Der E(xpectation-)M(aximization)-Algorithmus

von Peter Pfaffelhuber

Version: 11. Januar 2016

In bestimmten Fällen ist es schwierig, Maximum-Likelihood-Schätzer direkt anzugeben. Wir werden hier einen besonderen Fall diskutieren, in dem es einen Ausweg über den EM-Algorithmus gibt.

1 Maximum-Likelihood-Schätzer in Mischungsmodellen

Wir nehmen ein statistisches Modell

$$(X, \{\mathbb{P}_\theta = ((1-\pi)p_{\theta_0} + \pi p_{\theta_1})^n \cdot \lambda^n : \theta = (\pi, \theta_0, \theta_1), \pi \in [0,1], \theta_0, \theta_1 \in \mathcal{P}\}) \tag{1}$$

an. Das bedeutet: da $\mathbb{P}_\theta$ ein Produktmaß ist, sind die Datenpunkte $X_1, ..., X_n$ unabhängig. Die Verteilung von X_1 ist jedoch eine Mischung aus $p_{\theta_0} \cdot \lambda$ und $p_{\theta_1} \cdot \lambda$. Das stellt man sich am besten so vor: Sei $Z_1 = 1$ mit Wahrscheinlichkeit π und $Z_1 = 0$ mit Wahrscheinlichkeit $(1-\pi)$. Im Anschluss ziehen wir eine nach $p_{\theta_{Z_1}}$-verteilte Zufallsvariable X_1. Dann ist nämlich gerade $X_1 \sim ((1-\pi)p_{\theta_0} + \pi p_{\theta_1}) \cdot \lambda$.

Will man in einer solchen Situation einen Maximum-Likelihood-Schätzer für $\theta = (\pi, \theta_0, \theta_1)$ angeben, so hätte man

$$\ell(\theta; X) = \sum_{i=1}^n \log\big((1-\pi)p_{\theta_0}(X_i) + \pi p_{\theta_1}(X_i)\big)$$

zu maximieren, was wegen der Summe innerhalb von log nicht einfach ist. Viel einfacher wäre es, wenn wir über X_i wüssten, ob es sich um eine Ziehung aus $p_{\theta_0} \cdot \lambda$ oder aus $p_{\theta_1} \cdot \lambda$ handelt. Wäre nämlich $Z_i = 0$ oder $Z_i = 1$ je nachdem, welcher Fall eintritt, so ist die log-Likelihood

$$\ell'(\theta; X, Z) = \sum_{i=1}^n Z_i \log \pi + \sum_{i=1}^n (1-Z_i)\log(1-\pi) + \sum_{i=1}^n (1-Z_i)\log p_{\theta_0}(X_i) + Z_i p_{\theta_1}(X_i). \tag{2}$$

Hierbei meinen wir allerdings die Likelihood im statistischen Modell

$$((X,Z), \{\mathbb{P}_\theta = (p_\theta \cdot \lambda)^n : \theta = (\pi, \theta_0, \theta_1)\}) \text{ mit } p_\theta(x,z) = \pi^z(1-\pi)^{1-z}p_{\theta_z}(x). \tag{3}$$

Obwohl sicherlich die Maximierung im statistischen Modell 3 einfacher wird, kennen wir normalerweise Z nicht. Wir sprechen hier auch davon, dass Z eine *latente* oder *unbeobachtete* Variable ist. Der Trick, den der EM-Algorithmus verwendet, ist es, diese Variablen Z_i durch ihre Erwartung zu ersetzen, und anschließend die Likelihood zu maximieren.

Wir bemerken noch, dass es sich bei der Dichte im statistischen Modell (1) nicht um eine Exponentialfamilie handelt, in (3) aber schon. Auch dies spricht zumindest dafür, dass Berechnungen im Modell (3) einfacher sein werden.

2 Der EM-Algorithmus

Wir beginnen damit, in der obigen Situation den rekursiv definierten EM-Algorithmus anzugeben, wobei x die erhobenen Daten sind.

Der EM-Algorithmus

1. Starte für $t = 0$ mit Anfangswerten $\hat{\theta}^{(t=0)}$.
2. E-Schritt: Berechne für $t = 0, 1, 2, ...$
$$Q(\theta, \hat{\theta}^{(t)}; x) := \mathbb{E}_{\hat{\theta}^{(t)}}[\ell'(\theta, X, Z)|X = x].$$
3. M-Schritt: Berechne
$$\hat{\theta}^{(t+1)} := \arg\max\{Q(\theta, \theta^{(t)}; x) : \theta \in \mathcal{P}\}.$$
4. Iteriere 2. und 3. bis $\hat{\theta}^{(t)}$ konvergiert.

