Apprentissage statistique à l'ére du *Big Data* : stabilité et efficacité d'approximations de certains algorithmes

Florian Maire

Lundi 8 janvier 2018

Programme de cette présentation

Quelques informations sur mon parcours

Statistique computationelle : modélisation et apprentissage

Apprentissage statistique en présence d'un grand nombre de données : stabilité et efficacité d'approximations de certains algorithmes.

Quelques Illustrations

Plan

Quelques informations sur mon parcours

Statistique computationelle : modélisation et apprentissage

Apprentissage statistique en présence d'un grand nombre de données : stabilité et efficacité d'approximations de certains algorithmes.

Quelques Illustrations

Quelques dates

- 2005–2007, Classes prépa, Lycée Lavoisier, Paris
- 2007–2010, École d'ingénieur, Telecom Sudparis, Evry

- 2011–2014, Doctorat, Université Paris VI, Telecom Sudparis et ONERA

- 2014–2017, Post-doctorat, University College Dublin, Irlande

THESE DE DOCTORAT TELECOM SUDPARIS et UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Probabilités et Statistiques

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Florian MAIRE

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

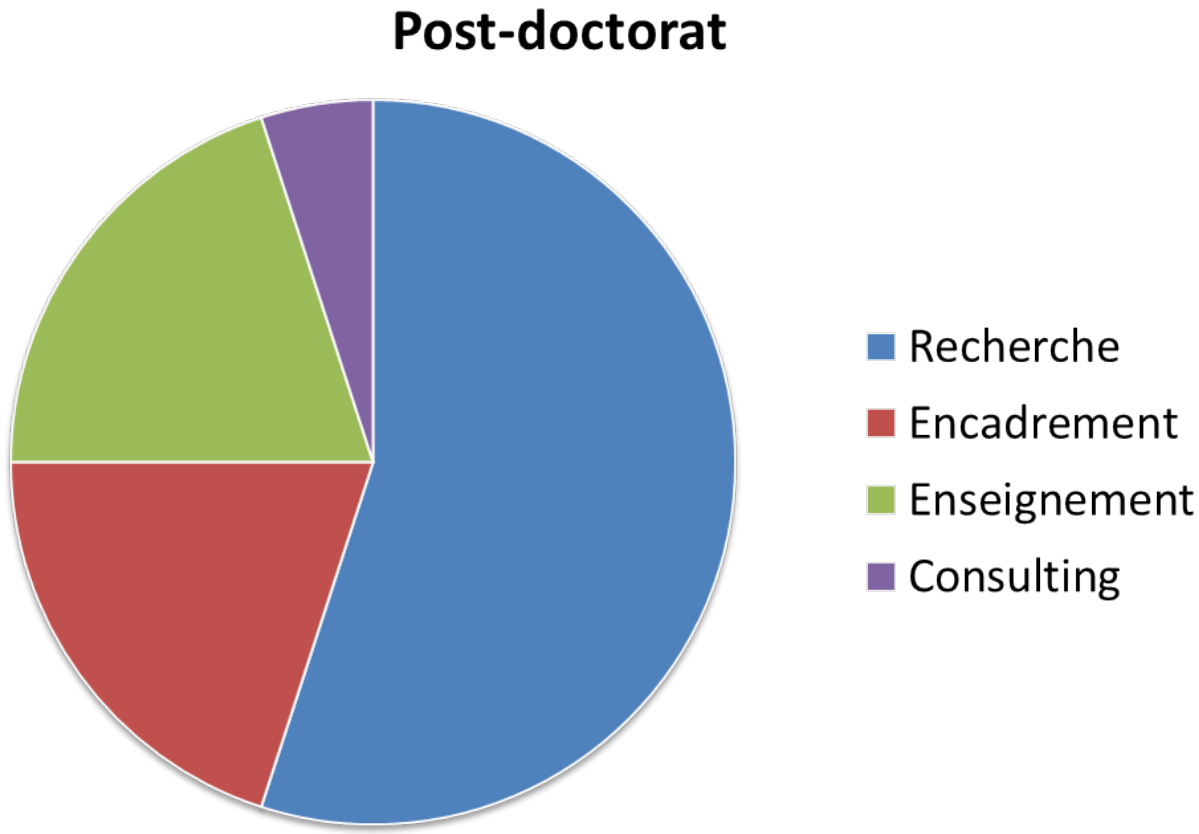
DETECTION ET CLASSIFICATION DE CIBLES MULTISPECTRALES DANS L'INFRAROUGE

dirigée par Randal DOUC, Professeur à Telecom SudParis,
co-dirigée par Eric MOULINES, Professeur à Telecom ParisTech,
encadrée par Sidonie LEFEBVRE, Ingénieur de Recherche à l'ONERA.

