16 Mixture Models

Goal

Mixture models for density estimation and the celebrated expectation-maximization algorithm.

Alert 16.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 16.2: Density estimation

The central problem of this note is to estimate a density function (or more generally a probability measure), through a finite training sample. Formally, we are interested in estimating a probability measure χ from a (non)parametric family $\{\chi_\theta\}_{\theta\in\Theta}$. A typical approach is to minimize some statistical divergence between a noisy version $\hat{\chi}$ and χ_θ:

$$\inf_{\theta\in\Theta} \mathsf{D}(\hat{\chi}\|\chi_\theta).$$

However, the minimization problem above may not always be easy to solve, and alternative (indirect) strategies have been developed. As mentioned in Remark 4.20, choosing the KL divergence corresponds to the maximum likelihood estimation procedure.

Algorithm 16.3: Expectation-Maximization (EM) (Dempster et al. 1977)

We formulate EM under the density estimation formulation in Definition 16.2, except that we carry out the procedure in a lifted space $\mathsf{X}\times\mathsf{Z}$ where Z is the space that some latent random variable $\mathbf{Z}$ lives in. Importantly, we do not observe the latent variable $\mathbf{Z}$: it is "artificially" constructed to aid our job. We fit our model with a prescribed family of joint distributions

$$\mu_\theta(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z}) = \zeta_\theta(\mathrm{d}\mathbf{z})\mathfrak{D}_\theta(\mathrm{d}\mathbf{x}|\mathbf{z}) = \chi_\theta(\mathrm{d}\mathbf{x})\mathfrak{E}_\theta(\mathrm{d}\mathbf{z}|\mathbf{x}), \quad \theta\in\Theta.$$

In EM, we typically specify the joint distribution μ_θ explicitly, and in a way that the posterior distribution $\mathfrak{E}_\theta(\mathrm{d}\mathbf{z}|\mathbf{x})$ can be easily computed. Similarly, we "lift" $\chi(\mathrm{d}\mathbf{x})$ (our target of estimation) to the joint distribution

$$\hat{\nu}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z}) = \hat{\chi}(\mathrm{d}\mathbf{x})\mathcal{E}(\mathrm{d}\mathbf{z}|\mathbf{x}).$$

(We use the hat notation to remind that we do not really have access to the true distribution χ but a sample from it, represented by the empirical distribution $\hat{\chi}$.) Then, we minimize the discrepancy between the joint distributions $\hat{\nu}$ and μ_θ, which is an *upper bound* of the discrepancy of the marginals $\mathsf{KL}(\hat{\chi}\|\chi_\theta)$ (Exercise 4.17):

$$\inf_{\theta\in\Theta}\ \inf_{\mathcal{E}(\mathrm{d}\mathbf{z}|\mathbf{x})}\ \mathsf{KL}(\hat{\nu}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})\|\mu_\theta(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})).$$

Note that there is no restriction on $\mathcal{E}$ (and do not confuse it with $\mathfrak{E}_\theta$, which is "prescribed").

The EM algorithm proceeds with alternating minimization:

- (E-step) Fix θ_t, we solve $\mathcal{E}_{t+1}$ by (recall Exercise 4.17)

$$\inf_{\mathcal{E}}\ \mathsf{KL}(\hat{\nu}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})\|\mu_{\theta_t}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})) = \mathsf{KL}(\hat{\chi}\|\chi_{\theta_t}) + \mathsf{E}_{\hat{\chi}}\mathsf{KL}(\mathcal{E}\|\mathfrak{E}_{\theta_t}),$$

  which leads to the "closed-form" solution:

$$\mathcal{E}_{t+1} = \mathfrak{E}_{\theta_t}.$$

- (M-step) Fix $\mathcal{E}_{t+1}$, we solve θ_{t+1} by

$$\inf_{\theta\in\Theta} \mathsf{KL}(\hat{\nu}_{t+1}(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})\|\mu_\theta(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{z})) = \underbrace{\mathsf{KL}(\hat{\chi}\|\chi_\theta)}_{\text{likelihood}} + \underbrace{\mathsf{E}_{\hat{\chi}}\mathsf{KL}(\mathcal{E}_{t+1}\|\mathfrak{E}_\theta)}_{\text{regularizer}}.$$

  For the generalized EM algorithm, we need only decrease the above (joint) KL divergence if finding a (local) minima is expensive. It may be counter-intuitive that minimizing the sum of two terms above can be easier than minimizing the first likelihood term only!

