An online EM algorithm in hidden (semi-)Markov models for audio segmentation and clustering

Alberto Bietti[1,2,3], Francis Bach[2,4], Arshia Cont[2,3]

[1]Quora, Inc. [2]INRIA [3]Ircam [4]Ecole Normale Supérieure

ICASSP 2015. April 2015, Brisbane.

Outline

1. Audio segmentation with hidden (semi-)Markov models

2. Online EM algorithm

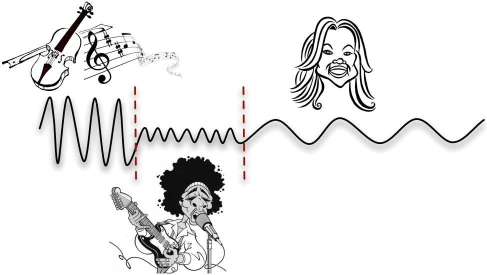
3. Experiments

Outline

1. Audio segmentation with hidden (semi-)Markov models
2. Online EM algorithm
3. Experiments

Audio segmentation

- **Goal**: segment audio signal into homogeneous chunks/segments
- Go from a signal representation to a symbolic representation
- Applications: music indexing, summarization, fingerprinting



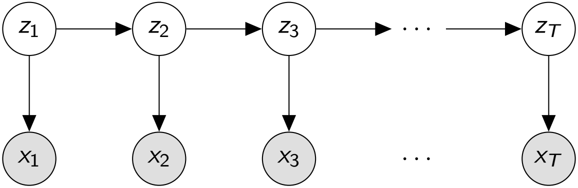
Audio segmentation: approach

- Most existing approaches: change-point detection, compute similarities separately
- Real-time approaches mainly for change detection, no clustering

Audio segmentation: approach

- Most existing approaches: change-point detection, compute similarities separately
- Real-time approaches mainly for change detection, no clustering
- **Our goal**: joint segmentation and clustering, unsupervised learning, online/real-time
- Online learning in **Hidden (semi-)Markov Models**

Hidden Markov Models (HMMs)



- K hidden states (cluster/segment IDs)
- Hidden chain of cluster ids: $z_{1:T} = (z_1, \ldots, z_T) \in \{1, \ldots, K\}^T$
  - Transition matrix: $A \in \mathbb{R}^{K \times K}$: $A_{ij} = p(z_t = j | z_{t-1} = i)$
- Sequence of observations $x_{1:T} = (x_1, \ldots, x_T)$, with $x_t \in \mathbb{R}^p$
  - Emission distribution in state i: $p(x_t | z_t = i; \mu_i)$

Audio representation and emission distributions

- $x_t = N|\hat{x}_t|/\|\hat{x}_t\|_1$
  - $|\hat{x}_t| \in \mathbb{R}_+^p$ from STFT
  - Normalized magnitude for invariance to volume
- Emission distributions: multinomials (N trials)
  - Parameterized by mean $\mu_i = \mathbb{E}[x|z = i]$
  - Corresponds to KL divergence, which performs well on audio

$$p(x|z = i) = h(x)\exp(-D_{KL}(x||\mu_i))$$

- Extends to other Bregman divergences (Banerjee et al., 2005) and exponential families

Duration distributions

- HMM
  - Segment length distributions are geometric:

$$p_i(d) = A_{ii}^{d-1}(1 - A_{ii})$$

  - Duration distribution learned implicitely through A_{ii}
- HSMM (explicit-duration HMM)
  - Model these duration distributions explicitely (e.g., Negative Binomial, Poisson)
  - Can help avoid short segments, encourage specific durations
  - Segment = (state i, length l), with $l \sim p_i(d)$
  - (Markov) transitions A_{ij} between segments
  - i.i.d. observations in each segment given state

