Global Learning Methods for Latent Variable Sequence Models

Cem Subakan
(PhD student)

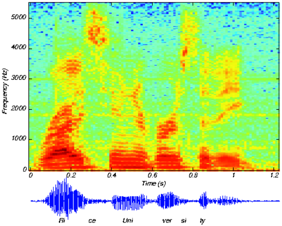
University of Illinois at Urbana-Champaign

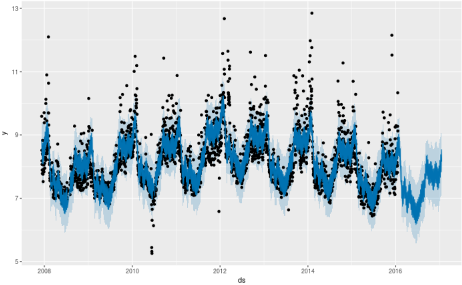
July 18'th, 2017

Outline

Sequence Modeling

- E.g. Speech, Handwriting, Music, Text, Finance, and
- **Uber**





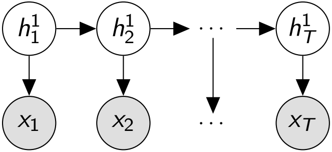

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.
```

Familiar Sequence Modeling Approaches

- Latent Variable Sequence Modeling
  - HMM/LDS: $p(x_{1:T}) = \sum_{h_{1:T}} \prod_t p(x_t|h_t)p(h_t|h_{t-1})$

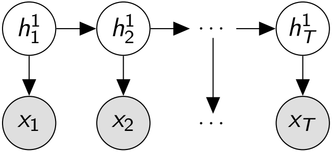
Familiar Sequence Modeling Approaches

- Latent Variable Sequence Modeling
  - HMM/LDS: $p(x_{1:T}) = \sum_{h_{1:T}} \prod_t p(x_t|h_t)p(h_t|h_{t-1})$



Familiar Sequence Modeling Approaches

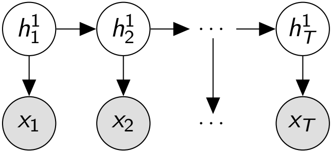
- Latent Variable Sequence Modeling
  - HMM/LDS: $p(x_{1:T}) = \sum_{h_{1:T}} \prod_t p(x_t | h_t) p(h_t | h_{t-1})$



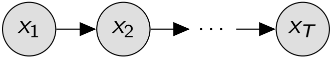
- Fully Observed Sequence Modeling

Familiar Sequence Modeling Approaches

- Latent Variable Sequence Modeling
  - HMM/LDS: $p(x_{1:T}) = \sum_{h_{1:T}} \prod_t p(x_t|h_t)p(h_t|h_{t-1})$

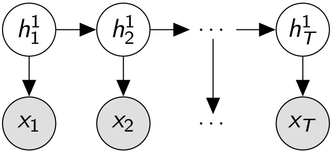


- Fully Observed Sequence Modeling
  - Markov Model $p(x_{1:T}) = \prod_t p(x_t|x_{t-1})$

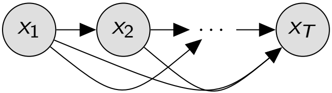


Familiar Sequence Modeling Approaches

- Latent Variable Sequence Modeling
  - HMM/LDS: $p(x_{1:T}) = \sum_{h_{1:T}} \prod_t p(x_t|h_t)p(h_t|h_{t-1})$

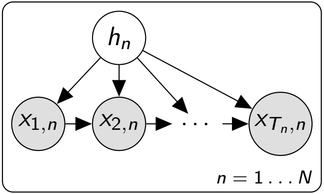


- Fully Observed Sequence Modeling
  - RNN $p(x_{1:T}) = \prod_t p(x_t|x_{1:t-1})$



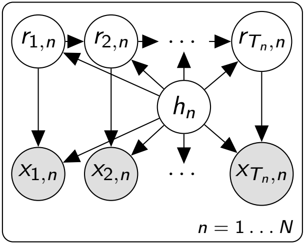
Familiar Sequence Models

Mixture of Markov Models



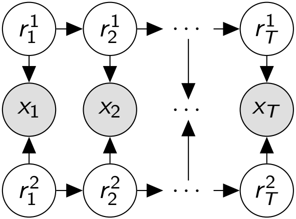
[Subakan, et al. 2013]

Mixture of HMMs

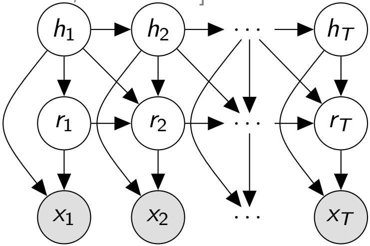


[Subakan, et al. 2014]

Factorial HMM



[Subakan, et al. 2015]



[Subakan, et al. 2015]

Maximum Likelihood via EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_\theta \log p(x|\theta) = \max_\theta \log \sum_h p(x, h|\theta)$$

Maximum Likelihood via EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_\theta \log p(x|\theta) = \max_\theta \log \sum_h p(x, h|\theta)$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h)$:

$$\log \sum_h p(x, h|\theta) \frac{q(h)}{q(h)} = \log \mathbb{E}_{q(h)} \left[\frac{p(x, h|\theta}{q(h)} \right]$$
$$\geq \mathbb{E}_{q(h)} [\log p(x, h|\theta)] + H_q$$

Maximum Likelihood via EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_{\theta} \log p(x|\theta) = \max_{\theta} \log \sum_{h} p(x, h|\theta)$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h)$:

$$\log \sum_{h} p(x, h|\theta) \frac{q(h)}{q(h)} = \log \mathbb{E}_{q(h)} \left[\frac{p(x, h|\theta}{q(h)} \right]$$
$$\geq \mathbb{E}_{q(h)} [\log p(x, h|\theta)] + H_q$$

- This objective is in general not jointly convex.

