GMDL212, HW3

Yuval Margalit

problem1:

נגדיר פונקציה חדשה

$$L(\theta,\theta^t,\lambda)=Q(\theta,\theta^t)-\lambda((\sum_{i=1}^{K}\pi_k)-1)=(\sum_{i=1}^{N}\sum_{k=1}^{K}r_{i,k}log\pi_k)+(\sum_{i=1}^{N}\sum_{k=1}^{K}r_{i,k}logp(x_i;\theta_k)-\lambda((\sum_{i=1}^{K}\pi_k)-1)=$$

$$=(\sum_{i=1}^{N}\sum_{k=1}^{K}r_{i,k}log\pi_k)+(\sum_{i=1}^{N}\sum_{k=1}^{K}r_{i,k}(log(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_k|^{\frac{1}{2}}})-\frac{||x_i-\mu_k||_{\Sigma_k^{-1}}}{2}))-\lambda((\sum_{i=1}^{K}\pi_k)-1)$$

(1)

יהי $1\leq k\leq K$. נמצא את הנגזרות החלקיות ונשווה ל0

$$\frac{\partial L}{\partial\pi_k}=\sum_{i=1}^{N}\frac{r_{i,k}}{\pi_k}-\lambda=0$$

$$\sum_{i=1}^{N}r_{i,k}=\lambda\pi_k\Leftrightarrow\pi_k=\frac{\sum\limits_{i=1}^{N}r_{i,k}}{\lambda}$$

$$\frac{\partial L}{\partial\lambda}=\sum_{k=1}^{K}\pi_k-1=0\Leftrightarrow\sum_{i=1}^{K}\pi_k=1$$

אז

$$\sum_{k=1}^{K}\pi_k=\sum_{k=1}^{K}\frac{\sum\limits_{i=1}^{N}r_{i,k}}{\lambda}=1$$

ולכן $\lambda=\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N}r_{i,k}=N$ אז $\pi_k=\frac{\sum_{i=1}^{N}r_{i,k}}{\lambda}=\frac{\sum_{i=1}^{N}r_{i,k}}{N}=\frac{N_k}{N}$

אנו מקבלים מקסימום כיוון וQ הינה concave (סכום של פונקציות concave הינו concave) והוספה של אילוץ לינארי לא משנה מבחינת הקמירות/קעירות של פונקציה.

(2)

נשים לב כי הרכיב היחיד שמכיל את μ_k בפונקציה הינו $-\sum_{i=1}^{N} r_{i,k}\frac{||x_i-\mu_k||_{\Sigma_k^{-1}}}{2} = -\sum_{i=1}^{N}\frac{1}{2}r_{i,k}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k)$
כיוון ומדובר במינוס סכום של מכפלת מספרים חיוביים נרצה למזער את הביטוי בסכום כדי למצוא את המקסימום של הפונקציה ביחס לμ_k.
כיוון וΣהיא SPD קיים לה פירוק צ'קולסקי $LL^T = \Sigma$ ולכן $\Sigma^{-1} = (LL^T)^{-1} = L^{-T}L^{-1}$

$$\sum_{i=1}^{N} r_{i,k}(x_i-\mu_k)^T\Sigma^{-1}(x_i-\mu_k) = \sum_{i=1}^{N} r_{i,k}(x_i-\mu_k)^T L^{-T}L^{-1}(x_i-\mu_k) = \sum_{i=1}^{N} r_{i,k}||L^{-1}(x_i-\mu_k)||_2$$

אז אנחנו מחפשים את $argmin_{\mu_k}\sum_{i=1}^{N} r_{i,k}||L(x_i-\mu_k)||_2$ נשים לי כי מדובר ב whigthed LLS ולכן קיים פתרון סגור.
נגדיר $W = diag(r_{1,k},...,r_{1,k}|zdtimes, r_{2,k},...,r_{2,k}|zdtimes, r_{N,k},...,r_{N,k}|zdtimes)$
$H = [L^{-1},...,L^{-1}]^T, y = [L^{-1}x_1,...,L^{-1}x_N]^T$
כיוון ולכל i,k מתקיים כי $r_{i,k} > 0$ (לגאוסיין תמד אינסופי) אז W הפיכה ולכן

$$\begin{aligned}
\mu_k &= (H^TWH)^{-1}H^TWy = \\
&= (\sum_{i=1}^{N} r_{i,k}L^{-T}L^{-1})^{-1}\sum_{i=1}^{N} r_{i,k}L^{-T}L^{-1}x_i = (\sum_{i=1}^{N} r_{i,k})^{-1}(L)(L^T)(L^{-T})(L^{-1})(\sum_{i=1}^{N} r_{i,k}x_i) = \\
&=(\sum_{i=1}^{N} r_{i,k})^{-1}(\sum_{i=1}^{N} r_{i,k}x_i) = \frac{\sum_{i=1}^{N} r_{i,k}x_i}{\sum_{i=1}^{N} r_{i,k}} = \frac{\sum_{i=1}^{N} r_{i,k}x_i}{N_k}
\end{aligned}$$

problem 2:

(1)

עבור $\alpha = \overrightarrow{1}$ תתקבל התפלגות אחידה מעל כל ההתפלגויות הקטגוריליות ממימד K. כיוון ולכל התפלגות קטוגריליותπ יתקיים כי

$$Dir(\pi;\overrightarrow{1}) \propto \prod_{k=1}^{K}\pi_k^0 = 1$$

כלומר לכל פאי קיימת אותה הסתברות להיבחר.

(2)

π אחידה אומר ש $\pi = \frac{1}{K}\times 1_{1\times K}$ כלומר הסיכוי להיות בכל קטגוריה הינו זהה.
לעומת זאת π שהוגרלה מהתפלגות אחידה אומר שלכל התפלגות קטגוריאלית $\pi^{'}$ קיים אותו סיכוי להידגם.

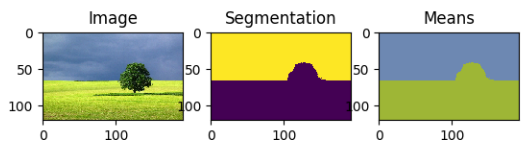
(3)

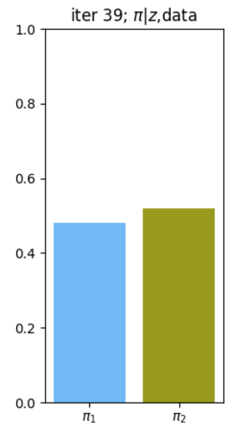
שימוש בפריור בהקשר של π מאפשר הכנסת ידע מוקדם או הנחות כלשהן על ההתפלגות וגם מאפשר טיפול טוב יותר ברעשים בדגימה.
בנוסף השימוש בפריור יכול למנוע הגעה למצבים מסוים (לדוגמא $\pi = [0.99, 0.01]$) שאינם רצויים.

Computer Exercise 1:

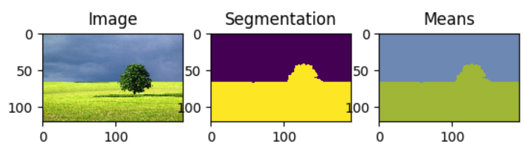
k=2

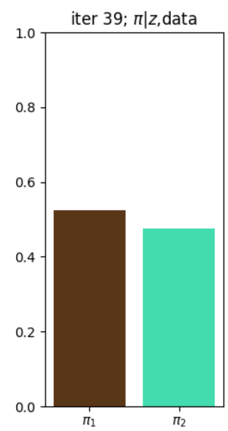
$\alpha = 1, \Psi = 0.01 I_{3\times3}, \nu = 3.1, \kappa = 1, m = 0.5$



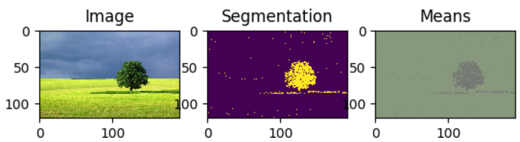


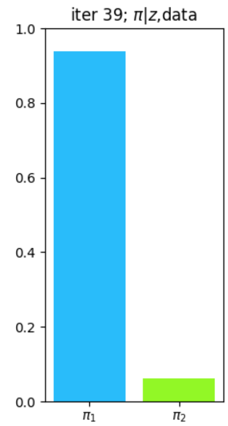
$\alpha = 1, \Psi = 0.01 I_{3\times3}, \nu = 1000, \kappa = 1, m = 0.5$



