Machine Learning for Data Science (CS 4786)

Lecture 16-17: EM Algorithm: Why EM works!, EM for Gaussian Mixture Models and Mixture of Mult

The text in black outlines high level ideas. The text in blue provides simple mathematical details to "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested.

1 EM Algorithm

Iteratively we repeat E-step (expectation step) and M-step MAximization step starting with $\theta^{(0)}$ a random initialization for parameter θ. The following are the E and M steps.

1.1 E-step

On iteration i, for each data point $t \in [n]$, set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)}) \propto p(x_t|c_t\theta^{(i-1)}) \times P(c_t|\theta^{(i-1)})$$

1.2 M-step for GMM

For the M-step (for MLE) we would like to find

$$\theta = \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q_t^{(i)}(c_t) \log P(x_t, c_t|\theta)$$

For MAP we add in a penalty of $+\log P(\theta)$ to the above maximization problem.

2 EM Algorithm: Why it works?

Log likelihood only decreases after one iteration of EM algorithm. Why?

We will show below that EM algorithm can never lead to a worsening of the objective in any step and can only imrpvie likelihood.

$$\begin{aligned}
\log P(x_1, \ldots, x_n|\theta^{(i+1)}) &= \sum_{t=1}^{n} \log P(x_t|\theta^{(i+1)}) && \text{(x's drawn independently)} \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} P(x_t, c_t|\theta^{(i+1)}) \right) && \text{(marginalizing over } c_t\text{'s)} \\
&= \sum_{t=1}^{n} \log \left(\sum_{c_t=1}^{K} \frac{Q^{(i+1)}(c_t)}{Q^{(i+1)}(c_t)} P(x_t, c_t|\theta^{(i+1)}) \right)
\end{aligned}$$

Logarithm is a concave function and by Jensen's inequality $\log(E[X]) \geq E[\log(X)]$ for any R.V. X. Treat the term in red as the random variable and the probability distribution is specified by $Q^{(i+1)}$, now using Jensen,

$$\begin{aligned}
&\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(\frac{P(x_t, c_t|\theta^{(i+1)})}{Q^{(i+1)}(c_t)} \right) \\
&= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i+1)}) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right)
\end{aligned}$$

Since in M-step $\theta^{(i+1)}$ is exactly the maximizer of $\sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i+1)}) \right)$, we conclude that this term is larger than $\sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i)}) \right)$ and so

$$\geq \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t, c_t|\theta^{(i)}) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right)$$

Now note that $P(x_t, c_t|\theta^{(i)}) = P(c_t|x_t, \theta^{(i)}) P(x_t|\theta^{(i)}) = Q^{(i+1)}(c_t) P(x_t|\theta^{(i)})$ and so,

$$\begin{aligned}
&= \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(P(x_t|\theta^{(i)}) \times Q^{(i+1)}(c_t) \right) - \sum_{t=1}^{n} \sum_{c_t=1}^{K} Q^{(i+1)}(c_t) \log \left(Q^{(i+1)}(c_t) \right) \\
&= \sum_{t=1}^{n} \log P(x_t|\theta^{(i)}) \\
&= \log P(x_1, \ldots, x_n|\theta^{(i)})
\end{aligned}$$

Hence we have shown that running the EM algorithm yields, $\log P(x_1, \ldots, x_n|\theta^{(i)}) \leq \log P(x_1, \ldots, x_n|\theta^{(i+1)})$, that is the Likelihood value never decreases and could only improve.

3 Gaussian Mixture Models

Each $\theta \in \Theta$ consist of mixture distribution π which is a distribution over the choices of the K clusters, $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ the choices of the K means for the corresponding gaussians and $\Sigma_1, \ldots, \Sigma_K$ the choices of the K covariance matrices. The latent variables are $c_1, \ldots, c_n$ the cluster assignments for the n points and $x_1, \ldots, x_n$ are the n observations.

