Stochastic Expectation-Maximization Methods for Sequential Inference in Missing Data Models

Florian Maire*, UCD

joint work with : Eric Moulines (Telecom Paris, Paris, France)
Sidonie Lefebvre (ONERA, Palaiseau, France)

*started at ONERA & Telecom Paris

Working Group on Statistical Learning, 26th of February 2014

Outlines

Outlines

Notations & main question

- **Observed data**:

$$\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{P}^\star$$

- $\mathbb{P}^\star$: **unknown** (hypothetical) probability distribution on $(\mathsf{Y}, \mathcal{Y})$
- An **observation model** on $(\mathsf{Y}, \mathcal{Y})$

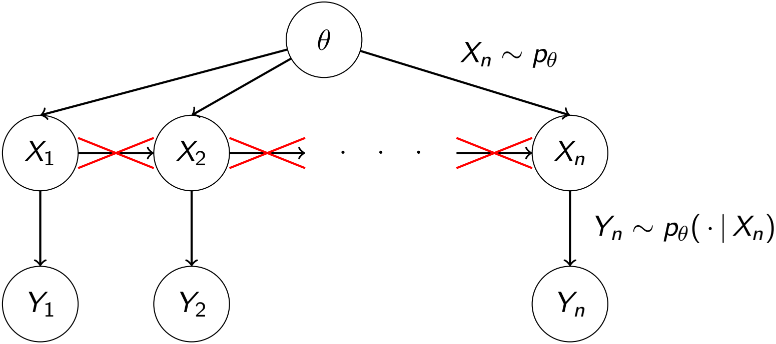
$$\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{P}_\theta$$

- $\mathbb{P}_\theta$ parameterized by $\theta \in \Theta$.

Question: how to find $\theta^\star \in \Theta$ s.t. $\mathbb{P}_{\theta^\star} \approx \mathbb{P}^\star$

Missing Data Models

- **Latent process**: $\{X_n,\ n \in \mathbb{N}\}$ on $(\mathsf{X}, \mathcal{X})$



- $\mathbb{P}_\theta$: an **intractable** marginal of the complete data model

$$\forall\, A \in \mathcal{Y}, \quad \mathbb{P}_\theta[Y \in A] = \int_A \int p_\theta(y \mid x) p_\theta(x) \mathrm{d}x \mathrm{d}y \ .$$

Estimating θ?

- Maximum Likelihood Estimator (MLE):

$$\theta^{\mathsf{MLE}} = \arg\max_{\theta \in \Theta} \mathbb{P}[\theta \mid Y_1, \dots, Y_n] = \arg\max_{\theta \in \Theta} \prod_{k=1}^{n} \int p_\theta(Y_k \mid X_k) p_\theta(X_k) \mathrm{d}X_k \,.$$

$\Longrightarrow$ Direct optimization methods cannot be used to reach θ^{MLE}

- Assume an Exponential Model *i.e* :

$$\log p_\theta(x, y) = \psi(\theta) + \langle S(x, y), \phi(\theta) \rangle \;,$$

  - $S : \mathsf{X} \times \mathsf{Y} \to \mathsf{S}$, vector of sufficient statistics
  - $\psi : \Theta \to \mathbb{R}$, $\phi : \Theta \to \mathsf{S}$, differentiable functions
  - $\langle \cdot, \cdot \rangle$: scalar product on S

Expectation-Maximization (Dempster et al., 1977)

- Assume $\{Y_1, \dots, Y_n\} \in \mathsf{Y}^n$ constantly available
- For all $(\theta, \theta') \in \Theta^2$

$$\mathbb{E}_{\theta}\left[\sum_{k=1}^n S(X_k, Y_k) \mid Y_k\right] \leq \mathbb{E}_{\theta'}\left[\sum_{k=1}^n S(X_k, Y_k) \mid Y_k\right]$$
$$\Downarrow$$
$$\mathbb{P}[\,\theta \mid Y_1, \dots, Y_n] \leq \mathbb{P}[\,\theta' \mid Y_1, \dots, Y_n]$$

- EM: starting from some $\theta_0 \in \Theta$, $\{\theta_k,\ k \in \mathbb{N}^*\}$

$$(i) \quad s_i = \bar{s}(\theta_{i-1}) = \sum_{k=1}^n \mathbb{E}_{\theta_{i-1}}\left[S(X_k, Y_k) \mid Y_k\right]$$

$$(ii) \quad \theta_i = \bar{\theta}(s_i) = \arg\max_{\theta \in \Theta} \psi(\theta) + \langle s_i, \phi(\theta) \rangle$$

