Algorytm EM

Algorytm EM

Algorytm EM ma wiele zastosowań, w sytuacji gdy bezpośrednia maksymalizacji funkcji log-wiarogodności jest trudna. W wielu sytuacjach, problem maksymalizacji można uprościć, jeżeli przestrzeń parametrów rozszerzy się o dodatkowe - nieobserwowane zmienne.

Typowe przykłady takich sytuacji to uzupełnianie braków w danych oraz badanie mieszanin rozkładów.

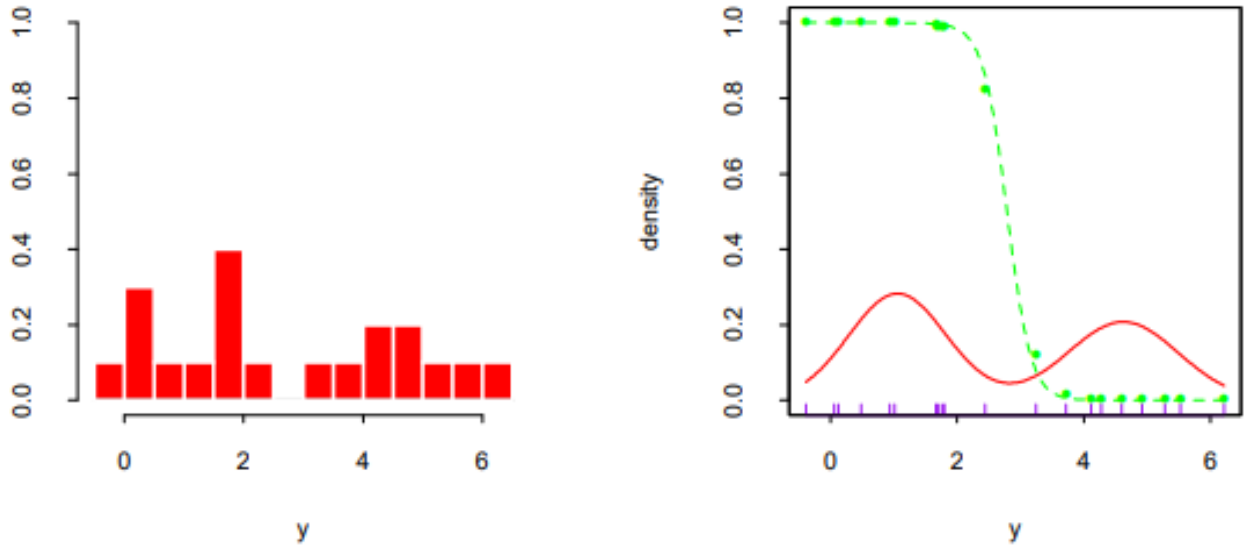


FIGURE 8.5. *Mixture example. (Left panel:) Histogram of data. (Right panel:) Maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation y, as a function of y.*

$$
\begin{aligned}
Y_1 &\sim N(\mu_1, \sigma_1^2), \\
Y_2 &\sim N(\mu_2, \sigma_2^2), \\
Y &= (1-\Delta) \cdot Y_1 + \Delta \cdot Y_2
\end{aligned}
$$

$$
\Delta \in \{0, 1\} \text{ with } \Pr(\Delta = 1) = \pi.
$$

Oznaczmy przez $\phi_\theta(x)$ gęstość rozkładu normalnego z parametrami $\theta = (\mu, \sigma^2)$.

Wtedy gęstość Y ma postać:

$$g_Y(y) = (1-\pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y)$$

Z parametrami $\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$.

Wtedy logarytm funkcji wiarogodności dany jest wzorem:

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \log[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)].$$

Załóżmy, że znamy wartości Δi, wtedy funkcja log-wiarogodności przybrałaby następującą postać:

$$\begin{aligned} \ell_0(\theta; \mathbf{Z}, \boldsymbol{\Delta}) &= \sum_{i=1}^{N} [(1-\Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i \log\phi_{\theta_2}(y_i)] \\ &\quad + \sum_{i=1}^{N} [(1-\Delta_i)\log(1-\pi) + \Delta_i \log\pi], \end{aligned}$$

- Ponieważ jednak wartości Δi są nieznane, stosujemy procedurę iteracyjną, podstawiając pod każde Δi jego wartość oczekiwaną:

$$\gamma_i(\theta) = \mathrm{E}(\Delta_i|\theta, \mathbf{Z}) = \Pr(\Delta_i = 1|\theta, \mathbf{Z}),$$

- Aby ustalić początkowe wartości parametrów jednym ze sposobów jest wybranie dwóch losowych yi i potraktowanie ich jako estymatory wartości oczekiwanej rozkładów. Jako startową wariancję możemy przyjąć wariancję z próbki oraz przyjąć parametr π=0.5.

