

פרק ג'

אמידת פרמטרים

3.1 פרמטרים

תארנו את בעיית ההסקה הסטטיסטית כמצב בו יש לנו נתונים שנוצרו מתוך התפלגות שעליה אנחנו רוצים להסיק מסקנות. בבעיות סטטיסטיות רבות ההתפלגות שיצרה את הנתונים ידועה לחלוטין למעט פרמטר בודד או יותר. מצב כזה קורה כאשר יש לנו הבנה של התהליכים היוצרים את התצפיות (למשל מודל פיזיקלי), אך לא ידועים לנו הפרמטרים של התהליך.

לדוגמא, במסגרת מחקר מבנה הסינפסה הוצע מודל פיזיקלי שלפיו וזיקולות של טרנסמיטור משתחררות באופן בלתי תלוי ממספר גדול של אתרי שחרור בעקבות הגעת פוטנציאל פעולה. המסקנה היא כי מספר הוזיקולות המשתחררות מתפלג בינומית, אך מספר אתרי השחרור n והסתברות השחרור p משתנים מסינפסה לסינפסה. כדי לקבוע את ערכי הפרמטרים האלו עבור סינפסה נתונה אנו עשויים לבצע ניסוי בו אנחנו חוזרים ומודדים את מספר הוזיקולות שמשתחררות בפועל, ולנסות להסיק מסקנות על ערכי שני הפרמטרים הנ"ל. למשל, אנחנו יכולים לנחש את הערך הכי סביר ל- p , או לקבוע טווח ערכים אפשריים.

נשים לב כי בהנתן ערכי הפרמטרים החסרים, ההתפלגויות שיוצרות את התצפיות ידועות לנו לחלוטין, אך עדיין מספר האפשרויות לערכי הפרמטרים (מספר "מצבי העולם"), יכול להיות אינסופי (ואף בעל עצמת רצף). במקרים מסובכים יותר לא ידועה לנו הצורה הפרמטרית של ההתפלגות, ואז נשתמש בשיטות א-פרמטריות שיתוארו בפרק הבא.

בפרק הנוכחי נטפל בהתפלגויות שבהן מספר הפרמטרים הוא קטן, ונסמן אותם ב- $\theta = (\theta_1, \theta_2, \dots, \theta_n)$. לדוגמא, ההתפלגות הבינומית עם הפרמטר הבודד

$$X \sim B(n, \theta) \quad P_n(m | \theta) = \binom{n}{m} \theta^m (1 - \theta)^{n-m}$$

או ההתפלגות הנורמלית, שלה שני פרמטרים

$$X \sim N(\theta_1, \theta_2^2) \quad P(x | \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi}\theta_2} \exp\left[-\frac{(x-\theta_1)^2}{2\theta_2^2}\right]$$

3.2 אמידה

נניח כי נתון לנו מדגם $X^{(n)} = \{x_1, \dots, x_n\}$ שנלקח מתוך התפלגות $f(x|\theta)$ שהיא ידועה למעט ערכו של פרמטר θ , ונניח כי ערכי הפרמטר θ (שיכול להיות רב מימדי) נלקחים מהקבוצה Ω . מטרת האמידה היא "לנחש" את הערך של הפרמטר באופן טוב ככל האפשר על סמך המדגם. **אומד**¹ (estimator) של הפרמטר θ המבוסס על מדגם (ופורמלית, מבוסס על המשתנים המקריים X_1, \dots, X_n) הוא פונקציה של המשתנים המקריים

$$(3.1) \quad \hat{\theta} = \varphi(X_1, \dots, X_n)$$

אשר מתאימה ערך משוער של θ לכל אוסף ערכי תצפיות $\{x_1, \dots, x_n\}$ שנפגוש. נשים לב כי היות והאומד הוא פונקציה של משתנים מקריים, הרי שגם הוא משתנה מקרי, ונוכל לחקור את ההתפלגות שלו ואת תכונותיו ההסתברותיות. בסעיפים הבאים נתאר שתי דרכים למציאת הפונקציות המתארות את האומד - אמידה בייסיאנית (סעיף 3.3), ואמידה בעזרת מקסימום נראות (סעיף 3.4), ובסעיף 3.5 נפנה לחקור את תכונות ההתפלגות של אומדים.

3.3 אמידה בייסיאנית

3.3.1 הגדרת האומד הבייסיאני

במסגרת הגישה הבייסיאנית, אמידת פרמטרים היא הרחבה טבעית לשיטות אותן תיארנו בפרק 2. בעוד שעד כה טיפלנו במקרה בו כל ההתפלגויות היו ידועות לנו בהינתן מצב העולם, הרי שכעת קיימת משפחה פרמטרית של התפלגויות אפשריות, וכל ערך של הפרמטר קובע מצב עולם. בהנתן מדגם $X^{(n)}$ עלינו להכריע באיזה מבין מצבי העולם השונים המוגדרים על ידי הערכים האפשריים השונים של הפרמטר θ אנחנו נמצאים.

כאשר הצורה הכללית של פונקציית ההתפלגות ידועה $P(x|\theta)$, אם נניח כי ידועות לנו גם ההסתברויות האפרוריות $P_0(\theta)$ יהיה בידינו מודל הסתברותי מלא של העולם, ונוכל לחשב את הסבירות למצב העולם המוגדר על ידי הפרמטר θ בהינתן המדגם $X^{(n)}$

¹ לעתים מתורגם המושג estimator כ- "משערך".

$$(3.2) \quad P(\theta | X^{(n)}) = P(X^{(n)} | \theta) \frac{P_0(\theta)}{P(X^{(n)})} = \frac{P_0(\theta) \cdot P(X^{(n)} | \theta)}{\int_{-\infty}^{\infty} P_0(\theta') P(X^{(n)} | \theta') d\theta'}$$

כאשר האינטגרל במכנה מבטא את העובדה שהסיכוי הכולל לקבל את סט התצפיות $X^{(n)}$ הוא סכום משוקלל של הסיכוי לקבל אותו עבור כל ערך אפשרי של θ (נוסחת ההסתברות השלמה).

כמו בפרק הקודם נגדיר גם כאן את פונקציית המחיר $\lambda(\hat{\theta}, \theta)$ הקובעת את המחיר שנשלם אם אמדנו את הפרמטר θ להיות $\hat{\theta}$ והמחיר הממוצע שנשלם על מדגם נתון יהיה

$$(3.3) \quad E[\lambda(\hat{\theta}, \theta) | X^{(n)}] = \int_{\Omega} \lambda(\hat{\theta}, \theta) P(\theta | X^{(n)}) d\theta$$

כעת נוכל להגדיר לבסוף את **האומד הבייסיאני** בתור הערך של הפרמטר θ המביא למינימום את תוחלת המחיר שנשלם

$$(3.4) \quad \theta^* = \underset{\hat{\theta}}{\operatorname{arg\,min}} E[\lambda(\hat{\theta}, \theta) | X^{(n)}]$$

האומד הבייסיאני תלוי אם כן גם בפונקציית המחיר שנבחר להשתמש בה וגם בהתפלגות האפריורית. נתאר כעת את השיטות החשובות ביותר לאמידת פרמטרים הנובעות מפונקציות מחיר שונות.

