Frontiers to the learning of nonparametric Hidden Markov Models

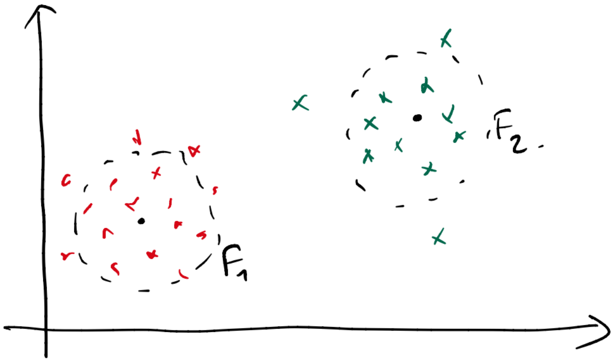
Kweku Abraham, Elisabeth Gassiat, **Zacharie Naulet**

https://arxiv.org/abs/2306.16293

Fréjus, September 18th, 2023

Mixture models

Mixture models are used to model data coming from unknown populations $F_1, F_2, \ldots, F_K$.



Conditional on the latent class $X_i \in \{1, \ldots, K\}$:

$$Y_i \mid X_i \overset{ind}{\sim} F_{X_i}$$

Widely used for model based clustering

[See Ibrahim Kaddouri's poster Tuesday night!]

Example: iid mixtures

Model:

$$X_1, \ldots, X_n \overset{iid}{\sim} (\pi_1, \ldots, \pi_K)$$
$$Y_i \mid X_i \overset{iid}{\sim} F_{X_i} \qquad i = 1, \ldots, n$$

Example: iid mixtures

Model:

$$X_1, \ldots, X_n \overset{iid}{\sim} (\pi_1, \ldots, \pi_K)$$
$$Y_i \mid X_i \overset{iid}{\sim} F_{X_i} \qquad i = 1, \ldots, n$$

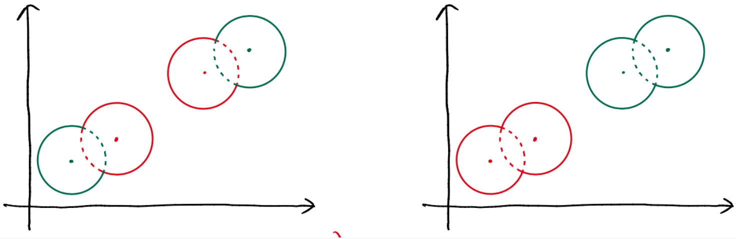
They are not nonparametrically identifiable.

eg. $K = 2$, model parameters are then $\theta = (\pi, 1-\pi, F_0, F_1)$.

Law of $(Y_1, \ldots, Y_n)$ under $\theta = (1/4, 3/4, F_0, F_1)$

$$\begin{aligned} P_\theta^{(n)}(A_1 \times \cdots \times A_n) &= \prod_{i=1}^n \left(\frac{1}{4} F_0(A_i) + \frac{3}{4} F_1(A_i) \right) \\ &= \prod_{i=1}^n \left(\frac{1}{2} \cdot \frac{F_0(A_i) + F_1(A_i)}{2} + \frac{1}{2} F_1(A_i) \right) \\ &= P_{\theta'}^{(n)}(A_1 \times \cdots \times A_n), \qquad \theta' = \left(1/2, 1/2, \frac{F_0 + F_1}{2}, F_1\right). \end{aligned}$$

Another example:



$$F_0 = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}4\\3\end{pmatrix}, I_2\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}8\\6\end{pmatrix}, I_2\right) \qquad F_0' = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}2\\2\end{pmatrix}, I_2\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}4\\3\end{pmatrix}, I_2\right)$$

$$F_1 = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}2\\2\end{pmatrix}, I_2\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}10\\7\end{pmatrix}, I_2\right) \qquad F_1' = \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}8\\6\end{pmatrix}, I_2\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}10\\7\end{pmatrix}, I_2\right)$$

$$P^{(n)}_{(1/2,1/2,F_0,F_1)} = P^{(n)}_{(1/2,1/2,F_0',F_1')}$$

Example: Hidden Markov Models (HMM)