Outline

1. Audio segmentation with hidden (semi-)Markov models

2. Online EM algorithm

3. Experiments

EM algorithm

- $\mathbf{x} = x_{1:T}$ observed variables, $\mathbf{z} = z_{1:T}$ hidden variables, θ parameter
- Goal: maximum likelihood $\max_\theta p(\mathbf{x};\theta)$

$$\log p(\mathbf{x};\theta) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})} \geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})} =: \hat{f}_q(\theta).$$

- E-step: maximize w.r.t. q. (forward-backward for H(S)MMs)

$$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta)$$

- M-step: maximize w.r.t. θ.

$$\hat{\theta} = \arg\max_\theta \mathbb{E}_q[\log p(\mathbf{z},\mathbf{x};\theta)]$$

EM algorithm

- $\mathbf{x} = x_{1:T}$ observed variables, $\mathbf{z} = z_{1:T}$ hidden variables, θ parameter
- Goal: maximum likelihood $\max_\theta p(\mathbf{x}; \theta)$

$$\log p(\mathbf{x}; \theta) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} =: \hat{f}_q(\theta).$$

- E-step: maximize w.r.t. q. (forward-backward for H(S)MMs)

$$q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x}; \theta)$$

- M-step: maximize w.r.t. θ.

$$\hat{\theta} = \arg\max_\theta \mathbb{E}_q[\log p(\mathbf{z}, \mathbf{x}; \theta)]$$

- Incremental EM for i.i.d. observations (Neal and Hinton, 1998)
  - Partial E-step: update q on a single observation

Incremental EM for HMMs

- Maximizer q takes the form:

$$p(z_{1:T}|x_{1:T};\theta) = p(z_1|x_{1:T}) \prod_{t\geq 2} p(z_t|z_{t-1}, x_{1:T})$$

- Consider q of the form:

$$q(z_{1:T}) = q_1(z_1) \prod_{t\geq 2} q_t(z_t|z_{t-1})$$

- Lower bound:

$$\begin{aligned}\hat{f}_T(\theta) &= \mathbb{E}_q \left[\log \frac{p_\theta(x_{1:T}, z_{1:T})}{q(z_{1:T})}\right] \\ &= \mathbb{E}_{q_1} \left[\log \frac{p_\theta(x_1, z_1)}{q_1(z_1)}\right] + \sum_{t=2}^{T} \mathbb{E}_q \left[\log \frac{p_\theta(x_t, z_t|z_{t-1})}{q_t(z_t|z_{t-1})}\right]\end{aligned}$$

Incremental EM for HMMs

Marginals:

- $\phi_t(z_t) = \sum_{z_{1:t-1}} q_1(z_1) \dots q_t(z_t|z_{t-1})$
- $q(z_{t-1}, z_t) = \phi_{t-1}(z_{t-1}) q_t(z_t|z_{t-1})$

Online algorithm

Initialize $\theta^{(1)}$, $\phi_1(i) = q_1(i) = p(z_1 = i|x_1; \theta)$

For $t = 2, \dots$

- (partial) E-step
  - $q_t(j|i) = \frac{1}{Z} A_{ij} p(x_t|z_t = j; \theta^{(t-1)})$
  - $\phi_t(j) = \sum_i \phi_{t-1}(i) q_t(j|i)$
  - Update expected sufficient statistics
- M-step: $\theta^{(t)} := \arg\max_\theta \hat{f}_t(\theta)$

Incremental EM for HMMs (Bregman emissions)

For Bregman emissions (e.g. multinomial, spherical Gaussian):

- Expected sums of sufficient statistics:

$$S_T^A(i,j) = \sum_{t=2}^{T} \phi_{t-1}(i) q_t(j|i)$$

$$S_T^{\mu,0}(i) = \sum_{t=1}^{T} \phi_t(i)$$

$$S_T^{\mu,1}(i) = \sum_{t=1}^{T} \phi_t(i) x_t$$

- M-step:

$$A_{ij}^{(T)} = \frac{S_T^A(i,j)}{\sum_{j'} S_T^A(i,j')} \quad \mu_i^{(T)} = \frac{S_T^{\mu,1}(i)}{S_T^{\mu,0}(i)}$$