Algorithm 8.1 *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).
2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \; i = 1, 2, \ldots, N. \qquad (8.42)$$

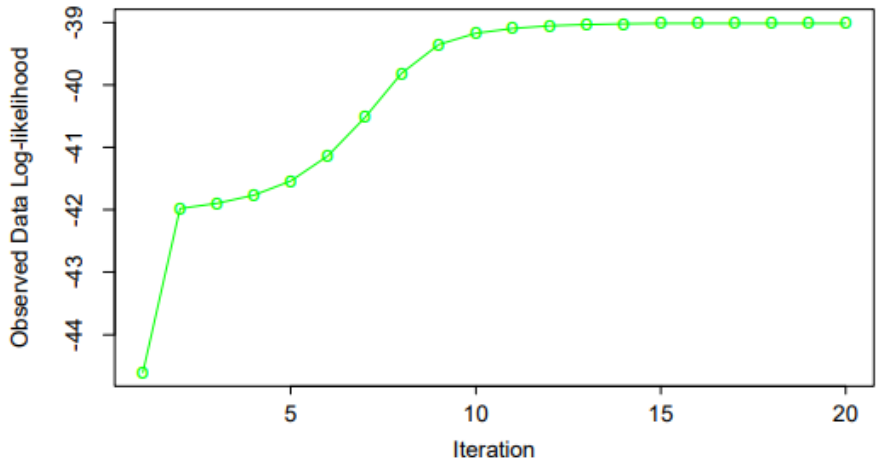
3. *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.



Uogólniony algorytm EM

- Oznaczmy przez **Z** obserwowane dane, a ich funkcję log-wiarogodności poprzez $\ell(\theta; \mathbf{Z})$.
- Ukryte lub brakujące dane oznaczamy poprzez $\mathbf{Z}^m$, a kompletne dane poprzez **T** = $(\mathbf{Z}, \mathbf{Z}^m)$, z funkcją log-wiarogodności $\ell_0(\theta; \mathbf{T})$, ℓ_0 na podstawie całościowej gęstości.

Algorithm 8.2 *The EM Algorithm.*

1. Start with initial guesses for the parameters $\hat{\theta}^{(0)}$.
2. *Expectation Step*: at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathrm{E}(\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)}) \tag{8.43}$$

   as a function of the dummy argument θ'.
3. *Maximization Step*: determine the new estimate $\hat{\theta}^{(j+1)}$ as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.
4. Iterate steps 2 and 3 until convergence.

Dlaczego ten algorytm działa?

- Ponieważ mamy $\Pr(\mathbf{Z}^m|\mathbf{Z},\theta') = \frac{\Pr(\mathbf{Z}^m,\mathbf{Z}|\theta')}{\Pr(\mathbf{Z}|\theta')}$, możemy zapisać $\Pr(\mathbf{Z}|\theta') = \frac{\Pr(\mathbf{T}|\theta')}{\Pr(\mathbf{Z}^m|\mathbf{Z},\theta')}$.
- Rozpatrując funkcję log-wiarogodności, mamy $\ell(\theta';\mathbf{Z}) = \ell_0(\theta';\mathbf{T}) - \ell_1(\theta';\mathbf{Z}^m|\mathbf{Z})$, gdzie ℓ_1 jest zależne od warunkowej gęstości $\Pr(\mathbf{Z}^m|\mathbf{Z},\theta')$
- Biorąc warunkową wartość oczekiwaną z uwagi na rozkład T|Z i parametr θ otrzymujemy:

$$\ell(\theta';\mathbf{Z}) = \mathrm{E}[\ell_0(\theta';\mathbf{T})|\mathbf{Z},\theta] - \mathrm{E}[\ell_1(\theta';\mathbf{Z}^m|\mathbf{Z})|\mathbf{Z},\theta] \equiv Q(\theta',\theta) - R(\theta',\theta)$$

- Zauważmy, że $R(\theta^*,\theta)$ jest wartością oczekiwaną log-wiarogodności gęstości (indeksowanej θ *) względem tej samej gęstości indeksowanej θ , stąd z nierówności Jensena jest maksymalizowana w punkcie $\theta^* = \theta$. Stąd mamy:

$$\begin{aligned}\ell(\theta';\mathbf{Z}) - \ell(\theta;\mathbf{Z}) &= [Q(\theta',\theta) - Q(\theta,\theta)] - [R(\theta',\theta) - R(\theta,\theta)] \\ &\geq 0.\end{aligned}$$

- Czyli algorytm EM nigdy nie obniża log-wiarogodności.

