Mixture models : static and dynamical models ; parametric and nonparametric cases

J. Rousseau

ENSAE - CEREMADE, Université Paris-Dauphine

Vienna, Joint work with E. Gassiat; Z. van Havre, K. Mengersen and E. Vernet

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of **p**

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of **p**

Parametric static and dynamic mixtures

- **Model :** Observations $Y_t, t \leq n$
  - Observations given latent stats X_t

$$[Y_t|X_t = j] \sim g_{\gamma_j}, \quad \gamma_j \in \Gamma$$

- Aim : estimation of θ : k is fixed

Parametric static and dynamic mixtures

- **Model :** Observations $Y_t, t \leq n$
  - Observations given latent stats X_t

$$[Y_t|X_t = j] \sim g_{\gamma_j}, \quad \gamma_j \in \Gamma$$

  - Latent states $X_t \in \{1, \cdots, k\}$

- Aim : estimation of θ : k is fixed

Parametric static and dynamic mixtures

- **Model :** Observations $Y_t, t \leq n$
  - Observations given latent stats X_t

$$[Y_t|X_t = j] \sim g_{\gamma_j}, \quad \gamma_j \in \Gamma$$

  - Latent states $X_t \in \{1, \cdots, k\}$
    - Static mixture :

$$X_t \overset{i.i.d}{\sim} \mathbf{p} = (p_1, \cdots, p_k)$$

- Aim : estimation of θ : k is fixed

Parametric static and dynamic mixtures

- **Model :** Observations $Y_t, t \leq n$
  - Observations given latent stats X_t

$$[Y_t|X_t = j] \sim g_{\gamma_j}, \quad \gamma_j \in \Gamma$$

  - Latent states $X_t \in \{1, \cdots, k\}$
    - Static mixture :

$$X_t \overset{i.i.d}{\sim} \mathbf{p} = (p_1, \cdots, p_k)$$

    - Dynamical : Hidden Markov model $(X_t)_t$ = MC with transition matrix $Q = (q_{i,j})_{i,j \leq k}$

$$P[X_t = j|X_{t-1} = i] = q_{i,j}$$

- Aim : estimation of θ : k is fixed

Parametric static and dynamic mixtures

- **Model :** Observations $Y_t, t \leq n$
  - Observations given latent stats X_t

$$[Y_t|X_t = j] \sim g_{\gamma_j}, \quad \gamma_j \in \Gamma$$

  - Latent states $X_t \in \{1, \cdots, k\}$
    - Static mixture :

$$X_t \overset{i.i.d}{\sim} \mathbf{p} = (p_1, \cdots, p_k)$$

    - Dynamical : Hidden Markov model $(X_t)_t$ = MC with transition matrix $Q = (q_{i,j})_{i,j \leq k}$

$$P[X_t = j|X_{t-1} = i] = q_{i,j}$$

  - **parameters** $\theta = (Q, \gamma_1, \cdot, \gamma_k)$ or $\theta = (\mathbf{p}, \gamma_1, \cdot, \gamma_k)$

- Aim : estimation of θ : k is fixed

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of **p**

Known results : $\theta^* = (\mathbf{p}^*, \gamma_1^*, \cdots, \gamma_{k_0}^*)$ true parameter

- **Weak identifiability** : labels non identifiable – switching issue.
- **If** $k = k_0$: Model is regular : For static and dynamic mixtures
  - MLE and posterior distribution are consistent $(1/\sqrt{n})$
  - BVM (de gunst et al. in HMMs)

$$[\sqrt{n}(\theta - \hat{\theta})|\, Y_{1:n}] \Rightarrow \mathcal{N}(0, I_0^{-1}), \quad P_{\theta_0}$$

- **If** $k > k_0$ Model Misspecification : Non identifiability (strong)
  - Emptying of extra states :

$$f^* = \sum_{j=1}^{k_0} p_j^* g_{\gamma_j^*} + \sum_{j=k_0+1}^{k} p_j g_{\gamma_j}, \quad p_{k_0+1} = \cdots = p_k = 0$$

Known results : $\theta^* = (\mathbf{p}^*, \gamma_1^*, \cdots, \gamma_{k_0}^*)$ true parameter

- **Weak identifiability** : labels non identifiable – switching issue.
- **If** $k = k_0$: Model is regular : For static and dynamic mixtures
- MLE and posterior distribution are consistent $(1/\sqrt{n})$
- BVM (de gunst et al. in HMMs)

$$[\sqrt{n}(\theta - \hat{\theta})|\, Y_{1:n}] \Rightarrow \mathcal{N}(0, I_0^{-1}), \quad P_{\theta_0}$$

- **If** $k > k_0$ Model Misspecification : Non identifiability (strong)
  - Emptying of extra states :

$$f^* = \sum_{j=1}^{k_0} p_j^* g_{\gamma_j^*} + \sum_{j=k_0+1}^{k} p_j g_{\gamma_j}, \quad p_{k_0+1} = \cdots = p_k = 0$$

  - Merging of extra states :

$$f^* = \sum_{j=1}^{k_0-1} p_j^* g_{\gamma_j^*} + \sum_{j=k_0}^{k} p_j g_{\gamma_j}, \quad \gamma_{k_0} = \cdots = \gamma_k = \gamma_{k_0}^*, \sum_{j=k_0}^{k} p_j = p_{k_0}^*$$

Asymptotic results if $k > k_0$ R. & Mengersen

$$f_\theta(x) = \sum_{j=1}^{k} p_j g_{\gamma_j}, \quad f^*(x) = \sum_{j=1}^{k_0} p_j^* g_{\gamma_j^*}, \quad k_0 < k$$

- **frequentist results** $\hat{\theta} \to \Theta^* = \{\theta; f_\theta = f_0\}$ but no more
- **Bayesian results**

• If $\mathbf{p} \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$ & $\gamma_j \in \mathbb{R}^d \sim \pi_\gamma$ i.i.d

If $\max_j \alpha_j < d/2$: Posterior concentrates on emptying the extra states

If $\min_j \alpha_j > d/2$: Posterior concentrates on merging the extra states

• Extension to $\mathbf{p} \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$ and repulsive $\pi_k(\gamma_1, \cdots, \gamma_k)$ (Dunson et al. 2011)

Hidden Markov models : what can we say ?