3.1 E-step

On iteration i, for each data point $t \in [n]$, set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned}
Q_t^{(i)}(c_t) &= P(c_t|x_t, \theta^{(i-1)}) \\
&\propto p(x_t|c_t \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\
&\propto \frac{1}{\sqrt{(2\pi)^d |\Sigma_{c_t}|}} \exp \left(-(x_t - \mu_{c_t})^\top \Sigma_{c_t} (x_t - \mu_{c_t})/2 \right) \pi_{c_t}
\end{aligned}$$

3.2 M-step for GMM

For the M-step (for MLE) we would like to find

$$\theta = \operatorname*{argmax}_{\theta\in\Theta} \sum_{t=1}^{n}\sum_{c_t=1}^{K} Q_t^{(i)}(c_t)\log P(x_t, c_t|\theta)$$

To this end note that

$$\begin{aligned}\sum_{t=1}^{n}\sum_{c_t=1}^{K} Q_t^{(i)}(c_t)\log P(x_t, c_t|\theta) &= \sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\left(\log\phi(x_t|\mu_k,\Sigma_k)+\log\pi_k\right)\\ &= \sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\left(\frac{1}{2}\log\left(\frac{1}{(2*3.14)^d|\Sigma_k|}\right)-\frac{1}{2}(x_t-\mu_k)^\top\Sigma_k^{-1}(x_t-\mu_k)+\log\pi_k\right)\\ &= \sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\left(-\frac{1}{2}\log\det(\Sigma_k)-\frac{1}{2}(x_t-\mu_k)^\top\Sigma_k^{-1}(x_t-\mu_k)+\log\pi_k\right)+\text{constant terms}\end{aligned}$$

For notational convenience define:

$$L(\mu_{1:K},\Sigma_{1:K},\pi) = \sum_{t=1}^{n}\sum_{k=1}^{K} Q_t^{(i)}(k)\left(-\frac{1}{2}\log\det(\Sigma_k)-\frac{1}{2}(x_t-\mu_k)^\top\Sigma_k^{-1}(x_t-\mu_k)+\log\pi_k\right)$$

Our goal is to find parameters that maximize $L(\mu_{1:K},\Sigma_{1:K},\pi)$.

M-step for mean: To optimize with respect to mean we take derivative and equate to 0,

$$\begin{aligned}\frac{\partial}{\partial\mu_k}L(\mu_{1:K},\Sigma_{1:K},\pi) &= -\frac{1}{2}\frac{\partial}{\partial\mu_k}\left(\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t-\mu_k)^\top\Sigma_k^{-1}(x_t-\mu_k)\right)\\ &= -\sum_{t=1}^{n} Q_t^{(i)}(k)\Sigma_k^{-1}(x_t-\mu_k) = -\Sigma_k^{-1}\left(\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t-\mu_k)\right)\end{aligned}$$

To maximize over μ_k we set derivative equal to 0. Hence

$$\sum_{t=1}^{n} Q_t^{(i)}(k)(x_t-\mu_k) = \sum_{t=1}^{n} Q_t^{(i)}(k)x_t - \mu_k\left(\sum_{t=1}^{n} Q_t^{(i)}(k)\right) = 0$$

Or equivallently:

$$\mu_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)x_t}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$

M-step for mixture distribution: Since we want to optimize over π subject to the constraint $\sum_{k=1}^{K}\pi_k = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we want to optimize the following term w.r.t. π_k and λ

$$L(\mu_{1:K},\Sigma_{1:K},\pi) + \lambda(1-\sum_{k=1}^{K}\pi_k)$$

Hence taking derivative of above w.r.t. π we get,

$$\frac{\partial}{\partial \pi_k}\left(L(\mu_{1:K},\Sigma_{1:K},\pi)+\lambda(1-\sum_{k=1}^{K}\pi_k)\right)=\frac{\partial}{\partial \pi_k}L(\mu_{1:K},\Sigma_{1:K},\pi)-\lambda$$

But,

$$\frac{\partial}{\partial \pi_k}L(\mu_{1:K},\Sigma_{1:K},\pi)=\frac{\partial}{\partial \pi_k}\sum_{t=1}^{n}Q_t^{(i)}(k)\log(\pi_k)=\frac{\sum_{t=1}^{n}Q_t^{(i)}(k)}{\pi_k}$$

