Lectures 8-9 CMS 165

Spectral Methods

Spectral Methods

- Utilize spectral decomposition of matrices (and tensors)
- Review of Eigen Decomposition

For a matrix S, u is an eigenvector if $Su = \lambda u$ and λ is eigenvalue.

- For symm. $S \in \mathbb{R}^{d \times d}$, there are d eigen values.
- $S = \sum_{i \in [d]} \lambda_i u_i u_i^\top$. U is orthogonal.

Rayleigh Quotient

For matrix S with eigenvalues $\lambda_1 \geq \lambda_2 \ldots \lambda_d$ and corresponding eigenvectors $u_1, \ldots u_d$, then

$$\max_{\|z\|=1} z^\top S z = \lambda_1, \quad \min_{\|z\|=1} z^\top S z = \lambda_d,$$

and the optimizing vectors are u_1 and u_d.

Optimal Projection

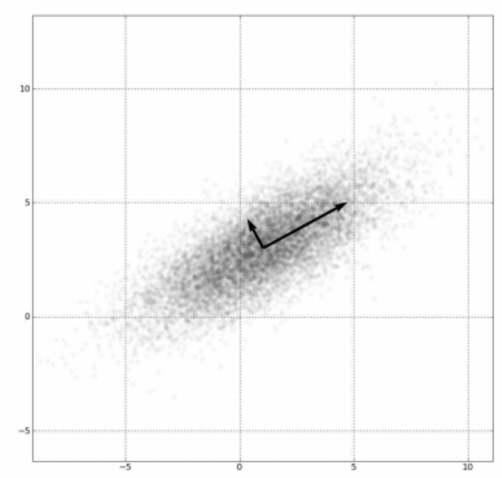
$$\max_{\substack{P:P^2=I \\ \text{Rank}(P)=k}} \text{Tr}(P^\top S P) = \lambda_1 + \lambda_2 \ldots + \lambda_k \text{ and } P \text{ spans } \{u_1, \ldots, u_k\}.$$

Simplest Spectral Method: PCA

Optimization problem

For (centered) points $x_i \in \mathbb{R}^d$, find projection P with Rank$(P) = k$ s.t.

$$\boxed{\min_{P \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i \in [n]} \|x_i - Px_i\|^2.}$$



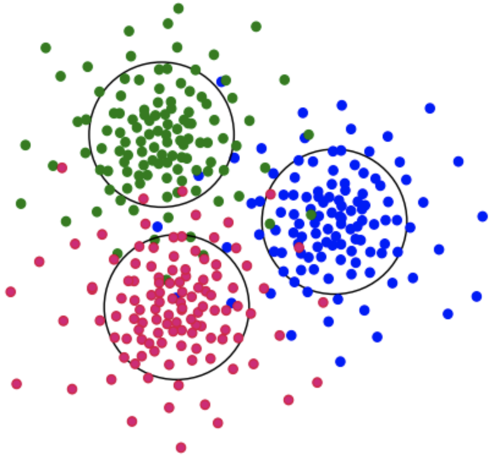
Result: If $S = \mathrm{Cov}(X)$ and $S = U \Lambda U^\top$ is eigen decomposition, we have $P = U_{(k)} U_{(k)}^\top$, where $U_{(k)}$ are top-k eigen vectors.

Proof

- By Pythagorean theorem: $\sum_i \|x_i - Px_i\|^2 = \sum_i \|x_i\|^2 - \sum_i \|Px_i\|^2$.
- Maximize: $\frac{1}{n} \sum_i \|Px_i\|^2 = \frac{1}{n} \sum_i \mathrm{Tr}\left[Px_i x_i^\top P^\top\right] = \mathrm{Tr}[PSP^\top]$.

PCA on Gaussian Mixtures

- k Gaussians: each sample is $x = Ah + z$.
- $h \in [e_1, \ldots, e_k]$, the basis vectors. $\mathbb{E}[h] = w$.
- $A \in \mathbb{R}^{d \times k}$: columns are component means.
- Let $\mu := Aw$ be the mean.
- $z \sim \mathcal{N}(0, \sigma^2 I)$ is white Gaussian noise.



$$\boxed{\mathbb{E}[(x-\mu)(x-\mu)^\top] = \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^\top + \sigma^2 I.}$$

How the above equation is obtained

$$\begin{aligned}\mathbb{E}[(x-\mu)(x-\mu)^\top] &= \mathbb{E}[(Ah-\mu)(Ah-\mu)^\top] + \mathbb{E}[zz^\top] \\ &= \sum_{i \in [k]} w_i (a_i - \mu)(a_i - \mu)^\top + \sigma^2 I.\end{aligned}$$