- **If** $k > k_0$ Model Misspecification : Non identifiability (strong)

empty state $\theta^* = (\gamma_1^*, \cdots, \gamma_{k_0}^*, \gamma_{k_0+1}, Q^*)$ with

$$Q_0 = \begin{pmatrix} Q^* & 0 \\ q_{k_0.} & 0 \end{pmatrix}, \quad Q^* = (q_{ij}^*, i, j \leq k_0)$$

same likelihood as merging states $\bar{\theta} = (\gamma_1^*, \cdots, \gamma_{k_0}^*, \gamma_{k_0}^*, \bar{Q})$ with

$$\bar{Q} = \begin{pmatrix} Q^*_{.[k_0-1]} & q_{.k_0} & q_{.k_0+1} \\ q^*_{k_0.} & q_{k_0k_0} & q_{k_0k_0+1} \\ q^*_{k_0.} & q_{k_0k_0} & q_{k_0k_0+1} \end{pmatrix}, \quad Q^*_{.[k_0-1]} \in [0,1]^{k_0 \times k_0 - 1},$$

$$q_{ik_0} + q_{ik_0+1} = q^*_{ik_0}$$

Understanding how the Bayesian penalization operates : case of static mixtures

$$f_\theta(x) = \sum_{j=1}^{k} p_j g_{\gamma_j}, \quad f^*(x) = \sum_{j=1}^{k_0} p_j^* g_{\gamma_j^*}, \quad k_0 < k$$

- **Posterior concentration to** f^* $B_n = \{\|f_\theta - f^*\|_1 \leq M_n/\sqrt{n}\}$

$$\pi\left(B_n | Y^n\right) = 1 + o_p(1)$$

- **Looking for sparsest path**

$$A_n = \{\theta; \theta \in B_n \,\&\, \sum_{j>k_0} p_j \lesssim M_n/\sqrt{n}\}$$

prior penalization :

$$\pi(A_n^c \cap B_n) = o(\pi(B_n))$$

Questions

- How can we extend this to HMMs ?

Questions

- How can we extend this to HMMs ?
- There is no asy. theory for MLE apart from consistency : Do we get posterior concentration ? to what ?

Questions

- How can we extend this to HMMs ?
- There is no asy. theory for MLE apart from consistency : Do we get posterior concentration ? to what ?
- Can we target the prior so that we empty/merge extra states ?

Questions

- How can we extend this to HMMs ?
- There is no asy. theory for MLE apart from consistency : Do we get posterior concentration ? to what ?
- Can we target the prior so that we empty/merge extra states ?
- Can we derive a procedure to estimate k_0 ?

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of **p**

Posterior concentration to marginals

- **framework :** 2 **marginals of** (Y_1, Y_2) . Set $pQ = p$

$$f_{2,\theta}(y_1, y_2) = \sum_{j_1, j_2 = 1}^{k} p_{j_1} q_{j_1 j_2} \prod_{t=1}^{2} g_{\gamma_{j_t}}(y_t)$$

$$f_2^*(y_1, y_2) = \sum_{j_1, j_2 = 1}^{k} p_{j_1}^* q_{j_1 j_2}^* \prod_{t=1}^{2} g_{\gamma_{j_t}^*}(y_t) \quad k > k_0$$

- prior $\pi(\theta) = \pi_Q(q_{ij}, i, j \leq k)\pi_\gamma(\gamma_1, ..., \gamma_k)$
- posterior

$$d\pi(\theta | Y_{1:n}) = \frac{e^{\ell_n(\theta)} d\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)} d\pi(\theta)}$$

- **Posterior concentration (penalized)**

$$\Pi\left((\rho_Q - 1) \| f_{2,\theta} - f_2^* \|_1 \lesssim \sqrt{\log n / n} \middle| Y_{1:n} \right) = 1 + o_p(1)$$

with ρ_Q : mixing coef of Q :

$$\| Q_{i.}^n - \mathbf{p} \|_1 = \sum_j |P[X_n = j | X_0 = i] - \mathbf{p}(j)| \lesssim \rho_Q^{-n}$$

Posterior concentration to f_2^*

$$\Pi\left((\rho_Q - 1)\|f_{2,\theta} - f_2^*\|_1 \lesssim \sqrt{\log n/n}\middle| Y_{1:n}\right) = 1 + o_p(1)$$

- **Need to get rid of** $(\rho_Q - 1)$ BUT $\exists\theta \approx \theta_0$ s.t. $\rho_Q \approx 1$

$$\rho_Q - 1 \gtrsim \sum_j \min_i q_{ij} \quad \Rightarrow \text{forbid small } \sum_j \min_i q_{ij}$$

- **result** If $\sum_j \alpha_j > k_0(k_0 - 1 + d) + k\sum_{j>k_0}\alpha_j$:
$\mathbf{q}_{i.} \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$

$$\Pi\left(\|f_{2,\theta} - f_2^*\|_1 \lesssim u_n\middle| Y_{1:n}\right) = 1 + o_p(1)$$

$$o(1) = u_n >> \log n/\sqrt{n}$$

Can we recover θ^* ?

