התאמת מודלים לנתוני תאומים עם מידע חסר

עפרא עראידה

עבודת גמר מחקרית (תיזה) המוגשת כמילוי חלק מהדרישות

לקבלת התואר "מוסמך האוניברסיטה".

אוניברסיטת חיפה

הפקולטה למדעי החברה

החוג לסטטיסטיקה

נובמבר 2014

התאמת מודלים לנתוני תאומים עם מידע חסר

מאת: עפרא עראידה

בהנחיית: ד"ר, יאיר גולדברג.

עבודת גמר מחקרית (תזה) המוגשת כמילוי חלק מהדרישות

לקבלת התואר "מוסמך האוניברסיטה".

אוניברסיטת חיפה

הפקולטה למדעי החברה

החוג לסטטיסטיקה

נובמבר 2014

מאושר על ידי:_____ תאריך:_____
(מנחה העבודה)

מאושר על ידי:_____ תאריך:_____
(יו"ר הוועדה החוגית לתואר שני)

תוכן עניינים

תקציר iii

רשימת איורים iv

1 הקדמה 1

2 מודל ACE 2

3 פונקציית הנראות 4

3.1 נראות מירבית Maximum Likelihood 4

3.2 הנראות עבור המודל המלא 4

3.3 הנראות עבור המודל החסר 5

4 אלגוריתם Expectation-Maximization Algorithm (EM) 7

4.1 הקדמה 7

4.2 הפעלת אלגוריתם EM ואלגוריתם ECM 8

4.2.1 סיכום האלגוריתם 11

5 סימולציות 14

5.1 הפעלת הסימולציות 14

5.1.1 התרחיש הראשון 14

5.1.2 התרחיש השני 18

5.2 תוצאות הסימולציות 22

6 סיכום 23

7 ביבליוגרפיה 24

8 נספחים 25

8.1 אומדי נראות מירבית עבור המודל המלא 25

8.2 הרחבה- אלגוריתם EM 28

8.3 אמידה כללית לווקטור המשקולות π 30

8.4 תרשימי Boxplot עבור הפרמטר μ 31

התאמת מודלים לנתוני תאומים עם מידע חסר

עפרא עראידה

תקציר

מחקרי תאומים הם כלי חשוב לחקירת השפעת התורשה והסביבה על שונות בין בני האדם. מחקרים רבים שמתייחסים לנתוני תאומים מניחים שהסיווג של התאומים נתון, כלומר המידע עבור אם זוג התאומים זהים או לא זהים הוא נתון. במאגרי נתונים רבים, המידע עבור הסיווג לא נתון באופן מלא. התאמת מודלים תחת ההנחה שאין נתונים חסרים, ואמידת הפרמטרים למודלים אלה יכולה להביא לאומדים מוטים. ניתוח והסתמכות על האומדים המוטים עלול להוביל למסקנות מוטעות. לפיכך צריך למצוא שיטות לטיפול בבעיה זו.

ישנם חוקרים שמנסים לטפל בבעיה של המידע החסר עבור סוג התאומים במתן משקולות קבועות מראש לסיכוי של זוג תאומים להיות זהים או לא זהים. בעבודה הזו אנו מציעים מודל תערובת של התפלגויות לניתוח נתוני תאומים, שהמידע עבור הסיווג של התאומים לא בהכרח ידוע באופן מלא. להבדיל משיטות קודמות, שבהם משלימים את המידע החסר בסיכויים ומשקולות קבועים מראש, אנו מציעים אמדי נראות מירבית לבעיה זו. בעבודה זו מציאת אומדי הנראות המירבית מהווה אתגר חישובי, מאחר שהמקסום מתבצע תחת אילוצים הנובעים ממבנה מטריצת השונות. למציאת אומדי הנראות המירבית אנו נעזרים באלגוריתם ECM, Expectation-Conditional Maximization, למימוש האלגוריתם פיתחנו קוד בשפת R המשתמש בשיטות נומריות. בעזרת קוד זה הרצנו סימולציות לבחינת המודל שהוצע. התוצאות שהתקבלו העידו שהשימוש במודל שהצגנו, מודל תערובת של התפלגויות עם אלגוריתם ECM למציאת אומדי הנראות המירבית, הוא מודל שמספק תוצאות טובות יותר מאשר השיטות הקודמות להשלמת המידע החסר.

רשימת איורים

1 איור 1 15
2 איור 2 16
3 איור 3 17
4 איור 4 19
5 איור 5 20
6 איור 6 21
7 איור 7 31

1 הקדמה

מודלים המסבירים את השונות בין בני האדם כפונקציה של הסביבה המשפחתית, התורשה, והסביבה הכללית חשובים בתחומים רבים, כגון הנדסה גנטית, פסיכולוגיה, ביולוגיה, רפואה, ביטוח, ועוד. הבנה טובה יותר של הגורמים לשונויות האלה תאפשר להתאים טיפולים רפואיים טובים יותר בהקשר הרפואי, תאפשר מציאת פתרונות לבעיות פסיכולוגיות בהקשר הפסיכולוגי, וכן בהקשר ביטוחי, תאפשר התאמת פרמיה מדויקת יותר למבוטחים. אחת הדרכים המקובלות להסברת השונויות היא על ידי התאמת מודלים פרמטריים. התאמת מודלים פרמטריים היא פעולה שבה מנסים להתאים התפלגות ידועה עם מספר סופי של פרמטרים שמתארת את הנתונים בצורה סבירה. לדוגמא, התאמת רגרסיה לינארית, או השימוש במודלים ליניאריים מוכללים (GLM).

חשוב לציין כי בניתוח נתונים בכל מיני תחומים, כמו בתחום האקטואריה וכן בתחום הרפואי, נתקלים פעמים רבות בבעיות של מידע חסר. לדוגמא: גיל המבוטח, הקשר המשפחתי, ועוד. השפעת הגיל על השינויים במצב הבריאותי היא משמעותית, על כן צריך לחפש דרכים שמתייחסות לנתון זה אפילו אם הוא חסר. הקשר המשפחתי והתורשה הגנטית אף הם משפיעים על המצב הבריאותי ועל הסיכויים להתפתחות מחלות. לכן באנליזה סטטיסטית צריך להתחשב בנתונים החסרים, כי התעלמות מהמידע החסר יכולה להביא לאמידה מוטה.

מחקרים רבים שמתעסקים במודלים של שונות בין בני האדם, מתעסקים בנתונים עבור תאומים. מחקרי תאומים (Twin Study) חושפים את החשיבות של השפעות הסביבה והגנטיקה (התורשה). תאומים הינם מקור חשוב להתבוננות ומחקר, מאחר והם מאפשרים למידה עבור ההשפעה של הסביבה ושל הגנים. אנו נתמקד במחקרי תאומים שבהם נחקור את השפעת הסביבה והגנטיקה על השונות בין בני האדם.

הרעיון הבסיסי שמשתמשים בו על מנת לחקור את השפעת הסביבה לעומת הגנטיקה במחקרי תאומים מסתמך על סוג התאומים: תאומים זהים או מונוזיגוטיים (MZ) ותאומים לא זהים או דיזיגוטיים (DZ). במאגרי נתונים שבהם יש נתוני משפחות מנסים לחפש את התאומים לצורך מחקר תאומים. בעיה המתעוררת במאגרים אלה, שהמידע עבור סוג התאומים לרוב לא קיים. אפשר למצוא במאגרי נתונים מידע על תאריך לידה שמאפשר איתור התאומים. אפשר למצוא מידע על המין, שזה מספק לנו מידע עבור הסוג לחלק מהתאומים, אבל עדיין אנו נתקלים בבעיית מידע חסר עבור חלק מהתצפיות. המטרה בעבודה זו היא לפתח שיטות להתאמת מודלים עם נתונים חסרים עבור מחקרי תאומים, ולהציע אומדי נראות מירבית המתחשבים במידע החסר.

2 מודל ACE

מודל ACE בודק את ההשפעה היחסית של גנים וסביבה על השונות בין בני האדם. במודל ACE אנו מסתכלים על השונות הכוללת כמורכבת משלושה גורמים: השונות האדיטיבית הגנטית (A- Additive genetic component), השונות של הסביבה המשותפת (C- The common shared environment) והשונות של הסביבה הייחודית (E- The unique environment). כלומר מסתכלים על השונות הכוללת כ־

$V = a^2 + c^2 + e^2$ כאשר a^2 היא השונות הגנטית, c^2 היא השונות עבור הסביבה המשותפת, ו־ e^2 היא השונות עבור הסביבה הייחודית. מודל ACE, כמו מודלים רבים במדעי החברה, שואף להסביר את השונות של האוכלוסייה כמורכבת מרכיבים שונים. באפן דומה לטכניקת ניתוח שונות (ANOVA), מודל ACE נותן אפשרות לעשות השוואה וניתוח שונויות בתוך הקבוצות וגם בין הקבוצות.

מחקרי תאומים, לעומת מחקרים שעוסקים בגנטיקה והתנהגות בין קשרי משפחה כלליים, מאפשרים אמידה פשוטה יותר למרכיבי השונות השונים. זה נובע כי תאומים זהים או מונוזיגוטיים (MZ) חולקים כמעט 100% מהגנים שלהם, לעומת תאומים לא זהים או דיזוגוטים (DZ) שחולקים רק כ־ 50% מהגנים שלהם. נוכל לדעת מהמידע הזה על ההבדל בקורלציה בין זוג תאומים זהים לבין זוג תאומים לא זהים. החלק של השונות הגנטית, a^2, בקורלציה בין תאומים זהים הוא 1, לעומת החלק של a^2 בקורלציה בין תאומים לא זהים הוא 0.5. החלק של השונות המוסברת על ידי הסביבה המשותפת, c^2, בקורלציה הוא 1 לשני סוגי התאומים DZ ו־ MZ, והמרכיב e^2 בין כל זוג תאומים לא משפיע על הקורלציה. מאחר ויש בידינו את המידע הזה עבור תאומים, אמידת הפרמטרים תהיה פשוטה יותר מאשר אמידת פרמטרים של שוניות במדגם שמכיל קשרים משפחתיים כלליים יותר [11].

תחת הנחה שכל זוג תאומים חיו באותה סביבה, הבדלים בין התאומים מאפשרים לנו לאמוד את מרכיב השונות עבור a^2 בין קבוצות של תאומים זהים ותאומים לא זהים. מהקשרים האלה נוכל לבנות משוואות ולאמוד מספר פרמטרים נמוך מאשר עבור המקרה של קשר משפחתי כללי יותר. אם נסמן את הקורלציה בין תאומים זהים ותאומים לא זהים ב־ r_{MZ} ו־ r_{DZ}, בהתאמה, נוכל להסתכל על המשוואות הבאות כאשר אנו מניחים כי $a^2 + c^2 + e^2 = 1$:

$$\begin{aligned} a^2 &= 2r_{MZ} - 2r_{DZ} \\ c^2 &= 2r_{DZ} - r_{MZ} \\ e^2 &= 1 - r_{MZ} \end{aligned}$$

כלומר נוכל להסתכל על מטריצת השונות עבור תאומים זהים באופן הבא:

$$\Sigma_{MZ} = \sigma^2 \begin{pmatrix} 1 & a^2 + c^2 \\ a^2 + c^2 & 1 \end{pmatrix}$$

ועבור תאומים לא זהים:

$$\Sigma_{DZ} = \sigma^2 \begin{pmatrix} 1 & 0.5a^2 + c^2 \\ 0.5a^2 + c^2 & 1 \end{pmatrix}$$

מחקרי תאומים רבים מניחים כי המידע על סוג התאומים ידוע. אך במציאות במאגרי מידע, האינפורמציה של סוג התאומים לא תמיד נתונה ולעיתים קשה לבדיקה, במיוחד לאחר שהנתונים כבר נאספו.

התעלמות מהמידע החסר הזה יכולה להביא לאמדים מוטים [2]. לכן אנו מנסים להתאים מודל שמתייחס לנתונים של תאומים גם בהיעדר המידע של הסיווג לכל תאום נתון.

בעיה מסוג זה ניתנת לפתרון מתמטי בעזרת תערובת של התפלגויות [10] ,[7]. מודל תערובת של התפלגויות הוא מודל הסתברותי פרמטרי לייצוג שילוב של תת־אוכלוסיות לאוכלוסייה כללית. מודל זה שימושי כאשר יש לנו מדגם שבו לא יודעים לאיזו תת־אוכלוסייה כל תצפית שייכת. במקרה הספציפי שלנו, כאשר אנו מניחים כי הנתונים באים מהתפלגות נורמלית, ניתן להתאים מודל (GMM Gaussian Mixture Model), תערובת של גאוסיאנים.