Hence,

$$\frac{\partial}{\partial \pi_k}\left(L(\mu_{1:K},\Sigma_{1:K},\pi)+\lambda(1-\sum_{k=1}^{K}\pi_k)+\sum_{i=1}^{K}\lambda_i\pi_i\right)=\frac{\sum_{t=1}^{n}Q_t^{(i)}(k)}{\pi_k}-\lambda$$

Setting derivative to 0 we discover that

$$\pi_k\propto\sum_{t=1}^{n}Q_t^{(i)}(k)$$

Since π needs to be a valid distribution, this yields that

$$\pi_k=\frac{\sum_{t=1}^{n}Q_t^{(i)}(k)}{\sum_{k=1}^{K}\sum_{t=1}^{n}Q_t^{(i)}(k)}$$

However notice that since $Q_t^{(i)}$ is a distribution over K clusters, $\sum_{k=1}^{K}\sum_{t=1}^{n}Q_t^{(i)}(k)=\sum_{t=1}^{n}1=n$. Hence,

$$\pi_k=\frac{\sum_{t=1}^{n}Q_t^{(i)}(k)}{n}$$

M-step for Covariance: This one needs being able to take derivative w.r.t. matrices and so I will only sketch the proof here. Let us consider optimizing w.r.t. some Σ_k. It makes the problem easier if we instead think of the problem as optimizing over Σ_k^{-1} and then invert the solution.

Here are two facts that come in handy:

$$\frac{\partial}{\partial \mathbf{X}}\log\det(\mathbf{X}^{-1})=-\mathbf{X}^{-1}$$

and for any vector v,

$$\frac{\partial}{\partial \mathbf{X}}v^\top\mathbf{X}v=vv^\top$$

Now note that

$$\begin{aligned}\frac{\partial}{\partial \Sigma_k}L(\mu_{1:K},\Sigma_{1:K},\pi)&=\frac{\partial}{\partial \Sigma_k}\left(\sum_{t=1}^{n}Q_t^{(i)}(k)\left(-\frac{1}{2}\log\det(\Sigma_k)-\frac{1}{2}(x_t-\mu_k)^\top\Sigma_k^{-1}(x_t-\mu_k)\right)\right)\\&=\left(\sum_{t=1}^{n}Q_t^{(i)}(k)\left(\frac{1}{2}(\Sigma_k^{-1})^{-1}-\frac{1}{2}(x_t-\mu_k)(x_t-\mu_k)^\top\right)\right)\end{aligned}$$

Hence equating to 0 we get that

$$\Sigma_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k)(x_t - \mu_k)(x_t - \mu_k)^\top}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

that is the weighted sample variance. (**there is a bit of a fudge here since μ_k is also an optimiation variable. But we skip the details of this for now.**)

3.3 EM for Mixture Models

For any mixture model with π as mixture distribution, and any arbitrary parameterization of likelihood of data given cluster assignment, one can write down a more detailed form for EM algorithm.

E-step On iteration i, for each data point $t \in [n]$, set

$$Q_t^{(i)}(c_t) = P(c_t|x_t, \theta^{(i-1)})$$

Note that

$$\begin{aligned} Q_t^{(i)}(c_t) &= P(c_t|x_t, \theta^{(i-1)}) \\ &\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\ &\propto p(x_t|c_t, \theta^{(i-1)}) \times P(c_t|\theta^{(i-1)}) \\ &= \frac{p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]} \end{aligned}$$

So all we need to fill out the $n \times K$ sized Q matrix is to have a current guess at π and the ability to compute $p(x_t|c_t, \theta^{(i-1)})$ up to multiplicative factor.

$$\begin{aligned} \theta &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log P(x_t, c_t = k|\theta) \\ &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log P(x_t|c_t = k, \theta) \times P(c_t = k|\theta) \\ &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(P(x_t|c_t = k, \theta) \times \pi[k]\right) \\ &= \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(P(x_t|c_t = k, \theta)\right) + \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(\pi[k]\right) \end{aligned}$$