Posterior concentration to $\Theta^* = \{\theta, f_\theta = f_2^*\}$

- **Uneven α_j's** : $p \leq k_0$

$$\bar{\alpha} = \alpha_1 = \cdots = \alpha_p > \alpha_{p+1} = \cdots = \alpha_k = \underline{\alpha}$$

If $\underline{\alpha}((k-k_0)^2 - (k-2k_0-1)) < d/2$ and
$p\bar{\alpha} + (k-p)\underline{\alpha} > F(k, k_0, d, \underline{\alpha})$

$$\pi\left(\sum_{j=k_0+1}^{k} p(j) \geq v_n | Y_{1:n}\right) = o_p(1), \quad o(1) = v_n >> n^{-1/2}$$

Emptying of the extra states : We recover θ_0

Undertanding what happens

- **Large** $\bar{\alpha}$

To avoid regions of Θ where Q non ergodic ($\rho_Q \approx 1$)

- **Small** $\underline{\alpha}$

To help $q_{ij} \approx 0$ $j \geq k_0 + 1$

- **Grey area**
  - What happens if all α_j's are small ?

Undertanding what happens

- **Large** $\bar{\alpha}$

To avoid regions of Θ where Q non ergodic ($\rho_Q \approx 1$)

- **Small** $\underline{\alpha}$

To help $q_{ij} \approx 0$ $j \geq k_0 + 1$

- **Grey area**
  - What happens if all α_j's are small ?
  - Diagonal configuration :

$$\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_2 \\ \alpha_2 & \alpha_1 & \cdots & \alpha_2 \\ \cdot & \cdots & \cdots & \cdots \\ \alpha_2 & \alpha_2 & \cdots & \alpha_1 \end{pmatrix}, \quad \alpha_1 >> 1, \quad \alpha_2 << 1$$

used in practice – no theory

Merging : case $k = 2$ with $k_0 = 1$

- **Transition matrix**

$$Q = \begin{pmatrix} 1 - q_1 & q_1 \\ q_2 & 1 - q_2 \end{pmatrix}, \mathbf{p} = \left(\frac{q_2}{q_1 + q_2}, \frac{q_1}{q_1 + q_2} \right), \quad \mathbf{p}Q = \mathbf{p}$$

- **prior**

$$q_1, q_2 \sim \text{Beta}(\alpha, \beta), \quad \text{ind}$$

- **Result** If $\alpha, \beta > 3d/4$, $\forall \epsilon_n = o(1)$

$$\pi \left(\mathbf{p}(1) \wedge \mathbf{p}(2) \leq \epsilon_n | Y_{1:n} \right) = o_p(1)$$

$$\pi \left(\|\gamma_1 - \gamma^0\| + \|\gamma_2 - \gamma^0\| \geq M_n n^{-1/4} | Y_{1:n} \right) = o_p(1)$$

Merging : same rate as static mixture

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of $\mathbf{p}$

Why does it work like that ?

- **Integration leads to penalisation : case emptying extra states**

$$P^{\pi}\left(B_n|\mathbf{x}^n\right) := \frac{N_n}{D_n} = \frac{\int_{B_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi(\theta)}{\int_{\Theta} e^{l_n(\theta)-l_n(\theta_0)} d\pi(\theta)}$$

$$D_n \gtrsim n^{-D/2}, \text{with proba. } \to 1$$

On $\Omega_n = \{\|f_{2,\theta} - f_2^*\|_1 \leq u_n\} \cap \{\rho_Q - 1 > v_n\}\ \forall S_n \subset \Omega_n$

$$\pi(S_n^c) = o(n^{-D/2}) \quad \Rightarrow \pi(S_n|\mathbf{x}^n) = 1 + o_p(1)$$

posterior concentrates on sparsest path (in terms of π)

- **Case of** $\bar{\alpha} >> \underline{\alpha}$ (emptying) $D = k_0(k_0 - 1 + d) + \underline{\alpha}k(k - k_0)$ ($\mathbf{p}Q = \mathbf{p}$)

$$S_n^c = \{\exists j > k_0; \mathbf{p}(j) > e_n\}, e_n = o(1)$$

A useful inequality : Gassiat , van Handel

$$\|f_{2,\theta} - f_{2,\theta^*}\|_1 \gtrsim \sum_{j:\|\gamma_j - \Gamma^*\|_1 > \epsilon} p(j)$$

$$+ \sum_{i_1,i_2} \Big| \sum_{j_1 \in A(i_1), j_2 \in (Ai_2)} p(j_1) q_{j_1,j_2} - p^*(i_1) q^*_{i_1,i_2} \Big|$$

$$+ \sum_{i_1,i_2} \left| \sum_{j_1 \in A(i_1), j_2 \in A(i_2)} p(j_1) q_{j_1,j_2} \begin{pmatrix} \gamma_{j_1} \\ \gamma_{j_2} \end{pmatrix} - p^*(i_1) q^*_{i_1,i_2} \begin{pmatrix} \gamma^*_{i_1} \\ \gamma^*_{i_2} \end{pmatrix} \right.$$

$$+ \sum_{i_1,i_2} \sum_{j_1 \in A(i_1), j_2 \in (Ai_2)} p(j_1) q_{j_1,j_2} \left(\|\gamma_{j_1} - \gamma^*_{i_1}\|^2 + \|\gamma_{j_2} - \gamma^*_{i_2}\|^2 \right)$$

Repulsive priors

$$\mathbf{q}_{i.} \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k), \quad \pi_k(\gamma_1, \cdots, \gamma_k) \propto \prod_{j=1}^{k} \pi_\gamma(\gamma_j) \times e^{-\frac{c}{\min |\gamma_i - \gamma_j|}}$$

or

$$\pi_k(\gamma_1, \cdots, \gamma_k) \propto \prod_{j=1}^{k} \pi_\gamma(\gamma_j) \times (\min |\gamma_i - \gamma_j|)^C$$