Für den speziellen Fall des obigen Mischungsmodells berechnen wir zunächst für $X = x$

$$\mathbb{E}_{\theta'}[\ell'(\theta; x, Z)] = n(\pi' \log \pi + (1 - \pi') \log(1 - \pi)) + \sum_{i=1}^{n} (1 - \pi') \log p_{\theta_0}(x_i) + \pi' \log p_{\theta_1}(x_i).$$

Nun ergibt sich folgendes:

Der EM-Algorithmus für das Mischungsmodell

1. Starte für $t = 0$ mit Anfangswerten $\hat{\theta}^{(t=0)} = (\hat{\pi}, \hat{\theta}_0, \hat{\theta}_1)$.

2. Erwartungswert-Schritt: Berechne

$$\hat{Z}_i^{(t+1)} = \mathbb{E}_{\hat{\theta}^{(t)}}[Z_i | X_i = x_i] = \frac{\hat{\pi}^{(t)} p_{\hat{\theta}_1^{(t)}}(x_i)}{(1-\hat{\pi}^{(t)}) p_{\hat{\theta}_0^{(t)}}(x_i) + \hat{\pi}^{(t)} p_{\hat{\theta}_1^{(t)}}(x_i)}, \qquad i = 1, ..., n$$

   Damit ergibt sich

$$Q(\theta, \theta^{(t)}; x) = \sum_{i=1}^{n} (\hat{Z}_i^{(t+1)} \log \pi + (1 - \hat{Z}_i^{(t+1)}) \log(1-\pi)) \\ + \sum_{i=1}^{n} (1 - \hat{Z}_i^{(t+1)}) \log p_{\theta_0}(x) + \hat{Z}_i^{(t+1)} \log p_{\theta_1}(x).$$

3. Maximierungs-Schritt: Wir führen zunächst die Maximierung bezüglich π durch. Wir erhalten als notwendige Bedingung

$$\sum_{i=1}^{n} (1-\pi) \hat{Z}_i^{(t+1)} = \sum_{i=1}^{n} (1 - \hat{Z}_i^{(t+1)}) \pi, \text{ also } \hat{\pi}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_i^{(t+1)}.$$

   Weiter berechnen wir (mit (2))

$$\hat{\theta}^{(t+1)} := \arg\max\{Q(\theta, \hat{\theta}^{(t)}; x) : \theta \in \mathcal{P}\}.$$

4. Iteriere 2. und 3. bis $\hat{\theta}^{(t)}$ konvergiert.

Beispiel 1 (Mischung aus Normalverteilungen). Hier sei für $\theta_i = (\mu_i, \sigma_i^2)$ die Verteilung $\mathbb{P}_{\theta_i} = \mathcal{N}(\mu_i, \sigma_i^2), i = 0, 1$. Hier ist also (wenn μ_0 und μ_1 genügend weit auseinander liegen) die Dichte $(1-\pi)p_{\theta_0} + \pi p_{\theta_1}$ bi-modal (d.h. sie hat zwei Maxima).

In Maximierungs-Schritt des EM-Algorithmus ist also noch

$$\sum_{i=1}^{n} (1 - \hat{Z}_i) \log p_{\theta_0}(x) + \hat{Z}_i \log p_{\theta_1}(x)$$

zu maximieren. Im Fall der Normalverteilungen müssen wir also

$$\sum_{i=1}^{n} (1 - \hat{Z}_i) \Big(\frac{1}{2} \log \sigma_0^2 + \frac{(x_i - \mu_0)^2}{2\sigma_0^2} \Big),$$
$$\sum_{i=1}^{n} \hat{Z}_i \Big(\frac{1}{2} \log \sigma_1^2 + \frac{(x_i - \mu_1)^2}{2\sigma_1^2} \Big)$$