Maximum Likelihood via EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_\theta \log p(x|\theta) = \max_\theta \log \sum_h p(x, h|\theta)$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h)$:

$$\log \sum_h p(x, h|\theta) \frac{q(h)}{q(h)} = \log \mathbb{E}_{q(h)} \left[\frac{p(x, h|\theta}{q(h)} \right]$$
$$\geq \mathbb{E}_{q(h)} [\log p(x, h|\theta)] + H_q$$

- This objective is in general not jointly convex.
- Is there an alternative method which yields a global solution for this?

Maximum Likelihood via EM

- Maximum Likelihood is the first thing that comes to mind:

$$\max_{\theta} \log p(x|\theta) = \max_{\theta} \log \sum_{h} p(x, h|\theta)$$

- We can use Jensen's inequality by injecting a logarithm, and the distribution $q(h)$:

$$\begin{aligned}\log \sum_{h} p(x, h|\theta) \frac{q(h)}{q(h)} &= \log \mathbb{E}_{q(h)} \left[\frac{p(x, h|\theta}{q(h)} \right] \\ &\geq \mathbb{E}_{q(h)} \left[\log p(x, h|\theta) \right] + H_q\end{aligned}$$

- This objective is in general not jointly convex.
- Is there an alternative method which yields a global solution for this?
- (Probably) No. ($\mathbf{P} \neq \mathbf{NP}$). But there are "close" problems which are easier to solve.

Global vs Local

Global vs Local

Disclaimer: We will not necessarily go to the red summit.

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0, \ \textit{diagonal}$$

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0, \ \textit{diagonal}$$

- Rayleigh Quotient

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0, \ \textit{diagonal}$$

- Rayleigh Quotient
- Procrustes Problem

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0, \ \textit{diagonal}$$

- Rayleigh Quotient
- Procrustes Problem
- Sinusoid Estimation (ESPRIT)

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0, \ \textit{diagonal}$$

- Rayleigh Quotient
- Procrustes Problem
- Sinusoid Estimation (ESPRIT)
- Method of Moments for Latent Variable Models

Example Problems with Global (Unique) Solutions

- Convex optimization problems. (Some convex problems are not poly-time solvable though)
- Non-Convex canonical example:

$$\min_{U,\Sigma,V} \|X - U\Sigma V^\top\|_F$$
$$U^\top U = I,$$
$$V^\top V = I,$$
$$\Sigma \geq 0,\ \textit{diagonal}$$

- Rayleigh Quotient
- Procrustes Problem
- Sinusoid Estimation (ESPRIT)
- Method of Moments for Latent Variable Models
- Dictionary learning. (under assumptions)

Plan

The other way: Method of Moments

- The idea is to estimate the models parameters $\mu_{1:K}$ by solving a system of non-linear equations formed with moments $\mathbb{E}[g_k(x)]$, $k \in \{1, \ldots K\}$:

$$\begin{aligned}\mathbb{E}[g_1(x)] =& f_1(\mu_{1:K}) \\ &\vdots \\ \mathbb{E}[g_K(x)] =& f_K(\mu_{1:K})\end{aligned}$$

The other way: Method of Moments

- The idea is to estimate the models parameters $\mu_{1:K}$ by solving a system of non-linear equations formed with moments $\mathbb{E}[g_k(x)]$, $k \in \{1, \ldots K\}$:

$$\begin{aligned} \mathbb{E}[g_1(x)] &= f_1(\mu_{1:K}) \\ &\vdots \\ \mathbb{E}[g_K(x)] &= f_K(\mu_{1:K}) \end{aligned}$$

- Canonical Example: $x \sim \mathcal{G}(a, b)$:

$$\begin{aligned} \mathbb{E}[x] &= ab \\ \mathbb{E}[x^2] &= ab^2 + a^2 b^2 \end{aligned} \quad \rightarrow \quad \begin{aligned} \widehat{b} &= (\mathbb{E}[x^2] - \mathbb{E}[x]^2)/\mathbb{E}[x] \\ \widehat{a} &= \mathbb{E}[x]^2/(\mathbb{E}[x^2] - \mathbb{E}[x]^2) \end{aligned}$$

The other way: Method of Moments

- The idea is to estimate the models parameters $\mu_{1:K}$ by solving a system of non-linear equations formed with moments $\mathbb{E}[g_k(x)]$, $k \in \{1, \ldots K\}$:

$$\begin{aligned}\mathbb{E}[g_1(x)] =& f_1(\mu_{1:K}) \\ &\vdots \\ \mathbb{E}[g_K(x)] =& f_K(\mu_{1:K})\end{aligned}$$

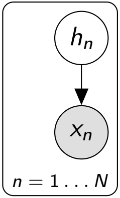
- Canonical Example: $x \sim \mathcal{G}(a, b)$:

$$\begin{aligned}\mathbb{E}[x] =& ab \\ \mathbb{E}[x^2] =& ab^2 + a^2 b^2\end{aligned} \quad \rightarrow \quad \begin{aligned}\widehat{b} =& (\mathbb{E}[x^2] - \mathbb{E}[x]^2)/\mathbb{E}[x] \\ \widehat{a} =& \mathbb{E}[x]^2/(\mathbb{E}[x^2] - \mathbb{E}[x]^2)\end{aligned}$$

- This is not as statistically efficient as ML (CRLB). But the problem is (usually) "easier".

