Lecture 4: Discrete Hidden Markov Models

Lester Mackey

April 9, 2014

Agenda

- Recap
- Discrete hidden Markov models for sequential clustered data

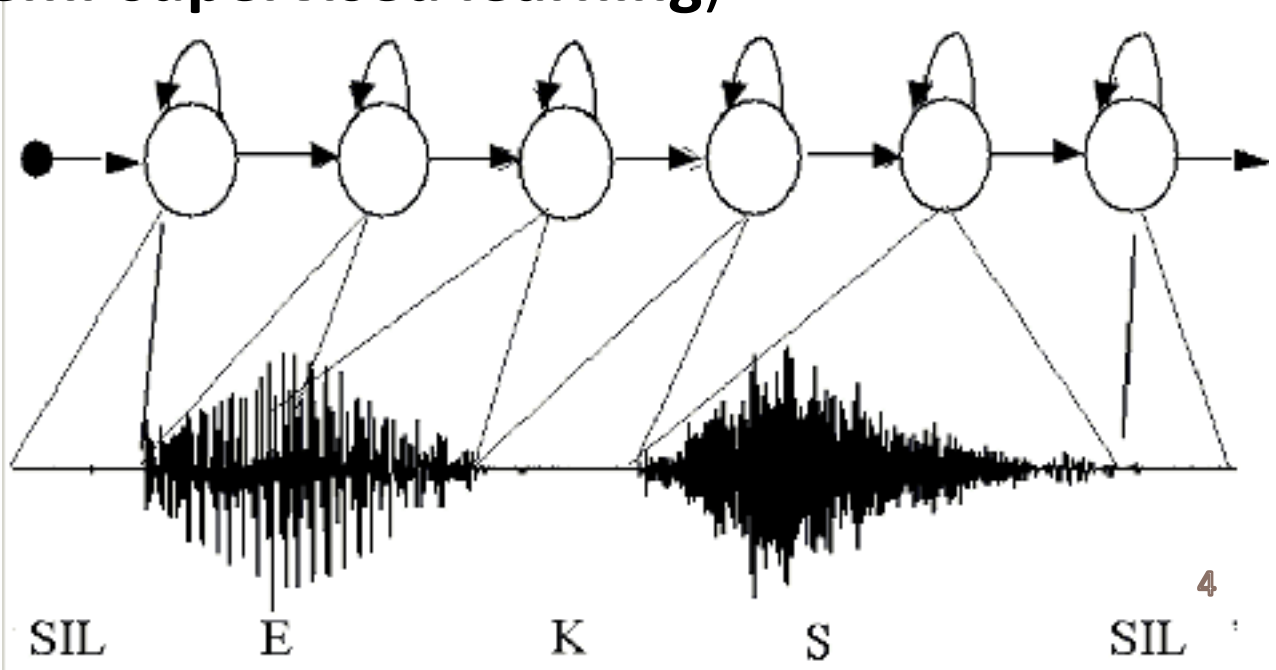
Recap

- Last time, we developed and analyzed the EM algorithm for carrying out approximate maximum likelihood estimation in latent variable models
  - Dempster/Laird/Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm"
  - “Proposed many times in special circumstances”
- Generic mixture modeling enables diverse applications
  - Email clustering for legal document review
  - Modeling unobserved disease status (cured vs. uncured) in cancer survival analysis (Yu & Tiwari, 2007)
  - Inferring community structure from network interactions (Newman, 2007)
  - Inferring test-taking behaviors from GRE response times (Schnipke & Scrams, 1997)
- However, mixture models assume *independent* draws.
- What if data has **temporal** or **sequential** structure?