3.3.2 פונקציות מחיר

שגיאה ריבועית והתוחלת המותנית

פונקציית מחיר מקובלת ביותר ונוחה לטיפול אנליטית היא השגיאה הריבועית (נורמה L_2) המוגדרת על ידי

$$(3.5) \quad \lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

מהצבת פונקציית מחיר זו במשוואה (3.3) וגזירה ביחס ל- $\hat{\theta}$ ניתן להראות (תרגיל 2) כי בהינתן מדגם נתון $X^{(n)}$ האומד האופטימלי תחת פונקציית המחיר של השגיאה הריבועית יהיה

$$(3.6) \quad \theta = E(\theta | X^{(n)}) = \int_{-\infty}^{\infty} \theta P(\theta | X^{(n)}) d\theta$$

גודל זה ניקרא **התוחלת המותנית** (conditional mean) של θ בהנתן המדגם $X^{(n)}$.

דוגמא: אמידת הפרמטר של התפלגות אקספוננציאלית

יהי X משתנה מקרי מפולג אקספוננציאלית עם פרמטר $\theta > 0$
 $X \sim \exp(\theta): f(x) = \theta e^{-\theta x} \quad \forall x \geq 0$

ולפרמטר התפלגות אפריורית

$$f_0(\theta) = e^{-\theta} \quad \forall \theta \geq 0$$

נחשב את התוחלת המותנית

$$\begin{aligned} \hat{\theta} &= \int_0^{\infty} \theta \cdot P(\theta | X^{(n)}) d\theta \\ &= \int_0^{\infty} \theta \left(\frac{\theta^n e^{-\theta \sum X_i} e^{-\theta}}{P(X^{(n)})} \right) d\theta \\ &= \frac{1}{P(X^{(n)})} \int_0^{\infty} \theta^{(n+1)} e^{-\theta(\sum X_i + 1)} d\theta \end{aligned}$$

נבצע החלפת משתנים: נסמן $y = 1 + \sum x_i$ ו- $t = y\theta$ ונקבל תוך שימוש בידוע לנו על פונקציות גאמא (סעיף 1.5.5) כי

$$\begin{aligned} \hat{\theta} &= \frac{1}{P(X^{(n)})} \int_0^{\infty} \left(\frac{t}{y} \right)^{n+1} e^{-t} dt \cdot \frac{1}{y} = \\ &= \frac{1}{P(X^{(n)})} \cdot \frac{1}{y^{(n+2)}} \cdot \Gamma(n+2) \\ &= \frac{1}{P(X^{(n)})} \cdot \frac{\Gamma(n+2)}{[\sum X_i + 1]^{n+2}} \\ &= \frac{[\sum X_i + 1]^{n+1}}{\Gamma(n+1)} \cdot \frac{\Gamma(n+2)}{[\sum X_i + 1]^{n+2}} = \frac{n+1}{\sum X_i + 1} \end{aligned}$$

וקיבלנו

$$\frac{1}{\hat{\theta}} = \frac{\sum X_i + 1}{n+1}$$

ונשים לב כי היות והתוחלת של מ"מ אקספוננציאלי היא $\frac{1}{\theta}$, הרי שהביטוי הנ"ל הוא מהצורה של הממוצע האמפירי, שאליו הוספנו דגימה אחת בעלת ערך 1.

שגיאה בערך מוחלט והחציון

פונקצית מחיר מקובלת נוספת היא השגיאה בערך מוחלט (נורמה L_1)

$$(3.7) \quad \lambda(\hat{\theta}, \theta) = |\hat{\theta} - \theta|.$$

תחת פונקצית מחיר זאת, האומדן הבייסיאני האופטימלי בהנתן מדגם $X^{(n)}$ יהיה החציון של המדגם (ראה תרגיל 2), כלומר אותו ערך של θ שעבורו

$$(3.8) \quad \int_{-\infty}^{\hat{\theta}} P(\theta | x^{(n)}) d\theta = \int_{\hat{\theta}}^{\infty} P(\theta | x^{(n)}) d\theta$$

התפלגות אפוסטריורית מקסימלית Maximum A Posteriori

כאשר פונקצית המחיר היא פונקצית דלתא של דיראק

$$(3.9) \quad \lambda(\hat{\theta}, \theta) = \begin{cases} 0 & \text{if } \hat{\theta} \neq \theta \\ 1 & \text{otherwise} \end{cases}$$

אז כפי שכבר ראינו, הכרעה בייסיאנית אופטימלית תהיה לבחור את "מצב העולם" הסביר ביותר בהנתן התצפיות - כלומר את הערך הסביר ביותר של θ אחרי שראינו את סדרת התצפיות $X^{(n)} = \{x_1, \dots, x_n\}$.

שיטה זאת נקראת שיטת מקסימום אפוסטריורי (Maximum A Posteriori MAP). ועל פיה, יש לבחור באומדן המביא למקסימום את $P(\theta | X^{(n)})$. פורמלית,

$$\hat{\theta}_{MAP} = \arg \max_{\theta} [P(\theta | X^{(n)})]$$

מאחר ופונקצית הלוגריתם היא מונוטונית - המקסימום האפשרי של הפונקציה והמקסימום של הלוגריתם שלה מתקבלים באותה נקודה ולכן

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} [P(\theta | X^{(n)})] \\ &= \arg \max_{\theta} [\log P(\theta | X^{(n)})] \\ &= \arg \max_{\theta} [\log P(X^{(n)} | \theta) + \log P_0(\theta) - \log P(X^{(n)})] \\ &= \arg \max_{\theta} [\log P_0(\theta) + \log P(X^{(n)} | \theta)] \end{aligned}$$

ואם θ הוא וקטור של פרמטרים רציפים ניתן למצוא את המקסימום של

$$\log P_0(\theta) + \log P(X^{(n)} | \theta)$$

ע"י גזירה לפי כל אחד ממרכיבי הוקטור .

אם המדגם שלנו גדול מספיק, אז האבר הראשון (לוג ההסתברות האפריורית) זניח לעומת השני (לוג הנראות של המדגם) ואז ניתן לאמוד את θ רק על סמך הנראות (או לוג הנראות) של המדגם. גישה זאת נקראת נראות מירבית (maximum likelihood) ונתאר אותה בסעיף הבא.