Method of Moments for LVMs

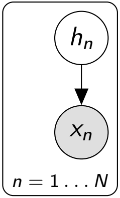
MoM for spherical GMM: [Hsu, Kakade 13]



$$
\begin{aligned}
h &\sim Cat(\pi) \\
x|h &\sim \mathcal{N}(\mu_h, \sigma^2 I)
\end{aligned}
$$

Method of Moments for LVMs

MoM for spherical GMM: [Hsu, Kakade 13]



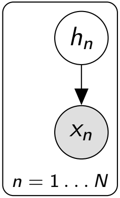
$$h \sim Cat(\pi)$$
$$x|h \sim \mathcal{N}(\mu_h, \sigma^2 I)$$

- Let's write down some moments:

$$\mathbb{E}[x] = \sum_{k=1}^{K} \mu_k \pi_k,$$

Method of Moments for LVMs

MoM for spherical GMM: [Hsu, Kakade 13]



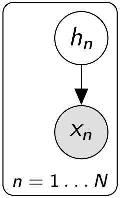
$$h \sim Cat(\pi)$$
$$x|h \sim \mathcal{N}(\mu_h, \sigma^2 I)$$

- Let's write down some moments:

$$\mathbb{E}[x] = \sum_{k=1}^{K} \mu_k \pi_k, \ \mathbb{E}[x \otimes x] = \sum_{k=1}^{K} \pi_k \ \mu_k \otimes \mu_k + \sigma^2 I$$

Method of Moments for LVMs

MoM for spherical GMM: [Hsu, Kakade 13]



$$h \sim Cat(\pi)$$
$$x|h \sim \mathcal{N}(\mu_h, \sigma^2 I)$$

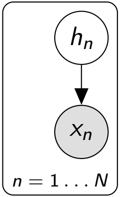
- Let's write down some moments:

$$\mathbb{E}[x] = \sum_{k=1}^{K} \mu_k \pi_k, \; \mathbb{E}[x \otimes x] = \sum_{k=1}^{K} \pi_k \; \mu_k \otimes \mu_k + \sigma^2 I$$

$$\mathbb{E}[x \otimes x \otimes x] = \sum_{k=1}^{K} \pi_k \; \mu_k \otimes \mu_k \otimes \mu_k$$
$$+ \sigma^2 \left(\sum_{l=1}^{L} \mathbb{E}[x] \otimes e_l \otimes e_l + e_l \otimes \mathbb{E}[x] \otimes e_l + e_l \otimes e_l \otimes \mathbb{E}[x] \right)$$

Method of Moments for LVMs

MoM for spherical GMM: [Hsu, Kakade 13]



$$h \sim Cat(\pi)$$
$$x|h \sim \mathcal{N}(\mu_h, \sigma^2 I)$$

$$\mathbb{E}[x] = \sum_{k=1}^{K} \mu_k \pi_k, \ \mathbb{E}[x \otimes x] = \sum_{k=1}^{K} \pi_k \ \mu_k \otimes \mu_k + \text{garbage}$$

$$\mathbb{E}[x \otimes x \otimes x] = \sum_{k=1}^{K} \pi_k \ \mu_k \otimes \mu_k \otimes \mu_k + \text{garbage}$$

Method of Moments for a GMM

- Form the system of equations:

$$M_2 := \mathbb{E}[x \otimes x] - \text{garbage} = \sum_{k=1}^{K} \pi_k \, \mu_k \otimes \mu_k$$

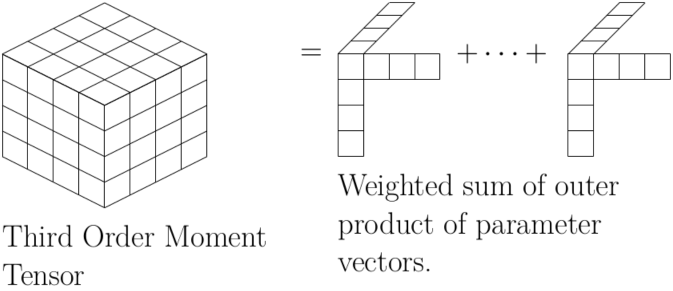
$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \text{garbage} = \sum_{k=1}^{K} \pi_k \, \mu_k \otimes \mu_k \otimes \mu_k$$

Method of Moments for a GMM

- Form the system of equations:

$$M_2 := \mathbb{E}[x \otimes x] - \text{garbage} = \sum_{k=1}^{K} \pi_k \, \mu_k \otimes \mu_k$$

$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \text{garbage} = \sum_{k=1}^{K} \pi_k \, \mu_k \otimes \mu_k \otimes \mu_k$$



Obtaining the parameters

- Whiten M_3 with a matrix W, such that:

$$W^\top M_2 W = I$$

.

Obtaining the parameters

- Whiten M_3 with a matrix W, such that:

$$W^\top M_2 W = I$$

- Then the eigenvectors of

$$\widetilde{M}_3 = \sum_{k=1}^{K} w_k (W^\top \mu) \otimes (W^\top \mu) \otimes (W^\top \mu)$$

are obtainable via power iterations.

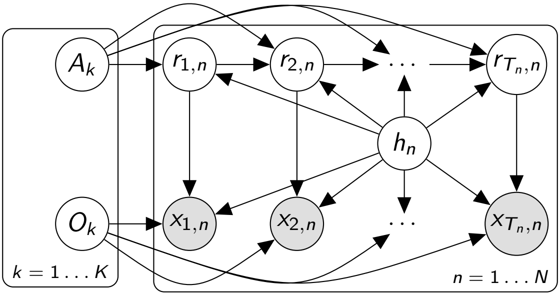
How general are moment methods?

