Minimax adaptive estimation of nonparametric hidden Markov models

Yohann De Castro, Elizabeth Gassiat, Claire Lacour

LMO, Université Paris-Sud

IHES, January 2016
Outline

Introduction

Spectral method

Penalized least squares

Final estimation
Introduction

Model
State of the art
Assumptions
Projection on an approximation space

Spectral method

Penalized least squares

Final estimation
Model

- \((X_i)\) Markov chain on \(\{1, \ldots, K\}\): non-observed transition \(Q\) (matrix \(K \times K\))

- \(Y_1, \ldots, Y_n\) 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 \(Y_i|X_i = k\):

\[ f_k(y)dy = \mathbb{P}(Y_i \in dy|X_i = k) \]

\(f\) unknown
Hidden Markov Model

\[
\begin{align*}
X_i & \xrightarrow{Q} X_{i+1} \\
Y_i & \xrightarrow{f_{X_i}} Y_{i+1} \\
X_{i+1} & \xrightarrow{Q} X_{i+2} \\
Y_{i+1} & \xrightarrow{f_{X_{i+1}}} Y_{i+2}
\end{align*}
\]

Observations: \( Y_1, \ldots, Y_n \)
Known parameter: \( K \)
To estimate: transition matrix \( Q \), initial distribution \( \pi \), emission functions \( f_1, \ldots, f_K \)
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)
Alexandrovič, Holzmann (2014)
Assumptions

\((H_1)\) Q has full rank

\((H_2)\) \((X_i)\) irreducible aperiodic

\((H_3)\) stationary Markov chain

\((H_4)\) \(f_1, \ldots, f_K\) linearly independent
Identifiability

Distribution of \((Y_1, Y_2, Y_3)\):

\[
g_{Q,f}(y) := \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)Q(k_1, k_2)Q(k_2, k_3)f_{k_1}(y_1)f_{k_2}(y_2)f_{k_3}(y_3)
\]

Lemma

Under \((H_1)-(H_4)\), there is identifiability, up to label switching, from three consecutive observations:

\[g_{Q,f+h} = g_{Q,f} \iff \exists \tau \text{ permutation such that } h_j = f_j - f_{\tau(j)}\]
Projection on an approximation space

Approximation of the $f_k$ : for $(\varphi_1, \ldots, \varphi_m)$ orthonormal basis

$$f_{k,m} = \sum_{i=1}^{m} \langle f_k, \varphi_i \rangle \varphi_i$$

Examples: Fourier basis, Piecewise polynomials, Wavelets

**Aim**

To estimate matrices $Q = (\mathbb{P}(X_2 = j|X_1 = k))_{k,j}$

and $F = (\langle f_k, \varphi_i \rangle)_{ik} = (\mathbb{E}[\varphi_i(Y_1)|X_1 = k])_{ik}$

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

Spectral method
  Matrix expressions
  Algorithm
  Result

Penalized least squares

Final estimation
Matrix expressions

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

Lemma

• \[ P_1 = F \pi \]

\[ Y_1 \quad Y_1|X_1 \quad X_1 \]

• \[ P_{12} = F \text{Diag}(\pi)Q F^T \]

\[ (Y_1, Y_2) \quad Y_1|X_1 \quad X_1 \quad X_2|X_1 \quad Y_2|X_2 \]

Csq : Knowing \( F, P_1, P_{12} \) allows to recover \( \pi \) and \( Q \)
A crucial Lemma

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

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

\[ B(j) = R \text{ Diag}(F(j,.)) R^{-1} \]
Consequence

\[ B(j) = (U^T P_{13} U)^{-1} U^T P_{123}(., j, .) U = R \text{ Diag}(F(j, .)) R^{-1} \]

⇒ Diagonalizing \( B(j) \), \( j = 1 \ldots, m \) allows to recover \( F \)

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 \( \Theta \) 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.
  $\hat{P}_{13}(a, c) := \frac{1}{n} \sum_{i=1}^{n-2} \phi_a(Y_i)\phi_c(Y_{i+2})$

- $\hat{U}$ 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 $\hat{F}(j, k)$

- $\tilde{\pi} = (\hat{U}^T \hat{F})^{-1} \hat{U}^T \hat{P}_1$ and
  $\tilde{Q} = (\hat{U}^T \hat{F} \text{Diag}(\tilde{\pi}))^{-1} \hat{U}^T \hat{P}_{12} \hat{U} (\hat{F}^T \hat{U})^{-1}$

- $\hat{Q}$ projection of $\tilde{Q}$ on the space of transition matrices, and $\hat{\pi}$ its stationary distribution
Performance of the spectral method

**Theorem**

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

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

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

where \( \alpha \) 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
Introduction

Spectral method

**Penalized least squares**

Joint law and conditional law
Estimation of the joint distribution
Results

Final estimation
Joint law and conditional law

Distribution of \((Y_1, Y_2, Y_3)\):

\[
g^{Q,f}(y) = \sum_{k_1, k_2, k_3 = 1}^{K} \pi(k_1) Q(k_1, k_2) Q(k_2, k_3) f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3) \]
Joint law and conditional law

