Weekly Exercise 5

Dr. Csaba Domokos
Technische Universität München, Computer Vision Group
May 29th, 2017 (submission deadline: May 29th, 2017)

The EM algorithm for mixtures of Gaussians (6 Points)

Exercise 1 (**E step**, 1 Point)**.** Consider a mixture of Gaussians with K component. Assume that we are given N data samples $\{\mathbf{x}_n\}_{n=1}^N$ and a current guess of parameters $\boldsymbol{\theta}^{\text{old}} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Show that

$$p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \triangleq \gamma_k(\mathbf{x}_n) \qquad \text{for all } n = 1, \ldots, N \; .$$

Solution. First we apply the Bayes' rule and substitute the likelihood $p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta})$, the joint distribution $p(\mathbf{z} \mid \boldsymbol{\theta})$ and the definition of the density function $p(\mathbf{x} \mid \boldsymbol{\theta})$ corresponding to the mixture of K Gaussians.

$$\begin{aligned} p(\mathbf{z}_n \mid \mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) &= \frac{p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\theta}^{\text{old}}) \, p(\mathbf{z}_n \mid \boldsymbol{\theta}^{\text{old}})}{p(\mathbf{x}_n \mid \boldsymbol{\theta}^{\text{old}})} \\ &= \frac{\prod_{k=1}^K \left(\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{nk}} \pi_k^{z_{nk}}}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \\ &= \frac{\pi_k \, \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \triangleq \gamma_k(\mathbf{x}_n) \; . \end{aligned}$$

Exercise 2 (**M step**, 5 Points)**.** Assume a mixture of Gaussians with K component and N data samples $\{\mathbf{x}_n\}_{n=1}^N$. The *log-likelihood* function is given as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^N \sum_{k=1}^K \gamma_k(\mathbf{x}_n) \big(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) \; .$$

a) Show that the optimal choice with respect to the *mean vectors* $\boldsymbol{\mu}_k$ for all $k = 1, \ldots, K$ is given as

$$\arg\max_{\boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = \frac{\sum_{n=1}^N \gamma_k(\mathbf{x}_n) \mathbf{x}_n}{\sum_{m=1}^N \gamma_k(\mathbf{x}_m)} \; .$$

Hint: for a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2 \mathbf{A} \mathbf{x} \; .$$

b) Show that the optimal choice with respect to the *covariance matrices* $\boldsymbol{\Sigma}_k$ for all $k = 1, \ldots, K$ is given as

$$\arg\max_{\boldsymbol{\Sigma}_k} \mathcal{L}(\boldsymbol{\theta}) = \frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)} \ .$$

Hint: for a symmetric matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ and vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \mathbf{X}} \mathbf{a}^T \mathbf{X}^{-1} \mathbf{b} = -\mathbf{X}^{-T} \mathbf{a} \mathbf{b}^T \mathbf{X}^{-T} \ ,$$

and for a non-singular matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$,

$$\frac{\partial}{\partial \mathbf{X}} |\mathbf{X}| = |\mathbf{X}| \ \mathbf{X}^{-1} \ .$$

Solution. a) We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$.

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \frac{1}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \ \frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \ .$$

Let us now consider the derivative of a Gaussian w.r.t. $\boldsymbol{\mu}_k$.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \ \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \Big) \\
&= -\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \Big) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \\
&= -\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \ .
\end{aligned}$$

By substituting back and setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\mu}_k$ to 0, we get

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathcal{L}(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x}_n)}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \\
\frac{\sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \ \mathbf{x}_n}{\sum_{m=1}^{N} \gamma_k(\mathbf{x}_m)} = \boldsymbol{\mu}_k \ .
\end{aligned}$$

b) We calculate the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \gamma_k(\mathbf{x}_n) \frac{1}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \ \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x})}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \ \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) \Big) \right) \ .
\end{aligned}$$

Let us first calculate the following derivatives:

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} = \frac{1}{(2\pi)^{\frac{D}{2}}} - \frac{1}{2} |\boldsymbol{\Sigma}_k|^{-\frac{3}{2}} |\boldsymbol{\Sigma}_k| \boldsymbol{\Sigma}_k^{-1} = \frac{-\boldsymbol{\Sigma}_k^{-1}}{2\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \ .$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \frac{-1}{2}(-\boldsymbol{\Sigma}_k^{-T})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_k^{-T} \\
&= \frac{1}{2}\exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \ .
\end{aligned}
$$

Now we are at the position to calculate the derivative of a Gaussian w.r.t. $\boldsymbol{\Sigma}_k$.