Speech recognition

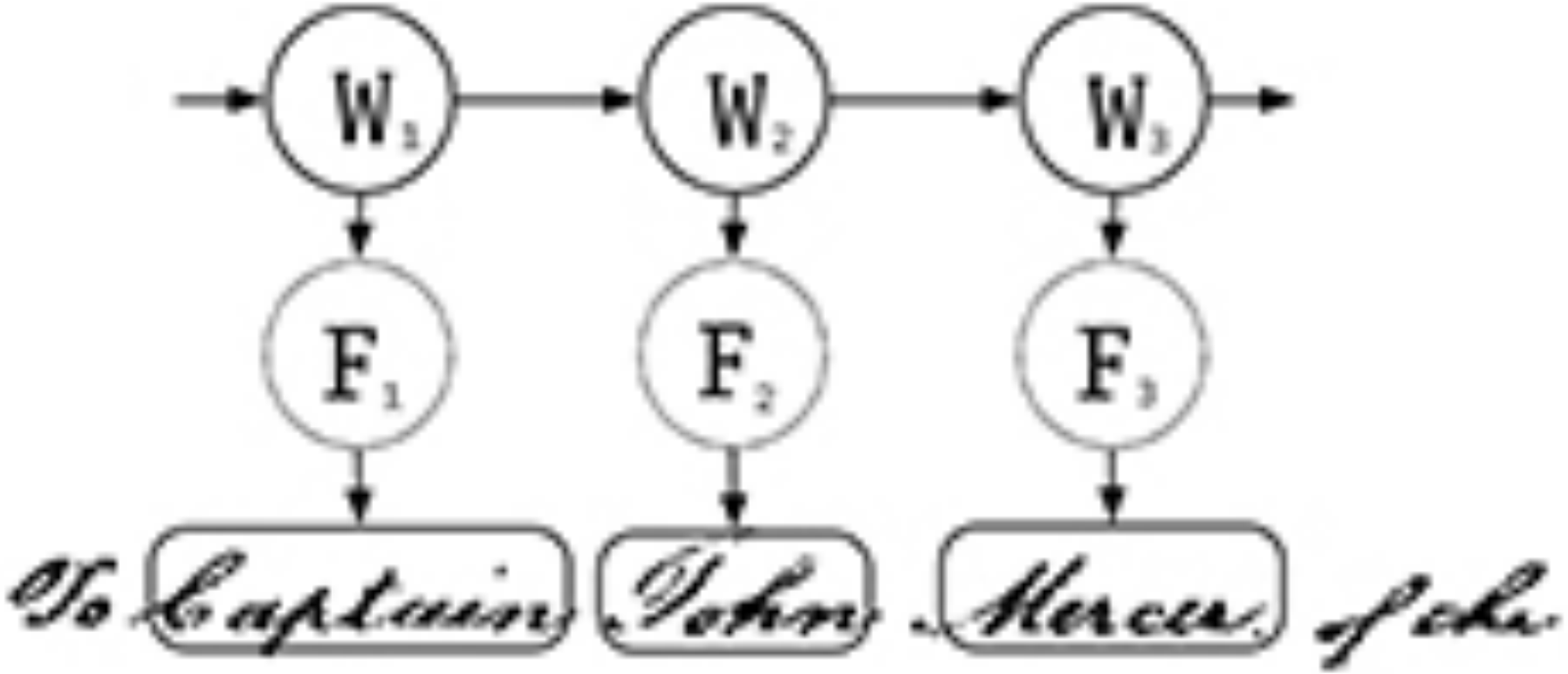
- Observe audio signal, segment and extract features
- **Goal:** Infer the latent words / phonemes responsible for each temporally indexed audio segment
- Benefit by modeling the dependence between current word or sound and adjacent words or sounds
- Can make use of expensive transcribed data and abundant unlabeled data (**semi-supervised learning**)

Speaker pronouncing "X" (SIL = silence)



Handwriting recognition

- Identify words or characters from handwritten text
- Adjacent characters and words provide important clues