דוגמא: אמידת הפרמטר של התפלגות אקספוננציאלית

נחזור על הדוגמא שהוצגה עבור פונקצית מחיר ריבועית

$$X \sim \exp(\theta): \quad f(x) = \theta e^{-\theta x} \quad \forall x \geq 0$$

$$f_0(\theta) = e^{-\theta} \quad \forall \theta \geq 0$$

נחפש מקסימום לפי θ

$$f(\theta | X^{(n)}) = e^{-\theta} \cdot \theta^n e^{-\theta \sum X_i}$$

$$\log(f(\theta | X^{(n)})) = -\theta + n \log(\theta) - \theta \sum X_i$$

$$= n \log(\theta) - \theta(\sum X_i + 1)$$

על ידי גזירה והשוואה לאפס נקבל

$$0 = \frac{\partial}{\partial \theta} \left(\log(\theta) - \frac{\theta}{n} (\sum X_i + 1) \right)$$

$$\frac{1}{\theta} = \frac{1}{n} (\sum X_i + 1)$$

ונשים לב שהאומד שקיבלנו שונה מזה שמתקבל בפונקצית מחיר ריבועית, שם קיבלנו

$$\frac{1}{\hat{\theta}} = \frac{1}{n+1} (\sum X_i + 1) .$$

3.3.3 גישות לבחירת התפלגות אפריורית

נתונות לנו דגימות מהתפלגות $f(X|\theta)$ ואנחנו רוצים לאמוד את θ בגישה הבייסיאנית. לצורך כך עלינו לבחור התפלגות אפריורית, ובסעיף זה נדון בגישות לבחירת התפלגות אפריורית כזאת.

התפלגויות צמודות: Conjugate prior

בתחילת פרק 2 ראינו כי בגישה הבייסיאנית אנחנו מקבלים תצפיות אשר מעדכנות את ההתפלגות האפריורית שלנו. נראה אם כן כי נוח יהיה אם ההתפלגות האפריורית שלנו באה ממשפחת התפלגויות כזאת, שלאחר עדכון ההתפלגות באמצעות התצפיות, ההתפלגות האפוסטריורית גם היא מאותה המשפחה, אבל

בפרמטרים אחרים. במלים אחרות, המבנה של הבעיה (הצורה הפונקציונלית של ההתפלגות) אינו תלוי בתצפיות המסוימות שראינו, ואלו קובעות רק את הפרמטרים של ההתפלגות. מסתבר שאכן קיימות התפלגויות של תצפיות שניתן למצוא להן התפלגות אפריורית שתקיים דרישה זאת ואז ההתפלגויות האפריוריות נקראות **התפלגויות צמודות** (conjugate priors). נתאר כעת צמדים כאלו של התפלגות תצפיות והתפלגות צמודה המתאימה לה.

דוגמא

נתונות דגימות מהתפלגות פואסונית

$$X \sim \text{Poisson}(\theta)$$

$$f(X^{(n)} | \theta) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \theta^{\sum x_i} e^{-n\theta} / \prod_{i=1}^n x_i!$$

נבחר פריור ל- θ , מהתפלגות גאמא

$$\theta \sim \text{Gamma}(\alpha, \beta)$$

$$P_0(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} e^{-\beta\theta} & \forall \theta \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

ונחשב את ההתפלגות האפוסטרירורית

$$P(\theta | X^n) = \frac{f(X^n | \theta) P_0(\theta)}{f(X^n)} \propto \theta^{\sum x_i} e^{-n\theta} \cdot \theta^{\alpha-1} e^{-\beta\theta} = \theta^{(\sum x_i + \alpha - 1)} e^{-(\beta+n)\theta}$$

וקיבלנו כי היא מצורת התפלגות גאמא

$$P(\theta | X^n) \sim \text{Gamma}(\alpha', \beta') \quad \text{with} \quad \begin{cases} \alpha' = (\alpha + \sum x_i) \\ \beta' = \beta + n \end{cases}$$

נשים לב כי ניתן לחשב את ההתפלגות השולית f(x) באופן הבא

$$f(x) = \int_{\theta} f(x, \theta) d\theta = \int_0^\infty \frac{\theta^{(\sum x_i + \alpha - 1)} e^{-(\beta+n)\theta}}{\Gamma(\alpha) \cdot \beta^{-\alpha}} \cdot \frac{1}{\prod (x_i!)} d\theta$$

$$= \frac{\Gamma(\sum x_i + \alpha) (\beta + n)^{-\sum x_i + \alpha}}{\Gamma(\alpha) \beta^{-\alpha} \prod (x_i!)}$$

התפלגות אפריורית בעלת מינימום הנחות: Non-Informative Prior

פעמים רבות אין לנו מידע מוקדם הקובע את ההתפלגות האפריורית, למעט אולי מספר תכונות (למשל שממוצע ההתפלגות צריך לקבל ערך מסוים). בפרק 7, נתאר עקרון המציע צורה פרמטרית להתפלגות האפריורית במקרה כזה. גישה אחרת היא לחפש תכונות אינווריאנטיות שההתפלגות אמורה לקיים (למשל אינווריאנטיות תחת כפל בקבוע שרירותי), ולגזור מהן אילוצים על צורת ההתפלגות. לא נפרט כאן על גישה זו, אך נושאים אלו נידונים למשל ב-Degroot.

3.4 אומד נראות מקסימלית

3.4.1 המגבלות של אומדים בייסיאניים

התיאוריה של אמידה בייסיאנית שהוצגה בפרקים הקודמים מהווה תיאוריה קוהרנטית לאמידת פרמטרים (וחסידי הפילוסופיה הבייסיאנית יטענו שזו התיאוריה הקוהרנטית היחידה). עם זאת, התיאוריה מצריכה להגדיר פונקצית מחיר והתפלגות אפריורית מפורשת, דבר שבבעיות מעשיות דורש לרוב משאבים גדולים ולפעמים אינו אפשרי. הבעיה מחריפה כאשר הפרמטר θ אותו יש לאמוד הוא וקטורי, כך שנדרשת התפלגות אפריורית משותפת של כל רכיביו. מן ההכרח לציין כי אין דרך טובה לפתור בעיות אלו: תיאוריות אחרות של אמידת פרמטרים סובלות מקשיים חמורים לא פחות. עם זאת, לעתים קרובות יהיה זה שימושי להשתמש בשיטה פשוטה יחסית כדי להעריך פרמטר, וזאת מבלי להידרש לאפיין התפלגות אפריורית. נתאר כעת לכן את שיטת אומד הנראות המקסימלית, שהוצגה על ידי פישר ב-1912.

