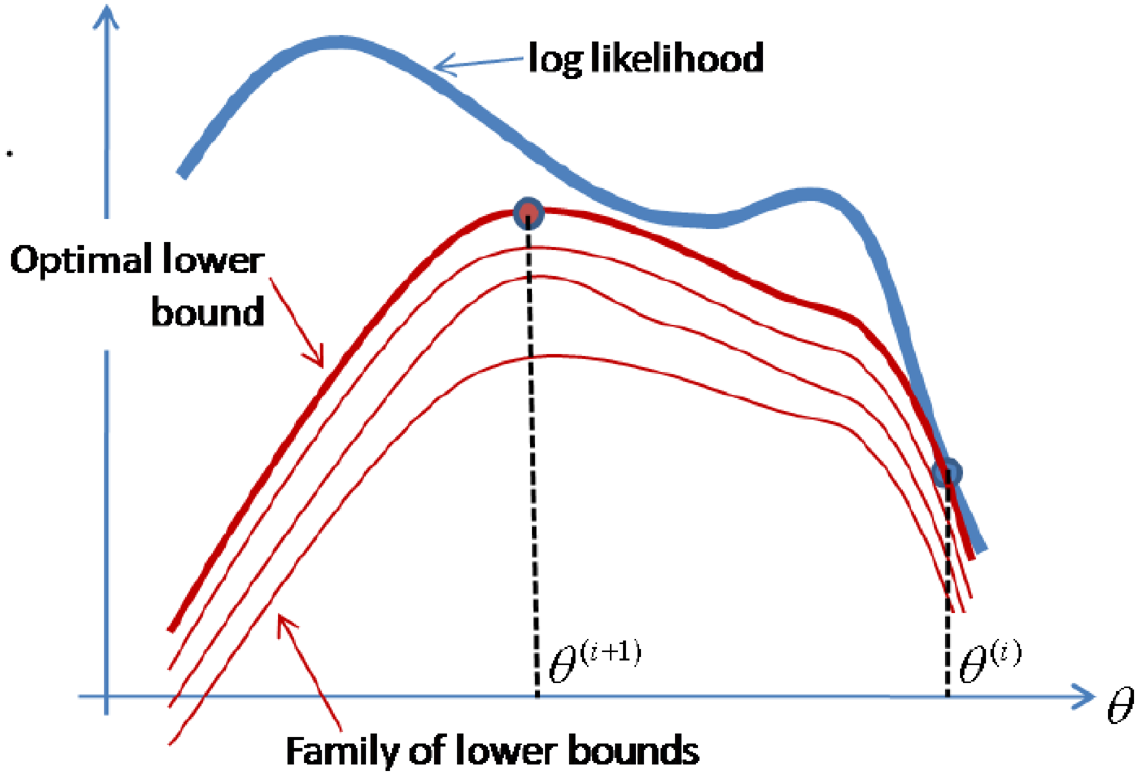
Christof Schütte

DFG Research Center Matheon
Mathematics for key technologies

Freie Universität Berlin

Successive lower bound maximization

$$B(\theta) \leq \log \mathcal{L}(\theta \mid x), \quad \forall \theta.$$



Algorithm 3 General EM-algorithm

Input: Time series $x = \{x_0 = X(t_0), \ldots, x_N = X(t_N)\}$, initial guess of parameters $\theta^{(0)}$, tolerance tol.

Output: Maximum likelihood estimate $\hat{\Theta}$.

(1) Set $i = 0$.

(2) $i := i + 1$.

(3) Expectation step (E-step):