Obviously, the EM algorithm monotonically decreases our (joint) KL divergence $\mathsf{KL}(\hat{\nu},\mu_\theta)$. Moreover, thanks to construction, the EM algorithm also ascends the likelihood:

$$\mathsf{KL}(\hat{\chi}\|\chi_{\theta_{t+1}}) \le \mathsf{KL}(\hat{\nu}_{t+1}\|\mu_{\theta_{t+1}}) \le \mathsf{KL}(\hat{\nu}_{t+1}\|\mu_{\theta_t}) = \mathsf{KL}(\hat{\chi}\|\chi_{\theta_t}).$$

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38.

Definition 16.4: Exponential family distribution

The exponential family distributions have the following density form:

$$\mathsf{p}(\mathbf{x}) = h(\mathbf{x})\exp\big(\langle\boldsymbol{\eta}, T(\mathbf{x})\rangle - A(\boldsymbol{\eta})\big),$$

where $T(\mathbf{x})$ is the sufficient statistics, $\boldsymbol{\eta}$ is the natural parameter, A is the log-partition function, and h represents the base measure. Since p integrates to 1, we have

$$A(\boldsymbol{\eta}) = \log\int \exp\big(\langle\boldsymbol{\eta}, T(\mathbf{x})\rangle\big)\cdot h(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

We verify that A is a convex function. (The cleanest way is perhaps through one of the rules in Exercise 3.13.)

Example 16.5: Gaussian distribution in exponential family

Recall from Example 4.7 that the multivariate Gaussian density is:

$$\begin{aligned}\mathsf{p}(\mathbf{x}) &= (2\pi)^{-d/2}[\det(\Sigma)]^{-1/2}\exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\\ &= \exp\left(\left\langle\begin{bmatrix}\mathbf{x}\\ -\frac{1}{2}\mathbf{x}\mathbf{x}^\top\end{bmatrix},\begin{bmatrix}\Sigma^{-1}\boldsymbol{\mu}\\ \Sigma^{-1}\end{bmatrix}\right\rangle - \tfrac{1}{2}(\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu} + d\log(2\pi) + \log\det\Sigma)\right).\end{aligned}$$

Thus, we identify

$$\begin{aligned}T(\mathbf{x}) &= (\mathbf{x}, -\tfrac{1}{2}\mathbf{x}\mathbf{x}^\top)\\ \boldsymbol{\eta} &= (\Sigma^{-1}\boldsymbol{\mu}, \Sigma^{-1}) =: (\boldsymbol{\xi}, S)\\ A(\boldsymbol{\mu},\Sigma) &= \tfrac{1}{2}(\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu} + d\log(2\pi) + \log\det\Sigma)\\ A(\boldsymbol{\eta}) &= A(\boldsymbol{\xi}, S) = \tfrac{1}{2}(\boldsymbol{\xi}^\top S^{-1}\boldsymbol{\xi} + d\log(2\pi) - \log\det S).\end{aligned}$$

Example 16.6: Bernoulli and Multinoulli in exponential family

The Bernoulli distribution is given as:

$$\mathsf{p}(z) = \pi^z(1-\pi)^{1-z} = \exp\left(z\log\pi + (1-z)\log(1-\pi)\right)$$

$$= \exp(z \log \tfrac{\pi}{1-\pi} + \log(1-\pi)).$$

Thus, we may identify

$$\begin{aligned} T(z) &= z \\ \eta &= \log \tfrac{\pi}{1-\pi} \\ A(\eta) &= \log(1+\exp(\eta)). \end{aligned}$$