3.4.2 הגדרת אומד נראות מקסימלית

נראות מקסימלית (Maximum Likelihood ML) היא השיטה המקובלת ביותר לאמידת פרמטרים של התפלגות "ממשפחה נתונה" מתוך מדגם נתון. על פי גישה זו אנו בוחרים בפשטות את הפרמטרים שיביאו למקסימום את הנראות של המדגם $X^{(n)}$ כפונקציה של סדרת פרמטרים $\theta = \{\theta_1, \theta_2, \dots, \theta_n\}$ (או, באופן שקול, יביאו למקסימום את הלוגריתם של הנראות). הנראות נתונה על ידי

$$(3.10) \quad L(X^{(n)} | \theta) = \prod_{i=1}^n P(x_i | \theta)$$

ואנו אומדים את הפרמטר על ידי

$$(3.11) \quad \hat{\theta} = \arg \max_{\theta} [L(X^{(n)} | \theta)]$$

דוגמא: התפלגות נורמלית עם סטיית תקן 1

$$L(X^{(n)} | \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i - \theta)^2 / 2} = \left(\frac{1}{\sqrt{2\pi}} \right)^n \exp \left[-\frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 \right]$$

$$\frac{\partial \log(L)}{\partial \theta} = \sum_{i=1}^n (x_i - \theta)$$

$$\frac{\partial \log(L)}{\partial \theta} = 0 \Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n (x_i)$$

$$\frac{\partial^2 \log(L)}{\partial \theta^2} = -n < 0$$

והאומד הוא הממוצע האמפירי

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

דוגמא: התפלגות בינומית

נגדיר

$$P(x | \theta) = \begin{cases} \theta^x (1-\theta)^{1-x} & \text{if } x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

(זאת אומרת הסיכוי של $x=1$ ("הצלחה") הוא θ , ושל $x=0$ ("כישלון") הוא $(1-\theta)$. הנראות של מדגם $X^{(n)} = \{x_1, x_2, \dots, x_n\}$ תהיה

$$L(X^{(n)} | \theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} \cdot (1-\theta)^{\sum_{i=1}^n (1-x_i)}$$

ולכן

$$\log(L(X^{(n)} | \theta)) = \sum_{i=1}^n x_i \log(\theta) + \sum_{i=1}^n (1-x_i) \log(1-\theta)$$

$$\begin{aligned} \frac{d \log(L)}{d\theta} = 0 &\Rightarrow \frac{\sum_{i=1}^n x_i}{\hat{\theta}} - \frac{\sum_{i=1}^n (1-x_i)}{1-\hat{\theta}} \\ &\Rightarrow \hat{\theta} \left(n - \sum_{i=1}^n x_i \right) - (1-\hat{\theta}) \sum_{i=1}^n x_i \\ &\Rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

ושב קבלנו את הממוצע האמפירי כאומד לסיכוי הצלחה.

3.4.3 מגבלות של אומד ML

אומד ML לא בהכרח קיים

נגזור אומד ML להתפלגות אחידה על הקטע הפתוח $(0, \theta)$

$$P(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

מאחר ובכל מדגם קיימות רק תצפיות שהסתברות לקיומן גדולה מאפס וההסתברות של כל תצפית היא $\frac{1}{\theta}$ אז הנראות של מדגם בן n תצפיות $X^{(n)}$ תהיה:

$$L(X^{(n)}|\theta) = \frac{1}{\theta} \cdot \dots \cdot \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n$$

על מנת למקסם את הנראות עלינו לקחת את ה- θ המינימלי האפשרי. מאחר ו- θ גדול מכל x_i אז אומד מקסימום נראות הוא המספר המינימלי שהנו גדול ממש מכל הדגימות x_i , אך בגלל רציפות הממשיים לא קיים מספר כזה, ולכל אומד נוכל למצוא אומד שהנראות שלו גבוהה יותר. קיבלנו כי במקרה כזה אומד ML אינו קיים.

אומד ML לא בהכרח יחיד

נביט בהתפלגות האחידה בקטע הסגור $[\theta, \theta+1]$. הפעם בדומה לדוגמה הקודמת עלינו לבחור θ שהוא קטן מכל ה- x_i אבל גדול מכל ה- $(x_i - 1)$, דהיינו מקיים

$$\max(x_i, \dots, x_n) - 1 \leq \theta \leq \min(x_i, \dots, x_n)$$

וכאשר מתקיים אי שוויון חזק $\max(x_1, \dots, x_n) - 1 < \min(x_1, \dots, x_n)$ אז כל θ באינטרוול יהיה אומד ML.

אומד ML עשוי להיות מוטה

בסעיף הבא (3.5.1) נתאר את המושג של אומדים חסרי הטיה, ונדגים שם כי אומד ML עשוי להיות מוטה.

3.5 תכונות של אומדים

3.5.1 הטיה וקונסיסטנטיות

הגדרה: אומד קונסיסטנטי

הגדרה: אומד קונסיסטנטי של הפרמטר הוא אומד המתכנס סטוכסטית לפרמטר, כלומר הוא משתווה לפרמטר בסיכויו "1" כאשר גודל המדגם שואף לאינסוף.

$$(3.12) \quad P(\hat{\theta}(X^{(n)}) = \theta) \xrightarrow{n \rightarrow \infty} 1$$

הגדרה: אומד מוטה ובלתי מוטה

אומד לא-מוטה (unbiased) של הפרמטר θ הוא אומד שהתוחלת שלו (הממוצע המשוקלל על פני כל המדגמים האפשריים) שווה לפרמטר θ

$$(3.13) \quad E_{X^{(n)}}[\hat{\theta}] = \int_{X^{(n)}} \hat{\theta}(X^{(n)}) P(X^{(n)}) dX^{(n)} = \theta$$

אם התוחלת של האומד אינה שווה לפרמטר אזי באופן טבעי נקרא האומד מוטה (biased).

דוגמא

בהתפלגות הבינומית והנורמלית קבלנו את הממוצע האמפירי כאומד לתוחלת. התוחלת של האומד היא

$$E[\hat{\mu}] = E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} \cdot n\mu = \mu$$

ואומד זה שווה לתוחלת ולכן איננו מוטה.

דוגמא

נגזור אומד ML להתפלגות אחידה על הקטע הסגור $[0, \theta]$

$$P(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

מאחר ובכל מדגם קיימות רק תצפיות שהסתברות לקיומן גדולה מאפס וההסתברות של כל תצפית היא $\frac{1}{\theta}$ אז הנראות של מדגם בן n תצפיות $X^{(n)}$ תהיה

$$L(X^{(n)}|\theta) = \frac{1}{\theta} \cdot \dots \cdot \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n$$

על מנת להביא למקסימום את הנראות עלינו לקחת את ה- θ המינימלי האפשרי. מאחר ו- θ גדול מכל x_i אז L יכול להיות לכל היותר $1/\max(x_i)^n$ ולכן $\hat{\theta} = \max(x_i)$ ישיג מקסימום לנראות. ידוע כי התוחלת של המקסימום עבור התפלגות אחידה היא

$$E[\hat{\theta}] = E\left[\max_i(x_i)\right] = \frac{n\theta}{n+1} < \theta$$

והאומד הוא לפיכך מוטה לכל n סופי. (לא נוכיח כאן את הביטוי לתוחלת המקסימום. הקורא המתעניין יוכל למצוא זאת כמעט בכל ספר בסיסי בהסתברות).

