רגרסיה ומודלים סטטיסטיים- תרגיל 2 תשפ"ה

שאלה 1- אלגברה לינארית:

א. יהיו $A, B \in R^{n\times n}$. תזכורת: $tr(A) \coloneqq \sum_{i=1}^{n} A_{ii}$
הוכיחו או הפריכו את התכונות הבאות:

1) $(AB)^T = B^T A^T$.

2) $(A+B)^T = A^T + B^T$

3) $tr(AB) = tr(A)tr(B)$

4) $tr(AB) = tr(BA)$

5) $tr(A+B) = tr(A) + tr(B)$

אם A, B הפיכות:
6) $(AB)^{-1} = B^{-1}A^{-1}$

7) $(A+B)^{-1} = A^{-1} + B^{-1}$

ב. הוכיחו כי $Ker(A) = \{v | v \in R^n, Av = 0\}$ ו-$IM(A) = \{w | w \in R^n, \exists v \in R^n | \ Av = w\}$ הם תתי מרחבים לינאריים.

ג. עבור $A \in R^{n\times p}$ הוכיחו כי $IM(A^T) = Ker(A)^{\perp}$.

פתרון:

א.

(1) $B^T A^T_{ij} = \sum_k B^T_{ik} A^T_{kj} = \sum_k B_{ki} A_{jk} = \sum_k A_{jk} B_{ki} = (AB)_{ji} = (AB)^T_{ij}$
(2) $(A+B)^T_{ij} = (A+B)_{ji} = A_{ji} + B_{ji} = A^T_{ij} + B^T_{ij}$
(3) קחו את המטריצות $A = diag(2,2,2)$, $B = diag(0.5,0.5,0.5)$. אז $AB = I$.
$tr(AB) = 3 \neq tr(A)tr(B) = 6 \cdot 1.5 = 9$. בסתירה.
(4) $tr(AB) = \sum_i \sum_k A_{ik} B_{ki} = \sum_k \sum_i B_{ki} A_{ik} = tr(BA)$
(5) $tr(A+B) = \sum_i (A+B)_{ii} = \sum_i A_{ii} + \sum_i B_{ii} = tr(A) + tr(B)$

(6) $ABB^{-1}A^{-1} = AIA^{-1} = AA^{-1} = I$
(7) קחו $A = B = I$ אז שתיהן הפיכות וגם ההופכיות שוות לעצמן,
$A + B = diag(2,2 \dots) = A^{-1} + B^{-1}$ וההופכית שלה היא $diag(0.5, \dots, 0.5)$. בסתירה.

ב. ניקח $k_1, k_2 \in R$,$v_1, v_2 \in Ker(A)$ אז:

$$A(k_1 v_1 + k_2 v_2) = k_1 A v_1 + k_2 A_2 = k_1 \cdot 0 + k_2 \cdot 0 = 0$$

כאשר השיוויון השלישי נובע מכך ש-$Av_1 = Av_2 = 0$ כיוון שבגרעין. ולכן $Ker(A)$ הוא תת מרחב (כמובן שוקטור האפס הוא הפתרון הטריוויאלי למערכת המשוואות ההומוגנית $Av = 0$ ולכן גם הוא בגרעין).

ניקח $w_1, w_2 \in IM(A)$. אז קיימים $u_1, u_2 \in R^n$ כך ש-$Au_1 = w_1, Au_2 = w_2$. נרצה להראות שקיים $u \in R^n$ כך ש-$Au = k_1 w_1 + k_2 w_2$.

נזכור של מטריצה מייצגת את ההעתקה הלינארית שהיא הכפלה בעצמה (ביחס לבסיס הסטנדרטי) ולכן מדובר בהעתקה לינארית. ואכן אם נבחר

$u = k_1u_1 + k_2u_2$

נקבל: $Au = k_1Au_1 + k_2Au_2 = k_1w_1 + k_2w_2$. ולכן גם התמונה היא תת מרחב לינארי.

ג. נניח $v \in IM(A^T)$. אז $v = A^Tw$ עבור $w \in R^n$. ניקח $u \in Ker(A)$. נקבל: $v^Tu = w^TAu = 0$ ולכן $v \in Ker(A)^\perp$.