מודל GMM הינו מודל הסתברותי פרמטרי, שפונקציית הצפיפות בו מיוצגת כסכום משוקלל של צפיפויות גאוסיאניות. אפשר להתייחס לכל מודל כזה, כמודל עם מידע חסר: המידע החסר הוא מאיזה התפלגות באה התצפית. נניח מדגם של n זוגות תאומים $x_i = (x_{i1}, x_{i2})$, $1 \leq i \leq n$ כל זוג מפולג נורמלי, אך הפרמטרים של ההתפלגות תלויים בתת־האוכלוסייה של תאומים זהים ותאומים לא זהים. נסמן $j = 1$ את המקרה של תאומים זהים, ו־ $j = 2$ את המקרה של תאומים לא זהים.
כך ש־ x_i שייך להתפלגות $N(\mu_j, \Sigma_j)$ בהסתברות π_j.
כלומר בהסתברות π_j, $j = 1, 2$

$$x_i \sim N(\mu_j, \Sigma_j)$$

עם הצפיפות:

$$f_j(x_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu_j)^t \cdot \Sigma_j^{-1} \cdot (x_i - \mu_j)\right)$$

עבור $j = 1$ התאומים זהים (MZ), והשונות והתוחלת הם (על פי מודל ACE):

$$\mu_1 \equiv \mu \quad , \quad \Sigma_{MZ} = \sigma^2 \begin{pmatrix} 1 & a^2 + c^2 \\ a^2 + c^2 & 1 \end{pmatrix}$$

עבור $j = 2$ התאומים לא זהים (DZ), והשונות והתוחלת הם (על פי מודל ACE):

$$\mu_2 \equiv \mu \quad , \quad \Sigma_{DZ} = \sigma^2 \begin{pmatrix} 1 & 0.5a^2 + c^2 \\ 0.5a^2 + c^2 & 1 \end{pmatrix}$$

כאשר $a^2 + c^2 + e^2 = 1$.
בפרק הבא נציג את פונקציית הנראות המתאימה לנתונים כאלה, עם ההנחה שסוג התאומים נתון. פונקציה זו תהיה הנראות המלאה בנוסף נציג את הנראות עבור המצב שאין בידינו את כל המידע של סוג התאומים. פונקציה זו תהיה הנראות עבור מידע חסר.

3 פונקציות הנראות

3.1 נראות מירבית (Maximum Likelihood)

גישת הנראות המירבית היא אחת הגישות הסטטיסטיות השימושיות ביותר לאמידת פרמטרים. בשיטה זו מתאימים מודל, ואומדים את הפרמטרים על פי מדגם מקרי של נתונים. הגישה של נראות מירבית מחפשת את הפרמטר שמסביר את הנתונים באופן הטוב ביותר מתוך כל הפרמטרים האפשריים. כלומר, אומד הנראות המירבית הוא הפרמטר שנותן את ההסתברות הגבוהה ביותר (המקסימלית) לקבלת המדגם במודל.
נניח כי יש לנו מודל פרמטרי, כאשר $f(\vec{x}|\theta)$ היא פונקציית הצפיפות או ההסתברות לנתונים $\vec{x}$, עם ווקטור הפרמטרים θ, השייכים לקבוצת הפרמטרים Θ. נגדיר את פונקציית הנראות באופן הבא:

הגדרה 3.1 פונקציית הנראות (the likelihood function): בהינתן ערכים של נתונים $\vec{x} = (x_1, ..., x_n)^t$, פונקציית הנראות $L(\theta|\vec{x})$ מוגדרת להיות הצפיפות (או ההסתברות) המשותפת: $L(\theta; x_1, x_2, ..., x_n) = f_{x_1,x_2,...,x_n}(x_1, ...x_n; \theta)$. בפרט, אם $x_1, ..., x_n$ הינו מדגם מקרי מצפיפות (או הסתברות) $f(\cdot; \theta)$ אזי פונקציית הנראות הינה:

$$L(\theta; x_1, x_2, ..., x_n) = f_{x_1,x_2,...,x_n}(x_1, ...x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

הגדרה 3.2 פונקציית לוג הנראות (the loglikelihood function) : פונקציית לוג הנראות, $l(\theta|\vec{x})$, מוגדרת להיות הלוגריתם הטבעי של פונקציית הנראות:

$$l(\theta|\vec{x}) = \ln\left(L(\theta; x_1, x_2, ..., x_n)\right)$$

הגדרה 3.3 אומד נראות מירבית: (A maximum likelihood (ML) estimator of θ): אומד הנראות המירבית הוא הערך של θ שממקסם את פונקציית הנראות, או באופן שקול, ממקסם את פונקציית הלוג נראות.

בפרק הקודם דנו בשני סוגי נתוני תאומים: נתונים בהם מופיע סוג התאומים באופן מלא, ונתונים בהם סוג התאומים חסר לפחות לחלק מהתצפיות. בתתי הפרקים הבאים נציג קודם את הנראות המלאה עבור נתונים עם מידע מלא, ולאחר מכן נציג את הנראות עבור המידע החסר.

3.2 הנראות עבור המודל המלא

נתון מדגם של n זוגות של תאומים $x_i = (x_{i1}, x_{i2})$, $1 \le i \le n$ כך שכל זוג מפולג נורמלי. ידוע לנו הסוג של כל זוג תאומים. נניח כי מספר התאומים הזהים הוא k.
לכן הנראות עבור התאומים הזהים היא:

$$\begin{aligned} L_{MZ}(x_1, ...x_k; \Sigma_{MZ}, \mu) &= \prod_{i=1}^{k} f_1(x_i; \Sigma_{MZ}, \mu) \\ &= \prod_{i=1}^{k} \frac{1}{(2\pi)^{1/2}|\Sigma_{MZ}|^{1/2}} \exp\left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma_{MZ}^{-1} \cdot (x_i - \mu)\right) \end{aligned}$$

ועבור התאומים הלא זהים:

$$\begin{aligned} L_{DZ}(x_{k+1},...x_n;\Sigma_{DZ},\mu) &= \prod_{i=k+1}^{n} f_2(x_i;\Sigma_{DZ},\mu) \\ &= \prod_{i=k+1}^{n} \frac{1}{(2\pi)^{1/2}|\Sigma_{DZ}|^{1/2}} \exp\left(-\frac{1}{2}(x_i-\mu)^t\cdot\Sigma_{DZ}^{-1}\cdot(x_i-\mu)\right) \end{aligned}$$

בשימוש באי תלות בין הנתונים נוכל לקבל את הנראות עבור כל המדגם:

$$L(\vec{x};\{\mu,a^2,c^2,\sigma^2\}) = \prod_{i=1}^{k} f_1(x_i;\Sigma_{MZ},\mu)\cdot\prod_{i=k+1}^{n} f_2(x_i;\Sigma_{DZ},\mu)$$

ולוג הנראות:

$$l(\vec{x};\{\mu,a^2,c^2,\sigma^2\}) = \sum_{i=1}^{k}\log\left(f_1(x_i;\Sigma_{MZ},\mu)\right) + \sum_{i=k+1}^{n}\log\left(f_2(x_i;\Sigma_{DZ},\mu)\right)$$

אמידת הפרמטרים במצב הזה היא יחסית פשוטה אף שאין ביטוי סגור. על מנת למקסם את $l(\vec{x},\{\mu,a^2,c^2,\sigma^2\})$ יש לגזור את הפונקציה עבור כל פרמטר ולהשוות לאפס (ראה פיתוח בנספח 8.1).

3.3 הנראות עבור המודל החסר

נציג את פונקציית הנראות עבור המודל תחת ההנחה שאין בידינו את כל המידע עבור סוג התאומים. כפי שהצגנו בפרק קודם ננסה להסתכל על הבעיה הזו עם המידע החסר, כתערובת של התפלגויות. נסמן את קבוצת הפרמטרים ב־ Θ. עבור כל זוג תאומים פונקציית הנראות נתונה על ידי:

$$L_i\left(\vec{x};\{\mu,a^2,c^2,\sigma^2,\vec{\pi}\}\right) = \pi_1 f_1(x_i,\theta)+\pi_2 f_2(x_i,\theta) = \sum_{j=1}^{2}\pi_j f_j(x_i,\theta)$$

כאשר נסמן: $\theta=\{\mu,a^2,c^2,\sigma^2,\pi\}$.
מתקיים גם $\sum_{j=1}^{2}\pi_j = 1$ כלומר: $\pi_2 = 1-\pi_1$, לכן אפשר לכתוב:

$$L_i\left(x_i;\{\mu,a^2,c^2,\sigma^2,\pi_1\}\right) = \pi_1 f_1(x_i,\theta)+(1-\pi_1)f_2(x_i,\theta)$$

כאשר $f_1(x_i,\theta)$ ו־ $f_2(x_i,\theta)$ תלויות רק בפרמטרים $\{\mu,a^2,c^2,\sigma^2\}$.
עבור תצפית של זוג תאומים זהים (MZ), מתקיים: $\pi_1=1$ ולכן $\pi_2=0$;

$$L_{MZ}\left(x_i;\{\mu,a^2,c^2,\sigma^2\}\right) = f_1(x_i,\theta)$$

ועבור תצפית של זוג תאומים לא זהים (DZ), מתקיים: $\pi_1=0$ ולכן $\pi_2=1$;

$$L_{DZ}\left(x_i;\{\mu,a^2,c^2,\sigma^2\}\right)=f_2(x_i,\theta)$$

עבור תצפית של זוג תאומים לא ידוע מה הסוג (NA) נקבל תערובת:

$$L_{NA}\left(x_i;\{\mu,a^2,c^2,\sigma^2,\pi_1\}\right)=\pi_1 f_1(x_i,\theta)+(1-\pi_1)f_2(x_i,\theta)$$

כאשר הפרמטרים הם: $\theta=\{\mu,a^2,c^2,\sigma^2,\pi_1\}$.

לכן הנראות של כל המדגם:

$$L\left(\vec{x};\{\mu,a^2,c^2,\sigma^2,\pi_1\}\right)=\prod_{i=1}^{n}\left(\pi_1 f_1(x_i,\theta)+(1-\pi_1)f_2(x_i,\theta)\right)$$

ולוג הנראות:

$$l\left(\vec{x};\{\mu,a^2,c^2,\sigma^2,\pi_1\}\right)=\sum_{i=1}^{n}\log\left(\pi_1 f_1(x_i,\theta)+(1-\pi_1)f_2(x_i,\theta)\right)$$

המטרה למצוא אומדים לפרמטרים על פי נתוני המדגם. מחפשים את האומד שממקסם את הפונקציה $L(\Theta)$. פונקציה כזו היא פונקציה לא לינארית בפרמטרים. מציאת אומדי נראות מירבית לא מתאפשרת כפתרון אנליטי. לכן נצטרך לפתח שיטות לפתרון בעיה מסוג זה.
בפרק הבא נציג אלגוריתם Expectation–Maximization (EM) ואת ההרחבה ECM, Expectation-Conditional Maximization) של אלגוריתם זה. שיעזרו לנו בפתרון בעיית מציאת אומד הנראות המירבית.

4 אלגוריתם Expectation–Maximization Algorithm (EM)

4.1 הקדמה

אלגוריתם EM [6] הוא שיטה איטרטיבית, השימושית בין השאר למציאת אומדי נראות מירבית עבור מידע חסר [1]. הרעיון העומד מאחורי EM, הוא לקחת את הנתונים הנצפים, להניח שיש לנו את הנתונים החסרים ואז להסתכל על כל הנתונים ביחד. נניח מדגם $\{x_1, x_2, ..., x_n\}$, מדגם נצפה שאינו בהכרח מלא כלומר ישנו מידע חסר לגבי נתונים אלה, נניח $\{y_1, y_2, ..., y_n\}$. במקרה שלנו $x_i = (x_{i1}, x_{i2})$ הוא ערכים עבור זוג של תאומים ו־y_i הוא סוג התאומים. באלגוריתם EM מניחים כי ההתפלגות של המדגם הנצפה ידועה, כלומר $P(X|Y,\theta)$ היא ידועה. במצב זה יש לאמוד את θ מתוך קבוצת הפרמטרים Θ, תוך התחשבות בנתונים החסרים. הרעיון שעומד מאחורי שיטת EM הוא להסתכל במדגם (X,Y) (המידע המלא) ולהשתמש בקשר:

$$P\left((X,Y)|\theta\right) = P\left(Y|X,\theta\right) \cdot P\left(X|\theta\right)$$

ממנו נוכל לקבל את המשוואה:

$$\log\left[P\left(X|\theta\right)\right] = \log\left[P\left((X,Y)|\theta\right)\right] - \log\left[P\left(Y|X,\theta\right)\right] \tag{1}$$

המטרה למצוא את הפרמטרים שממקסמים את הפונקציה הזו. על מנת למקסם את הפונקציה נסתכל בתוחלת של אגף ימין של (1), וננסה לחפש את המקסימום שלה (להרחבה, ראה נספח 8.2).
כדי למצוא את המקסימום, נגדיר את הפונקציה הבאה:

$$Q\left(\theta,\theta^{(m)}\right) = E_{\left(Y|X,\theta^{(m)}\right)}\left[\log\left(L\left((X,Y)|\theta\right)\right)|X,\theta^{(m)}\right] = \int \log\left(L\left((X,Y)|\theta\right)\right) \cdot f\left(Y|X,\theta^{(m)}\right) dY$$

שבעזרתה מבצעים את האלגוריתם. האלגוריתם מתבצע בשני צעדים והם:
צעד ה־E: חישוב תוחלת הנראות כלומר $Q\left(\theta,\theta^{(m)}\right)$ כפונקציה של θ.
צעד ה־M: מציאת אומד של θ שממקסם את $Q\left(\theta,\theta^{(m)}\right)$, כלומר, מחפשים $\theta^{(m+1)}$ כך ש־

$$Q\left(\theta^{(m+1)},\theta^{(m)}\right) \geq Q\left(\theta,\theta^{(m)}\right), \text{ לכל } \theta.$$

האלגוריתם מתבצע באופן רקורסיבי בתחילה בוחרים $\theta^{(0)}$, ועל סמך $\theta^{(0)}$ מבצעים את שני הצעדים של האלגוריתם ומחשבים את $\theta^{(1)}$. ממשיכים כך, משתמשים ב־ $\theta^{(m)}$ לאמידת $\theta^{(m+1)}$ עד להתכנסות. אלגוריתם EM כפי שהוגדר לא תמיד עובד במצבים שיש אילוצים על הפרמטרים (להרחבה ראה [5]). המודל שאנו מציגים, כפי שהוגדר בפרק 2, מניח כי התוחלות בשני המצבים עבור תאומים זהים ולא זהים הן שוות וגם כי ישנו קשר בין השונויות. אפשר להסתכל על הקשרים האלה באופן הבא:

$$\mu_1 = \mu_2 = \mu \quad ; \quad \Sigma_{MZ} \succ \Sigma_{DZ}$$

הקשרים האלה הינם אילוצים שנצטרך להתחשב בהם בעת חיפוש הפתרון, לכן שימוש באלגוריתם EM נאיבי הוא לא מספיק. הסיבה לכך היא שצעד ה־M הופך למסובך מדי, ולא ניתן לקבל נוסחה

סגורה לפרמטרים. לכן, נצטרך להשתמש באחת ההרחבות של אלגוריתם EM שנקראת אלגוריתם ECM, Expectation-Conditional Maximization.