soutenue le : 14 Février 2014

devant le jury composé de :

M[me] Stéphanie ALLASSONNIÈRE	Ecole Polytechnique	examinateur
M. Jacques BLANC-TALON	DGA	examinateur
M. Randal DOUC	Telecom SudParis	directeur
M[me] Sidonie LEFEBVRE	ONERA	encadrant
M. Jean-Michel MARIN	Université de Montpellier 2	rapporteur
M. Eric MOULINES	Telecom ParisTech	co-directeur
M. Christian ROBERT	Université Paris-Dauphine	examinateur
M. Olivier WINTENBERGER	Université Paris 6	examinateur



- Recherche : Apprentissage en statistique Bayesienne (groupe de Prof. Nial Friel)
- Encadrement : PhD de Lampros Bourannis (avec N. Friel), 2 thèses de Masters et 3 étudiants de Masters
- Enseignement : Probability Theory, Time Series, Monte Carlo methods

Plan

Quelques informations sur mon parcours

Statistique computationelle : modélisation et apprentissage

Apprentissage statistique en présence d'un grand nombre de données : stabilité et efficacité d'approximations de certains algorithmes.

Quelques Illustrations

Notations

Les données $\{Y_n,\ n \in \mathbb{N}\}$ sont des observations d'un phénomènes aléatoires $\mathcal{Y}$ d'intérêt

$$\mathcal{Y} \rightsquigarrow Y_1, Y_2, \ldots$$

Exemples : signal, image, graphe

- les données sont généralement dans un espace euclidien de dimension d

$$Y_1 \in \mathsf{Y} \subseteq \mathbb{R}^d\,.$$

- Un modèle est une distribution de probabilité f sur l'espace mesurable $(\mathsf{Y}, \mathscr{Y})$
- $f \in \mathscr{F}$ où $\mathscr{F}$ est une famille de distributions de proba. sur $(\mathsf{Y}, \mathscr{Y})$

Apprentissage : trouver un "bon" modèle f^* appartenant à la famille $\mathscr{F}$

ayant observé $Y = Y_1, Y_2, \ldots, Y_N$

$$\Pr(Y \in A) \approx f^*(A) \qquad \forall\, A \in \mathscr{Y}$$

Statistique paramétrique

Hypothèse : $\mathscr{F}$ est une famille de distributions paramétrisée par un vecteur θ

$$\theta \in \Theta \subseteq \mathbb{R}^p \ (p < \infty)$$

$$\Rightarrow \forall f \in \mathscr{F},\ \exists \theta \in \Theta,\ f \equiv f(\theta)$$

$\Rightarrow$ Passage à un problème en dimension fini

Apprentissage : trouver un “bon" vecteur $\theta^* \in \Theta$ de sorte à ce qu'ayant observé

$$Y = Y_1, Y_2, \ldots, Y_N$$

$$\Rightarrow \Pr(Y \in A) \approx f(A \,|\, \theta^*), \quad A \in \mathscr{Y}$$

Exemple 1 : Regression logistique

Application : Détecter la présence d'une maladie chez un patient

- Les données sont des réponses bouléennes $Y_1 \in \{0, 1\}$
  patient 1 est malade ou sain
- Variables explicatives (covariates) $X_1 = (X_{1,1}, \ldots, X_{1,r})$
  patient 1 a des données auxiliaires : age, sexe, indice de santé,...

Le modéle s'écrit

$$f(Y_1 \mid X_1, \theta) = \left\{ \frac{1}{1 + \exp(-\theta^T X_1)} \right\}^{Y_1} \left\{ 1 - \frac{1}{1 + \exp(-\theta^T X_1)} \right\}^{1-Y_1}.$$

Apprentissage : consiste a estimer le paramètre $\theta \in \Theta$
sachant $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$

Exemple 2 : Modèle à templates déformables

Application : Classifier un chiffre manuscrit

- l'image d'un chiffre $\mathscr{Y}_n$ est modélisée par une collection de fonctions **deterministes** (les **templates**) $\mathscr{T}^{(0)}, \ldots, \mathscr{T}^{(9)}$:

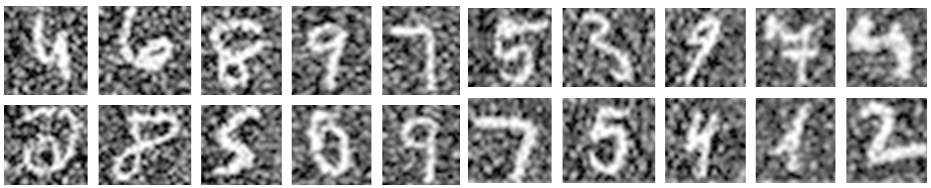
$$\mathscr{T}^{(i)} : \mathbb{R}^2 \to \mathbb{R}$$