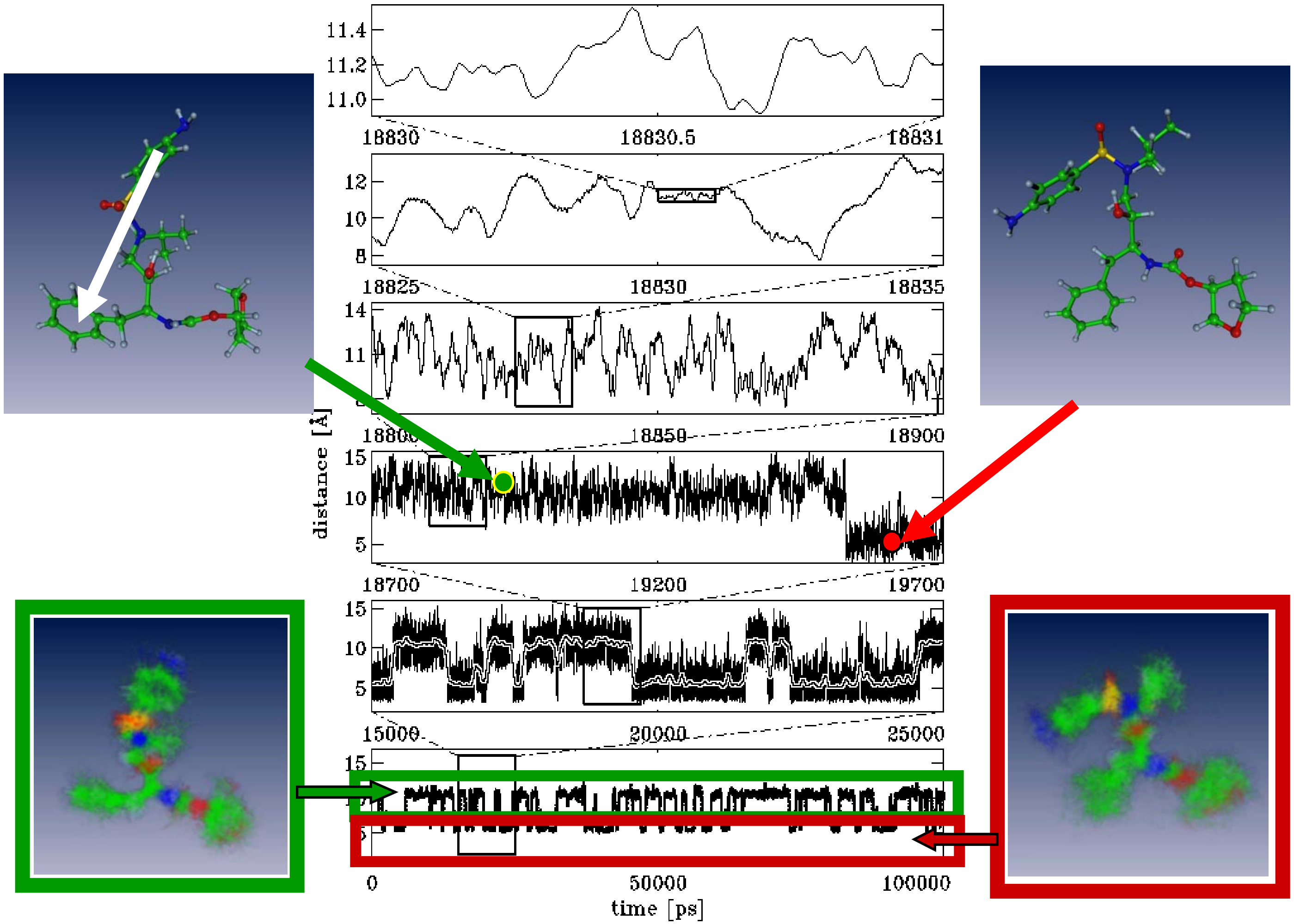
Compute the function $\mathcal{G}(\theta, \theta^{(i)})$.

(4) Maximization step (M-Step):

$\theta^{(i+1)} = \arg\max_\theta \mathcal{G}(\theta, \theta^{(i)})$

(5) If $\Delta\mathcal{L}^{(i)} > \text{tol}$, go to Step (2). Otherwise, terminate.

$$\begin{aligned}
\mathcal{G}(\theta, \theta^{(i)}) &= E_{p(Y|x,\theta^{(i)})}\left[\log p(x, Y \mid \theta)\right] \\
&= \int_{S_Y} p(y \mid x, \theta^{(i)}) \log p(x, y \mid \theta) dy.
\end{aligned}$$


11.4
11.2
11.0
18830
18830.5
18831
12
10
8
18825
18830
18835
14
11
18800
18850
18900
15
10
5
18700
19200
19700
15
10
5
15000
20000
25000
15
0
50000
100000
distance [Å]
time [ps]