Model:

$$X_1, X_2, \ldots \sim \mathrm{Markov}(Q, \pi)$$

$$Y_i \mid X_i \overset{ind}{\sim} F_{X_i} \qquad i = 1, \ldots, n$$

Example: Hidden Markov Models (HMM)

Model:

$$X_1, X_2, \ldots \sim \mathrm{Markov}(Q, \pi)$$
$$Y_i \mid X_i \overset{ind}{\sim} F_{X_i} \qquad i = 1, \ldots, n$$

They are nonparametrically identifiable!

eg. $K = 2$, model parameters are $\theta = (Q, \pi, F_0, F_1)$, and law of $(Y_1, \ldots, Y_n)$ is:

$$P_\theta^{(n)}(A_1 \times \cdots \times A_n) = \sum_{x \in \{0,1\}^n} \pi(x_1) \prod_{i=1}^{n-1} Q(x_i, x_{i+1}) \prod_{i=1}^{n} F_{x_i}(A_i)$$

Example: Hidden Markov Models (HMM)

Model:

$$X_1, X_2, \ldots \sim \mathrm{Markov}(Q, \pi)$$
$$Y_i \mid X_i \overset{ind}{\sim} F_{X_i} \qquad i = 1, \ldots, n$$

They are nonparametrically identifiable!

eg. $K = 2$, model parameters are $\theta = (Q, \pi, F_0, F_1)$, and law of $(Y_1, \ldots, Y_n)$ is:

$$P_\theta^{(n)}(A_1 \times \cdots \times A_n) = \sum_{x \in \{0,1\}^n} \pi(x_1) \prod_{i=1}^{n-1} Q(x_i, x_{i+1}) \prod_{i=1}^{n} F_{x_i}(A_i)$$

Theorem 1 (Allman, Matias, and Rhodes 2009).

If $n \geq 3$ and $(Y_1, \ldots, Y_n)$ are "truly dependent" then θ is identifiable from $P_\theta^{(n)}$ up to label switching [ie $\theta \mapsto \mathbb{P}_\theta^{(n)}$ is invertible up to permutation of the population labels].

Settings: two-state HMM

Binary latent variables $\mathbf{X} = (X_1, X_2, \dots) \in \{0,1\}^{\mathbb{N}}$,

$$\mathbf{X} = (X_n)_{n\in\mathbb{N}} \sim \text{Stat. Markov}(Q), \qquad Q = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

$$Y_n \mid X_n \sim F_{X_n}$$

We denote $\theta = (p, q, F_0, F_1)$,

$$\pi_0 := \mathbb{P}_\theta\left(X_1 = 0\right) = \frac{q}{p+q}, \qquad \pi_1 := \mathbb{P}_\theta\left(X_1 = 1\right) = \frac{p}{p+q}.$$

HMM: identifiability

From the identifiability Theorem, $(Y_1, \ldots, Y_n)$ are independent iff one of the three condition holds:

1. $(X_1, \ldots, X_n)$ are independent $\iff 1 - p - q = 0$;
2. $X_1 = X_2 = \cdots = X_n$ almost-surely $\iff p = 0$ or $q = 0$;
3. $F_0 = F_1$.

HMM: identifiability

From the identifiability Theorem, $(Y_1, \dots, Y_n)$ are independent iff one of the three condition holds:

1. $(X_1, \dots, X_n)$ are independent $\iff 1 - p - q = 0$;
2. $X_1 = X_2 = \cdots = X_n$ almost-surely $\iff p = 0$ or $q = 0$;
3. $F_0 = F_1$.

If none of the above hold, then the model is identifiable.

BUT

Can we estimate θ (modulo label-switching) from $(Y_1, \dots, Y_n) \sim P_\theta^{(n)}$?

