Nonparametric mixture models with finite state space

Judith Rousseau, with E. Gassiat and E. Vernet

CEREMADE, Université Paris-Dauphine & CREST - ENSAE

Luminy

Outline

1. Non parametric mixture models : static and dynamic mixtures
2. Various results on estimation
3. Case of static mixture : semiparametric estimation of **p**

Nonparametric mixture models - static and dynamic

- **Model**
  - Observations $Y_t|X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$

- **Parameters**

• Parameters from the emissions $Y|X : F_j, j = 1, \cdots, K$
• Parameters of the latent process X_t : $\mathbf{p}$ or Q.

Nonparametric mixture models - static and dynamic

- **Model**
  - Observations $Y_t|X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$
  - Hidden states $X_t \in \{1, \cdots, K\}$.

- **Parameters**

• Parameters from the emissions $Y|X$: $F_j, j = 1, \cdots, K$
• Parameters of the latent process X_t : $\mathbf{p}$ or Q.

Nonparametric mixture models - static and dynamic

- **Model**
  - Observations $Y_t|X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$
  - Hidden states $X_t \in \{1, \cdots, K\}$.
    - Static mixtures : $X_t \overset{iid}{\sim} \mathbf{p} = (p(1), \cdots, p(K))$

- **Parameters**

• Parameters from the emissions $Y|X$: $F_j, j = 1, \cdots, K$

• Parameters of the latent process X_t : $\mathbf{p}$ or Q.

Nonparametric mixture models - static and dynamic

- **Model**
  - Observations $Y_t | X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$
  - Hidden states $X_t \in \{1, \cdots, K\}$.
    - Static mixtures : $X_t \overset{iid}{\sim} \mathbf{p} = (p(1), \cdots, p(K))$
    - Dynamic : $(X_t)_t$= MC (Q) or asy. stationary
- **Parameters**

• Parameters from the emissions $Y|X$: $F_j, j = 1, \cdots, K$

• Parameters of the latent process X_t : $\mathbf{p}$ or Q.

Parametric mixture models : $Y \sim \sum_{j=1}^{k} p_j F_{\theta_j}$

- Observations $Y_t | X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$

Parametric mixture models : $Y \sim \sum_{j=1}^{k} p_j F_{\theta_j}$

- Observations $Y_t | X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$
- Hidden states $X_t \in \{1, \cdots, K\}$.

Parametric mixture models : $Y \sim \sum_{j=1}^{k} p_j F_{\theta_j}$

- Observations $Y_t | X_t = j \sim F_j(Y_t)$ $t = 1, \cdots, n$
- Hidden states $X_t \in \{1, \cdots, K\}$.
- $$F_j = F_{\theta_j}, \quad \text{e.g.} \quad \mathcal{N}(\mu_j, \sigma_j^2)$$

Identifiability issues– $Y \sim G_{p,F} = \sum_{j=1}^{K} p(j) F_j$

- **static mixtures** Non identifiability : (Allman et al.) but if

$$Y = (y_1, y_2, y_3) \quad \& \quad F_j = F_{j1} \otimes F_{j2} \otimes F_{j3}$$

with $(F_{j,\ell})_j$ linearly indpt and $p(j) > 0 \ \forall j$

$$\sum_{j=1}^{k} p(j) F_j = \sum_{j=1}^{k} p(j)' F_j' \quad \Rightarrow p(j) = p(j)' \quad F_j = F_j'$$

Dynamic mixtures

- **Location mixtures** Gassiat & R. stationarity &

$$Y_t = m_{X_t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} F, X_t \in \{1, \cdots, K\}$$

$$(Y_1, Y_2) \sim G^{(2)}_{Q,F} = \sum_{j_1,j_2} Q(X_1 = j_1, X_2 = j_2) F(. - m_{j_1}) F(. - m_{j_2})$$

If $det(Q), det(Q') > 0$ & $m_j \neq m_i$ Then

$$G^{(2)}_{Q,\mathbf{m}} = G^{(2)}_{Q',\mathbf{m}'} \quad \Rightarrow Q = Q' \quad m_j = m'_j \quad \forall j, \quad K = K', \quad F = F'$$

- **General HMMs** Gassiat et al. if (X_t) MC (Q) Then if If $det(Q) > 0$ & linear indpdc of $(F_j)_j$

$$G^{(3)}_{Q,F} = G^{(3)}_{Q',F'} \quad \Rightarrow Q = Q' \quad F_j = F'_j \quad \forall j, \quad K = K'$$