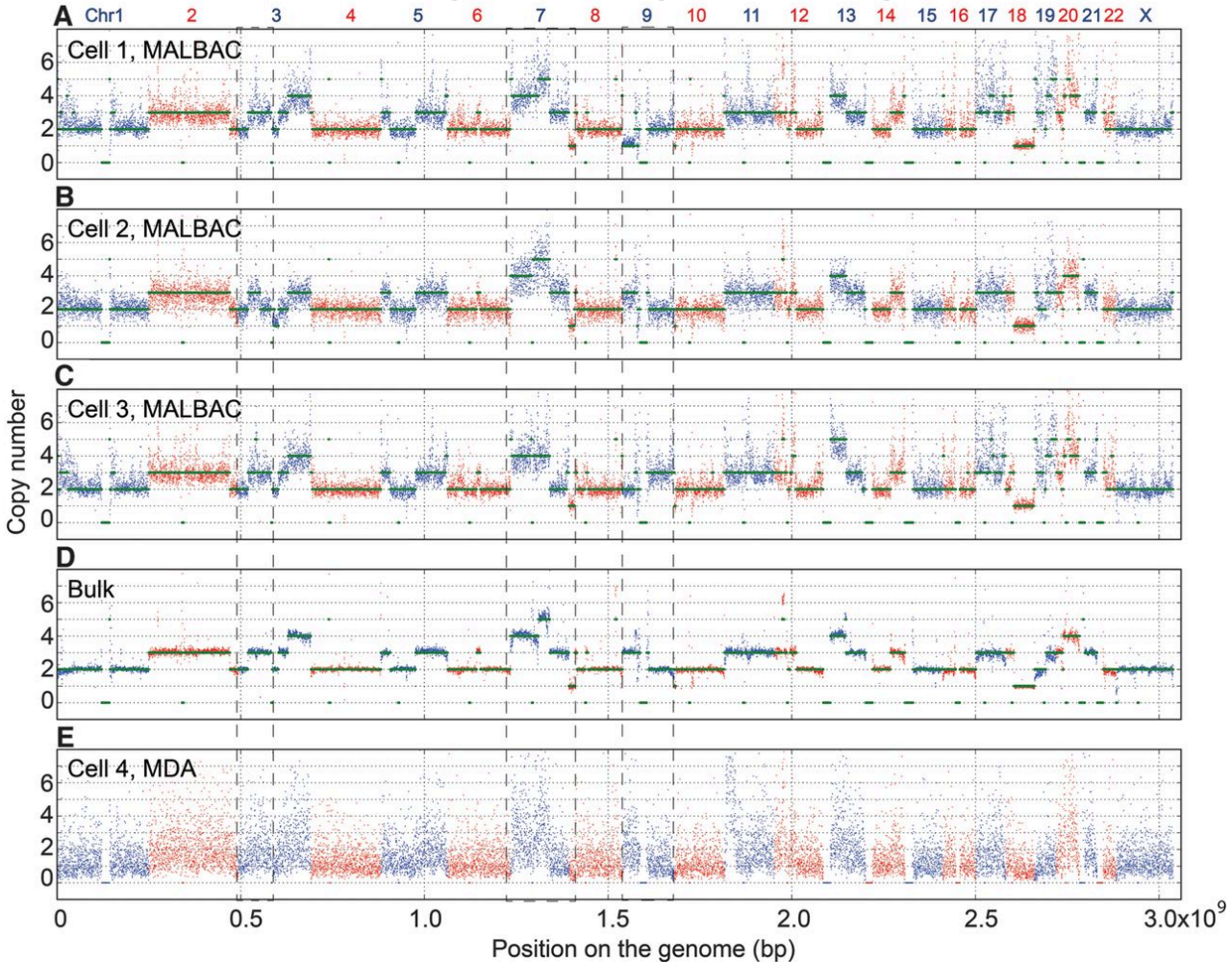
(Hard to read? That's the point!)

http://homepages.inf.ed.ac.uk/vlavrenk/

Copy number segmentation

Zong et al., http://www.sciencemag.org/content/338/6114/1622.full

- Estimate true DNA copy number from normal or tumor samples given noisy intensity measurements
- Copies often occur in large contiguous regions



Sequential clustered data

- Such sequential data examples call for methods that
  1. Uncover the latent state associated with each datapoint
  2. Take into account the sequential dependence structure
- Today, we will take a probabilistic modeling approach to clustering sequential data with **discrete hidden Markov models (HMMs)**

Blackboard discussion

- See lecture notes