HMM: identifiability

From the identifiability Theorem, $(Y_1, \ldots, Y_n)$ are independent iff one of the three condition holds:

1. $(X_1, \ldots, X_n)$ are independent $\iff 1 - p - q = 0$;
2. $X_1 = X_2 = \cdots = X_n$ almost-surely $\iff p = 0$ or $q = 0$;
3. $F_0 = F_1$.

If none of the above hold, then the model is identifiable.

BUT

Can we estimate θ (modulo label-switching) from $(Y_1, \ldots, Y_n) \sim P_\theta^{(n)}$?

We analyze the minimax risk over

$$\Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R) := \Big\{\theta \, : \, p, q \geq \delta, \; |1 - p - q| \geq \epsilon, \; \|f_0 - f_1\|_{L^2} \geq \zeta, \; \|f_i\|_{B^{s_i}_{2,\infty}} \leq R\Big\}.$$

Rough statement of the results

[We ignore label-switching issues for simplification]

Estimation of Q

$$\inf_{\hat{Q}} \sup_{\theta \in \Theta^{s_0, s_1}_{\delta, \epsilon, \zeta}(R)} \mathbb{E}_\theta \left(\|\hat{Q} - Q\|^2 \right) \asymp \frac{\max(\delta, \epsilon\zeta)^2}{\delta^2 \epsilon^4 \zeta^6} \frac{1}{n}.$$

Rough statement of the results

[We ignore label-switching issues for simplification]

Estimation of Q

$$\inf_{\hat{Q}} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{Q} - Q\|^2\right) \asymp \frac{\max(\delta, \epsilon\zeta)^2}{\delta^2 \epsilon^4 \zeta^6} \frac{1}{n}.$$

Estimation of F_0 and F_1 on $[0, 1]$

The minimax rate for estimating the densities exhibit a transition:

- If $s_0 = s_1 = s$:

$$\inf_{\hat{f}_j} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{f}_j - f_j\|^2_{L^2}\right) \asymp \left(\frac{1}{\delta^2 \epsilon^2 \zeta^2 n}\right)^{2s/(2s+1)} + \frac{1}{\delta^2 \epsilon^2 \zeta^4 n}$$

Rough statement of the results

[We ignore label-switching issues for simplification]

Estimation of Q

$$\inf_{\hat{Q}} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{Q} - Q\|^2\right) \asymp \frac{\max(\delta, \epsilon\zeta)^2}{\delta^2\epsilon^4\zeta^6}\frac{1}{n}.$$

Estimation of F_0 and F_1 on $[0, 1]$

The minimax rate for estimating the densities exhibit a transition:

- If $s_0 = s_1 = s$:

$$\inf_{\hat{f}_j} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{f}_j - f_j\|^2_{L^2}\right) \asymp \left(\frac{1}{\delta^2\epsilon^2\zeta^2 n}\right)^{2s/(2s+1)} + \frac{1}{\delta^2\epsilon^2\zeta^4 n}$$

- If $s_0 > s_1$:

$$\inf_{\hat{f}_0} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{f}_0 - f_0\|^2_{L^2}\right) \asymp \left(\frac{1}{\delta^2\epsilon^2\zeta^2 n}\right)^{2s_0/(2s_0+1)} + \frac{1}{\delta^2\epsilon^2\zeta^4 n},$$

$$\inf_{\hat{f}_1} \sup_{\theta \in \Theta^{s_0,s_1}_{\delta,\epsilon,\zeta}(R)} \mathbb{E}_\theta\left(\|\hat{f}_1 - f_1\|^2_{L^2}\right) \asymp \left(\frac{1}{\delta^2 n}\right)^{2s_1/(2s_1+1)} + \frac{1}{\delta^2\epsilon^2\zeta^4 n}.$$

Some remarks:

1. Unexpected transition in the rates when $s_0 \neq s_1$.
2. If the latent variables $\mathbf{X} = (X_1, X_2, \dots)$ were known, then the minimax rates would be in any cases:

$$\left(\frac{1}{\delta n}\right)^{2s/(2s+1)}$$

where δn corresponds to the worse average size of the smallest cluster.

