שיטות סטוכסטיות- משימת MCMC

Part A- modeling

בחלק זה, נדרש מאיתנו לבחור מודל לחיזוי מספר מקרי הקורונה החדשים בכל יום. בצפייה בגרף של מספר המקרים היומי החדש לפי חלון נע של שבוע, היה ניתן להבחין בירידה לינארית ועל כן, המודל שבחרתי הינו מודל רגרסיה לינארית מרובה משתנים. המודל הינו מרובה משתנים שכן ישנו שימוש ב-3 משתנים: חותמת זמן (מספר בין 1 ל-60), מספר המקרים שהיה לפני שבוע בדיוק והיום בשבוע (מספר בין 0 ל-6 כך ש-0 מייצג את יום שני). משוואת המודל הנבחר המתואר לעיל הינה:

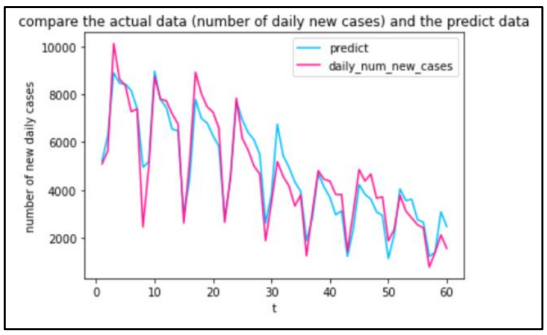
$$\mu = a_0 + a_1 \cdot time + a_2 \cdot cases\_before\_7days + a_3 \cdot day + \varepsilon$$

כך ש-a_0, a_1, a_2, a_3 הם הפרמטרים של המודל ו-ε זו שגיאת המודל אשר מתפלגת נורמלית. כמו כן, הבחירה במשתנים אלו היא כך שלחומת הזמן יש השפעה, שכן אנו מצפים כי עם התקדמות הזמן התחלואה תרד (כפי שנצפה בטווח התאריכים שלי), נצפה גם כי ליום בשבוע ישנה השפעה כך שלמשל בסופי שבוע ישנה ירידה בבדיקות ובאופן דומה גם למספר הבדיקות בדיוק בשבוע שלפני (אותו היום) ישנה השפעה כך שנצפה שיהיה מספר נמוך יותר עקב הירידה הנצפית.

המודל הנבחר מכיל מספר משתנים מסבירים לערך הרצוי/המוסבר (Y), וממנו ניתן למצוא את הפרמטרים המגדירים את הקשר הלינארי בין המשתנים המסבירים למוסבר.

למודל זה ישנן כמה הנחות:

- השגיאה במודל עבור כל תצפית הינה נורמלית $\varepsilon_i \sim N(0, \sigma^2)$
- הקשר בין כל משתנה מסביר x_i ל-Y הוא לינארי
- הקבועים a_i אינם ידועים.
- לכלל התצפיות ישנה שונות אחידה ללא תלות בערכי המשתנים המסבירים.
- מודל זה מתפלג באופן הבא: $(y_i | x_{i1}, \dots, x_{ip}) \sim N\left(a_0 + \sum_{j=1}^{p} a_j x_{ij}, \sigma^2\right)$



תרשים 1- המודל החזוי אל מול הנתונים האמיתיים

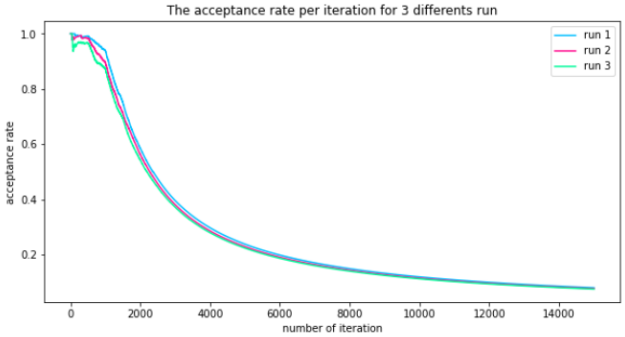
על מנת לקבל את מקדמי המודל, החלטתי לבצע מזעור של ה-MSE כלומר סכום ריבועי השגיאות בין הנתונים החזויים ממשוואת המודל הנתונה אל בין האמיתיים. מזעור ה-MSE שקול למיקסום ה-likelihood ולכן מתאפשר. באמצעות שימוש בפונקציית minimize מספריית scipy.optimize חיפשתי את המקדמים הטובים ביותר בהינתן פונקציית ה-MSE שתמוזער וניחוש התחלתי למקדמים (הסבר לניחוש מצורף בקוד). המקדמים הטובים ביותר שהתקבלו מכך הם: $a_0 = 2289.658, a_1 = -30.703, a_2 = 0.689, a_3 = -120.695$. ערך המטרה של הפונקציה הוא 32270066.647.