3.5.2 Minimum variance

בסעיפים הקודמים דנו באומדים אופטימליים עבור מדגם נתון. במונחים בייסיאנים, דיברנו על אומד שמביא למינימום את הסיכון המותנה שבבחירת האומד $\hat{\theta}$ כאשר ראינו את המדגם $X^{(n)}$ עבור פונקצית המחיר $\lambda(\hat{\theta}, \theta)$

$$(3.14) \quad R(\hat{\theta}|X^{(n)}) = \int \lambda(\hat{\theta}, \theta) P(\theta|X^{(n)}) d\theta$$

נניח כעת שאנו רוצים לקבוע שיטת שיערוך אופטימלית, כלומר פונקציה המגדירה את האומד עבור כל מדגם אפשרי: $\hat{\theta} = \hat{\theta}(X^{(n)})$ (במונחים בייסיאנים "קביעה של אסטרטגיה"). עלינו למצוא פונקציה $\hat{\theta}(X^{(n)})$ שתביא למינימום את הסיכון הכולל בממוצע על-פני כל המדגמים האפשריים

$$(3.15) \quad R[\hat{\theta}(X^{(n)})] = \int_{X^{(n)}} R(\hat{\theta}|X^{(n)}) P(X^{(n)}) dX^{(n)}$$

נניח כעת שהעולם נמצא במצב-עולם מוגדר θ שהוא מוגדר אך לא ידוע לנו, ונחשב את הסיכון הכולל ביחס למצב עולם זה (כלומר את הסיכון המשוקלל על-פני אוסף כל המדגמים האפשריים בגודל n , $\{X^{(n)}\}$ שאנו עשויים להתקל בהם במצב-עולם θ). עבור פונקציית המחיר הריבועית: $\lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ נקבל

$$\begin{aligned} R[\hat{\theta}(X^{(n)})] &= \int_{X^{(n)}} (\hat{\theta}(X^{(n)}) - \theta)^2 P(X^{(n)}) dX^{(n)} \\ &= E_{X^{(n)}} (\hat{\theta}(X^{(n)}) - \theta)^2 \\ &= E_{X^{(n)}} (\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= E(\hat{\theta} - E(\hat{\theta}))^2 + E(E(\hat{\theta}) - \theta)^2 \\ &= \text{Var}(\hat{\theta}) + \underbrace{(E(\hat{\theta}) - \theta)^2}_{\text{bias}} \end{aligned}$$

נגדיר את ההטיה (bias) $b(\theta) = (E(\hat{\theta}) - \theta)$ ונקבל

$$(3.16) \quad E_{(X^{(n)}, \theta)} (\hat{\theta}(X^{(n)}) - \theta)^2 = \text{Var}(\hat{\theta}) + b(\theta)^2$$

כלומר: על-מנת להביא למינימום את הסיכון הכולל בבחירת אומדן עבור פונקציית מחיר ריבועית עלינו להביא למינימום את הסכום של השונות של האומדן (מחושבת על-פני כל המדגמים האפשריים) וריבוע ההטיה (ה- bias) של האומדן. עבור אומדן לא-מוטה $E_{X^{(n)}}[\hat{\theta}(X^{(n)})] = \theta$ האומדן בעל הסיכון הכולל המינימאלי כאשר הפרמטר האמיתי הוא θ יהיה האומדן בעל השונות המינימאלית (Minimum Variance)

$$(3.17) \quad \hat{\theta} = \arg \min_{\hat{\theta}} E_{X^{(n)}} \left[(\hat{\theta}(X^{(n)}) - \theta)^2 \right]$$

הדרישה לאומדן לא-מוטה היא היסטורית במקורה ונובעת גם מ"דעות קדומות" - אומדים מוטים עשויים במקרים רבים להיות טובים יותר במובן הבייסיאני מאומדים לא מוטים.

3.5.3 אינפורמציות פישר ויעילות של אומדים

ראינו כי עבור אומדים לא-מוטים האומד שמביא למינימום את הסיכון הריבועי (משוקלל על פני כל המדגמים האפשריים) הוא האומד שהשונות שלו $Var(\hat{\theta}) = E_{X^{(n)}}[(\hat{\theta} - \theta)^2]$ היא מינימאלית. היות והאומד הוא פונקציה של המדגם, הרי שהשונות שלו תלויה בהתפלגות היוצרת את הדגימות. נחקור כעת את התלות הזאת.

נטפל תחילה במקרה של פרמטר יחיד (סקלר ולא וקטור) ונגדיר ציון S (Fisher score) של מדגם $X^{(n)}$ המתפלג על פי $P(x|\theta)$ בתור הנגזרת ביחס לפרמטר θ של לוג הנראות

$$(3.18) \quad S(X^{(n)}, \theta) = \frac{\partial}{\partial \theta} \log L(X^{(n)} | \theta)$$

הציון S מודד את הרגישות של ההסתברות לקבל את המדגם $X^{(n)}$ לשינוי בערך הפרמטר θ : אם שינוי קטן בערך הפרמטר גורר שינוי גדול בהסתברות לראות את המדגם, הרי שהציון יהיה גבוה. במקרה ש- θ הוא פרמטר רב ממדי, הרי ש-S יהיה וקטור הנגזרות החלקיות.

הציון S הוא פונקציה של המדגם ולכן גם הוא משתנה מקרי. נחקור את ההתפלגות של S על ידי שנחשב את המומנטים הראשונים של ההתפלגות.

ראשית, התוחלת של S מקיימת

$$\begin{aligned} E_{X^{(n)}}(S) &= \int \frac{\partial}{\partial \theta} \log(P(X^{(n)} | \theta)) \cdot P(X^{(n)} | \theta) dX^{(n)} \\ &= \int \frac{\frac{\partial}{\partial \theta} P(X^{(n)} | \theta)}{P(X^{(n)} | \theta)} \cdot P(X^{(n)} | \theta) dX^{(n)} \\ (3.19) \quad &= \int \frac{\partial}{\partial \theta} P(X^{(n)} | \theta) dX^{(n)} \\ &= \frac{\partial}{\partial \theta} \int P(X^{(n)} | \theta) dX^{(n)} \\ &= \frac{\partial}{\partial \theta} (1) \\ &= 0. \end{aligned}$$

כלומר, תוחלת הציון היא תמיד אפס ולפיכך, $E[S^2] = Var(S)$. השונות של S היא בעלת חשיבות מיוחדת, ונתאר אותה להלן.

