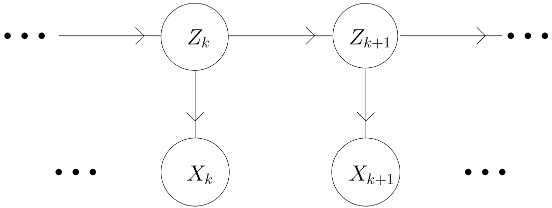
Inference in non parametric Hidden Markov Models

Elisabeth Gassiat

Université Paris-Sud (Orsay) and CNRS

Van Dantzig Seminar, June 2017

Hidden Markov models (HMMs)



Observations $(X_k)_{k\geq 1}$ are independent conditionnally to $(Z_k)_{k\geq 1}$

$$\mathcal{L}\left((X_k)_{k\geq 1}|(Z_k)_{k\geq 1}\right) = \bigotimes_{k\geq 1} \mathcal{L}\left(X_k|Z_k\right)$$

Latent (unobserved) variables $(Z_k)_{k\geq 1}$ form a Markov chain

Finite state space stationary HMMs

The Markov chain is stationary, has finite state space $\{1, \ldots, K\}$ and transition matrix Q. The stationary distribution is denoted μ.

Conditionnally to $Z_k = j$, X_k has emission distribution F_j.

Finite state space stationary HMMs

The Markov chain is stationary, has finite state space $\{1, \ldots, K\}$ and transition matrix Q. The stationary distribution is denoted μ.

Conditionnally to $Z_k = j$, X_k has emission distribution F_j.

The marginal distribution of any X_k is

$$\sum_{j=1}^{K} \mu(j) F_j$$

A finite state space HMM is a finite mixture with Markov regime

The use of hidden Markov models

Modeling dependent data arising from heterogeneous populations.

The use of hidden Markov models

Modeling dependent data arising from heterogeneous populations.

Markov regime : leads to efficient algorithms to compute :

- Filtering/prediction/smoothing/ probabilities (Forward/Backward recursions) : given a set of observations, the probability of hidden states.
- Maximum a posteriori (prediction of hidden states) ; Viterbi's algorithm.
- Likelihoods and EM algorithms : estimation of the transition matrix Q and the emission distributions $F_1, \ldots, F_K$
- MCMC Bayesian methods

The parametric/non parametric story

The inference theory is well developed in the parametric situation where for all j, $F_j \in \{F_\theta, \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$.
But parametric modeling of emission distributions may lead to poor results in particular applications.

Motivating example : DNA copy number variation using DNA hybridization intensity along the genome

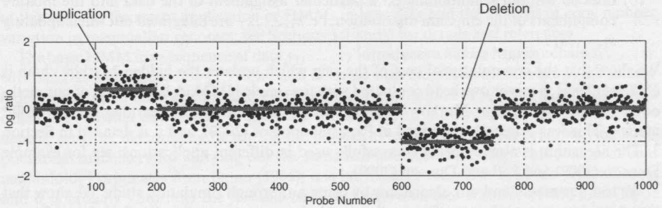


Fig. 1. Example array CGH data set: this data sets shows a copy number gain (duplication) and a copy number loss (deletion) which are characterized by relative upward and downward shifts in the log-intensity-ratio respectively; the probe number here indicates the chromosomal location

Popular approach : HMM with emission distributions $\mathcal{N}(m_j;\sigma^2)$ for state j.

Sensitivity to outliers, skewness or heavy tails that may lead to large numbers of false copy number variants detected.

$\rightarrow$ Non parametric Bayesian algorithms : Yau, Papaspiliopoulos, Roberts, Holmes JRSSB 2011)

Other examples in which the use of nonparametric algorithms improves performances

- Bayesian methods
  - Climate state identification (Lambert et al. 2003)
- EM-style algorithms
  - Voice activity detection (Couvreur et al., 2000)
  - Facial expression recognition (Shang et al. 2009)

Finite state space non parametric HMMs

The marginal distribution of any X_k is $\sum_{j=1}^{K} \mu(j) F_j$

Non parametric mixtures are not identifiable with no further assumptions

$$
\begin{aligned}
&\mu(1) F_1 + \mu(2) F_2 + \ldots + \mu(K) F_K \\
&= (\mu(1)+\mu(2)) \left[\frac{\mu(1)}{\mu(1)+\mu(2)} F_1 + \frac{\mu(2)}{\mu(1)+\mu(2)} F_2 \right] + \ldots + \mu(K) F_K \\
&\qquad = \frac{\mu(1)}{2} F_1 + \frac{\left[\frac{\mu(1)}{2} F_1 + \mu(2) F_2 \right]}{\frac{\mu(1)}{2} + \mu(2)} + \ldots + \mu(K) F_K
\end{aligned}
$$

Why do non parametric HMM algorithms work ? ? ? ?