(and where do I come in)

- PCA
- ICA papers from 90s [mainly Cardoso]
- System ID literature from 90s. (Kalman Filters)
- Inference in HMMs [Hsu et al. 09]
- Parameter Estimation in HMMs [Anandkumar et al. 12, 14]
- Multiview Discrete/Mixture Models [Anandkumar et al. 12]
- Inference in general trees [Parikh et al. 12]
- Single View Spherical GMMs [Hsu, Kakade, 13]
- Parameter estimation in somewhat general graphs [Chaganty, Liang 14]
- Framework for HMMs with special transition structures. [Me et al., 14,15]
- Attempts on Neural Networks [Anandkumar, 15,16]

Spectral Learning of Mixture of HMMs

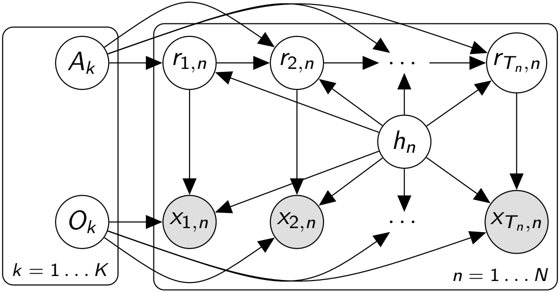
[Smyth, 97]



$$h_n \sim Categorical(\pi_n)$$
$$\mathbf{x}_n \sim HMM(A_n, O_n)$$

Spectral Learning of Mixture of HMMs

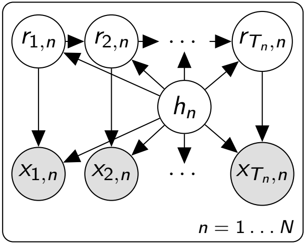
[Smyth, 97]



$$h_n \sim Categorical(\pi_n)$$
$$\mathbf{x}_n \sim HMM(A_n, O_n)$$

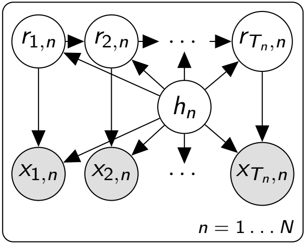
- Learning Goal: Estimate π_n, A_n, O_n, given $\mathbf{x}_{1:N}$

Mixture of HMMs



$$
\begin{aligned}
h_n &\sim Categorical(\pi_n) \\
\mathbf{x}_n | h_n &\sim HMM(\rho_{h_n}, A_{h_n}, O_{h_n})
\end{aligned}
$$

Mixture of HMMs



$$h_n \sim Categorical(\pi_n)$$
$$\mathbf{x}_n | h_n \sim HMM(\rho_{h_n}, A_{h_n}, O_{h_n})$$

$$\begin{aligned}
\mathbb{E}[x_2 \otimes x_1] =& \sum_{h,r_1} \rho_h \pi_h \left(\mathbb{E}[x_2 | r_1, h] \otimes \mathbb{E}[x_1 | r_1, h] \right) \\
=& \sum_{h,r_1} \rho_h \pi_h \left(\sum_{r_2} A(r_1, r_2, h) \mu_{r_2,h} \right) \otimes \mu_{r_1,h} \\
=& O_{flat} A_{bdiag} \text{diag}(\rho \otimes \pi) O_{flat}^\top
\end{aligned}$$

Problem: The moment estimator is agnotic to the block structure of the model.

Mixture of HMMs

- An MHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, \pi)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & A_2 & \dots & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & \mathbf{0} & \dots & A_K \end{bmatrix}, \quad \bar{\nu} = \begin{bmatrix} \pi_1 \nu_1 \\ \pi_2 \nu_2 \\ \vdots \\ \pi_K \nu_K \end{bmatrix} .$$

- How to impose this structural constraint on the estimator?

Two stage estimation for HMMs

HMM-Mixture model equivalence, [Kontorovich et al., 13]

An HMM with state marginals $p(h_t)$ is equivalent to a mixture model with mixing weights $\pi := \frac{1}{T}\sum_{t=1}^{T} p(h_t)$, and the same emission parameters.

Two stage estimation for HMMs

HMM-Mixture model equivalence, [Kontorovich et al., 13]

An HMM with state marginals $p(h_t)$ is equivalent to a mixture model with mixing weights $\pi := \frac{1}{T}\sum_{t=1}^{T} p(h_t)$, and the same emission parameters.

- First compute (estimate) $\widehat{O}$, and $\widehat{p}i$.
- Then solve the convex problem:

$$\min_{A} \|M_2 - \widehat{O} A \text{diag}(\widehat{\pi}) \widehat{O}\|_F$$
$$s.t.\ 1^\top A = 1^\top,$$
$$A \geq 0.$$

Two stage estimation framework with structural constraints:

Two stage estimation framework

- Get rough/permuted estimates for the parameters $\widehat{O}, \widehat{A}, \widehat{\pi}$.
- De-permute A. (Solve the graph problem dictated by model)
- Solve:

$$\begin{aligned} &\min_{A} \|M_2 - \widehat{O} A \mathrm{diag}(\widehat{\pi}) \widehat{O}\|_F \\ &s.t.\ 1^\top A = 1^\top, \\ &\quad A \geq 0. \\ &\quad f(\mathcal{M}, A) = 0 \end{aligned}$$

- f, and $\mathcal{M}$ depend on the model.

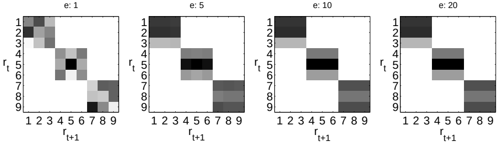
Structural Constraints wrt. Model

The framework handles these models:

- **MHMM**: $f(\mathcal{M}, A) = A \odot (1 - \mathcal{M}) = \mathbf{0}$. $\mathcal{M}$ is block diagonal.
- **SHMM**: $f(\mathcal{M}, A) = A \odot (1 - \mathcal{M}) - \widehat{B} \otimes \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^\top = \mathbf{0}$. $\mathcal{M}$ is block diagonal.
- **Left-to-Right HMM**: $f(\mathcal{M}, A) = A \odot (1 - \mathcal{M}) = 0$, estimate $\mathcal{M}$ with a greedy graph traversal algorithm. $\mathcal{M}$ is lower triangular.
- **Bakis HMM**: $f(\mathcal{M}, A) = A \odot (1 - \mathcal{M}) = 0$, $\mathcal{M}$ corresponds to an Hamiltonian circuit (TSP approximation). $\mathcal{M}$ is binary lower first uni-triangular.
- **HMM with mixture emissions**: $f(\mathcal{M}, A^{i,j}) = A^{i,j}\mathbf{1}^\top$.

