תרגיל 1 | IML (67577)

שם: יואב שפירא| ת`ז: 312492838

26 במרץ 2022

חלק 1 תאורטי

1. **אזכיר כי $||x||^2 = x^T x$. ומכאן:**

$$\begin{aligned}||Ax||^2 &= (Ax)^T (Ax)\\ &\overset{1}{=} x^T A^T Ax\\ &\overset{2}{=} x^T I x\\ &\overset{3}{=} x^T x = ||x||^2\end{aligned}$$

כש: 1 תכונת שחלוף, 2 תכונת של מטריצה אורתוגונלית, ו3 מהגדרת $||\cdot||$.
מכיוון שנורמה היא אי שלילית נובע כי $||Ax|| = ||x||$.

2. **$A \in \mathbb{R}_{2\times 3}$, ולכן יש לה 2 ערכים סינגולריים $\sigma_{1,2}$, וקודם נמצא אותם: הם למעשה השורשים של הע"ע של המטריצה הסימטרית AA^T:**

$$AA^T = \begin{bmatrix}1 & 1 & 0\\ 1 & -1 & 2\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -1\\ 2 & 0\end{bmatrix} = \begin{bmatrix}2 & 0\\ 0 & 6\end{bmatrix}$$

יצא קל, המטריצה אלכסונית וזה אומר שהע"ע הם הכניסות על האלכסון. כלומר קיבלנו ש $\sigma_1 = \sqrt{2}$ ו$\sigma_2 = \sqrt{6}$.
נמצא מטריצה U שמהווה בסיס אורתונורמלי לAA^T. גם כאן במקרה הוא קל וניתן לבחור פשוט $U = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$.
נמצא את V: תחילה נוכיח את הנוסחה הבאה: לכל $i \le n$ (במקרה הזה $n = 2$) מתקיים:

$$\overline{v}_i = \frac{1}{\sigma_i} A^T \overline{u}_i$$

הוכחה: תהא $A \in \mathbb{R}_{n\times m}$, נסתכל על הפירוק SVD

$$A = U\Sigma V^T \Longrightarrow A^T = V\Sigma^T U^T$$

מאורתוגונליות של U נקבל כי

$$A^TU = V\Sigma^T$$

נסתכל על המטריצה $V\Sigma^T \in \mathbb{R}_{m\times n}$. במטריצה Σ יש את הערכים הסינגולריים של A על האלכסון הראשי ובכל השאר 0, ולכן לכל $i \le n$ מתקיים:

$$\overline{v}_i \Sigma_i^T = \sigma_i \overline{v}_i$$

כלומר המטריצה נראית כך:

$$V\Sigma^T = \begin{bmatrix} \vdots & \cdots & \vdots \\ \sigma_1 \overline{v}_1 & \cdots & \sigma_n \overline{v}_n \\ \vdots & \cdots & \vdots \end{bmatrix}$$

ואם נסתכל על כל וקטור בנפרד, נקבל מ(*) ש $\overline{v}_i = \frac{1}{\sigma_i} A^T \overline{u}_i$ כמו שרצינו.
עכשיו נמצא את V בשיטה הזאת. עבור σ_1:

$$\overline{v}_1 = \frac{1}{\sigma_1} A^T \overline{u}_1 = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

עבור σ_2:

$$\overline{v}_2 = \frac{1}{\sigma_2} A^T \overline{u}_2 = \frac{\sqrt{6}}{6} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 2 \end{bmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{\sqrt{6}}{6} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

כעת נמצא את $\overline{v}_3$ ע"י מציאת וקטור המאונך לשני הקודמים וננרמל אותו. נחפש $\overline{v}_3 = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$ כך ש:

$$\left\langle \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\rangle = 0 \Longrightarrow x = -y$$

וגם

$$\left\langle \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} \right\rangle = 0 \Longrightarrow x = y - 2z$$

ונקבל כי $y = z$, כלומר הוקטור הוא מהצורה $\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$ ולכן נבחר את וקטור היחידה $\hat{v}_3 = \frac{\sqrt{3}}{3}\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$.

