A New Approach for the Identification of Hidden Markov Models

Bart Vanluyten Jan C. Willems Bart De Moor

Department of Electrical Engineering
Katholieke Universiteit Leuven
Belgium

46the Conference on Decision and Control
12-14/12/2007

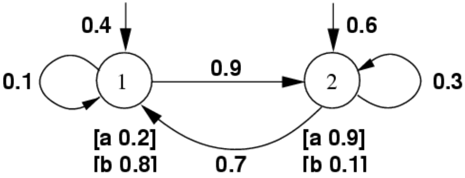
Outline

Outline

Problem statement

abababaababababbababbbabbbabbababababbbabababbbbbababaa...

⇓



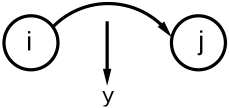
- Approximate identification of hidden Markov models
  - Baum-Welch
  - New technique inspired by subspace identification for LTI systems

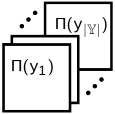
Outline

1. Problem statement
2. Hidden Markov models
3. Baum-Welch identification
4. Nonnegative matrix factorization
5. New identification procedure for HMMs
6. Simulation Example
7. Conclusions

Hidden Markov models

- Underlying state process and output process take values from the finite sets $\mathbb{X} = \{1, 2, \ldots, |\mathbb{X}|\}$ and $\mathbb{Y}$ with cardinality $|\mathbb{Y}|$
  - Mealy model $(\mathbb{X}, \mathbb{Y}, \Pi, \pi(1))$
    - $\Pi(\mathrm{y})_{ij} = P(x(t+1) = j, y(t) = \mathrm{y} | x(t) = i)$
    - $\pi(1)_i = P(x(1) = i)$





- Define $\Pi_{\mathbb{X}} := \sum_{\mathrm{y} \in \mathbb{Y}} \Pi(\mathrm{y})$
- Assume stationarity:

$$\pi(1)\Pi_{\mathbb{X}} = \pi(1)$$

Hidden Markov models

- Define $\mathbb{Y}^*$ as the set of finite words with symbols from the alphabeth $\mathbb{Y}$
- String probabilities $\mathcal{P} : \mathbb{Y}^* \mapsto [0,1]$

$$\mathcal{P}(\mathbf{u}) := P(y(1) = u_1, y(2) = u_2, \ldots, y(|\mathrm{y}|) = u_{|\mathrm{y}|})$$

- String probabilities generated by HMM $(\mathbb{X}, \mathbb{Y}, \Pi, \pi(1))$

$$\mathcal{P}(\mathbf{u}) = \pi(1)\Pi(\mathbf{u})e$$

where $\Pi(\mathbf{u}) := \Pi(u_1)\Pi(u_2)\ldots\Pi(u_{|\mathbf{u}|})$

- Identification problem
  - *Given*: an output string $y_1 y_2 \ldots y_T$ of length T
  - *Find*: HMM $(\mathbb{X}, \mathbb{Y}, \Pi, \pi(1))$ that models the output string approximately

Outline

Baum-Welch identification

- Classical approach based on Maximum Likelihood
- Solved by Expectation Maximization: recursive improvement of the model

Outline

Nonnegative matrix factorization (1)

- The new identification method makes use of the nonnegative matrix factorization
- Nonnegative matrix factorization:
  Given $M \in \mathbb{R}_+^{m_1 \times m_2}$, find $V \in \mathbb{R}_+^{m_1 \times a}$ and $H \in \mathbb{R}_+^{a \times m_2}$ such that $M = VH$ with a as small as possible
- Minimal $a =:$ positive rank of M
- Approximate nonnegative matrix factorization [Lee, Sueng]:
  Given $M \in \mathbb{R}_+^{m_1 \times m_2}$ and given a, minimize $D(M||VH)$ with respect to V and H, subject to the constraints $V, H \geq 0$
  where $D(A||B)$ is the Kullback-Leibler divergence defined as

$$D(A||B) = \sum_{ij} (A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij})$$

Nonnegative matrix factorization (2)

- The algorithm is a special (to preserve nonnegativity) gradient type algorithm

Theorem (Lee, Sueng)

The divergence $D(M||VH)$ is nonincreasing under the update rules

$$H_{il} \leftarrow H_{il} \frac{\sum_\mu V_{\mu i} \frac{M_{\mu l}}{(VH)_{\mu l}}}{\sum_\mu V_{\mu i}} \qquad V_{ki} \leftarrow V_{ki} \frac{\sum_\nu H_{i\nu} \frac{M_{k\nu}}{(VH)_{k\nu}}}{\sum_\nu H_{i\nu}}$$

(V,H) is invariant under these updates if and only if (V,H) is a stationary point of the divergence.