Distribution of \((Y_1, Y_2, Y_3)\):

\[
g^{Q,f}(y) = \sum_{k_1, k_2, k_3 = 1}^{K} \pi(k_1)Q(k_1, k_2)Q(k_2, k_3)f_{k_1}(y_1)f_{k_2}(y_2)f_{k_3}(y_3)
\]

\((H_5)\) \(P(Q, \langle f_k, f_l \rangle) \neq 0\) \hspace{1cm} \(P\) polynomial

\(\rightarrow\) generically satisfied

\(\rightarrow\) always satisfied if \(K = 2\)
Joint law and conditional law

Distribution of \((Y_1, Y_2, Y_3)\):

\[
g^{Q,f}(y) = \sum_{k_1, k_2, k_3=1}^K \pi(k_1)Q(k_1, k_2)Q(k_2, k_3)f_{k_1}(y_1)f_{k_2}(y_2)f_{k_3}(y_3)
\]

\((H_5)\) \(P(Q, \langle f_k, f_l \rangle) \neq 0\)

\[\rightarrow \text{generically satisfied}\]
\[\rightarrow \text{always satisfied if } K = 2\]

Theorem (De Castro, Gassiat, L. 2016)

Under \((H1)-(H5)\), there exists \(C > 0\) such that

\[
\|g^{Q,f} - g^{Q,\hat{f}}\|_2 \geq C \sum_{k=1}^K \|f_k - \hat{f}_k\|_2
\]
**Detail of \((H5)\)**

\[ G(f)_{i,j} := \langle f_i, f_j \rangle, \quad A := \text{Diag}(\pi). \text{ If } U \text{ matrix s.t. } U1_K = 0, \]

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

defines a semidefinite positive quadratic form \(\mathcal{D}\) in the coefficients \(U_{i,j}\), \(i = 1, \ldots, K, \quad j = 1, \ldots, K - 1\). 

\(P(Q, G(f)) := \text{the numerator of the determinant of } \mathcal{D}\)
Contrast minimization

We are looking for a function $t$ minimizing

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

$$
\implies \hat{g}_m = \arg\min_{t \in S} \frac{1}{n} \sum_{i=1}^{n-2} \left( \|t\|^2 - 2t(Y_i, Y_{i+1}, Y_{i+2}) \right)
$$
Approximation space

We are looking for an estimator among functional space

$$S_{m,Q} = \left\{ t : \mathbb{R}^3 \to \mathbb{R}, \quad t(y) = \sum_{k_1,k_2,k_3=1}^{K} \pi(k_1)Q(k_1,k_2)Q(k_2,k_3) \right\}$$

$$\sum_{j_1,j_2,j_3=1}^{m} a_{j_1k_1} a_{j_2k_2} a_{j_3k_3} \varphi_{j_1}(y_1) \varphi_{j_2}(y_2) \varphi_{j_3}(y_3)$$

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

$mK$ coefficients $(a_{jk})$ to estimate
Model selection

Collection of estimators:
\[ \hat{g}_m = \arg \min_{t \in S_{m, \hat{Q}}} \frac{1}{n} \sum_{i=1}^{n-2} (\|t\|^2 - 2t(Y_i, Y_{i+1}, Y_{i+2})) \]

Choice of \( m \): Birgé-Massart model selection
\[ \hat{m} = \arg \min_{1 \leq m \leq n} \{ -\|\hat{g}_m\|^2 + \text{pen}(m) \} \]

Finally \( \hat{g} = \hat{g}_{\hat{m}} \)
then \( \hat{f}_k \) such that \( \hat{g} = g^{\hat{Q}, \hat{f}} \)
Oracle inequality and rate of convergence

Theorem (De Castro, Gassiat, L. 2016)

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

\[
\sum_{k=1}^{K} \mathbb{E} \| f_k - \hat{f}_k \|_2^2 \leq C \min_m \{ \| f_k - f_{k,m} \|_2^2 + \frac{m \log n}{n} \} + \frac{\log n}{n}
\]

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

Quasi-optimal rate of convergence

Proof requires concentration inequality for dependent variables, and control of the complexity of \( S_{m,Q} \) 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 $\hat{Q}$ and $\hat{f}_k$

2. Use $\hat{Q}$ to define $S_{m,\hat{Q}}$ and $\hat{f}_k$ as initial point of the contrast 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 distributions
  De Castro, Gassiat, Lecorff (2016)

same model, distribution of \( X_i | Y_{1:i} \) and \( X_i | Y_{1:n} \) using \( \hat{Q} \) and \( \hat{f} \)
Future works

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

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

- Estimation of $K$: Lehéricy (2016)

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

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

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

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

- Estimation of $K$: Lehéricy (2016)

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

  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...