ובסה"כ קיבלנו את פירוק הSVD הבא:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{6}}{6} & \frac{\sqrt{6}}{3} \\ \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \end{bmatrix}$$

3. **משפט שימושי מליניארית:** תהא $C \in \mathbb{R}^{n\times n}$ מטריצה לכסינה ע"י מטריצה אורתוגונלית P כך ש $C = PDP^T$. אז ניתן לרשום את C כסכום ההטלות שלה על P, כשכל הטלה מתוחה ע"י הע"ע המתאים. כלומר:

$$C = \sum_{i=1}^{n} \lambda_i p_i p_i^T$$

נשתמש בזה בהמשך.

כעת נתון לנו כי $C_0 = A^T A$, ואנחנו יודעים אם כך שהע"ע של C_0 הם הערכים הסינגולריים של A בריבוע כלומר $\lambda_i = \sigma_i^2$, וכי אם $A = U\Sigma V^T$ פירוק SVD של A , אנחנו יודעים כי $C_0 = U\Sigma\Sigma^T V^T$. נסמן

$$\Sigma\Sigma^T = diag\left(\lambda_1, ..., \lambda_n\right)$$

מכיוון שזו מטריצה ריבועית קל לנו להעלות בחזקה:

$$\left(\Sigma\Sigma^T\right)^k = diag\left(\lambda_1^k, ..., \lambda_n^k\right)$$

וגם נעלה כך את C_0:

$$C_0^k = V diag\left(\lambda_1^k, ..., \lambda_n^k\right) V^T$$

נבחין כי C_0^k היא לכסינה ומתאימה לשימוש במשפט מלמעלה, נרשום אותה כך:

$$C_0^k = \sum_{i=1}^{n} v_i \lambda_i^k v_i^T = \sum_{i=1}^{n} \lambda_i^k v_i v_i^T$$

עכשיו נחשב את b_{k+1}: זהו למעשה וקטור מנורמל, אז אתייחס רק אל חישוב הכיוון (מונה) שלו ומשם נסיק על המכנה שלו. המונה שלו מתקיים:

$$C_0 b_k = C_0 C_0 b_{k-1} = ... = C_0^k b_0$$

נציב את C_0^k שחישבנו, ואת b_0 הנתון לנו:

$$C_0 b_k = \sum_{i=1}^{n} \lambda_i^k v_i v_i^T \cdot \sum_{i=1}^{n} a_i v_i = \sum_{i=1}^{n} a_i \lambda_i^k v_i v_i^T v_i = \sum_{i=1}^{n} a_i \lambda_i^k v_i$$

כשמהעבר האחרון נובע מאורתוגונליות של V. כעת, מהנתון ש $\lambda_1 >> \lambda_j$ ומהנתון ש $a_1 \neq 0$ נקבל ש

$$C_0^k b_0 \overset{k\to\infty}{\longrightarrow} a_1 \lambda_1^k v_1$$

b_{k+1} הוא וקטור מנורמל, כלומר מחולק במגניטודה שלו. נחשב אותה:

$$||C_0^k b_0|| \overset{k\to\infty}{\longrightarrow} ||a_1 \lambda_1^k v_1|| = \left|a_1 \lambda_1^k\right| \cdot ||v_1|| = \left|a_1 \lambda_1^k\right|$$

ובסה"כ נקבל:

$$b_{k+1} = \frac{C_0 b_k}{||C_0 b_k||} = \frac{C_0^k b_0}{||C_0^k b_0||} \overset{k\to\infty}{\longrightarrow} \frac{a_1 \lambda_1^k v_1}{||a_1 \lambda_1^k v_1||} = \frac{a_1 \lambda_1^k v_1}{\left|a_1 \lambda_1^k\right| \cdot ||v_1||} = \pm v_1$$

כדרוש.

4. U היא מטריצה אורתוגונלית ו$diag(\sigma)$ היא מטריצה אלכסונית, ולכן $Udiag(\sigma)U^T$ נותנת לנו מטריצה לכסינה (זה ליטרלי ההגדרה). מהמשפט השימושי בסעיף הקודם, נובע שניתן לרשום גם:

$$Udiag(\sigma)U^T = \sum_{i=1}^{n} \sigma_i u_i u_i^T$$

כלומר נקבל כי

$$f(\sigma) = \sum_{i=1}^{n} \sigma_i u_i u_i^T x$$

ככה קל לגזור את f לפי σ :

$$\frac{\partial f(\sigma)}{\partial \sigma_i} = u_i u_i^T \overline{x} = u_i \langle u_i, x \rangle$$

