Hw 1

1 a) WE HAVE TO SHOW THAT $\underbrace{\sum_{x \in C_j} d(x : m_J)^2}_{*} = \frac{1}{2} \frac{1}{|C_J|} \sum_{x \in C_J} \sum_{x' C_J} d(x : x')^2$

NOW: $\frac{1}{|C_J|} \sum_{x \in C_J} \sum_{x' \in C_J} \underbrace{\|x\|^2 + \|x'\|^2}_{\text{SAME TERM GIVEN THE DOUBLE SUM}} - 2|x \cdot x'| = \frac{2}{|C_J|} \sum_{x \in C_J} \sum_{x' \in C_J} 2|x|^2 - 2\, x \cdot x' \quad (1)$

ON THE OTHER HAND

$$\sum_{x \in C_j} d(x ; m)^2 = \sum_{x \in C_J} \left(|x|^2 + \left(\frac{1}{|C_J|} \sum_{x' \in C_J} x' \right)^2 - 2\, x \cdot \frac{1}{|C_J|} \sum_{x' \in C_J} x' \right) =$$

$$= \sum_{x \in C_J} \left(|x|^2 + \frac{1}{|C_J|^2} \sum_{x' \in C_J} \sum_{x'' \in C_J} x' \cdot x'' - 2 x \cdot \frac{1}{|C_J|} \sum_{x' \in C_J} x' \right)$$

BROKE SUM $\sum_i (a_i + b_i) = \sum a_i + \sum b_i$

$$= \left(\sum_{x \in C_J} |x|^2 \right) + \frac{C_J}{|C_J|^2} \sum_{x' \in C_J} \sum_{x'' \in C_J} x' \cdot x'' - \left(\frac{2}{|C_J|} \sum_{x \in C_J} \sum_{x' \in C_J} x \cdot x' \right)$$

$\frac{\sum^{C_J} A}{C_J} = A$

IF A CONST

$$= \sum_{x \in J} |x|^2 - \frac{1}{C_J} \sum_{x \in C_J} \sum_{x' \in C_J} x' \cdot x''$$

$$= \sum_{x \in C_J} \sum_{x' \in C_J} \frac{1}{C_J} \left(|x|^2 - x' \cdot x \right) \quad (2)$$

COMPARING (1) AND (2) $\rightarrow$ $J_{IC} = 2\, J_{avg^2}$

1 (b) $J_{AVG^2} = \sum_{i=1}^{m} d(x_i\, m_{\gamma_i})^2$

- BY DEFINITION WHEN $(m_1 \cdots m_K)$ IS FIXED $\gamma_i \leftarrow \underset{j\in\{1,\cdots k\}}{\arg\min}\, d(x_i; m_j)$

MINIMIZE $\sum d(x_i; m_{\gamma_i})$; BUT SINCE $(\ )^2$ IS MONOTONE IT ALSO

MINIMIZE $\sum d(x; m_{\gamma_i})^2$

- TO MINIMIZE $\sum_{i=1}^{m} d(x_i; m_{\gamma_i})$ WITH γ_i FIXED $\frac{\partial}{\partial m} \sum_{i=1}^{m} |x_i|^2 + |m_{\gamma_i}|^2 - 2\, x_i m_{\gamma_i} =$

$\sum_{i=1}^{m} 2|m_{\gamma_i}| - 2x_i = 0 \longrightarrow \sum_{i:\gamma_i=j} x_i = |C_j|\, m_j$ WHICH CORRESPONDS

TO THE UPDATING RULE; NOTE THAT THIS IS BASICALLY THE DEFINITION OF THE MEAN (.

c) THE K MEANS IS BASED ON AN ITERATIVE STEP WITH THE UPDATE RULES IN EXERCISE 1b; SINCE BOTH MINIMIZE J KEEPING A SET OF ITS ARGUMENT FIXED THE FULL STEP MINIMIZE $J_{AVG^2}(\gamma_i; m_i)$

d) THE WORST CASE SCEARIO WOULD BE WHEN THE ALGORITH HAS TO TRY ALL THE POSSIBLE ASSIGNMENTS OF EACH DATA TO A CLUSTER BEFORE IT FINDS THE RIGHT ONE