- Convergence

$$\theta_i \to \{\theta \in \Theta,\ \nabla_\theta \mathbb{P}[\,\theta \mid Y_1, \dots, Y_n] = 0\}\,.$$

Stochastic Approximation EM (Delyon et al., 1999)

- In many situations $s_i = \mathbb{E}_{\theta_{i-1}}[S(X,Y)\,|\,Y]$ is intractable
- SAEM: given $\hat{\theta}_0 \in \Theta$, a stochastic sequence $\{\hat{\theta}_i,\ i \in \mathbb{N}^*\}$

$$(i) \quad \hat{s}_i = \hat{s}_{i-1} + \rho_i \left(\underbrace{\frac{1}{L}\sum_{k=1}^{n}\sum_{\ell=1}^{L} S(X_k^{(\ell)}, Y_k)}_{\substack{\downarrow\, L\to\infty \\ \sum_{k=1}^n \mathbb{E}_{\hat{\theta}_{i-1}}[S(X_k,Y_k)\,|\,Y_k]}} - \hat{s}_{i-1} \right), \quad X_k^{(\ell)} \sim p_{\hat{\theta}_{i-1}}(\,\cdot\,|\,Y_k)\,,$$

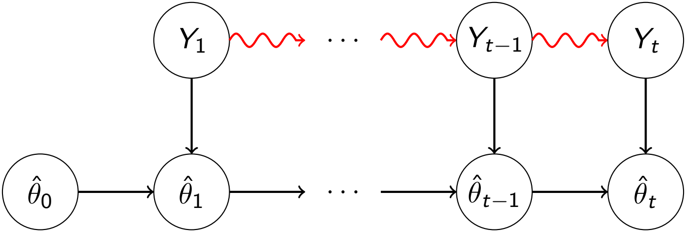
$$(ii) \quad \hat{\theta}_i = \bar{\theta}(\hat{s}_i)$$

- Convergence: under mild assumptions on $\{\rho_i,\ i \in \mathbb{N}\}$

$$\hat{\theta}_i \to \{\theta \in \Theta,\ \nabla_\theta \mathbb{P}[\theta\,|\,Y_1,\ldots,Y_n] = 0\}\,.$$

Sequential Inference Framework

- *Sequential* (or *Online*) as opposed to *Batch* (or *Bloc*) methods
- Only one observation at a time $\{Y_t,\ t \in \mathbb{N}\}$



- the t-th iteration will happen when Y_t becomes available...
- ... and will produce a new estimate $\hat{\theta}_t$ *"better"* than $\hat{\theta}_{t-1}$

Motivations behind Sequential Inference?

- Storage issue (especially for big data s.t. images...)
- Computational aspect:
  - in batch setting: compute (or estimate) $\underline{n}$ conditional expectations
- Complexity of the method independent of *n*!
- First EM iteration will use <u>all</u> the data while $\hat{\theta}_0$ is a random guess!
- Tracking any trend on the data...

Online EM (Cappé & Moulines, 2008)

- Online EM: given $\hat{\theta}_0 \in \Theta$, a stochastic sequence $\{\hat{\theta}_t,\ t \in \mathbb{N}^*\}$

$$(i) \quad \hat{s}_t = \hat{s}_{t-1} + \rho_t \left(\mathbb{E}_{\hat{\theta}_{t-1}} \left[S(X_t, Y_t) \mid Y_t \right] - \hat{s}_{t-1} \right) ,$$

$$(ii) \quad \hat{\theta}_t = \bar{\theta}(\hat{s}_t) .$$

- MLE non-sense in sequential context: other dissimilarity measure between $\mathbb{P}^\star$ and $\mathbb{P}_\theta$

$$\mathrm{KL}\left(\mathbb{P}^\star \,\|\, \mathbb{P}_\theta\right) = \mathbb{E}^\star \left[\frac{\mathbb{P}^\star(Y)}{\mathbb{P}_\theta(Y)} \right]$$

- Convergence of Online EM (under *"reasonable"* conditions)

$$\hat{\theta}_t \to \{\theta \in \Theta,\ \nabla_\theta \mathrm{KL}\left(\mathbb{P}^\star \,\|\, \mathbb{P}_\theta\right) = 0\}$$

More questions...

- Online EM allows sequential inference in missing data models...
- ... but what if $\mathbb{E}_{\hat{\theta}_{t-1}}[S(X_t, Y_t) \mid Y_t]$ is intractable?
- Any possible extension of the Online EM?
- Theoretical justification behind it?