We can also consider the multinoulli distribution:

$$\begin{aligned} \mathsf{p}(\mathbf{z}) &= \prod_{k=1}^{c} \pi_k^{z_k} = \exp\left(\langle \mathbf{z}, \log \boldsymbol{\pi} \rangle\right) \\ &= \exp\left(\tilde{\mathbf{z}}^\top \log \tfrac{\tilde{\boldsymbol{\pi}}}{1-\langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle} + \log(1 - \langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle)\right), \end{aligned}$$

where recall that $\mathbf{z} \in \{0,1\}^c$ is one-hot (i.e. $\mathbf{1}^\top \mathbf{z} = 1$), and we use the tilde notation to denote the subvector with the last entry removed. Thus, we may identify

$$\begin{aligned} T(\tilde{\mathbf{z}}) &= \tilde{\mathbf{z}} \\ \tilde{\boldsymbol{\eta}} &= \log \tfrac{\tilde{\boldsymbol{\pi}}}{1-\langle \mathbf{1}, \tilde{\boldsymbol{\pi}} \rangle} \\ A(\tilde{\boldsymbol{\eta}}) &= \log(1 + \langle \mathbf{1}, \exp(\tilde{\boldsymbol{\eta}}) \rangle). \end{aligned}$$

(Here, we use the tilde quantities to remove one redundancy since $\mathbf{1}^\top \mathbf{z} = \mathbf{1}^\top \boldsymbol{\pi} = 1$).

Exercise 16.7: Mean parameter and moments

Prove that for the exponential family distribution,

$$\begin{aligned} \nabla A(\boldsymbol{\eta}) &= \mathsf{E}[T(\mathbf{X})] \\ \nabla^2 A(\boldsymbol{\eta}) &= \mathsf{E}[T(\mathbf{X}) \cdot T(\mathbf{X})^\top] - \mathsf{E}[T(\mathbf{X})] \cdot \mathsf{E}[T(\mathbf{X})]^\top = \mathtt{Cov}(T(\mathbf{X})), \end{aligned}$$

where the last equality confirms again that the log-partition function A is convex (since the covariance matrix is positive semidefinite).

Exercise 16.8: Marginal, conditional and product of exponential family

Let $\mathsf{p}(\mathbf{x}, \mathbf{z})$ be a joint distribution from the exponential family. Prove the following:

- The marginal $\mathsf{p}(\mathbf{x})$ need not be from the exponential family.
- The conditional $\mathsf{p}(\mathbf{z}|\mathbf{x})$ is again from the exponential family.
- The product of two exponential family distributions is again in exponential family.

Exercise 16.9: Exponential family approximation under KL

Let $\mathsf{p}(\mathbf{x})$ be an arbitrary distribution and $\mathsf{q}_{\boldsymbol{\eta}}(\mathbf{x})$ from the exponential family with sufficient statistics T and log-partition function A. Then,

$$\boldsymbol{\eta}^* := \operatorname*{argmin}_{\boldsymbol{\eta}} \ \mathsf{KL}(\mathsf{p} \| \mathsf{q}_{\boldsymbol{\eta}})$$

is given by moment-matching:

$$\mathsf{E}_{\mathsf{p}} T(\mathbf{X}) = \mathsf{E}_{\mathsf{q}_{\boldsymbol{\eta}}} T(\mathbf{X}) = \nabla A(\boldsymbol{\eta}), \quad \textit{i.e.}, \quad \boldsymbol{\eta} = \nabla A^{-1}(\mathsf{E}_{\mathsf{p}} T(\mathbf{X})).$$

Exercise 16.10: EM for exponential family

Prove that the M-step of EM simplifies to the following, if we assume the joint distribution $\mu_{\boldsymbol{\eta}}$ is from the exponential family with natural parameter $\boldsymbol{\eta}$, sufficient statistics T and log-partition function A:

$$\boldsymbol{\eta}_{t+1} = \nabla A^{-1}(\mathsf{E}_{\hat{\nu}_{t+1}}(T(\mathbf{X}))).$$