SO WE HAVE $O(K^m)$; THE ALGORITH WON'T VISIT THE SAME COMBINATION TWICE BECAUSE IT IS DESIGNED TO MINIMIZE THE COST FUNCTION MONOTONICALLY; THERE ARE STRONGER LIMITS IN THE LITERATURE:

× $O(m^{kd})$ in d dimension FROM VORONOI TASSELLATION CONSIDERATION INABA et AL

× $O(m\Delta^2)$ WITH Δ SPREAD OF POINT $\Delta = \left(\frac{\text{LARGEST PAIRWISE DISTANCE}}{\text{SMALLER PAIRWISE DISTANCE}}\right)$

3) AS A FUNCTION OF THE PARAMETER THE LIKELIHOOD CAN BE SEEN AS $P(X|\theta)$; $P(X|\theta) = \prod_{i=1}^{m} \pi_{z_i} N(x_i | \mu_i ; z_i)$;

SO $\log P(X|\theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_i | \mu_k ; \Sigma_{ik}) \right\}$;

THIS IS THE LIKELIHOOD FOR THE COMPLETE DATA SET $P(X|\pi, \mu\Sigma)$ WHERE X ALSO CONTAINS THE LATENT VARIABLE z $X = (x; \bar{z})$

b) THIS CAN BE OBTAINED USING THE BAYES THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(A)}.$$

WE WANT $P(z_{ij} = 1 | x_i) = \dfrac{P(x_i | z_{ij} = 1) P(z_{ij} = 1)}{P(x_i)}$;

NOW, $P(z_k) = \pi_k$ AND $P(x_i | z_{ij} = 1)$ WAS DEFINED IN 3a

$$P(z_{ij} = 1 | x_i) := P_{i,j} = \frac{\pi_j N(x_i ; \mu_j ; \Sigma_j)}{\sum_{j'} \pi_{j'} N(x_i ; \mu_{j'} ; \Sigma_{j'})}$$

c) NOW GIVEN THE LATENT VARIABLE z WE HAVE

$P(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ AND $P(X|z) = \prod_{k=1}^{K} N(x ; \mu_k ; \Sigma_k)^{z_k}$

SO $\ln P(x; z | \mu ; \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \ln (x_n | \mu_k ; \Sigma_k) \right\}$

WITH z_{nk} k COMPONENT OF $\bar{z}$; NOW WITH BAYES THEOREM

$P(z | X; \mu ; \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\pi_k N(x_n | \mu_k \Sigma_k) \right]^{z_{nk}}$ SO EASILY $E[z_{nk}] = P_{nk}$

THIS LEADS TO $E_z[\ln P(x; z | \mu \Sigma \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} P_{nk} \left\{ \ln \pi_k + \ln N(x ; \mu ; \Sigma) \right\}$

3 d) WE ARE LOOKING FOR $\frac{\partial}{\partial \pi_J} \bar{\ell}_{OLD}(\theta) = 0$; TO IMPOSE THE CONSTRAINT $\sum_{J=1}^{K} \pi_J = 1$ WE CAN USE THE LAGRANGE MULTIPLIERS

$$\sum_{i}^{m} \sum_{J}^{K} P_{iJ} (\log \pi_J) + \lambda \left(\sum_{J} \pi_J - 1 \right)$$

$$\downarrow \partial / \partial \pi_J$$

$$\sum_{i=1}^{m} \frac{P_{iJ}}{\pi_J} + \lambda = 0 \rightarrow \pi_J = - \frac{\sum_i P_{iJ}}{\lambda} \quad ①$$

$\frac{\partial}{\partial \lambda} = 0$

$$\rightarrow \sum_{j} \pi_j = 1 \rightarrow - \sum_{J} \sum_{i} \frac{P_{i,J}}{\lambda} \rightarrow \lambda = -m \quad ②$$

$$\left[\pi_J = 1/m \sum_i P_{i;J} \right] \Delta \text{ COMBINING } ① \ ②$$

3e) $\boxed{\mu}$ WE WANT $\frac{\partial}{\partial \mu} = 0$ OF THE LOG LIKELIHOOD

$$\frac{\partial}{\partial \mu_k} \sum_{i}^{N} \sum_{K}^{K} P_{Jk} \log N(x_i | \mu_k \Sigma_k) = 0$$ SINCE π_K AND P_{iJ} DOES NOT DEPEND ON μ.