נשים לב שהנגזרת היא וקטור - זהו בדיוק הוקטור הi הוא העמודה הi ביעקביאן. נקבל שהיעקביאן נראה כך:

$$J(f_\sigma) = \begin{bmatrix} \vdots & & \vdots \\ u_1 \langle u_1, x \rangle & \cdots & u_n \langle u_n, x \rangle \\ \vdots & & \vdots \end{bmatrix}$$

5. לפי הבנתי f היא אותה הפונקצייה מהשאלה הקודמת. הגרדיאנט הוא מהגדרה:

$$\nabla h(\sigma) = \begin{bmatrix} \frac{\partial h(\sigma)}{\partial \sigma_1} \\ \vdots \\ \frac{\partial h(\sigma)}{\partial \sigma_n} \end{bmatrix}$$

נגזור בכל כיוון i , לפי כלל השרשרת:

$$\begin{aligned}\frac{\partial h}{\partial \sigma_i} &= \frac{\partial}{\partial \sigma_i}\frac{1}{2}(f(\sigma)-y)^2 = 2\cdot\frac{1}{2}(f(\sigma)-y)\cdot\frac{\partial}{\partial \sigma_i}f(\sigma)-y \\ &= (f(\sigma)-y)\cdot u_i\left\langle u_i, x\right\rangle - 0\end{aligned}$$

כשהמעבר האחרון הוא מהנגזרת שחישבנו בסעיף הקודם. מכאן נקבל כי:

$$\nabla h(\sigma) = \begin{bmatrix}(f(\sigma)-y)\cdot u_1\left\langle u_1, x\right\rangle \\ \vdots \\ (f(\sigma)-y)\cdot u_n\left\langle u_n, x\right\rangle\end{bmatrix} = (f(\sigma)-y)\cdot\begin{bmatrix}u_1\left\langle u_1, x\right\rangle \\ \vdots \\ u_n\left\langle u_n, x\right\rangle\end{bmatrix}$$

6. נסמן $x = \begin{pmatrix}x_1 \\ \vdots \\ x_d\end{pmatrix} \in \mathbb{R}^d$. נמצא את כל הנגזרות $\frac{\partial S(x)_j}{\partial x_i}$ ונרכיב את היעקוביאן:

$$\frac{\partial S(x)_j}{\partial x_i} = \frac{\partial}{\partial x_i}\frac{e^{x_j}}{\sum_{k=1}^d e^{x_k}}$$

נחשב כל איבר במונה ובמכנה בנפרד:

$$\frac{\partial}{\partial x_i}e^{x_j} = \begin{cases} e^{x_i} & i=j \\ 0 & i\neq j\end{cases}$$

$$\frac{\partial}{\partial x_i}\sum_{k=1}^{d} e^{x_k} = e^{x_i}$$

נטפל קודם במקרה שבו $i=j$. נרכיב את הנגזרת של כל השבר:

$$\begin{aligned}\frac{\partial}{\partial x_i}\frac{e^{x_j}}{\sum_{k=1}^d e^{x_k}} &= \frac{e^{x_i}\left(\sum_{k=1}^d e^{x_k}\right) - e^{x_j}e^{x_i}}{\left(\sum_{k=1}^d e^{x_k}\right)^2} \\ &= \frac{e^{x_i}\cancel{\left(\sum_{k=1}^d e^{x_k}\right)}}{\left(\sum_{k=1}^d e^{x_k}\right)^{\cancel{2}}} - \frac{e^{x_j}e^{x_i}}{\left(\sum_{k=1}^d e^{x_k}\right)^2} \\ &= S(x)_i - S(x)_i S(x)_j = S(x)_i\left(1 - S(x)_j\right)\end{aligned}$$

במקרה בו $i\neq j$:

$$\frac{\partial}{\partial x_i}\frac{e^{x_j}}{\sum_{k=1}^d e^{x_k}} = \frac{-e^{x_j}e^{x_i}}{\left(\sum_{k=1}^d e^{x_k}\right)^2} = -S(x)_i S(x)_j$$

נרכיב את היעקביאן:

$$\left[J_{S(x)}\right]_i^j=\begin{cases}S(x)_i\left(1-S(x)_j\right) & i=j\\ -S(x)_iS(x)_j & i\neq j\end{cases}$$