Using $\Theta^{\setminus \pi}$ to denote the set of parameters excluding π,

$$\begin{aligned} &= \operatorname*{argmax}_{\theta \in \Theta^{\setminus \pi}, \pi} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(P(x_t|c_t = k, \theta)\right) + \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(\pi_k\right) \right) \\ &= \left(\operatorname*{argmax}_{\theta \in \Theta^{\setminus \pi}} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(P(x_t|c_t = k, \theta)\right) \right), \operatorname*{argmax}_{\pi} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log \left(\pi_k\right) \right) \right) \end{aligned}$$

Notice that the term in red is exactly the optimization we solved for in GMM example. We know this already! The solution is:

$$\pi_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

and this is the same for any mixture model.

On the other hand, the optimization problem,

$$\underset{\theta \in \Theta \setminus \pi}{\text{argmax}} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \log \left(P(x_t | c_t = k, \theta) \right) \right)$$

is simply a weighted version of MLE when our observation includes c_t's the hidden or latent variables. In the M-step, this is the only portion that changes the mixture distribution solution has same form always.

4 Mixture of Multinomials

Each $\theta \in \Theta$ consist of mixture distribution π which is a distribution over the choices of the K clusters or types, $p_1, \ldots, p_K$ are K distributions over the d items. The latent variables are $c_1, \ldots, c_n$ the cluster assignments for the n points indicating that the t^{th} data point was drawn using distribution p_{c_t}. $x_1, \ldots, x_n$ are the n observations.

Story: You own a grocery store and multiple customers walk in to your store and buy stuff. You want group customers into K group based on distribution over the d products/choices in your store. Think of customers as being independently drawn and they each belong to one of K groups. We will first start with a simple scenario and build up to a more general one. To start with, say each day a customer walks in to your store and buys $m = 1$ product. The generative story then is that we first draw customer type $c_t \sim \pi$ from a mixture distribution π, next associated with type c_t, there is a distribution p_{c_t} over products the customer would buy. We draw $x_t \in [d]$ the product the customer bought as $x_t \sim p_{c_t}$. That is

$$p(x_t | c_t = k, \theta) = p_{c_t}[x_t]$$

Next we can move to a slightly more complex scenario where the customer on every round buys (fixed) $m > 1$ products by drawing x_t as m samples from the multinomial distribution. That is,

$$p(x_t | c_t = k, \theta) = \frac{m!}{x_t[1]! \cdot \ldots \cdot x_t[d]!} p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]}$$

where $x_t[j]$ indicates the amount of product j bought by the customer t.

4.1 Mixture of Multinomials (Primer $m = 1$)

E-step On iteration i, for each data point $t \in [n]$, set

$$\begin{aligned} Q_t^{(i)}(c_t) &= \frac{p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p(x_t|c_t, \theta^{(i-1)}) \cdot P(c_t|\theta^{(i-1)})} \\ &= \frac{p_{c_t}^{(i-1)}[x_t] \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p(x_t|c_t, \theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]} \end{aligned}$$

M-step As we already saw, we set

$$\pi_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

Now as for the remaining parameters, we want to maximize

$$\underset{p_1,\ldots,p_K}{\text{argmax}} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(p_k[x_t]\right) \right)$$

Define $L(p_1, \ldots, p_K) = \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(p_k[x_t]\right)$. We want to optimize $L(p_1, \ldots, p_K)$ w.r.t. $p_1, \ldots, p_k$ s.t. each p_k is a valid probability distribution over $\{1, \ldots, d\}$. As an example, to find the optimal p_k, we want to optimize over p_k subject to the constraint $\sum_{j=1}^d p_k[j] = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we find $p_k[j]$'s by taking derivative and equating to 0 the following Lagrangian objective,

$$L(p_1, \ldots, p_K) + \lambda_k(1 - \sum_{j=1}^d p_k[j])$$

Taking derivative and equating to 0, we want to find p_k s.t.,

$$\sum_{t=1}^n Q_t^{(i)}(k) \frac{1}{p_k[x_t]} - \lambda_k = 0$$

In other words, for every $j \in [d]$,

$$\sum_{t:x_t=j} Q_t^{(i)}(k) \frac{1}{p_k[j]} - \lambda_k = 0$$

Hence we conclude that

$$p_k[j] \propto \sum_{t:x_t=j} Q_t^{(i)}(k)$$

Hence,

$$p_k[j] = \frac{\sum_{t:x_t=j} Q_t^{(i)}(k)}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

