Minimax adaptive estimation of nonparametric hidden Markov models

Yohann De Castro, Elizabeth Gassiat, Claire Lacour

LMO, Université Paris-Sud

IHES, january 2016

Outline

Model

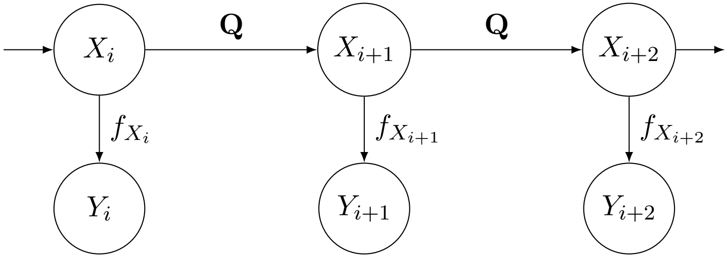
- (X_i) Markov chain on $\{1, \dots, K\}$: non-observed — K known
  transition $\mathbf{Q}$ (matrix $K \times K$) — $\mathbf{Q}$ unknown

- $Y_1, \dots, Y_n$ in $\mathbb{R}$: observations
  - Y_i are independent given $(X_i)_{i \geq 1}$
  - the distribution of Y_i only depends on X_i

  Conditional distribution of $Y_i | X_i = k$:

  $$\mathbf{f}_k(y)dy = \mathbb{P}(Y_i \in dy | X_i = k)$$ — $\mathbf{f}$ unknown

Hidden Markov Model



Observations : $Y_1, \dots, Y_n$
Known parameter : K
To estimate : transition matrix $\mathbf{Q}$, initial distribution $\boldsymbol{\pi}$,
emission functions $\mathbf{f}_1, \dots, \mathbf{f}_K$

State of the art

Until very recently, theoretical results only in the *parametric* setting

Nonparametric: Gassiat, Rousseau (2015) Dumont, Lecorff (2016)

Identifiability:
Allman, Matias, Rhodes (2009)
Hsu, Kakade, Zhang (2012)
Gassiat, Cleynen, Robin (2015)
Alexandrovich, Holzmann (2014)

Assumptions

(H_1) $\mathbf{Q}$ has full rank

(H_2) (X_i) irreducible aperiodic

(H_3) stationary Markov chain

(H_4) $\mathbf{f}_1, \ldots, \mathbf{f}_K$ linearly independent

Identifiability

Distribution of (Y_1, Y_2, Y_3):

$$g^{\mathbf{Q},\mathbf{f}}(y) := \sum_{k_1,k_2,k_3=1}^{K} \boldsymbol{\pi}(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3)\mathbf{f}_{k_1}(y_1)\mathbf{f}_{k_2}(y_2)\mathbf{f}_{k_3}(y_3)$$

Lemma

Under (H_1)–(H_4), there is identifiability, up to label switching, from three consecutive observations:
$g^{\mathbf{Q},\mathbf{f}+h} = g^{\mathbf{Q},\mathbf{f}} \Leftrightarrow \exists\tau$ *permutation such that* $h_j = \mathbf{f}_j - \mathbf{f}_{\tau(j)}$

Projection on an approximation space

Approximation of the $\mathbf{f}_k$: for $(\varphi_1, \ldots, \varphi_m)$ orthonormal basis

$$\mathbf{f}_{k,m} = \sum_{i=1}^{m} \langle \mathbf{f}_k, \varphi_i \rangle \varphi_i$$

Examples : Fourier basis, Piecewise polynomials, Wavelets

Aim

To estimate matrices $\mathbf{Q} = (\mathbb{P}(X_2 = j | X_1 = k]))_{kj}$
and $\mathbf{F} = (\langle \mathbf{f}_k, \varphi_i \rangle)_{ik} = (\mathbb{E}[\varphi_i(Y_1) | X_1 = k])_{ik}$

choice of m ? $\rightarrow$ m fixed for now

Matrix expressions

$$P_1(a) := \mathbb{E}[\varphi_a(Y_1)] \qquad 1 \leq a \leq m$$
$$P_{12}(a,b) := \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)] \qquad 1 \leq a,b \leq m$$

Lemma

- $P_1 = \mathbf{F}\,\boldsymbol{\pi}$

  $P_1 \downarrow Y_1$, $\mathbf{F} \swarrow Y_1|X_1$, $\boldsymbol{\pi} \searrow X_1$