Bayesian nonparametric estimation in HMMs : E. Vernet

$$Y_t | X_t = j \sim f_j, \quad (X_t) = CM(Q)$$

- General posterior concentration theorem :

$$\Pi(\|g_{Q,f} - g_{Q',f'}\|_1 \leq \epsilon_n | Y_{1:n}) = 1 + o_p(1)$$

$$g_{Q,f}(Y_1, Y_2) = \sum_{j_1, j_2} Q(X_1 = j_1, X_2 = j_2) f_{j_1}(Y_1) f_{j_2}(Y_2)$$

- Issues : What about

$$\|Q - Q'\|?, \quad \|f_j - f_j'\|_1?$$

Not trivial

Frequentist results on $\mathbf{p}$ or Q - moment and spectral methods

- **Mixtures** Bonhomme et al. , Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{\mathbf{p}} - \mathbf{p}^*\|\right) = O(1/\sqrt{n})$$

- **HMMs** Gassiat & R. , Cleynen, Gassiat & Robin, Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{Q} - Q^*\|\right) = O(1/\sqrt{n})$$

- **Questions :**
  - Construction on Bayesian estimators of $\mathbf{p}$ and Q with rate $1/\sqrt{n}$?

Frequentist results on $\mathbf{p}$ or Q - moment and spectral methods

- **Mixtures** Bonhomme et al. , Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{\mathbf{p}} - \mathbf{p}^*\|\right) = O(1/\sqrt{n})$$

- **HMMs** Gassiat & R. , Cleynen, Gassiat & Robin, Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{Q} - Q^*\|\right) = O(1/\sqrt{n})$$

- **Questions :**
  - Construction on Bayesian estimators of $\mathbf{p}$ and Q with rate $1/\sqrt{n}$?
  - Asymptotic normality ?

Frequentist results on **p** or Q - moment and spectral methods

- **Mixtures** Bonhomme et al. , Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{\mathbf{p}} - \mathbf{p}^*\| \right) = O(1/\sqrt{n})$$

- **HMMs** Gassiat & R. , Cleynen, Gassiat & Robin, Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{Q} - Q^*\| \right) = O(1/\sqrt{n})$$

- **Questions :**
  - Construction on Bayesian estimators of **p** and Q with rate $1/\sqrt{n}$?
  - Asymptotic normality ?
  - BvM ?

Frequentist results on **p** or Q - moment and spectral methods

- **Mixtures** Bonhomme et al. , Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{\mathbf{p}} - \mathbf{p}^*\|\right) = O(1/\sqrt{n})$$

- **HMMs** Gassiat & R. , Cleynen, Gassiat & Robin, Anandkumar et al.

$$\mathbb{E}^* \left(\|\hat{Q} - Q^*\|\right) = O(1/\sqrt{n})$$

- **Questions :**
  - Construction on Bayesian estimators of **p** and Q with rate $1/\sqrt{n}$?
  - Asymptotic normality ?
  - BvM ?
  - efficiency ?

Use of the identifiability result of Allman et al.

$$Y = (y_1, y_2, y_3) \overset{iid}{\sim} g_{\mathbf{p},F} = \sum_{j=1}^{K} p(j) f_j^{(1)} \otimes f_j^{(2)} \otimes f_j^{(3)}$$

$$\text{case} : f_j^{\otimes 3}(y) = f_j(y_1) f_j(y_2) f_j(y_3), \quad y = (y_1, y_2, y_3)$$

- **Prior model** Piecewise constant densities

- Let $\mathcal{I}(L) = (I_1, \cdots, I_L)$ be an *admissible* partition of $[0, 1]$, s.t

$$\text{rank} \begin{pmatrix} F_1^*(I_1) & \cdots & F_1^*(I_L) \\ F_2^*(I_1) & \cdots & F_2^*(I_L) \\ \ldots & \cdots & \cdots \\ F_K^*(I_1) & \cdots & F_K^*(I_L) \end{pmatrix} = K$$

- Parameters given $\mathcal{I}$:

$$f_j(y) = \sum_{\ell=1}^{L} \frac{w_{j,\ell}}{|I_\ell|} \mathbb{1}_{y \in I_l}, \quad \sum_{\ell} w_{j,\ell} = 1, \quad w_{j,\ell} > 0, \quad \forall j \leq K$$