- ces templates sont observés à travers un processus aléatoire
  - (i) de déformation du plan $D_n : \mathbb{R}^2 \to \mathbb{R}^2$,
  - (ii) et un sous bruit additif $\mathscr{W}_n : \mathbb{R}^2 \to \mathbb{R}$
- Modèle:

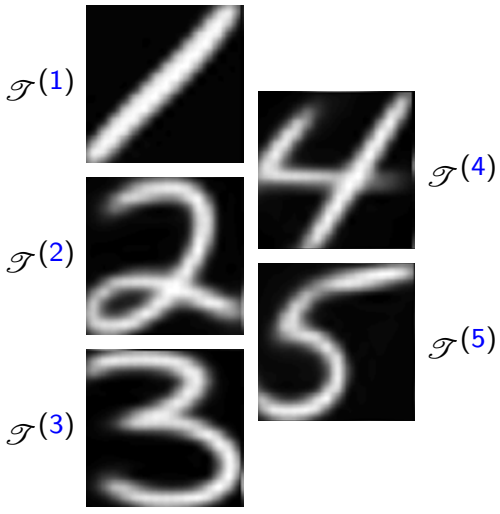
$$\forall u \in \mathbb{R}^2, \qquad \mathscr{Y}_n(u) = \mathscr{T}^{(i)} \circ D_n(u) + \mathscr{W}_n(u) \; .$$

- Les données sont des functions $\{\mathscr{Y}_n : \mathbb{R}^2 \to \mathbb{R},\, n \leq N\}$ discrétisées :

$$\{\mathscr{Y}_n,\, n \leq N\} \xrightarrow[\text{sur une grille de pixels}]{\text{discretisation}} \{Y_n,\, n \leq N\}$$



Exemple 2 : Modèle à templates déformables



5 templates

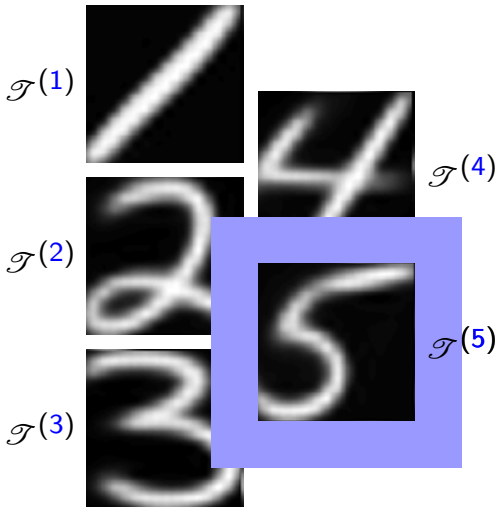
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



5 templates

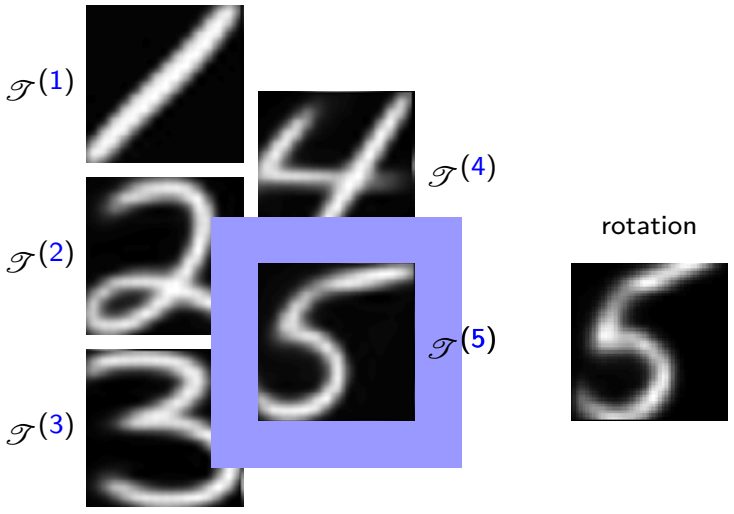
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