אלגוריתם ECM הינו מחלקה של אלגוריתמים שהוצגו על ידי החוקרים Rubin ו־Meng [9]. אלגוריתם ECM מחליף את צעד ה־M "המסובך" בצעדים יותר פשוטים שבהם מעדכנים את הפרמטרים באופן הדרגתי שנקראים צעדי CM-Conditional Maximization, או בעברית מקסום מותנה. זאת אומרת, במקום לחשב את ווקטור הפרמטרים שממקסם את הפונקציה בצעד ה־M בשלב אחד, מחלקים את הצעד הזה ל־S צעדים. בכל צעד מחשבים פרמטר אחד שממקסם את הפונקציה בהינתן שאר הפרמטרים, כלומר מקבעים את כל הפרמטרים פרט לאחד, ומחפשים את המקסימום עבור אותו הפרמטר. במקרה שלנו, נחלק את צעד ה־M לשני צעדים, כלומר $S = 2$. בצעד ראשון, נעדכן את התוחלת והמשקולות, לאחר מכן נעדכן את השונות על סמך העדכון שנעשה לתוחלת וכך הלאה. בצעד ה־M של אלגוריתם זה, לאחר העדכון הראשון, חוזרים ומחשבים את הפונקציה שבצעד ה־E על סמך העדכון הראשון שנעשה. כלומר צעד ה־M מכיל בתוכו את צעד ה־E. לכן אלגוריתם ECM מצריך חישוב אחד נוסף יותר מאלגוריתם EM, ולפיכך התכנסותו יכולה להיות איטית יותר מהתכנסות אלגוריתם EM (ולהרחבה על התכנסות שני האלגוריתמים ראה [13], [8]).

4.2 הפעלת אלגוריתם EM ואלגוריתם ECM

בבעיה שלנו אנו חושבים על המודל כתערובת של התפלגויות. בפרט, יש לנו את המדגם הנצפה $\{x_1, x_2, ..., x_n\}$, כך ש־ $x_i = (x_{i1}, x_{i2})$ $1 \leq i \leq n$ וישנו מידע חסר לגבי נתונים אלה שהוא הסוג של כל אחד מהתאומים $\{y_1, y_2, ..., y_n\}$. עבור המידע המלא, נסמן את האינדיקטור z_{ij} באופן הבא:

$$1 \leq j \leq 2 \text{ עבור } ,z_{ij} = \begin{cases} 1 & x_i \sim N(\mu_j, \Sigma_j) \\ 0 & otherwise \end{cases}$$

כלומר x_i זוג תאומים שייך להתפלגות $N(\mu_1, \Sigma_1)$ בהסתברות π_1, ושייך להתפלגות $N(\mu_2, \Sigma_2)$ בהסתברות $\pi_2 = 1 - \pi_1$, כאשר: $\Sigma_1 \equiv \Sigma_{MZ}$ ו־ $\Sigma_2 \equiv \Sigma_{DZ}$.

ההתפלגות של המדגם הנצפה ידועה, כלומר $P(X|Y, \theta)$ הינה התפלגות נורמלית. במצב זה יש לאמוד את θ מקבוצת הפרמטרים $\Theta = \{(\mu, a^2, \sigma^2, c^2, \pi_1)\}$, תחת האילוצים $\mu_1 = \mu_2$ ו־ $\Sigma_{MZ} \succ \Sigma_{DZ}$, ותוך התחשבות בנתונים החסרים.

נפעיל את האלגוריתם:

בונים את הפונקציה:

$$
\begin{aligned}
Q(\theta,\theta^{(m)}) &= E_{Z|X,\theta^{(m)}}\left[\log\left(L(Z,X|\Theta)|X,\theta^{(m+1)}\right)\right] \\
&= E_{Z|X,\theta^{(m)}}\left[\log\prod_{i=1}^{n}\left(L(\theta,x_i,z_i)\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2}P\left(z_{ij}|x_i,\theta^{(m)}\right)\log\left(\pi_j f_j(x_{i,},\theta)\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{2}P\left(z_{ij}|x_i,\theta^{(m)}\right)\left[\log\pi_j-\frac{1}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_j|-\frac{1}{2}\left(x_i-\mu_j\right)^t\cdot\Sigma_j^{-1}\cdot\left(x_i-\mu_j\right)\right]
\end{aligned}
\tag{2}
$$

כאשר על פי נוסחת בייס:

$$
P(z_{ij}|x_i,\theta^{(m)})=\frac{\pi_j^{(m)}f_j(x_i|\theta^{(m)})}{\pi_1^{(m)}f_1(x_i|\theta^{(m)})+\pi_2^{(m)}f_2(x_i|\theta^{(m)})}
$$

נסמן: $p_{ij}^{(m)}=P(z_{ij}|x_i,\theta^{(m)})$

כלומר נקבל כי,

$$
p_{i1}^{(m)}=P(z_{i1}|y_i,\theta^{(m)})=\frac{\pi_1^{(m)}f_1(x_i|\theta^{(m)})}{\pi_1^{(m)}f_1(x_i|\theta^{(m)})+(1-\pi_1^{(m)})f_2(x_i|\theta^{(m)})} \tag{3}
$$

וגם,

$$
p_{i2}^{(m)}=P(z_{i2}|y_i,\theta^{(m)})=\frac{\pi_2^{(m)}f_2(x_i|\theta^{(m)})}{\pi_1^{(m)}f_1(x_i|\theta^{(m)})+(1-\pi_1^{(m)})f_2(x_i|\theta^{(m)})}=1-p_{i1}^{(m)}
$$

לכן הפונקציה (2),

$$
\begin{aligned}
Q\left(\theta,\theta^{(m)}\right) &= \sum_{i=1}^{n}p_{i1}^{(m)}\left[\log\pi_1-\frac{1}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_1|-\frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_1^{-1}\cdot\left(x_i-\mu\right)\right] \\
&\quad+\sum_{i=1}^{n}p_{i2}^{(m)}\left[\log\pi_2-\frac{1}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_2|-\frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_2^{-1}\cdot\left(x_i-\mu\right)\right] \\
&= \sum_{i=1}^{n}p_{i1}^{(m)}\left[\log\pi_1-\frac{1}{2}\log 2\pi-\frac{1}{2}]\log|\Sigma_1|-\frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_1^{-1}\cdot\left(x_i-\mu\right)\right] \\
&\quad+\sum_{i=1}^{n}\left(1-p_{i1}^{(m)}\right)\left[\log(1-\pi_1)-\frac{1}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_2|-\frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_2^{-1}\cdot\left(x_i-\mu\right)\right]
\end{aligned}
\tag{4}
$$

נאמוד את הפרמטרים על ידי גזירה של הפונקציה $Q\left(\theta,\theta^{(m)}\right)$:

נתחיל מהפרמטר של המשקולות π_1 שלא תלוי בתוחלת ולא בשונות. נגזור את (4) לפי π_1 ונקבל:

$$d\frac{\Theta\left(\theta,\theta^{(m)}\right)}{d\pi_1} = \sum_{i=1}^{n} p_{i1}^{(m)} \cdot \tfrac{1}{\pi_1} - \sum_{i=1}^{n}\left(1-p_{i1}^{(m)}\right)\cdot\tfrac{1}{1-\pi_1}$$

(לדרך כללית יותר לקבלת האומד למשקולות ראה נספח 8.3)

נשווה לאפס ונחלץ את π_1:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} p_{i1}^{(m)} \cdot \frac{1}{\pi_1} - \sum_{i=1}^{n}\left(1-p_{i1}^{(m)}\right)\cdot\frac{1}{1-\pi_1} \\
\Leftrightarrow 0 &= (1-\pi_1)\cdot\sum_{i=1}^{n} p_{i1}^{(m)} - \pi_1\cdot\sum_{i=1}^{n}\left(1-p_{i1}^{(m)}\right) \\
\Leftrightarrow 0 &= \sum_{i=1}^{n} p_{i1}^{(m)} - \pi_1\cdot\sum_{i=1}^{n} p_{i1}^{(m)} - \pi_1 n + \pi_1\cdot\sum_{i=1}^{n} p_{i1}^{(m)} \\
\Leftrightarrow \pi_1 n &= \sum_{i=1}^{n} p_{i1}^{(m)}
\end{aligned}$$

$$\boxed{\pi_1^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{i1}^{(m)}}$$

כלומר:

$$\begin{aligned}
\pi_1^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n} p_{i1}^{(m)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_1^{(m)} f_1(x_i|\theta^{(m)})}{\sum_{r=1}^{2}\pi_r^{(m)} f_r(x_i|\theta^{(m)})} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\pi_1^{(m)} f_1(x_i|\theta^{(m)})}{\pi_1^{(m)} f_1(x_i|\theta^{(m)}) + \left(1-\pi_1^{(m)}\right) f_2(x_i|\theta^{(m)})}
\end{aligned}$$

עד עכשיו קיבלנו נוסחה לחישוב המשקולות. נעבור מכאן לחישוב התוחלת תוך התחשבות באילוץ. נגזור את (4) לפי $\vec{\mu}$ ונקבל:

$$d\frac{Q\left(\theta,\theta^{(m)}\right)}{d\mu} = \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\tfrac{1}{2}\cdot(-2)\Sigma_j^{-1}\left(x_i-\mu\right)$$

נשווה לאפס ונחלץ $\vec{\mu}$:

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}x_i - \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}\mu$$

נחשב:

$$\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}x_i = \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}\mu$$

נקבל:

$$\mu^{(m+1)} = \left(\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}x_i\right)$$

ולכן נסכם:

$$\begin{aligned}\mu^{(m+1)} &= \left(\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}x_i\right)\\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}x_i}{\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i,\theta^{(m)}\right)\Sigma_j^{-1}}\end{aligned}$$

קיבלנו נוסחה סגורה לתוחלת.

עבור המקרה שלנו מציאת אומד לכל אחד מהפרמטרים: a^2, c^2 ו־ σ^2, או לשונות באופן כללי, אינה מתאפשרת מבחינה אנליטית. אם נגזור את הפונקציה (4) לפי הפרמטרים של השונות לא נוכל לחלץ את הפרמטרים ולקבל נוסחה סגורה בגלל התלות בין שתי השונויות Σ_1, Σ_2. לכן נשתמש בשיטות נומריות של הנראות. לקבלת אומדים ל־a^2, c^2 ו־ σ^2.

4.2.1 סיכום האלגוריתם

נתון המדגם $\vec{x} = (x_1, ..., x_n)$, שמכיל n זוגות של תאומים. נרצה להתאים לו מודל של תערובת התפלגויות, ולאמוד את הפרמטרים של הנראות. כלומר יש לנו $(x_1, ..., x_n)$ זוגות של תאומים. מסתכלים על זה כמטריצה $X_{n\times 2}$, (כל תצפית הינה זוג של תאומים). משלימים את הנתונים האלה בווקטור $Y_{n\times 1}$, כך שווקטור זה מייצג את הסוג של התאומים. y_i שווה ל־MZ, DZ או NA.