Outlines

Monte Carlo Online EM (MCoEM)

- Monte Carlo EM: given $\tilde{\theta}_0 \in \Theta$, a stochastic sequence $\{\tilde{\theta}_t,\ t \in \mathbb{N}^*\}$

$$(i) \quad \tilde{s}_t = \tilde{s}_{t-1} + \rho_t \left(\underbrace{\frac{1}{L}\sum_{\ell=1}^{L} S(X_t^{(\ell)}, Y_t)}_{\displaystyle \Big\downarrow\, L \to \infty \atop \displaystyle \mathbb{E}_{\tilde{\theta}_{t-1}}[S(X_t, Y_t) \mid Y_t]} - \tilde{s}_{t-1} \right), \quad X_t^{(\ell)} \sim p_{\tilde{\theta}_{t-1}}(\,\cdot \mid Y_t)\,,$$

$$(ii) \quad \tilde{\theta}_t = \bar{\theta}(\tilde{s}_t)\,.$$

- MCoEM: equivalent of the SAEM for the sequential settings.

Illustration on a Mixture of Gaussian regression

- Let $\begin{cases} I \in \{1,2\} & \text{mixture index} \\ \beta \in \mathbb{R} & \text{auxiliary variable} \\ Y \in \mathbb{R} & \text{observation} \end{cases}$
- At time t, we simulate
  (i) $I_t = i \sim \omega_i$
  (ii) $\beta_t \mid I_t = i \sim \mathcal{N}(\mu_i, \gamma^2)$
  (iii) $Y_t \mid I_t = i, \beta_t \sim \mathcal{N}(\Phi_{\beta_t}\alpha_i, \sigma_i^2)$
  with $\Phi_\beta = (1, \beta, \beta^2/10)$ and $\{\alpha_i \in \mathbb{R}^3\}_{i=1}^2$
- Sufficient Statistics $S(i, \beta, Y) = \begin{pmatrix} \delta_{i=1}\left(\Phi_\beta^T Y, \Phi_\beta^T \Phi_\beta\right)^T \\ \delta_{i=2}\left(\Phi_\beta^T Y, \Phi_\beta^T \Phi_\beta\right)^T \end{pmatrix}$
- Two learning setups
  LS-1 Observations: (Y_n, β_n) - Missing data: I_n
  LS-2 Observations: Y_n - Missing data: (β_n, I_n)

Comparison Online EM / MCoEM (LS-1)

- Online EM

$$\hat{s}_t = \hat{s}_{t-1} + \rho_t \left(\mathbb{E}_{\hat{\theta}_{t-1}} [S(I, \beta, Y) \mid Y_t, \beta_t] - \hat{s}_{t-1} \right) ,$$

$$\mathbb{E}_{\hat{\theta}_{t-1}} [S_j(I, \beta, Y) \mid Y_t, \beta_t] \propto \exp \left\{ -(1/2) \frac{(Y_t - \Phi_{\beta_t} \hat{\alpha}_{j,t-1})^2}{\sigma_j^2} \right\} S^*(Y_t, \beta_t)$$

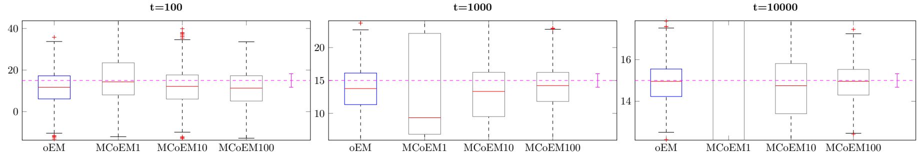
- Monte Carlo Online EM

$$\tilde{s}_t = \tilde{s}_{t-1} + \rho_t \left(\frac{1}{L} \sum_{\ell=1}^{L} S(I_t^{(\ell)}, \beta_t, Y_t) - \tilde{s}_{t-1} \right) , \quad I_t^{(\ell)} \overset{i.i.d.}{\sim} \mathbb{P}_{\tilde{\theta}_{t-1}} [\cdot \mid \beta_t, Y_t] ,$$

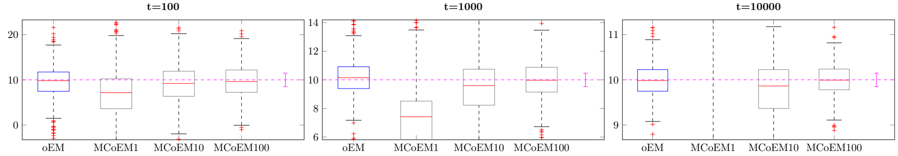
Comparison Online EM / MCoEM (LS-1)