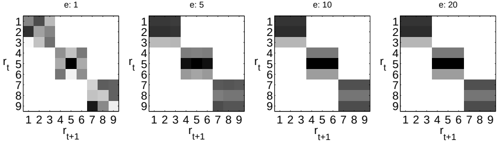
Mixture of HMMs: De-permutation

- $\lim_{e\to\infty} \bar{A}^e = [\bar{v}_1 1_M^\top,\ \bar{v}_2 1_M^\top,\ \ldots,\ \bar{v}_K 1_M^\top]$, where $\bar{v}_k$ is the k'th eigenvector of $\bar{A}$.

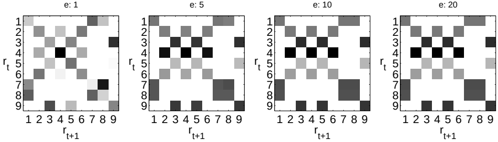


Mixture of HMMs: De-permutation

- $\lim_{e\to\infty} \bar{A}^e = [\bar{v}_1 1_M^\top,\ \bar{v}_2 1_M^\top,\ \ldots,\ \bar{v}_K 1_M^\top]$, where $\bar{v}_k$ is the k'th eigenvector of $\bar{A}$.

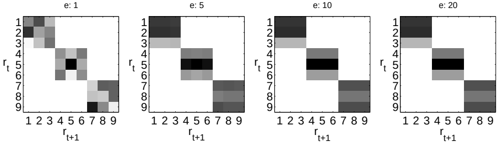


- What happens for $\mathcal{P}(\bar{A})$:

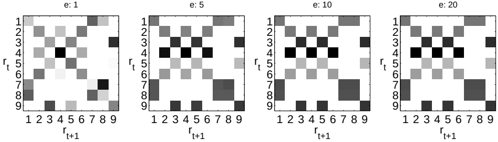


Mixture of HMMs: De-permutation

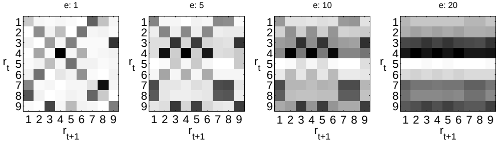
- $\lim_{e \to \infty} \bar{A}^e = [\bar{v}_1 1_M^\top,\ \bar{v}_2 1_M^\top,\ \ldots,\ \bar{v}_K 1_M^\top]$, where $\bar{v}_k$ is the k'th eigenvector of $\bar{A}$.



- What happens for $\mathcal{P}(\bar{A})$:

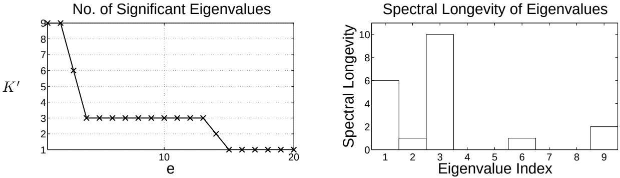


- What happens in practice:



MHMM De-permutation Continued

- But we can estimate the number of HMMs:



- Then form rank-$\widehat{K}$ reconstruction A^r:

$$A^r = V_{1:\widehat{K}} \Lambda_{1:\widehat{K}} V^{-1}$$

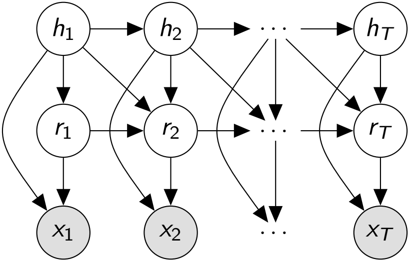
- Then Cluster. (A La Spectral Clustering)

Experimental Results

Digit clustering with MHMMs:

Algorithm	1v2	1v3	1v4	1v5	2v3	2v4	2v5
Spectral	100	70	54	55	83	99	99
EM init. w/ Spectral	100	99	100	100	96	100	100
EM init. at Random	96	99	98	54	83	100	100

Switching HMM



$$\begin{aligned} h_t|h_{t-1} &\sim Cat(B(:, h_{t-1})) \\ r_t|r_{t-1}, h_t, h_{t-1} &\sim [h_t = h_{t-1}]Cat(A(:, r_{t-1}, h_t)) \\ &\quad + [h_t \neq h_{t-1}]\mathcal{U}(.) \\ x_t|h_t, r_t &\sim p(x_t|h_t, r_t) \end{aligned}$$

Switching HMM

- An SHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, B)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \bar{A} = \begin{bmatrix} B_{1,1} A_1 & B_{1,2} \frac{\mathbf{1}\mathbf{1}^\top}{M} & \dots & B_{1,M} \frac{\mathbf{1}\mathbf{1}^\top}{M} \\ B_{2,1} \frac{\mathbf{1}\mathbf{1}^\top}{M} & B_{2,2} A_2 & \dots & B_{2,M} \frac{\mathbf{1}\mathbf{1}^\top}{M} \\ & & \ddots & \\ B_{M,1} \frac{\mathbf{1}\mathbf{1}^\top}{M} & B_{M,2} \frac{\mathbf{1}\mathbf{1}^\top}{M} & \dots & B_{M,M} A_K \end{bmatrix},$$

$$\bar{\nu} = \begin{bmatrix} \pi_1 \nu_1 & \pi_2 \nu_2 & \dots & \pi_K \nu_K \end{bmatrix}^\top.$$

- How to impose this structural constraint on the estimator?