- Prior :

$$\mathbf{w}_j \overset{iid}{\sim} \pi_w, \quad \mathbf{p} \sim \pi_p$$

First simple result : fixed $\mathcal{I}$, non efficient BvM

If $L \geq K$ and $\mathcal{I}$ is admissible and $p(j) > 0 \ \forall j$,

$$\Pi\left(\sqrt{n}(\mathbf{p} - \hat{\mathbf{p}_{\mathcal{I}}}) \leq t | Y_{1:n}, \mathcal{I}\right) \to Pr\left(\mathcal{N}(0, J_{\mathcal{I}}^{-1}) \leq t\right)$$

with
$\hat{\mathbf{p}_{\mathcal{I}}} = MLE$ in model $f_j(x) = \sum_\ell \frac{w_{j,\ell}}{|I_\ell|} \mathbb{1}_{x \in I_l}$
$J_{\mathcal{I}} := J_{\mathcal{I}}(\mathbf{p}^*, \mathbf{f}^*)$ = Fisher info

$$\sqrt{n}(\mathbf{p}^* - \hat{\mathbf{p}_{\mathcal{I}}}) \to \mathcal{N}(0, J_{\mathcal{I}}^{-1}), \quad G_{\mathbf{p}^*, \mathbf{f}^*}$$

- **So** BvM and

$$\mathbb{E}^*\left(\|\hat{\mathbf{p}} - \mathbf{p}^*\|\right) = O(1/\sqrt{n})$$

Comments : $Y_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})$

$$n_{\underline{\ell}} = \sum_{i=1}^{n} \mathbb{1}_{Y_{i,1} \in I_{\ell_1}} \mathbb{1}_{Y_{i,2} \in I_{\ell_2}} \mathbb{1}_{Y_{i,3} \in I_{\ell_3}}, \quad \underline{\ell} = (\ell_1, \ell_2, \ell_3)$$

- fixed $\mathcal{I}$: Simple case since regular parametric model with data $\mathbf{N} = (n_{\underline{\ell}}, \underline{\ell} \in \{1, \cdots, L\}^3)$,

- No model mis-specification but *data reduction*

- Behaviour of $J_{\mathcal{I}}$ when $\mathcal{I}$ varies ? when $|\mathcal{I}|$ increases ?

- How can we choose $\mathcal{I}$?
- How can we choose L ?

Efficient estimation of $\mathbf{p}$

For any sequence of embedded partitions $(\mathcal{I}_L)_L$
For any $L_n \to +\infty$

$$J_{\mathcal{I}_{L_n}} \to J_0 \quad \text{efficient Fisher info}$$

Therefore choosing $L_n \to +\infty$ slowly

- Asymptotic normality of the MLE $\hat{\mathbf{p}}_{\mathcal{I}_{L_n}}$ + efficiency

$$\sqrt{n} J_0^{1/2} (\hat{\mathbf{p}}_{\mathcal{I}_{L_n}} - \mathbf{p}^*) \Rightarrow \mathcal{N}(0, id), \quad P_{\mathbf{p}^*, \mathbf{f}^*}$$

Efficient estimation of $\mathbf{p}$

For any sequence of embedded partitions $(\mathcal{I}_L)_L$
For any $L_n \to +\infty$

$$J_{\mathcal{I}_{L_n}} \to J_0 \quad \text{efficient Fisher info}$$

Therefore choosing $L_n \to +\infty$ slowly

- Asymptotic normality of the MLE $\hat{\mathbf{p}}_{\mathcal{I}_{L_n}}$ + efficiency

$$\sqrt{n} J_0^{1/2}(\hat{\mathbf{p}}_{\mathcal{I}_{L_n}} - \mathbf{p}^*) \Rightarrow \mathcal{N}(0, id), \quad P_{\mathbf{p}^*, \mathbf{f}^*}$$

- BvM

$$\left[\sqrt{n} J_0^{1/2}(\mathbf{p} - \hat{\mathbf{p}}_{\mathcal{I}_{L_n}}) | Y_{1:n}, \mathcal{I}_{L_n}\right] \Rightarrow \mathcal{N}(0, id),$$