$$\sum_{i}^{N} \sum_{K} P_{ik} \cdot \frac{1}{2} \frac{\partial}{\partial \mu_k} \left[(\bar{x} - \mu_k)^T \Sigma_k^{-1} (\bar{x} - \mu_k) \right] = \sum_{i}^{N} P_{ik} - \frac{1}{2} \left(-2 \Sigma_i^{-1} (\bar{x} - \mu) \right)$$

$$= \sum_{i}^{N} P_{ik} \Sigma_i^{-1} (\bar{x} - \mu_k)$$

USE
$\partial(\bar{A} \cdot \bar{B})$
$= \partial(\bar{A}) \bar{B} +$
$\bar{A} \, \partial(\bar{B})$

NOW $\frac{\partial}{\partial \mu} (\ ) = 0 \quad \mu_k = \frac{\sum_{i=1}^{m} P_{ik} x_i}{\sum_{i=1}^{m} P_{i;J}}$

3 e) $\boxed{\Sigma_i}$ $\frac{\partial}{\partial \Sigma_i}() = \sum_i^m P_{ik} \frac{\partial}{\partial \Sigma_{ik}} \left\{ -\frac{1}{2} \ln(|\Sigma|) - \frac{1}{2} (x_m - \mu)^T \Sigma_k (x_m - \mu) \right\}$

NOW $\frac{\partial}{\partial \Sigma_i} \ln(|\Sigma_i|) = \left(\Sigma_i^{-1}\right)^T$

AND $\frac{\partial}{\partial \Sigma_i} (x-\mu)^T \Sigma_i^{-1} (x-\mu) = \left(\Sigma_i^{-1} (x_i - \mu)(x_i - \mu)^T \Sigma_i^{-1} \right)^T$

THE LATTER CAN BE SEEN USING $Tr[AB] = Tr[BA]$

$\sum_{m=1}^{N} (x-\mu)^T \Sigma^{-1} (x-\mu) =$

SO $\frac{\partial}{\partial \Sigma_i}() \propto \sum_{i=1}^{m} P_{ik} \left\{ \left(\Sigma_{in}^{-1}\right)^T - \left(\Sigma_n^{-1} (x_i - \mu_n)(x_i - \mu_n)^T \Sigma_n^{-1} \right)^T \right\}$

$\propto \left(\Sigma_n^{-1} \right)^T \sum_{i=1}^{m} P_{ij} \left\{ 1 - \left(\Sigma_n^{-1} (x_i - \mu_n)(x_i - \mu_n)^T \right)^T \right\}$

WHERE WE NEGLECT CONSTANT BECAUSE WE ARE INTERESTED IN $\frac{\partial}{\partial \Sigma}() = 0$

NOW $\frac{\partial}{\partial \Sigma_i}() = 0$ $\sum_i P_{is} = \Sigma^{-1} \sum_i P_{ik} (x_i - \mu_n)(x_i - \mu_n)^T$

$\Downarrow$

$$\boxed{\Sigma_n = \frac{\sum P_{in} (x_i - \mu_n)(x_i - \mu_n)^T}{\sum P_{in}}}$$

3(f) GAUSSIAN MIXTURE (GM | K-MEANS

GAUSSIAN MIXTURE (GM)

- START WITH GUESS FOR π_{iJ} μ_J σ_J

- ASSIGN A WEIGH TO EACH POINT TO BELONG TO SAME CLUSTER WITH $P_{iJ} = \pi_J N(x_i \mid \mu_J \sigma_J)$

- NOW MAXIMIZE GIVEN THE PARAMETERS P_{iJ};

$$\pi_J \leftarrow \frac{1}{m} \sum P_{iJ} \; ; \; \mu_J = \frac{\sum_i^m P_{iJ} x_J}{\sum_i^m P_{iJ}}$$

$$\Sigma_i \leftarrow \frac{\sum_{i=1}^{M} P_{i;J} (x_i - \mu_J)(x_i - \mu_J)^T}{\sum_{i=1}^{m} P_{i;J}}$$

K-MEANS

- INITIALIZE RANDOMLY THE K CENTROID μ SAMPLE OR USING K++

- ASSIGN EACH POINT TO A CLUSTER BY MINIMIZING $\sum_i^N d(x_i m_{\gamma_i})^2$ WITH $\bar{m}_i$ FIXED