משתמשים באלגוריתם ECM:

צעד ה־ E:

מחשבים את הפונקציה:

$$\begin{aligned}&Q\left(\theta,\theta^{(m)}\right) =\\ &=\sum_{i=1}^{n} p_{ij}^{(m)}\left[\log\pi_1 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x_i-\mu)^t\cdot\Sigma_1^{-1}\cdot(x_i-\mu)\right]\\ &+\sum_{i=1}^{n}\left(1-p_{ij}^{(m)}\right)\left[\log(1-\pi_1) - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x_i-\mu)^t\cdot\Sigma_2^{-1}\cdot(x_i-\mu)\right]\end{aligned}$$

כאשר:

$$p_{i1}^{(m)} = P(z_{i1}|x_i, \theta^{(m)}) = \frac{\pi_1^{(m)} f_1(x_i|\theta^{(m)})}{\pi_1^{(m)} f_1(x_i|\theta^{(m)}) + (1-\pi_1^{(m)}) f_2(x_i|\theta^{(m)})}$$

צעד CM הראשון:

מחשבים את המשקולות:

$$\pi_1^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_1^{(m)} f_1(x_i|\theta^{(m)})}{\pi_1^{(m)} f_1(x_i|\theta^{(m)}) + \left(1-\pi_1^{(m)}\right) f_2(x_i|\theta^{(m)})}$$

ואת התוחלת:

$$\mu^{(m+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i, \theta^{(m)}\right)\left(\Sigma_j^{(m)}\right)^{-1} x_i}{\sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|x_i, \theta^{(m)}\right)\left(\Sigma_j^{(m)}\right)^{-1}}$$

צעד CM השני:

נחזור ונעדכן את הפונקציה בצעד ה־ E, כלומר נחשב את (3) ו־(4) עם העידכון החדש של התוחלת והמשקולת. ונקבל:

$$p_{i1}^{(m+\frac{1}{2})} = P(z_{i1}|x_i, \theta^{(m)}) = \frac{\pi_1^{(m+1)} f_1(x_i|\theta^{(m+\frac{1}{2})})}{\pi_1^{(m+1)} f_1(x_i|\theta^{(m+\frac{1}{2})}) + \pi_2^{(m+1)} f_2(x_i|\theta^{(m+\frac{1}{2})})}$$

$$Q\left(\theta, \theta^{(m+\frac{1}{2})}\right) =$$

$$= \sum_{i=1}^{n} p_{i1}^{(m+\frac{1}{2})} \left[\log \pi_1 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x_i-\mu)^t \cdot \Sigma_1^{-1} \cdot (x_i-\mu)\right]$$

$$+ \sum_{i=1}^{n}\left(1 - p_{i1}^{(m+\frac{1}{2})}\right)\left[\log(1-\pi_1) - \frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(x_i-\mu)^t \cdot \Sigma_2^{-1} \cdot (x_i-\mu)\right]$$

חישוב הפרמטרים $(a^2)^{(m+1)}, (c^2)^{(m+1)}, (\sigma^2)^{(m+1)}$ שממקסמים את הפונקציה $Q\left(\theta, \theta^{(m+\frac{1}{2}))}\right)$ מתבצע על ידי שיטה נומרית.

ולכן הלולאה שנבנה תסתכם בשלבים הבאים:

1) בוחרים תנאים התחלתיים $\mu^{(0)}, \pi_1^{(0)}, \Sigma_j^{(0)}$, כאשר $\Sigma_j^{(0)}$ תלויה ב־ $a^{2^{(0)}}, c^{2^{(0)}}, \sigma^{2^{(0)}}$.

2) בונים את הפונקציה: $Q\left(\theta, \theta^{(m)}\right)$ ומחשבים:

$$p_{ij}^{(m)} = P(z_{ij}|x_i, \theta^{(m)}) = \frac{\pi_j^{(m)} f_j(x_i|\theta^{(m)})}{\sum_{r=1}^{k} \pi_r^{(m)} f_r(x_i|\theta^{(m)})}$$

3) מעדכנים את הפרמטר $\vec{\pi}$:

$$\pi_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(m)}$$

4) מעדכנים את התוחלת $\vec{\mu}$:

$$\mu^{(m+1)} = \frac{\sum_i \sum_j p_{ij}^{(m)} \left(\Sigma_j^{(m)}\right)^{-1} x_i}{\sum_i \sum_j p_{ij}^{(m)} \left(\Sigma_j^{(m)}\right)^{-1}}$$

5) חוזרים ומעדכנים את p_{ij}:

$$p_{ij}^{(m+\frac{1}{2})} = P(z_{ij}|x_i, \Sigma_j^{(m)}, \mu^{(m+1)}, \pi_j^{(m+1)}) = \frac{\pi_j^{(m+1)} f_j(x_i|\Sigma_j^{(m)}, \mu^{(m+1)})}{\sum_{r=1}^{2} \pi_r^{(m+1)} f_r(x_i|\Sigma_r^{(m)}, \mu^{(m+1)})}$$

6) ממקסמים את הפונקציה $Q\left(\theta, \theta^{(m+\frac{1}{2})}\right)$ בעזרת שיטות נומריות (למשל פונקציית Optim ב־R) ומקבלים את האומדים: $(a^2)^{(m+1)}, (c^2)^{(m+1)}, (\sigma^2)^{(m+1)}$.

7) חוזרים לשלב 2 עד לקירוב הרצוי.

5 סימולציות

5.1 הפעלת הסימולציות

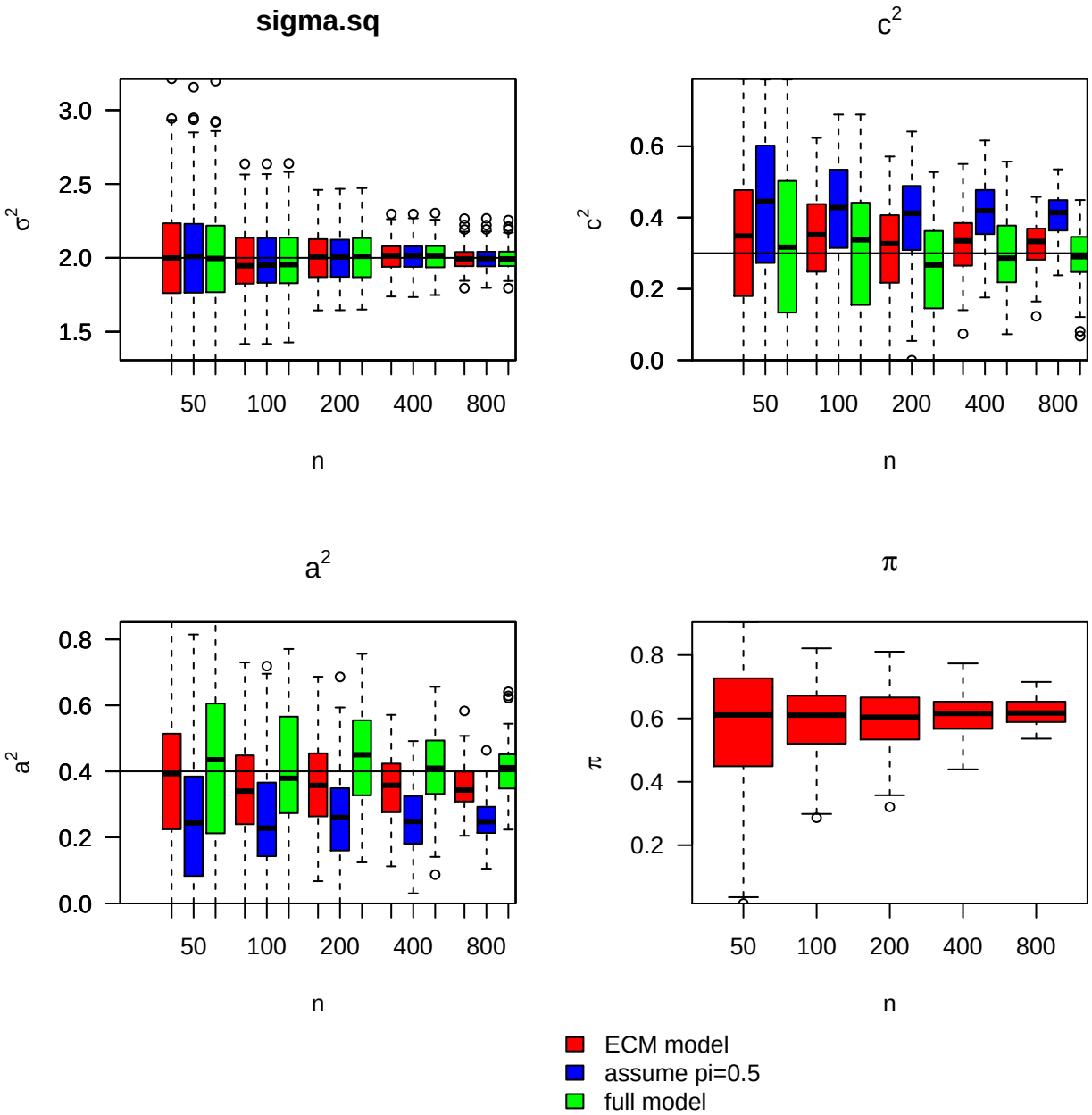
לבדיקת המודל שהצענו, ביצענו מספר סימולציות. כל הסימולציות והחישובים לבדיקה התבצעו בתוכנת R. לחישוב הפונקציה נעזרנו בחבילת mvtnorm [3]. נציג את הסימולציות עבור שני תרחישים.

5.1.1 התרחיש הראשון

בתרחיש הראשון הגרלנו מדגם אקראי של נתונים, כך שבממוצע חצי מהתאומים יהיו זהים והחצי השני לא זהים, באופן הבא: הגרלנו 30% תאומים זהים ו־ 30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כי 50% הם זהים (MZ) ו־ 50% הם לא זהים (DZ). ביצענו מספר סימולציות עבור תרחיש זה: בכל פעם הגרלנו את הנתונים מהתפלגות שבה קבענו ערכים שונים לפרמטרים: μ, a^2, c^2 ו־ σ^2. לכל ערך של ווקטור הפרמטרים שקבענו הגרלנו נתונים עבור גדלי המדגם הבאים:
$n = 50, 100, 200, 400, 800$. עבור כל n הרצנו את הסימולציה 100 פעמים עבור שלוש שיטות. בשיטה הראשונה הנחנו עבור חלק מהנתונים שסוג התאומים נתון וידוע, ועבור חלק מהנתונים שהסוג לא ידוע. לכן השתמשנו בשיטה של המידע החסר שהוצעה בתת־פרק 4.2 שבו אמדנו את חמשת הפרמטרים: μ, σ^2, a^2, c^2 ו־π_1. השיטה השנייה היא השיטה שבה עבור התאומים שלא ידוע הסוג שלהם הנחנו כי $\pi_1 = 0.5$, כלומר הנחנו משקולות נתונות מראש, ואמדנו את הפרמטרים μ, a^2, c^2 ו־ σ^2. השיטה השלישית היא אמידת נראות מירבית עבור המודל המלא. בשיטה זו, בניגוד לשתי השיטות הקודמות המידע הכיל את סוג התאומים עבור כל התצפיות. השיטה השלישית נועדה לצורך השוואה.
באיור 1 אפשר לראות את התוצאות בתרשימי BOXPLOTS עבור הפרמטרים:
$a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. ניתן לראות שהאומד של σ^2 הופך למדויק יותר ככל שמגדילים את גודל המדגם. האומדים לפרמטרים a^2 ו־ c^2 פחות קרובים מאשר האומד של σ^2, אבל ניתן לראות שהשיטה הראשונה שבה לקחנו את הנראות עבור המידע החסר (הקופסא השמאלית האדומה) טובה יותר מהשיטה שבה הנחנו $\pi_1 = 0.5$ (הקופסא האמצעית הכחולה). האומד של μ יצא קרוב מאוד וכמעט זהה בכל התרחישים [1].
הרצנו עוד מספר פעמים את התרחיש הזה, בכל פעם שינינו את הערכים של פרמטרים. באיור 2 התוצאות מופיעות בתרשימי BOXPLOTS עבור הפרמטרים:
$a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. גם בסימולציה זו רואים שהאומד של σ^2 מתקרב ככל שמגדילים את גודל המדגם. האומדים לפרמטרים a^2 ו־ c^2 במצב שבו $a^2 = c^2$ טובים יותר ממה שראינו בסימולציה הקודמת באיור 1.
באיור 3 רואים את ההרצה לאחר שינוי נוסף בערך של σ^2. הפרמטרים שנבחרו כעת הם:
$a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 1.33, \mu{=}5$. שוב נוכל לראות שהאומד של σ^2 כמעט זהה בשלושת בשיטות. אך כאן רואים שהאומדים לפרמטרים a^2 ו־ c^2 פחות טובים ממצבים אחרים, אבל עדיין יותר קרובים בשיטה הראשונה שהוצעה בעבודה מאשר מהשיטה השנייה.