Definition 16.11: Mixture Distribution

We define the joint distribution over a discrete latent random variable $Z \in \{1, \dots, c\}$ and an observed random variable $\mathbf{X}$:

$$\mathsf{p}(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^{c} [\pi_k \cdot \mathsf{p}_k(\mathbf{x}; \theta_k)]^{z_k},$$

where we represent $\mathbf{z}$ using one-hot encoding. We easily obtain the marginal and conditional:

$$\begin{aligned} \mathsf{p}(\mathbf{x}) &= \sum_{k=1}^{c} \pi_k \cdot \mathsf{p}_k(\mathbf{x}; \theta_k) \qquad (16.1) \\ \mathsf{p}(\mathbf{z} = \mathbf{e}_k) &= \pi_k \\ \mathsf{p}(\mathbf{x} | \mathbf{z} = \mathbf{e}_k) &= \mathsf{p}_k(\mathbf{x}; \theta_k) \\ \mathsf{p}(\mathbf{z} = \mathbf{e}_k | \mathbf{x}) &= \frac{\pi_k \cdot \mathsf{p}_k(\mathbf{x}; \theta_k)}{\sum_{j=1}^{c} \pi_j \cdot \mathsf{p}_j(\mathbf{x}; \theta_j)}. \end{aligned}$$

The marginal distribution $\mathsf{p}(\mathbf{x})$ can be interpreted as follows: There are c component densities p_k. We choose a component p_k with probability π_k and then we sample $\mathbf{x}$ from the resulting component density. However, in reality we do not know which component density an observation $\mathbf{x}$ is sampled from, i.e., the discrete random variable $\mathbf{Z}$ is not observed (missing).

Let $\mathsf{p}_k(\mathbf{x}; \theta_k)$ be multivariate Gaussian (with θ_k denoting its mean and covariance) we get the popular Gaussian mixture model (GMM).

Algorithm 16.12: Mixture density estimation – ML

Replacing the parameterization χ_θ with the mixture model in (16.1) we get a direct method for estimating the density function χ based on a sample:

$$\min_{\boldsymbol{\pi} \in \Delta, \boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathsf{KL}(\hat{\chi} \| \mathsf{p}), \quad \mathsf{p}(\mathbf{x}) = \sum_{k=1}^{c} \pi_k \cdot \mathsf{p}_k(\mathbf{x}; \theta_k)$$

where Δ denotes the simplex constraint (i.e., $\boldsymbol{\pi} \geq \mathbf{0}$ and $\mathbf{1}^\top \boldsymbol{\pi} = 1$). The number of components c is a hyperparameter that needs to be determined *a priori*. We may apply (projected) gradient descent to solve $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. However, it is easy to verify that the objective function is nonconvex hence convergence to a reasonable solution may not be guaranteed.

We record the gradient here for later comparison. We use $\mathsf{p}_W(\mathbf{x}, \mathbf{z})$ for the joint density whose marginalization over the latent $\mathbf{z}$ gives $\mathsf{p}(\mathbf{x})$. For mixtures, the parameter W includes both $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$.

$$\frac{\partial}{\partial W} = -\mathsf{E}_{\hat{\mathsf{p}}_W(\mathbf{z}, \mathbf{x})} \frac{\partial \log \mathsf{p}_W(\mathbf{x}, \mathbf{z})}{\partial W}, \quad \text{where} \quad \hat{\mathsf{p}}_W(\mathbf{z}, \mathbf{x}) := \hat{\chi}(\mathrm{d}\mathbf{x}) \cdot \mathsf{p}_W(\mathbf{z} | \mathbf{x}).$$