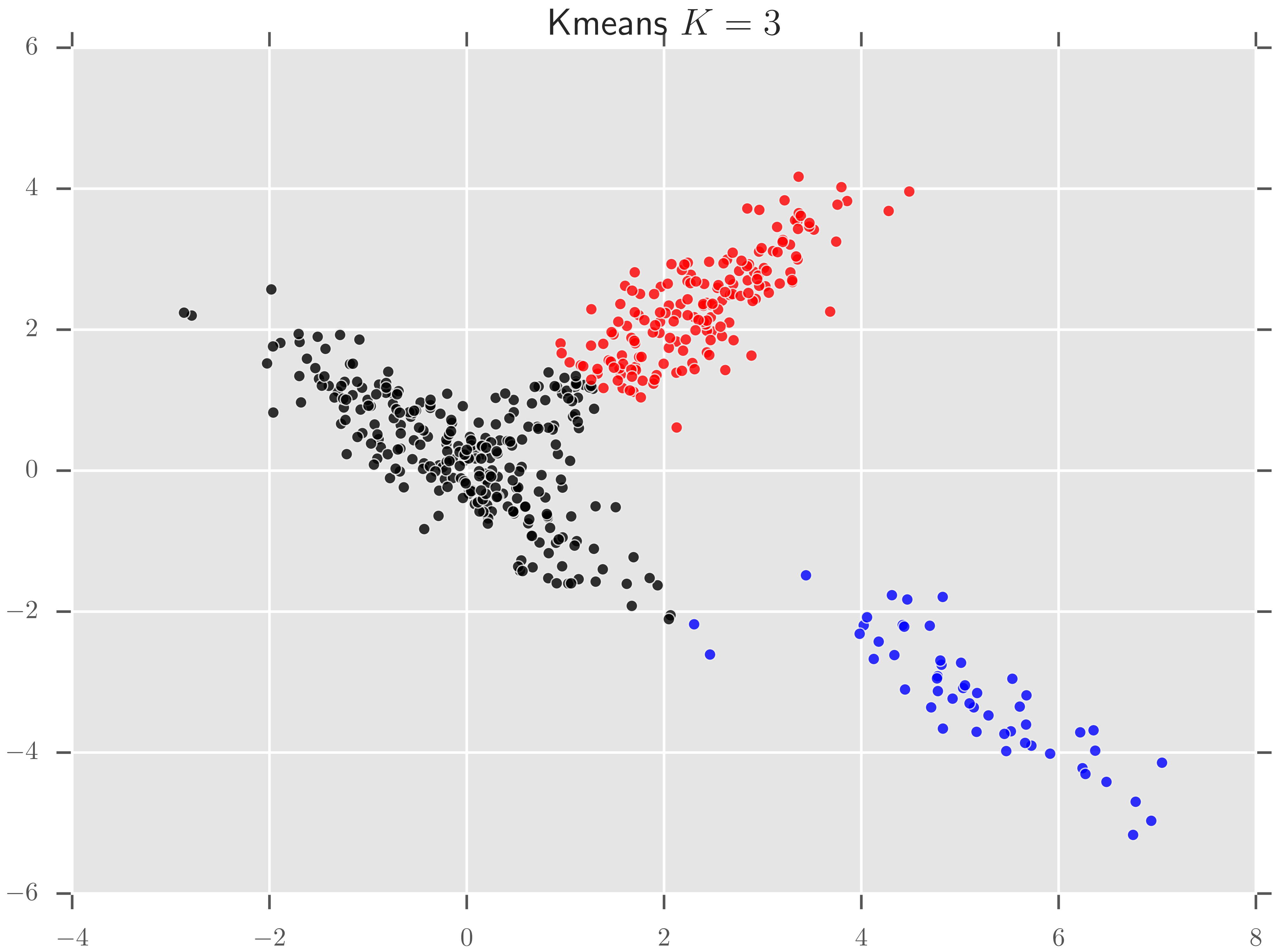
  ↳ THIS IS A ONE VS ALL ASSIGNMENT; IN GM YOU ASSIGN EACH POINT TO ALL OF THE GAUSSIAN WITH DIFFERENT WEIGHTS