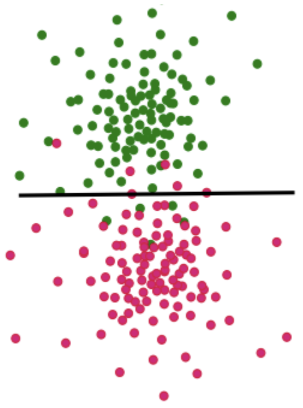
PCA on Gaussian Mixtures Cont.

$$\boxed{\mathbb{E}[(x-\mu)(x-\mu)^\top] = \sum_{i\in[k]} w_i(a_i-\mu)(a_i-\mu)^\top + \sigma^2 I.}$$

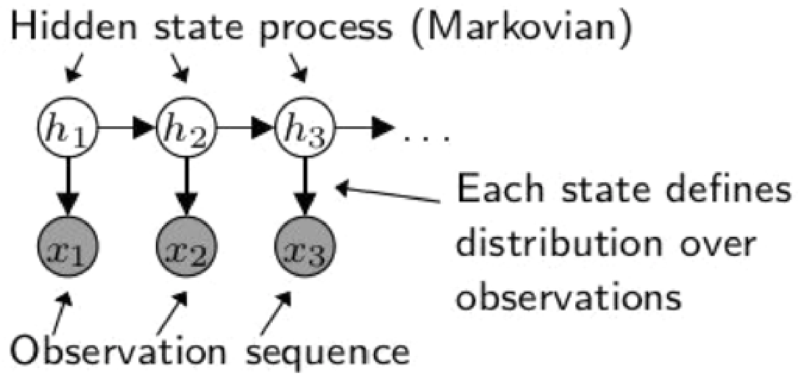
- The vectors $\{a_i - \mu\}$ are linearly dependent: $\sum_i w_i(a_i - \mu) = 0$. The PSD matrix $\sum_{i\in[k]} w_i(a_i-\mu)(a_i-\mu)^\top$ has rank $\le k-1$.
- $(k-1)$-PCA on covariance matrix $\cup\{\mu\}$ yields span(A).

Learning A through Spectral Clustering

- Project samples x on to span(A).
- Distance-based clustering (e.g. k-means).
- A series of works, e.g. Vempala & Wang.



Hidden Markov Models

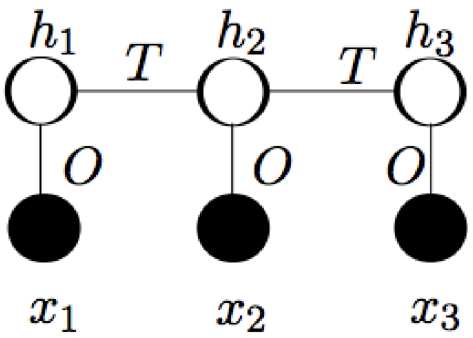


- Why HMMs?
  - Handle temporally-dependent data
  - Succinct "factored" representation when state space is low-dimensional (*c.f.* autoregressive model)
- Some uses of HMMs:
  - Monitor "belief state" of dynamical system
  - Infer latent variables from time series
  - Density estimation

Source: slides from Daniel Hsu

Discrete Hidden Markov Models

- $\mathbb{P}[h_{t+1} = i | h_t = j] = T_{i,j}$.
- $\mathbb{E}[x_t | h_t = j] = O e_j$.
- π: Initial distribution (of x_1).
- Three view model. $w := T\pi$.



$$\mathbb{E}[x_1 | h_2] = O \mathrm{Diag}(\pi) T^\top \mathrm{Diag}(w)^{-1} h_2$$
$$\mathbb{E}[x_2 | h_2] = O h_2$$
$$\mathbb{E}[x_3 | h_2] = O T h_2.$$

Condition for non-degeneracy

- $O \in \mathbb{R}^{d \times k}$ has full column rank.
- T is invertible, π and $T\pi$ have positive entries.

Observable operator in HMM

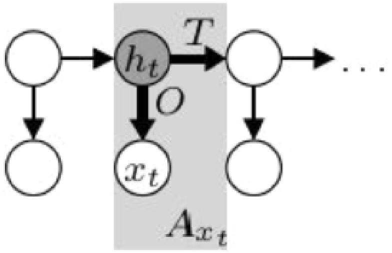
Discrete HMMs: observation operators

For $x \in \{1, \ldots, n\}$: define

$$A_x \triangleq \begin{bmatrix} & & \\ & T & \\ & & \end{bmatrix} \begin{bmatrix} O_{x,1} & & 0 \\ & \ddots & \\ 0 & & O_{x,m} \end{bmatrix} \in \mathbb{R}^{m \times m}$$

$$[A_x]_{i,j} \quad = \quad \Pr[\, h_{t+1} = i \ \wedge \ x_t = x \mid h_t = j \,].$$

The $\{A_x\}$ are *observation operators* (Schützenberger, ’61; Jaeger, ’00).



Source: slides from Daniel Hsu

Observable operator in HMM contd.