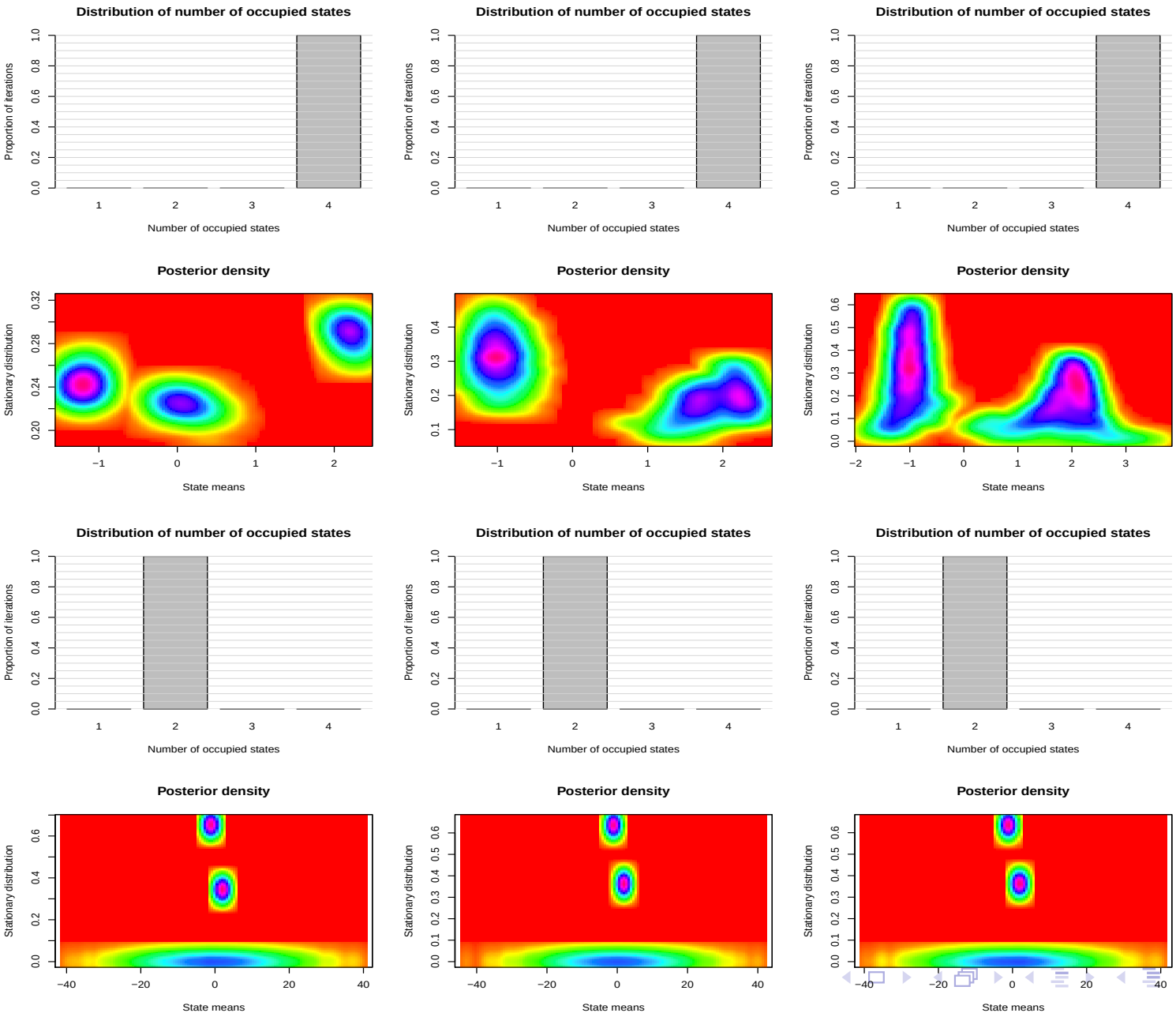
- **case 1** : If $\sum_i \alpha_i > k(k-1+d)$ & $\forall c > 0$ emptying the extra states
- **case 2** : If $\sum_i \alpha_i > k(k-1+d)$ & $C > C_0(k)$ emptying the extra states

Simulation study 1

- **Emission distribution** $k = 5$; $g_\gamma = \mathcal{N}(\gamma, 1)$, $(d = 1)$
- Prior $q_{i.} \sim \mathcal{D}(\bar{\alpha}, \underline{\alpha}, \cdots, \underline{\alpha})$
- $\gamma_j \overset{i.i.d}{\sim} \mathcal{N}(\bar{y}, 100)$.
- **True model** $k_0 = 2$,

$$Q^* = \begin{pmatrix} 0.6 & 0.4 \\ 0.7 & 0.3 \end{pmatrix}, \quad \gamma^* = (-1, 3)$$

Top : $\bar{\alpha} = \underline{\alpha} = 172, 4, 1$; bottom : $\bar{\alpha} = 172, 4, 1; \underline{\alpha} = 1/n$



Questions

- **Suboptimal Constraints** in particular $\bar{\alpha}$? ? ? : very informative prior
- **Diagonal priors ?**

$$\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_2 \\ \alpha_2 & \alpha_1 & \cdots & \alpha_2 \\ \cdot & \cdots & \cdots & \cdots \\ \alpha_2 & \alpha_2 & \cdots & \alpha_1 \end{pmatrix}, \quad \alpha_1 >> 1, \quad \alpha_2 << 1$$

Diagonal prior : not to be trusted ?

Sim 2 $\gamma^*_{S2} = (-5, 5, 9)$, $\mathbf{p}^*_{S2} = (0.56, 0.18, 0.26)$, and

$$Q^*_{S2} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix},$$

Diagonal prior : not to be trusted ?