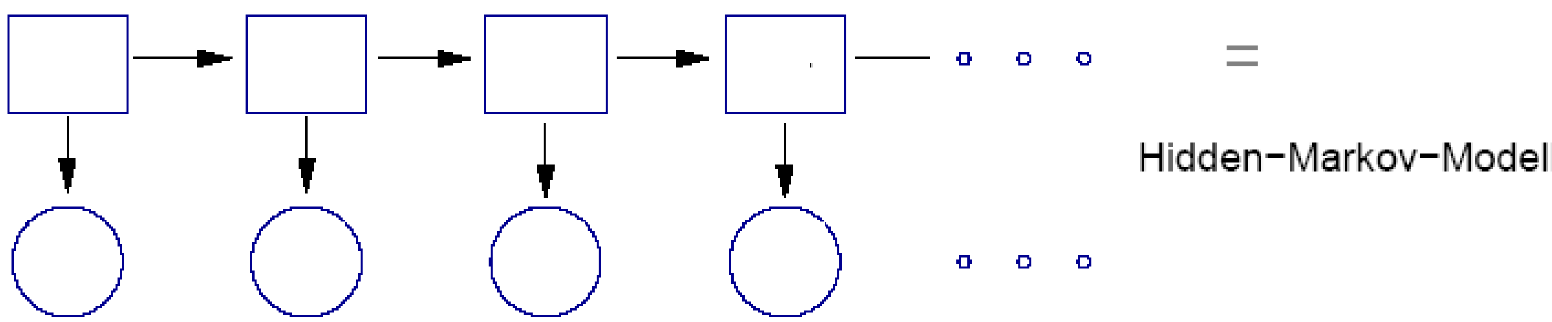
state (observable state , hidden state)

KxK transition matrix between metastable states

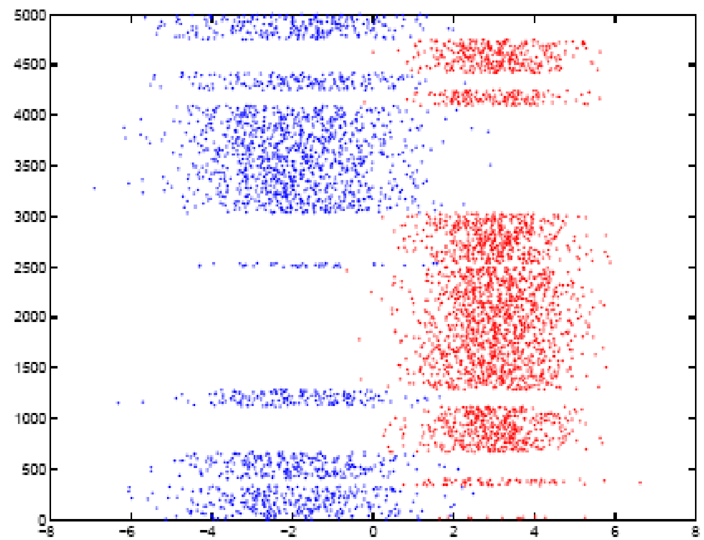
b: probability density for output from hidden state

e.g., Gaussian distribution

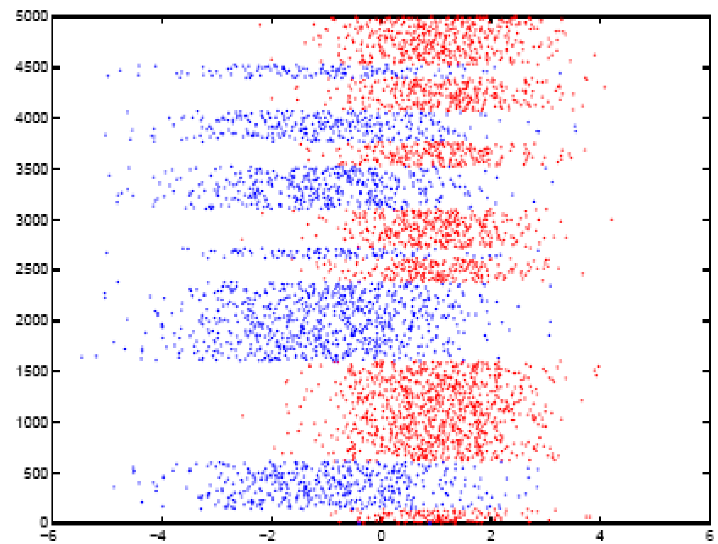
probability distribution of initial state



	Modellparameter
	$\begin{pmatrix} 0.997 & 0.003 \\ 0.002 & 0.998 \end{pmatrix}$
b_1	$\mathcal{N}(-2, 1.5)$
b_2	$\mathcal{N}(3, 1)$



	Modellparameter
	$\begin{pmatrix} 0.997 & 0.003 \\ 0.002 & 0.998 \end{pmatrix}$
b_1	$\mathcal{N}(-1, 1.5)$
b_2	$\mathcal{N}(1, 1)$