Dependence of observed variables has to help !

Basic questions

Denote $\mathbb{F} = (F_1, \ldots, F_K)$.
For m an integer, let $\mathbb{P}^{(m)}_{K;Q;\mathbb{F}}$ be the distribution of $(X_1, \ldots, X_m)$.

The sequence of observed variables has mixing properties : adaptive estimation of $\mathbb{P}^{(m)}_{K;Q;\mathbb{F}}$ is possible. Can one get information on K, Q and $\mathbb{F}$ from an estimator $\widehat{\mathbb{P}^{(m)}}$ of $\mathbb{P}^{(m)}_{K;Q;\mathbb{F}}$?

- Identifiability : for some m,

$$\mathbb{P}^{(m)}_{K_1;Q_1;\mathbb{F}_1} = \mathbb{P}^{(m)}_{K_2;Q_2;\mathbb{F}_2} \Longrightarrow K_1 = K_2,\ Q_1 = Q_2,\ \mathbb{F}_1 = \mathbb{F}_2.$$

- Inverse problem : Build estimators $\widehat{K}$, $\widehat{Q}$ and $\widehat{\mathbb{F}}$ such that one may deduce consistency/rates from those of $\widehat{\mathbb{P}^{(m)}}$ as an estimator of $\mathbb{P}^{(m)}_{K;Q;\mathbb{F}}$.

Joint work with Judith Rousseau *(translated emission distributions ; Bernoulli 2016)*

Joint work with Alice Cleynen and Stéphane Robin *(General identifiability ; Stat. and Comp. 2016)*,
Yohann De Castro and Claire Lacour *(Adaptive estimation via model selection and least squares ; JMLR 2016)*,
Yohann De Castro and Sylvain Le Corff *(Spectral estimation and estimation of filtering/smoothing probabilities ; IEEE IT to appear)*,

Work by Elodie Vernet *(Bayesian estimation ; consistency EJS 2015 and rates Bernoulli in revision)*

Work by Luc Lehéricy *(Estimation of K ; submitted ; state by state adaptivity ; submitted)*

Work by Augustin Touron *(Climate applications ; PHD in progress)*

Identifiability/inference theoretical results in nonparametric HMMs

1. Identifiability in non parametric finite translation HMMs and extensions
2. Identifiability in non parametric general HMMs
3. Generic methods
4. Inverse problem inequalities
5. Further works

Identifiability/inference theoretical results in nonparametric HMMs

Translated emission distributions

Here we assume that there exists a distribution function F and real numbers $m_1, \ldots, m_K$ such that

$$F_j(\cdot) = F(\cdot - m_j),\ j = 1, \ldots, K.$$

The observations follow

$$X_t = m_{Z_t} + \epsilon_t,\ \ t \geq 1,$$

where the variables ϵ_t, $t \geq 1$, are i.i.d. with distribution function F, and are independent of the Markov chain $(Z_t)_{t \geq 1}$.

Previous work : independent variables ; $K \leq 3$; symmetry assumption on F : Bordes, Mottelet, Vandekerkhove (Annals of Stat. 2006) ; Hunter, Wang, Hettmansperger (Annals of Stat. 2007) ; Butucea, Vandekerkhove (Scandinavian J. of Stat, to appear).

Identifiability : assumptions

For $K \geq 2$, let Θ_k be the set of $\theta = \big(m, (\mathcal{Q}_{i,j})_{1\leq i,j\leq K,(i,j)\neq(K,K)}\big)$ satisfying :

- $\mathcal{Q}$ is a probability mass function on $\{1,\ldots,K\}^2$ such that $\det(\mathcal{Q}) \neq 0$,
- $m \in \mathbb{R}^K$ is such that $m_1 = 0 < m_2 < \ldots < m_k$.

For any distribution function F on $\mathbb{R}$, denote $\mathbb{P}^{(2)}_{(\theta,F)}$ the law of (X_1, X_2) :

$$\mathbb{P}^{(2)}_{(\theta,F)}(A \times B) = \sum_{i,j=1}^{K} \mathcal{Q}_{i,j} F(A - m_i) F(B - m_i).$$

Identifiability result

Theorem [EG, J. Rousseau (Bernoulli 2016)]

Let F and $\tilde{F}$ be distribution function on $\mathbb{R}$, $\theta \in \Theta_K$ and $\tilde{\theta}$ in $\Theta_{\tilde{K}}$. Then

$$\mathbb{P}^{(2)}_{\theta,F} = \mathbb{P}^{(2)}_{\tilde{\theta},\tilde{F}} \Longrightarrow K = \tilde{K},\ \theta = \tilde{\theta} \text{ and } \mathrm{F} = \tilde{\mathrm{F}}.$$

- No assumption on F !
- HMM not needed ; dependent (stationary) state variables suffice.
- Extension (by projections) to multidimensional variables.
- Identification of ℓ-marginal distribution, i.e. the law of $(Z_1, \ldots, Z_\ell)$, K and F using the law of $(X_1, \ldots, X_\ell)$.