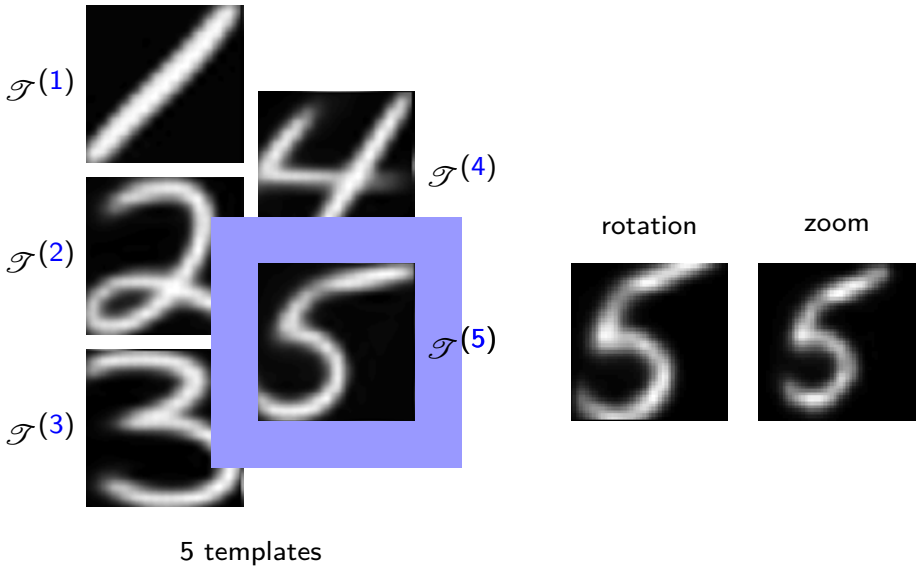
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



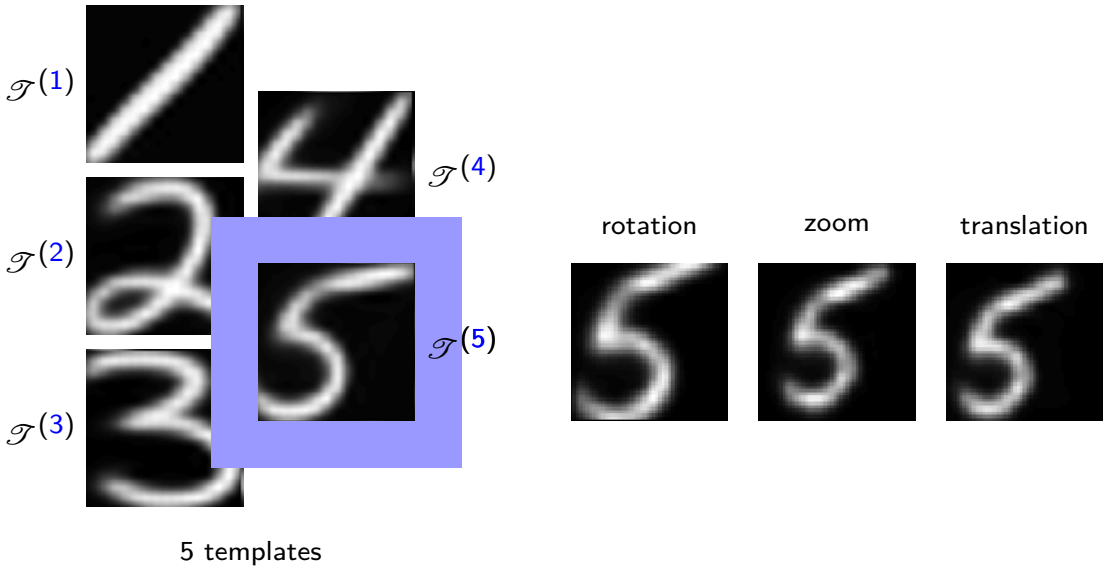
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



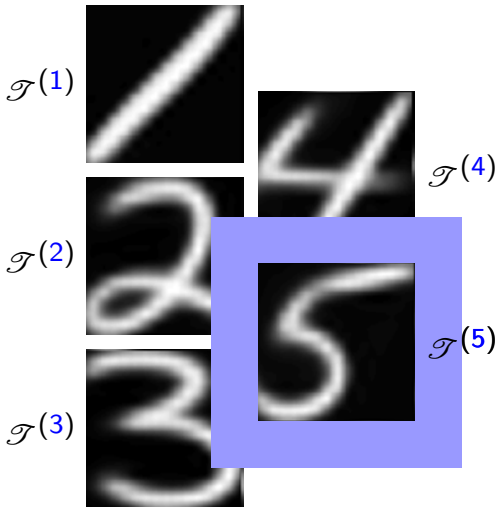
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

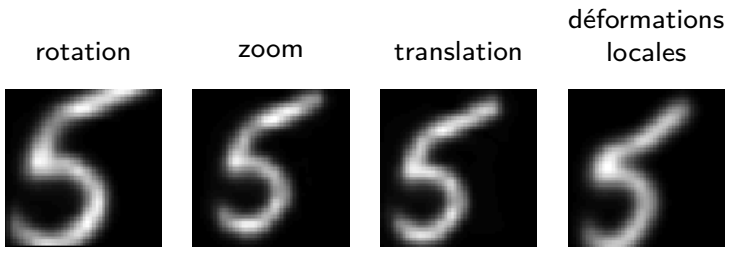
et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