- MAXIMIZE GIVEN THE ASSIGNMENT γ_i

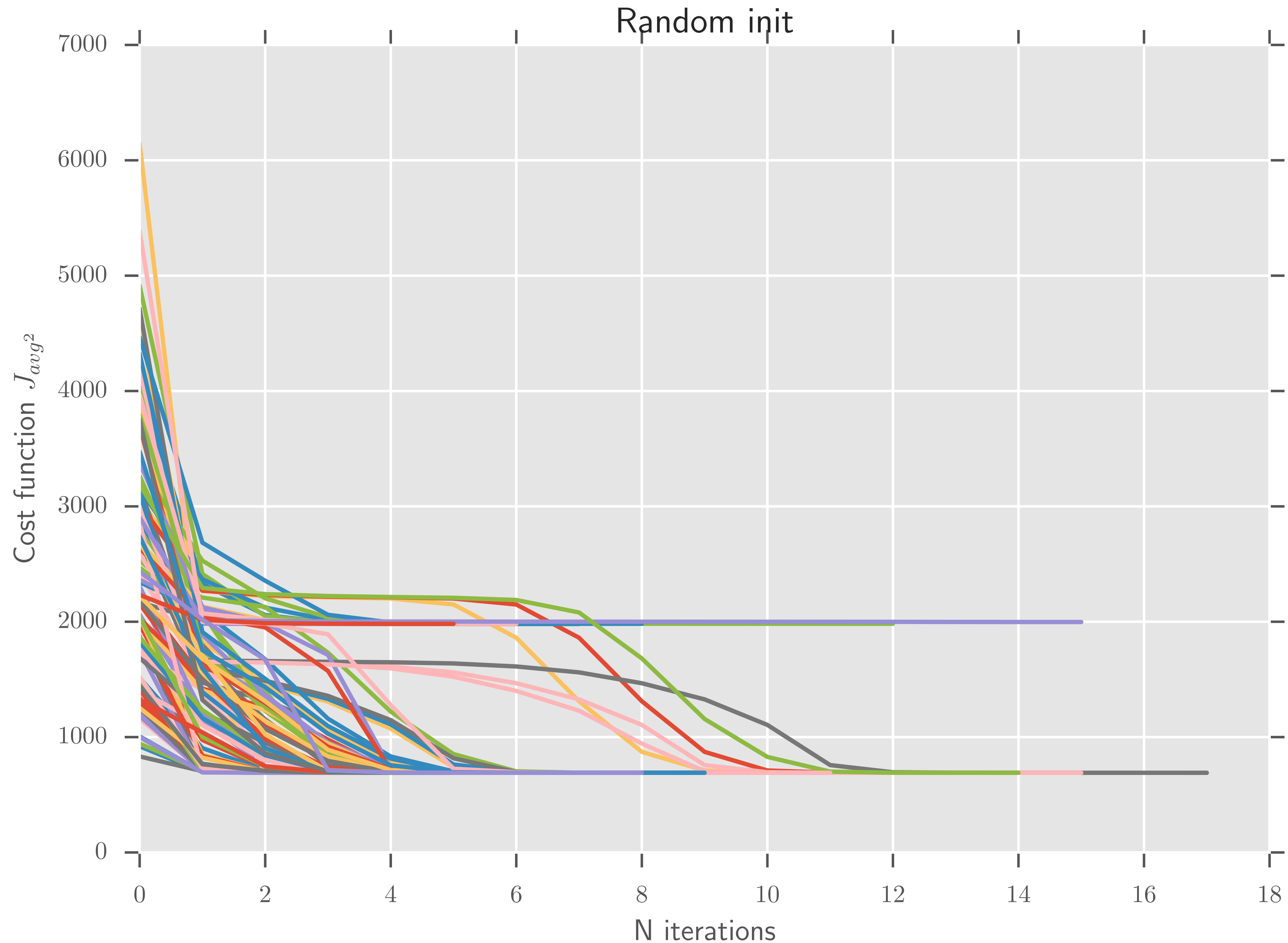
$$m_J = \frac{1}{|C_J|} \sum_{i : \gamma_i = J} x_i$$

THEY ARE SIMILAR BUT 1) IN GM YOU HAVE A MODEL; SO YOU HAVE TO MAXIMIZE 3 PARAMS π, σ, μ INSTEAD OF THE ONLY m_γ YOU HAVE IN K MEANS

