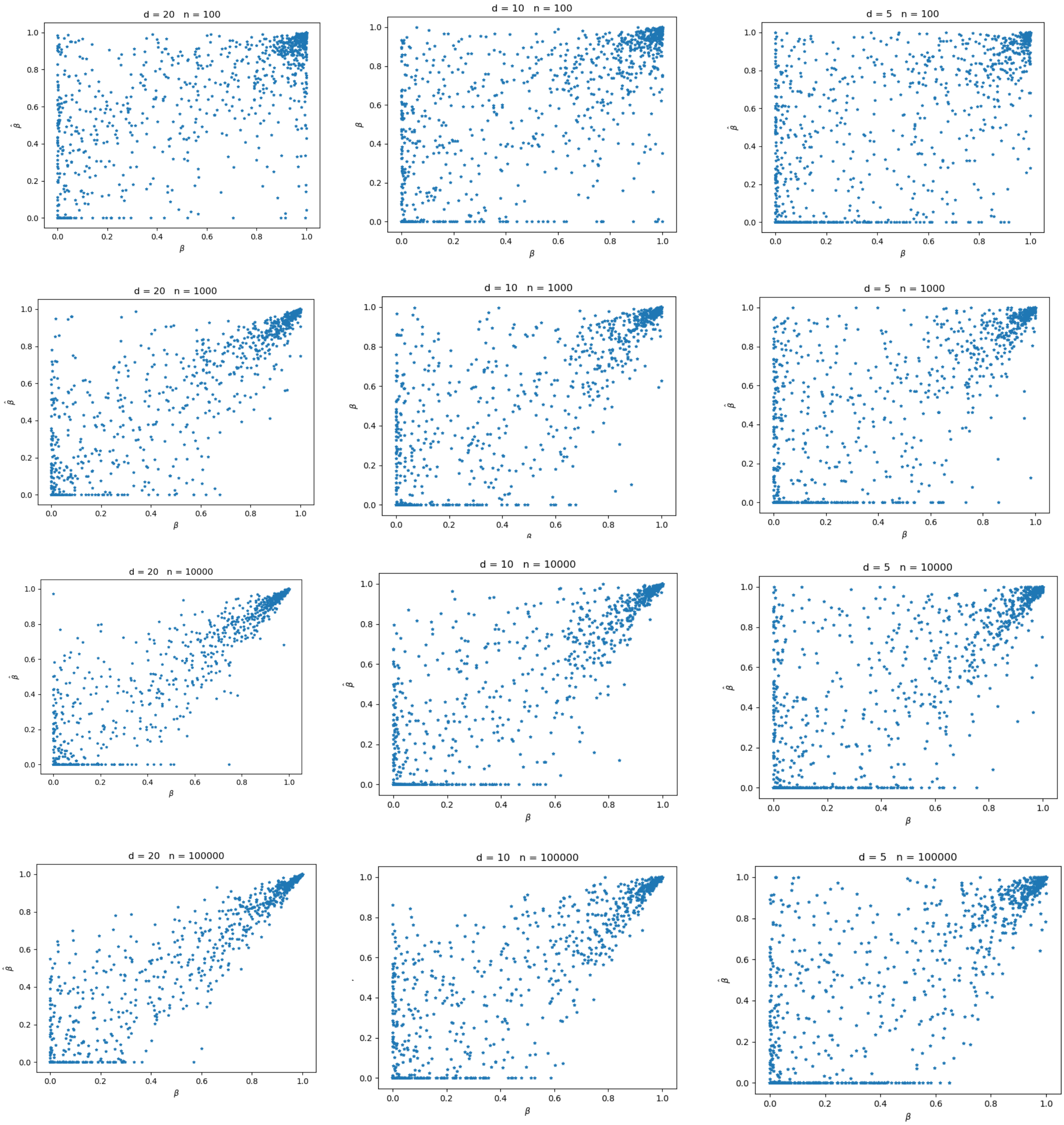
אכן, ניתן לראות כי תוצאות אלו קרובות לתוצאות המאמר. גם השחזור הנ"ל מחזק את ההערכתנו כי מימושנו לשיטה אכן נכון ועובד כפי שהתכוונו מחברי המאמר.