Part B- MCMC

בחלק זה היה עלינו להריץ את מודל MCMC ולשם כך נדרשו אתחולים לפרמטרים שונים שאיתם המודל מופעל.

- scaling factor: חושב לפי הנוסחה הבאה $\frac{2.4}{num\ of\ parameters=4}$ ולכן 0.6.
- covariance matrix: עבור J שיהווה התפלגות של בחירת הפרמטרים החדשים בהינתן הנוכחיים: מטריצה אלכסונית בגודל 4×4 עם ערכים 0.005 באלכסון. כלומר גודל צעד התחלתי זה 0.005 שנבחר על מנת לא "לקפוץ" יותר מידי ולפספס נקודות.
- ניחוש התחלתי לפרמטרים: נדרשו 3 ריצות ולכן ישנם 3 אתחולים אשר נלקחו מהתפלגות נורמלית מרובת משתנים עם תוחלת של המקדמים הטובים ביותר שנמצאו בחלק הקודם ומטריצת covariance אלכסונית שבה הערכים יהיו שורש המקדמים הטובים ביותר שנמצאו בחלק הקודם. 3 הסטים שהתקבלו הם:
  - $a_0 = 2293.093, a_1 = -29.178, a_2 = 2.077, a_3 = -121.153$
  - $a_0 = 2287.653, a_1 = -33.472, a_2 = 1.848, a_3 = -119.688$
  - $a_0 = 2280.772, a_1 = -33.272, a_2 = 1.533, a_3 = -124.873$

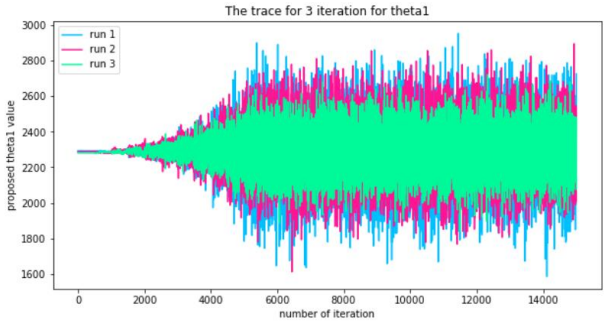
<u>הצגת הגרפים והסברים</u>

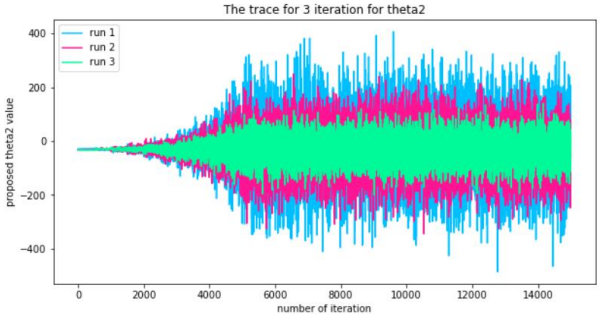


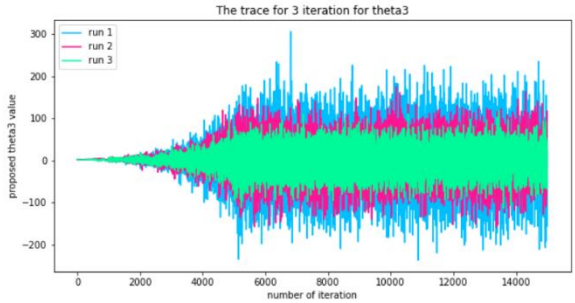
the acceptance rate per iteration:

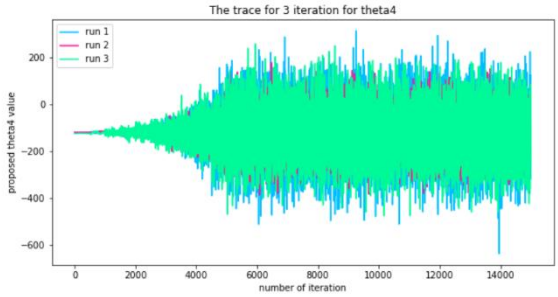
ניתן לראות מגרף זה כי עבור 3 הריצות מתקבלת התנהגות דומה לפיה מתחילים בקצב קבלה גבוה (1) וככל שעובר הזמן ישנה ירידה עד לערך של כ-0.1. כלומר עם הזמן פחות נקבל את הפרמטרים המוצעים בכל איטרציה כך שהמדגם המוצע פחות סביר מאשר הנוכחי.