5 templates



processus de déformation aléatoire

$\mathscr{T}^{(5)} \rightarrow \mathscr{T}^{(5)} \circ D_n^{(5)}$

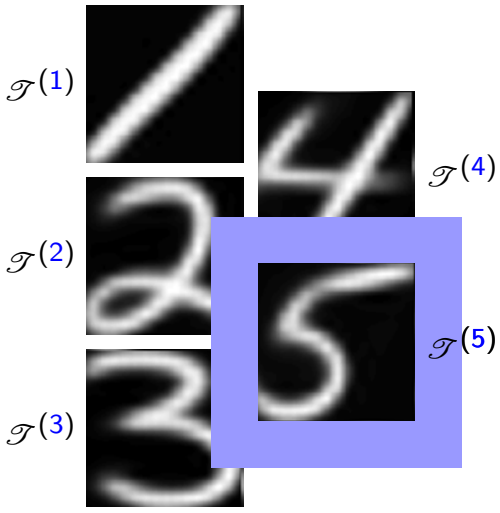
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

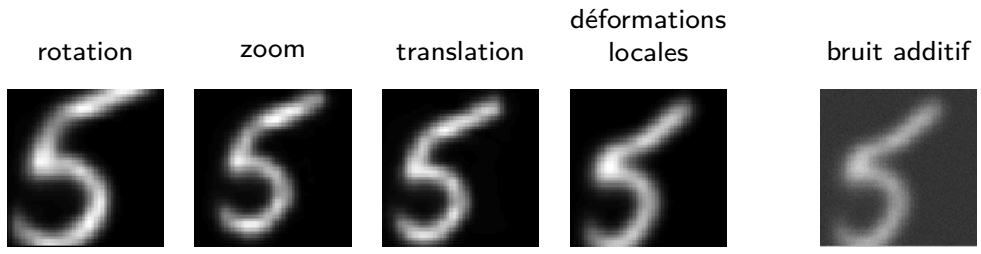
et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



5 templates



processus de déformation aléatoire

$\mathscr{T}^{(5)} \to \mathscr{T}^{(5)} \circ D_n^{(5)}$

$\mathscr{Y}_n$

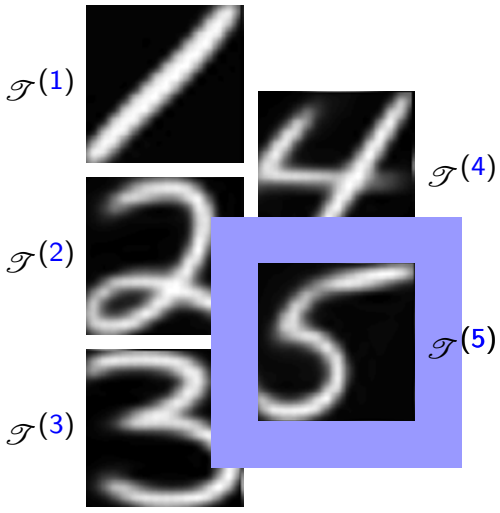
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

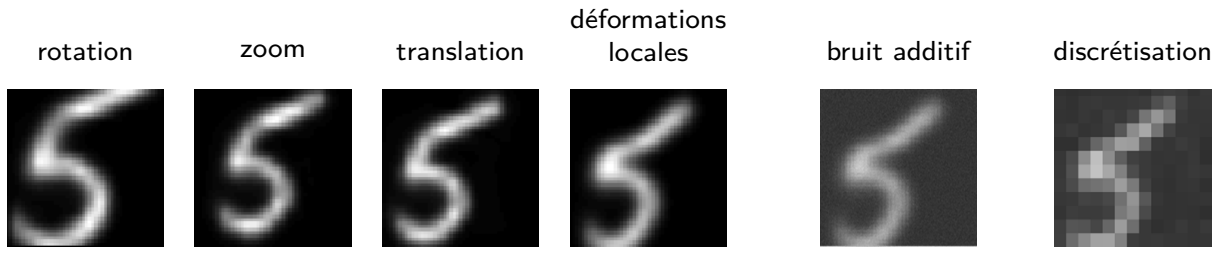
et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réécrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