$\Longrightarrow$ *Effective sample size* goes from δn when $\mathbf{X}$ is known to $\delta^2\epsilon^2\zeta^2 n$ when $\mathbf{X}$ unknown (much harder!)

Estimation strategy (ie. proving the upper bounds) when $s_0 = s_1$

We construct a wavelet estimator using the CDV[1] basis $(\Psi_{jk})_{jk}$.

For simplicity we identify in the next $f_m \equiv (f_m^{\Psi_{jk}})_{jk}$.

[1]Cohen, Daubechies, and Vial 1993
[2]Reminiscent to the spectral method of Anandumar et al. 2014

Estimation strategy (ie. proving the upper bounds) when $s_0 = s_1$

We construct a wavelet estimator using the CDV[1] basis $(\Psi_{jk})_{jk}$.

For simplicity we identify in the next $f_m \equiv (f_m^{\Psi_{jk}})_{jk}$.

Inspired by the identifiability Theorem, for any h the map

$$\mathcal{M}_h : (p, q, (f_0^{\Psi_{jk}}), (f_1^{\Psi_{jk}})) \mapsto \Big(\mathbb{E}_\theta(\cdot),\ \mathbb{E}_\theta(h \otimes \cdot),\ \mathbb{E}_\theta(h \otimes 1 \otimes h),\ \mathbb{E}_\theta(h \otimes h \otimes h)\Big)$$

can be inverted (modulo label-switching) provided $\langle h, f_0 - f_1 \rangle \neq 0$[2].

[1] Cohen, Daubechies, and Vial 1993

[2] Reminiscent to the spectral method of Anandumar et al. 2014

The estimation strategy then goes as follows:

1. Find a good h.
2. Using the method of moments, we obtain estimators of $(p, q, (f_0^{\Psi_{jk}}), (f_1^{\Psi_{jk}}))$ by letting

$$(\hat{p}, \hat{q}, (\hat{f}_0^{\Psi_{jk}}), (\hat{f}_1^{\Psi_{jk}})) = \mathcal{M}_h^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i},\ \frac{1}{n-1}\sum_{i=1}^{n-1}h(Y_i)\delta_{Y_{i+1}},\right.$$
$$\left.\frac{1}{n-3}\sum_{i=1}^{n-2}h(Y_i)h(Y_{i+2}),\ \frac{1}{n-3}\sum_{i=1}^{n-2}h(Y_i)h(Y_{i+1})h(Y_{i+2})\right).$$

3. Construct block-thresholded wavelet estimators $\hat{f}_0$ and $\hat{f}_1$. [not so easy! in contrast with density estimation the optimal thresholds depend on the parameters].

The estimation strategy then goes as follows:

1. Find a good h.
2. Using the method of moments, we obtain estimators of $(p, q, (f_0^{\Psi_{jk}}), (f_1^{\Psi_{jk}}))$ by letting

$$(\hat{p}, \hat{q}, (\hat{f}_0^{\Psi_{jk}}), (\hat{f}_1^{\Psi_{jk}})) = \mathcal{M}_h^{-1}\Bigg(\frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i},\ \frac{1}{n-1}\sum_{i=1}^{n-1}h(Y_i)\delta_{Y_{i+1}},$$
$$\frac{1}{n-3}\sum_{i=1}^{n-2}h(Y_i)h(Y_{i+2}),\ \frac{1}{n-3}\sum_{i=1}^{n-2}h(Y_i)h(Y_{i+1})h(Y_{i+2})\Bigg).$$

3. Construct block-thresholded wavelet estimators $\hat{f}_0$ and $\hat{f}_1$. [not so easy! in contrast with density estimation the optimal thresholds depend on the parameters].