Switching HMM

- An SHMM with *local* parameters $\theta_{1:K} = (O_{1:K}, A_{1:K}, \nu_{1:K}, B)$ is an HMM with *global* parameters $\bar{\theta} = (\bar{O}, \bar{A}, \bar{\nu})$, where:

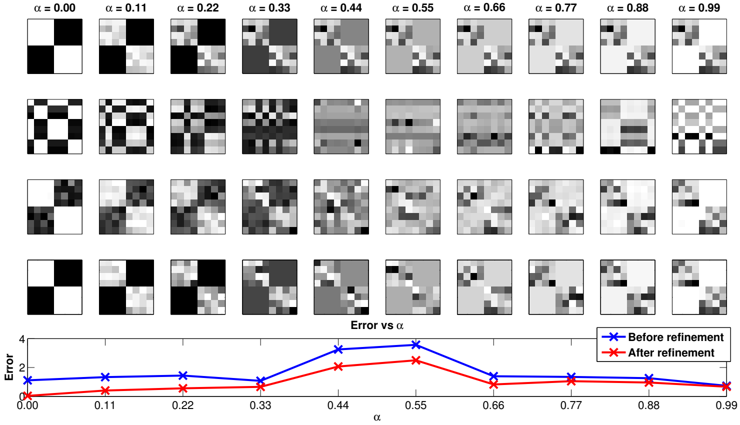
$$\bar{O} = \begin{bmatrix} O_1 & \dots & O_K \end{bmatrix}, \bar{A} = \begin{bmatrix} B_{1,1}A_1 & B_{1,2}\frac{\mathbf{1}\mathbf{1}^\top}{M} & \dots & B_{1,M}\frac{\mathbf{1}\mathbf{1}^\top}{M} \\ B_{2,1}\frac{\mathbf{1}\mathbf{1}^\top}{M} & B_{2,2}A_2 & \dots & B_{2,M}\frac{\mathbf{1}\mathbf{1}^\top}{M} \\ & & \ddots & \\ B_{M,1}\frac{\mathbf{1}\mathbf{1}^\top}{M} & B_{M,2}\frac{\mathbf{1}\mathbf{1}^\top}{M} & \dots & B_{M,M}A_K \end{bmatrix},$$

$$\bar{\nu} = \begin{bmatrix} \pi_1\nu_1 & \pi_2\nu_2 & \dots & \pi_K\nu_K \end{bmatrix}^\top.$$

- How to impose this structural constraint on the estimator?
- Use the same de-permutation method as MHMM.

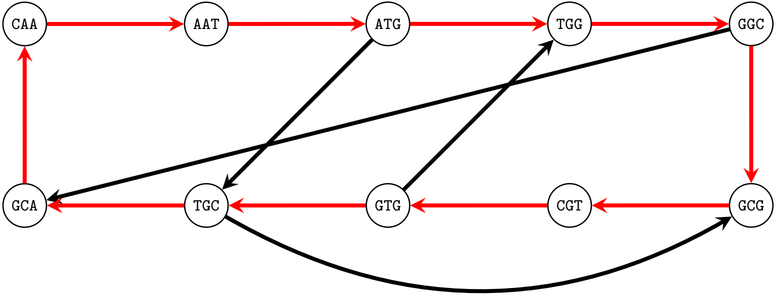
MHMM-SHMM spectrum

- $A_{i,i} \sim \text{Dirichlet}(1, \ldots, 1)$, $B = \begin{bmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{bmatrix}$.

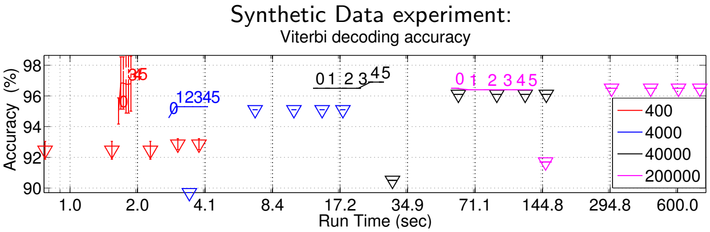


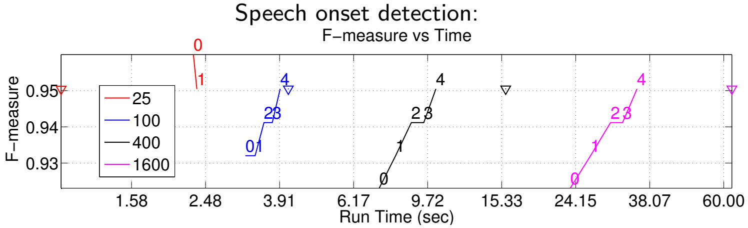
Bakis-HMM

- Is an HMM that can only move one state at a time.
- $$A = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & \dots & 0 & 0 \\ 0 & \dots & \ddots & \dots & 0 \\ 0 & \dots & 1 & 1 & 0 \\ 0 & \dots & 0 & 1 & 1 \end{bmatrix}$$
- Every state is visited exactly once.
- **Depermutation:** Find a maximum weight Hamiltonian circuit on $\widehat{A}$. (Traveling Salesman problem)



Experimental Results





Impressions on MoM:

- Good:
  - **Global**
  - **Initialization:** No need to worry about initialization. Also can initialize EM.
  - **Scalable:** Computationally cheap: Gather the moments, factorize a small matrix.
  - **Interesting/Theoretical:** Bounds.
  - **Subroutine:** Potentially can be used as a subroutine under EM.

Impressions on MoM:

- Good:
  - **Global**
  - **Initialization:** No need to worry about initialization. Also can initialize EM.
  - **Scalable:** Computationally cheap: Gather the moments, factorize a small matrix.
  - **Interesting/Theoretical:** Bounds.
  - **Subroutine:** Potentially can be used as a subroutine under EM.
- Bad:
  - **Model Mismatch:** Horrible in regards to model mismatch. (Hard assumption on model Unlike ML, which minimizes $KL(p\|q)$.
  - Not as statistically efficient as ML.

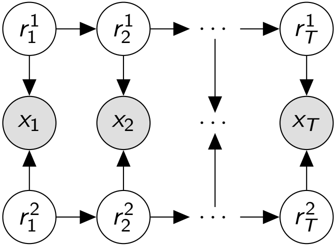
Impressions on MoM:

- Good:
  - **Global**
  - **Initialization:** No need to worry about initialization. Also can initialize EM.
  - **Scalable:** Computationally cheap: Gather the moments, factorize a small matrix.
  - **Interesting/Theoretical:** Bounds.