EM jako procedura Maximization–Maximization

- Rozważmy poniższą funkcję, gdzie $\tilde{P}(\mathbf{Z}^m)$ jest dowolnym rozkładem ukrytej zmiennej $\mathbf{Z}^m$

$$F(\theta', \tilde{P}) = \mathrm{E}_{\tilde{P}}[\ell_0(\theta'; \mathbf{T})] - \mathrm{E}_{\tilde{P}}[\log \tilde{P}(\mathbf{Z}^m)].$$

- W przykładzie mieszanki rozkładów $\tilde{P}(\mathbf{Z}^m)$ zawiera zbiór prawdopodobieństw $\gamma_i = \Pr(\Delta_i = 1|\theta, \mathbf{Z})$

- Zauważmy, że funkcja F w $\tilde{P}(\mathbf{Z}^m) = \Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')$ jest funkcją log-wiarogodności obserwowanych danych. Funkcja F rozszerza dziedzinę log-wiarogodności, aby ułatwić jego maksymalizację.

- Algorytm EM może być widziany jako metoda wspólnej maksymalizacji funkcji F nad θ' i $\tilde{P}(\mathbf{Z}^m)$, poprzez ustalenie jednego z argumentów i maksymalizowanie po drugim.
- Maksymalizatorem nad $\tilde{P}(\mathbf{Z}^m)$ dla ustalonego θ' jest $\tilde{P}(\mathbf{Z}^m) = \Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')$ Jest to rozkład obliczony przez krok E, na przykład w przykładzie modelu mieszaniny.

- W kroku M maksymalizujemy $F(\theta', \tilde{P})$ nad θ' dla ustalonego $\tilde{P}(\mathbf{Z}^m)$.
- Ostatecznie, skoro $F(\theta', \tilde{P})$ i funkcja log-wiarogodności obserwowanych danych zgadzają się w punkcie $\tilde{P}(\mathbf{Z}^m) = \Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')$ maksymalizacja pierwszego osiąga maksymalizację drugiego.

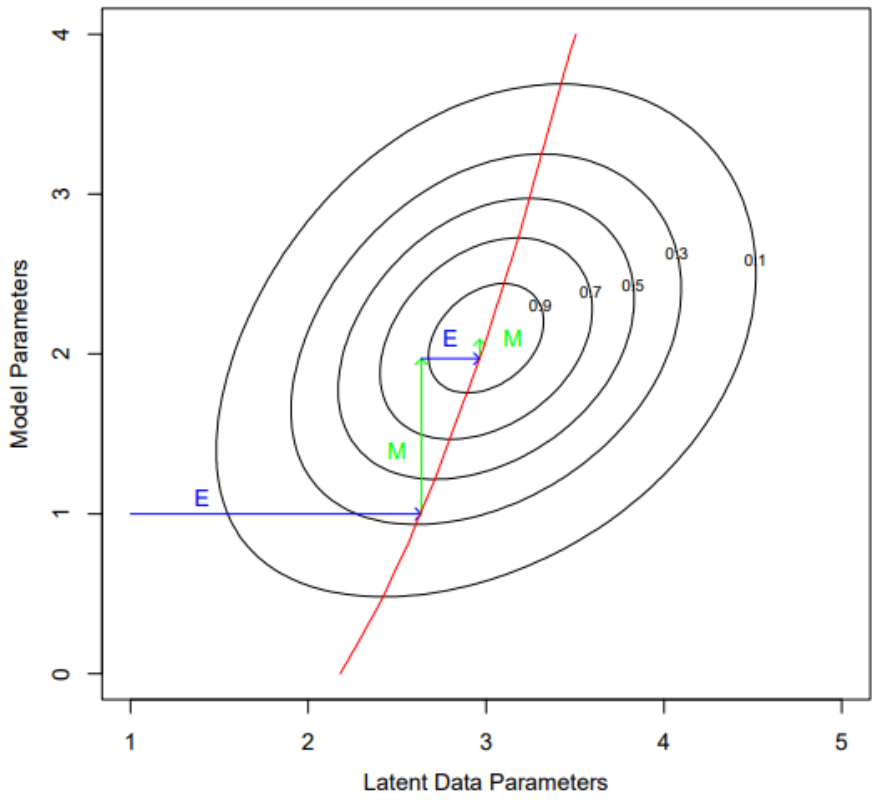


FIGURE 8.7. *Maximization–maximization view of the EM algorithm. Shown are the contours of the (augmented) observed data log-likelihood* $F(\theta', \tilde{P})$. *The E step is equivalent to maximizing the log-likelihood over the parameters of the latent data distribution. The M step maximizes it over the parameters of the log-likelihood. The red curve corresponds to the observed data log-likelihood, a* profile *obtained by maximizing* $F(\theta', \tilde{P})$ *for each value of* θ'.

Dziękuje za uwagę