<u>ניסויים שמבוססים על הערות מן המאמר:</u>

<u>ניסוי 1: השפעת *regularization*:</u>

בניסוי זה אנו מעוניינים לבחון את ההשפעה של שימוש ב*regularization* על תוצאות השיטה לאמידת עוצמת הערפול המוצגת במאמר. כותבי המאמר מציינים כי שימוש ב*regularization* לא נבדק ויכול להיות כיוון מחקר עתידי. כמו כן, הכותבים מציינים כי יש לבחור בשיטת *regularization* שתשמר את הנחת הסימטריה מהמאמר.
בניית הניסוי:
כדי שנוכל לבחון את השפעת ה*regularization* לבדו שימרנו את בניית הנתונים והמודל מהשחזור שביצענו לניסוי 6.1. למעשה השינוי היחיד שעשינו הוא הוספת *regularization* לאמידת $\hat{a}$. בחרנו ב*Ridge* בשל הסימטריה שלו ומימוש נוח יחסית שלו ב*sklearn* נציג תוצאות עם פרמטר $\lambda = 0.5$ שנתן תוצאות דומות לערכים אחרים שבדקנו.
השערה: אנו משערים כי מאחר ו*regularization* אמור לעזור להתעלם מהרעש ולקרב יותר נכון את המקדמים אנו מאמינים כי שימוש בו יחזק את השיטה וייצר גרפים קרובים יותר לליניאריים.

תוצאות:

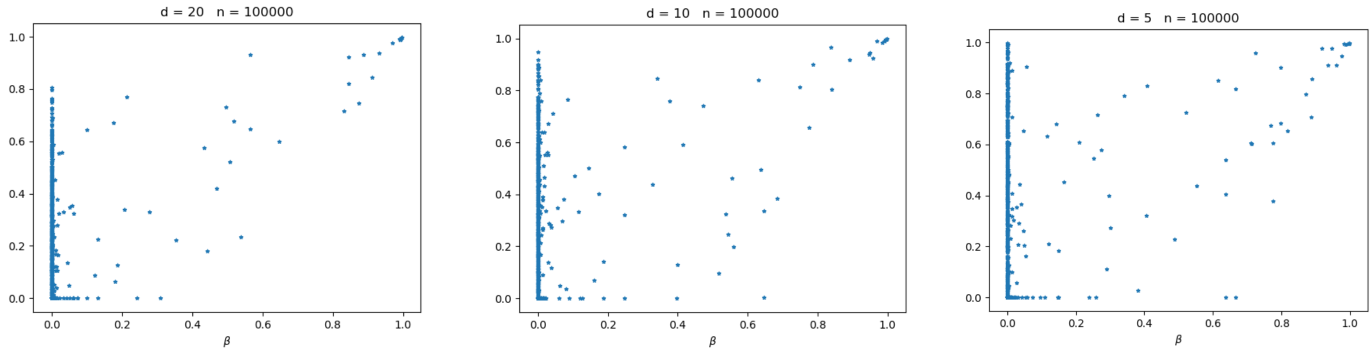


ניתן לראות כי אכן מתקבלות תוצאות טובות אם כי ההשוואה בין התוצאות עם וללא *regularization* היא מעט בעייתית שכן התוצר שלנו לא נומרי בניסוי זה. עם זאת אנו חושבים כי אכן רואים שיפור מסוים בפיזור ולכל הפחות לא נראה כי קיימת פגיעה בביצועי השיטה.