מצד שני, על פי משפט המימדים:

$$n = \dim\big(Ker(A)\big) + \dim\big(IM(A)\big) = \dim\big(Ker(A)\big) + \dim\big(IM(A^T)\big)$$

כאשר השיוויון השני נובע מכך שדרגת השורות שווה לדרגת העמודות. כעת ממשפט הסכום הישר נקבל:

$$n = \dim(Ker(A)) + \dim(Ker(A)^\perp) \Rightarrow \dim(Ker(A)^\perp) = \dim\big(IM(A^T)\big)$$

מסעיף ב׳ מדובר בתתי מרחבים ומהחלק הראשון $IM(A^T) \subseteq Ker(A)^\perp$. מהמשפט שראינו בתגבור- אם V,U תתי מרחבים בעלי מימדים שווים ומתקיים $V \subseteq U$ אז $V = U$- נובע כי $IM(A^T) = Ker(A)^\perp$.

שאלה 2- מטריצת הטלה:

מטריצה איידמפוטנטית היא מטריצה $A \in R^{n\times n}$ שדרגתה r המקיימת $A = A^2$.

מטריצה סימטרית ואיידמפוטנטית נקראת **מטריצת הטלה אורתוגונלית.**

א. הוכיחו כי הע״ע של מטריצת הטלה הם 1, בריבוי כדרגת המטריצה, ו-0 בריבוי השווה למימד של גרעין המטריצה.
ב. תהי $X \in R^{n\times p}$ מטריצה מדרגה מלאה ונגדיר $P_X = X(X^TX)^{-1}X^T$. הראו כי P_X היא מטריצת הטלה למרחב הנפרש על ידי העמודות של X. כלומר:

P_X סימטרית, איידמפוטנטית ומתקיים שלכל $v \in R^n$: $P_Xv \in IM(X)$.

ג. מצאו את $trace(P_X)$.

פתרון:

א. יהי λ ע״ע של A, אז עבור $v \in R^n$ שונה מ-0:
$Av = \lambda v = A^2v = A\lambda v = \lambda Av = \lambda^2 v \Leftrightarrow \lambda(1-\lambda)v = 0$ כלומר $\lambda \in \{0,1\}$.
כעת, עבור $\lambda = 0$, מתקיים: $Av = 0$. כלומר הווקטורים אשר פותרים משוואה זו אלו הוקטורים העצמיים של המטריצה, ומצד שני אלו גם הוקטורים השייכים ל-$Ker(A)$.
ממשפט המימדים יש בדיוק $n-r$ כאלו. מכאן שישנם r וקטורים עצמיים השייכים לע״ע 1.

ב. הוכחה:
(1) $P_X^T = (X(X^TX)^{-1}X^T) = X^{T^T}(X^TX)^{-1^T}X^T = X(X^TX)^{T^{-1}}X^T = X(X^TX)^{-1}X^T = P_X$
(2) $P_X^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = P_X$
(3) $P_Xv = X(X^TX)^{-1}X^Tv$. נסמן $\tilde{v} := (X^TX)^{-1}X^Tv$. אז: $X\tilde{v} = \sum_{i=1}^{p} X_i\tilde{v}_i$ וזהו צירוף לינארי של עמודות X ולכן בתמונה.

ג. כפי שראינו בתרגול $A \in R^{p\times p} = \sum_i \lambda_i$ עבור מטריצה לכסינה ש-λ_i הוא הע"ע שלה. כיוון ש-P_X סימטרית אז בהכרח לכסינה. מסעיף א' נקבל $trace(P_X) = r$.

שאלה 3- יישומים של ליכסון אורתוגונלי:

<u>חלק 1</u>

תהי $A \in R^{n\times p}$ מטריצה כלשהי.

א. הראו של- AA^T ולַ- A^TA יש את אותם הע"ע. (תזכורת: בתרגיל הקודם הוכחתם שהם גם אי שליליים).
ב. השתמשו בכך כדי להראות שניתן לכתוב כל מטריצה $A = USV^T$ עבור U ו-V מטריצות ריבועיות מהממדים המתאימים ו-$S \in R^{n\times p}$ כך ש-$S_{ii} \geq 0$ וכן $S_{ij} = S_{ji} = 0$ לכל $i \neq j$. פירוק זה נקרא פירוק ה- SVD של A.