Identifiability : sketch of proof

ϕ_F : characteristic function of F ; $\phi_{\tilde{F}}$: c.f. of $\tilde{F}$;
$\phi_{\theta,i}$: ($\phi_{\tilde{\theta},i}$) c.f. of the law of m_{Z_i} under $P_{\theta,F}$, (under $P_{\tilde{\theta},\tilde{F}}$) ;
Φ_θ : ($\Phi_{\tilde{\theta}}$) c.f. of the law of (m_{Z_1}, m_{Z_2}) under $P_{\theta,F}$ (under $P_{\tilde{\theta},\tilde{F}}$).

The c.f. of the law of X_1, of X_2, then of (X_1, X_2), give

$$\phi_F(t)\,\phi_{\theta,1}(t) = \phi_{\tilde{F}}(t)\,\phi_{\tilde{\theta},1}(t)\,,$$
$$\phi_F(t)\,\phi_{\theta,2}(t) = \phi_{\tilde{F}}(t)\,\phi_{\tilde{\theta},2}(t)\,,$$

$$\phi_F(t_1)\,\phi_F(t_2)\,\Phi_\theta(t_1,t_2) = \phi_{\tilde{F}}(t_1)\,\phi_{\tilde{F}}(t_2)\,\Phi_{\tilde{\theta}}(t_1,t_2)\,.$$

We thus get for all $(t_1, t_2) \in \mathbb{R}^2$,

$$\phi_F(t_1)\,\phi_F(t_2)\,\Phi_\theta(t_1,t_2)\,\phi_{\tilde{\theta},1}(t_1)\,\phi_{\tilde{\theta},2}(t_2) = \phi_F(t_1)\,\phi_F(t_2)\,\Phi_{\tilde{\theta}}(t_1,t_2)\,\phi_{\theta,1}(t_1)\,\phi_{\theta,2}(t_2)\,.$$

Identifiability : sketch of proof

Thus on a neighborhood of 0 in which ϕ_F is non zero :

$$\Phi_\theta\left(t_1, t_2\right) \phi_{\tilde{\theta},1}\left(t_1\right) \phi_{\tilde{\theta},2}\left(t_2\right) = \Phi_{\tilde{\theta}}\left(t_1, t_2\right) \phi_{\theta,1}\left(t_1\right) \phi_{\theta,2}\left(t_2\right).$$

Then

- Equation extended to the complex plane (entire functions).
- The set of zeros of $\phi_{\theta,1}$ coincides with the set of zeros of $\phi_{\tilde{\theta},1}$ (here $\det(\mathrm{Q}) \neq 0$ is used).
- Hadamard's factorization theorem allows to prove that $\phi_{\theta,1} = \phi_{\tilde{\theta},1}$.
- Same proof for $\phi_{\theta,2} = \phi_{\tilde{\theta},2}$, leading to $\Phi_\theta = \Phi_{\tilde{\theta}}$, and then $\phi_F = \phi_{\tilde{F}}$

Finally the characteristic function characterizes the law, so that $K = \tilde{K}$, $\theta = \tilde{\theta}$ and $F = \tilde{F}$.

Identifiability : estimation of θ

$$\Phi_{\theta}(t_1,t_2)\,\phi_{X_1}(t_1)\,\phi_{X_2}(t_2)-\Phi_{(X_1,X_2)}(t_1,t_2)\,\phi_{\theta,1}(t_1)\,\phi_{\theta,2}(t_2)=0.$$

- Replace $\phi_{X_1}(t_1)$, $\phi_{X_2}(t_2)$ and $\Phi_{(X_1,X_2)}(t_1,t_2)$ by estimators (ex : empirical estimators) to get an empirical contrast (take the square of the modulus and integrate).
- Preliminar estimator : penalize to get consistent estimators of K and θ satisfying the assumptions.
- $\widehat{\theta}_n$ minimize the contrast over a suitable compact.

