Normal hidden Markov models

Stanisław Galus

13 January 2013

1 Forward-backward variables

Let $(X_t)_{t=1}^T$ be a homogeneous Markov chain over the state space $S = \{1, \dots, s\}$ with transition matrix $P = [p_{ij}]$, $i, j \in S$, and initial state distribution $p = [p_i]$, $i \in S$. Then, for each $i_1, \dots, i_T \in S$,

$$P(X_1 = i_1, X_2 = i_2, \dots, X_T = i_T) = p_{i_1} p_{i_1 i_2} \dots p_{i_{T-1} i_T}.$$

For each state $i \in S$, let $f_i(y, \theta_i)$ be a corresponding probability density function. In each moment $t = 1, \dots, T$, a value y_t of a random variable Y_t is observed which comes from the density f_{i_t}. The likelihood of the sample $y_1, \dots, y_T$ is

$$\begin{aligned}
\mathcal{L} &= L(p, P, \theta_1, \dots, \theta_s) = \\
&= \sum_{i_1, \dots, i_T = 1}^{s} p_{i_1} f_{i_1}(y_1, \theta_{i_1}) p_{i_1 i_2} f_{i_2}(y_2, \theta_{i_2}) \dots p_{i_{T-1} i_T} f_{i_T}(y_T, \theta_{i_T}) = \\
&= \sum_{i_1=1}^{s} p_{i_1} f_{i_1}(y_1, \theta_{i_1}) \sum_{i_2=1}^{s} p_{i_1 i_2} f_{i_2}(y_2, \theta_{i_2}) \dots \sum_{i_T=1}^{s} p_{i_{T-1} i_T} f_{i_T}(y_T, \theta_{i_T}).
\end{aligned}$$

The last expression can be calculated using forward variables

$$\alpha_1(j) = p_j f_j(y_1, \theta_j), \quad j \in S, \tag{1}$$

$$\alpha_t(j) = \sum_{i=1}^{s} (\alpha_{t-1}(i) p_{ij}) f_j(y_t, \theta_j), \quad j \in S, \quad t = 2, \dots, T \tag{2}$$

or backward variables

$$\beta_T(i) = 1, \quad i \in S, \tag{3}$$

$$\beta_t(i) = \sum_{j=1}^{s} p_{ij} f_j(y_{t+1}, \theta_j) \beta_{t+1}(j), \quad i \in S, \quad t = T-1, \dots, 1 \tag{4}$$

or both as

$$\mathcal{L} = \sum_{i=1}^{s} \alpha_T(i) = \sum_{i=1}^{s} p_i f_i(y_1, \theta_i) \beta_1(i) = \sum_{i=1}^{s} \alpha_t(i) \beta_t(i), \quad t = 1, \dots, T. \tag{5}$$

Moreover, if we define

$$\begin{aligned}\gamma_t(i) &= \alpha_t(i)\beta_t(i)/\mathcal{L}, \quad t=1,\ldots,T, \quad i\in S, & (6)\\ \xi_t(i,j) &= \alpha_t(i)\beta_{t+1}(j)p_{ij}f_j(y_{t+1},\theta_j)/\mathcal{L}, \quad t=1,\ldots,T-1, \quad i,j\in S & (7)\end{aligned}$$

the following interpretations are possible:

$$\begin{aligned}\alpha_t(i) &= P(Y_1=y_1,\ldots,Y_t=y_t,X_t=i),\\ \beta_t(i) &= P(Y_{t+1}=y_{t+1},\ldots,Y_T=y_T,X_t=i),\\ \gamma_t(i) &= P(Y_1=y_1,\ldots,Y_T=y_T,X_t=i),\\ \xi_t(i,j) &= P(Y_1=y_1,\ldots,Y_T=y_T,X_t=i,X_{t+1}=j),\end{aligned}$$

where P denotes likelihood of the respective event.

2 Baum-Welch algorithm

If $f_i(y,\theta_i)=\phi((y-\mu_i)/\sigma_i)$, where ϕ is standard normal probability density, the following formulas can be used for $i,j\in S$ to increase the likelihood $\mathcal{L}$:

$$\begin{aligned}\overline{p}_i &= \gamma_1(i), & (8)\\ \overline{p}_{ij} &= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}, & (9)\\ \overline{\mu}_i &= \frac{\sum_{t=1}^{T}\gamma_t(i)y_t}{\sum_{t=1}^{T}\gamma_t(i)}, & (10)\\ \overline{\sigma}_i^2 &= \frac{\sum_{t=1}^{T}\gamma_t(i)(y_t-\overline{\mu}_i)^2}{\sum_{t=1}^{T}\gamma_t(i)}. & (11)\end{aligned}$$

3 Viterbi algorithm

Having found p, P, θ_1, $\ldots$, θ_s, one may need to find the best sequence of states, that is a sequence

$$i_1,\ldots,i_T \tag{12}$$

which maximizes

$$p_{i_1}f_{i_1}(y_1,\theta_{i_1})p_{i_1i_2}f_{i_2}(y_2,\theta_{i_2})\ldots p_{i_{T-1}i_T}f_{i_T}(y_T,\theta_{i_T}). \tag{13}$$

The Viterbi algorithm proceeds as follows. Let

$$\delta_1(i)=p_if_i(y_1,\theta_i), \quad \psi_1(i)=0, \quad i\in S.$$

For $t=2,\ldots,T$, let

$$\delta_t(j)=\max_{i\in S}(\delta_{t-1}(i)p_{ij})f_j(y_t,\theta_j), \quad \psi_t(j)=\operatorname*{argmax}_{i\in S}(\delta_{t-1}(i)p_{ij}), \quad i\in S.$$