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left(\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big)\right) \\
=&\frac{-\boldsymbol{\Sigma}_k^{-1}}{2\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \\
&+ \frac{1}{2}\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} \exp\Big(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
=& -\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \frac{1}{2}\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
=&\frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{2}\boldsymbol{\Sigma}_k^{-1}\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \ .
\end{aligned}
$$

Therefore, the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$ is as follows.

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \frac{\gamma_k(\mathbf{x}_n)}{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{2}\boldsymbol{\Sigma}_k^{-1}\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \\
&= \frac{\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) \ .
\end{aligned}
$$

Setting the derivative of $\mathcal{L}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\Sigma}_k$ to 0, we get that

$$
\begin{aligned}
\frac{\boldsymbol{\Sigma}_k^{-1}}{2}\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)\left((\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - 1\right) &= 0 \\
\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} &= \sum_{m=1}^{N}\gamma_k(\mathbf{x}_m) \\
\frac{\sum_{n=1}^{N}\gamma_k(\mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{m=1}^{N}\gamma_k(\mathbf{x}_m)} &= \boldsymbol{\Sigma}_k \ .
\end{aligned}
$$

Minimum cut and maximum flow (7 Points)

Exercise 3 (**Flow**, 4 Points)**.** Show that the following two definitions are equivalent.

a) Let $(\mathcal{V}, \mathcal{E}, c, s, t)$ be a flow network with non-negative edge weights. A function $f : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is called a flow if it satisfies the following properties:

i) Capacity constraint: $f(i,j) \leq c(i,j)$ for all $i,j \in \mathcal{V}$.

ii) Skew-symmetry: $f(i,j) = -f(j,i)$ for all $i,j \in \mathcal{V}$.

iii) Flow conservation: $\sum_{j \in \mathcal{V}} f(i,j) = 0$ for all $i \in \mathcal{V} \setminus \{s,t\}$.

b) Let $(\mathcal{V}, \mathcal{E}, c, s, t)$ be a flow network with non-negative edge weights. A function $f : \mathcal{E} \to \mathbb{R}^+$ is called a flow if it satisfies the following two properties:

i) $f(i,j) \leq c(i,j)$ for all $(i,j) \in \mathcal{E}$.

ii) For all $i \in \mathcal{V} \setminus \{s,t\}$

$$\sum_{(i,j)\in\mathcal{E}} f(i,j) = \sum_{(j,i)\in\mathcal{E}} f(j,i) \ .$$

Solution. $a) \Rightarrow b)$ Suppose we are given a flow f satisfying the definition given in a). Let us introduce $f' : \mathcal{E} \to \mathbb{R}^+$ such that $f'(i,j) := \max(0, f(i,j))$ for all $(i,j) \in \mathcal{E}$.

i) $0 \leq f'(i,j) = \max(0, f(i,j)) \leq \max(0, c(i,j)) = c(i,j)$ for all $(i,j) \in \mathcal{E}$.

ii) For all $i \in \mathcal{V} \setminus \{s,t\}$, we have

$$\begin{aligned}
\sum_{(i,j)\in\mathcal{E}} f'(i,j) &= \sum_{(i,j)\in\mathcal{E},\, f(i,j)\geq 0} f(i,j) \\
&= \sum_{(i,j)\in\mathcal{E},\, f(j,i)\leq 0} f(i,j) \\
&= \sum_{(j,i)\in\mathcal{E},\, f(j,i)\geq 0} f(j,i) \\
&= \sum_{(j,i)\in\mathcal{E}} f'(j,i) \ .
\end{aligned}$$

$b) \Rightarrow a)$ Suppose we are given a flow f satisfying the definition given in b). Let us introduce $f' : \mathcal{E} \to \mathbb{R}$ such that

$$f'(i,j) = \begin{cases} f(i,j) & \text{if } (i,j) \in \mathcal{E} \\ -f(i,j) & \text{if } (j,i) \in \mathcal{E} \ . \end{cases}$$

i) $c(i,j) \geq f(i,j) \geq f'(i,j)$ for all $(i,j) \in \mathcal{E}$.

ii) By definition $f'(i,j) = -f'(j,i)$ for all $(i,j) \in \mathcal{E}$.

iii) For all $i \in \mathcal{V} \setminus \{s,t\}$, we have

$$\sum_{(i,j)\in\mathcal{E}} f'(i,j) = \sum_{(i,j)\in\mathcal{E}} f(i,j) = \sum_{(j,i)\in\mathcal{E}} f(j,i) = -\sum_{(i,j)\in\mathcal{E}} f'(i,j) \ ,$$

which completes the proof.