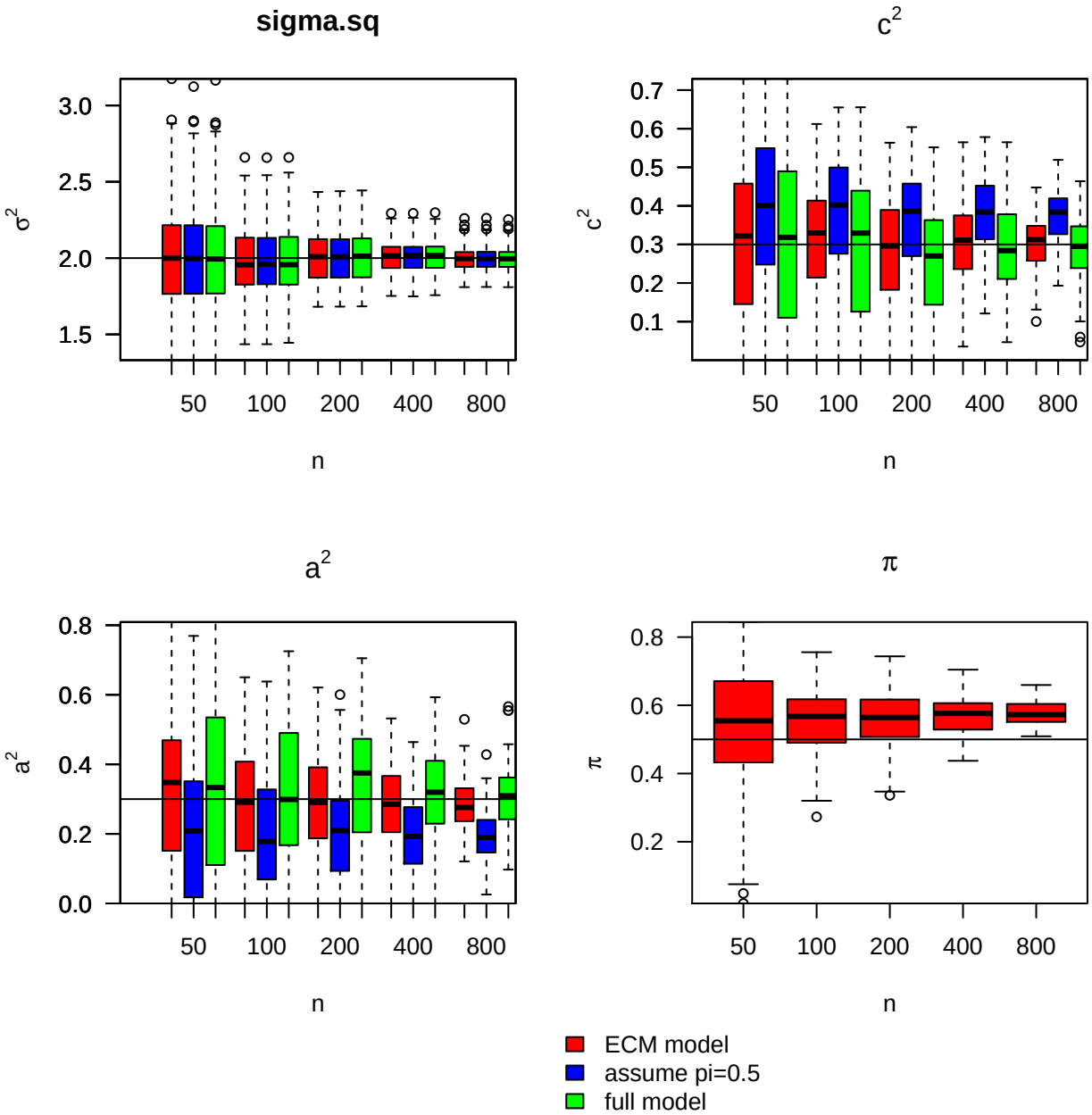
[1] לתרשימי BOXPLOT של μ ראה נספח 8.4

איור 1:



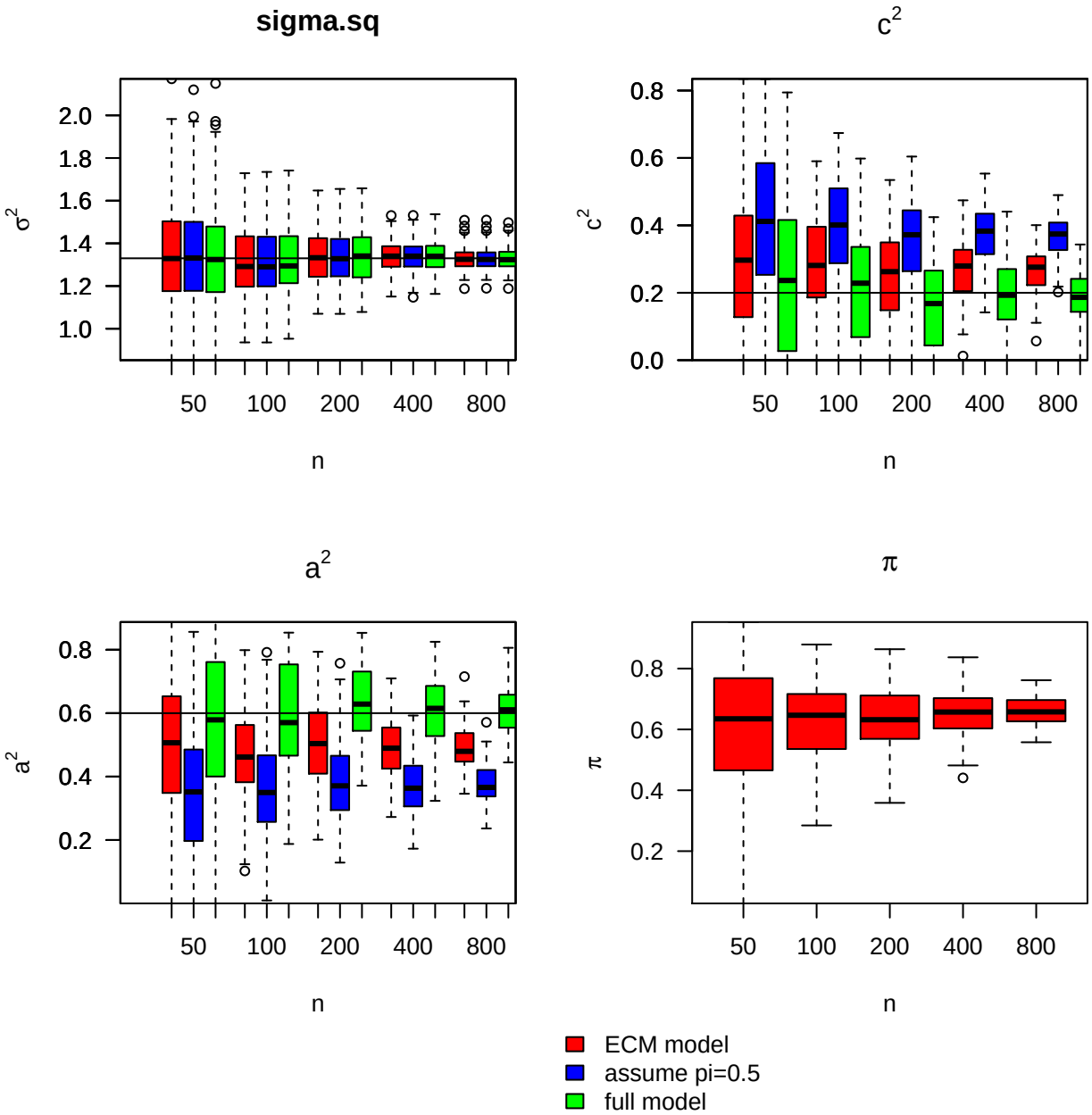
הפרמטרים שנבחרו הם $a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.
הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

איור 2:



הפרמטרים שנבחרו הם $a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$.הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.
הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

איור 3:



הפרמטרים שנבחרו הם $a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 2, \mu{=}5$.
הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.
הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

5.1.2 התרחיש השני

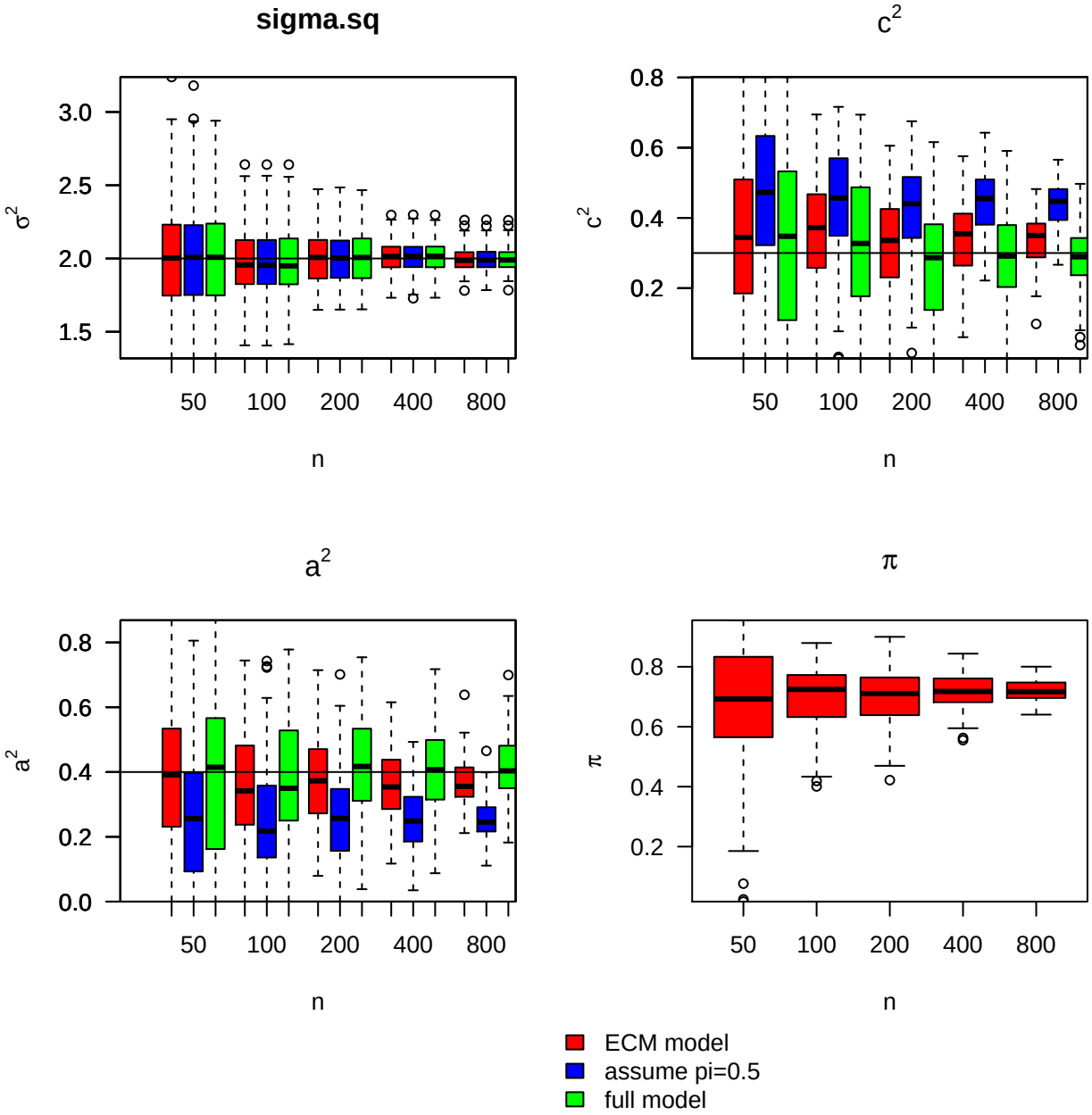
בתרחיש השני יצרנו מדגם אקראי של נתונים, כפי שנעשה בתרחיש הראשון. אך כאן שינינו את אופן קביעת התאומים הזהים והלא זהים. בעוד שבתרחיש הראשון לקחנו כחצי מהנתונים להיות תאומים זהים והחצי השני להיות תאומים לא זהים. כאן שינינו את האחוז הזה באופן הבא. בהתחלה הגרלנו אחוז מסוים של תאומים זהים ואחוז של תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הגרלנו אחוז מסוים שיהיו זהים (MZ) והשאר יהיו לא זהים (DZ). שוב הגרלנו את הנתונים מהתפלגות שבה בחרנו ערכים שונים לפרמטרים: μ, a^2, c^2 ו־ σ^2. ולקחנו ערכים שונים של n: $n = 50, 100, 200, 400, 800$. עבור כל n הרצנו את הסימולציה 100 פעמים עבור שלושה שיטות שנזכרו בתרחיש 1. ניתן לראות מהתוצאות כי האומד של σ^2 הוא קרוב וכמעט זהה ככל שמגדילים את גודל המדגם. האומדים לפרמטרים a^2 ו־ c^2 פחות קרובים אבל עדיין השיטה הראשונה שבו לקחנו את הנראות עבור המידע החסר (הקופסא האדומה השמאלית) טובה יותר מהשיטה השנייה שבה מניחים כי המשקולות ידועות מראש (הקופסא הכחולה האמצעית).