- $P_{12} = \mathbf{F}\mathrm{Diag}(\boldsymbol{\pi})\mathbf{Q}\,\mathbf{F}^T$

  $P_{12} \downarrow (Y_1, Y_2)$, $\mathbf{F} \swarrow Y_1|X_1$, $\mathrm{Diag}(\boldsymbol{\pi}) \downarrow X_1$, $\mathbf{Q} \downarrow X_2|X_1$, $\mathbf{F}^T \searrow Y_2|X_2$

Csq : Knowing $\mathbf{F}, P_1, P_{12}$ allows to recover $\boldsymbol{\pi}$ and $\mathbf{Q}$

A crucial Lemma

$$
\begin{aligned}
P_{123}(a,b,c) &:= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)] \\
P_{12}(a,b) &:= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)] \\
P_{13}(a,c) &:= \mathbb{E}[\varphi_a(Y_1)\varphi_c(Y_3)]
\end{aligned}
$$

easily estimable

Lemma

Let U be the $m \times K$ matrix of right singular vectors of P_{13}. Then $U^T P_{13} U$ is invertible and if

$$B(j) := (U^T P_{13} U)^{-1} U^T P_{123}(.,j,.) U$$

then there exists R not depending on j such that

$$B(j) = R \,\mathrm{Diag}(\mathbf{F}(j,.))\, R^{-1}$$

Consequence

$$B(j) = (U^T P_{13} U)^{-1} U^T P_{123}(., j, .) U = R \, \mathrm{Diag}(\mathbf{F}(j, .)) \, R^{-1}$$

$\Rightarrow$ Diagonalizing $B(j), j = 1 \ldots, m$ allows to recover $\mathbf{F}$

Remark: Instead of diagonalizing B, random mixtures of the $B(j)$ in order to separate the eigenvalues:

$$C(k) = \sum_{j=1}^{m} (U\Theta)(j, k) B(j)$$

with Θ random unitary matrix

Algorithm (inspired from Anandkumar, Hsu, Kakade (2012))

- Estimate $P_1, P_{12}, P_{13}, P_{123}$ by their empirical equivalent e.g. $\hat{P}_{13}(a,c) := \frac{1}{n}\sum_{i=1}^{n-2} \varphi_a(Y_i)\varphi_c(Y_{i+2})$
- $\hat{U}$ matrix $m \times K$ of right singular vectors of $\hat{P}_{13}$ corresponding to the K largest singular values
- $\hat{B}(j) := (\hat{U}^T \hat{P}_{13} \hat{U})^{-1} \hat{U}^T \hat{P}_{123}(., j, .)\hat{U}$
- Diagonalize $\hat{B}$: eigenvalues provide $\hat{\mathbf{F}}(j,k)$
- $\tilde{\boldsymbol{\pi}} = (\hat{U}^T \hat{\mathbf{F}})^{-1} \hat{U}^T \hat{P}_1$ and $\tilde{\mathbf{Q}} = (\hat{U}^T \hat{\mathbf{F}} \mathrm{Diag}(\tilde{\boldsymbol{\pi}}))^{-1} \hat{U}^T \hat{P}_{12} \hat{U} (\hat{\mathbf{F}}^T \hat{U})^{-1}$
- $\hat{\mathbf{Q}}$ projection of $\tilde{\mathbf{Q}}$ on the space of transition matrices, and $\hat{\boldsymbol{\pi}}$ its stationnary distribution

Performance of the spectral method

Theorem

Under $(H1)$–$(H4)$, up to label switching,

$$\mathbb{E}\|\mathbf{Q}-\hat{\mathbf{Q}}\|^2 \leq C\frac{m^3\log(n)}{n}$$

$$\mathbb{E}\|\mathbf{f}_k-\hat{\mathbf{f}}_k\|_2^2 \leq \|\mathbf{f}_k-\mathbf{f}_{k,m}\|_2^2 + C\frac{m^3\log(n)}{n} \leq C'm^{-2\alpha} + C\frac{m^3\log(n)}{n}$$

where α regularity of functions $\mathbf{f}_k$

- for $\mathbf{Q}$: quasi-parametric rate of convergence
- for $\mathbf{f}_k$: rate of convergence $(n/\log(n))^{-\alpha/(2\alpha+3)}$
  $\rightarrow$ non optimal