Exercise 4 (Flow, 3 Points). Let $G = (\mathcal{V}, \mathcal{E}, c, s, t)$ be a flow network, and let f be a flow in G. Show that the following equalities hold:

a) For all $X \subseteq \mathcal{V}$, we have $f(X, X) = 0$.

b) For all $X, Y \subseteq \mathcal{V}$, we have $f(X, Y) = -f(Y, X)$.

c) For all $X, Y, Z \subseteq \mathcal{V}$ with $X \cap Y = \varnothing$, we have the sums

$$f(X \cup Y, Z) = f(X, Z) + f(Y, Z) \quad \text{and} \quad f(Z, X \cup Y) = f(Z, X) + f(Z, Y) .$$

Solution. a) Assume that b) is already held. Then $f(X, X) = -f(X, X)$ for all $X \subseteq \mathcal{V}$, which means that $f(X, X) = 0$.

b) For all $X, Y \subseteq \mathcal{V}$, we have

$$f(X, Y) = \sum_{a \in X} \sum_{b \in Y} f(a, b) = \sum_{a \in X} \sum_{b \in Y} -f(b, a) = -\sum_{b \in Y} \sum_{a \in X} f(b, a) = -f(Y, X) .$$

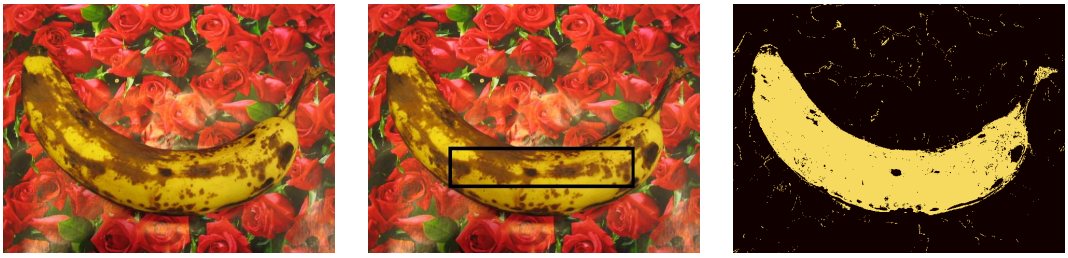
c) Suppose that $X, Y \subseteq \mathcal{V}$ and $X \cap Y = \emptyset$. Then for any $Z \subseteq \mathcal{V}$, we have

$$f(X \cup Y, Z) = \sum_{a \in X \cup Y} \sum_{b \in Z} f(a, b) = \sum_{a \in X} \sum_{b \in Z} f(a, b) + \sum_{a \in Y} \sum_{b \in Z} f(a, b) = f(X, Z) + f(Y, Z) .$$

One can prove similarly the second equality, too.

Programming (6 Points)

Exercise 5 (**Gaussian Mixture Model Estimation**, 6 Points)**.** Estimate two mixtures of Gaussians to model the densities of the foreground and background pixels based on their intensity. Apply the estimated models to the input image and segment the foreground and background regions. An exemplar test image is shown in Figure 1 (a) (you can find the image in the supplementary material `in2329-exercise_05_supp.zip`). The specific requirement is as follows.



(a) (b) (c)

Figure 1: (a) test image. (b) an exemplar bounding box. (c) foreground segmentation based the trained model $p_F(I)$.

1. Write a program taking an *image* and a *bounding box* of the foreground region, given by the coordinates of the top-left and bottom-right corners, as input argument. An example is shown Figure 1 (b). Estimate the density of the foreground $f_{\text{fg}}(I), I \in \mathbb{R}^3$ using all pixels inside the bounding box, where $f_{\text{fg}}(I)$ is a mixture of Gaussians with $K = 5$ components and I is the RGB values of a pixel.

2. Based on the same bounding box, estimate the density of the background $p_{\text{bg}}(I)$ using all pixels outside the bounding box, where $f_{\text{bg}}(I)$ is a mixture of Gaussians with $K = 5$ components.

3. Compute the binary segmentation of the input image by making use of the two estimated densities. Check the results.

Hints: the input bounding box should mostly contain banana, i.e. the foreground region. You may initialize the mixture of Gaussians with random Gaussian kernels. Note that the covariance matrix should not be singular, during the estimation. If the covariance matrix does become singular, you may restart the estimation from a different initialization all over again. Alternatively, can you think of way to better initialize the optimization?