HMM parameters:

$$
\begin{aligned}
b(x;\vartheta_j) &= p(X_t = x \mid Y_t = j), \\
p_{ij} &= p(Y_t = j \mid Y_{t-1} = i), \\
\rho_j &= p(Y_0 = j).
\end{aligned}
$$

$$
\theta = \Big(\vartheta_j, p_{ij}, \rho_j\Big) = \Big(\vartheta_1, \ldots, \vartheta_K, p_{11}, \ldots, p_{KK}, \rho_1, \ldots, \rho_K\Big).
$$

$$
\begin{aligned}
p(X_t \mid Y_t, X_{1:t-1}, X_{t+1:T}, Y_{0:t-1}, Y_{t+1:T}) &= p(X_t \mid Y_t), \\
p(Y_t \mid X_{1:t-1}, Y_{0:t-1}) &= p(Y_t \mid Y_{t-1}), \\
p(X_{t:t+h} \mid Y_{t:t+h}) &= \prod_{s=t}^{t+h} p(X_s \mid Y_s).
\end{aligned}
$$

$$p(X_{1:t}, Y_{0:t}) \overset{(CR)}{=} p(X_t \mid Y_t, X_{1:t-1}, Y_{0:t-1}) \cdot p(Y_t \mid X_{1:t-1}, Y_{0:t-1}) \cdot p(X_{1:t-1}, Y_{0:t-1}),$$

$$= p(X_t \mid Y_t) \cdot p(Y_t \mid Y_{t-1}) \cdot p(X_{1:t-1}, Y_{0:t-1}).$$

$$= \prod_{s=1}^{t} p(X_s \mid Y_s) \cdot p(Y_s \mid Y_{s-1}) \cdot p(Y_0).$$

$$\mathcal{L} = \sum_{j_{0:T}} p(X_{1:T}, Y_{0:T} = j_{0:T})$$

$$= \sum_{j_{0:T}} \prod_{s=1}^{T} p(X_s \mid Y_s = j_s) \cdot p(Y_s = j_s \mid Y_{s-1} = j_{s-1}) \cdot p(Y_0 = j_0)$$

HMM Likelihood

$$
\begin{aligned}
\mathcal{L} &= \sum_{j_{0:T}} p(X_{1:T}, Y_{0:T} = j_{0:T}) \\
&= \sum_{j_{0:T}} \prod_{s=1}^{T} p(X_s \mid Y_s = j_s) \cdot p(Y_s = j_s \mid Y_{s-1} = j_{s-1}) \cdot p(Y_0 = j_0)
\end{aligned}
$$

$$
\sum_{j_{0:T}} = \sum_{j_0=1}^{K} \sum_{j_1=1}^{K} \cdots \sum_{j_T=1}^{K}
$$

HMM parameters:

$$
\begin{aligned}
b(x; \vartheta_j) &= p(X_t = x \mid Y_t = j), \\
p_{ij} &= p(Y_t = j \mid Y_{t-1} = i), \\
\rho_j &= p(Y_0 = j).
\end{aligned}
$$

$$
\mathcal{L}(\theta) = \mathcal{L}(\theta; X_{1:T} = x_{1:T}) = \sum_{j_{0:T}} \prod_{s=1}^{T} b(x_s; \vartheta_{j_s}) \cdot p_{j_{s-1}, j_s} \cdot \rho_{j_0},
$$

$$\begin{aligned}\mathcal{G}(\theta, \theta^{(i)}) &= E_{p(Y_{0:T}|X_{1:T},\theta^{(i)})}\left[\log p(X_{1:T}, Y_{0:T} \mid \theta)\right] \\ &= \sum_{j_{0:T}} p(Y_{0:T} = j_{0:T} \mid X_{1:T}, \theta^{(i)}) \cdot \log p(X_{1:T}, Y_{0:T} = j_{0:T} \mid \theta))\end{aligned}$$

$$\begin{aligned}\gamma_j(t) &= p(Y_t = j \mid X_{1:T}, \theta) \\ \xi_{ij}(t) &= p(Y_{t+1} = j, Y_t = i \mid X_{1:T}, \theta) \\ \eta_{ij}(t) &= p(Y_{t+1} = j \mid Y_t = i, X_{1:T}, \theta).\end{aligned}$$

$$\begin{aligned}
\gamma_j(t) &= p(Y_t = j \mid X_{1:T}, \theta) \\
\xi_{ij}(t) &= p(Y_{t+1} = j, Y_t = i \mid X_{1:T}, \theta) \\
\eta_{ij}(t) &= p(Y_{t+1} = j \mid Y_t = i, X_{1:T}, \theta).
\end{aligned}$$

$$\begin{aligned}
\sum_i \gamma_i(t) &= 1 \\
\sum_j \xi_{ij}(t) &= \gamma_i(t) \\
\sum_i \xi_{ij}(t) &= \gamma_j(t+1) \\
\sum_j \eta_{ij}(t) &= 1 \\
\xi_{ij}(t) &= \eta_{ij}(t)\gamma_i(t),
\end{aligned}$$