$\widehat{\theta}_n$ is $\sqrt{n}$-consistent + asymptotic distr. + deviation inequalities [G. , Rousseau (Bernoulli 2016)]

Identifiability/inference theoretical results in nonparametric HMMs

1. Identifiability in non parametric finite translation HMMs and extensions
2. Identifiability in non parametric general HMMs
3. Generic methods
4. Inverse problem inequalities
5. Further works

Finite state space HMM : Connexion with mixtures of independent variables

The distribution of (X_1, X_2, X_3) may be written as

$$
\begin{aligned}
\mathbb{P}^{(3)}_{Q,\mathbb{F}} &= \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{m=1}^{K} \mu(i)\, Q_{i,j} Q_{j,m} F_i \otimes F_j \otimes F_m \\
&= \sum_{j=1}^{K} \mu(j) \left(\sum_{i=1}^{K} \frac{\mu(i)\, Q_{i,j}}{\mu(j)} F_i \right) \otimes F_j \otimes \left(\sum_{m=1}^{K} Q_{j,m} F_m \right) \\
&= \sum_{j=1}^{K} \mu(j)\, G_{j,1} \otimes G_{j,2} \otimes G_{j,3}
\end{aligned}
$$

which is a mixture of K populations, in each population the observation is that of independent variables.

Z_1 and Z_3 are independent conditionally to Z_2.

→ Use results about mixtures of independent variables.

An old result by Kruskal

Kruskal's algebraic result (1977) : 3-way contingency tables are identifiable (up to label switching) under some Kruskal's rank assumption.

Kruskal + adequate approximation argument : Non parametric mixtures in which, conditionally to the population, at least 3 variables are independent, are identifiable under some linear independence assumption of the conditional probability distributions of those variables. (Allman et al. , 2009)

Theorem (A. Cleynen, S. Robin, EG, 2016 Stat. and Comput.)

Assume that the probability measures $F_1, \ldots, F_K$ are linearly independent and that Q has full rank. Then the parameters K, Q and $F_1, \ldots, F_K$ are identifiable from the distribution of 3 consecutive observations X_1, X_2, X_3, up to label swapping of the hidden states.

Mixtures of independent variables : spectral analysis

Works by Anandkumar, Dai, Hsu, Kakade, Song, Zhang, Xie.

Let $X = (X_1; X_2; X_3)$ have distribution $\otimes_{d=1}^{3} G_{j,d}$ conditionally to $Z = j$ so that X has distribution

$$\sum_{j=1}^{K} \mu(j) \otimes_{d=1}^{3} G_{j,d}$$

Let $\varphi_1, \ldots, \varphi_M$ be M real valued functions.
For $d = 1, 2, 3$, define $A^{(d)}$ the $M \times K$ matrix such that

$$A_{l,j}^{(d)} = \int \varphi_l dG_{j,d} = E\left[\varphi_l(X_d) | Z = j\right]$$

$$A^{(d)} = \begin{pmatrix} \int \varphi_1 dG_{1,d} & \cdots & \int \varphi_1 dG_{K,d} \\ \vdots & \vdots & \vdots \\ \int \varphi_M dG_{1,d} & \cdots & \int \varphi_M dG_{K,d} \end{pmatrix}$$

Mixtures of independent variables : spectral analysis

Let $D = Diag(\mu(1), \cdots, \mu(K))$.

Let S the $M \times M$ matrix such that $S_{l,m} = E[\varphi_l(X_1)\varphi_m(X_2)]$.
Then,

$$S = A^{(1)} D (A^{(2)})^T.$$

If for all $d = 1, 2, 3$, $G_{1,d}, \ldots, G_{K,d}$ are linearly independent, then for large enough M, $rank(A^{(d)}) = K$ and

$$rank\,(S) = K.$$

Let U_1 and U_2 be $M \times K$ matrices such that $U_1^T S U_2$ is invertible (may be found by SVD of S).

$$U_1^T S U_2 = \left(U_1^T A^{(1)}\right) D \left((A^{(2)})^T U_2\right).$$

Mixtures of independent variables : spectral analysis

Define T be the $M \times M \times M$ tensor such that

$$T(l_1, l_2, l_3) = E[\varphi_{l_1}(X_1)\varphi_{l_2}(X_2)\phi_{l_3}(X_3)].$$

Let $V \in \mathbb{R}^M$, and define $T[V]$ the $M \times M$ matrix such that

$$T[V]_{l,m} = E[\varphi_l(X_1)\varphi_m(X_2)\langle V, \Phi(X_3)\rangle]$$

where $\Phi(X_3) = (\varphi_h(X_3))_{1 \le h \le M}$. Then

$$T[V] = A^{(1)} D \cdot Diag\left((A^{(3)})^T V\right)(A^{(2)})^T$$

Define

$$B(V) = (U_1^T T[V] U_2)(U_1^T S U_2)^{-1}.$$

Then, one has

$$B(V) = (U_1^T A^{(1)}) Diag\left((A^{(3)})^T V\right)(U_1^T A^{(1)})^{-1}.$$

Mixtures of independent variables : spectral analysis

$$U_1^T S U_2 = \left(U_1^T A^{(1)}\right) D \left((A^{(2)})^T U_2\right)$$

$$\left(U_1^T S U_2\right)^{-1} = \left((A^{(2)})^T U_2\right)^{-1} D^{-1} \left(U_1^T A^{(1)}\right)^{-1}$$