Joint law and conditional law

Distribution of (Y_1, Y_2, Y_3):

$$g^{\mathbf{Q},\mathbf{f}}(y) = \sum_{k_1,k_2,k_3=1}^{K} \boldsymbol{\pi}(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3)\mathbf{f}_{k_1}(y_1)\mathbf{f}_{k_2}(y_2)\mathbf{f}_{k_3}(y_3)$$

Joint law and conditional law

Distribution of (Y_1, Y_2, Y_3):

$$g^{\mathbf{Q},\mathbf{f}}(y) = \sum_{k_1,k_2,k_3=1}^{K} \boldsymbol{\pi}(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3)\mathbf{f}_{k_1}(y_1)\mathbf{f}_{k_2}(y_2)\mathbf{f}_{k_3}(y_3)$$

(H_5) $P(\mathbf{Q}, \langle \mathbf{f}_k, \mathbf{f}_l \rangle) \neq 0$ $\quad$ P polynomial

$\rightarrow$ generically satisfied

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

Joint law and conditional law

Distribution of (Y_1, Y_2, Y_3):

$$g^{\mathbf{Q},\mathbf{f}}(y) = \sum_{k_1,k_2,k_3=1}^{K} \boldsymbol{\pi}(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3)\mathbf{f}_{k_1}(y_1)\mathbf{f}_{k_2}(y_2)\mathbf{f}_{k_3}(y_3)$$

(H_5) $P(\mathbf{Q}, \langle \mathbf{f}_k, \mathbf{f}_l \rangle) \neq 0$ P polynomial

$\rightarrow$ generically satisfied

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

Theorem (De Castro, Gassiat, L. 2016)

Under $(H1)$–$(H5)$, there exists $C > 0$ such that

$$\|g^{\mathbf{Q},\mathbf{f}} - g^{\mathbf{Q},\hat{\mathbf{f}}}\|_2 \geq C \sum_{k=1}^{K} \|\mathbf{f}_k - \hat{\mathbf{f}}_k\|_2$$

Detail of $(H5)$

$G(\mathbf{f})_{i,j} := \langle \mathbf{f}_i, \mathbf{f}_j \rangle$, $A := Diag(\boldsymbol{\pi})$. If U matrix s.t. $U\mathbf{1}_K = 0$,

$$\mathcal{D} := \sum_{i,j=1}^{K} \Big\{ \left(\mathbf{Q}^T AUG(\mathbf{f})U^T A\mathbf{Q}\right)_{i,j} \left(G(\mathbf{f})\right)_{i,j} \left(\mathbf{Q}G(\mathbf{f})\mathbf{Q}^T\right)_{i,j}$$
$$+ \left(\mathbf{Q}^T AG(\mathbf{f})A\mathbf{Q}\right)_{i,j} \left(UG(\mathbf{f})U^T\right)_{i,j} \left(\mathbf{Q}G(\mathbf{f})\mathbf{Q}^T\right)_{i,j}$$
$$+ \left(\mathbf{Q}^T AG(\mathbf{f})A\mathbf{Q}\right)_{i,j} \left(G(\mathbf{f})\right)_{i,j} \left(\mathbf{Q}UG(\mathbf{f})U^T\mathbf{Q}^T\right)_{i,j} \Big\}$$
$$+2 \sum_{i,j} \Big\{ \left(\mathbf{Q}^T AUG(\mathbf{f})A\mathbf{Q}\right)_{i,j} \left(UG(\mathbf{f})\right)_{j,i} \left(\mathbf{Q}G(\mathbf{f})\mathbf{Q}^T\right)_{i,j}$$
$$+ \left(\mathbf{Q}^T AUG(\mathbf{f})A\mathbf{Q}\right)_{i,j} \left(\mathbf{Q}UG(\mathbf{f})\mathbf{Q}^T\right)_{j,i} \left(G(\mathbf{f})\right)_{i,j}$$
$$+ \left(UG(\mathbf{f})\right)_{i,j} \left(\mathbf{Q}UG(\mathbf{f})\mathbf{Q}^T\right)_{j,i} \left(\mathbf{Q}^T AG(\mathbf{f})A\mathbf{Q}\right)_{i,j} \Big\}$$

defines a semidefinite positive quadratic form $\mathcal{D}$ in the coefficients $U_{i,j}$, $i = 1, \dots, K$, $j = 1, \dots, K-1$.
$P(\mathbf{Q}, G(\mathbf{f})) :=$ the numerator of the determinant of $\mathcal{D}$