$$\mathcal{G}(\theta,\theta^{(i)}) = E_{p(Y_{0:T}|X_{1:T},\theta^{(i)})}\left[\log p(X_{1:T},Y_{0:T}\mid\theta)\right]$$
$$= \sum_{j_{0:T}} p(Y_{0:T}=j_{0:T}\mid X_{1:T},\theta^{(i)}) \cdot \log p(X_{1:T},Y_{0:T}=j_{0:T}\mid\theta)$$

$$\begin{aligned}
p(Y_{0:T}=j_{0:T}\mid X_{1:T},\theta) &= p(Y_{1:T}=j_{1:T}\mid Y_0=j_0,X_{1:T},\theta)\cdot p(Y_0=j_0\mid X_{1:T},\theta)\\
&= p(Y_{1:T}=j_{1:T}\mid Y_0=j_0,X_{1:T},\theta)\cdot \gamma_0(j_0)\\
&= p(Y_{2:T}=j_{2:T}\mid Y_0=j_0,Y_1=j_1,X_{1:T},\theta)\\
&\quad \cdot p(Y_1=j_1\mid Y_0=j_0,X_{1:T},\theta)\cdot \gamma_0(j_0)\\
&= p(Y_{2:T}=j_{2:T}\mid Y_1=j_1,X_{1:T},\theta)\cdot \eta_{j_0,j_1}(0)\cdot \gamma_0(j_0)\\
&= \prod_{t=1}^{T}\eta_{j_{t-1},j_t}(t)\cdot \gamma_0(j_0)
\end{aligned}$$

G-function for E-step

$$\mathcal{G}(\theta,\theta^{(i)}) = E_{p(Y_{0:T}|X_{1:T},\theta^{(i)})}\left[\log p(X_{1:T}, Y_{0:T} \mid \theta)\right]$$

$$= \sum_{j_{0:T}} p(Y_{0:T}=j_{0:T} \mid X_{1:T},\theta^{(i)}) \cdot \log p(X_{1:T}, Y_{0:T}=j_{0:T} \mid \theta)$$

$$p(X_{1:T}=x_{1:T}, Y_{0:t}=j_{0:t}) = \prod_{s=1}^{T} b(x_s;\vartheta_{j_s}) \cdot p_{j_{s-1},j_s} \cdot \rho_{j_0}.$$

$$= \prod_{t=1}^{T} \eta_{j_{t-1},j_t}(t) \cdot \gamma_0(j_0)$$

Lemma *When using a superindex on the ξ variables in order to indicate that these variables correspond to the parameter set $\theta^{(i)}$, and keeping in mind that the p_{ij} and b_j come from the parameter set θ, we get*

$$\begin{aligned}
\mathcal{G}(\theta, \theta^{(i)}) &= \mathcal{G}_1(\theta, \theta^{(i)}) + \mathcal{G}_2(\theta, \theta^{(i)}) + \mathcal{G}_3(\theta, \theta^{(i)}) \\
\mathcal{G}_1(\theta, \theta^{(i)}) &= \sum_{t=0}^{T-1} \sum_{k,l} \xi_{k,l}^{(i)}(t) \cdot \log p_{k,l}, \\
\mathcal{G}_2(\theta, \theta^{(i)}) &= \sum_{t=0}^{T-1} \sum_{l} \gamma_l^{(i)}(t+1) \cdot \log b(x_{t+1}; \vartheta_l), \\
\mathcal{G}_3(\theta, \theta^{(i)}) &= \sum_{k} \gamma_k^{(i)}(0) \cdot \log \rho_k.
\end{aligned}$$

General M-step problem:

Corollary *Let* $\lambda = (\lambda_1, \ldots, \lambda_K)$, *and* κ *be Lagrange multipliers. The M-step of the EM-algorithms has to maximize wrt.* $\theta = (p_{ij}, \vartheta_j)$, κ, *and* λ:

$$\mathcal{G}_c(\theta, \kappa, \lambda; \theta^{(i)}) = \mathcal{G}(\theta, \theta^{(i)}) + \sum_k \lambda_k (\sum_l p_{kl} - 1) + \kappa (\sum_k \rho_k - 1) \quad (50)$$

under additional non-negativity constraints (49). This maximization problem can be decomposed into two independent problems:

$$\begin{aligned}
\mathcal{G}_c(\theta, \lambda; \theta^{(i)}) &= \mathcal{G}_{c,1}(p_{ij}, \lambda; \theta^{(i)}) + \mathcal{G}_2(\vartheta_j; \theta^{(i)}) + \mathcal{G}_{c,3}(\rho_j, \kappa; \theta^{(i)}) \\
\mathcal{G}_{c,1}(p_{ij}, \lambda; \theta^{(i)}) &= \sum_{t=0}^{T-1} \sum_{k,l} \xi_{kl}^{(i)}(t) \cdot \log p_{kl} + \sum_k \lambda_k (\sum_l p_{kl} - 1). \\
\mathcal{G}_2(\vartheta_j; \theta^{(i)}) &= \sum_{t=1}^{T} \sum_l \gamma_l^{(i)}(t) \cdot \log b(x_t; \vartheta_l), \\
\mathcal{G}_{c,3}(\rho_j, \kappa; \theta^{(i)}) &= \sum_l \gamma_l^{(i)}(0) \cdot \log \rho_l + \kappa (\sum_l \rho_l - 1),
\end{aligned}$$

where $\mathcal{G}_{c,1}$, $\mathcal{G}_2$ *and* $\mathcal{G}_{c,3}$ *have to be maximized on their respective orthogonal subspaces under the respective additional non-negativity constraints for* $\mathcal{G}_{c,1}$, *and* $\mathcal{G}_{c,3}$.

Result of M-step for Gaussian HMMs

$$b(x;\vartheta_j) = G(x;\mu_j,\Sigma_j).$$

$$\hat{\rho}_r = \frac{\gamma_r^{(i)}(0)}{\sum_r \gamma_r^{(i)}(0)}.$$

$$\hat{p}_{rq} = \frac{\sum_{t=0}^{T-1} \xi_{rq}^{(i)}(t)}{\sum_{t=0}^{T-1} \gamma_r^{(i)}(t)}.$$

$$\hat{\mu}_r = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t) x_t$$

$$\hat{\Sigma}_r = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t)(x_t-\mu_r)(x_t-\mu_r)^\top.$$

$$\gamma_r^{(i)} = \sum_{t=1}^{T} \gamma_r^{(i)}(t).$$

Result of M-step for Gaussian HMMs

$$b(x;\vartheta_j) = G(x;\mu_j,\Sigma_j).$$

$$\hat{\rho}_r = \frac{\gamma_r^{(i)}(0)}{\sum_r \gamma_r^{(i)}(0)}.$$

$$\hat{p}_{rq} = \frac{\sum_{t=0}^{T-1} \xi_{rq}^{(i)}(t)}{\sum_{t=0}^{T-1} \gamma_r^{(i)}(t)}.$$

$$\hat{\mu}_r = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t) x_t$$

$$\hat{\Sigma}_r = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t)(x_t-\mu_r)(x_t-\mu_r)^\top.$$

Done??

$$\gamma_r^{(i)} = \sum_{t=1}^{T} \gamma_r^{(i)}(t).$$

Direct evaluation infeasible

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; X_{1:T} = x_{1:T}) = \sum_{j_{0:T}} \prod_{s=1}^{T} b(x_s; \vartheta_{j_s}) \cdot p_{j_{s-1},j_s} \cdot \rho_{j_0},$$

Using this formula:

Computation of likelihood requires operations

Similarly bad scaling for etc.

Forward-Backward

$$\alpha_j(t) = p(Y_t = j,\ X_{1:t}),$$
$$\beta_j(t) = p(X_{t+1:T} \mid Y_t = j),$$

Lemma *The following three formula allow to compute* $\gamma_j(t)$, $\xi_{jk}(t)$, *and the likelihood* $\mathcal{L}$ *from the forward-backward variables and the respective parameter set* p_{jk} *and* ϑ_k*:*

$$\mathcal{L} = \sum_t \sum_j \alpha_j(t)\beta_j(t)$$

$$\gamma_j(t) = \frac{\alpha_j(t)\beta_j(t)}{\sum_j \alpha_j(t)\beta_j(t)}$$

$$\xi_{jk}(t) = \frac{\alpha_j(t) p_{jk}\, b(x_{t+1};\vartheta_k)\beta_k(t+1)}{\sum_{jk} \alpha_j(t) p_{jk} b(x_{t+1};\vartheta_k)\beta_k(t+1)}.$$