Algorithm 16.13: Mixture density estimation – EM

Let us now apply the EM Algorithm 16.3 to the mixture density estimation problem. As mentioned before, we minimize the upper bound:

$$\min_{\boldsymbol{\pi}\in\Delta,\boldsymbol{\theta}\in\boldsymbol{\Theta}} \min_{\mathcal{E}} \ \mathsf{KL}(\hat{\nu}(\mathbf{x},\mathbf{z})\|\mathsf{p}(\mathbf{x},\mathbf{z})), \quad \hat{\nu}(\mathbf{x},\mathbf{z}) = \hat{\chi}(\mathbf{x})\mathcal{E}(\mathbf{z}|\mathbf{x}), \ \mathsf{p}(\mathbf{x},\mathbf{z}) = \prod_{k=1}^{c}[\pi_k \cdot \mathsf{p}_k(\mathbf{x};\theta_k)]^{z_k},$$

with the following two steps alternated until convergence:

- E-step: Fix $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$, we solve

$$\mathcal{E}_{t+1} = \mathsf{p}^{(t)}(\mathbf{z}=\mathbf{e}_k|\mathbf{x}) = \frac{\pi_k^{(t)} \cdot \mathsf{p}_k(\mathbf{x};\theta_k^{(t)})}{\sum_{j=1}^{c}\pi_j^{(t)} \cdot \mathsf{p}_j(\mathbf{x};\theta_j^{(t)})} =: r_k^{(t+1)}(\mathbf{x}). \tag{16.2}$$

- M-step: Fix $\mathcal{E}_{t+1}$ hence $\hat{\nu}_{t+1}$, we solve

$$\begin{aligned}\min_{\boldsymbol{\pi}\in\Delta}\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathsf{KL}(\hat{\nu}_{t+1}(\mathbf{x},\mathbf{z})\|\mathsf{p}(\mathbf{x},\mathbf{z})) &\equiv \max_{\boldsymbol{\pi}\in\Delta}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathsf{E}_{\hat{\chi}}\mathsf{E}_{\mathcal{E}_{t+1}}\left[\langle\mathbf{z},\log\boldsymbol{\pi}\rangle + \langle\mathbf{z},\log\mathbf{p}(\mathbf{X};\boldsymbol{\theta})\rangle\right] \\ &= \max_{\boldsymbol{\pi}\in\Delta}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathsf{E}_{\hat{\chi}}\left[\left\langle\mathbf{r}^{(t+1)}(\mathbf{X}),\log\boldsymbol{\pi}\right\rangle + \left\langle\mathbf{r}^{(t+1)}(\mathbf{X}),\log\mathbf{p}(\mathbf{X};\boldsymbol{\theta})\right\rangle\right].\end{aligned}$$

  It is clear that the optimal

$$\boldsymbol{\pi}^{(t+1)} = \mathsf{E}_{\hat{\chi}}\mathbf{r}^{(t+1)}(\mathbf{X}) = \sum_{i=1}^{n}\hat{\chi}(\mathbf{x}_i)\cdot\mathbf{r}^{(k+1)}(\mathbf{x}_i). \tag{16.3}$$

  (For us the empirical training distribution $\hat{\chi} \equiv \frac{1}{n}$, although we prefer to keep everything abstract and general.)

  The θ_k's can be solved independently:

$$\max_{\theta_k\in\Theta_k} \mathsf{E}_{\hat{\chi}}\left[r_k^{(t+1)}(\mathbf{X})\cdot\log\mathsf{p}_k(\mathbf{X};\theta_k)\right] \quad\equiv\quad \min_{\theta_k\in\Theta_k} \mathsf{KL}(\hat{\chi}_k^{(t+1)}\|\mathsf{p}_k(\cdot;\theta_k)), \tag{16.4}$$

  where we define $\hat{\chi}_k^{(t+1)} \propto \hat{\chi}\cdot r_k^{(t+1)}$. (This is similar to Adaboost, where we reweigh the training examples!)