über (μ_0, σ_0^2) bzw. (μ_1, σ_1^2) maximieren. Wir berechnen

$$\sigma_0^2 \frac{\partial}{\partial \mu_0} \sum_{i=1}^{n} (1 - \hat{Z}_i) \Big(\frac{1}{2} \log \sigma_0^2 + \frac{(x_i - \mu_0)^2}{2\sigma_0^2} \Big) = \sum_{i=1}^{n} (1 - \hat{Z}_i)(x_i - \mu_0),$$
$$2\sigma_0^2 \frac{\partial}{\partial \sigma_0^2} \sum_{i=1}^{n} (1 - \hat{Z}_i) \Big(\frac{1}{2} \log \sigma_0^2 + \frac{(x_i - \mu_0)^2}{2\sigma_0^2} \Big) = \sum_{i=1}^{n} (1 - Z_i) \Big(1 - \frac{(x_i - \mu_0)^2}{2\sigma_0^2} \Big)$$

und damit sind die Maximierer

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} (1 - \hat{Z}_i) x_i}{\sum_{i=1}^{n} (1 - \hat{Z}_i)},$$
$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n} (1 - \hat{Z}_i)(x_i - \hat{\mu}_0)^2}{\sum_{i=1}^{n} (1 - \hat{Z}_i)}.$$

Analog ergeben sich

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} \hat{Z}_i x_i}{\sum_{i=1}^{n} \hat{Z}_i},$$
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^{n} \hat{Z}_i (x_i - \hat{\mu}_0)^2}{\sum_{i=1}^{n} \hat{Z}_i}.$$

Beispiel 2 (Gauß'sches Mischungsmodell). Für $\pi = 0.5$ wählen wir $(\mu_0, \sigma_0^2) = (-2, 1)$ und $(\mu_1, \sigma_1^2) = (2, 1)$.

```
n=1000
pi=0.5
z<-rbinom(n,1,pi)
x<-0*z

for(i in 1:n) {
  if(z[i]==0) x[i]<-rnorm(1, mean=-2, sd=1)
  else x[i]<-rnorm(1, mean=2, sd=1)
}
```

Nachdem unsere Daten nun in `x` gespeichert sind, wollen wir die Parameter schätzen.

```
# theta = c(pi, mu0, sigma20, mu1, sigma21) = (0.5,-2,1,2,1)
thetaold=0
theta = c(0.2, -1, 1, 1, 1)

eps=10^(-4)

while(norm(as.matrix(theta-thetaold))>eps) {
  thetaold=theta
  hatZ<-(theta[1]*dnorm(x, mean=theta[4], sd=sqrt(theta[5])))/
      (theta[1]*dnorm(x, mean=theta[4], sd=sqrt(theta[5]))
      + (1-theta[1])*dnorm(x, mean=theta[2], sd=sqrt(theta[3])))
```

```
  theta[1] = mean(hatZ)
  theta[2] = sum((1-hatZ)*x)/sum(1-hatZ)
  theta[3] = sum((1-hatZ)*(x-theta[2])^2)/sum(1-hatZ)
  theta[4] = sum(hatZ*x)/sum(hatZ)
  theta[5] = sum(hatZ*(x-theta[4])^2)/sum(hatZ)
  cat(theta, "\n")
}
0.4080493 -1.657099 1.873421 2.146217 1.0034
0.4221161 -1.757708 1.571493 2.157208 0.8661055
0.4324705 -1.818895 1.393268 2.143772 0.8500342
0.4408061 -1.860213 1.292241 2.121252 0.8673303
0.4472676 -1.889556 1.228207 2.099996 0.8913398
0.4521783 -1.910826 1.184815 2.082438 0.9139556
0.4558773 -1.926362 1.154515 2.06858 0.9331244
0.4586528 -1.937761 1.133011 2.057859 0.9486591
0.4607324 -1.946157 1.117571 2.049651 0.9609445
0.4622902 -1.952364 1.106378 2.043406 0.9705129
0.4634575 -1.956969 1.098198 2.038673 0.9778897
0.4643328 -1.960395 1.092182 2.035094 0.983537
0.4649893 -1.96295 1.087734 2.032393 0.987839
0.465482 -1.96486 1.084432 2.030356 0.9911048
0.465852 -1.966289 1.081973 2.028822 0.9935777
0.4661299 -1.967359 1.080138 2.027666 0.9954469
0.4663386 -1.968162 1.078766 2.026797 0.9968578
0.4664955 -1.968764 1.077738 2.026142 0.9979217
0.4666134 -1.969217 1.076968 2.02565 0.9987234
0.466702 -1.969556 1.076391 2.025279 0.9993273
0.4667686 -1.969812 1.075957 2.025001 0.9997818
0.4668187 -1.970003 1.075631 2.024791 1.000124
0.4668563 -1.970147 1.075387 2.024634 1.000381
0.4668847 -1.970256 1.075203 2.024515 1.000575
0.4669059 -1.970337 1.075065 2.024426 1.000721
0.4669219 -1.970398 1.074961 2.024359 1.00083
0.466934 -1.970445 1.074883 2.024309 1.000913
0.466943 -1.970479 1.074824 2.024271 1.000975
0.4669498 -1.970505 1.07478 2.024242 1.001021
0.4669549 -1.970525 1.074747 2.024221 1.001056
0.4669588 -1.970539 1.074722 2.024205 1.001083
```