הערה: מכאן והלאה לא נציג את כל הגרפים אלה רק את אלו בהם $n = 100000$ כי הקשר בין הגרפים זהה בכל הניסויים ועבור ערך n זה מתקבלות התוצאות הטובות ביותר. נציג את שלושת ערכי הd כי במאמר ניתן משקל מיוחד למספר הממדים (ככל שמספר הממדים גדול יותר השיטה, תאורטית, מדויקת יותר) ולכן נרצה להדגיש את השפעת ערך זה.

ניסוי 2.1: שינוי G בלבד:

בניסוי זה אנו מעוניינים לבדוק את ההשפעה של שינוי Σ_{EE} על ידי שינוי G בכך למעשה אנו משנים את Σ_{XX} וכן משנים את משקל ההשפעה של הערפול Z על X. במאמר מציינים כי Σ_{EE} הינו פרמטר של השיטה ועל כן ניתן לשנות אותו כרצוננו. לא נבדק לאורך המאמר האם לשינויים בפרמטר הנ"ל יש השפעה על תוצאות השיטה. כדי לבחון זאת בנינו את הנתונים באופן זהה לניסוי 6.1 כאשר את הכניסות בעמודות של G דגמנו באופן בלתי תלוי מהתפלגות נורמלית עם $variance$ משתנה במקום באופן בלתי תלוי מהתפלגות נורמלית סטנדרטית.
השערה: מאחר וG היינו משתנה של המודל אנו לא מאמינים כי שינוי שלו אמור להשפיע על התוצאות באופן דרסטי.
תוצאות:



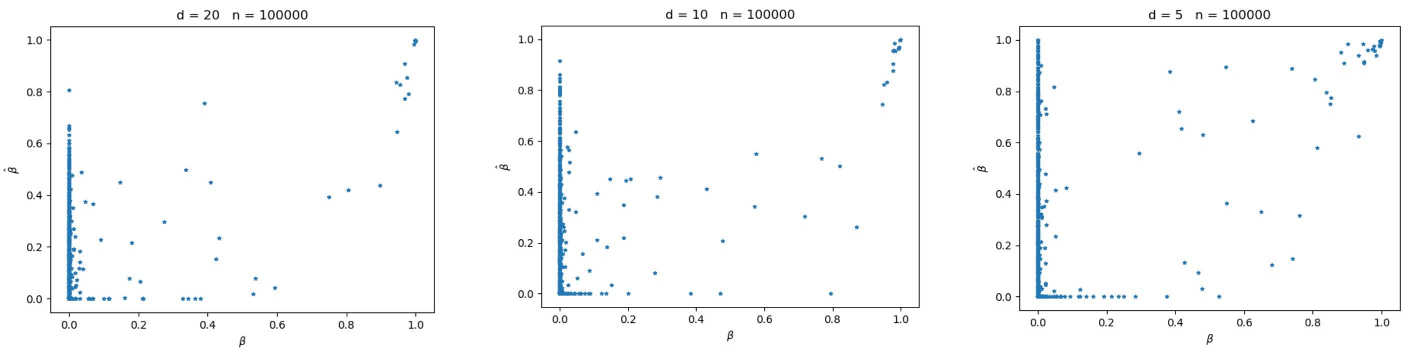
נראה כי אף על פי שהשיטה עדיין עובדת התוצאות גרועות יותר מהמצב של שונות אחידה. תוצאות אלו לא תואמות את ההשערה שלנו. קראנו שנית את דוגמאות הקצה שדיברו על כיצד הערפלן גורם לתיאום בין וקטור המשקולות לוקטורים העצמיים הקיצוניים. מאחר וההפרשים בין הערכים העצמיים גדולים בהרבה מ$||b||$ לפי מקרי הקצה נצפה לתאימות בין המשקולות לוקטורים העצמיים הקטנים יותר כאשר β גדול. אכן, ניתן לראות כי, כאשר β גדול השיטה עובדת טוב למדי. אך, בשל ההשפעה הקטנה יחסית של b על הערכים העצמיים של X מתקבל כי β לרוב קטן למעשה לרוב מתקבל $\beta = 0.0$. עובדה זו גרמה לנו לבחון את תוצאות השיטה בעבור ערכי β קטנים בסימולציה המקורית ושמנו לב כי לשיטה המוצגת במאמר קשה יותר לשערך את β ככל שהנ"ל קטן יותר באופן כללי. אנו מאמינים כי הסיבה לפגיעה בתוצאות נעוצה בחולשה זו של השיטה ונטיית השונות המשתנה לייצר ערכי β קטנים.