Cost: $O(K^2 + Kp)$ per observation

Semi-Markov extension

- Parameterize as HMM with two-variable hidden chain
  - z_t: state of current segment
  - z_t^D: counter of time steps since the start of the segment
- Use quantities $q_t(z_t, z_t^D | z_{t-1}, z_{t-1}^D)$ and $\phi_t(z_t, z_t^D)$
- Thanks to sparsity in the transitions, $O(K^2 D + KDp)$ cost per observation
- Similar approach can be used for mixture model emissions

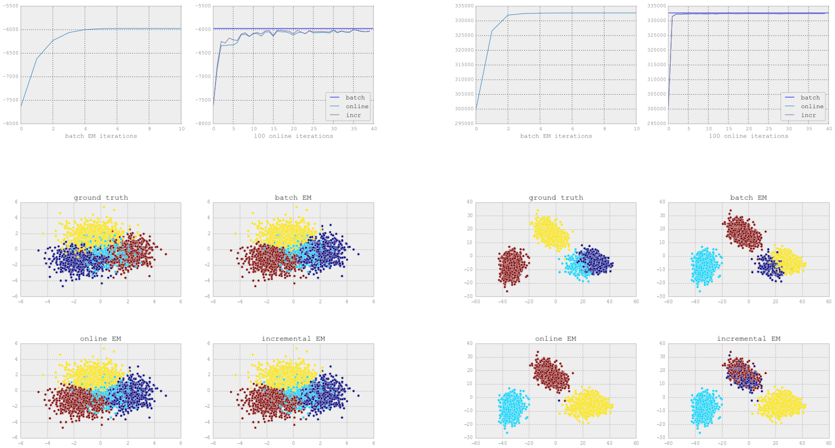
Outline

Comparison with Cappé (2011)

- Online EM algorithm based on stochastic approximation, tries to reach a *limiting EM* recursion (infinite observations)
- $O(K^4 + K^3 p)$ per observation ($O(K^4 D + K^3 D p)$ for HSMM)

	HMM	HSMM	HMM long sequence
Batch EM (5 iter.)	0.60s	1.06s	18s
Cappé (2011)	1.52s	108.6s	231s
Incremental	0.51s	0.91s	10.5s

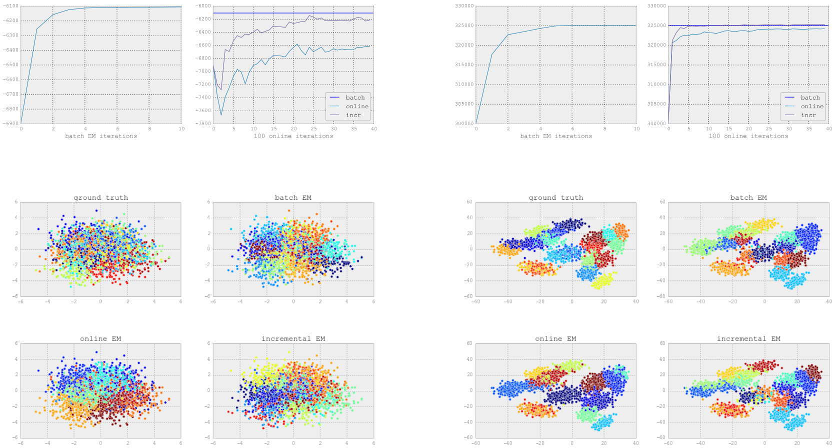
Comparison on synthetic data



Spherical Gaussian (left) and Multinomial (right).