הגדרה: אינפורמציות פישר

השונות של $S(X^{(n)}, \theta)$ נקראת אינפורמציות פישר $J_n(\theta)$

$$(3.20) \quad J_n(\theta) = E \left[\left(\frac{\partial}{\partial \theta} \log P(X^{(n)} | \theta) \right)^2 \right] = \int \left(\frac{\partial}{\partial \theta} \log P(X^{(n)} | \theta) \right)^2 P(X^{(n)} | \theta) dX^{(n)}$$

במקרים בהם ההתפלגות מאופיינת ע"י מספר פרמטרים $\underline{\theta} = (\theta_1, \theta_2, \dots, \theta_n)$, אז גם הציון S הוא משתנה מקרי רב מימדי. באופן טבעי, במקום פרמטר שונות יחיד נשתמש במטריצת השונות המשותפת (ראה סעיף 1.5.7). אינפורמציות-פישר תהיה לכן המטריצה הבאה

$$(3.21) \quad J_n(\theta_i, \theta_j) = E_{X^{(n)}} \left[\frac{\partial \log P(X^{(n)} | \underline{\theta})}{\partial \theta_i} \cdot \frac{\partial \log P(X^{(n)} | \underline{\theta})}{\partial \theta_j} \right]$$

וניתן גם להראות (ראה תרגיל) כי

$$(3.22) \quad J_n(\theta_i, \theta_j) = -E_{X^{(n)}} \left[\frac{\partial^2 \log P(X^{(n)} | \underline{\theta})}{\partial \theta_i \partial \theta_j} \right]$$

מטריצה זאת נקראת גם מטריצת ה-Hessian של לוג ההתפלגות.

אם X_1, X_2, \dots, X_n מתפלגים i.i.d. קל לראות כי

$$(3.23) \quad J_n(\theta) = nJ_1(\theta)$$

אינפורמציות פישר מהווה מדד לכמות האינפורמציה על הפרמטר θ שיש במדגם. אינטואיטיבית, ככל שיש יותר נקודות ששעבורן ערכו המוחלט של הציון גבוה, המדגם רגיש יותר לשינויים ב- θ וקל יותר לשערך את θ מתוך המדגם.

דוגמא: אינפורמציות פישר עבור דגימה בודדת מהתפלגות אקספוננציאלית

ההתפלגות של דגימה מהתפלגות אקספוננציאלית היא

$$P(x|\theta) = \frac{1}{\theta} e^{-x/\theta}$$

הציון S הוא

$$S(x|\theta) = \frac{\partial \log P(x|\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \left(-\log(\theta) - \frac{x}{\theta} \right) = -\frac{1}{\theta} + \frac{x}{\theta^2}$$

ואינפורמציות-פישר (התוחלת של S^2) היא

$$\begin{aligned} J(\theta) &\equiv E_x[S^2] = E_x \left[\frac{1}{\theta^2} - 2\frac{x}{\theta^3} + \frac{x^2}{\theta^4} \right] = \\ &= \left(\frac{1}{\theta} \right)^2 - 2\frac{\theta}{\theta^3} + \frac{2\theta^2}{\theta^4} = \left(\frac{1}{\theta} \right)^2 \end{aligned}$$

כלומר הכמות של אינפורמציות-פישר על θ שיש בדגימה x מתנהגת כמו $\frac{1}{\theta^2}$.

וניתן גם לחשב תוך שימוש בנגזרת השניה

$$\begin{aligned} J(\theta) &\equiv E_x \left[-\frac{\partial^2 \log(P(x|\theta))}{\partial \theta^2} \right] = \\ &= -E_x \left[\frac{\partial S}{\partial \theta} \right] = -E_x \left[\frac{1}{\theta^2} - 2\frac{x}{\theta^3} \right] = \frac{2\theta}{\theta^3} - \frac{1}{\theta^2} = \frac{1}{\theta^2} \end{aligned}$$

אי-שויון קרמר-ראו (Cramer-Rao) עבור אומד לא - מוטה

נראה כעת תוצאה חשובה המקשרת את השונות של אומד לא מוטה לאינפורמציות פישר.

נביט בתוחלת המכפלה של האומד $\hat{\theta}(X^{(n)})$ (אומד לא מוטה) ושל הציון

$$S(X^{(n)})$$

$$\begin{aligned}
 [E(S \cdot \hat{\theta})]^2 &= \\
 &_1 = [E(S \cdot \hat{\theta}) - E(S)E(\hat{\theta})]^2 \\
 &_2 = [COV(S, \hat{\theta})]^2 \\
 (3.24) \quad &_3 = [\rho \cdot \sqrt{Var(S) \cdot Var(\hat{\theta})}]^2 \\
 &_4 \leq Var(S) \cdot Var(\hat{\theta}) \\
 &_5 = J(\theta) \cdot Var(\hat{\theta})
 \end{aligned}$$

נסביר את המעברים אחד לאחד: (1) נובע מכך שתוחלת הציון S היא אפס. (2) הגדרת השונות המשותפת. (3) הגדרת מקדם המתאם. (4) נובע מאי שוויון קושי שוורץ, או לחלופין מכך שמקדם המתאם מקבל ערכים בין -1 ל-1. (5) הגדרת אינפורמציה פישר $J(\theta) = Var(S)$.

במקרה של פרמטר רב ממדי, אז כאמור גם S הוא רב ממדי, והמעברים לעיל מתקיימים עבור כל אחד מהרכיבים $E^2(S_i \cdot \hat{\theta}_i) \leq J_{ii}(\theta) \cdot Var(\hat{\theta}_i)$ כאשר J_{ii} הוא האיבר ה- i באלכסון מטריצת אינפורמציה פישר.

מצד שני

$$\begin{aligned}
 E_{X^{(n)}}(S \cdot \hat{\theta}) &= \\
 &= \int \frac{\frac{\partial}{\partial \theta} (P(X^{(n)} | \theta))}{P(X^{(n)} | \theta)} \cdot \hat{\theta}(X^{(n)}) P(X^{(n)} | \theta) dX^{(n)} \\
 (3.25) \quad &= \frac{\partial}{\partial \theta} \int \hat{\theta}(X^{(n)}) \cdot P(X^{(n)} | \theta) dX^{(n)} \\
 &= \frac{\partial}{\partial \theta} E_{X^{(n)}}(\hat{\theta}) = \frac{\partial}{\partial \theta} \theta = 1
 \end{aligned}$$

קיבלנו אם כן כי

$$J(\theta) \cdot Var(\hat{\theta}) \geq 1$$

או

$$(3.26) \quad Var(\hat{\theta}) \geq \frac{1}{J(\theta)}$$

במקרה של משתנה מקרי רב ממדי נקבל באופן דומה אי שוויון בין מטריצות

$$(3.27) \quad C(\hat{\theta}) \geq (J(\theta))^{-1}$$

כאשר $C(\hat{\theta})$ היא מטריצת השונות המשותפת של המשתנה הוקטורי $\hat{\theta}$. אי השוויון בין המטריצות מוגדר באופן הבא: מטריצה A נקראת חיובית, ונסמן $A > 0$ אם מתקיים $x^T A x > 0$ לכל וקטור x . נאמר ש- $A > B$ אם $A - B > 0$. נשים לב כי היות ומותר לבחור עבור x את וקטורי הבסיס הסטנדרטי, אז בפרט מתקיים עבור כל רכיב של הפרמטר הוקטורי

$$(3.28) \quad \text{Var}(\hat{\theta}_i) \geq (J(\theta))^{-1}_{ii}$$

כאשר הביטוי מימין פירושו האיבר ה- i באלכסון של המטריצה $(J(\theta))^{-1}$.