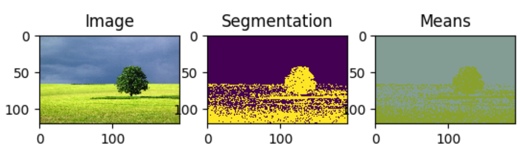


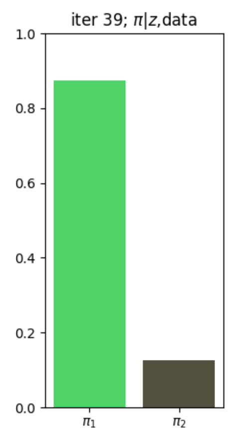
$\alpha = 100, \Psi = 0.01 I_{3\times3}, \nu = 1000, \kappa = 10000, m = 0.5$





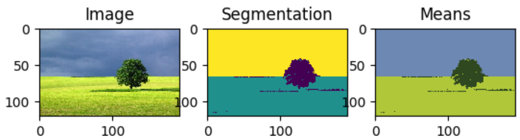
$\alpha = [30000, 1], \Psi = 0.01 I_{3\times3}, \nu = 3.1, \kappa = 1, m = 0.5$

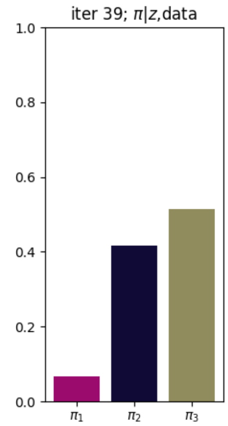




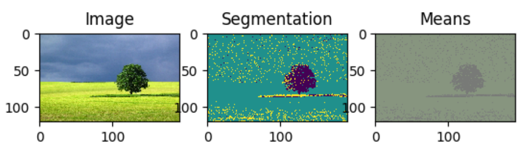
k=3

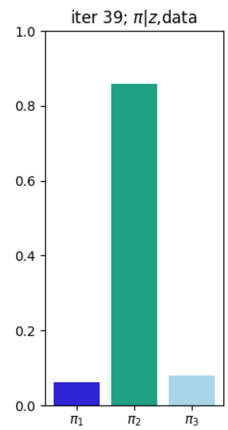
$\alpha = 1, \Psi = 0.01 I_{3\times3}, \nu = 1000, \kappa = 1, m = 0.5$



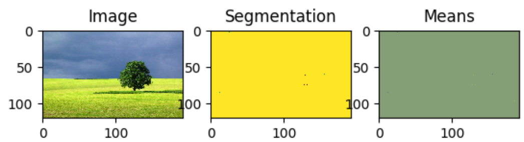


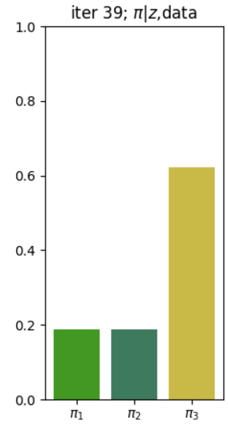
$\alpha = 100, \Psi = 0.5 I_{3\times3}, \nu = 1000, \kappa = 15000, m = 0.5$



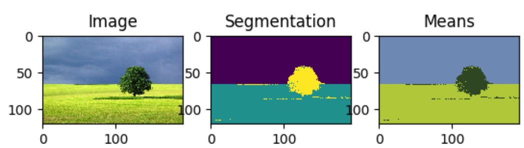


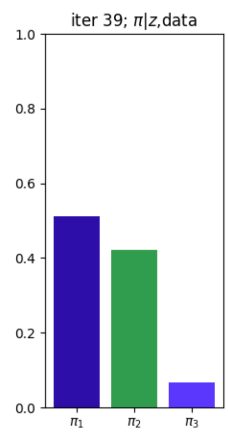
$\alpha = 10000, \Psi = I_{3\times3}, \nu = 1000, \kappa = 1, m = 1$





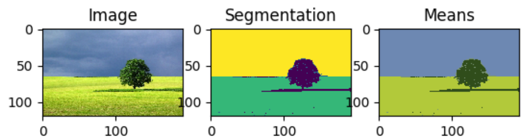
$$\alpha = 100, \Psi = \begin{bmatrix} 0.02 & 0 & 0 \\ 0 & 0.0025 & 0 \\ 0 & 0 & 0.0025 \end{bmatrix}, \nu = 1000, \kappa = 1, m = 0.5$$