באיור 4 אפשר לראות את תרשימי ה־ BOXPLOTS של תוצאות הנתונים שהוגרלו באופן הבא: 30% תאומים זהים ו־ 30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כי 90% הם זהים (MZ) ו־ 10% הם לא זהים (DZ). עבור הפרמטרים: $a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$.

באיור 5 אפשר לראות את תרשימי ה־ BOXPLOTS של תוצאות הנתונים שהוגרלו באופן הבא: 30% תאומים זהים ו־ 30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כעת כי 10% הם זהים (MZ) ו־ 90% הם לא זהים (DZ). עבור הפרמטרים:
$a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$.

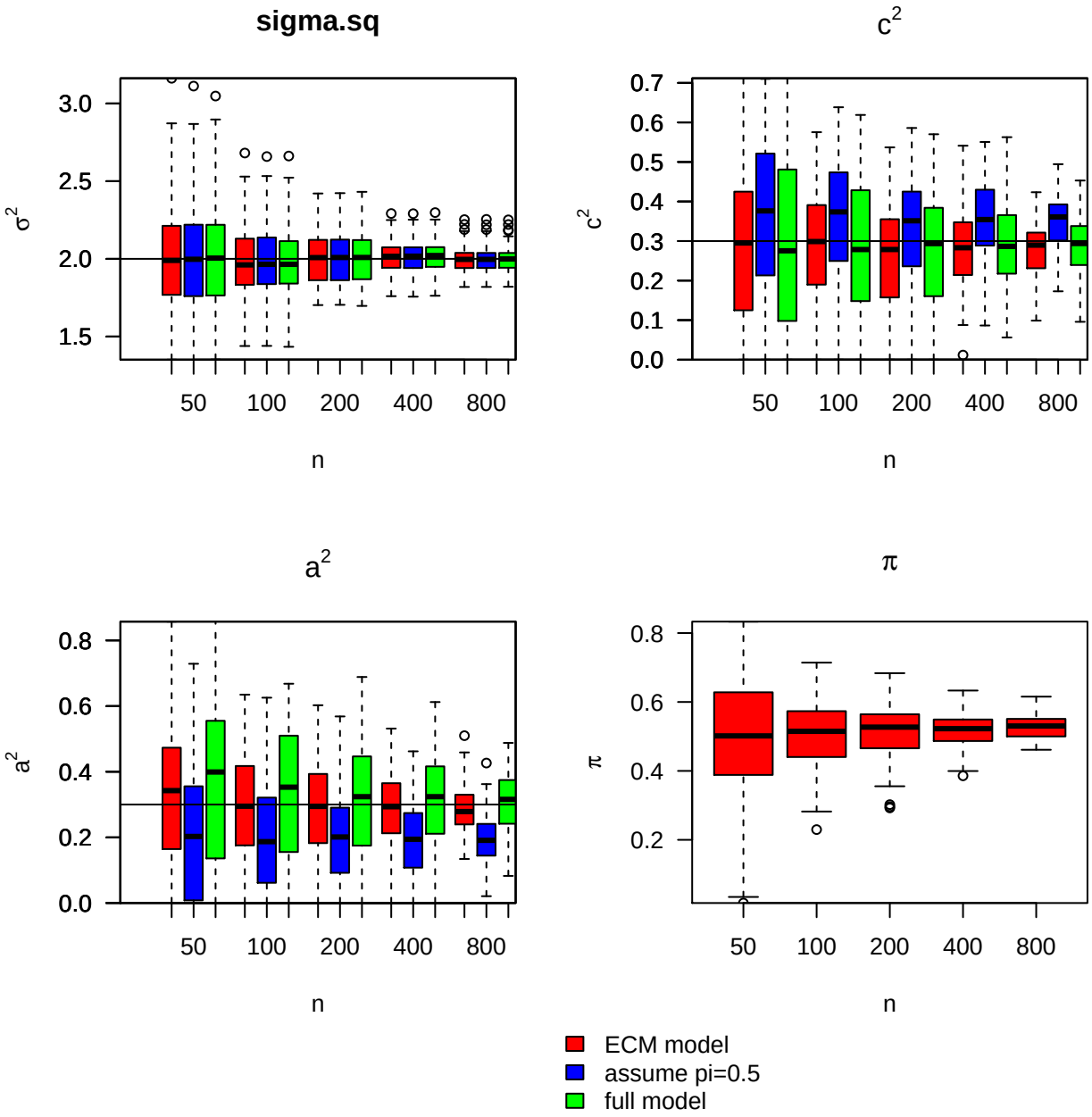
באיור 6 מוצגים תרשימי ה־BOXPLOTS של תוצאות הנתונים שהוגרלו באופן הבא: 40% תאומים זהים ו־ 30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כעת כי 15% הם זהים (MZ) ו־ 85% הם לא זהים (DZ). עבור הפרמטרים: $a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 1.33, \mu{=}5$.

איור 4:



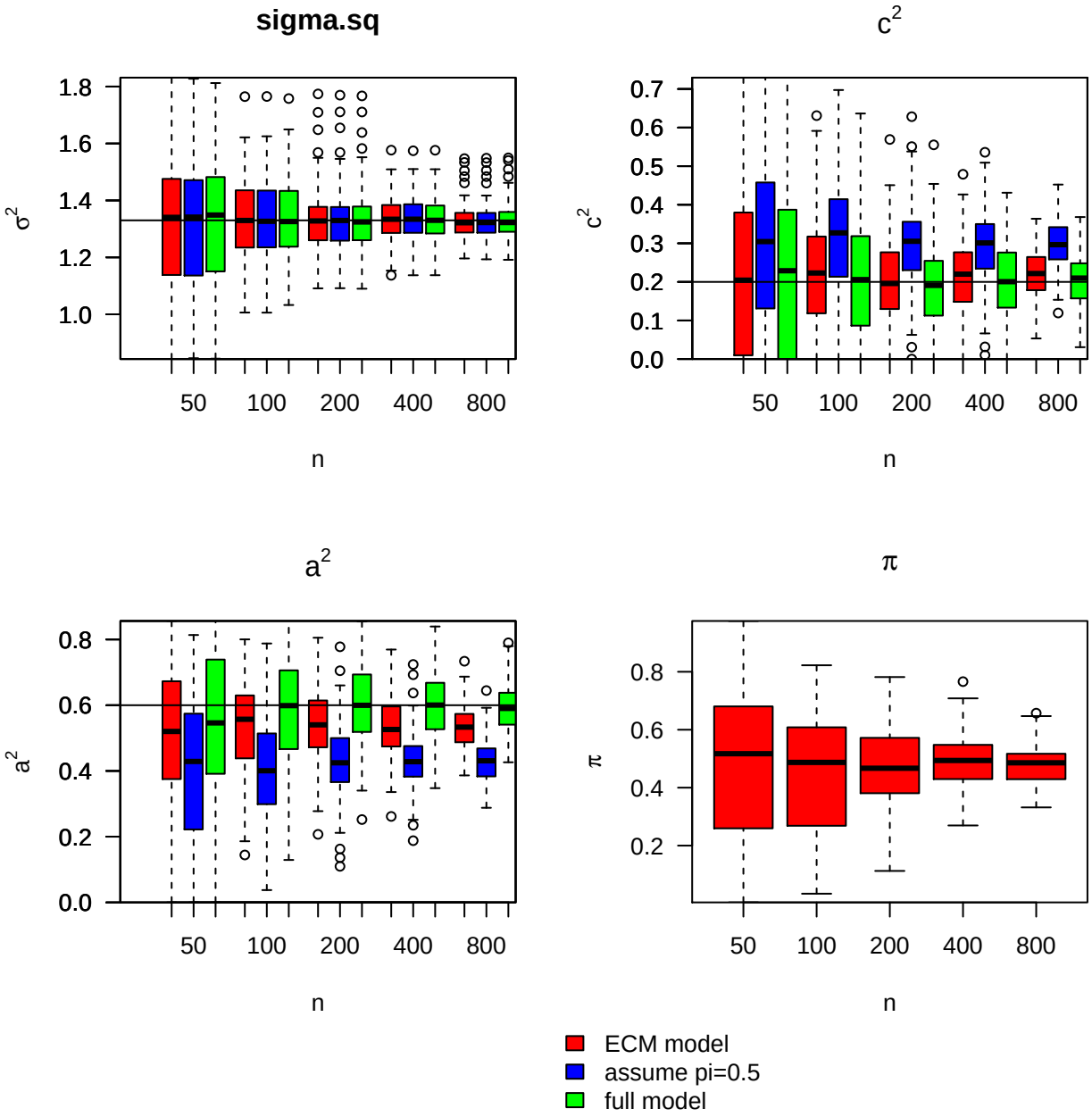
הפרמטרים שנבחרו הם $a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.
הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

איור 5:



הפרמטרים שנבחרו הם $a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2,\ \mu = 5$.הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.
הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

איור 6:



הפרמטרים שנבחרו הם $a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 1.33,\ \mu = 5$. הקופסא האדומה השמאלית מתארת את השיטה הראשונה שבה הפעלנו את המודל החסר ואמדנו את π.

הקופסא הכחולה האמצעית היא השיטה השנייה, והקופסא הירוקה הימנית היא השיטה השלישית שבה לקחנו את המודל המלא להשוואה.

5.2 תוצאות הסימולציות

על פי תרשימי ה־BOXPLOTS, התוצאות שהתקבלו מהרצת הסימולציות מעידות שהשיטה, שהוצעה בתת־פרק 4.2, למציאת אומדי הנראות המירבית עבור מודל עם מידע חסר תוך השימוש במודל של תערובת התפלגויות עבור נתוני תאומים עם מידע חסר, סיפקה תוצאות יותר טובות מהמצב שבו מניחים משקולות ידועות מראש, למשל בחירת $\pi_1 = 0.5$ לכל התצפיות.
בתרחיש הראשון, בכל המצבים שהרצנו, קיבלנו שהאומדים של μ ו־ σ^2 הם די קרובים ככל שמגדילים את גודל המדגם. לעומת זאת האומדים לפרמטרים a^2 ו־ c^2 פחות קרובים אבל עדיין השיטה הראשונה טובה יותר מאשר השיטה השנייה שבה לקחנו משקולות נתונות. אפשר גם לראות שכאשר בחרנו ערכים שונים עבור הפרמטרים a^2 ו־ c^2. ככל שההבדל בין a^2 ל־ c^2 היה קטן יותר, התוצאות היו יותר קרובות בשיטה שהצענו, לערכים האמיתיים מהתוצאות של המודל "הקלאסי", שהוא המודל שבו מניחים את המשקולות כידועות מראש.
בתרחיש השני שבו שינינו את המשקולות ההתחלתיות, אנו רואים שוב כי האומדים של התוחלת μ ו־σ^2 מתקרבים ומתכנסים לערכים האמיתיים ככל שמגדילים את n. בשלושת השיטות. האומדים של הפרמטרים a^2 ו־ c^2 הם טובים יותר מאשר עבור השיטה השנייה של משקולות קבועות מראש. אנו יכולים לזהות כי האומדים של π_1 הם פחות טובים: בתרחיש הראשון הערך האמיתי של π_1 הוא חצי, אך על פי התוצאות שקיבלנו האומדים לא כל כך תואמים לערך האמיתי. גם בתרחיש השני, באיור 6, אפשר לראות כי האומדים של π_1 אינם במגמת התכנסות. אפשרות לבדיקת נקודה זו יכולה להיות בהגדלת מספר התצפיות, ובדיקת הסימולציה מחדש. למרות שבחלק מההרצות יצא כי האומדים של המודל המוצע בעבודה זו לא ממש תואמים לערכים האמיתיים, עדיין המודל הזה מספק תוצאות קרובות יותר לערכים האמיתיים משיטה המקובלת של קביעת משקולות מראש.

6 סיכום

בעבודה זו הצגנו מודל של תערובת התפלגויות, שמתאים לנתוני תאומים עם מידע סיווג חסר, ומתחשב באילוצים בין הפרמטרים. כמו כן הצענו דרך לאמידת הפרמטרים של המודל בשיטת הנראות המירבית.

בפרק 2 הוצג מודל של ACE, מודל שהוא שימושי לחקירת שונות בין בני אדם באופן כללי. על פי מודל ACE לנתוני תאומים, ניתן היה לראות כי ישנו קשר בין השונויות של תאומים זהים ולא זהים, וכמו כן יש קשר בין התוחלות. הקשרים האלה נחשבים לאילוצים שיש לקחת אותם בחשבון, בעת חישוב ומציאת האומדים. המודל שהתאמנו לנתוני התאומים הוא מודל תערובת GMM. לאחר התאמת המודל, בפרק 3, הוצגה שיטת הנראות המירבית לאמידת הפרמטרים, עם המידע המלא והמידע החסר. ניתן היה לראות כי פונקציית הנראות הינה פונקציה לא לינארית בפרמטרים, שמצריכה שיטות פתרון לא טריוויאליות למציאת האומדים. בפרק 4 הוצגו אלגוריתם EM וההרחבה שלו, אלגוריתם ECM, לאמידת הפרמטרים, תוך התחשבות בשני דברים: האילוצים, והמידע החסר. לאחר הצגת האלגוריתמים אמדנו את הפרמטרים. לתוחלת ולמשקולות התקבלה נוסחה סגורה, אך למרכיבי השונות לא התאפשר לקבל נוסחה סגורה, לכן היה צורך לשימוש בשיטות נומריות על מנת לאמוד את מרכיבי השונות. בפרק 5 בדקנו את המודל שהצגנו על ידי סימולציות. ולאחר מכן דנו בתוצאות שהתקבלו. הצלחנו להראות, בעזרת הסימולציות שבוצעו בתוכנת R, כי המודל שהצגנו הוא מודל טוב יותר מאשר המודל הקלאסי, המודל שבו מניחים את המשקולות כידועות מראש.