5 templates



processus de déformation aléatoire

$\mathscr{T}^{(5)} \to \mathscr{T}^{(5)} \circ D_n^{(5)}$

$\mathscr{Y}_n$ Y_n

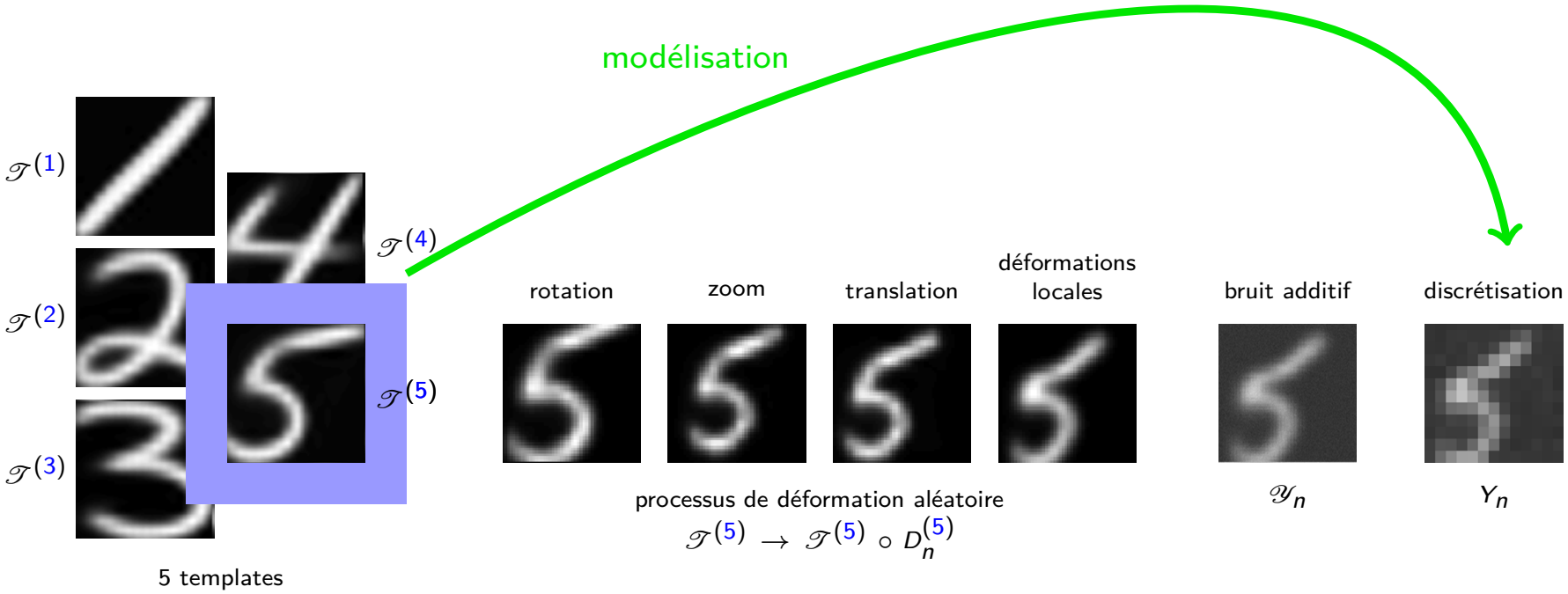
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



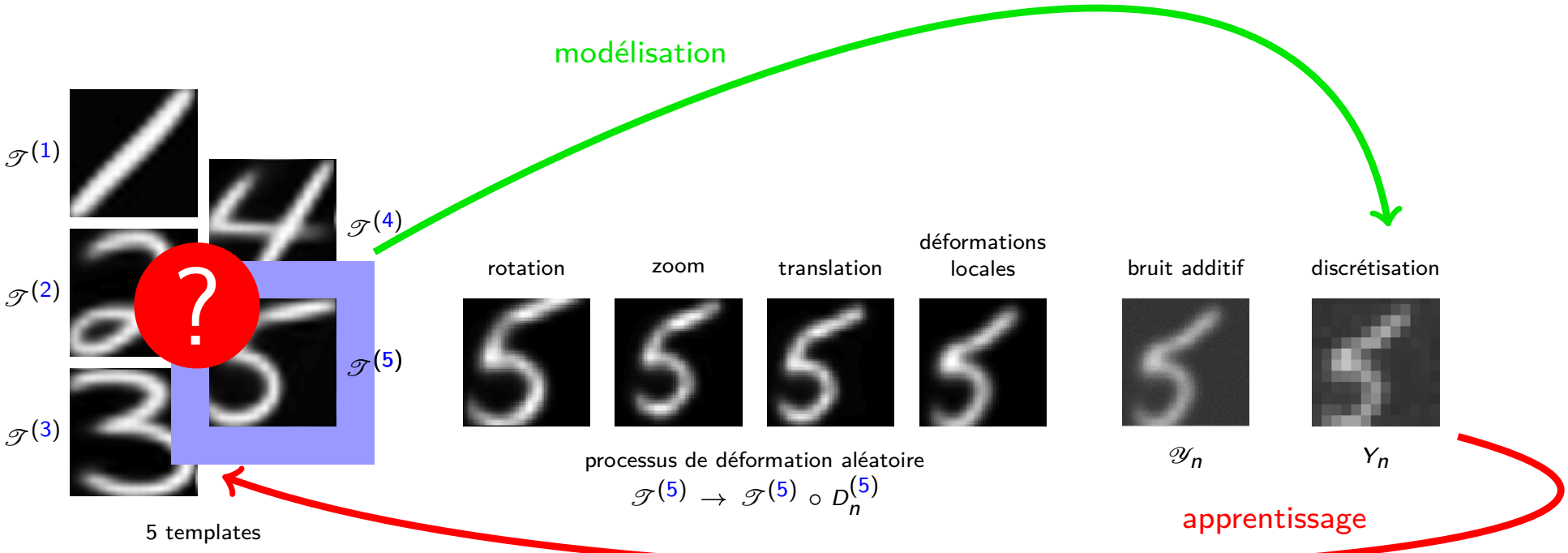
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réécrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