ה-trace ל-3 הריצות לכל פרמטר בנפרד:



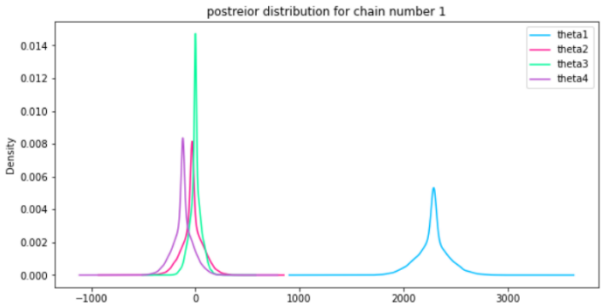


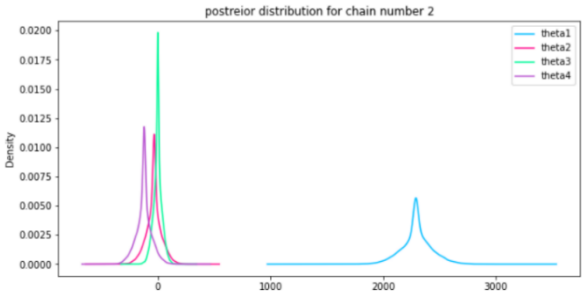


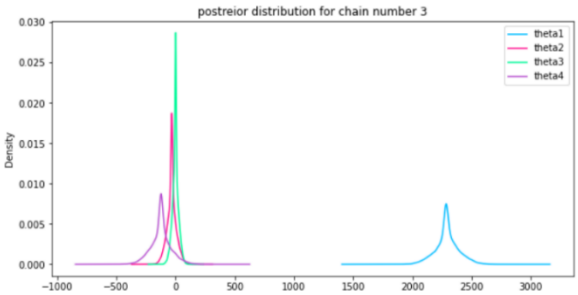


בגרפים לעיל ניתן לראות הרצה של אלגוריתם ה-MCMC לאורך 3 הרצות שונות לצורך בדיקת התכנסות. כלומר רואים כי ישנה התכנסות לכל מקדם במודל הנבחר לאורך 3 איטרציות שכן עבור כל האיטרציות ישנה התמקדות בחלק דומה בגרף ואין איטרציה אחת שנמצאת ממש מעל/מתחת לאחרות. כמו כן זה עומד בקנה אחד עם תוצאות מבחן ההתכנסות של Gelman-Rubin שעליו ידובר בהמשך.

התפלגות ה-posterior לכל פרמטר עבור כל הרצה בנפרד:







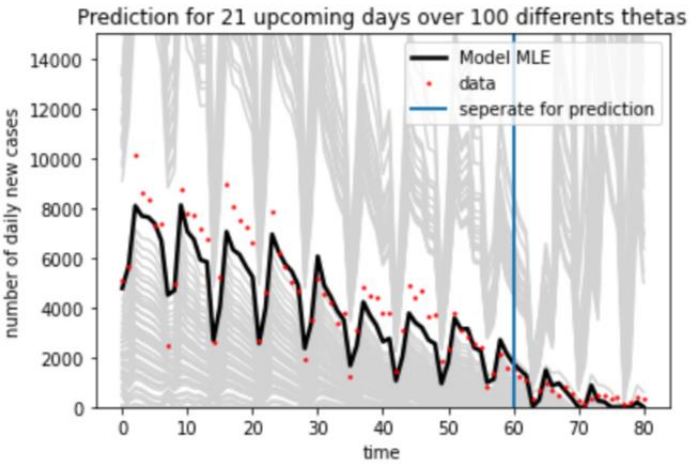
מ-3 הגרפים, ניתן לראות כי ישנה התפלגות דומה לכלל המקדמים על כל 3 הריצות. כלומר, ניתן לראות כי מקבלים התפלגות posterior דומה לכל מקדם במודל הרגרסיה הלינארית המרובה וכמו כן ההתפלגות הזו מדמה התפלגות נורמלית והיא המשוערכת לכל מקדם. כמו כן המקדם הראשון הינו החיתוך של המשוואה כך שהגיוני כי נמצא בטווח ערכים גבוה יותר על מנת לבצע "העלאה" של מספר המקרים לטווח הנכון, ושאר המקדמים בטווח יותר קטן שכן באים "לכפר" על ירידה/העלאה בהתאם למשתנה המסביר (כדוגמת היום בשבוע שנצפה בו להשפעה קטנה).