$$T[V] = A^{(1)} D \cdot Diag\left((A^{(3)})^T V\right) (A^{(2)})^T$$

$$\begin{aligned}
B(V) &= (U_1^T T[V] U_2)(U_1^T S U_2)^{-1} \\
&= U_1^T A^{(1)} D \cdot Diag\left((A^{(3)})^T V\right) (A^{(2)})^T U_2 (U_1^T S U_2)^{-1} \\
&= U_1^T A^{(1)} Diag\left((A^{(3)})^T V\right) \cdot D (A^{(2)})^T U_2 (U_1^T S U_2)^{-1} \\
&= (U_1^T A^{(1)}) Diag\left((A^{(3)})^T V\right) (U_1^T A^{(1)})^{-1}.
\end{aligned}$$

Mixtures of independent variables : spectral analysis

Recall

$$B(V) = (U_1^T T[V] U_2)(U_1^T S U_2)^{-1} = (U_1^T A^{(1)}) Diag\left((A^{(3)})^T V\right)(U_1^T A^{(1)})$$

All matrices $B(V)$ have the same eigenvectors, and eigenvalues the coordinates of $(A^{(3)})^T V$.

By exploring various vectors V, one may recover $A^{(3)}$. The eigenvectors stay the same when permuting coordinates 2 and 3 of the observed variable, so that one may recover $A^{(2)}$, and thus also $A^{(1)}$. Recovering D is then also possible. Then, by taking M to infinity, one may recover the whole distributions $G_{1,j}$, $G_{2,j}$ and $G_{3,j}$, $j = 1, \ldots, K$.

One may recover $\mu(1), \ldots, \mu(K)$ and $G_{1,j}$, $G_{2,j}$ and $G_{3,j}$, $j = 1, \ldots, K$ using Singular Value/ Eigen Value decompositions of matrices built from the distribution of $X = (X_1, X_2, X_3)$.

Spectral analysis : estimation

Emission distributions with densities $f_j^\star$, $j = 1, \ldots, K$ in $\mathbf{L}^2(\mathcal{X})$.

- Use a sieve of finite dimensional subspaces with orthonormal basis $\Phi_M := \{\varphi_1, \ldots, \varphi_M\}$.
  Examples : histograms ; splines ; Fourier ; wavelets.
- Estimation of $Q^\star$ and $\langle f_j^\star, \varphi_m \rangle$, $j = 1, \ldots, K$, $m = 1, \ldots, M$ on the basis of the empirical distribution of the three-dimensional marginal, i.e. the distribution of (X_1, X_2, X_3)
  Uses only one SVD, matrix inversions and one diagonalization.

$$\|\widehat{Q} - Q^\star\|^2 \text{ and } \|\widehat{f}_{M,j} - f_{M,j}^\star\|^2 \text{ are } O_P\left(\frac{M^3}{n}\right)$$

(De Castro, G., Le Corff, IEEE IT to appear)

Identifiability/inference theoretical results in nonparametric HMMs

Model selection via penalized contrast

Define a contrast function $\gamma_n(g)$, g a possible density such that $\gamma_n(g) - \gamma_n(g^\star)$ has positive limit for $g \neq g^\star$, $g^\star$ being the true density.

The possible densities g have a particular form depending on the emission densities and a parametric part : $g := g_{\theta,F}$.

A sieve for the emission distributions leads to sieves on the possible densities $\mathcal{S}(\theta, M)$.

For the parametric part, we have in hand an estimator $\widehat{\theta}$ that converges at parametric (or nearly parametric) rate.

For each M, define $\widehat{g}_M$ as the minimizer of $\gamma_n(g)$ for $g \in \mathcal{S}(\widehat{\theta}, M)$.