500 runs of 10000 iterations of the two methods $\begin{cases} \text{Online EM} \\ \text{MC Online EM with } L = 1,\ 10\,\&\,100 \end{cases}$

- Estimation of $\alpha_{1,1}$



- Estimation of $\alpha_{2,1}$



MCoEM (LS-2)

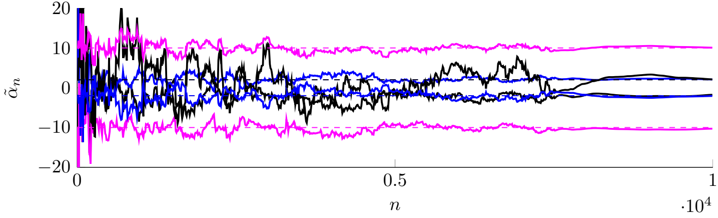
- In this case, the Online EM cannot be implemented...

$$e.g \quad : \qquad \mathbb{E}_{\hat{\theta}_{t-1}}\left[\mathbb{1}_{\{j\}}(I)\Phi_\beta^T\Phi_\beta \mid Y_t\right] = \int \Phi_\beta^T\Phi_\beta p_{\hat{\theta}_{t-1}}(j, \mathrm{d}\beta \mid Y_t)$$

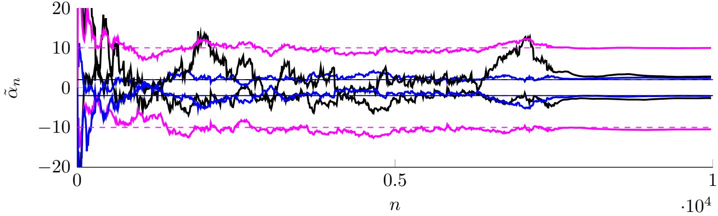
- ... but MCoEM might still be!
  - $(I_t^{(\ell)}, \beta_t^{(\ell)}) \overset{i.i.d.}{\not\sim} p_{\hat{\theta}_{t-1}}(\cdot \mid Y_t)$
  - $(I_t^{(\ell)}, \beta_t^{(\ell)}) \sim K_{\hat{\theta}_{t-1}}(I_t^{(\ell-1)}, \beta_t^{(\ell-1)}; \cdot \mid Y_t)$
- K: Markov transition kernel on $\mathsf{X} \times \mathcal{X}$:
  - Metropolis-within-Gibbs,
  - Carlin & Chib,
  - ...

Sampling path

- Online EM (LS-1)

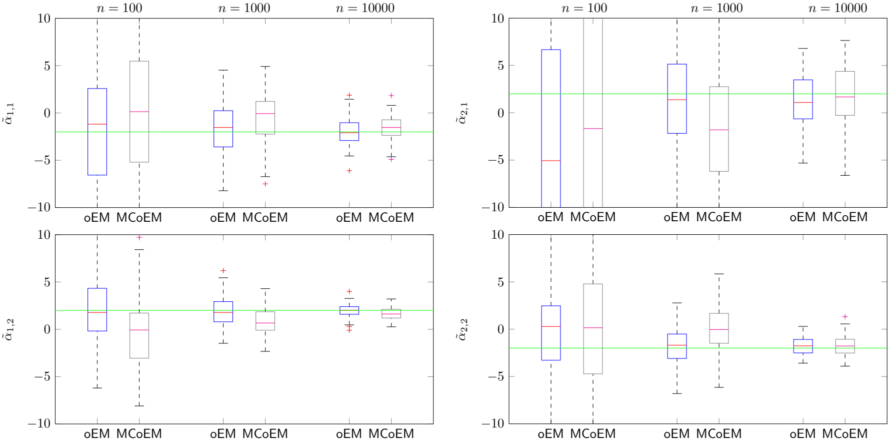


- MC Online EM + Carlin & Chib (LS-2)



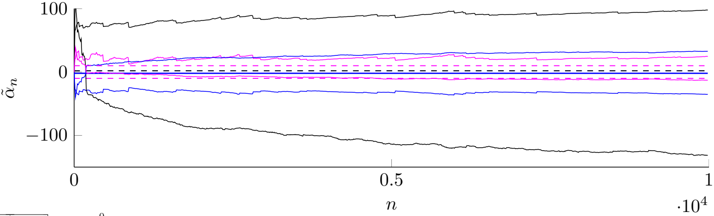
oEM (LS-1) vs MCoEM + Carlin & Chib (LS-2)