.7

$$f\left(x,y\right)=x^3-5xy-y^5$$

נחשב את הנגזרות החלקיות הראשונות:

$$\frac{\partial f}{\partial x}=3x^2-5y$$

$$\frac{\partial f}{\partial y}=-5x-5y^4$$

נחשב את הנגזרות החלקיות השניות:

$$\frac{\partial f^2}{\partial x^2}=6x$$

$$\frac{\partial f^2}{\partial xy}=-5$$

$$\frac{\partial f^2}{\partial y^2}=-20y^3$$

נרכיב את ההסיאן:

$$H\left[f(x,y)\right]=\begin{bmatrix}6x & -5\\ -5 & -20y^3\end{bmatrix}$$

8. הרעיון הוא להראות שהאומד מתכנס מסביב לערך הממוצע ככל שכמות הסמפלים גדלה. כלומר, $Var(\hat{\mu}_n)\propto\frac{1}{n}$.
השיטה היא לחסום את סטיית התקן עבור כל $n\in\mathbb{N}$:
יהי $n\in\mathbb{N}$, ונבחר $\varepsilon>0$ איזשהי מידת דיוק רצויה. מכיוון ש$Var(x_i)<\infty$ וגם $\varepsilon>0$ ניתן להשתמש בא"ש צ'בישב:

$$P\left(|\hat{\mu}_n-\mathbb{E}[\hat{\mu}_n]|\geq\varepsilon\right)\leq\frac{Var(\hat{\mu}_n)}{\varepsilon^2}$$

$|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]|$ היא בדיוק סטיית התקן של $\hat{\mu}_n$. נמשיך בפיתוח הא"ש:

$$\begin{aligned}\frac{Var(\hat{\mu}_n)}{\varepsilon^2} &= \frac{Var(\frac{1}{n}\sum_{i=1}^{n} x_i)}{\varepsilon^2}\\ {}_1 &= \frac{Var(\sum_{i=1}^{n} x_i)}{n^2\varepsilon^2}\\ {}_2 &= \frac{nVar(x_i)}{n^2\varepsilon^2} = \frac{Var(x_i)}{n\varepsilon^2}\end{aligned}$$

כשהמעבר הראשון מהגדרת $\hat{\mu}_n$, 1 ו2 מתכונות של שונות.
מכיוון שממוצע הוא אומד לא מוטה אז $\mathbb{E}[\hat{\mu}_n] = \mathbb{E}[x_i]$. כלומר קיבלנו כי עבור כל n, ε, הסיכוי שהאומד רחוק מהממוצע ביותר מε הולך וקטן, כלומר האומד מתכנס מסביב לערך הממוצע שלו.

9. סימנתי כאן את התצפיות x_i כוקטורים: $\overline{x_i}$, לנוחות שלי.
ה$likelihood$ הוא בעצם הביטוי המביע את הסיכוי לקבל את $\overline{x_i}$ אם הוא היה מתפלג ע"י $N\left(\overline{\mu}, \Sigma\right)$:

$$P\left(\overline{x_i}\right) \quad s.t \;\; \overline{x_i} \sim N\left(\overline{\mu}, \Sigma\right) \quad \left(\overline{x_i} \in \mathbb{R}^d\right)$$

מכיוון ש$P(\overline{x}) \propto p\overline{(x)}$ והביטוי הזה נדרש ל$argMax$ אחר כך (p היא פונקציית הצפיפות של ההתפלגות) אז $likelihood$ מחושב ע"י פונקציית הצפיפות בעצמה. במקרה שלנו, ה$likelihood$ לכל תצפית בנפרד נתונה ע"י:

$$p\left(\overline{x_i}\right) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot exp\left(-\frac{1}{2}\left(\overline{x} - \overline{\mu}\right)^T \Sigma^{-1}\left(\overline{x} - \overline{\mu}\right)\right)$$

מכיוון שכל m התצפיות נדגמו באופן בת"ל מאותה התפלגות אז ה$likelihood$ בסה"כ נתון ע"י:

$$P\left(\overline{x}_1, \overline{x}_2 ... \overline{x}_m | \overline{\mu}, \Sigma\right) = \prod_{i=1}^{m} P\left(\overline{x}_i | \overline{\mu}, \Sigma\right)$$