Sim 2 $\gamma^*_{S2} = (-5, 5, 9)$, $\mathbf{p}^*_{S2} = (0.56, 0.18, 0.26)$, and

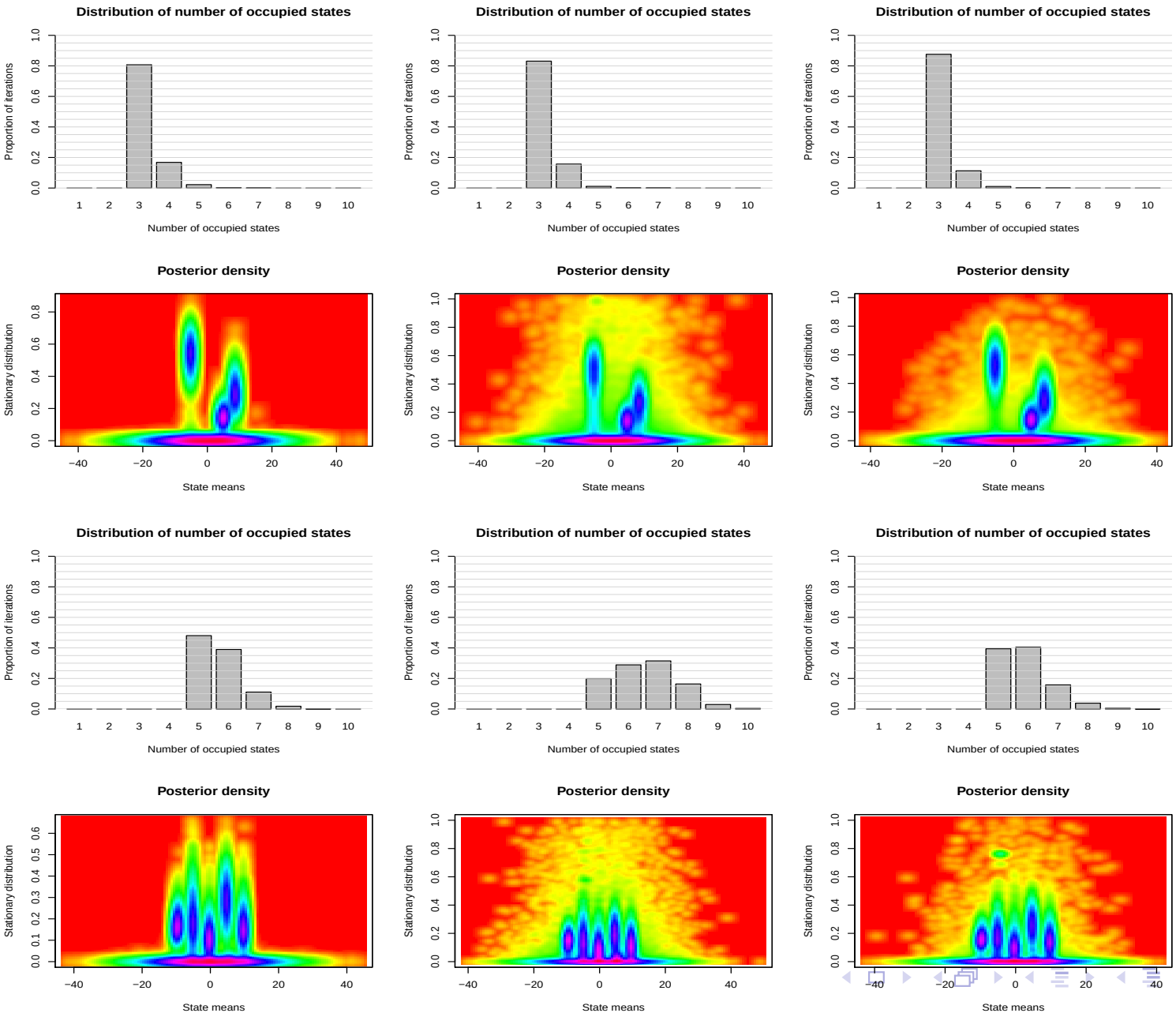
$$Q^*_{S2} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.4 & 0.4 \\ 0.3 & 0.2 & 0.5 \end{bmatrix},$$

Sim 3 $\gamma^*_{S3} = (-10, -5, 0, 5, 10)$,

$\mathbf{p}^*_{S3} = (0.11, 0.24, 0.20, 0.22, 0.22)$,

$$\text{and } Q^*_{S3} = \begin{bmatrix} 0.2 & 0.3 & 0.1 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.6 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.6 \end{bmatrix}.$$

n=100, Column, diagonal, mixture, $\bar{\alpha} = 1$ $\underline{\alpha} = 1/n$



Robustness issues

• Results are often non robust to mis-specification of emission distribution

- **Non parametric models**
  - 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.

Robustness issues

• Results are often non robust to mis-specification of emission distribution

- **Non parametric models**
  - 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.

Robustness issues

• Results are often non robust to mis-specification of emission distribution

- **Non parametric models**
  - 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.

Robustness issues

• Results are often non robust to mis-specification of emission distribution

- **Non parametric models**
  - 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.

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 $(F_j)_j$

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

Outline

1. Parametric mixture models : static and dynamic
   - Models
   - Known results
2. Posterior concentration to marginals
   - results
   - why ?
3. Semi parametric mixture models
   - Various results on estimation
4. Case of static mixture : semiparametric estimation of $\mathbf{p}$

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 **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 $1/\sqrt{n}$?

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 ?