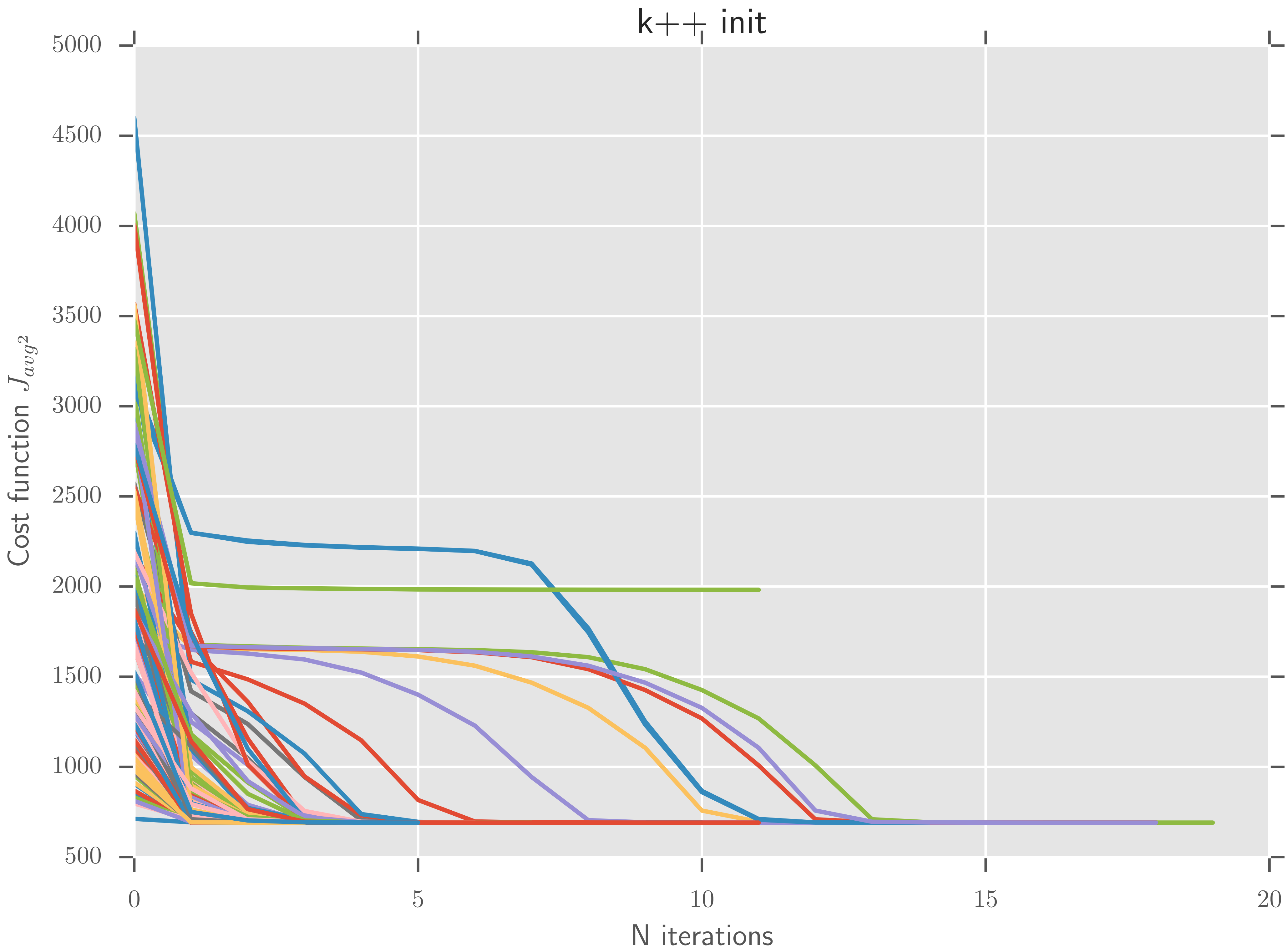
2) IN GM ALL THE POINTS BELONG TO ALL THE GAUSSIAN INSTEAD OF THE BINARY ASSIGNATION YOU HAVE IN K MENS