Set a penalty function $pen(n, M)$ and choose

$$\widehat{M} = \arg\min_{M=1,\ldots,n} \left\{\gamma_n(\widehat{g}_M) + pen(n, M)\right\}.$$

Then the estimator of $g^\star$ is $\widehat{g} = \widehat{g}_{\widehat{M}}$, and the estimator of $F^\star$ is $\hat{F}$ such that

$$\widehat{g} = g_{\widehat{\theta},\widehat{F}}.$$

Model selection via penalized contrast

Translation mixtures with dependent regime

Recall that the observations follow :

$$X_t = m_{Z_t} + \epsilon_t,\ t \geq 1,$$

where the variables ϵ_t, $t \geq 1$, are i.i.d. with distribution function F, and are independent of the Markov chain $(Z_t)_{t\geq 1}$.
When $\theta = ((m_j)_j, (Q_{i,j})_{i,j})$ is known, one may recover F from the marginal density $g_{\theta,F}$ of X_t.
If F has density f, then $g_{\theta,f} := g_{\theta,F}$ is given by :

$$g_{\theta,f}(x) = \sum_{j=1}^{K} \mu(j)\, f(x - m_j)\,.$$

where $\mu(i) = \sum_{j=1}^{K} Q_{i,j}$. Given the estimator $\widehat{\theta}_n = ((\widehat{m}_i)_{1\leq i\leq k^\star}, (\widehat{Q}_{i,j})_{(i,j)\neq(k^\star,k^\star)})$, denote $\widehat{\mu}(i) = \sum_{j=1}^{k^\star} \widehat{Q}_{i,j}$.

Model selection via penalized contrast

Maximum marginal-likelihood :

$$\gamma_n(g) = -\frac{1}{n}\sum_{i=1}^{n} \log g(X_i).$$

The sieve $\mathcal{S}(\widehat{\theta}, M)$ is the set of functions $g = \sum_{j=1}^{K} \widehat{\mu}(j) f(x - \widehat{m}_j)$ where $f \in \mathcal{F}_M$:

$$\mathcal{F}_M = \left\{ \sum_{i=1}^{M} \pi_i \varphi_{\beta_i}(x - \alpha_i),\ \alpha_i \in [-A_M, A_M],\ \beta_i \in [b_M, B], \right.$$
$$\left. \pi_i \geq 0,\ i = 1, \ldots, p,\ \sum_{i=1}^{p} \pi_i = 1 \right\}$$

with φ_β the centered gaussian density with variance β^2.

Model selection via penalized contrast

General finite state space HMMs

Here $\theta = Q$ the transition matrix of the hidden Markov chain. For $F = (f_1, \ldots, f_K)$ emission densities, if π is the stationary distribution of Q, the density of (X_1, X_2, X_3) is given by

$$g_{\theta,F}(x_1, x_2, x_3) = \sum_{j_1,j_2,j_3=1}^{K} \pi(j_1) Q(j_1, j_2) Q(j_2, j_3) f_{j_1}(x_1) f_{j_2}(x_2) f_{j_3}(x_3).$$

Least squares :

$$\gamma_n(g) = \|g\|_2^2 - \frac{2}{n} \sum_{s=1}^{n-2} g(X_s, X_{s+1}, X_{s+2}).$$

As n tends to infinity, $\gamma_n(g) - \gamma_n(g^\star)$ converges almost surely to $\|g - g^\star\|_2^2$.

The sieve $\mathcal{S}(\widehat{\theta}, M)$ is the set of functions $g_{\widehat{\theta},F}$ such that

$$\forall j = 1, \ldots, K, \ \exists (a_{mj})_{1 \leq m \leq M} \in \mathbb{R}^M, \ f_j = \sum_{m=1}^{M} a_{mj} \varphi_m.$$

Oracle inequalities (in general)

There exist constants κ, C and n_0 such that : if

$$pen(n,M) \geq \kappa \text{ complexity}(M) \frac{\log n}{n},$$

then for all $x > 0$, for all $n \geq n_0$, with probability $1 - e^{-x}$, it holds

$$D^2(\widehat{g}, g^\star) \leq C \left\{ \inf_M \left[d^2(g^\star_M, g^\star) + pen(n,M) \right] + \text{small terms} \right\}.$$

- Proof : concentration inequality + control of the complexity of the Sieve (ex : using bracketing entropy).
- Adaptive rates ; automatic best compromise bias/variance.
- Penalty in practice : slope heuristics.

Oracle inequalities : Translation mixtures and HMMs

Additional difficulty : deal with $\widehat{\theta}$ in γ_n.
C depends here on the hidden chain (concentration inequality for dependent variables).

Translation mixtures with dependent regime
Oracle inequality using penalized m.l.e (G. , Rousseau [Bernoulli 2016]).
$D^2(\widehat{g}, g^\star)$: Hellinger's distance.
$d^2(g^\star_M, g^\star)$: Kullback's divergence.

General finite state space HMMs
Oracle inequality using least squares (De Castro, G. Lacour [JMLR 2016]).
$D^2(\widehat{g}, g^\star)$ and $d^2(g^\star_M, g^\star)$: L_2-square distance.