הדרכה:

א. כתבו $AA^T = U\Lambda U^T$. מדוע קיים הפירוק הזה? כעת על פי הסעיף הקודם, בחרו את V המתאימה. מדוע היא אורתוגונלית?
ב. השתמשו בכך וכתבו $V^TV = I$ כדי לקבל $U\Lambda U^T = USV^T(USV^T)^T$. מי אלו איברי S במונחי איברי המטריצה Λ? רמז: נמקו מדוע איברי Λ אי שליליים.
ג. איך ניתן לבטא את העמודה ה-i של U באמצעות העמודה ה-i של V, A ואיברי S? הסיקו כי ניתן לכתוב $AV = US$ ומשם מצאו את הפירוק של A .

אם אתם משתמשים בהדרכה עליכם לענות על השאלות בדרך.

<u>חלק 2</u>

א. בעבור מטריצה $X \in R^{n\times p}$ נגדיר:
הגדרה ראשונה[1]:

$$PC_1^{var} = \underset{\substack{w\in R^p \\ ||w||=1}}{\operatorname{argmax}} \sum_{i=1}^{n} \left|w^Tx_i\right|^2 = \underset{\substack{w\in R^p \\ ||w||=1}}{\operatorname{argmax}} \left||Xw|\right|^2 = \underset{\substack{w\in R^p \\ ||w||=1}}{\operatorname{argmax}} w^TX^TXw$$

הגדרה שניה - שגיאת ריבועים פחותים מינימאלית

$$\mathrm{PC}_1^{\mathrm{LS}} = \underset{\substack{w\in\mathbb{R}^p \\ \|w\|_2=1}}{\arg\min} \sum_{i=1}^{n} \mathrm{dist}(x_i, w)^2,$$

כאשר

$$\mathrm{dist}(x_i, w) = \|x_i - P_w(x_i)\|_2$$

ו $P_w(x)$ הוא ההיטל האורתוגונלי של הנק' x על תת המרחב הנפרש על ידי הוקטור w.

הניחו כי $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. הראו ש-$PC_1^{var} = v_1$ כאשר u_1 הוא העמודה הראשונה של U מהפירוק הספקטרלי: $X^TX = U\Lambda U^T$.

ב. הראו כי עבור כל וקטור w כך ש $\|w\|_2 = 1$ מתקיים

$$\sum_{i=1}^{n} \mathrm{dist}(x_i, w)^2 = \|X\|_{\mathrm{F}}^2 - \sum_{i=1}^{n} |w^Tx_i|^2.$$

---

[1] בהמשך נראה שתחת הנחות מסויימות, מדובר בוקטור המנורמל ששונותו היא הגבוהה ביותר מבין כל הוקטורים המכפילים מימין את מטריצת הנתונים.

והסיקו כי $PC_1^{var} = PC_1^{LS}$. (הוא נקרא "המרכיב הראשי הראשון" של A, והתהליך למציאתו נקרא PCA עליו תרחיבו בהמשך התואר).

הערה: פירוק ה-SVD שימושי מאוד בסטטיסטיקה, עיבוד תמונה ולמידת מכונה, ואף בבעיות רגרסיה כאשר עמודות X תלויות לינארית.
תוכלו לקרוא על חלק מהשימושים [כאן](#) ו-[כאן](#).

פתרון:
נניח בה"כ כי $n \geq p$.

א. יהי λ ע"ע של A^TA המתאים לו"ע $v \in R^p$ ונניח שהוא עם נורמה 1 (אחרת נוכל לנרמל- בכל מקרה הוא במרחב העצמי של A^TA). אז: $A^TAv = \lambda v$ ולכן: $AA^TAv = A\lambda v = \lambda Av$. לכן ע"ע עם הו"ע $Av \in R^n$.
ב. A^TA סימטרית אי שלילית ולכן לכסינה. מכאן של-A^TA ישנם p וקטורים עצמיים, ולכל היותר p ע"ע גדולים ממש מ-0. נסדר אותם בסדר יורד. נוכל לכתוב:

$$AA^T = U\Lambda U^T = USIS^TU^T = USV^TVS^TU^T := (*)$$

כיוון שדרגת השורות של מטריצה שווה לדרגת העמודות וכן בדרך דומה לזו שהשתמשתם בה בהוכחת הגרירה הראשונה בתרגיל 1 בשאלה 2, אז דרגתה של AA^T היא לכל היותר p. כיוון שמצאנו בסעיף א' ש-$Av_i \in R^n$ הוא ו"ע של AA^T המתאים לערך העצמי λ_i. נקבל כי:

$$\left||Av_i|\right|^2 = v_i^TA^TAv_i = v_i^T\lambda_iv_i = \lambda_i\left||v_i|\right|^2 = \begin{cases} 0, if\ v_i \in Ker(A^TA) \\ \lambda_i, O.W \end{cases}$$

נניח שיש $r \leq p$ וקטורים ע"ע המתאימים לערכים העצמיים החיוביים ממש (כדרגת המטריצה A^TA). כלומר הם בתמונה של A^TA. אזי כדי לקבל (בסופו של דבר) בסיס אורתונורמלי למרחב העצמי של AA^T, ננרמל אותם ונקבל כי הוקטורים:

$$u_1, \dots, u_r = \frac{Av_1}{\sqrt{\lambda_1}}, \dots, \frac{Av_r}{\sqrt{\lambda_r}}, \lambda_i > 0, \forall r \geq i \geq 1$$

שאר הוקטורים העצמיים של AA^T הם וקטורים המתאימים ל-n-r הע"ע השווים ל-0. כלומר הוקטורים הפורשים את מרחב הפתרונות $AA^T = 0$ (ולכן שייכים ל-$Ker(AA^T)$) ומכאן שגם ל- $Ker(A)$
כעת נוכל להגדיר ב-$(*)$ את $S \in R^{n\times p} |\ S_{ij} = \begin{cases} \sqrt{\lambda_i}, & i \leq r, and\ i = j \\ 0, O.W \end{cases}$ ולקבל את המבוקש. לסיכום, מתקיים ב-$(*)$:
U היא המטריצה שעמודתה ה-i היא $\frac{Av_i}{S_{ii}}$ עבור $r \geq i \geq 1$ והעמודות האחרות הן בסיס אורתונורמלי ל-$Ker(A)$ ומתקיים עבורן ש-$Av_i = 0$. לכן נוכל לכתוב: $AV = US$. כיוון ש-V אורתוגונלית נקבל:

$$A = USV^T$$

ג. $\underset{\substack{w\in\mathbb{R}^p \\ \|w\|=1}}{\operatorname{argmax}} \|Xw\|^2 = \underset{\|w\|=1}{\operatorname{argmax}} \sum_{i=1}^n \left(\sum_{j=1}^p w_jX_{ij}\right)^2 = \underset{\|w\|=1}{\operatorname{argmax}}\, w^TX^TXw = \underset{\|w\|=1}{\operatorname{argmax}}\, w^TU\Lambda U^Tw$

נסמן $U^Tw := c$. שימו לב שבהצבת האילוץ וכיוון ש-U אורתוגונלית, נקבל ש-
$\left||c|\right|^2 = w^TUU^Tw = w^Tw = \left||w|\right|^2 = 1$
ולכן: $w^TU\Lambda U^Tw = c^T\Lambda c = \sum_{i=1}^p \lambda_i c_i^2 \leq \lambda_1 \sum_{i=1}^p c_i^2 = \lambda_1$

כעת אם ניקח $w = u_1$ נקבל $c = e_1$ ומכאן שהחסם מושג וזהו הממקסם.