חשוב לציין כי בעבודה הזו, הנתונים הסתמכו רק על נתוני תאומים והתאמת מודל תערובת שמתייחס לנתונים אלה ספציפית, תוך טיפול בשתי הבעיות: האילוצים בין הפרמטרים, והמידע החסר עבור סוג של כל זוג תאומים. כעבודה עתידית אפשר לקחת את המודל שהותאם כאן בכיוון נוסף, למשל להתאים מודל לינארי מוכלל (GLM), לאמידת כל מיני הסתברויות וסיכויים. למשל, אמידת סיכויים להתפתחות מחלה מסוימת על סמך ידע נתון על אחד מהתאומים. בנוסף אפשר להרחיב את מה שנעשה בעבודה על ידי בחירת נתונים עם קשרים משפחתיים כלליים יותר, כלומר קשרים משפחתיים כדוגמת: אב ובנו, אחים ואחיות, בני דודים וכו'.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[2] Mark M Fredrickson. Ace in the hole: Correlation and classical twin studies. Unpublished manuscript, 2009.

[3] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. mvtnorm: Multivariate normal and t distributions. 2012. *URL http://CRAN. R-project. org/package= mvtnorm. R package version 0.9-9992.*

[4] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[5] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data (Wiley Series in Probability and Statistics)*. Wiley, second edition, 2002.

[6] G. J. MacLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[7] J. J. McArdle and C. A. Prescott. Mixed-effects variance components models for biometric family analyses. *Behavior genetics*, 35(5):631–652, 2005.

[8] X. L. Meng. On the rate of convergence of the ECM algorithm. *The Annals of Statistics*, 22(1):326–339, 1994.

[9] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[10] M. C. Neale. A finite mixture distribution model for data collected from twins. *Twin Research and Human Genetics*, 6(03):235–239, 2003.

[11] M. C. Neale and L. Cardon. *Methodology for Genetic Studies of Twins and Families*. Springer, 1992.

[12] F. Vaida. Parameter convergence for EM and MM algorithms. *Statistica Sinica*, 15(3):831, 2005.

[13] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1):95–103, 1983.

8 נספחים

8.1 אומדי נראות מירבית עבור המודל המלא

נחשב את אומדי הנראות המירבית עבור המודל המלא. כפי שראינו בתת פרק 3.2 פונקציית לוג הנראות עבור המודל המלא היא:

$$l(\vec{x};\{\mu,a^2,c^2,\sigma^2\}) = \sum_{i=1}^{k}\log\left(f_1(x_i;\Sigma_{MZ},\mu)\right) + \sum_{i=k+1}^{n}\log\left(f_2(x_i;\Sigma_{DZ},\mu)\right)$$

כלומר:

$$\begin{aligned} l(\vec{x};\{\mu,a^2,c^2,\sigma^2\}) &= \sum_{i=1}^{k}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{MZ}| - \frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_{MZ}^{-1}\cdot\left(x_i-\mu\right)\right) \\ &+ \sum_{i=k+1}^{n}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{DZ}| - \frac{1}{2}\left(x_i-\mu\right)^t\cdot\Sigma_{DZ}^{-1}\cdot\left(x_i-\mu\right)\right) \end{aligned} \quad (5)$$

נסמן: $\Sigma_1 = \Sigma_{MZ}$ ו- $\Sigma_2 = \Sigma_{DZ}$
למציאת אומד נראות מירבית עבור μ, נגזור את משוואה (5) לפי μ.
ונקבל:

$$\begin{aligned} \frac{\partial l(\vec{x};\{\mu,a^2,c^2,\sigma^2\})}{\partial\mu} &= \sum_{i=1}^{k}\left(\Sigma_1^{-1}\left(x_i-\mu\right)\right) + \sum_{i=k+1}^{n}\left(\Sigma_2^{-1}\cdot\left(x_i-\mu\right)\right) \\ &= \sum_{i=1}^{k}\left(\Sigma_1^{-1}\begin{pmatrix} x_{i1}-\mu \\ x_{i2}-\mu\end{pmatrix}\right) + \sum_{i=k+1}^{n}\left(\Sigma_2^{-1}\cdot\begin{pmatrix} x_{i1}-\mu \\ x_{i2}-\mu\end{pmatrix}\right) \\ &= \sum_{i=1}^{k}\left(\Sigma_1^{-1}\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right) - k\Sigma_1^{-1}\begin{pmatrix}1\\1\end{pmatrix}\mu + \sum_{i=k+1}^{n}\left(\Sigma_2^{-1}\left[\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right]\right) \\ &\quad -(n-k)\Sigma_2^{-1}\begin{pmatrix}1\\1\end{pmatrix}\mu \end{aligned}$$

נשווה לאפס למציאת μ:

$$\frac{\partial l(\vec{x};\{\mu,a^2,c^2,\sigma^2\})}{\partial\mu} = 0$$

$$k\Sigma_1^{-1}\begin{pmatrix}1\\1\end{pmatrix}\mu + (n-k)\Sigma_2^{-1}\begin{pmatrix}1\\1\end{pmatrix}\mu = \sum_{i=1}^{k}\left(\Sigma_1^{-1}\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right) + \sum_{i=k+1}^{n}\left(\Sigma_2^{-1}\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right)$$

נחלץ את μ:

$$\hat{\mu} = \left(\left[(n-k)\Sigma_2^{-1}\begin{pmatrix}1\\1\end{pmatrix} + k\Sigma_1^{-1}\begin{pmatrix}1\\1\end{pmatrix}\right]^{-1}\right)^t\left(\sum_{i=1}^{k}\left(\Sigma_1^{-1}\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right) + \sum_{i=k+1}^{n}\left(\Sigma_2^{-1}\begin{pmatrix} x_{i1} \\ x_{i2}\end{pmatrix}\right)\right)$$

עכשיו נסתכל בשאר הפרמטרים. קודם נסדר את הפונקציה (5).
על פי מודל ACE:

$$\Sigma_2 = \Sigma_{DZ} = \sigma^2 \begin{pmatrix} 1 & 0.5a^2 + c^2 \\ 0.5a^2 + c^2 & 1 \end{pmatrix} \quad \Sigma_1 = \Sigma_{MZ} = \sigma^2 \begin{pmatrix} 1 & a^2 + c^2 \\ a^2 + c^2 & 1 \end{pmatrix}$$

כל תצפית: $x_i = (x_{i1}, x_{i2})$
נסמן

$$\begin{aligned} \rho_1 &= a^2 + c^2 \\ \rho_2 &= 0.5a^2 + c^2 \end{aligned}$$

נציב ב־ (5) ונרשום את החלק התלוי בפרמטרים $\tau = \{\rho_1, \rho_2, \sigma^2\}$:

$$\begin{aligned} \zeta(\tau) = &\sum_{i=1}^{k} \left(-\frac{1}{2}\log\left(\sigma^4(1-\rho_1^2)\right) - \frac{1}{2} \cdot \frac{1}{\sigma^4(1-\rho_1^2)} \left[(x_{i1}-\mu)^2 - 2\rho_1 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i2}-\mu)^2 \right] \right) \\ &+ \sum_{i=k+1}^{n} \left(-\frac{1}{2}\log\left(\sigma^4(1-\rho_2^2)\right) - \frac{1}{2} \cdot \frac{1}{\sigma^4(1-\rho_2^2)} \left[(x_{i1}-\mu)^2 - 2\rho_2 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i2}-\mu)^2 \right] \right) \end{aligned} \tag{6}$$

נסמן: $v = \frac{1}{\sigma^4}$
נגזור את (6) לפי v:

$$\begin{aligned} \frac{\partial \zeta(\{\mu, \rho_1, \rho_2, \sigma^2\})}{\partial v} &= \sum_{i=1}^{k} \left(\frac{1}{2}\cdot\frac{1}{v} - \frac{1}{2}\cdot\frac{1}{(1-\rho_1^2)} \left[(x_{i1}-\mu)^2 - 2\rho_1 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i1}-\mu)^2 \right] \right) \\ &+ \sum_{i=k+1}^{n} \left(\frac{1}{2}\cdot\frac{1}{v} - \frac{1}{2}\cdot\frac{1}{(1-\rho_2^2)} \left[(x_{i1}-\mu)^2 - 2\rho_2 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i2}-\mu)^2 \right] \right) \end{aligned}$$

נשווה לאפס ונקבל:

$$\begin{aligned} 0 &= \frac{k}{v} - \sum_{i=1}^{k} \frac{1}{2(1-\rho_1^2)} \left[(x_{i1}-\mu)^2 - 2\rho_1 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i1}-\mu)^2 \right] \\ &+ \frac{n-k}{v} - \sum_{i=k+1}^{n} \frac{1}{2(1-\rho_2^2)} \left[(x_{i1}-\mu)^2 - 2\rho_2 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i2}-\mu)^2 \right] \end{aligned}$$

נסדר:

$$\begin{aligned} \frac{n}{v} &= \sum_{i=1}^{k} \frac{1}{2(1-\rho_1^2)} \left[(x_{i1}-\mu)^2 - 2\rho_1 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i1}-\mu)^2 \right] \\ &+ \sum_{i=k+1}^{n} \frac{1}{2(1-\rho_2^2)} \left[(x_{i1}-\mu)^2 - 2\rho_2 \cdot (x_{i1}-\mu)\cdot(x_{i2}-\mu) + (x_{i2}-\mu)^2 \right] \end{aligned}$$

לכן:

$$\begin{aligned}\frac{1}{v} &= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{2(1-\rho_1^2)}\left[(x_{i1}-\mu)^2-2\rho_1\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i1}-\mu)^2\right]\right.\\ &+ \left.\sum_{i=k+1}^{n}\frac{1}{2(1-\rho_2^2)}\left[(x_{i1}-\mu)^2-2\rho_2\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i2}-\mu)^2\right]\right)\end{aligned}$$

קיבלנו:

$$\begin{aligned}\hat{\sigma}^4 &= \frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{2(1-\rho_1^2)}\left[(x_{i1}-\mu)^2-2\rho_1\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i1}-\mu)^2\right]\right.\\ &+ \left.\sum_{i=k+1}^{n}\frac{1}{2(1-\rho_2^2)}\left[(x_{i1}-\mu)^2-2\rho_2\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i2}-\mu)^2\right]\right)\end{aligned}$$

נסכם:

$$\begin{aligned}\hat{\sigma}^2 &= \left[\frac{1}{n}\left(\sum_{i=1}^{k}\frac{1}{2(1-\rho_1^2)}\left[(x_{i1}-\mu)^2-2\rho_1\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i1}-\mu)^2\right]\right.\right.\\ &+ \left.\left.\sum_{i=k+1}^{n}\frac{1}{2(1-\rho_2^2)}\left[(x_{i1}-\mu)^2-2\rho_2\cdot(x_{i1}-\mu)\cdot(x_{i2}-\mu)+(x_{i2}-\mu)^2\right]\right)\right]^{\frac{1}{2}}\end{aligned}$$

נשאר לנו לאמוד את הפרמטרים ρ_1, ρ_2. לפרמטרים האלה לא נוכל לחשב את האומדים בצורה אנליטית כי הפונקציה (6) מסובכת בפרמטרים אלה כך שלא נוכל לקבל נוסחה סגורה. השיטה החילופית היא השיטה הנומרית. נוכל לאמוד את הפרמטירם בעזרת שיטות נומריות כגון ניוטון רפסון או אחרות.