The forward-backward variables satisfy the following recursion formula:

$$\alpha_j(t) = b(X_t, \vartheta_j) \sum_k p_{kj}\, \alpha_k(t-1),$$

$$\beta_j(t) = \sum_k b(X_{t+1}, \vartheta_k)\, p_{jk}\, \beta_k(t+1),$$

where the forward recursion for $\alpha_j(t)$ *has the initial condition*

$$\alpha_j(1) = b(X_1, \vartheta_j) \cdot \rho_j,$$

while the backward recursion for $\beta_j(t)$ *starts with*

$$\beta_j(T) = 1.$$

The forward-backward variables satisfy the following recursion formula:

$$\alpha_j(t) = b(X_t, \vartheta_j) \sum_k p_{kj}\, \alpha_k(t-1),$$

$$\beta_j(t) = \sum_k b(X_{t+1}, \vartheta_k)\, p_{jk}\, \beta_k(t+1),$$

where the forward recursion for $\alpha_j(t)$ *has the initial condition*

$$\alpha_j(1) = b(X_1, \vartheta_j) \cdot \rho_j,$$

while the backward recursion for $\beta_j(t)$ *starts with*

$$\beta_j(T) = 1.$$

Algorithm 5 HMM-Gauss: EM-algorithm for HMM with Gaussian output probabilities

Input: Time series $X_{1:T} = x_{1:T}$, initial guess of parameters

$$\theta^{(0)} = (p_{kl}^{(0)}, \vartheta_k^{(0)}, \rho_k^{(0)}), \quad \vartheta_k^{(0)} = (\mu_k^{(0)}, \Sigma_k^{(0)}), \quad k, l = 1, \ldots, K,$$

tolerance tol.

Output: Maximum likelihood estimate $\hat{\Theta}$.

(1) Set $i = -1$. Set formally $\mathcal{L}^{(-1)} = 0$.

(2) $i := i + 1$.

(3) Expectation step (E-step):

(3.1) Compute forward-backward variables $\alpha_j^{(i)}(t)$ and $\beta_j^{(i)}(t)$ via forward-backward recursion as given in Lemma 4.4 using parameter set $(p_{kl}^{(i)}, \vartheta_k^{(i)}, \rho_k^{(i)})$

$$\alpha_j(t) = b(X_t, \vartheta_j) \sum_k p_{kj}\, \alpha_k(t-1),$$

$$\beta_j(t) = \sum_k b(X_{t+1}, \vartheta_k)\, p_{jk}\, \beta_k(t+1),$$

(3.3) Compute

$$\mathcal{L} = \sum_t \sum_j \alpha_j(t)\beta_j(t)$$

$$\gamma_j(t) = \frac{\alpha_j(t)\beta_j(t)}{\sum_j \alpha_j(t)\beta_j(t)}$$

$$\xi_{jk}(t) = \frac{\alpha_j(t) p_{jk}\, b(x_{t+1}; \vartheta_k)\beta_k(t+1)}{\sum_{jk} \alpha_j(t) p_{jk} b(x_{t+1}; \vartheta_k)\beta_k(t+1)}.$$

(4) Convergence check:
If $\mathcal{L}^{(i)} - \mathcal{L}^{(i-1)} > \text{tol}$, go to Step (5). Otherwise, terminate and set $\hat{\Theta} = \theta^{(i)}$.

(5) Maximization step (M-Step):
Compute the new optimal parameter estimates

$$\hat{p}_{rq}^{(i+1)} = \frac{\sum_{t=0}^{T-1} \xi_{rq}^{(i)}(t)}{\sum_{t=0}^{T-1} \gamma_r^{(i)}(t)},$$

$$\hat{\rho}_r^{(i+1)} = \frac{\gamma_r^{(i)}(0)}{\sum_r \gamma_r^{(i)}(0)},$$

$$\hat{\mu}_r^{(i+1)} = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t) x_t$$

$$\hat{\Sigma}_r^{(i+1)} = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t)(x_t - \mu_r)(x_t - \mu_r)^\top.$$

(6) Go to Step (2).

EM algorithm for Gaussian HMMs

Algorithm 5 HMM-Gauss: EM-algorithm for HMM with Gaussian output probabilities

Input: Time series $X_{1:T} = x_{1:T}$, initial guess of parameters

$$\theta^{(0)} = (p_{kl}^{(0)}, \vartheta_k^{(0)}, \rho_k^{(0)}), \quad \vartheta_k^{(0)} = (\mu_k^{(0)}, \Sigma_k^{(0)}), \quad k, l = 1, \ldots, K,$$

tolerance tol.

Output: Maximum likelihood estimate $\hat{\Theta}$.

(1) Set $i = -1$. Set formally $\mathcal{L}^{(-1)} = 0$.

(2) $i := i + 1$.