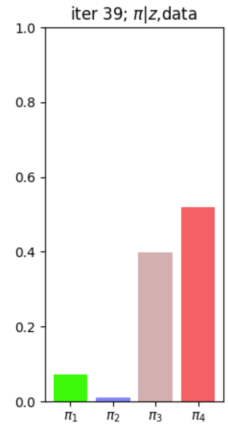




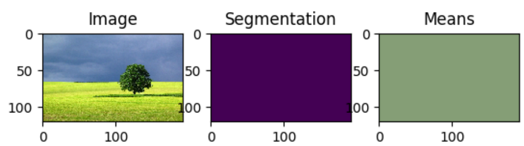
k=4

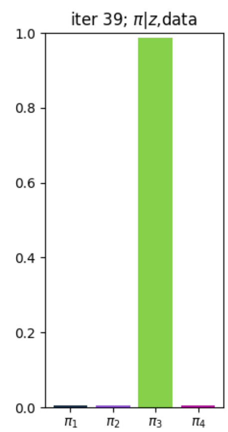
$\alpha = 1, \Psi = 0.01 I_{3\times3}, \nu = 50, \kappa = 1, m = 0.5$



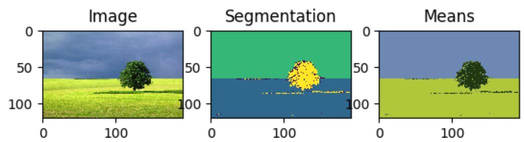


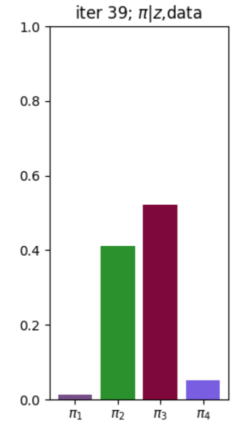
$\alpha = 1, \Psi = I_{3\times3}, \nu = 1000, \kappa = 1, m = 0.5$



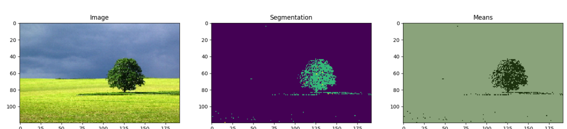


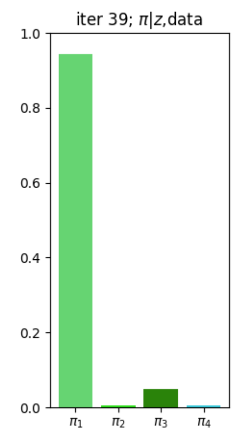
$\alpha = 1, \Psi = 0.01 I_{3\times 3}, \nu = 1000, \kappa = 1, m = 0.5$





$\alpha = 100, \Psi = 0.1 I_{3\times 3}, \nu = 1000, \kappa = 1, m = 0.5$





בגלל טווח הערכים הקטן בחירה של Ψ כמטריצה בעלת ערכים יחסית "גדולים" ($\sigma : 3\sigma \sim 0.3$ לכל צבע) מביא לכך שחלק נרחב מהטווח של הצבעים בתמונה מכוסה ע"י גאוסיין יחיד ולכן גאוסיין יחיד נבחר כמעט לכל הנקודות ובכך נוצרת סגמנטציה "רעה" כאשר הגואיסיינים האחרים נבחרים רק בנקודות "קצה" (הופכים לקבועים/כמעט קבועים).
בנוסף בחירה של α באופן היוצר הבדל גדול ומשמעותי בין הגאוסייינים גם היא מייצרת מצב זהה.
ניתן לראות שלprior יש השפעה על איכות התוצאה ולכן בחירת prior שהינו בעל ערכים שאינם מתאמים עם pseudo count גדול יתן סגמנטציות שאינן טובות.

Problem 3:

(2)

הבחירה המתאימה של ההיפר פרמטרים תהייה כזאת הנותנת משקל חזק לlikelihood על פני הpostirior. אזי נרצה pseudo count קטן ככל הניתן.

פרמטרים המתאימים לכך יהיו $m = \vec{0.5}, \kappa = 0.1, \nu = 2.01, \alpha = \vec{0.1}$ ו $\Psi = 0.001I$.

(1)

בכדי לקבל תוצאות דומות לk-means clustering אבחר את ההיפר פרמטרים הנותנים משקל חזק לlikelihood על פני הpostirior וגם בסיכוי גבוהה $\Sigma_k = \Sigma = \sigma I_{3\times3}$ כאשר $\sigma \to 0$ כיוון ובחירה זה מממשת k-means ב non baysian EM GMM.

היפר פרמטרים מתאים יהיו $\kappa = 0.1, \Psi = 0.000001 I_{3\times3}, \kappa = 0.1, \nu = 2.01, \alpha = \vec{0.1}$ ו $m = \vec{0.5}$.