This will attain:

$$\inf_{\hat{f}_j} \sup_{\theta \in \Theta_{\delta,\epsilon,\zeta}^{s_0,s_1}(R)} \mathbb{E}_\theta\left(\|\hat{f}_j - f_j\|_{L^2}^2\right) \lesssim \left(\frac{1}{\delta^2\epsilon^2\zeta^2 n}\right)^{2s_j/(2s_j+1)} + \frac{1}{\delta^2\epsilon^2\zeta^4 n}$$

What does it mean to find a good h?

1. Find a good h.

The inverse map $\mathcal{M}_h^{-1}$ is unstable for poor choice of h.

[3] Anandumar et al. 2014; Moss and Rousseau 2022; Abraham, Castillo, and Gassiat 2021; Lehéricy 2018; etc.

What does it mean to find a good h?

1. Find a good h.

The inverse map $\mathcal{M}_h^{-1}$ is unstable for poor choice of h.

To avoid instabilities and achieve optimality we need $c \geq 1/2$ such that

$$|\langle f_0 - f_1, h\rangle| \geq c\|f_0 - f_1\|_{L^2}\|h\|_{L^2}. \qquad \text{(Separating Hyperplane Condition)}$$

[3] Anandumar et al. 2014; Moss and Rousseau 2022; Abraham, Castillo, and Gassiat 2021; Lehéricy 2018; etc.

What does it mean to find a good h?

1. Find a good h.

The inverse map $\mathcal{M}_h^{-1}$ is unstable for poor choice of h.

To avoid instabilities and achieve optimality we need $c \geq 1/2$ such that

$$|\langle f_0 - f_1, h\rangle| \geq c\|f_0 - f_1\|_{L^2}\|h\|_{L^2}. \qquad \text{(Separating Hyperplane Condition)}$$

Previous works[3] on estimation in HMM faces similar issues and suggest to choose h at random...

This will eventually work for fixed f_0, f_1 but cannot be minimax optimal.

[3] Anandumar et al. 2014; Moss and Rousseau 2022; Abraham, Castillo, and Gassiat 2021; Lehéricy 2018; etc.

What does it mean to find a good h?

1. Find a good h.

The inverse map $\mathcal{M}_h^{-1}$ is unstable for poor choice of h.

To avoid instabilities and achieve optimality we need $c \geq 1/2$ such that

$$|\langle f_0 - f_1, h \rangle| \geq c \|f_0 - f_1\|_{L^2} \|h\|_{L^2}. \qquad \text{(Separating Hyperplane Condition)}$$

Previous works[3] on estimation in HMM faces similar issues and suggest to choose h at random...

This will eventually work for fixed f_0, f_1 but cannot be minimax optimal.

h must be estimated from the data!

[3] Anandumar et al. 2014; Moss and Rousseau 2022; Abraham, Castillo, and Gassiat 2021; Lehéricy 2018; etc.

Failure of the separating hyperplane condition

Suppose $(e_j)_{j \geq 1}$ is an orthonormal basis for $L^2[0,1]$ and choose

$$h = \sum_{k=1}^{d} W_k e_k \qquad\qquad W_k \overset{iid}{\sim} \mathcal{N}(0,1).$$

Failure of the separating hyperplane condition

Suppose $(e_j)_{j\geq 1}$ is an orthonormal basis for $L^2[0,1]$ and choose

$$h = \sum_{k=1}^{d} W_k e_k \qquad\qquad W_k \overset{iid}{\sim} \mathcal{N}(0,1).$$

Then with Π_d orthogonal projection onto $\mathrm{span}(e_1, \ldots, e_d)$:

$$\frac{\|h\|_{L^2}}{\sqrt{d}} \xrightarrow[d\to\infty]{as} 1, \qquad\qquad \langle f_0 - f_1, h\rangle \sim \mathcal{N}\big(0, \|\Pi_d(f_0 - f_1)\|_{L^2}^2\big)$$

so

$$\langle f_0 - f_1, h\rangle \approx \frac{\mathcal{N}(0,1)}{\sqrt{d}} \|\Pi_d(f_0 - f_1)\|_{L^2} \|h\|_{L^2}.$$