ניסוי 2.2: נורמליזציה על X שנוצר מG:

כאשר נשנה את G נקבל עמודות בסדרי גודל שונים, במצב כזה לרוב נבצע נורמליזציה. לכן, בחנו את התוצאות עם G כפי שקבענו בניסוי הקודם אך בתוספת נורמליזציה.

ההשערה שלנו הייתה כי הנורמליזציה תשחית את התוצאות מאחר כי היא תקטין את ההשפעה של b על כל משתנה בפקטור שונה מה שישנה את β.

תוצאות:



כפי שניתן לראות, שנית שגינו בהשערתנו, נראה כי הנרמול שיפר את התוצאות בהשוואה ל2.1 ועזר לשערך יותר במדויק ערכי β נמוכים. נראה כי, הגודל היחסי של התרומה של b הוא מה שמשפיע על התוצאה. כמו כן, מאחר וערכים עצמיים לוקחים חלק בשערוך התאימות בין הוקטורים העצמיים לוקטור המשקולות $\hat{a}$, ערכים עצמיים שהינם בסדרי גודל שונים מטים את ההערכה. לכן, נרמול והבאת הערכים העצמיים לאותם סדרי גודל משפר את ההערכה כפי שניתן לראות בתוצאות.

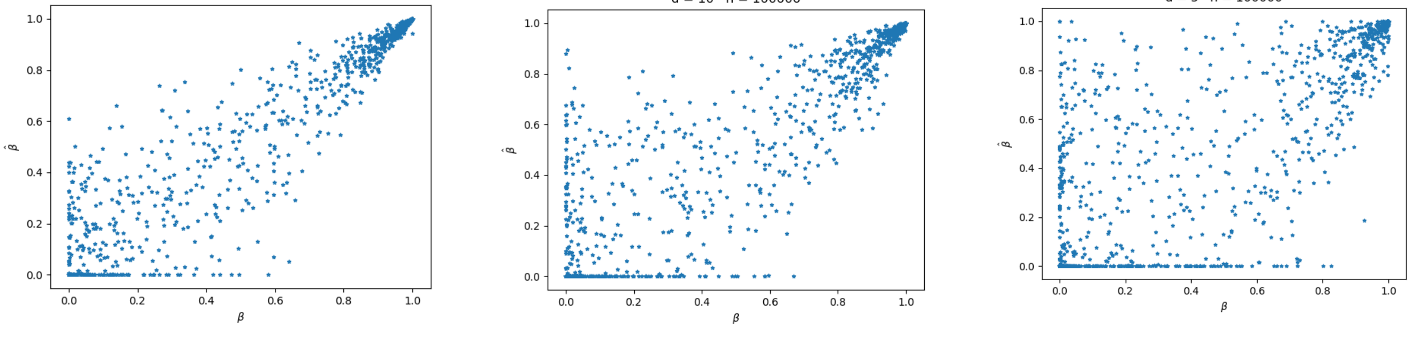
ניסוי 3: "המרת יחידות" + נורמליזציה:

בשל התוצאות שלנו עניין אותנו לבדוק מצב בוא הנתונים נוצרים מראש לפי הנחות המודל (כפי שנוצרו בניסוי 6.1) אך, מתקבלים ביחידות אחרות. במקרה זה ההשפעה של b למעשה קבועה ולא משתנה כאשר מגדילים את השונות של המשתנים ולכן גם ערכי β לא משתנים.