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 $\mathcal{I}$:

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

- 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 $f_j(x) = \sum_\ell \frac{w_{j,\ell}}{|I_\ell|} \mathbb{I}_{x \in I_l}$
$J_{\mathcal{I}} := J_{\mathcal{I}}(\mathbf{p}^*, \mathbf{f}^*)$ = 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 : $Y_i = (Y_{i,1}, Y_{i,2}, Y_{i,3})$

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

- 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

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

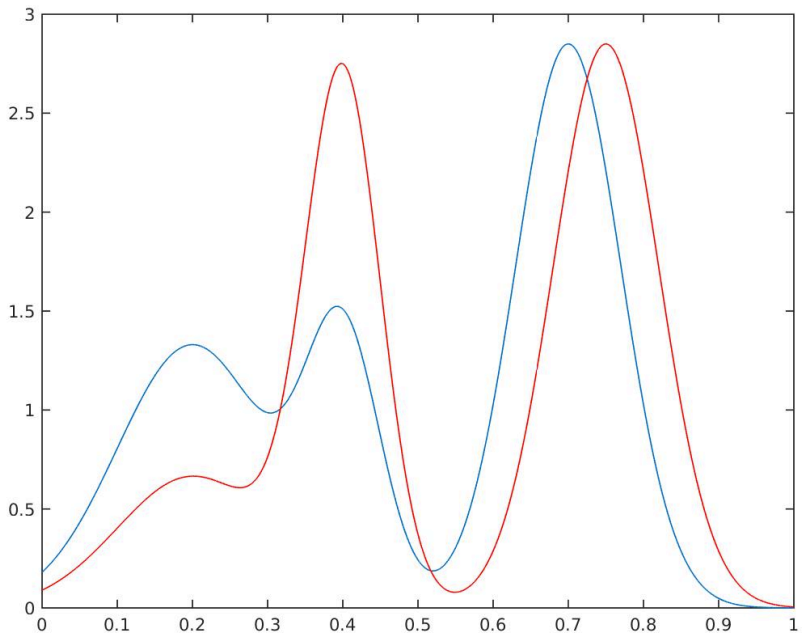
Some simulation results : K=2

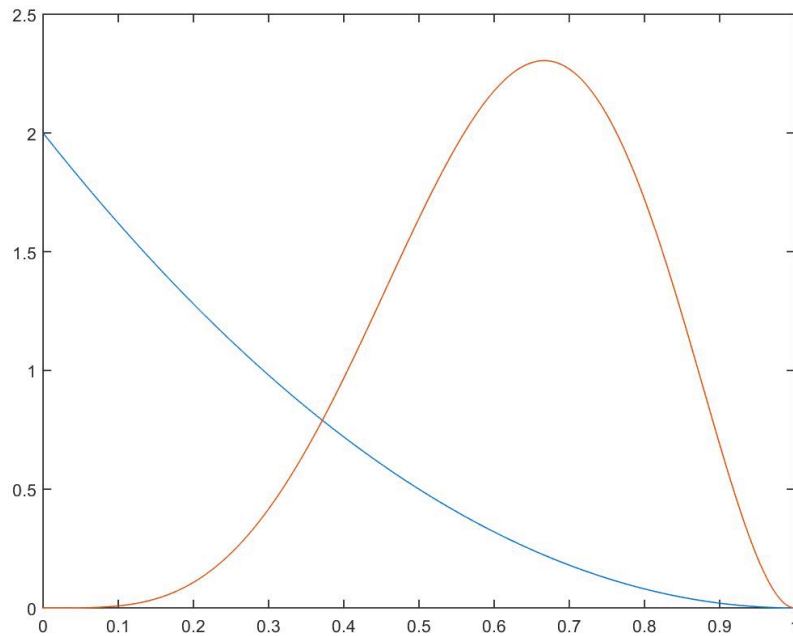
- **Data 1 :** $p = 0.3$ (difficult)

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

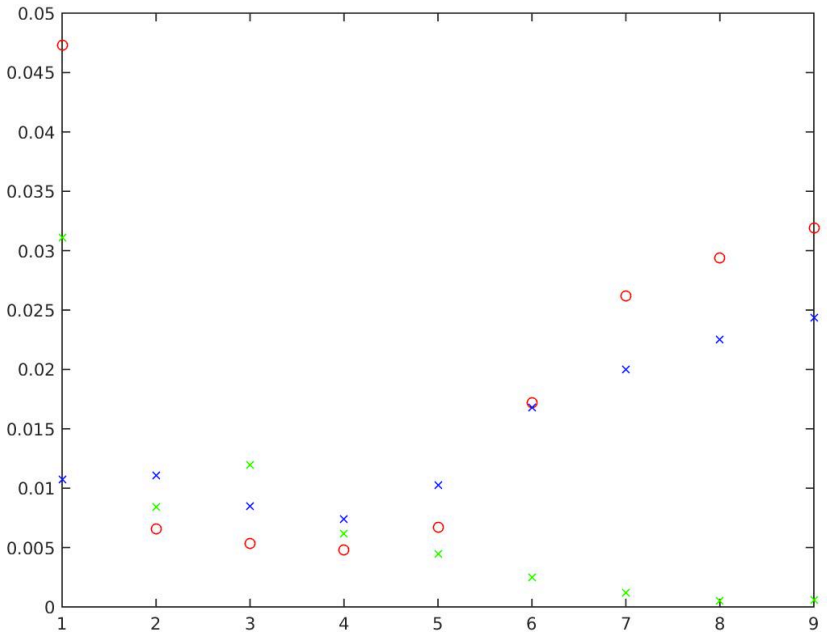
$$f_2 = \frac{1}{3} * \mathcal{N}(0.2, 0.01)\mathbb{1}_{|\cdot|\leq 1} + \frac{1}{2} * \mathcal{N}(0.77, 0.07^2)\mathbb{1}_{|\cdot|\leq 1} + \frac{1}{6}\mathcal{N}(0.4, 0.05)\mathbb{1}_{|\cdot|\leq 1}$$