Failure of the separating hyperplane condition

Suppose $(e_j)_{j\geq 1}$ is an orthonormal basis for $L^2[0,1]$ and choose

$$h = \sum_{k=1}^{d} W_k e_k \qquad\qquad W_k \overset{iid}{\sim} \mathcal{N}(0,1).$$

Then with Π_d orthogonal projection onto $\mathrm{span}(e_1,\dots,e_d)$:

$$\frac{\|h\|_{L^2}}{\sqrt{d}} \xrightarrow[d\to\infty]{as} 1, \qquad\qquad \langle f_0 - f_1, h\rangle \sim \mathcal{N}\big(0, \|\Pi_d(f_0 - f_1)\|_{L^2}^2\big)$$

so

$$\langle f_0 - f_1, h\rangle \approx \frac{\mathcal{N}(0,1)}{\sqrt{d}} \|\Pi_d(f_0 - f_1)\|_{L^2} \|h\|_{L^2}.$$

Problem: having $\|\Pi_d(f_0 - f_1)\|_{L^2} \approx \|f_0 - f_1\|_{L^2}$ can require d large.

Key ideas to compute $\mathcal{M}_h^{-1}$

2. Method of moments, invert $\mathcal{M}_h$

(Invertible) Reparameterization: $\theta \mapsto (\phi_1, \phi_2, \phi_3, \psi_1, \psi_2)$ such that

- sparsity $\iff |\phi_1|$ near 1,
- near independence of $\mathbf{X} \iff |\phi_2|$ near 0,
- populations not well separated $\iff |\phi_3|$ near 0,
- ψ_1 is the invariant distribution of the X_i's,
- ψ_2 is a direction.

Key ideas to compute $\mathcal{M}_h^{-1}$

2. Method of moments, invert $\mathcal{M}_h$

(Invertible) Reparameterization: $\theta \mapsto (\phi_1, \phi_2, \phi_3, \psi_1, \psi_2)$ such that

- sparsity $\iff |\phi_1|$ near 1,
- near independence of $\mathbf{X} \iff |\phi_2|$ near 0,
- populations not well separated $\iff |\phi_3|$ near 0,
- ψ_1 is the invariant distribution of the X_i's,
- ψ_2 is a direction.

The magic formula:

$$\begin{aligned} p_\theta^{(3)} = \psi_1 \otimes \psi_1 \otimes \psi_1 &+ \frac{1}{4}(1-\phi_1^2)\phi_2\phi_3^2\Big(\psi_2 \otimes \psi_2 \otimes \psi_1 + \psi_1 \otimes \psi_2 \otimes \psi_2\Big) \\ &+ \frac{1}{4}(1-\phi_1^2)\phi_2^2\phi_3^2 \cdot \psi_2 \otimes \psi_1 \otimes \psi_2 \\ &- \frac{1}{4}(1-\phi_1^2)\phi_1\phi_2^2\phi_3^3 \cdot \psi_2 \otimes \psi_2 \otimes \psi_2 . \end{aligned}$$

From here we can easily extract $(1-\phi_1^2)\phi_2\phi_3^2$, $(1-\phi_1^2)\phi_2^2\phi_3^2$, $(1-\phi_1^2)\phi_1\phi_2^2\phi_3^3$ as well as the wavelets coefficients of ψ_1 and ψ_2.

Key ideas to compute $\mathcal{M}_h^{-1}$

2. Method of moments, invert $\mathcal{M}_h$

(Invertible) Reparameterization: $\theta \mapsto (\phi_1, \phi_2, \phi_3, \psi_1, \psi_2)$ such that

- sparsity $\iff |\phi_1|$ near 1,
- near independence of $\mathbf{X} \iff |\phi_2|$ near 0,
- populations not well separated $\iff |\phi_3|$ near 0,
- ψ_1 is the invariant distribution of the X_i's,
- ψ_2 is a direction.