Wir wollen nun noch begründen, warum der EM-Algorithmus (zumindest lokale) Maxima der Likelihood ansteuert.

Proposition 3. *Für*

$$\ell(\theta; x) := \log \int p_\theta(x, z')\lambda(dz')$$

gilt

$$\ell(\hat{\theta}^{(t+1)}; x) \geq \ell(\hat{\theta}^{(t)}; x)$$

mit "=" genau dann, wenn $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)}$.

Beweis. Wir schreiben zunächst

$$p_{\theta,x}(z) := \frac{p_\theta(x,z)}{\int p_\theta(x,z')\lambda(dz')}$$

als die bedingte Dichte von p_θ gegeben $X = x$. Dann ist nämlich

$$\ell(\theta; x) = \log p_\theta(x,z) - \log p_{\theta,x}(z).$$

Nehmen wir nun auf der rechten Seite Erwartungswerte bezüglich der Verteilung mit Dichte $p_{\hat\theta^{(t)},x}(z)$, so folgt (mit $R(\theta,\theta';x) := \mathbb{E}_{\theta'}[\log p_{\theta,X}(Z)|X=x]$)

$$\begin{aligned}\ell(\theta;x) &= \mathbb{E}_{\hat\theta^{(t)}}[\log p_\theta(X,Z)|X=x] - \mathbb{E}_{\hat\theta^{(t)}}[\log p_{\theta,X}(Z)|X=x] \\ &= Q(\theta,\hat\theta^{(t)};x) - R(\theta,\hat\theta^{(t)};x).\end{aligned}$$

Nun ist

$$\ell(\hat\theta^{(t+1)};x) - \ell(\hat\theta^{(t)};x) = Q(\hat\theta^{(t+1)},\hat\theta^{(t)};x) - Q(\hat\theta^{(t)},\hat\theta^{(t)};x) - (R(\hat\theta^{(t+1)},\hat\theta^{(t)};x) - R(\hat\theta^{(t)},\hat\theta^{(t)};x))$$

und wir haben

$$Q(\hat\theta^{(t+1)},\hat\theta^{(t)};x) - Q(\hat\theta^{(t)},\hat\theta^{(t)};x) \geq 0$$

da $\hat\theta^{(t+1)}$ der Maximierer von $\theta \mapsto Q(\theta,\hat\theta^{(t)};x)$ ist, und für jedes θ,θ' ist

$$\begin{aligned}R(\theta',\theta) - R(\theta,\theta) &= \mathbb{E}_\theta[\log p_{\theta',X}(Z)|X=x] - \mathbb{E}_\theta[\log p_{\theta,X}(Z)|X=x] \\ &= \mathbb{E}_\theta\Big[\log\frac{p_{\theta',X}(Z)}{p_{\theta,X}(Z)}|X=x\Big] \leq \log\mathbb{E}_\theta\Big[\frac{p_{\theta',X}(Z)}{p_{\theta,X}(Z)}|X=x\Big] \\ &= \log\int\frac{p_{\theta',x}(z)}{p_{\theta,x}(z)}p_{\theta,x}(z)\lambda(dz) = 0.\end{aligned}$$

mit "=" genau dann, wenn $\theta = \theta'$. Daraus folgen nun alle Behauptungen. □

Der EM-Algorithmus konvergiert wegen der Beschränktheit der Likelihood immer. Allerdings ist unklar, ob er nicht in einem lokalen Maximum der Likelihood konvergiert. Eine mögliche Strategie, dies zu umgehen, ist es, in verschiedenen Startpunkten zu beginnen, und dann die maximale Likelihood auszuwählen.