- **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



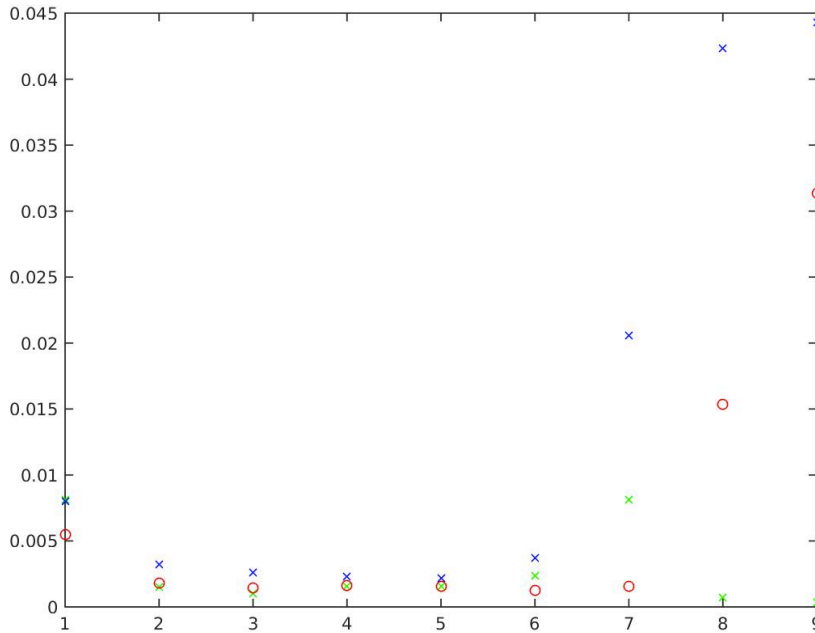
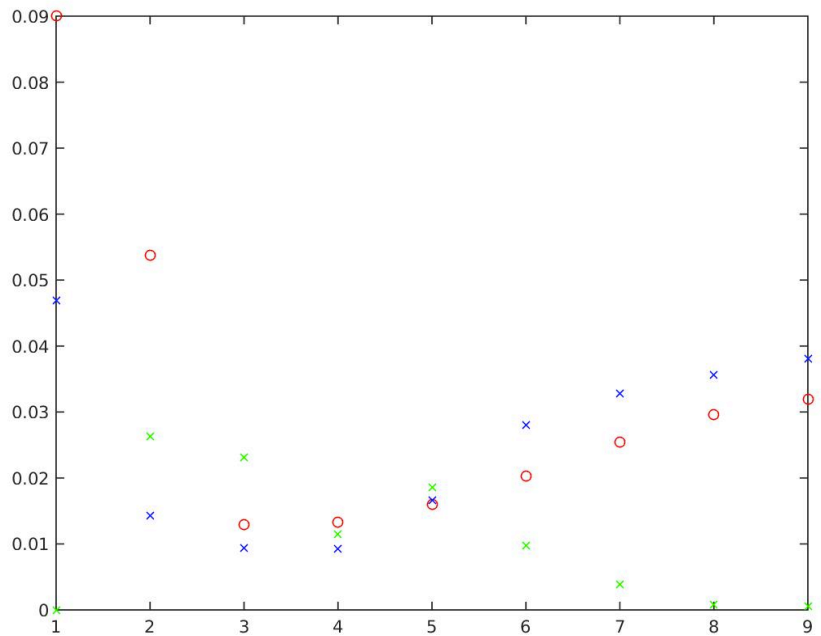


FIG.: Data 2, n=100 (left), n= 500 (right)

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



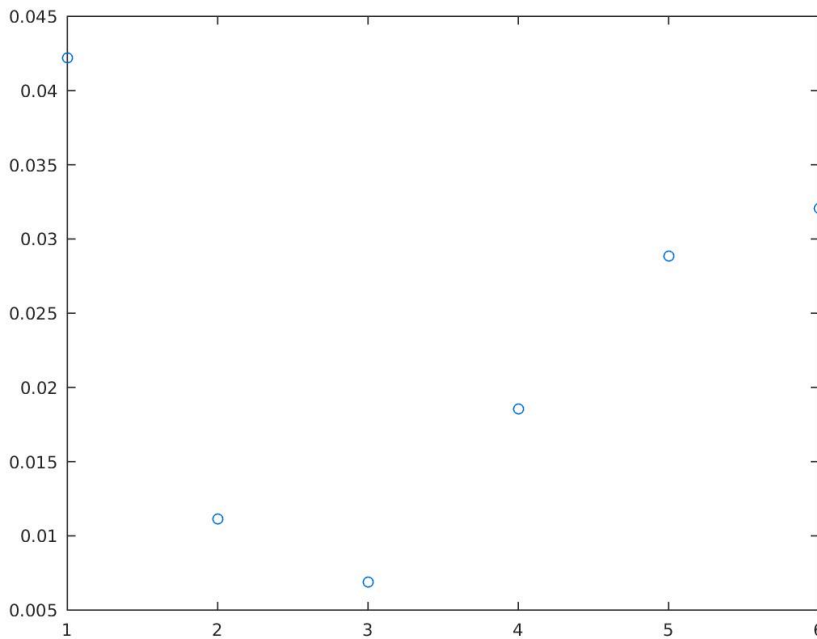


FIG.: Data 1, n=100 : left = pfixed partition, right = empirical

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

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

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

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

- **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)

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

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

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

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

- **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"**

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

$$n_\epsilon^{(j)} = \sum \mathbb{1}_{y_{i,j} \in B_\epsilon} \mathbb{1}_{x_i = j}$$