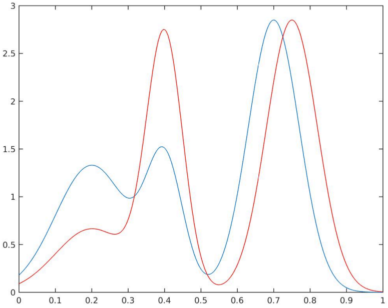
Some simulation results : K=2

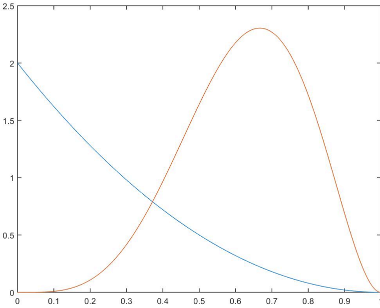
- **Data 1 :**$p = 0.3$ (difficult)

$$f_1 = \frac{1}{3}*\mathcal{N}(0.2, 0.01)\mathbb{1}_{|\cdot|\leq 1}+\frac{1}{2}*\mathcal{N}(0.7, 0.07^2)\mathbb{1}_{|\cdot|\leq 1}+\frac{1}{6}\mathcal{N}(0.4, 0.05)\mathbb{1}_{|\cdot|\leq 1}$$

$$f_2 = \frac{1}{3}*\mathcal{N}(0.2, 0.01)\mathbb{1}_{|\cdot|\leq 1}+\frac{1}{2}*\mathcal{N}(0.77, 0.07^2)\mathbb{1}_{|\cdot|\leq 1}+\frac{1}{6}\mathcal{N}(0.4, 0.05)\mathbb{1}_{|\cdot|\leq 1}$$

- **Data 2 :** $p = 0.3$ $f_1 = \text{Beta}(1, 2)$, $f_2 = \text{Beta}(5, 3)$ (easy)





Results : $\mathbb{E}^*(p^* - \hat{p})^2$, $\hat{p} = E[p|\mathbf{y}^n]$. First easy

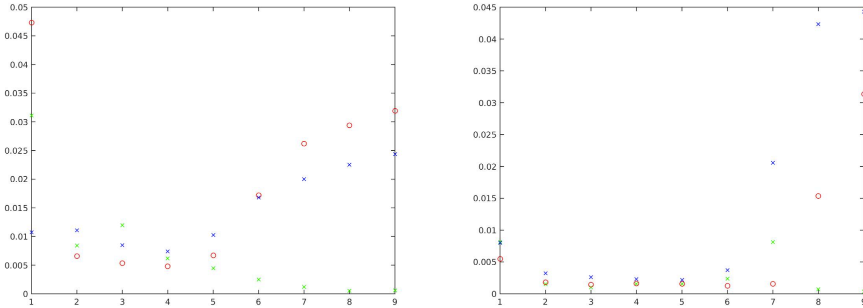
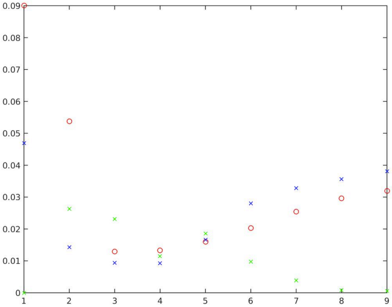


FIG.: Data 2, n=100 (left), n= 500 (right)

Results : $\mathbb{E}^*(p^* - \hat{p})^2$, $\hat{p} = E[p|\mathbf{y}^n]$.



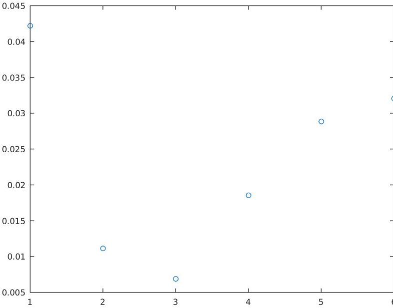


FIG.: Data 1, n=100 : left = pfixed partition, right = empirical

Criteria to select L_n

- **Sequence of embedded partition** $(\mathcal{I}_L)_L$

$$R(p^*, L) = \mathbb{E}^*[\|p^* - \hat{p}_L\|^2], \quad \hat{p}_L = E^\pi(p|\mathbf{y}^n, \mathcal{I}_L), \quad L \geq K$$

Choose L that minimizes $R(p^*; L) \Rightarrow$ Need to estimate $R(p^*; L)$.
Let $L_0 > K$ small, random split of the sample $y_1, \cdots, y_n$ in two, $b = 1, \cdots, B$

$$\hat{R}(p^*, L) = B^{-1} \sum_{b=1}^{B} (\hat{p}_{L_0}(-b) - \hat{p}_L(b))^2$$

- **Theory : on going work**

Some practical choices for $\mathcal{I}_{L_n}$

- We can choose a sequence of embedded partition and select L_n using a criteria

Some practical choices for $\mathcal{I}_{L_n}$

- We can choose a sequence of embedded partition and select L_n using a criteria
- Empirical partition : unconditional or conditional