כדי למדל זאת בנינו את X כפי שנבנה בניסוי 6.1 והכפלנו אותו במטריצה אלכסונית $scale$ לשינוי "יחידות"

ההשערה שלנו לניסוי זה היא כי נקבל תוצאה דומה לסימולציה המקורית (ניסוי 6.1) ואולי אף נראה שיפור בשל נרמול השונות.

תוצאות:



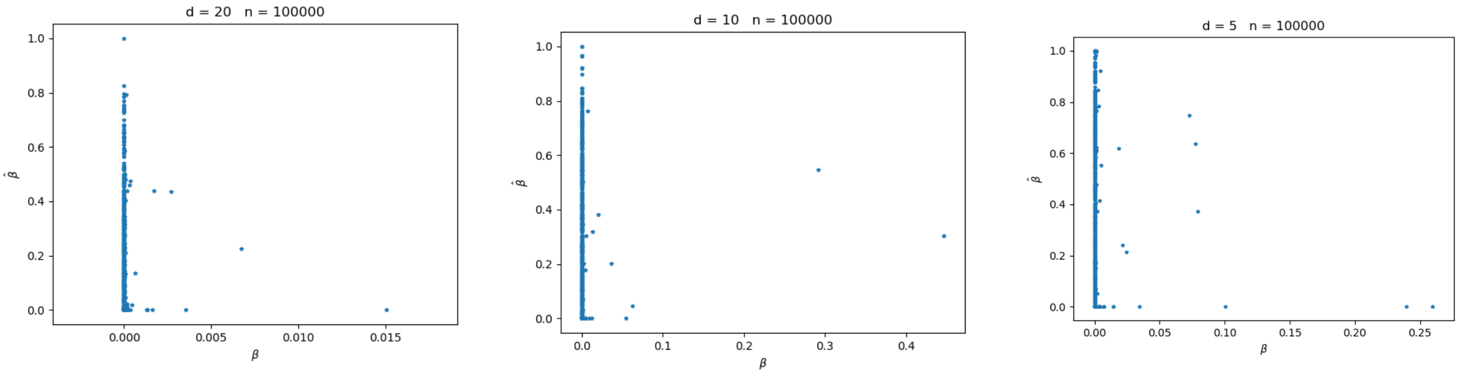
נראה כי השערתנו נכונה במקרה זה. אכן קיבלנו תוצאות דומות לתוצאות השחזור של ניסוי 6.1. עובדה זו מחזקת את הערת כותבי המאמר כי נראה כי אמפירית שימוש בנורמליזציה לא מזיק לתוצאות. תוצאות אלו לא מפתיעות כיוון שצורת הנרמול שבחרנו (חלוקה במטריצת ה$covariance$) כוללת הכפלה במטריצה ההופכית של ה$scale$ ולכן, סביר כי אין לה השפעה דרמטית על התוצאות.

<u>ניסוי 4: שימוש בשיטות $kernel$ כדי להתחמק מהנחת ה$linearity$:</u>

בניסוי זה בדקנו האם אנו מסוגלים להשתמש בשיטות $kernel$ כדי לעקוף את הנחת הלינאריות. הסתמכנו על הערה שנתנו מחברי המאמר בו ציינו כי למיטב ידיעתם ניתן להשתמש בשיטות $kernel$ כדי לעקוף את הנחת הלינאריות. בחרנו להשתמש ב$kernel$ ריבועי העלנו כל כניסה בX בחזקת 2. מאחר והחזקה משפיעה על המודל לא נוכל לחשב את β במדויק ונאלץ לשערך אותו על ידי $\beta' = \Sigma_{XX}^{-1} cov(X, cZ)$.

ההשערה שלנו לניסוי זה היא שנצליח להשתמש בשיטות $kernel$ ולשמר את איכות התוצאות.