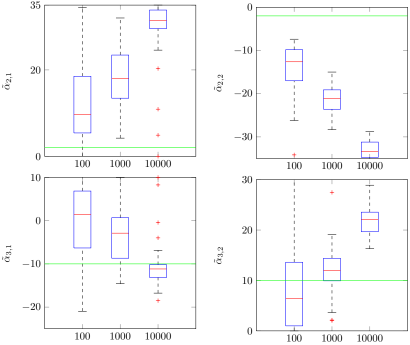
- $L = 500$ transitions of the Markov chains (incl. 100 burning iterations)
- Laplace approximations for the pseudo-prior (Carlin & Chib sampler)



MC Online EM + Gibbs (LS-2)

Sample path:





Variance for 4 parameters

$\Rightarrow$ No convergence!!

$\Longrightarrow$ Any theoretical justification of the MCoEM convergence?

$\Longrightarrow$ MC online EM: a *noisy* Online EM...

$\Longrightarrow$ Does the noise added by the MCoEM to the Online EM sequence affect its convergence?

A general framework to prove convergence of Stochastic EM methods...

Revisiting EM methods

-1- Define the distribution of the complete data *viewed by the algorithm*

$$\pi_\theta(X, Y) = p_\theta(X \mid Y)\pi(Y)$$

-2- Define

$$\bar{s}_\pi(\theta) = \mathbb{E}_{\pi_\theta}[S(X, Y)] = \int_{\mathsf{Y}} \mathbb{E}_\theta\left[\, S(X, Y) \mid Y = y \,\right] \pi(\mathrm{d}y)$$

-3- Any EM iteration (batch or online) is the deterministic mapping

$$\theta_{i+1} = \bar{\theta} \circ \bar{s}_\pi(\theta_i) \qquad \left(\text{or equivalently} \quad s_{i+1} = \bar{s}_\pi \circ \bar{\theta}(s_i)\right)$$

-4- EM methods search for the roots of the function

$$\begin{cases} \bar{h} : \mathsf{S} \to \mathsf{S} \\ \bar{h} = \bar{s}_\pi \circ \bar{\theta} - \mathrm{Id} \end{cases}$$

Indeed...

- In batch setting *i.e* $(Y_1, \ldots, Y_n)$

$$\pi(y) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{y_k\}}(y)$$

and

$$\bar{s}_\pi(\theta) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_\theta \left[S(X_k, Y_k) \mid Y_k \right]$$

$\Rightarrow$ Original EM (Dempster et al., 1977)

- In sequential setting

$$\pi(y) = \mathbb{P}^\star(\mathrm{d}y)$$

$$\bar{s}_\pi(\theta) = \int_{\mathsf{Y}} \mathbb{E}_\theta \left[S(X, Y) \mid Y = y \right] \mathbb{P}^\star(\mathrm{d}y)$$

$\Rightarrow$ The *exact* Online EM does not exist!!

Stochastic Approximations

- Allow finding the roots of a function $h : \mathsf{S} \to \mathsf{S}$ s.t.
  (i) h analytical expression is unknown
  (ii) noisy observations of h are available for any data points

$$\hat{h}(s) = h(s) + \zeta$$

- The method: recursively compute the stochastic sequence $\{\hat{s}_n,\ n \in \mathbb{N}\}$

$$\hat{s}_n = \hat{s}_{n-1} + \rho_n \hat{h}(\hat{s}_{n-1})$$

  $\{\rho_n,\ n \in \mathbb{N}\}$ is a decreasing sequence of positive stepsize
- Convergence of $\{\hat{s}_n,\ n \in \mathbb{N}\}$ to the set of roots (Andrieu et al., 2005)
  H-1 Mild conditions on $\{\rho_n,\ n \in \mathbb{N}\}$
  H-2 Existence of a Lyapounov function for h
  H-3 Condition on the noise process $\{\zeta_n,\ n \in \mathbb{N}\}$

$$\lim_{k\to\infty} \sup_{\ell>k} \left| \sum_{n=k}^{\ell} \rho_n \zeta_n \right| = 0$$

Link with Stochastic EM...