  - **Subroutine:** Potentially can be used as a subroutine under EM.
- Bad:
  - **Model Mismatch:** Horrible in regards to model mismatch. (Hard assumption on model Unlike ML, which minimizes $KL(p\|q)$.
  - Not as statistically efficient as ML.
- Ugly:
  - You can get complex numbers for parameter estimates/likelihoods.

Plan

Factorial HMM

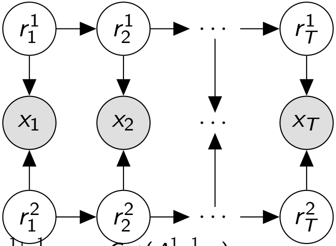
[Ghahramani, Jordan; 97]



$$
\begin{aligned}
r_t^1 | r_{t-1}^1 &\sim Cat(A^1 r_{t-1}^1) \\
&\vdots \\
r_t^K | r_{t-1}^K &\sim Cat(A^K r_{t-1}^K) \\
x_t | r_t^1, \ldots, r_t^K &\sim \mathcal{N}\left([O^1, \ldots, O^K] \begin{bmatrix} r_t^1 \\ \ldots \\ r_t^K \end{bmatrix}, \sigma^2 I\right)
\end{aligned}
$$

Factorial HMM

[Ghahramani, Jordan; 97]



$$
\begin{aligned}
r_t^1 | r_{t-1}^1 &\sim Cat(A^1 r_{t-1}^1) \\
&\vdots \\
r_t^K | r_{t-1}^K &\sim Cat(A^K r_{t-1}^K) \\
x_t | r_t^1, \dots, r_t^K &\sim \mathcal{N}([O^1, \dots, O^K] \begin{bmatrix} r_t^1 \\ \dots \\ r_t^K \end{bmatrix}, \sigma^2 I)
\end{aligned}
$$

$$X = \underbrace{O}_{\text{The dictionary}} \underbrace{R}_{\text{Activations}} + \underbrace{\epsilon}_{\text{noise}}$$

Some Dictionary Learning Perspective..

- General Dictionary Learning

$$\min_{O,R} \|X - \underbrace{O}_{Dictionary} \underbrace{R}_{Activations} \|_F$$

  - **PCA:** Both O and R are orthogonal.
  - **ICA:** Solvable if R has independent coordinates.
  - **Mixture Model:** R is one sparse. Solvable is O has full column rank.
  - **Sparse Dictionary Learning:** Solvable if O is square and R is sparse Bernouilli-Gaussian. [Spielman et al. 12]

Some Dictionary Learning Perspective..

- General Dictionary Learning

$$\min_{O,R} \|X - \underbrace{O}_{Dictionary} \underbrace{R}_{Activations} \|_F$$

  - **PCA:** Both O and R are orthogonal.
  - **ICA:** Solvable if R has independent coordinates.
  - **Mixture Model:** R is one sparse. Solvable is O has full column rank.
  - **Sparse Dictionary Learning:** Solvable if O is square and R is sparse Bernouilli-Gaussian. [Spielman et al. 12]

- Factorial Models:

$$O = \begin{bmatrix} O^1 & \dots & O^K \end{bmatrix},\ R = \begin{bmatrix} R^1 \\ \vdots \\ R^K \end{bmatrix}$$

  - No constraint on O, columns of R are block-K sparse.
  - No Unique Solution!!!

FHMM Identifiability

Rank Deficiency

$\mathbf{rank}(R) \leq MK - (K - 1)$

FHMM Identifiability

Rank Deficiency

$\mathbf{rank}(R) \leq MK - (K-1)$

Proof Sketch:

$$\dim(\text{null}(R^\top)) \geq K - 1.$$

Therefore from rank-nullity theorem $\mathbf{rank}(R) = MK - (K-1)$.

FHMM is unidentifiable

For a given assignment matrix $R \in \mathbb{R}^{KM \times T}$ There exists $O_1 \neq O_2$ such that $\prod_t \mathcal{N}(x_t | O_1 R, \sigma^2 I) = \prod_t \mathcal{N}(x_t | O_2 R, \sigma^2 I)$.

FHMM Identifiability

Rank Deficiency

$\mathbf{rank}(R) \leq MK - (K - 1)$

Proof Sketch:

$$\dim(\text{null}(R^\top)) \geq K - 1.$$

Therefore from rank-nullity theorem $\mathbf{rank}(R) = MK - (K - 1)$.

FHMM is unidentifiable

For a given assignment matrix $R \in \mathbb{R}^{KM \times T}$ There exists $O_1 \neq O_2$ such that $\prod_t \mathcal{N}(x_t | O_1 R, \sigma^2 I) = \prod_t \mathcal{N}(x_t | O_2 R, \sigma^2 I)$.

Proof: Since $\dim(\text{null}(R^\top)) \geq K - 1$, $(O_1 - O_2)R = 0$, for $O_1 \neq O_2$.

FHMM Identifiable Alternative 1

Shared Component FM

$$\forall k,\ O^k = \begin{bmatrix} | & | & | & | & | \\ \mu_k^1 & \mu_2^k & \dots & \mu_{M-1}^k & s \\ | & | & | & | & | \end{bmatrix}$$

SC-FM is identifiable

Given an assignment matrix $\widetilde{R}$ which is rank $MK - (K-1)$, the emission matrix of an SC-FM is identifiable.

Proof Sketch:

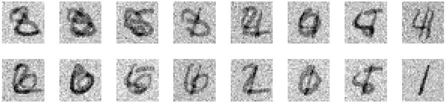
$$\dim(\text{null}(R^\top)) = 0.$$

Therefore $(O_1 - O_2)R \neq 0,\ \forall\ O_1 \neq O_2$.

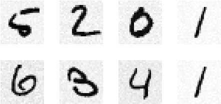
Learning Example for Shared Component Factorial Model

- **Gist:** If the shared component s is incoherent, then we can identify it, and reveal the other components.