תוצאות:



לאור התוצאות אנו לא בטוחים כי אכן שילבנו את השימוש ב$kernel$ בצורה נכונה. הגורם שגרם לנו לפקפק בתוצאותינו הוא שאיפת המודל לשערך את β כ0 ככל שגדל מספר הממדים והדוגמאות. אנו לא מוצאים סיבה סבירה לשערוך זה. לכן, בחנו מספר שיטות שונות לשערוך β ולמציאת המקום המתאים לשימוש ב$kernel$ (על X על E רק בחישוב Y וכו') ללא הצלחה משמעותית בהבנת המקור להתנהגות זו של β.

<u>סיכום וכיווני המשך:</u>

בעבודה זו, סקרנו את המאמר Detecting Confounding in Multivariate Linear Models via Spectral Analysis המציע שיטה לזיהוי נוכחות ערפלן חבוי במודל לינארי תחת הנחות שונות על בסיס אנליזה ספקטרלית דרך הגדרת אוריינטציה גנרית. אנו, בחרנו לשחזר שני ניסויים משמעותיים מן המאמר ולבחון בניסויים הערות וכיווני המשך שנתנו על ידי מחברי המאמר. אכן, בעזרת השחזור אימתנו את מהימנות הניסויים והגענו לתוצאות מעניינות בנוגע להתנהגות השיטה המוצגת במאמר. בין היתר תוצאות הניסויים כללו:

- ראיה לכך ששימוש בregularization אינו פוגע ביעילות השיטה.
- ראיה לכך שהשיטה רגישה להגדלה משמעותית בשונות עמודות במטריצה G.

- ראיה לכך שכאשר השונות גדולה ולא אחידה שימוש בנורמליזציה משפר את יעילות השיטה.
- השיטה טובה יותר בזיהוי ושיערוך מקורב של ערכי β גדולים וחלשה יותר עבור ערכי β קטנים.

לאורך כל העבודה הנוכחית השתמשנו בגרף פיזור ערכי $(\beta, \hat{\beta})$ כמדד לאיכות הביצועים. עובדה זו הקשתה על השוואת תוצאות הניסויים בגרסאות השונות. אי לכך, היינו שמחים לראות מדד נומרי לאיכות התוצאות שיהיה קל יותר להשוואה במקרים בהם ההבדלים בין התוצאות הגרפיות לא בולטים לעין.

כמו כן, מאחר ואיננו בטוחים בתוצאות ניסוי 4 נשמח לראות ניסויים ועבודות המשך הנוגעות ומתעמקות בנושא. תוצאות הניסויים שעשינו מעידות כי השיטה עובדת אך רגישה לשינויים בהתפלגות הנתונים ועל כן, דעתנו שימוש בשיטה זו על נתונים אמיתיים צריך להילקח בערבון מוגבל. עם זאת, השיטה והמאמר מהווים בסיס תאורטי נרחב לזיהוי ערפלנים ואנו מאמינים שניתן יהיה לפתח שיטות נוספות על בסיס זה.

מבואות

[1] Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M., “Estimation of causal effects using linear non-Gaussian causal models with hidden variables”, *International Journal of Approximate Reasoning*, pp. 49(2), 362-378, 2008.

[2] Janzing, D., Peters, J., Mooij, J., & Schölkopf, B, “Identifying confounders using additive noise models”, 2012.

[3] Janzing, D., Sgouritsa, E., Stegle, O., Peters, J., & Schölkopf, B., “Detecting low-complexity unobserved causes”, 2012.

[4] Janzing D, Schölkopf B, “Causal inference using the algorithmic Markov condition”, *IEEE Trans Inf Theo*, p. 56:5168–5194, 2010.

[5] Lemeire J, Janzing D, “Replacing causal faithfulness with algorithmic independence of conditionals”, *Minds Mach*, p. 227–249, 2012.

[6] Reed M, Simon B., “Functional Analysis”, *San Diego, California: Academic Press*, 1980.