Using observation operators

Matrix multiplication handles "local" marginalization of hidden variables: *e.g.*

$$\begin{aligned}\Pr[x_1, x_2] &= \sum_{h_1} \Pr[h_1] \cdot \sum_{h_2} \Pr[h_2|h_1] \Pr[x_1|h_1] \cdot \sum_{h_3} \Pr[h_3|h_2] \Pr[x_2|h_2] \\ &= \vec{1}_m^\top A_{x_2} A_{x_1} \vec{\pi}\end{aligned}$$

where $\vec{1}_m \in \mathbb{R}^m$ is the all-ones vector.

<u>Upshot</u>: The $\{A_x\}$ contain the same information as T and O.

Source: slides from Daniel Hsu

Learning Observable Operators in HMM

Key rank condition: require $T \in \mathbb{R}^{m \times m}$ and $O \in \mathbb{R}^{n \times m}$ to have rank m (rules out pathological cases from hardness reductions)

Define $P_1 \in \mathbb{R}^n$, $P_{2,1} \in \mathbb{R}^{n \times n}$, $P_{3,x,1} \in \mathbb{R}^{n \times n}$ for $x = 1, \ldots, n$ by

$$\begin{aligned} [P_1]_i &= \Pr[x_1 = i] \\ [P_{2,1}]_{i,j} &= \Pr[x_2 = i, x_1 = j] \\ [P_{3,x,1}]_{i,j} &= \Pr[x_3 = i, x_2 = x, x_1 = j] \end{aligned}$$

(probabilities of singletons, doubles, and triples).

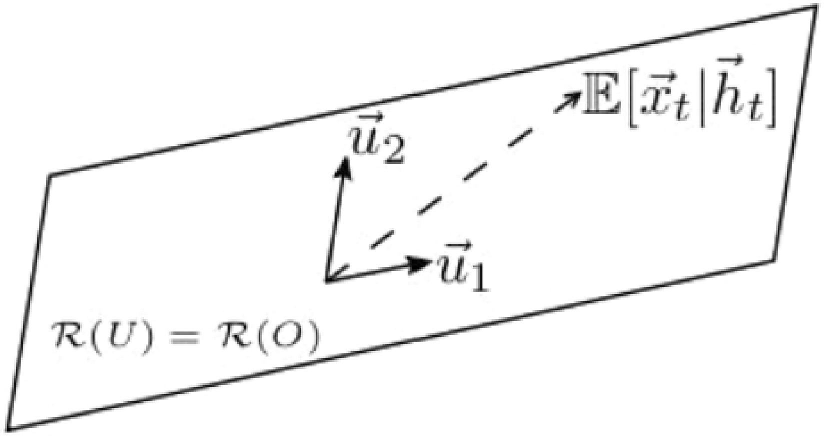
Claim: Can recover equivalent HMM parameters from P_1, $P_{2,1}$, $\{P_{3,x,1}\}$, and *these quantities can be estimated from data.*

Source: slides from Daniel Hsu

Learning Observable Operators in HMM cont.

"Thin" SVD: $P_{2,1} = U \Sigma V^\top$ where $U = [\vec{u}_1 | \dots | \vec{u}_m] \in \mathbb{R}^{n \times m}$

Guaranteed m non-zero singular values by rank condition.



New parameters (based on U) implicitly transform hidden states

$$\vec{h}_t \quad \mapsto \quad (U^\top O)\vec{h}_t \quad = \quad U^\top \mathbb{E}[\vec{x}_t | \vec{h}_t]$$

(*i.e.* change to coordinate representation of $\mathbb{E}[\vec{x}_t | \vec{h}_t]$ w.r.t. $\{\vec{u}_1, \dots, \vec{u}_m\}$).

Source: slides from Daniel Hsu

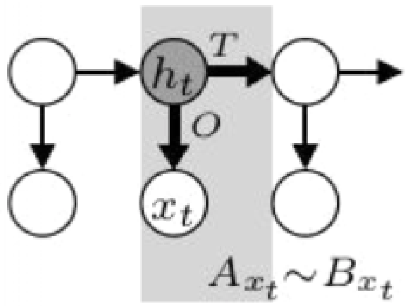
Learning Observable Operators in HMM cont.