Thus for the M-step when we are dealing with the mixture model with exactly $m = 1$ purchase on every round, we get that, for every $k \in [K]$ and every $j \in [d]$,

$$p_k[j] = \frac{\sum_{t:x_t=j} Q_t^{(i)}(k)}{\sum_{t=1}^n Q_t^{(i)}(k)}$$

4.2 Mixture of Multinomials ($m > 1$)

E-step On iteration i, for each data point $t \in [n]$, set

$$\begin{aligned} Q_t^{(i)}(c_t) &= \frac{p(x_t|c_t,\theta^{(i-1)}) \cdot \pi^{(i-1)}[c_t]}{\sum_{k=1}^K p(x_t|k,\theta^{(i-1)}) \cdot P(k|\theta^{(i-1)})} \\ &= \frac{p_{c_t}[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi^{(i-1)}[c_t]}{\sum_{c_t=1}^K p_{c_t}[1]^{x_t[1]} \cdot \ldots \cdot p_{c_t}[d]^{x_t[d]} \cdot \pi^{(i-1)}[k]} \end{aligned}$$

M-step For mixture distribution, as usual,

$$\pi_k = \frac{\sum_{t=1}^n Q_t^{(i)}(k)}{n}$$

Now as for the remaining parameters, we want to maximize

$$\begin{aligned} &\operatorname*{argmax}_{p_1,\ldots,p_K} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(P(x_t|c_t = k,\theta)\right) \right) \\ &= \operatorname*{argmax}_{p_1,\ldots,p_K} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \log\left(p_k[1]^{x_t[1]} \cdot \ldots \cdot p_k[d]^{x_t[d]}\right) \right) \\ &= \operatorname*{argmax}_{p_1,\ldots,p_K} \left(\sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \sum_{j=1}^d x_t[j] \log\left(p_k[j]\right) \right) \end{aligned}$$

Again to solve this, define $L(p_1,\ldots,p_K) = \sum_{t=1}^n \sum_{k=1}^K Q_t^{(i)}(k) \sum_{j=1}^d x_t[j] \log\left(p_k[j]\right)$. We want to optimize $L(p_1,\ldots,p_K)$ w.r.t. $p_1,\ldots,p_K$ s.t. each p_k is a valid probability distribution over $\{1,\ldots,d\}$. As an example, to find the optimal p_k, we want to optimize over p_k subject to the constraint $\sum_{j=1}^d p_k[j] = 1$ (ie. its a distribution), we do so by introducing Lagrange variables. That is we find $p_k[j]$'s by taking derivative and equating to 0 the following Lagrangian objective,

$$L(p_1,\ldots,p_K) + \lambda_k(1 - \sum_{j=1}^d p_k[j])$$

Taking derivative and equating to 0, we want to find p_k s.t.,

$$\sum_{t=1}^n Q_t^{(i)}(k) \sum_{j=1}^d x_t[j] \frac{1}{p_k[j]} - \lambda_k = 0$$

In other words, for every $j \in [d]$,

$$\sum_{t=1}^n Q_t^{(i)}(k) \frac{x_t[j]}{p_k[j]} - \lambda_k = 0$$

Hence we conclude that

$$p_k[j] \propto \sum_{t=1}^n Q_t^{(i)}(k) x_t[j]$$

Hence,

$$p_k[j] = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{\sum_{j=1}^{d} \sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{\sum_{t=1}^{n} Q_t^{(i)}(k) \left(\sum_{j=1}^{d} x_t[j] \right)} = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) x_t[j]}{m \sum_{t=1}^{n} Q_t^{(i)}(k)}$$