המשמעות של אי שוויון קרמר-ראו היא שיש חסם תחתון על השגיאה שבה אנחנו יכולים לצפות לאמוד נכונה את θ . היות והדגימות נוצרות באופן סטוכסטי, אזי

רעש הדגימה יגרום לשונות באומד $\hat{\theta}$, שערכה לכל הפחות $\frac{1}{J(\theta)}$.

משפט קרמר-ראו מאפשר לנו לכן, להעריך את מידת הטעות (השונות) באמידת הפרמטר θ .

אי-שויון קרמר-ראו עבור אומד מוטה

נרשום

$$E(\hat{\theta}) = \int_{-\infty}^{\infty} \hat{\theta}(X^{(n)}) P(X^{(n)} | \theta) dX^{(n)} = \theta + b(\theta)$$

כאשר $b(\theta)$ היא ההטיה (bias), ונזכור כי מתקיים

$$1 = \int_{-\infty}^{\infty} P(X^{(n)} | \theta) dX^{(n)}$$

נגזור כעת את שני האגפים (ונסמן $b'(\theta) = \partial b / \partial \theta$)

$$\begin{aligned} 1 + b'(\theta) &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \hat{\theta}(X^{(n)}) P(X^{(n)} | \theta) dX^{(n)} = \\ &= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta} (P(X^{(n)} | \theta))}{P(X^{(n)} | \theta)} \cdot \hat{\theta}(X^{(n)}) \cdot P(X^{(n)} | \theta) dX^{(n)} \\ &= \int_{-\infty}^{\infty} \hat{\theta} \frac{\partial \ln P}{\partial \theta} P(X^{(n)} | \theta) dX^{(n)} \\ &= E(S \cdot \hat{\theta}) \end{aligned}$$

$$0 = \int \frac{\partial \ln P}{\partial \theta} P(X^{(n)} | \theta) dX^{(n)} = \left\langle \frac{\partial \ln P}{\partial \theta} \right\rangle_{P(X^{(n)} | \theta)} \Rightarrow E(S) = 0$$

ומאותם שיקולים כמו עבור האומד הלא-מוטה

$$(3.29) \quad \begin{aligned} \left(E(S \cdot \hat{\theta})\right)^2 &= (1 + b'(\theta))^2 \leq \text{Var}(\hat{\theta}) \text{Var}\left(\frac{\partial \ln P}{\partial \theta}\right) = \text{Var}(\hat{\theta}) \cdot J(\hat{\theta}) \\ &\Rightarrow \text{Var}(\hat{\theta}) \geq \frac{(1 + b'(\theta))^2}{J(\theta)} \end{aligned}$$

כלומר: בדיוק כמו עבור אי-שוויון קרמר-ראו המקורי, כאשר במונה מופיע לנו $(1 + b'(\theta))^2$ במקום "1" (שימו לב שכעת החסם יכול להיות קטן יותר ואפילו שווה ל-0, אם $b'(\theta) = -1$).

הגדרה - אומדים יעילים

אומד יקרא יעיל אם מתקיים עבורו שוויון באי-שוויון קרמר-ראו, כלומר

$$(3.30) \quad \text{Var}(\hat{\theta}) = \frac{1}{J(\theta)}$$

הראנו כי קיים חסם תחתון על השונות של האומד, ונרצה לאפיין מתי חסם תחתון זה הוא הדוק. כלומר, עבור אילו התפלגויות ניתן למצוא אומד יעיל הממצה את המידע שיש לנו במדגם. הסעיף הבא ידון בהתפלגויות עבורן קיימים אומדים יעילים.

3.5.4 התפלגויות אקספוננציאליות

על מנת שאומד יהיה יעיל, הרי שהוא צריך לקיים שוויון באי-שוויון קרמר ראו. במקרה של פרמטר סקלרי, תכונה זו מתקיימת אם ורק אם

$$(3.31) \quad \left(E_{X^{(n)}}(S \cdot \hat{\theta}) \right)^2 = \text{Var}(S) \cdot \text{Var}(\hat{\theta}) = J(\hat{\theta}) \cdot \text{Var}(\hat{\theta})$$

ושוויון כזה מתקיים אם ורק אם הערך המוחלט של מקדם המתאם הוא 1 (או לחלופין ישנו שוויון באי שוויון קושי שווארץ), כלומר אם ורק אם $S(X^{(n)}, \theta)$ ו- $\hat{\theta}(X^{(n)})$ תלויים לינארית, כלומר קיים סקלר $\lambda(\theta)$ כך שמתקיים

$$S(X^{(n)}) = \frac{\partial}{\partial \theta} \log(P(X^{(n)} | \theta)) = \lambda(\theta) \cdot \hat{\theta}(X^{(n)}) + c, \quad \forall X^{(n)}$$

מהאדיטיביות של S נובע

$$\sum_{i=1}^n \frac{\partial}{\partial \theta} \log(P(x_i | \theta)) = \lambda(\theta) \cdot \hat{\theta}(X^{(n)}) + c'$$

ואם $\log(P(x | \theta))$ גזיר לפי x אז מאחר ובאגף שמאל של המשוואה רק האיבר שמכיל את x_i בסכום תלוי ב- x_i אנו מקבלים

$$\frac{\partial^2}{\partial x_i \partial \theta} \log(P(x_i | \theta)) = \lambda(\theta) \cdot \frac{d\hat{\theta}}{dx_i}$$

זאת אומרת, באופן כללי

$$\frac{\partial^2}{\partial x_i \partial \theta} \log(P(x_i | \theta)) = \lambda(\theta) a(x_i)$$

ומאינטגרציה על θ אנו מקבלים

$$\frac{\partial}{\partial x_i} \log(P(x_i | \theta)) \equiv a(x_i) b(\theta) + c(x_i)$$

ואינטגרציה נוספת על x_i תיתן

$$\log P(x_i | \theta) = A(x_i) \cdot B(\theta) + C(x_i) + D(\theta)$$

כאשר "קבועי-האינטגרציה" A ו- C הן פונקציות של x בלבד, ואינן תלויות ב- θ . קיבלנו אם כן כי ההתפלגות היא בהכרח מהצורה הבאה

$$(3.32) \quad P(X^{(n)} | \theta) = \exp \left[A(X^{(n)}) B(\theta) + C(X^{(n)}) + D(\theta) \right]$$

משפחת ההתפלגויות שניתן להציג אותן בצורה זו קיבלה לכן שם מיוחד.