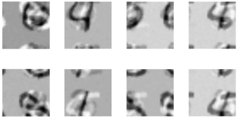
Example Observations



Obtained Components with SC-FM



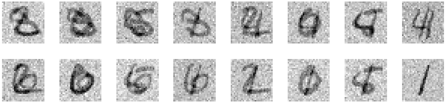
Components with regular model-EM



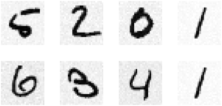
Learning Example for Shared Component Factorial Model

- **Gist:** If the shared component s is incoherent, then we can identify it, and reveal the other components.

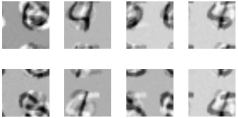
Example Observations



Obtained Components with SC-FM



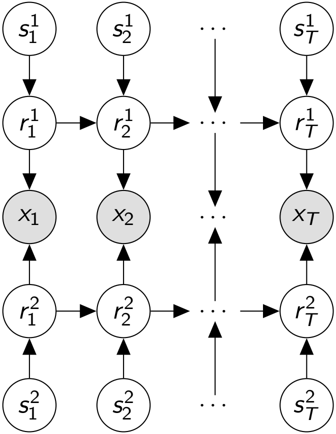
Components with regular model-EM



- The shared component + incoherence assumption a bit too restrictive. Can we think of another model?

FHMM Identifiable Alternative 2

$$\begin{aligned} s_t^k &\sim Bernouilli(\pi),\; k \in \{1, \dots, K\} \\ r_t^1 | r_{t-1}^1 &\sim s_t^1 Cat(A^1 r_{t-1}^1) \\ &\vdots \\ r_t^K | r_{t-1}^K &\sim s_t^K Cat(A^K r_{t-1}^K) \\ x_t | r_t^1, \dots, r_t^K &\sim \mathcal{N}([O^1, \dots, O^K] \begin{bmatrix} r_t^1 \\ \dots \\ r_t^K \end{bmatrix}, \sigma^2 I) \end{aligned}$$



- Identifiability follows similarly from the activation matrix R.

Revealing FHMM Practical Algorithm

Practical Algorithm for Revealing FHMM

- Cluster the data matrix $X \in \mathbb{R}^{L\times T}$ into clusters $X^c \in \mathbb{R}^{L\times C}$.
- Solve:

$$\begin{aligned} \min_H \|X^c - X^c H\|_F^2 + \beta \|H\|_1, \\ s.t.\ H_{i,i} = 0,\ \text{for } 1 \leq i \leq C, \\ H \geq 0, \end{aligned}$$

  where $H \in \mathbb{R}^{C\times C}$.
- Construct a bi-partite graph by reading the solution for H.

Revealing FHMM Practical Algorithm

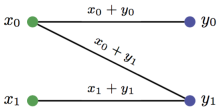
Practical Algorithm for Revealing FHMM

- Cluster the data matrix $X \in \mathbb{R}^{L \times T}$ into clusters $X^c \in \mathbb{R}^{L \times C}$.
- Solve:

$$\begin{aligned} &\min_H \|X^c - X^c H\|_F^2 + \beta \|H\|_1, \\ &s.t.\ H_{i,i} = 0,\ \text{for } 1 \leq i \leq C, \\ &H \geq 0, \end{aligned}$$

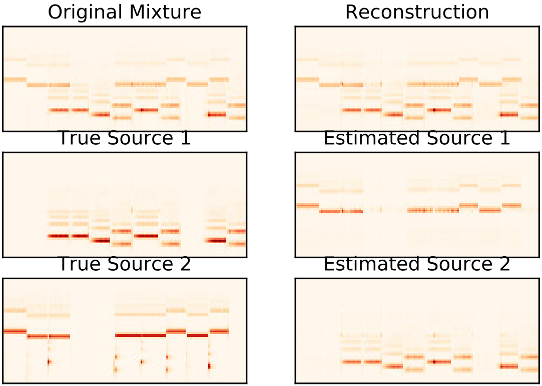
  where $H \in \mathbb{R}^{C \times C}$.
- Construct a bi-partite graph by reading the solution for H.

- **Condition for learnability:** Let $O_1 = [x_0, x_1]$, $O_2 = [y_0, y_1]$. Observed combinations needs to form a connected bi-partite graph (Connectivity) *(linear number of edges in number of components, not quadratic)*, and we need to observe all nodes and edges (Observability).



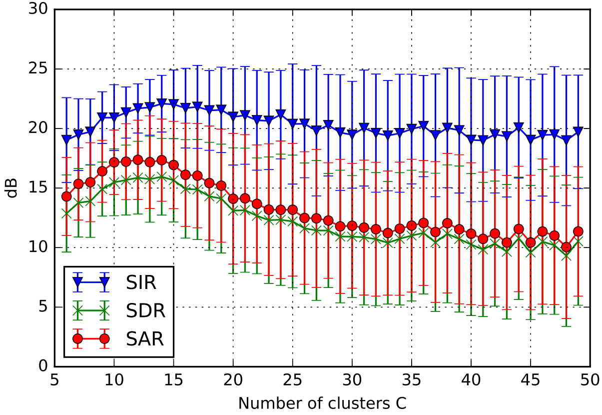
Unsupervised audio source separation example

- We mixed recording of double bass and flute (at 0dB).
- The observed mixtures satisfy the connectivity constraint.



- We obtain almost perfect source separation.

Sensitivity on number of clusters



- The algorithm is robust to the choice of number of clusters C.

Conclusions on FHMM

- **Identifiability:** The original FHMM model is unidentifiable.
- **Identifiable Alternatives:** There exists identifiable alternatives which are globally learnable under stringent assumptions.
- **Unsupervised Source Separation:** Revealing FHMM works well under the connectedness and observability assumptions.
- **Future work:**
  - Can we relax the observability assumption so that we only require to observe less nodes in the connectivity graph?
  - Potential application in semi-supervised source separation.