- Applications: e.g. in image representations

Outline

New identification procedure for HMMs

Idea behind ”subspace identification”:

1. Determine state sequence directly from data
2. Determine system parameters from state/output sequence

Proposed HMM identification method:

1. Estimate of the state sequence
   $\longrightarrow$ from the given output string
2. Estimate of the system matrices
   $\longrightarrow$ from the obtained state sequence and the given output sequence

Estimating the state sequence (1)

- Given the model, determination of state sequence corresponding to a particular output sequence
- Suppose we are given:
  - the matrix $M(i_1,i_2)$ of the underlying model $(\mathbb{X},\mathbb{Y},\Pi,\pi(1))$

$$M(i_1,i_2)_{kl} = \mathcal{P}(\mathbf{u}_k\mathbf{v}_l)$$

with $\mathbf{u}_k$ the k-th length-i_1 string and $\mathbf{v}_l$ the l-th length i_2 string

$$M(i_1,i_2) = \begin{bmatrix} \mathcal{P}(\mathbf{u}_1\mathbf{v}_1) & \mathcal{P}(\mathbf{u}_1\mathbf{v}_2) & \dots & \mathcal{P}(\mathbf{u}_1\mathbf{v}_{|\mathbb{Y}|^{i_2}}) \\ \mathcal{P}(\mathbf{u}_2\mathbf{v}_1) & \mathcal{P}(\mathbf{u}_2\mathbf{v}_2) & \dots & \mathcal{P}(\mathbf{u}_2\mathbf{v}_{|\mathbb{Y}|^{i_2}}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{P}(\mathbf{u}_{|\mathbb{Y}|^{i_2}}\mathbf{v}_1) & \mathcal{P}(\mathbf{u}_{|\mathbb{Y}|^{i_2}}\mathbf{v}_2) & \dots & \mathcal{P}(\mathbf{u}_{|\mathbb{Y}|^{i_2}}\mathbf{v}_{|\mathbb{Y}|^{i_2}}) \end{bmatrix}$$

Estimating the state sequence (2)

- the nonnegative decomposition of $M(i_1, i_2)$ given by

$$M(i_1, i_2) = VH \tag{1}$$

with

$$V = \begin{bmatrix} \pi(1)\Pi(\mathbf{u}_1) \\ \pi(1)\Pi(\mathbf{u}_2) \\ \vdots \\ \pi(1)\Pi(\mathbf{u}_{|\mathbb{Y}|^{i_1}}) \end{bmatrix}$$

$$H = \begin{bmatrix} \Pi(\mathbf{v}_1)e & \Pi(\mathbf{v}_2)e & \dots & \Pi(\mathbf{v}_{|\mathbb{Y}|^{i_2}})e \end{bmatrix}$$

- Rows of V can be seen to be

$$V_{k,i} = P(y(1,2,\dots,i_1) = \mathbf{u}_k, x(i_1+1) = i)$$

Estimating the state sequence (3)

- Using V, compute $P(x(t)|y(t-i_1)\dots y(t-1))$ for $t = i_1+1, i_1+2, \dots T$
- Now replace the probabilities by a maximum likelihood estimate $\hat{x}(t) \in \{0,1\}^{1\times|\mathbb{X}|}$

$$\begin{aligned} \hat{x}(t)_i &= 1 \\ \hat{x}(t)_j &= 0, \quad j = 1,\dots,i-1, i+1,\dots|\mathbb{X}| \end{aligned}$$

  where i is the most likely state estimate for $x(t)$ given the observations of $y(t-i_1)\dots y(t-1)$
- Define the estimated state sequence matrix $\hat{X}_{i_1} \in \{0,1\}^{(T-i_1)\times|\mathbb{X}|}$ as

$$\hat{X}_{i_1} = \begin{bmatrix} \hat{x}(i_1+1) \\ \hat{x}(i_1+2) \\ \vdots \\ \hat{x}(T) \end{bmatrix}$$

Estimating the state sequence (4)

- The next step is to estimate the state at time $t+1$ based on $y(t-i_1)\ldots y(t-1)$
- This is done in a similar way using $M(i_1+1, i_2)$
- We obtain

$$\hat{X}^+_{i_1+1} = \begin{bmatrix} \hat{x}^+(i_1+2) \\ \hat{x}^+(i_1+3) \\ \vdots \\ \hat{x}^+(T+1) \end{bmatrix}.$$

Estimating the string probabilities

- The matrices $M(i_1, i_2)$ and $M(i_1+1, i_2)$ contain string probabilities of length i_1+i_2 and i_1+i_2+1
- String probabilities can be estimated from the given output sequence yielding $\hat{M}(i_1, i_2)$ and $\hat{M}(i_1+1, i_2)$
- The nonnegative matrix factorization yields

$$\begin{aligned} \hat{M}(i_1, i_2) &= \hat{V}\hat{H} \\ \hat{M}(i_1+1, i_2) &= \hat{W}\hat{H} \end{aligned}$$

Estimating the system matrices

- We now calculate the system matrices using

$$P(x(t+1), y(t) = \mathrm{y} \mid y(t-i_1,\ldots,t-1) = u_{t-i_1}\ldots u_{t-1}) = P\left(x(t) \mid y(t-i_1,\ldots,t-1) = u_{t-i_1}\ldots u_{t-1}\right)\Pi(\mathrm{y}),$$