הגדרה

התפלגות נקראת **מהמשפחה האקספוננציאלית** אם קיימות פונקציות A, B, C ו-D (שהן סקלריות או וקטוריות) כך שניתן לכתוב את פונקציית ההתפלגות בצורה הבאה

$$(3.33) \quad P(x|\theta) = \exp[A(x) + B(\theta) \cdot C(x) + D(\theta)].$$

המשפחה האקספוננציאלית היא משפחה רחבה, המכילה את רוב ההתפלגויות המופיעות בספרי הלימוד. למשל, ההתפלגות הנורמלית היא מהמשפחה האקספוננציאלית, את ההתפלגות הבינומית אפשר לרשום בצורה אקספוננציאלית

$$\begin{aligned} P_n(m) &= \binom{n}{m} \theta^m (1-\theta)^{n-m} \\ &= \exp\left[\log \frac{n!}{m!(n-m)!}\right] \cdot \exp[m \cdot \log \theta + (n-m) \log(1-\theta)] \\ &\approx \exp\left[\frac{1}{2} \log \frac{n}{2\pi m(n-m)} - n \left(\log \frac{m}{n} + \log \frac{n-m}{n}\right)\right] \\ &\quad \cdot \exp[m \cdot \log \theta + (n-m) \log(1-\theta)] \end{aligned}$$

וכמובן גם ההתפלגות האקספוננציאלית $\theta \exp(-x\theta)$ היא מהמשפחה האקספוננציאלית.

לעומת זאת, ההתפלגויות בעולם האמיתי הן, פעמים רבות, לא אקספוננציאליות. לדוגמה, התפלגות של ציונים במבחן המשקף הבנה של חומר קשה היא פעמים רבות, תערובת של שני גאוסיאנים: גאוסיאן אחד מתאר את אוכלוסיית הסטודנטים שהבינו את החומר, והגאוסיאן השני מתאר את אוכלוסיית הסטודנטים שלא הבינו את החומר. ממוצע הציונים הכיתתי וסטית התקן אינם משקפים את ההתפלגות האמיתית היות וצריך להתייחס לכל אוכלוסייה בנפרד. התפלגויות מסוג זה יהיו במוקד הפרק הבא.

לסיכום ננסה במפורש את המשפט העיקרי שהוכחנו בפתח הסעיף הנוכחי:

משפט:

להתפלגות $P(x|\theta)$ קיים אומד יעיל אם ורק אם $P(x|\theta)$ היא מהצורה האקספוננציאלית.

3.5.5 אופטימליות אסימפטוטית

בגבול של מדגם גדול אומד ML הוא אומד יעיל. לא נביא כאן הוכחה מלאה לכך, אלא נוכיח תוצאה אחרת המספקת אינטואיציה מעניינת. המעוניין בהוכחה יוכל למצוא אותה למשל בספר של Kay.

ראשית נזכיר כי אומד ML הוא קונסיסטנטי, דהיינו, שואף לערך האמיתי של הפרמטר בגבול של מדגם גדול $P(\theta_{ML}(X^{(n)}) = \theta) \xrightarrow{n \rightarrow \infty} 1$. זאת היות ובשל חוק

המספרים הגדולים הנראות של המדגם שואפת לתוחלת הנראות

$$\frac{1}{n} \log(P(X^{(n)} | \theta)) = \frac{1}{n} \sum_{i=1}^n \log(P(X_i | \theta)) \rightarrow \sum_{\{x\}} P(X_i | \theta) \log(P(X_i | \theta))$$

והנראות על פי הפרמטר האמיתי θ^* גדולה מהנראות על פי כל פרמטר אחר θ היות ו-

$$\begin{aligned} & \sum_{\{x\}} P(X_i | \theta^*) \log(P(X_i | \theta^*)) - \sum_{\{x\}} P(X_i | \theta^*) \log(P(X_i | \theta)) = \\ & = D_{KL}[P(X_i | \theta^*) \| P(X_i | \theta)] > 0 \end{aligned}$$

נטפל כעת בנראות המותנית

$$P(X^{(n)} | \theta) = \exp(\log(P(X^{(n)} | \theta)))$$

נפתח טור טיילור של $\log(P(X^{(n)} | \theta))$ בסביבת הערך של אומד הנראות המקסימלית

$$\begin{aligned} \log(P(X^{(n)} | \underline{\theta})) &= \log(P(X^{(n)} | \underline{\theta}_{ML})) \\ &+ \nabla_{\underline{\theta}} \log(P(X^{(n)} | \underline{\theta}_{ML})) (\underline{\theta}_{ML} - \underline{\theta}) \\ &+ \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2 \log(P(X^{(n)} | \underline{\theta}_{ML}))}{\partial \theta_i \partial \theta_j} (\underline{\theta}_{ML} - \underline{\theta})^2 \\ &+ O((\underline{\theta}_{ML} - \underline{\theta})^3) \end{aligned}$$

כעת נשים לב כי האיבר השני (הלינארי) בטור מכיל את הנגזרות של הנראות בנקודה של θ_{ML} , אך היות ו- θ_{ML} נבחר כדי להביא למקסימום את הנראות הרי שאיבר זה מתאפס. בנוסף לכך, האיבר האחרון זניח בגבול של מדגם גדול. האיבר המעניין ביותר הוא האיבר הריבועי, והוא שווה לאינפורמציה פשוט על הפרמטר לפי ההגדרה (ראה 3.22). קיבלנו אם כן כי בגבול של מדגם גדול, לוג הנראות הוא מהצורה

$$\log(P(X^{(n)} | \underline{\theta})) \approx \log(P(X^{(n)} | \underline{\theta}_{ML})) - n \frac{1}{2} J(\theta_{ML})(\theta_{ML} - \theta)^2$$

דהיינו לנראות המותנית כפונקצית של הפרמטר יש צורה של התפלגות נורמלית עם תוחלת θ_{ML} ושונות $1/nJ(\theta_{ML})$

$$P(X^{(n)} | \underline{\theta}) \approx P(X^{(n)} | \underline{\theta}_{ML}) \exp\left[-n \frac{1}{2} J(\theta_{ML})(\theta_{ML} - \theta)^2\right]$$

כאמור, ניתן גם להראות תוצאה משלימה, שעל פיה אומד ML הוא יעיל בגבול של מדגם גדול.