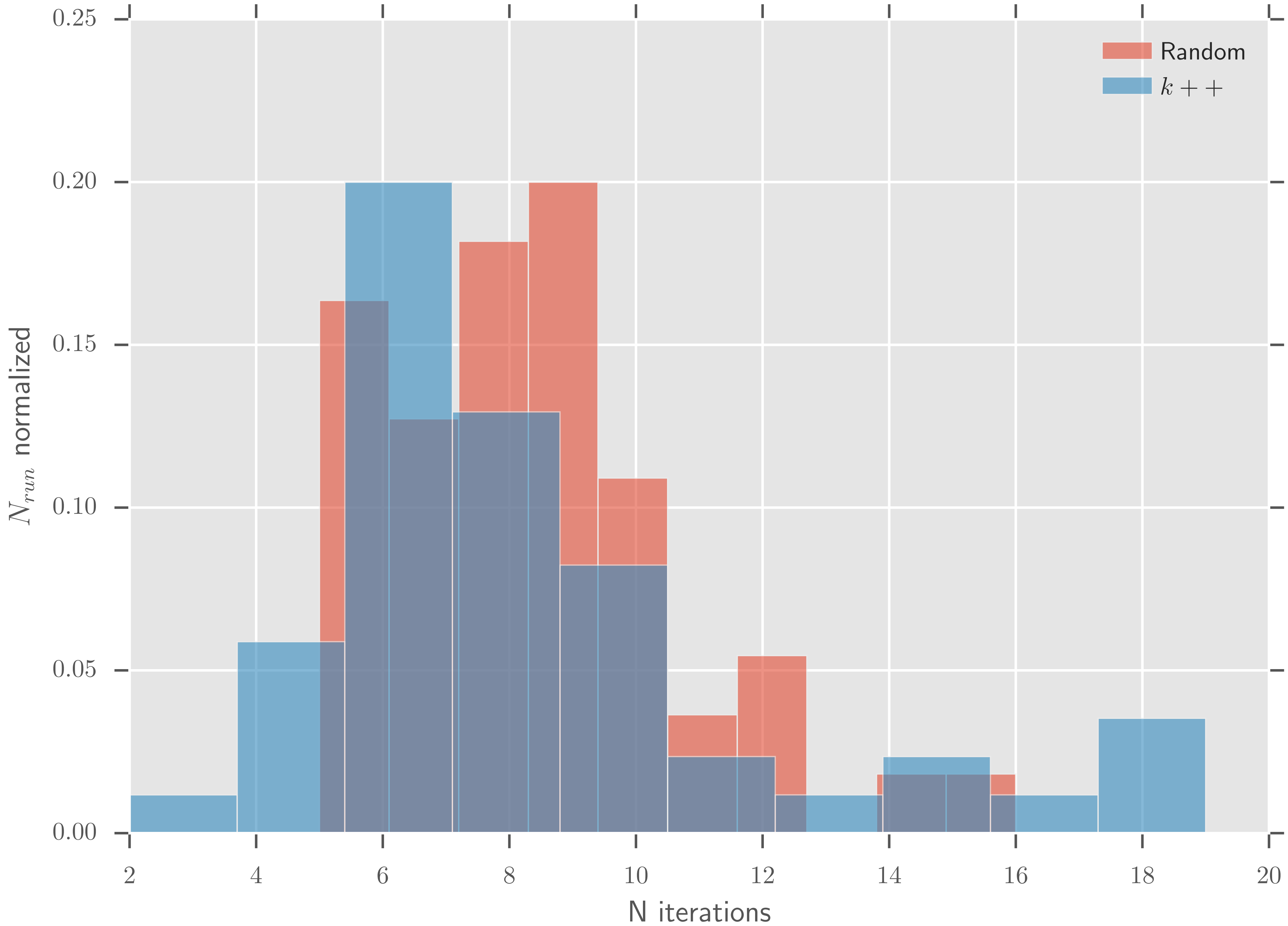
Results of the K-means algorithm. Each color represents a different cluster