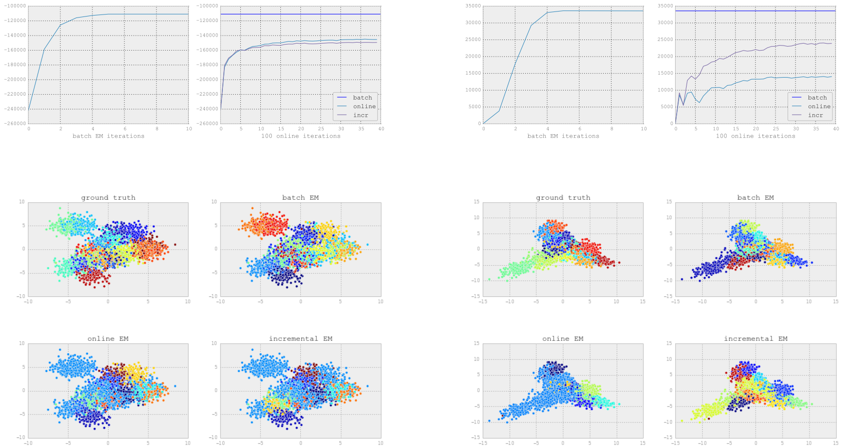
$K = 4,\ p = 5.$

Comparison on synthetic data



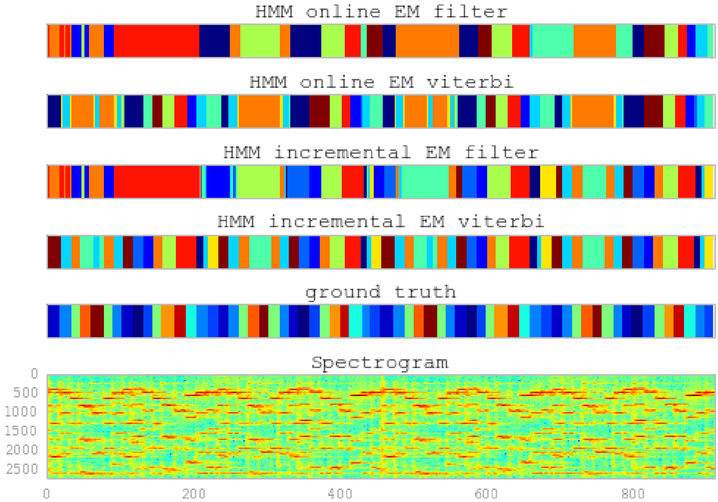
Spherical Gaussian (left) and Multinomial (right).
$K = 20$, $p = 5$.

Comparison on synthetic data



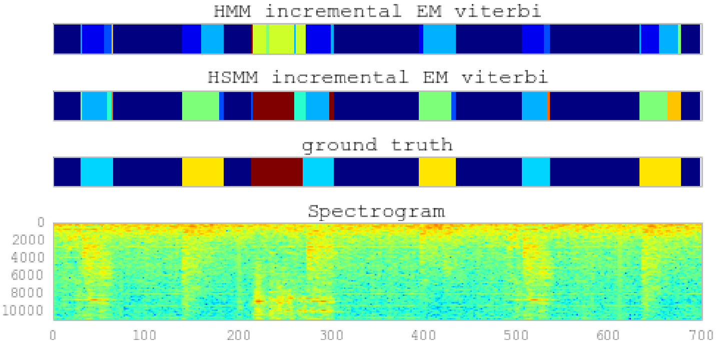
Spherical Gaussian (left) and Multinomial (right).
$K = 20,\ p = 100.$

Musical note segmentation (JS Bach violin sonata)



Cappé vs our incremental EM for HMMs on a musical note segmentation task. $K = 10$ different notes, $T = 910$.

Acoustic scene segmentation



Keys dropping and door slam from Office Live dataset. $K = 10$, HSMM durations: $NB(5, 0.2)$ (mean $= 20$).

Conclusion

- First attempt at incremental lower bound maximization in HMMs
  - Faster iterations, better convergence (empirically) than (Cappé, 2011)
  - Same complexity as one single batch EM iteration
- Works well for real-time/streaming setting
- Can accelerate learning on very long sequences
- Good results for real-time audio segmentation
- But: no theoretical guarantees

References

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6: 1705–1749, Dec. 2005.

O. Cappé. Online EM algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749, Jan. 2011.

R. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.