Some practical choices for $\mathcal{I}_{L_n}$

- We can choose a sequence of embedded partition and select L_n using a criteria
- Empirical partition : unconditional or conditional
- data dependent partition based on risk minimization

Empirical partition : the unconditional approach

- **empirical quantiles** marginal density

$$f^*(y) = \sum_{j=1}^{K} p_j^* f_j^*(y), \quad q_{t,L} : \; F(q_{t,L}) = \sum_j p_j^* F_j^*(q_{t,L}) = \frac{t}{L}, \quad t \leq L-1$$

$$B_{t,L}^* = q_{t,L} - q_{t-1,L} \quad \text{replaced by} \quad \hat{B}_{t,L}^* = \hat{q}_{t,L} - \hat{q}_{t-1,L}$$

$$\text{empirical quantiles} : \frac{1}{3n} \sum_{i=1}^{n} \sum_{s=1}^{3} \mathbb{1}_{y_{i,s} \leq \hat{q}_{t,L}} = \frac{t}{L}$$

- **Unconditional approach** pretend $\hat{B}_{.,L}$ does not depend on the data.

"BvM"

$$\left[\sqrt{n} J_0^{1/2} (\mathbf{p} - \hat{\mathbf{p}}_{\mathcal{I}_{L_n}}) | Y_{1:n}, \hat{\mathcal{I}}_{L_n}\right] \Rightarrow \mathcal{N}(0, id),$$

but

$$\sqrt{n} J_0^{1/2} (\hat{\mathbf{p}}_{\mathcal{I}_{L_n}} - \mathbf{p}^*) \Rightarrow \mathcal{N}(0, id), \quad P_{\mathbf{p}^*, \mathbf{f}^*} ???$$

Why BvM and not MLE ?

- For "BvM" : Enough to have consistency +

$$\frac{1}{n} \sup_{|p-p^*|<\epsilon; |w-w^*|<\epsilon} \left| D^2 \ell_n(p, w|\mathcal{I}_L) - D^2 \ell_n(p, w|\hat{\mathcal{I}}_L) \right| = o_p(1)$$

true because

$$|\hat{q}_{t,L} - q^*_{t,L}| = O_p(n^{-1/2})$$

Why BvM and not MLE ?

- For "BvM" : Enough to have consistency +

$$\frac{1}{n} \sup_{|p-p^*|<\epsilon; |w-w^*|<\epsilon} \left| D^2 \ell_n(p, w|\mathcal{I}_L) - D^2 \ell_n(p, w|\hat{\mathcal{I}}_L) \right| = o_p(1)$$

  true because

$$|\hat{q}_{t,L} - q^*_{t,L}| = O_p(n^{-1/2})$$

- For asymp normality of MLE

$$\frac{1}{\sqrt{n}} \left| D\ell_n(p^*, w^*|\mathcal{I}_L) - D\ell_n(p^*, w^*|\hat{\mathcal{I}}_L) \right| = o_p(1)$$

Empirical partition : conditional approach : Polya tree prior Holmes et al.

- **Polya tree prior** (Holmes et al. 2013)

$$\mathcal{T} = \Big\{(B_0, B_1), (B_{0,0}, B_{0,1}, B_{1,0}, B_{1,1}), \cdots, (B_\epsilon, \epsilon \in \{0,1\}^k), \quad k \in \mathbb{N}^*\Big\}$$

$$F \Leftrightarrow \left(\theta_{\epsilon,0} = F(B_{\epsilon,0}|B_\epsilon), \epsilon \in \{0,1\}^m, m \geq 0\right)$$

- **At level** $m+1$ **:** $\epsilon \in \{0,1\}^m$,

$$\theta_{\epsilon,0} := F(B_{\epsilon,0}|B_\epsilon) \sim \mathrm{Beta}(\alpha_k, \alpha_k), \quad \alpha_k = a(k+1)^c, \quad c > 1$$

- **Truncated Polya tree** We stop at level M.
- **Here** $F_j \overset{iid}{\sim} PT(\mathcal{T}_{[M]}, \underline{\alpha})$, $(p_1, \cdots, p_k) \sim \mathcal{D}(a_1, \cdots, a_k)$.