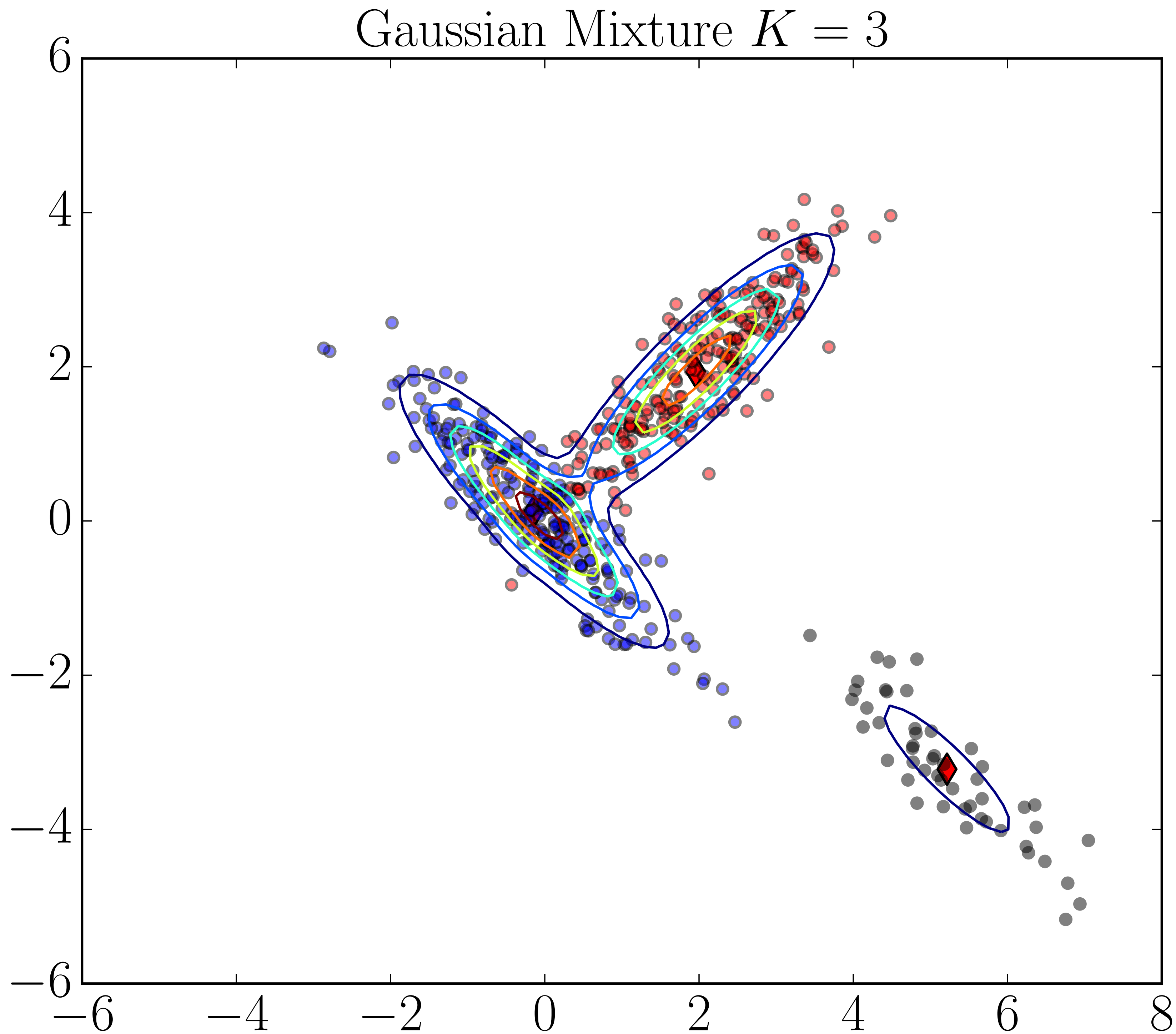
Cost function as a function of the iteration for the Kmeans algorithm (100 runs). As expected the function is monotonically decreasing at each step and most of the curves converge on the right minima of J~680. Some of the curves does not converge to the minima because they are stuck in local minima. This is why one should always run the K-mean algorithm multiple time and pick the configuration that minimize J among them



Cost function as a function of the iteration for the Kmeans algorithm (100 runs) with a k++inizialization. The same comments of the previous figures apply here. Note that it can be see by eye that the cost function start at lower values because of the clever initialization and that the curves converge a little bit faster than in the random init case. See next figure on this.



Even if the number of runs of the algorithm is not enough to have a statistical sample it can be see with K++ it converges a little bit faster than with a random init. Indeed the distribution is peaked a little bit on the left. Also the problem here is not hard enough to see the full power of the k++ initialization



Results of the gaussian mixture algorithm. The red diamonds correspond to the mean of the three gaussian and the contour plot allows to have an intuition of the different weights and the direction of the covariance matrices. The weights are 0.44321029, 0.10000401, 0.4567857