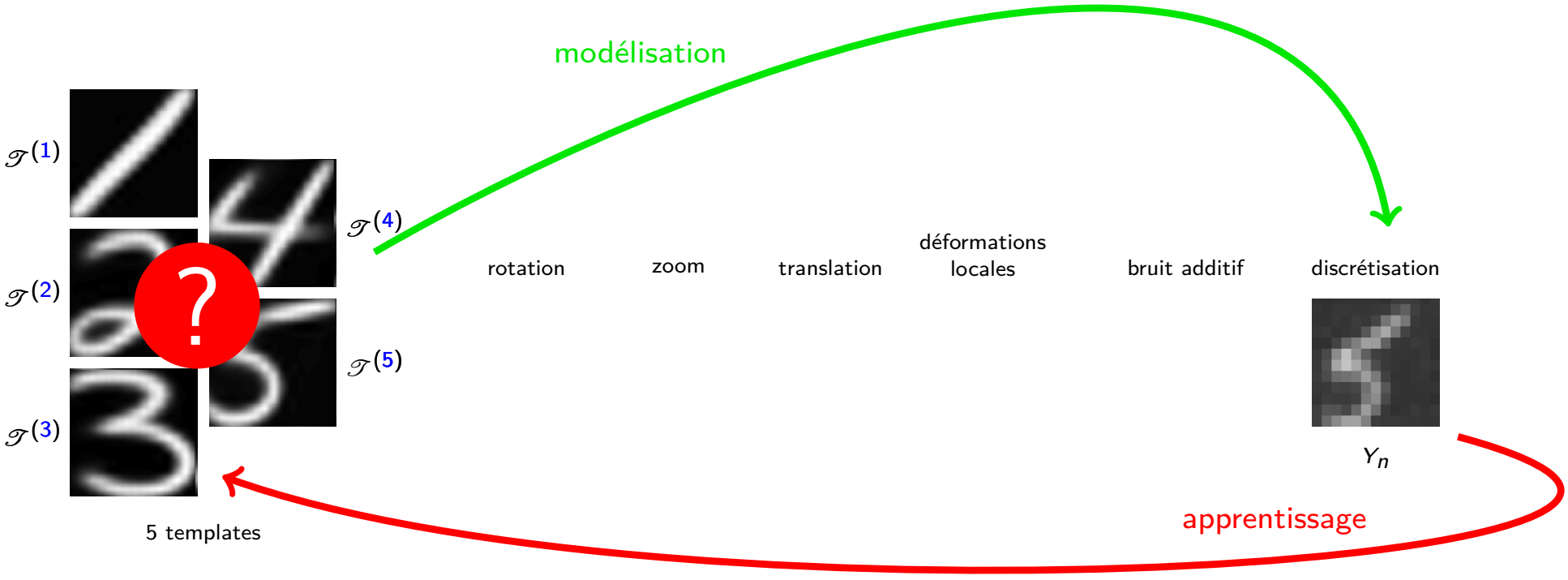
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réecrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