ד. יהי w וקטור כך ש-$\|w\|_2 = 1$. מתקיים כי:

$$\sum_{i=1}^{n} dist(x_i,w)^2 = \sum_{i=1}^{n}\|x_i - P_w(x_i)\| = \sum_{i=1}^{n}\left\|x_i - \frac{x_i^T w}{\|w\|}\cdot w\right\|^2 \overset{(*)}{=} \sum_{i=1}^{n}\|x_i - (x_i^T w)\ \cdot w\|_2^2$$
$$= \sum_{i=1}^{n}\left(x_i - (x_i^T w)\cdot w\right)^T\left(x_i - (x_i^T w)\cdot w\right)$$
$$= \sum_{i=1}^{n} x_i^T x_i - 2(x_i^T w)^2 + (x_i^T w)^2\cdot\|w\| \overset{(*)}{=} \sum_{i=1}^{n}\|x_i\|^2 - |w^T x_i|^2$$
$$= \|X\|_F^2 - \sum_{i=1}^{n}|w^T x_i|^2$$

$(*)$ - $\|w\| = 1$ וסימטריות של מכפלה פנימית.

ונסיק:

$$PC_1^{LS} = \underset{\substack{\|w\|=1\\ w\in\mathbb{R}^p}}{argmin}\sum_{i=1}^{n} dist(x_i,w)^2 = \underset{\substack{\|w\|=1\\ w\in\mathbb{R}^p}}{argmin}\|X\|_F^2 - \sum_{i=1}^{n}|w^T x_i|^2$$

נשים לב שזה שקול ל-

$$\|X\|_F^2 - \underbrace{\underset{\substack{\|w\|=1\\ w\in\mathbb{R}^p}}{argmax}\sum_{i=1}^{n}|w^T x_i|^2}_{(**)}$$

$(**)$ - אבל זאת בדיוק ההגדרה של PC_1^{var}, לכן ישנה התלכדות בין שתי ההגדרות.

שאלה 4- הקדמה למודל הלינארי

יהי Y משתנה מקרי כלשהו עם תוחלת ושונות סופיים.

א. נגדיר את הפונקציה - $M(a) = E[(Y\text{-}a)^2]$.
לכל a , הראו כי M מקבלת מינימום בערך $\hat{a} = E[Y]$.

ב. X ו-Y הם משתנים מקריים בעלי תוחלות ושונויות סופיות. לכל פונקציה f נגדיר – $MSE = E[(Y\text{-}f(x))^2]$. היעזרו בסעיף א' כדי להסיק ש-MSE ממוזער עבור $\widehat{f(x)} = E[Y|X = x]$.

<u>הדרכה:</u>
השתמשו בנוסחת התוחלת השלמה וכתבו – $MSE = E[g(x)]$, כאשר מתקיים –
$g(x) = E[(Y\text{-}f(x))^2|X = x]$
והפעילו את סעיף א' על Y <u>בהינתן</u> X = x.

פתרון:

א). לפי הגדרה, $M(a) := E(Y-a)^2 = E[Y^2]+a^2-2aE[Y]$.

נגזור:
$$\frac{d}{da}M(a) = 2a-2E[Y]$$

כשמשווים ל-0 מקבלים $a = E[Y]$. בנוסף, זה בהכרח נקודת מינימום כי $M(a)$ היא פרבולה ב-a עם מקדם >0 של האיבר הריבועי.

ב). נגדיר
$$f^*(x) := E[Y|X=x].$$

לפי נוסחת התוחלת השלמה,
$$\begin{aligned} MSE &:= E[(Y-f(X))^2] \\ &= E\{E[(Y-f(X))^2 | X]\} \\ &= E[g_f(X)] \end{aligned}$$

כאשר
$$g_f(x) := E[(Y-f(X))^2 | X=x] = E[(Y-f(x))^2|X=x],$$
ושימו לב לאופרטור של X (הגדול) ו-x (ערך אפשרי שהמ"מ X יכול לקבל).

עכשיו, אם מפעילים את סעיף א' <u>בהינתן $X=x$</u>, מקבלים
$$\arg\min_{f(x)} g_f(x) = \arg\min_a [(Y-a)^2|X=x] = E[Y|X=x]=f^*(x),$$
כלומר הפונקציה $f^*(\cdot)$ ממזערת עבורנו את $g_f(x)$ לכל x. מכאן שהפונקציה הנ"ל ממזערת גם את $E[g_f(X)]$ (בשלב האחרון נעזר במונוטוניות התוחלת: לכל Z,W מ"מ כלשהם, אזי $Z\le W \Rightarrow E[Z]\le E[W]$).