Contrast minimization

We are looking for a function t minimizing

$$\begin{aligned} \|t - g^{\mathbf{Q},\mathbf{f}}\|^2 &= \|t\|^2 - 2\langle t, g^{\mathbf{Q},\mathbf{f}}\rangle + \|g^{\mathbf{Q},\mathbf{f}}\|^2 \\ &= \|t\|^2 - 2\mathbb{E}[t(Y_i, Y_{i+1}, Y_{i+2})] + \cancel{\|g^{\mathbf{Q},\mathbf{f}}\|^2} \end{aligned}$$

$$\Longrightarrow \hat{g}_m = \operatorname*{argmin}_{t \in S} \frac{1}{n} \sum_{i=1}^{n-2} \left(\|t\|^2 - 2t(Y_i, Y_{i+1}, Y_{i+2}) \right)$$

Approximation space

We are looking for an estimator among functional space

$$S_{m,\mathbf{Q}} = \left\{ t : \mathbb{R}^3 \to \mathbb{R}, \quad t(y) = \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)\mathbf{Q}(k_1,k_2)\mathbf{Q}(k_2,k_3) \right.$$
$$\left. \sum_{j_1,j_2,j_3=1}^{m} a_{j_1k_1}a_{j_2k_2}a_{j_3k_3}\varphi_{j_1}(y_1)\varphi_{j_2}(y_2)\varphi_{j_3}(y_3) \right\}$$

i.e. $\hat{\mathbf{f}}_k \in \mathrm{Vect}\{\varphi_1, \ldots, \varphi_m\}$

mK coefficients (a_{jk}) to estimate

Model selection

Collection of estimators:

$$\hat{g}_m = \operatorname*{argmin}_{t \in S_{m,\hat{\mathbf{Q}}}} \frac{1}{n} \sum_{i=1}^{n-2} \left(\|t\|^2 - 2t(Y_i, Y_{i+1}, Y_{i+2}) \right)$$

Choice of m: Birgé-Massart model selection

$$\hat{m} = \operatorname*{argmin}_{1 \leq m \leq n} \{ -\|\hat{g}_m\|^2 + \mathrm{pen}(m) \}$$

Finally $\hat{g} = \hat{g}_{\hat{m}}$

then $\hat{\mathbf{f}}_k$ such that $\hat{g} = g^{\hat{\mathbf{Q}}, \hat{\mathbf{f}}}$

Oracle inequality and rate of convergence

Theorem (De Castro, Gassiat, L. 2016)

If $\mathrm{pen}(m) = \rho \frac{m \log n}{n}$ *then, up to label switching,*

$$\sum_{k=1}^{K} \mathbb{E}\|\mathbf{f}_k - \hat{\mathbf{f}}_k\|_2^2 \leq C \min_m \{\|\mathbf{f}_k - \mathbf{f}_{k,m}\|_2^2 + \frac{m \log n}{n}\} + \frac{\log n}{n}$$

$$\leq C' \left(\frac{n}{\log n} \right)^{-2\alpha/(2\alpha+1)}$$

Quasi-optimal rate of convergence

Proof requires concentration inequality for dependent variables, and control of the complexity of $S_{m,\mathbf{Q}}$ with bracket entropy

Introduction

Spectral method

Penalized least squares

Final estimation

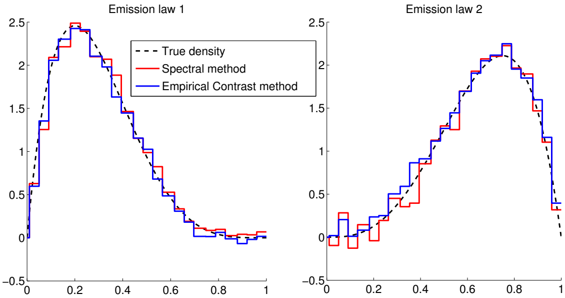
- Combination of both methods
- Simulations
- Prospects