Identifiability/inference theoretical results in nonparametric HMMs

General question

Consistent estimation of $g^\star$ translates to consistent estimation of $F^\star$.

Do adaptive minimax rates for the estimation of $g^\star$ translate to adaptive minimax rates for the estimation of $F^\star$?

Inverse problem : translation mixtures

Recall $g^\star = \sum_{j=1}^K \mu^\star(j) f^\star\left(x - m_j^\star\right)$.

G., Rousseau, Bernoulli 2016

If $f^\star$ has bounded derivative,

$$\left(2\max_j \widehat{\mu}(j) - 1\right)\left\|\widehat{f} - f^\star\right\|_1 \leq 2h(g^\star, \widehat{g}) + (1 + \|(f^\star)'\|_\infty)\|\widehat{\theta}_n - \theta^\star\|.$$

Consequence : if $\max_j \mu^\star(j) > \frac{1}{2}$, results on $h^2(g^\star, \widehat{g})$ and $\|\widehat{\theta}_n - \theta^\star\|$ translate to results on $\left\|\widehat{f} - f^\star\right\|_1$.

Remark : $\phi_{g^\star} = \phi_{f^\star}\phi_{\theta^\star}$ with $\phi_{\theta^\star}(t) = \sum_{j=1}^K \mu^\star(j) e^{im_j^\star t}$, and $\phi_{\theta^\star}(t) \neq 0$ for all t if and only if $\max_j \mu^\star(j) > \frac{1}{2}$ (Moreno 1973).

Proof

Proof : starts from $\|g^\star - \widehat{g}\|_1^2 \leq 4h^2(g^\star, \widehat{g})$. Then,

$$\begin{aligned}
\|g^\star - \widehat{g}\|_1 &= \|\sum_{j=1}^{K} \mu^\star(j) f^\star(y - m_j^\star) - \sum_{j=1}^{K} \widehat{\mu}(j) \widehat{f}(\cdot - \widehat{m}_j)\|_1 \\
&\geq \|\sum_{j=1}^{K} \widehat{\mu}(j)(\widehat{f} - f^\star)(\cdot - \widehat{m}_j)\|_1 \\
&\quad - \|\sum_{j=1}^{K} \mu^\star(j) f^\star(y - m_j^\star) - \sum_{j=1}^{K} \widehat{\mu}(j) f^\star(\cdot - \widehat{m}_j)\|_1 \\
&\geq \|\sum_{j=1}^{K} \widehat{\mu}(j)(\widehat{f} - f^\star)(\cdot - \widehat{m}_j)\|_1 - (1 + \|(f^\star)'\|_\infty)\|\widehat{\theta}_n - \theta^\star\|
\end{aligned}$$

Then using the triangle inequality,

$$\|\sum_{j=1}^{K} \widehat{\mu}(j)(\widehat{f} - f^\star)(\cdot - \widehat{m}_j)\|_1 \geq \left(2 \max_j \widehat{\mu}(j) - 1\right) \left\|\widehat{f} - f^\star\right\|_1 .$$

Inverse problem : non parametric HMMs

Recall that for $F = (f_1, \ldots, f_K)$ emission densities and Q a transition matrix with stationary distribution π,

$$g_{Q,F}(x_1, x_2, x_3) = \sum_{j_1, j_2, j_3 = 1}^{K} \pi(j_1) Q(j_1, j_2) Q(j_2, j_3) f_{j_1}(x_1) f_{j_2}(x_2) f_{j_3}(x_3).$$

Assumption : $P(Q^\star, \langle f_j^\star, f_l^\star \rangle) \neq 0$ $\qquad$ P polynomial

$\rightarrow$ generically satisfied

$\rightarrow$ always satisfied if $K = 2$

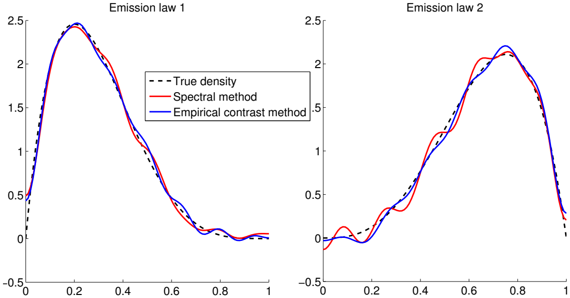
Theorem (Y. de Castro, EG, C. Lacour, JMLR 2016)