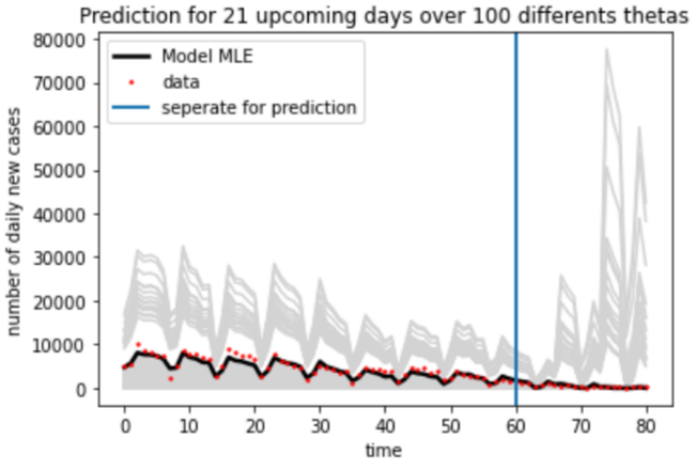
<u>הסבר על תוצאות מבחן Gelman-Rubin</u>

נדרשנו לבצע מבחן זה על מנת לבדוק את ההתכנסות. ממבחן זה קיבלתי את התוצאות הבאות: עבור a_0 1.002, עבור a_1 1.0014, עבור a_2 1.0003 ועבור a_3 1.0001. במבחן זה ישנה בדיקה כמה השונות בתוך כל שרשרת שונה מהשונות בין השרשראות. מבצעים חלוקה של היחס על מנת לבדוק את ההתכנסות ובמידה והיחס גדול אין התכנסות. במקרה שלי, התוצאות כולן קרובות לערך 1, מה שמעיד כי השונות בתוך כל שרשרת דומה לשונות בין השרשראות. מכאן התוצאות מעידות על התכנסות מה שעומד בקנה אחד עם גרפי ה-trace אשר מראים התכנסות.

Part C- 21 days projection

בחלק זה היה עלינו לחזות את 21 הימים הבאים תוך שימוש ב-MCMC. כתוצאה מכך נוצרו הגרפים המוצגים מטה. הגרפים הינם זהים כך שבימין הגרף המוצג הוא ללא הגבלה בציר ה-Y והשמאלי עם הגבלה ל-15000 על מנת לראות בצורה יותר טובה את החיזוי.





הרציונל של חלק זה היה להוציא את ההתפלגות הפוסטריורית של המקדמים אשר קיבלנו מה-MCMC ולהגריל 100 סטים של מקדמים מהתפלגות זו. עבור כל סט של מקדמים בוצעה הרצה של המודל (רגרסיה לינארית מרובה) באופן רגיל עבור 60 הרשומות מהתאריכים שניתנו לי והרצה של המודל עם מרכיב של שגיאה שמתפלגת נורמלית עבור 21 הימים הבאים. כמו כן חושב מדד ה-likelihood בהתאם לסט המקדמים והנתון האמיתי של 60 הרשומות על מנת למצוא את ה-MLE. יש לציין כי עקב העמודות החדשות שנוספו (היום בשבוע ומספר המקרים לפני שבוע) היה אסור להשתמש בנתונים האמיתיים העתידיים ולכן במידת האפשר השתמשתי בנתונים של 60 הרשומות מהעבר ובהמשך לכך השתמשתי בחיזוי שקיבלתי ל-21 הימים החדשים. למשל עבור נתון 68, הנתון שמתאים למספר המקרים לפני שבוע הוא 61 והיה אסור לנו לגעת בנתון האמיתי שלו ולכן השתמשתי בחיזוי שלו. לבסוף בוצעה הצגה של 100 הריאליזציות (בקווים אפורים), הריאליזציה עם ה-MLE הגבוה ביותר (בשחור), הנקודות האמיתיות של 81 התאריכים (באדום) וקו מפריד בין החיזוי של 21 הימים קדימה לבין 60 הימים שניתנו לי מההתחלה.