(3) Expectation step (E-step):

(3.1) Compute forward-backward variables $\alpha_j^{(i)}(t)$ and $\beta_j^{(i)}(t)$ via forward-backward recursion as given in Lemma 4.4 using parameter set $(p_{kl}^{(i)}, \vartheta_k^{(i)}, \rho_k^{(i)})$

$$\alpha_j(t) = b(X_t, \vartheta_j) \sum_k p_{kj}\, \alpha_k(t-1),$$

$$\beta_j(t) = \sum_k b(X_{t+1}, \vartheta_k)\, p_{jk}\, \beta_k(t+1),$$

(3.3) Compute

$$\mathcal{L} = \sum_t \sum_j \alpha_j(t)\beta_j(t)$$

$$\gamma_j(t) = \frac{\alpha_j(t)\beta_j(t)}{\sum_j \alpha_j(t)\beta_j(t)}$$

$$\xi_{jk}(t) = \frac{\alpha_j(t) p_{jk}\, b(x_{t+1}; \vartheta_k)\beta_k(t+1)}{\sum_{jk} \alpha_j(t) p_{jk} b(x_{t+1}; \vartheta_k)\beta_k(t+1)}.$$

(4) Convergence check:
If $\mathcal{L}^{(i)} - \mathcal{L}^{(i-1)} > \text{tol}$, go to Step (5). Otherwise, terminate and set $\hat{\Theta} = \theta^{(i)}$.

(5) Maximization step (M-Step):
Compute the new optimal parameter estimates

$$\hat{p}_{rq}^{(i+1)} = \frac{\sum_{t=0}^{T-1} \xi_{rq}^{(i)}(t)}{\sum_{t=0}^{T-1} \gamma_r^{(i)}(t)},$$

$$\hat{\rho}_r^{(i+1)} = \frac{\gamma_r^{(i)}(0)}{\sum_r \gamma_r^{(i)}(0)},$$

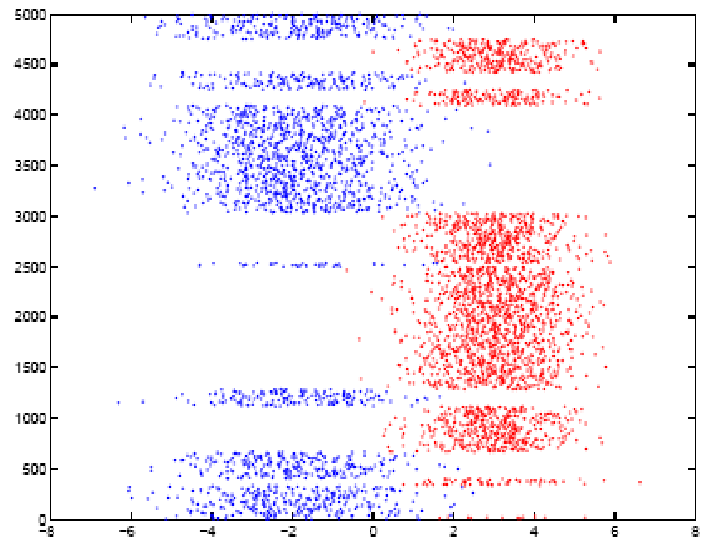
$$\hat{\mu}_r^{(i+1)} = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t) x_t$$

$$\hat{\Sigma}_r^{(i+1)} = \frac{1}{\gamma_r^{(i)}} \sum_{t=1}^{T} \gamma_r^{(i)}(t)(x_t - \mu_r)(x_t - \mu_r)^\top.$$

(6) Go to Step (2).

Hidden Markov Models (HMMs): Example

	Startwerte EM	Ausgabe EM	Modellparameter
	$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.9969 & 0.0031 \\ 0.0026 & 0.9974 \end{pmatrix}$	$\begin{pmatrix} 0.997 & 0.003 \\ 0.002 & 0.998 \end{pmatrix}$
b_1	$\mathcal{N}(-3, 2)$	$\mathcal{N}(-1.9, 1.50)$	$\mathcal{N}(-2, 1.5)$
b_2	$\mathcal{N}(3, 2)$	$\mathcal{N}(3.01, 1.01)$	$\mathcal{N}(3, 1)$



	Startwerte EM	Ausgabe EM	Modellparameter
	$\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.9967 & 0.0033 \\ 0.0023 & 0.9977 \end{pmatrix}$	$\begin{pmatrix} 0.997 & 0.003 \\ 0.002 & 0.998 \end{pmatrix}$
b_1	$\mathcal{N}(-3, 2)$	$\mathcal{N}(-0.995, 1.47)$	$\mathcal{N}(-1, 1.5)$
b_2	$\mathcal{N}(3, 2)$	$\mathcal{N}(1.01, 0.998)$	$\mathcal{N}(1, 1)$