ניקח log:

$$log\left(P\left(\overline{x}_1, \overline{x}_2 ... \overline{x}_m | \overline{\mu}, \Sigma\right)\right) = \sum_{i=1}^{m} log\left(P\left(\overline{x}_i | \overline{\mu}, \Sigma\right)\right) = logL$$

אני סימנתי את הביטוי ב$logL$. נפתח את הביטוי:

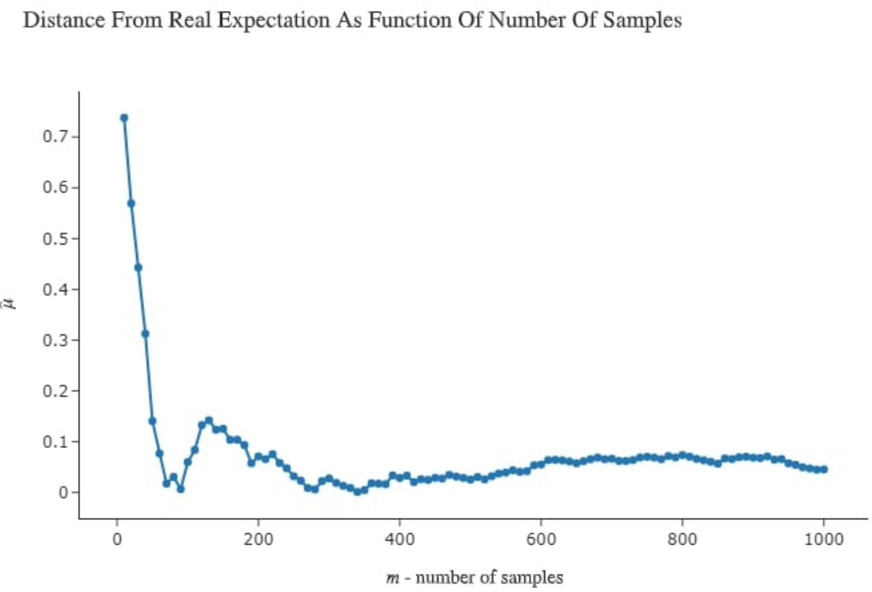
$$\begin{aligned}logL &= \sum_{i=1}^{m} log\left(\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \cdot exp\left(-\frac{1}{2}\left(\overline{x_i} - \overline{\mu}\right)^T \Sigma^{-1}\left(\overline{x_i} - \overline{\mu}\right)\right)\right)\\ {}_1 &= \sum_{i=1}^{m}\left(-log\left(\sqrt{(2\pi)^d|\Sigma|}\right) - \frac{1}{2}\left(\overline{x_i} - \overline{\mu}\right)^T \Sigma^{-1}\left(\overline{x_i} - \overline{\mu}\right)\right)\\ {}_2 &= -\sum_{i=1}^{m}\frac{1}{2}\left(log\left((2\pi)^d\right) + log\left(|\Sigma|\right) + \left(\overline{x_i} - \overline{\mu}\right)^T \Sigma^{-1}\left(\overline{x_i} - \overline{\mu}\right)\right)\\ {}_3 &= -\frac{md \cdot log\left(2\pi\right)}{2} - \frac{-log\left(|\Sigma|\right)}{2} - \sum_{i=1}^{m}\frac{1}{2}\left(\overline{x_i} - \overline{\mu}\right)^T \Sigma^{-1}\left(\overline{x_i} - \overline{\mu}\right)\end{aligned}$$

כשב1: יישום חוקי לוגריתם, 2: הוצאת שורש מהלוג ואז גורם משותף, 3: הוצאת איברים שלא תלויים בסיגמא.