מהגרפים המוצגים ניתן לראות כי עבור 100 ההרצות עם מקדמים שונים מההתפלגות הפוסטריורית ישנן 2 קבוצות של חיזויים, האחת קופצנית המגיעה לערכים גבוהים במיוחד והשנייה קבוצה תחתונה המדמה בצורה טובה את הדאטה האמיתי. את הקבוצה הקופצנית ניתן להסביר על ידי התפלגות המקדמים אשר קיבלנו ב-3 ההרצות של ה-MCMC. בהתפלגות זו ניתן לראות כי ישנן מקדמים אשר מגיעים לערכים גבוהים במיוחד בהסתברות כלשהי (על אף שקטנה יותר) וכמו כן מקדמים אלה מוכפלים בנתוני הדאטה x_i

אשר יכולים גם כן להגיע לערכים גבוהים כדוגמת עמודת מספר המקרים היומיים שהיה לפני שבוע בדיוק. מכל הנאמר, ניתן להסיק כי ייתכנו ריאליזציות עם מקדמים גבוהים אשר יחזו ערך גבוה מהצפוי.

כפי שנאמר לעיל, החלק התחתון של הריאליזציות המתקבלות קרובות למציאות, כלומר לנקודות האדומות ובפרט גרף ה-MLE קרוב מאוד לכלל הנתונים ובפרט עבור 21 הימים הבאים שעבורם מבצעים חיזוי. על כן ניתן להגיד כי החיזוי בוצע באופן טוב המדמה את המציאות ומזהה את השינויים בו. כמו כן אין זה מפתיע שהתוצאות טובות, זאת שכן כפי שציפיתי ישנה השפעה מועטה ושלילית לחותמת הזמן כלומר ירידה בתחלואה. בנוסף לכך ליום בשבוע (בעיקר בסופי השבוע) גם כן יש השפעה שלילית מה שמשפיע על הירידה וכמו כן מספר המקרים לפני שבוע תוך הכפלה במקדם קטן תורם ומדמה בצורה טובה את הירידה/עלייה (עקב קפיצות) בתחלואה בהתאם לדאטה האמיתי שקיבלתי ובהתאם להכנסה של לוגיקה זו למודל הרגרסיה הלינארית המרובה.

כמו כן, אציין כי הנתונים שקיבלתי הם התאריכים בהם הייתה תחלואה נוספת גדולה במיוחד שהגיעה בשיאה עד לכ-10 אלף חולים ביום. עם זאת עם ההגעה לשיא החלה ירידה בתחלואת החולים ככל שהזמן עבר עד לירידה מתחת ל-2000 חולים ביום. כלל הירידה הזו במודל באה לידי ביטוי במודל כפי שהוסבר לעיל.

לבסוף, על מנת לשפר את החיזוי שביצעתי ניתן להוסיף עוד נתונים ובכך לקבל מקדמים טובים יותר המותאמים באופן טוב יותר לנתונים האמיתיים. עבור חיזויים ידוע כי ככל שיש יותר נתונים זה יותר טוב לבניית מודל מדויק יותר. בחלק זה כל אחד עבד על 60 נתונים בלבד וכמות זו מהווה כמות מזערית ביחס לחקר בעולם האמיתי. בנוסף לכך ניתן לחשוב על פיצ'רים נוספים אשר בעלי חשיבות בנושא התפשטות מחלות מעבר למשתנים המסבירים אשר השתמשתי בהם במודל שלי. במודל זה השתמשתי ב-3 משתנים מסבירים בלבד ובמציאות ככל הנראה נצטרך הרבה יותר משתנים מסבירים כדי לבצע חיזוי שינבא תוצאות מדויקות ויביא למדדים גבוהים יותר בנכונות החיזוי. למשל היה ניתן להוסיף פיצ'רים המעידים על סגרים, הגבלות כמו חבישת מסיכה או בידוד שמשפיעים בעיקר על הירידה בתחלואה (והשפעה על מדד R למשל). עם זאת פיצ'רים כאלה יכולים להיות מוכפלים במקדם שלילי במודל אשר יתרום לירידה בתחלואה בהינתן ואכן היו הגבלות. דבר נוסף שיכול לשפר את החיזוי הוא בחינת מודלים שונים ולאו דווקא מודל רגרסיה לינארית מרובה משתנים, כדוגמת מודל SIR אשר מיועד גם לבדיקת התפשטות מחלות. זאת שכן, במציאות אנו בודקים מודלים רבים עד להגעה למודל הנבחר אשר יהיה זה בעל המדדים הטובים ביותר בחיזוי הנדרש.