There exists $C > 0$ such that for all Q in a neighborhood of $Q^\star$,

$$\|g_{Q,F^\star} - g_{Q,F}\|_2 \geq C \sum_{j=1}^{K} \|f_j^\star - f_j\|_2.$$

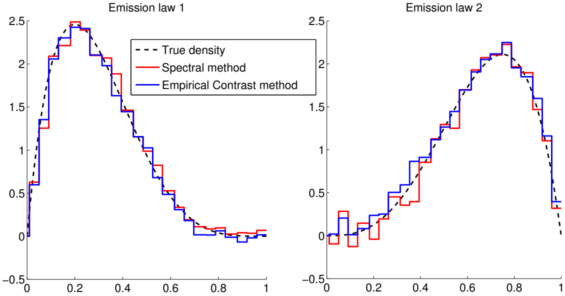
Thus, results on $\|g^\star - \widehat{g}\|_2$ translate to results on $\sum_{j=1}^{K} \|f_j^\star - \widehat{f}_j\|_2$.

Simulations : K=2



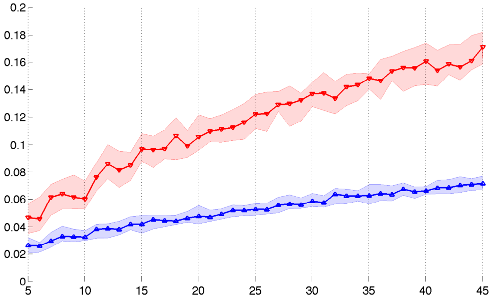
Reconstruction of densities f_1 and f_2 (Beta distributions) with spectral and least squares methods ($N = 50000$, trigonometric basis)

Simulations : K=2



Reconstruction of densities f_1 and f_2 (Beta distributions) with spectral and least squares methods ($N = 50000$, histogram basis)

Simulations : K=2

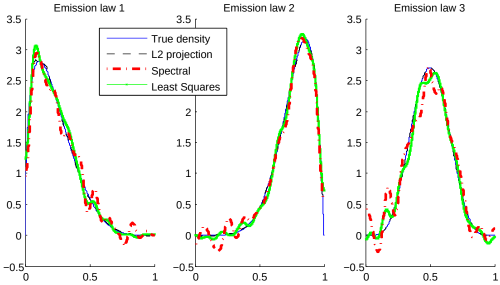


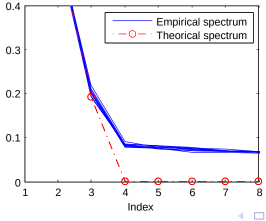
Integrated variance $\sum_{j=1}^{2} E\|\widehat{f_j} - f_{M,j}\|^2$ of spectral and least squares estimators, as a function of M ($N = 50000$, histogram basis)

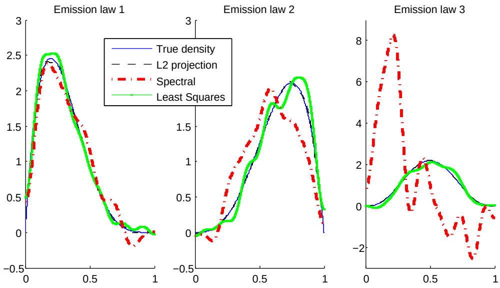
Identifiability/inference theoretical results in nonparametric HMMs

Sensitivity to the linear dependence assumption

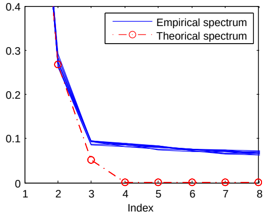
(L. Lehéricy, mémoire de M2, 2015).







Emission law 1
Emission law 2
Emission law 3
True density
L2 projection
Spectral
Least Squares
3
2.5
2
1.5
1
0.5
0
−0.5
0
0.5
1
8
6
4
2
−2



Empirical spectrum
Theorical spectrum
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
Index

Likelihood methods

Back to Kruskal : identifiability holds when Q is full rank and $F_1, \dots, F_K$ are distinct probability distributions, but on the basis of the $(2K+1)[(K^2-2K+2)+1]$-th marginal distribution. (Alexandrovitch et al., 2016)

$\rightarrow$ Full likelihood methods

(Oracle inequalities, L. Lehéricy, on going work)

Others

- Bayesian methods E. Vernet : consistency of the posterior distribution (EJS 2015) ; rates of concentration for the posterior distribution (Bernoulli, in revision).
- Clustering/Estimation of the filtering and marginal smoothing distibutions (Y. De Castro, EG, S. Le Corff, IEEE IT, to appear)
- Estimation of K (L. Lehéricy, 2016, submitted)
- Adaptive estimation of each emission density using Lepski's method (L. Lehéricy, on going work)
- Seasonal HMMs and climate applications (A. Touron, work in progress)

Thank you for your attention !