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

Some simulations : conditional approach

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

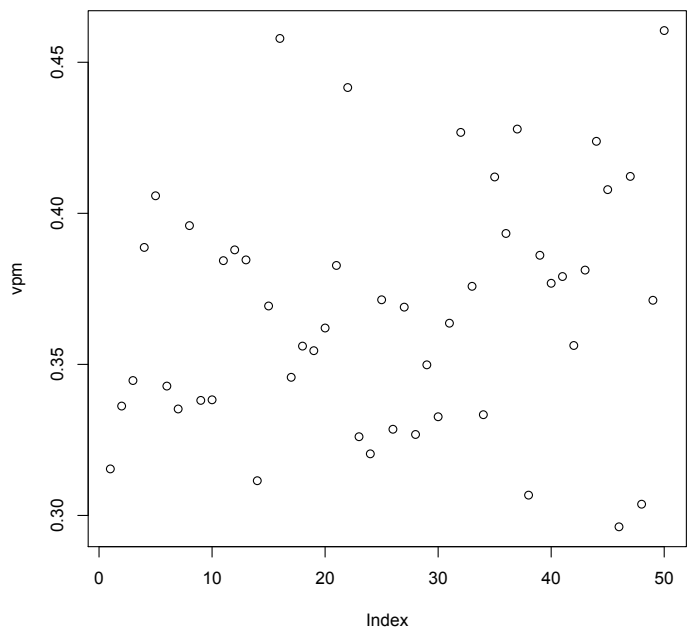


FIG.: n=100, 50 replicates, mean = 0.367975

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

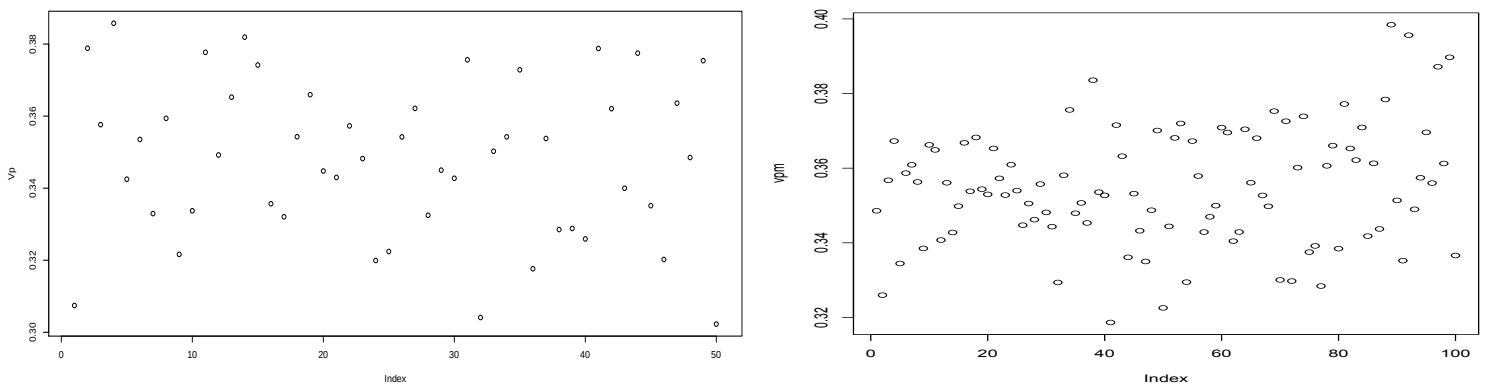


FIG.: left : n=500, 50 replicates, $\hat{p}$ = 0.348, right : n= 1000, 100 replicates , $\hat{p}$ = 355

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

- **Penalisation via priors helps in over-identifiable parametric models**
- **HMM models** The picture not complete yet : both parametric and nonparametric
  - Can we get rid of $\rho_Q - 1$ in concentration rate ? (parametric)
  - Can we control the likelihood in regions with $\rho_Q - 1 \approx 0$?
- **These results also lead to BF consistency** In practice is it best to use BF ?
  - allows to reduce the set of candidate k's.
- **NP mixtures**
  - efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j

Conclusion

- **Penalisation via priors helps in over-identifiable parametric models**
- **HMM models** The picture not complete yet : both parametric and nonparametric
  - Can we get rid of $\rho_Q - 1$ in concentration rate ? (parametric)
  - Can we control the likelihood in regions with $\rho_Q - 1 \approx 0$?
- **These results also lead to BF consistency** In practice is it best to use BF ?
  - allows to reduce the set of candidate k's.
- **NP mixtures**
  - efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
  - Shall we mix $\pi(p|y^n, \mathcal{I})$ with NP $\pi(f_1, \cdots, f_K|p, y^n)$?

Conclusion

- **Penalisation via priors helps in over-identifiable parametric models**
- **HMM models** The picture not complete yet : both parametric and nonparametric
  - Can we get rid of $\rho_Q - 1$ in concentration rate ? (parametric)
  - Can we control the likelihood in regions with $\rho_Q - 1 \approx 0$?
- **These results also lead to BF consistency** In practice is it best to use BF ?
  - allows to reduce the set of candidate k's.
- **NP mixtures**
  - efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
  - Shall we mix $\pi(p|y^n, \mathcal{I})$ with NP $\pi(f_1, \cdots, f_K|p, y^n)$?
  - Semi - parametric problems : targeted likelihood.

Conclusion

- **Penalisation via priors helps in over-identifiable parametric models**
- **HMM models** The picture not complete yet : both parametric and nonparametric
  - Can we get rid of $\rho_Q - 1$ in concentration rate ? (parametric)
  - Can we control the likelihood in regions with $\rho_Q - 1 \approx 0$?
- **These results also lead to BF consistency** In practice is it best to use BF ?
  - allows to reduce the set of candidate k's.
- **NP mixtures**
  - efficient estimation of the weights – no biais despite misspecified for f_j : but stupid model for f_j
  - Shall we mix $\pi(p|y^n, \mathcal{I})$ with NP $\pi(f_1, \cdots, f_K|p, y^n)$?
  - Semi - parametric problems : targeted likelihood.
  - Shall we change likelihood for different parameter of interests

Thank you

Conclusion