How do we choose $\mathcal{T}_{[M]}$?

Conditional approach on the empirical partition

$$\mathbf{y} = (Y_{i,j}, i \leq n, j \leq 3), \quad \text{Empirical quantiles on } \mathbf{y}$$

$$\Downarrow$$

$$\hat{\mathcal{T}} = \hat{B}_\epsilon, \epsilon \in \{0,1\}^m, \quad m \leq M$$

- **Full conditional "likelihood"**

$$L(\mathbf{y}, \mathbf{x} | \hat{\mathcal{T}}) = \prod_{m \leq M-1} \prod_{\epsilon \in \{0,1\}^m} EHG(\mathbf{n}_{\epsilon,0}^{(j)}, j \leq k | \mathbf{n}_{\epsilon,0}, \mathbf{n}_\epsilon^{(j)}, \theta_{\epsilon,0}^{(j)}, j \leq k)$$

$$n_\epsilon^{(j)} = \sum \mathbb{I}_{y_{i,j} \in B_\epsilon} \mathbb{I}_{x_i = j}$$

- **Bayesian approach** $\theta_\epsilon^{(j)} \overset{ind}{\sim} \text{Beta}(\alpha_m, \alpha_m)$
- **For the moment : no theory**

Some simulations : conditional approach

$$F = 0.35 * \mathcal{N}(0.5, 0.01) * \mathbb{1}_{|\mathcal{N}(0.5,0.01)| \leq 1} + 0.65 * \mathcal{U}(0, 1)$$

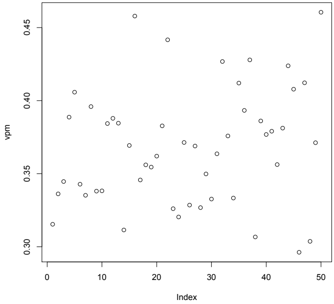


FIG.: n=100, 50 replicates, mean = 0.367975

$$F = 0.35 * \mathcal{N}(0.5, 0.01) * \mathbb{I}_{|\cdot| \leq 1} + 0.65 * \mathcal{E}(1) * \mathbb{I}_{|\mathcal{E}(1)| \leq 1}$$

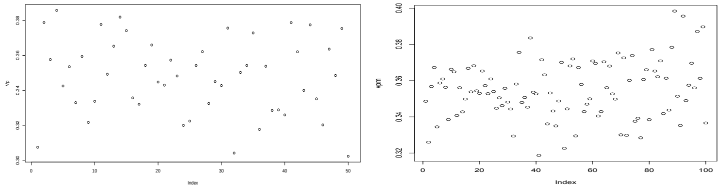


FIG.: left : n=500, 50 replicates, $\hat{p}$ = 0.348, right : n= 1000, 100 replicates , $\hat{p}$ = 355

Open questions – on going work

- Prove the theoretical properties of $\hat{R}_L(p^*)$: but at best for $n/2$ individuals : to be on the safe side

Open questions – on going work

- Prove the theoretical properties of $\hat{R}_L(p^*)$: but at best for $n/2$ individuals : to be on the safe side
- alternative : Bootstrap approach ?

Open questions – on going work

- Prove the theoretical properties of $\hat{R}_L(p^*)$: but at best for $n/2$ individuals : to be on the safe side
- alternative : Bootstrap approach ?
- Understand the behaviour of the conditional empirical approach

Conclusion : A stone in the pond – flexibility can be misleading

- NP mixtures : efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j

Conclusion : A stone in the pond – flexibility can be misleading

- NP mixtures : efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
- Can we generalize to other models ?

Conclusion : A stone in the pond – flexibility can be misleading

- NP mixtures : efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
- Can we generalize to other models ?
- Semi - parametric problems : targeted likelihood.

Conclusion : A stone in the pond – flexibility can be misleading

- NP mixtures : efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
- Can we generalize to other models ?
- Semi - parametric problems : targeted likelihood.
- Shall we change likelihood for different parameter of interests

Conclusion : A stone in the pond – flexibility can be misleading

- NP mixtures : efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
- Can we generalize to other models ?
- Semi - parametric problems : targeted likelihood.
- Shall we change likelihood for different parameter of interests
- Shall we mix $\pi(p|y^n, \mathcal{I})$ with NP $\pi(f_1, \cdots, f_K|p, y^n)$?

Thank you