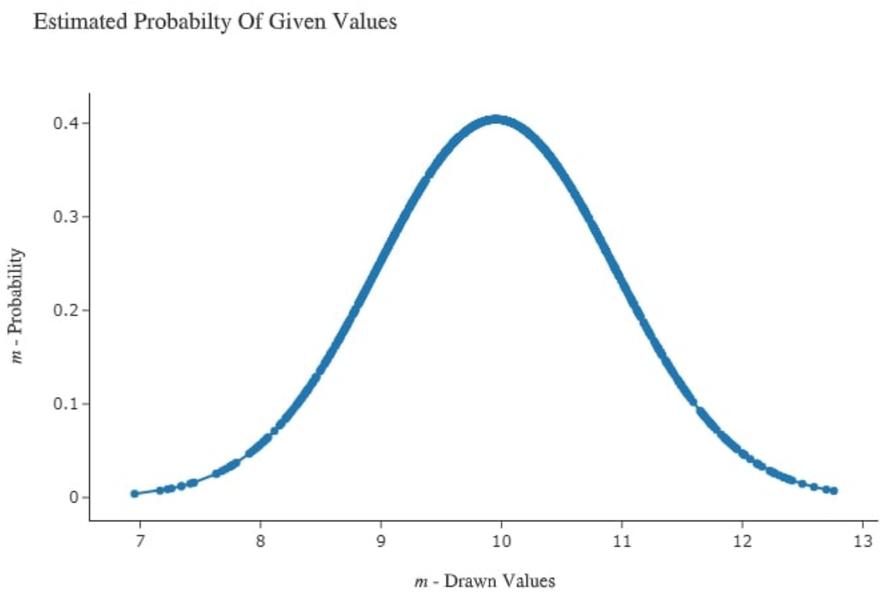
חלק 3 מעשי

1. קוד.

2. **בגרף הבא, לכל $10 \le m \le 1000$ התאמתי מודל של אומד גאוסיאני חדש (ע"י $.fit(X[m:])$). לכל m חישבתי את המרחק מהתוחלת האמיתית, והדפסתי את המרחק כפונקצייה של m:**



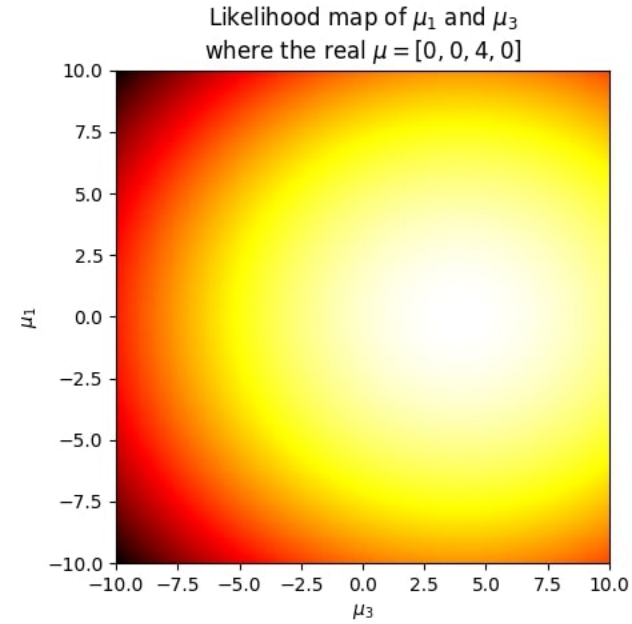
3. בשאלה זו התבקשנו להציג את הPDF שחושבה ע"י האומד כפונקצייה של הערכים שהגרלנו. זוהי למעשה פונקציית הצפיפות שהיינו מצפים למצוא. מכיוון ש זו התפלגות נורמלית עם תוחלת 10 ושונות 1 היינו מצפים לראות פעמון עגלגל מעל הערך 10. זה אכן המצב:



4. (שאלת קוד). בשאלה זו אימנו את האומד שלנו עבור דאטא שהוגרל מתוך התפלגות נורמלית שנתונה ע"י:

$$\overline{\mu} = \begin{pmatrix} 0 \\ 0 \\ 4 \\ 0 \end{pmatrix}, \Sigma = \begin{bmatrix} 1 & 0.2 & 0 & 0.5 \\ 0.2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}$$

5. בשאלה זו השתמשנו באומד שאימנו, ובדקנו מה ה$Likelihood$ של וקטור התוחלת עבור $\overline{\mu} = \begin{pmatrix} \mu_1 \\ 0 \\ \mu_3 \\ 0 \end{pmatrix}$ כאשר $\mu_{1,3}$ הם נעלמים בתוך הקטע $[-10, 10]$. הדפסנו עבור כל צמד ערכים את ה$likelihood$ בצורת $heatmap$. המפה שיצאה:



6. ניתן לראות בבירור שיש העדפה לערכים באזור $\mu_1 = 0$ ו-$\mu_3 = 4$, כמו שהיינו מצפים. מבדיקה מדוייקת בקוד, הערכים הם (בקירוב):

$$\mu_1 = 0.050 \quad , \mu_3 = 3.969$$