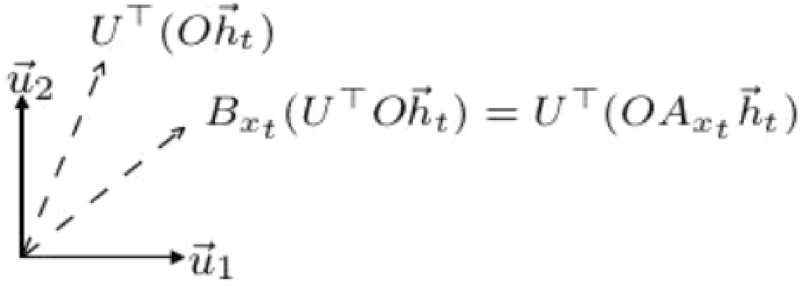
For each $x = 1, \ldots, n$,

$$
\begin{aligned}
B_x &\triangleq (U^\top P_{3,x,1})\,(U^\top P_{2,1})^+ && (X^+ \text{ is pseudoinv. of } X)\\
&= (U^\top O)\,A_x\,(U^\top O)^{-1}\,. && \text{(algebra)}
\end{aligned}
$$

The B_x operate in the coord. system defined by $\{\vec{u}_1, \ldots, \vec{u}_m\}$ (columns of U).

$$\Pr[x_{1:t}] \;=\; \vec{1}_m^\top A_{x_t} \ldots A_{x_1} \vec{\pi} \;=\; \vec{1}_m^\top (U^\top O)^{-1} B_{x_t} \ldots B_{x_1} (U^\top O) \vec{\pi}$$





Upshot: Suffices to learn $\{B_x\}$ instead of $\{A_x\}$.

Source: slides from Daniel Hsu

Learning Algorithm for HMM

1. Look at triples of observations (x_1, x_2, x_3) in data;
   estimate frequencies $\widehat{P}_1$, $\widehat{P}_{2,1}$, and $\{\widehat{P}_{3,x,1}\}$

2. Compute SVD of $\widehat{P}_{2,1}$ to get matrix of top m singular vectors $\widehat{U}$
   ("subspace identification")

3. Compute $\widehat{B}_x \triangleq (\widehat{U}^\top \widehat{P}_{3,x,1})(\widehat{U}^\top \widehat{P}_{2,1})^+$ for each x
   ("observation operators")

4. Compute $\widehat{b}_1 \triangleq \widehat{U}^\top \widehat{P}_1$ and $\widehat{b}_\infty \triangleq (\widehat{P}_{2,1}^\top \widehat{U})^+ \widehat{P}_1$

- Joint probability calculations:

$$\widehat{\Pr}[x_1, \ldots, x_t] \triangleq \widehat{b}_\infty^\top \widehat{B}_{x_t} \ldots \widehat{B}_{x_1} \widehat{b}_1.$$

- Conditional probabilities: Given $x_{1:t-1}$,

$$\widehat{\Pr}[x_t | x_{1:t-1}] \triangleq \widehat{b}_\infty^\top \widehat{B}_{x_t} \widehat{b}_t$$

  where

$$\widehat{b}_t \triangleq \frac{\widehat{B}_{x_{t-1}} \ldots \widehat{B}_{x_1} \widehat{b}_1}{\widehat{b}_\infty^\top \widehat{B}_{x_{t-1}} \ldots \widehat{B}_{x_1} \widehat{b}_1} \approx (U^\top O)\mathbb{E}[\vec{h}_t | x_{1:t-1}].$$

  "Belief states" $\widehat{b}_t$ linearly related to conditional hidden states.
  (b_t live in hypercube $[-1, +1]^m$ instead simplex Δ^m)

Source: slides from Daniel Hsu

Learning Guarantees

Sample complexity bound

Joint probability accuracy: with probability $\geq 1 - \delta$,

$$O\left(\frac{t^2}{\epsilon^2} \cdot \left(\frac{m}{\sigma_m(O)^2 \sigma_m(P_{2,1})^4} + \frac{m \cdot n_0}{\sigma_m(O)^2 \sigma_m(P_{2,1})^2}\right) \cdot \log \frac{1}{\delta}\right)$$

observation triples sampled from the HMM suffices to guarantee

$$\sum_{x_1,\ldots,x_t} |\Pr[x_1, \ldots, x_t] - \widehat{\Pr}[x_1, \ldots, x_t]| \leq \epsilon.$$

- m: number of states
- n_0: number of observations that account for most of the probability mass
- $\sigma_m(M)$: mth largest singular value of matrix M

Also have a sample complexity bound for conditional probability accuracy.

Source: slides from Daniel Hsu

Lots of other applications of spectral methods

- Extending HMMs to Partially observed Markov decision processes (POMDP) and Predictive state representations (PSR): passive vs active.
- POMDP: Action based on each observation and can influence Markovian evolution of hidden state
- PSR: No explicit Markovian assumption on hidden state. Directly predicts future (tests) based on past observations and actions (For linear PSR, similar to spectral updates in HMM)
- Stochastic bandits in a low rank subspace (ask TA Sahin about it)

References

- Matrix computations (textbook) by Golub and Van Loan
- A spectral algorithm for learning hidden Markov models by Hsu, Kakade and Zhang.
- Spectral Approaches to Learning Predictive Representations by Byron Boots (PhD thesis) https://apps.dtic.mil/dtic/tr/fulltext/u2/a566112.pdf