The magic formula:

$$
\begin{aligned}
p_\theta^{(3)} = \psi_1 \otimes \psi_1 \otimes \psi_1 &+ \frac{1}{4}(1-\phi_1^2)\phi_2\phi_3^2\Big(\psi_2 \otimes \psi_2 \otimes \psi_1 + \psi_1 \otimes \psi_2 \otimes \psi_2\Big) \\
&+ \frac{1}{4}(1-\phi_1^2)\phi_2^2\phi_3^2 \cdot \psi_2 \otimes \psi_1 \otimes \psi_2 \\
&- \frac{1}{4}(1-\phi_1^2)\phi_1\phi_2^2\phi_3^3 \cdot \psi_2 \otimes \psi_2 \otimes \psi_2 .
\end{aligned}
$$

From here we can easily extract $(1-\phi_1^2)\phi_2\phi_3^2$, $(1-\phi_1^2)\phi_2^2\phi_3^2$, $(1-\phi_1^2)\phi_1\phi_2^2\phi_3^3$ as well as the wavelets coefficients of ψ_1 and ψ_2.

Exponential deviations for moments: Paulin 2015.

Estimation strategy when $s_0 > s_1$ (rough)

Previous estimators not always optimal!

Stationnary distribution of $(Y_1, Y_2, \dots)$:

$$\psi_1 = \pi_0 f_0 + (1 - \pi_0) f_1$$

so

$$f_1 = \frac{1}{1 - \pi_0}\Big(\psi_1 - \pi_0 f_0\Big)$$

Estimation strategy when $s_0 > s_1$ (rough)

Previous estimators not always optimal!

Stationnary distribution of $(Y_1, Y_2, \ldots)$:

$$\psi_1 = \pi_0 f_0 + (1 - \pi_0) f_1$$

so

$$f_1 = \frac{1}{1 - \pi_0} \Big(\psi_1 - \pi_0 f_0 \Big)$$

When $s_0 > s_1$:

- ψ_1 has (morally) smoothness s_1 and can be easily estimated using your favorite density estimator;
- f_0 can be estimated at a much faster rate than ψ_1 since it is smoother;

Estimation strategy when $s_0 > s_1$ (rough)

Previous estimators not always optimal!

Stationnary distribution of $(Y_1, Y_2, \dots)$:

$$\psi_1 = \pi_0 f_0 + (1 - \pi_0) f_1$$

so

$$f_1 = \frac{1}{1 - \pi_0} \Big(\psi_1 - \pi_0 f_0 \Big)$$

When $s_0 > s_1$:

- ψ_1 has (morally) smoothness s_1 and can be easily estimated using your favorite density estimator;
- f_0 can be estimated at a much faster rate than ψ_1 since it is smoother;

So when $s_0 > s_1$ we introduce the *rough estimator* based on the above heuristic:

$$\hat{f}_1^R$$

Proofs of the lower bounds

We use the "traditional" Fano-Birgé device.

Main challenge is computing $\mathrm{KL}(P_\theta^{(n)}; P_{\tilde\theta}^{(n)})$; $\theta = (p, q, f_0, f_1)$, $\tilde\theta = (\tilde p, \tilde q, \tilde f_0, \tilde f_1)$

Proofs of the lower bounds

We use the "traditional" Fano-Birgé device.