Then the maximized probability (13) is equal to $\max_{i \in S} \delta_T(i)$ and the best sequence (12) can be backtracked by

$$i_T = \operatorname*{argmax}_{i \in S} \delta_T(i), \quad i_t = \psi_{t+1}(i_{t+1}), \quad t = T-1, \dots, 1.$$

4 Scaling

If the forward and backward variables are scaled, i. e.

$$\begin{aligned} \hat{\alpha}_1(j) &= c_1 \alpha_1(j), \quad j \in S, & (14)\\ \hat{\alpha}_t(j) &= c_t \sum_{i=1}^{s} (\hat{\alpha}_{t-1}(i) p_{ij}) f_j(y_t, \theta_j), \quad j \in S, \quad t = 2, \dots, T, & (15) \end{aligned}$$

and

$$\begin{aligned} \hat{\beta}_T(i) &= d_T \beta_T(i), \quad i \in S, & (16)\\ \hat{\beta}_t(i) &= d_t \sum_{j=1}^{s} p_{ij} f_j(y_{t+1}, \theta_j) \hat{\beta}_{t+1}(j), \quad i \in S, \quad t = T-1, \dots, 1 & (17) \end{aligned}$$

are calculated instead of (1–4), where

$$\begin{aligned} c_1^{-1} &= \sum_{j=1}^{s} \alpha_1(j), & (18)\\ c_t^{-1} &= \sum_{j=1}^{s} \sum_{i=1}^{s} (\hat{\alpha}_{t-1}(i) p_{ij}) f_j(y_t, \theta_j), \quad t = 2, \dots, T, & (19)\\ d_T^{-1} &= \sum_{i=1}^{s} \beta_T(i) = s, & (20)\\ d_t^{-1} &= \sum_{i=1}^{s} \sum_{j=1}^{s} p_{ij} f_j(y_{t+1}, \theta_j) \hat{\beta}_{t+1}(j), \quad t = T-1, \dots, 1, & (21) \end{aligned}$$

then

$$\begin{aligned} \hat{\alpha}_t(j) &= c_1 \dots c_t \alpha_t(j) = \frac{\alpha_t(j)}{\sum_{j=1}^{s} \alpha_t(j)}, & (22)\\ \hat{\beta}_t(i) &= d_T \dots d_t \beta_t(i) = \frac{\beta_t(i)}{\sum_{i=1}^{s} \beta_t(i)}, & (23) \end{aligned}$$

for $i, j \in S$, $t = 1, \dots, T$. The logarithm of likelihood may be calculated using the first equality (5) and (22) for $t = T$ as

$$\log \mathcal{L} = -\sum_{t=1}^{T} \log c_t, \qquad (24)$$

since $\sum_{i=1}^{s} \alpha_T(i) = (c_1 \dots c_T)^{-1}$. The values (6) and (7) may be calculated as

$$\gamma_t(i) = \frac{\hat{\alpha}_t(i)\hat{\beta}_t(i)}{\sum_{i=1}^{s} \hat{\alpha}_t(i)\hat{\beta}_t(i)}, \quad t = 1, \dots, T, \tag{25}$$

$$\xi_t(i,j) = d_t \frac{\hat{\alpha}_t(i)\hat{\beta}_{t+1}(j) p_{ij} f_j(y_{t+1}, \theta_j)}{\sum_{i=1}^{s} \hat{\alpha}_t(i)\hat{\beta}_t(i)}, \quad t = 1, \dots, T-1 \tag{26}$$

for $i, j \in S$.

The Baum-Welch adjustments (8–11) can be calculated as above except for (9), which should be calculated as

$$\overline{p}_{ij} = \frac{\sum_{t=1}^{T-1} \frac{\hat{\alpha}_t(i)\hat{\beta}_{t+1}(j) p_{ij} f_j(y_{t+1}, \theta_j)}{\sum_{i=1}^{s} \hat{\alpha}_t(i)\hat{\beta}_t(i)} d_t}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{27}$$

for $i, j \in S$.

The Viterbi algorithm needs not scaling, but what should be maximized is logarithm of (13) rather than (13) itself.

References: [3], [1], [2].

5 Forecast normal pseudo-residuals

Forecast normal pseudo-residuals are defined as follows [4, p. 97]. If X_t is a continuous random variable with distribution function F_{X_t}, then $F_{X_t}(X_t)$ is uniformly distributed on $(0, 1)$ and $u_t = P(X_t \leq x_t) = F_{X_t}(x_t)$ is the uniform pseudo-residual. The random variable $\Phi^{-1}(F_{X_t}(X_t))$ is distributed standard normal and

$$z_t = \Phi^{-1}(u_t) = \Phi^{-1}(F_{X_t}(x_t))$$

is the normal pseudo-residual. If we take

$$F_{X_t}(x_t) = P(X_t \leq x_t \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}),$$

we get forecast normal pseudo-residuals, while taking

$$F_{X_t}(x_t) = P(X_t \leq x_t \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}),$$

we get ordinary normal pseudo-residuals. Therefore, we calculate density of forecast according to formula

$$P(X_t = x \mid \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}) = \frac{\alpha_{t-1} \Gamma P(x) 1^T}{\alpha_{t-1} 1^T}.$$

References

[1] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[2] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, New York, 2005.

[3] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[4] Walter Zucchini and Iain L. MacDonald. *Hidden Markov Models for Time Series. An Introduction Using R*. Chapman and Hall/CRC, Boca Raton, 2009.