Implementation

1. With spectral method, we obtain estimators $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}_k$

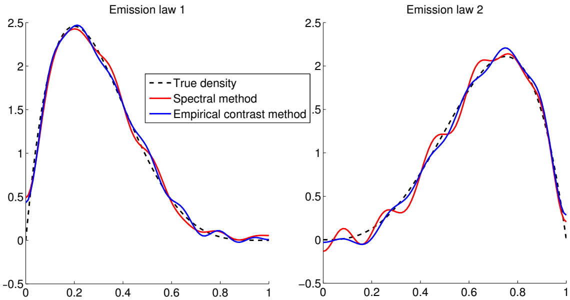
2. Use $\hat{\mathbf{Q}}$ to define $S_{m,\hat{\mathbf{Q}}}$ and $\hat{\mathbf{f}}_k$ as initial point of the constrast minimization
   (calibration of the penalty with slope heuristic of Birgé-Massart)

Simulations for $K = 2$



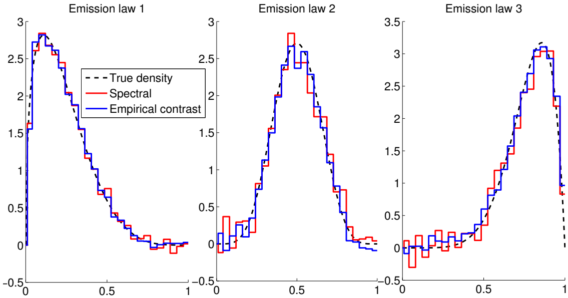
Reconstruction of densities f_1 and f_2 (Beta distributions) with spectral and least squares methods ($n = 50000$, histogram basis)

Simulations for $K = 2$



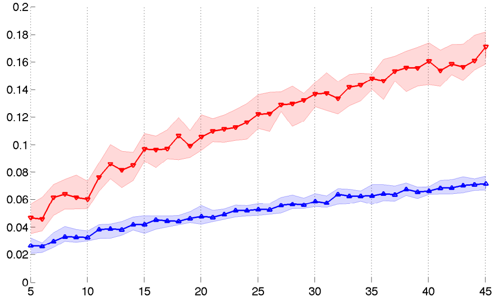
Reconstruction of densities f_1 and f_2 (Beta distributions) with spectral and least squares methods
($n = 50000$, trigonometric basis)

Simulations for $K = 3$



Reconstruction of densities f_1, f_2, f_3 (Beta distributions) with spectral and least squares methods ($n = 50000$, histogram basis)

Simulations for $K = 2$



Integrated variance $\mathbb{E}\|\hat{f}_k - f_{k,m}\|^2$ of spectral and least squares estimators, as a function of m ($n = 50000$, histogram basis)

Future works

- Estimation of the filtering and marginal smoothing distibutions
  De Castro, Gassiat, Lecorff (2016)

  same model, distribution of $X_i|Y_{1:i}$ and $X_i|Y_{1:n}$ using $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}$

Future works

- Estimation of the filtering and marginal smoothing distibutions De Castro, Gassiat, Lecorff (2016)

  same model, distribution of $X_i|Y_{1:i}$ and $X_i|Y_{1:n}$ using $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}$
- Estimation of K: Lehéricy (2016)

$$(\hat{K}, \hat{M}) = \underset{K \leq \log n, m \leq n}{\text{argmin}} \{-\|\hat{g}_{K,m}\|^2 + \text{pen}(K, m)\}$$

  with $\text{pen}(K, m) = (mK + K^2 - 1)\log(n)/n$

Future works

- Estimation of the filtering and marginal smoothing distibutions De Castro, Gassiat, Lecorff (2016)

  same model, distribution of $X_i|Y_{1:i}$ and $X_i|Y_{1:n}$ using $\hat{\mathbf{Q}}$ and $\hat{\mathbf{f}}$
- Estimation of K: Lehéricy (2016)

$$(\hat{K}, \hat{M}) = \underset{K \leq \log n, m \leq n}{\text{argmin}} \{-\|\hat{g}_{K,m}\|^2 + \text{pen}(K, m)\}$$

  with $\text{pen}(K, m) = (mK + K^2 - 1)\log(n)/n$
- $Y_i = f(X_i) + \varepsilon_i$ with X_i non-observed Markov chain Dumont Lecorff (2016)

  Rates of convergence to find...