- Define

$$\hat{x}^{\mathrm{y}}(t+1) = \begin{cases} \hat{x}^{+}(t+1) & u_t = \mathrm{y} \\ \begin{bmatrix} 0 & 0 & \ldots & 0 \end{bmatrix} & \text{else} \end{cases}$$

- Now $\hat{\Pi}(\mathrm{y}), \mathrm{y} \in \mathbb{Y}$ can be calculated from the least squares problem

$$\begin{bmatrix} \hat{X}^{\mathrm{y}_1}_{i_1+1} & \hat{X}^{\mathrm{y}_2}_{i_1+1} & \ldots & \hat{X}^{\mathrm{y}_{|\mathbb{Y}|}}_{i_1+1} \end{bmatrix} = \hat{X}_{i_1} \begin{bmatrix} \hat{\Pi}(\mathrm{y}_1) & \hat{\Pi}(\mathrm{y}_2) & \ldots & \hat{\Pi}(\mathrm{y}_{|\mathbb{Y}|}) \end{bmatrix}$$

- It can be proven that the matrices $\hat{\Pi}(\mathrm{y}), \mathrm{y} \in \mathbb{Y}$ are elementwise nonnegative (due to ML approach)
- The initial state distribution $\hat{\pi}(1)$ is taken equal to the normalised left eigenvector of $\sum_{\mathrm{y}} \hat{\Pi}(\mathrm{y})$ corresponding to the eigenvalue 1

Outline

Simulation Example (1)

- Given an output string $y_1 \ldots y_{1000}$ generated by $\lambda_{\text{true}} = (\{1,2\},\{1,2\},\Pi_{\text{true}},\pi_{\text{true}}(1))$ where

$$\begin{aligned}\Pi_{\text{true}}(1) &= \begin{bmatrix} 0.20 & 0.40 \\ 0.00 & 0.20 \end{bmatrix} \\ \Pi_{\text{true}}(2) &= \begin{bmatrix} 0.10 & 0.30 \\ 0.80 & 0.00 \end{bmatrix} \\ \pi_{\text{true}} &= \begin{bmatrix} 0.53 & 0.47 \end{bmatrix}\end{aligned}$$

- Identification ($|\mathbb{X}_{\text{SS}}| = 2$) with new method with $i_1 = i_2 = 3$

$$\begin{aligned}\Pi_{\text{SS}}(1) &= \begin{bmatrix} 0.0699 & 0.2574 \\ 0.5651 & 0.0000 \end{bmatrix} \\ \Pi_{\text{SS}}(2) &= \begin{bmatrix} 0.1342 & 0.5386 \\ 0.4349 & 0.0000 \end{bmatrix} \\ \pi_{\text{SS}} &= \begin{bmatrix} 0.5568 & 0.4432 \end{bmatrix}\end{aligned}$$

- Identification ($|\mathbb{X}_{\text{BW}}| = 2$) with Baum-Welch

$$\begin{aligned}\Pi_{\text{BW}}(1) &= \begin{bmatrix} 0.0736 & 0.0986 \\ 0.5311 & 0.1415 \end{bmatrix} \\ \Pi_{\text{BW}}(2) &= \begin{bmatrix} 0.0751 & 0.7526 \\ 0.2424 & 0.0850 \end{bmatrix} \\ \pi_{\text{BW}} &= \begin{bmatrix} 0 & 1 \end{bmatrix}\end{aligned}$$

Simulation Example (2)

- Distance between HMMs

$$D(\lambda_{\text{true}}||\lambda_{\text{approx}}) = \sum_{\mathbf{u}\in\mathbb{Y}^*} \mathcal{P}(\mathbf{u}|\lambda_{\text{true}}) \log \frac{\mathcal{P}(\mathbf{u}|\lambda_{\text{true}})}{\mathcal{P}(\mathbf{u}|\lambda_{\text{approx}})}$$

- In practice

$$D_\Delta(\lambda_{\text{true}}||\lambda_{\text{approx}}) = \sum_{\mathbf{u}\in\mathbb{Y}^*, |\mathbf{u}|\leq\Delta} \mathcal{P}(\mathbf{u}|\lambda_{\text{true}}) \log \frac{\mathcal{P}(\mathbf{u}|\lambda_{\text{true}})}{\mathcal{P}(\mathbf{u}|\lambda_{\text{approx}})}$$

- Results for $\Delta = 8$

$$\begin{aligned} D_\Delta(\lambda_{\text{true}}||\lambda_{\text{SS}}) &= 0.3876 \\ D_\Delta(\lambda_{\text{true}}||\lambda_{\text{BW}}) &= 1.8955 \end{aligned}$$

Outline

Conclusions

- New identification procedure for hidden Markov models in two steps
  - Calculate the state sequence directly from output sequence
  - Estimate system matrices from output sequence and estimated state sequence
- Method gives good results on a simulation example