Main challenge is computing $\mathrm{KL}(P_\theta^{(n)}; P_{\tilde\theta}^{(n)})$; $\theta = (p, q, f_0, f_1)$, $\tilde\theta = (\tilde p, \tilde q, \tilde f_0, \tilde f_1)$

We use one of our earlier result that if $\min(f_0, f_1, \tilde f_0, \tilde f_1) \geq c$, then

$$\mathrm{KL}(P_\theta^{(n)}; P_{\tilde\theta}^{(n)}) \asymp n\|p_\theta^{(3)} - p_{\tilde\theta}^{(3)}\|^2.$$

and then we use the magic formula to control

$$\begin{aligned}
\|p_\theta^{(3)} - p_{\tilde\theta}^{(3)}\| \asymp\, & |(1-\phi_1^2)\phi_2\phi_3^2 - (1-\tilde\phi_1^2)\tilde\phi_2\tilde\phi_3^2| \\
& + |(1-\phi_1^2)\phi_2^2\phi_3^2 - (1-\tilde\phi_1^2)\tilde\phi_2^2\tilde\phi_3^2| \\
& + |(1-\phi_1^2)\phi_1\phi_2^2\phi_3^3 - \mathrm{sgn}(\langle\psi_2, \tilde\psi_2\rangle)(1-\tilde\phi_1^2)\tilde\phi_1\tilde\phi_2^2\tilde\phi_3^3| \\
& + \|\psi_1 - \tilde\psi_1\|_{L^2} \\
& + \max\Big(|(1-\phi_1^2)\phi_2\phi_3^2|,\ |(1-\tilde\phi_1^2)\tilde\phi_2\tilde\phi_3^2|\Big)\|\psi_2 - \mathrm{sgn}(\langle\psi_2, \tilde\psi_2\rangle)\tilde\psi_2\|_{L^2}.
\end{aligned}$$

Conclusion, perspectives

Take home message

- HMMs are mixture models with Markov regime that can be identified without any assumption on the population distributions as soon as they are distinct and the Markov has invertible transition thus not i.i.d.
- For 2 states HMMs, we identify how the minimax rates depend on n and being far from the non-identifying region with parameters describing the "distance" to independence.

Conclusion, perspectives

Take home message

- HMMs are mixture models with Markov regime that can be identified without any assumption on the population distributions as soon as they are distinct and the Markov has invertible transition thus not i.i.d.
- For 2 states HMMs, we identify how the minimax rates depend on n and being far from the non-identifying region with parameters describing the "distance" to independence.

Further questions

- Extension to more than two latent states?
- Algorithms: robustness; detection of problematic regions?
- Non parametric clustering for HMMs? (See Ibrahim's poster!)
- Model selection: can we choose between $(\hat{f}_0, \hat{f}_1)$, $(\hat{f}_0, \hat{f}_1^R)$, $(\hat{f}_0^R, \hat{f}_1)$, and $(\hat{f}_0^R, \hat{f}_1^R)$?
- Secondary adaptation questions...
- ...

Thank you!

References

Abraham, Kweku, Ismael Castillo, and Elisabeth Gassiat (2021). **Multiple Testing in Nonparametric Hidden Markov Models: An Empirical Bayes Approach.** arXiv: `2101.03838 [math.ST]`.

Allman, E. S., C. Matias, and J. A. Rhodes (2009). **"Identifiability of parameters in latent structure models with many observed variables".** In: *Ann. Statist.* 37.6A, pp. 3099–3132.

Anandumar, A. et al. (2014). **"Tensor decompositions for learning latent variable models".** In: *J. Mach. Learn. Res.* 15, pp. 2773–2832.

Cohen, Albert, Ingrid Daubechies, and Pierre Vial (1993). **"Wavelets on the interval and fast wavelet transforms".** In: *Appl. Comput. Harmon. Anal.* 1.1, pp. 54–81. ISSN: 1063-5203. DOI: `10.1006/acha.1993.1005`. URL: `https://doi-org.ezproxy.universite-paris-saclay.fr/10.1006/acha.1993.1005`.

Lehéricy, Luc (2018). **"Estimation adaptative pour les modèles de Markov cachés non paramétriques".** PhD thesis. Université Paris-Saclay (ComUE).

Moss, Daniel and Judith Rousseau (2022). **Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models.** arXiv: `2203.06081 [math.ST]`.

Paulin, D. (2015). **"Concentration inequalities for Markov chains by**