If we choose the component density $\mathsf{p}_k(\mathbf{x};\theta_k)$ from the exponential family with sufficient statistics T_k and log-partition function A_k, then from Exercise 16.9 we know (16.4) can be solved in closed-form:

$$\theta_k^{(t+1)} = \nabla A_k^{-1}\left[\mathsf{E}_{\hat{\chi}_k^{(t+1)}}T_k(\mathbf{X})\right] \tag{16.5}$$

Alert 16.14: implicit EM vs. explicit ML

We now make an important connection between EM and ML. We follow the notation in Algorithm 16.12. For the joint density $\mathsf{p}_W(\mathbf{x},\mathbf{z})$, EM solves

$$W_{t+1} = \operatorname*{argmin}_W \mathsf{KL}(\hat{\mathsf{p}}_{t+1}\|\mathsf{p}_W), \qquad \text{where} \quad \hat{\mathsf{p}}_{t+1}(\mathbf{x},\mathbf{z}) := \hat{\mathsf{p}}_{W_t}(\mathbf{x},\mathbf{z}) = \hat{\chi}(\mathrm{d}\mathbf{x})\cdot\mathsf{p}_{W_t}(\mathbf{z}|\mathbf{x}).$$

In particular, at a minimizer W_{t+1} the gradient vanishes:

$$-\mathsf{E}_{\hat{\mathsf{p}}_{t+1}}\frac{\partial\log\mathsf{p}_W(\mathbf{x},\mathbf{z})}{\partial W} = \mathbf{0}.$$

In other words, EM solves the above nonlinear equation (in W) to get W_{t+1} while ML with gradient descent simply performs one fixed-point iteration.

Example 16.15: Gaussian Mixture Model (GMM) – EM

Using the results in multivariate Gaussian Example 16.5 we derive:

$$\nabla A(\boldsymbol{\eta}) = \begin{bmatrix} S^{-1}\boldsymbol{\xi} \\ -\frac{1}{2}(S^{-1}\boldsymbol{\xi}\boldsymbol{\xi}^\top S^{-1} + S^{-1}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ -\frac{1}{2}(\boldsymbol{\mu}\boldsymbol{\mu}^\top + \Sigma) \end{bmatrix},$$

$$\mathsf{E}_{\hat{\chi}^{(t+1)}} T(\mathbf{X}) = \begin{bmatrix} \mathsf{E}_{\hat{\chi}^{(t+1)}} \mathbf{X} \\ -\frac{1}{2}\mathsf{E}_{\hat{\chi}^{(t+1)}} \mathbf{X}\mathbf{X}^\top \end{bmatrix} \propto \sum_{i=1}^{n} (\hat{\chi} \cdot r^{(t+1)})(\mathbf{x}_i) \begin{bmatrix} \mathbf{x}_i \\ -\frac{1}{2}\mathbf{x}_i\mathbf{x}_i^\top \end{bmatrix},$$

where we have omitted the component subscript k. Thus, from (16.5) we obtain:

$$\boldsymbol{\mu}_k^{(t+1)} = \mathsf{E}_{\hat{\chi}_k^{(t+1)}} \mathbf{X} = \sum_{i=1}^{n} \frac{(\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_i)}{\sum_{\iota=1}^{n} (\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_\iota)} \cdot \mathbf{x}_i \tag{16.6}$$

$$\Sigma_k^{(t+1)} = \mathsf{E}_{\hat{\chi}_k^{(t+1)}} \mathbf{X}\mathbf{X}^\top - \boldsymbol{\mu}_k^{(t+1)} \boldsymbol{\mu}_k^{(t+1)\top} = \sum_{i=1}^{n} \frac{(\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_i)}{\sum_{\iota=1}^{n} (\hat{\chi} \cdot r_k^{(t+1)})(\mathbf{x}_\iota)} \cdot (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^\top, \tag{16.7}$$

where we remind that the empirical training distribution $\hat{\chi} \equiv \frac{1}{n}$.

The updates on "responsibility" $\mathbf{r}$ in (16.2), mixing distribution $\boldsymbol{\pi}$ in (16.3), on the means $\boldsymbol{\mu}_k$ in (16.6), and on the covariance matrices S_k in (16.7), consist of the main steps for estimating a GMM using EM.