- All the Stochastic EM features different **noisy** observations of $\bar{s}_\pi \circ \bar{\theta}$
  - SAEM

$$\hat{s}_\pi \circ \bar{\theta}(s) = \frac{1}{nL} \sum_{k=1}^{n} \sum_{\ell=1}^{L} S(X_k^{(\ell)}, Y_k) \approx \underbrace{\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{\hat{\theta}_{i-1}} [S(X_k, Y_k) \mid Y_k]}_{\mathbb{E}_n \left[\mathbb{E}_{\hat{\theta}_{i-1}} [S(X_k, Y_k) \mid Y_k]\right]} = \bar{s}_\pi \circ \bar{\theta}(s)$$

  - Online EM

$$\hat{s}_\pi \circ \bar{\theta}(s) = \mathbb{E}_{\bar{\theta}(s)} [\, S(X_t, Y_t) \mid Y_t = y \,] \approx \underbrace{\int_{\mathsf{Y}} \mathbb{E}_\theta [\, S(X, Y) \mid Y = y \,] \pi(\mathrm{d}y)}_{\mathbb{E}^\star [\mathbb{E}_\theta [\, S(X,Y) \mid Y=y \,]]}$$

- Actually both of these algorithms approximate either
  - the expectation against the missing data measure
  - the expectation against the observed data measure

What about the MCoEM?

- Both expectations are actually intractable...
- The approximation is twofold

$$\tilde{s}_\pi \circ \bar{\theta}(s) = \frac{1}{L}\sum_{\ell=1}^{L}\left[\, S(X_t^{(\ell)}, Y_t) \mid Y_t = y \,\right] \approx \underbrace{\int_{\mathsf{Y}} \mathbb{E}_\theta\left[\, S(X, Y) \mid Y = y \,\right]\pi(\mathrm{d}y)}_{\mathbb{E}^\star[\mathbb{E}_\theta[\, S(X,Y) \mid Y=y\,]]}$$

- A doubly Stochastic approximation method?
- Convergence in the *i.i.d.* case follows the footstep of the proof of the Online EM

  $\Rightarrow$ Only the noise boundedness proof H-3 is different
- No proof at the moment when there is a markovian dependance between the simulated missing data...

Outlines

Functional Data Analysis with Deformable Template Model

- Data are discretized functions $\{\mathscr{Y}_n : \mathbb{R}^2 \to \mathbb{R},\ n \in \mathbb{N}\}$ t.q. :

$$\{\mathscr{Y}_n,\ n \in \mathbb{N}\} \quad \xRightarrow[\text{on a lattice}]{\text{discretization}} \quad \{Y_n,\ n \in \mathbb{N}\}$$

- $\forall n \in \mathbb{N}$, $\mathscr{Y}_n$ originates from the **deterministic** function (**template**)

$$\mathscr{T} : \mathbb{R}^2 \to \mathbb{R}$$

- and are observed through:
  (i) a random plan deformation $D_n : \mathbb{R}^2 \to \mathbb{R}^2$,
  (ii) an additive noise process $\mathscr{W}_n : \mathbb{R}^2 \to \mathbb{R}$

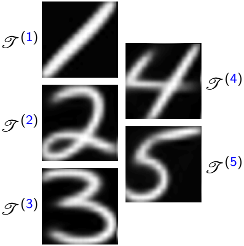
$$\forall u \in \mathbb{R}^2, \quad \mathscr{Y}_n(u) = \mathscr{T} \circ D_n(u) + \mathscr{W}_n(u)\ .$$

Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:



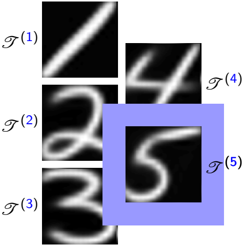
5 templates

Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\, i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)} \,.$$

- Illustration for a 5-class mixture model:



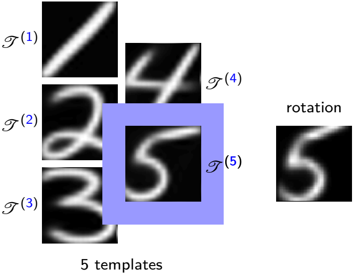
5 templates

Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)}, i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)} .$$

- Illustration for a 5-class mixture model:

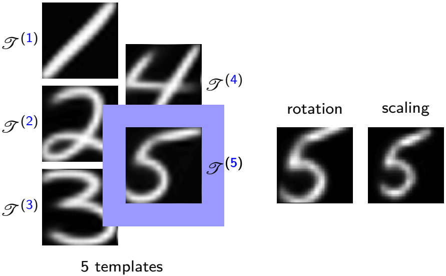


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:

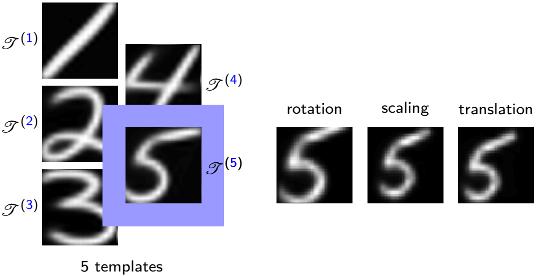


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\, i \in \mathsf{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:

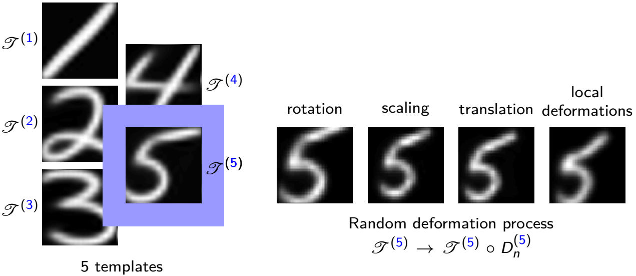


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:

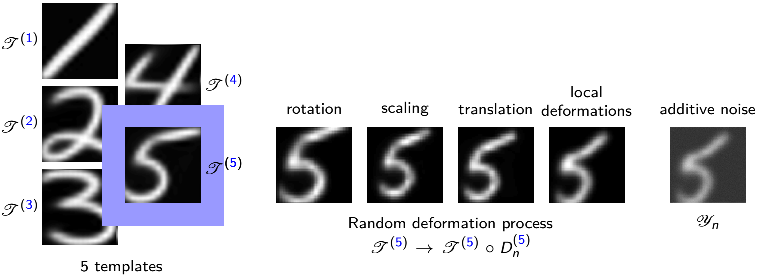


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathsf{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)} .$$

- Illustration for a 5-class mixture model:

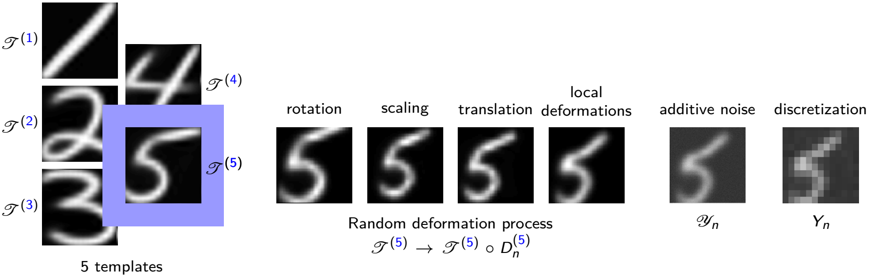


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:

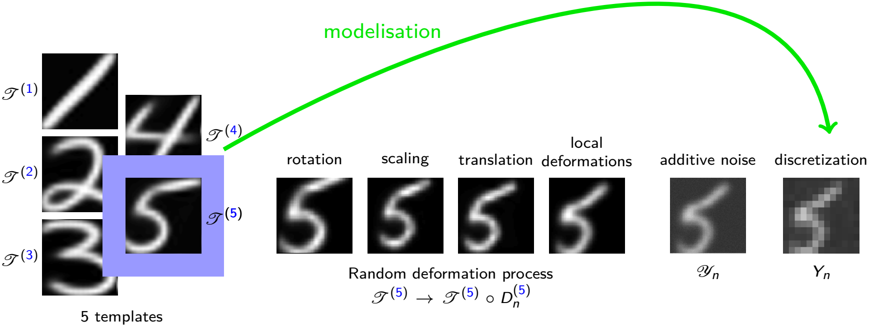


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathsf{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:

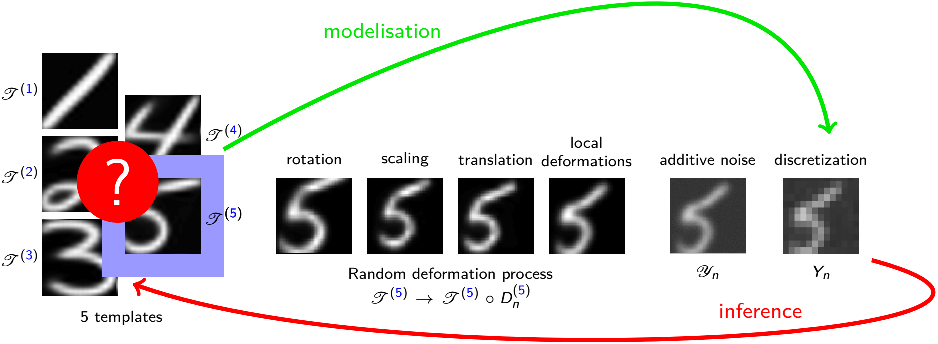


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)} \,.$$

- Illustration for a 5-class mixture model:

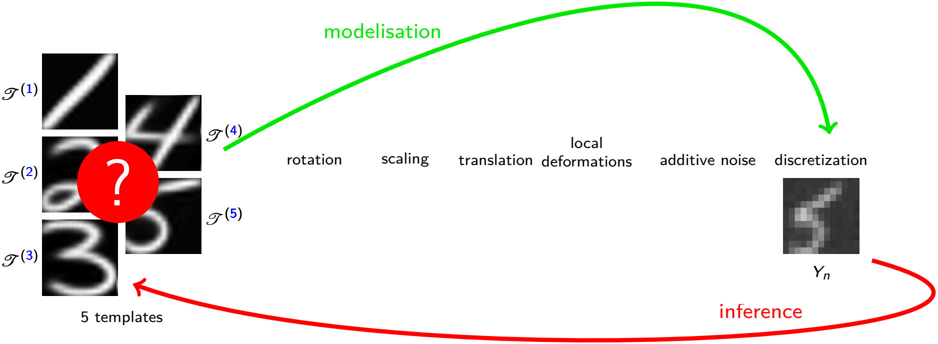


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)} \,.$$

- Illustration for a 5-class mixture model:

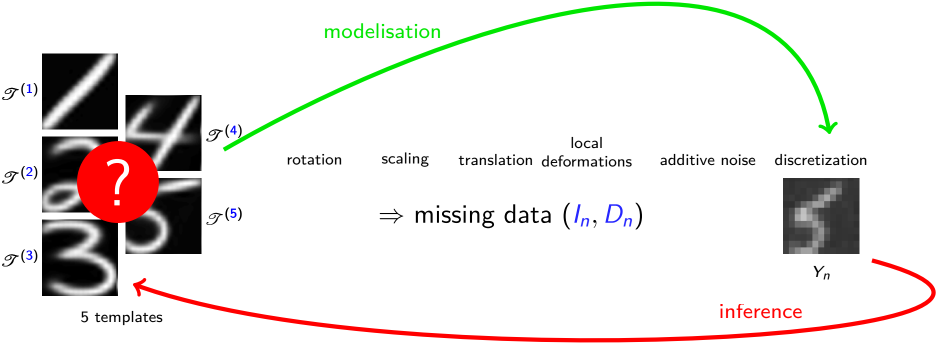


Toward a Mixture Model...

- Existence of **several** templates $\{\mathscr{T}^{(i)},\ i \in \mathrm{I}\}$

$$\textbf{given } I_n = i, \qquad \mathscr{Y}_n = \mathscr{T}^{(i)} \circ D_n^{(i)} + \mathscr{W}_n^{(i)}\ .$$

- Illustration for a 5-class mixture model:



Template estimation

- Under parametrization, the model may be rewritten as:

$$\text{Given, } I_n = i \quad Y_n = \Phi_{\beta_n}\alpha_i + \sigma W_n, \quad \begin{cases} W_n \sim \mathcal{N}_{|\mathrm{Y}|}(0, \mathrm{Id}_{|\mathrm{Y}|}) \\ \beta_n \mid I_n = i \sim \mathcal{N}_{|\mathrm{X}|}(0, \Gamma_i) \end{cases}$$

- $\beta \to \Phi_\beta$ is a non-linear mapping without any Gaussian approximation...
- Neither the Original EM nor the Online EM allow parameter estimation
- SAEM or MCoEM are possible solutions...

Computational (un)efficiency

- Recall that at each iteration the SAEM needs to estimate n cnd. expectations:
  $\Rightarrow$ Simulate n Markov chains $\{(\beta_n^{(\ell)}, I_n^{(\ell)}),\ \ell \in \mathbb{N}\}$ targeting $p_\theta(\,\cdot \mid Y_n)$
  $\Rightarrow$ dim $\beta \approx 100$ is prohibitive...
- Instead MC Online EM needs only one Markov chain per iteration...
- Simulations!

MCoEM has some drawbacks as well...

- Unrobustness to outliers
  - $\Rightarrow$ Especially in mixture models...(degenerescence)
  - $\Rightarrow$ A pre-processing step?
- Requires an *efficient* MCMC method
  - $\Rightarrow$ Such that Carlin & Chib (inducing extra computing cost...)
- SAEM less affected by these drawbacks
  - $\Rightarrow$ The batch set of data allows to balance mis-estimations...

Works in progress...

- Convergence proof of the MCoEM coupled with a MCMC...
  - (*i*) Metropolis-Hastings sampler
  - (*ii*) More sophisticated kernels (s.t. Carlin & Chib)
- Possible extension to state-space models...