8.2 הרחבה - אלגוריתם EM

אלגוריתם EM מסתמך על הקשר (1).
בהינתן קבוצת פרמטרים $\theta^{(m)}$, ועל ידי הכפלת המשוואה (1) ב- $P(Y|\theta^{(m)})$ נקבל:

$$\log\left[P\left(X|\theta\right)\right]\cdot P(Y|\theta^{(m)})=\log\left[P\left((X,Y)|\theta\right)\right]\cdot P(Y|\theta^{(m)})-\log\left[P\left(Y|X,\theta\right)\right]\cdot P(Y|\theta^{(m)})$$

נניח Y בדיד נסכום על כל האפשריות לערכי Y ונקבל:

$$\log\left[P\left(X|\theta\right)\right]=\sum_{y}\log\left[P\left((X,Y)|\theta\right)\right]\cdot P(Y|\theta^{(m)})-\sum_{y}\log\left[P\left(Y|X,\theta\right)\right]\cdot P(Y|\theta^{(m)})$$

נסמן:

$$Q\left(\theta|\theta^{(m)}\right)=\sum_{y}\log\left[P\left((X,Y)|\theta\right)\right]\cdot P(Y|\theta^{(m)})$$

$$H\left(\theta|\theta^{(m)}\right)=-\sum_{y}\log\left[P\left(Y|X,\theta\right)\right]\cdot P(Y|\theta^{(m)})$$

קיבלנו:

$$\log\left[P\left(x|\theta\right)\right]=Q\left(\theta|\theta^{(m)}\right)+H\left(\theta|\theta^{(m)}\right)$$

במקום למקסם את פונקציית המטרה $\log\left[P\left(X|\theta\right)\right]$ אנו ממקסימים את הפונקציה $Q\left(\theta|\theta^{(m)}\right)$.
על פי אי שוויון ינסן אפשר לקבל כי:

$$H(\theta^{(m)}|\theta^{(m)})\leq H(\theta|\theta^{(m)})$$

כלומר:

$$Q\left(\theta|\theta^{(m)}\right)+H\left(\theta^{(m)}|\theta^{(m)}\right)\leq Q\left(\theta|\theta^{(m)}\right)+H\left(\theta|\theta^{(m)}\right)$$

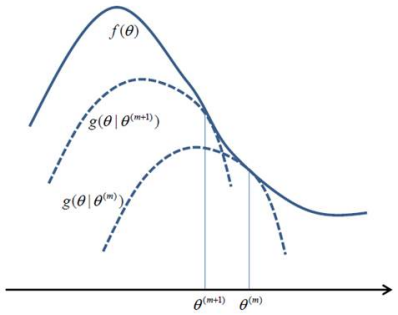
ולכן הפונקציה הנוצרת בצעד ה- E, $Q\left(\theta|\theta^{(m)}\right)$, עד הקבוע: $c=H(\theta^{(m)}|\theta^{(m)})$ נחשבת לפונקציה חילופית שמחפשים את המקסימום של פונקציה זו במקום הפונקציה המקורית.
כלומר אלגוריתם EM משתמש ברעיון של שיטת כללית יותר שנקראת שיטת MM (MM= Minoriza-tion – Maximization), נגדיר את השיטה הזו:
MM הינה שיטה שבעזרתה בונים אלגוריתם לפתרון בעיות אופטימיזציה באופן נומרי (היא נקראת גם "אלגוריתם MM" [4]). שיטת MM מציעה פונקציה "תחליפית" לפונקציה "מסובכת" שנרצה לעשות לה אופטימיזציה. כלומר היא הופכת את הבעיה לצורה "פשוטה" יותר. במקרה שלנו אנו מתעניינים בחיפוש אומד נראות מירבית כלומר הבעיה היא בעיית מקסמיזציה, נגדיר את MM בבעיית מקסמיזציה:
בבעיות מקסמיזציה: MM= Minorization – Maximization
ההגדרה של פעולת Minorization

הגדרה 8.1 A function $g\left(\theta|\theta^{(m)}\right)$ is said to minorize a real-valued function $f\left(\theta\right)$ at point $\theta^{(m)}$ provided:
$\forall\theta \quad g\left(\theta|\theta^{(m)}\right) \leq f(\theta)\,,\, g(\theta^{(m)}|\theta^{(m)}) = f(\theta^{(m)})$

אלוגריתם MM מתבצע בשני צעדים:

1. מציאת את הפונקציה $g(\theta|\theta^{(m)})$ "המנורייזר"
2. $\theta^{(m+1)}=\underset{\theta}{argmax}\, g(\theta|\theta^{(m)})$

ממשיכים כך עד להתכנסות להרחבה על התכנסות של אלגוריתם MM ראה [12].
בשרטוט



אנו רואים ששני הצעדים של אלגוריתם MM מקבילים לשני הצעדים של אלגוריתם EM. במקרה של EM הפוקנציה $Q\left(\theta|\theta^{(m)}\right)$ היא המינורייזר של הפונקציה $\log\left[P\left(x|\theta\right)\right] - H\left(\theta|\theta^{(m)}\right)$. ולכן ניתן להסתכל על כל אלגוריתם EM כמקרה פרטי של אלגוריתם כללי יותר MM.

8.3 אמידה כללית לווקטור המשקולות π

החישוב שהוצג בפרק 4.2 לאמידת המשקולות π, עובד רק במצב שיש שתי אפשריות לתערובות כלומר רק למצב הספיציפי שלנו. אפשר להסתכל על החישוב באופן אחר, שיתאים למקרה הכללי. נציג כאן את החישוב הכללי הזה.

נגזור (4) לפי ווקטור המשקולות $\overrightarrow{\pi}$ תוך התחשבות בתנאי $\sum_j \pi_j = 1$.

החישוב הכללי:

על מינת לטפל באילוץ $\sum_j \pi_j = 1$ נוסיף כופל לגרנג λ.

כלומר נוסיף: $\lambda\left(\sum_j \pi_j - 1\right) = \lambda\left(\pi_1 + \pi_2 - 1\right)$ עכשיו נצטרך לגזור את:

$$\xi\left(\theta, \theta^{(m)}\right) = Q\left(\theta, \theta^{(m)}\right) + \lambda\left(\sum_j \pi_j - 1\right)$$

$$d\frac{\xi\left(\theta,\theta^{(m)}\right)}{d\pi} = \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|y_i,\theta^{(m)}\right)\frac{1}{\pi_j} + \lambda$$

נשווה לאפס:

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{2} P\left(z_{ij}|y_i,\theta^{(m)}\right)\frac{1}{\pi_j} + \lambda$$

עבור j בודד:

$$0 = \sum_{i=1}^{n} P\left(z_{ij}|y_i,\theta^{(m)}\right)\frac{1}{\pi_j} + \lambda \tag{7}$$

נסכום את המשוואות לכל j , במצב הזה נסכום את שתי המשוואות, על מנת לקבל אומד ל- λ:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} P\left(z_{i1}|y_i,\theta^{(m)}\right)\frac{1}{\pi_1} + \lambda + \sum_{i=1}^{n} P\left(z_{i2}|y_i,\theta^{(m)}\right)\frac{1}{\pi_2} + \lambda \\
0 &= \frac{1}{\pi_1}\cdot\pi_1 n + \lambda + \frac{1}{\pi_2}\cdot\pi_2 n + \lambda \\
0 &= 2n + 2\lambda \\
-n &= \lambda
\end{aligned}$$

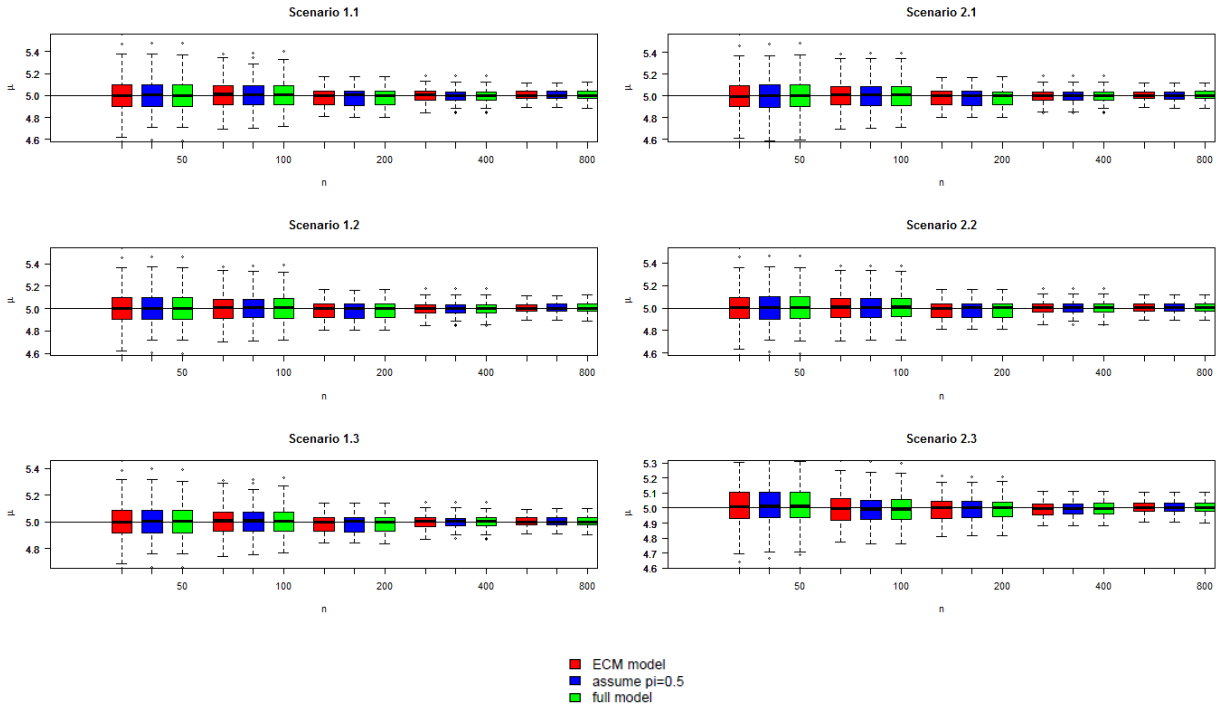
לכן אם נחזור ל-(7) ונציב את מה שקיבלנו במשוואה האחרונה נקבל:

$$\pi_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(m)}$$

ראינו כי עבור המקרה הספיציפי שלנו קיבלנו את אותה התוצאה.

8.4 תרשימי Boxplot עבור הפרמטר μ

איור 7:



ב־Scenario 1.1 אפשר לראות את תרשימי ה־BOXPLOTS של תוצאות הנתונים שהיו בתרחיש 1 בסימולציה הראשונה שבה הפרמטרים הוגרלו באופן הבא: $a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. ב־Scenario 1.2 אפשר לראות את תרשימי ה־BOXPLOTS של תוצאות הנתונים שהוגרלו בתרחיש 1 בסימולציה השנייה שהוצגה, עבור הפרמטרים: $a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. ב־Scenario 1.3 מוצגים תרשימי ה־BOXPLOTS של תוצאות הנתונים שהוגרלו בתרחיש 1 בסימולציה השלישית שהוצגה. עבור הפרמטרים: $a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 1.33, \mu{=}5$. ב־Scenario 2.1 אפשר לראות את תרשימי ה־BOXPLOTS של תרחיש 2 עבור תוצאות הנתונים שהוגרלו באופן הבא: 30% תאומים זהים ו־30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כי 90% הם זהים (MZ) ו־10% הם לא זהים (DZ). עבור הפרמטרים: $a^2 = 0.4,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. ב־Scenario 2.2 אפשר לראות את תרשימי ה־BOXPLOTS של תרחיש 2 עבור תוצאות הנתונים שהוגרלו באופן הבא: 30% תאומים זהים ו־30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כעת כי 10% הם זהים (MZ) ו־90% הם לא זהים (DZ). עבור הפרמטרים: $a^2 = 0.3,\ c^2 = 0.3,\ \sigma^2 = 2, \mu{=}5$. ב־Scenario 2.3 מוצגים תרשימי ה־BOXPLOTS של תרחיש 2 עבור תוצאות הנתונים שהוגרלו באופן הבא: 40% תאומים זהים ו־30% תאומים לא זהים והשאר לא ידוע (NA). עבור אלה הלא ידועים הנחנו כעת כי 15% הם זהים (MZ) ו־85% הם לא זהים (DZ). עבור הפרמטרים: $a^2 = 0.6,\ c^2 = 0.2,\ \sigma^2 = 1.33, \mu{=}5$.

Fitting Models for Twins Data with Missing Data

By: Afraa Araidy

Supervised by: Dr Yair Goldberg

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa

Faculty of Social Science

Department of Statistics

November 2014

Fitting Models for Twins Data with Missing Data

Afraa Araidy

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa

Faculty of Social Science

Department of Statistics

November 2014

Fitting Models for Twins Data with Missing Data

Afraa Araidy

Abstract

Twin studies are an important tool for investigating the influence of heredity and environment on variation in human population. The majority of twin studies assume that the zygosity is given, i.e., the information for each pair of twins whether they are monozygotic (MZ) or dizygotic (DZ) is given. However, this can be missing. Matching models under the wrong assumption that there is no missing data, and estimation of the parameters of these models can lead to biased estimators. Analysis on biased estimators may lead to erroneous conclusions. For this reason we need to find methods to solve this problem.

Some researchers tried to solve the problem of missing data by providing predetermined weights for the probability of twins to be either MZ or DZ.

In this work we propose a model of mixture distributions for twin data analysis without knowing the zygosity for some of the twins. In other words, we consider a model that deals with missing data. In contrast to previous methods which complement the missing information with probability and predetermined weights, we offer a maximum likelihood estimator (MLE). Finding the MLE in this problem is challenging since this maximization include constraints which arise from the structure of the variance matrices. To find the maximum likelihood estimator we use ECM algorithm (Expectation-Conditional Maximization). For implementing of the algorithm we developed an R code using numerical methods. In this work, we ran simulations to examine the proposed method. The results showed that using the method presented here, i.e. the model of mixture distributions with ECM algorithm for finding the maximum likelihood estimator is a model that provides better results than the previous methods which complement the missing data in this context.