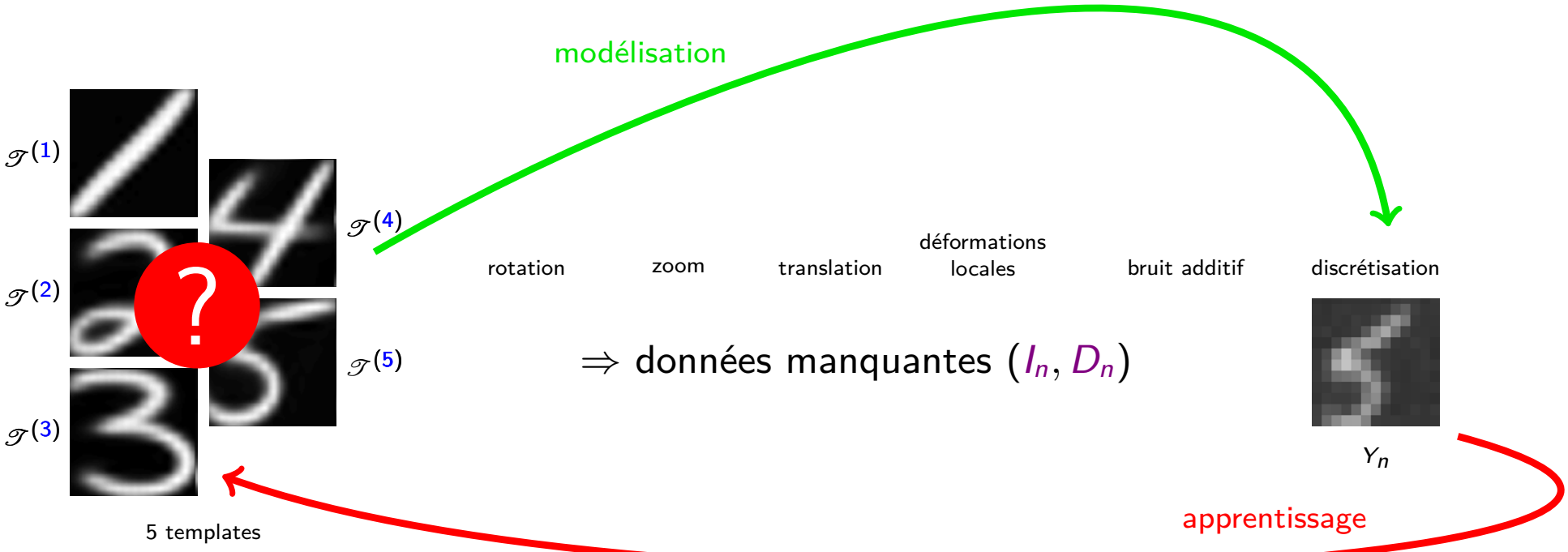
En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réécrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Exemple 2 : Modèle à templates déformables



En paramétrisant

- la famille de templates $\mathscr{T}$ par $\theta \in \Theta$
- le champ de déformation D_n par $\beta_n \in \mathbb{B}$

et en discrétisant $\mathscr{Y}_n$ sur une grille, le modèle peut se réécrire comme :

$$Y_n = \Phi_{\beta_n} \theta_{i_n} + \zeta_n, \qquad \zeta_n \sim \mathcal{N}(0, \Sigma)\,.$$

Apprentissage paramètrique

Deux écoles

	fréquentiste	**Bayésienne**
modèle	$f(\cdot \mid \theta)$, $\theta \in \Theta$	$f(\cdot \mid \theta)$, $\theta \in \Theta$
source d'information	données $Y = Y_1, Y_2, \ldots$	données $Y = Y_1, Y_2, \ldots$ $+$ une distribution *apriori* $\theta \sim p$
apprentissage	trouver l'ensemble Θ^* $\arg\min_{\theta \in \Theta} \mathrm{d}(Y, f(\cdot \mid \theta))$ pour une distance d	trouver la distribution π $\pi(\theta \mid Y) :\propto f(Y \mid \theta) p(\theta)$
nature du problème	optimisation	échantillonnage
objectif	$\Pr(\theta \in A \mid Y)$ intervalle de confiance	$\Pr(\theta \in A) = \pi(A \mid Y)$ intervalle de crédibilité
exemple	Maximum de vraisemblance	méthodes de Monte Carlo

Exemple 1 (régression logistique) : apprentissage Bayésien

- $N = 10^7$ données $Y = (Y_1, \ldots, Y_N)$
- covariates $X = (X_1, \ldots, X_N)$
- loi apriori $p(\theta) = \text{Laplace}(\theta; 1) \propto \exp\{-\|\theta\|_1\}$
- loi aposteriori

$$\pi(\theta \mid Y, X) \propto \prod_{n=1}^{N} \left\{ \frac{1}{1 + \exp(-\theta^T X_n)} \right\}^{Y_n} \left\{ 1 - \frac{1}{1 + \exp(-\theta^T X_n)} \right\}^{1-Y_n} e^{-\|\theta\|_1}$$

- π n'est pas une distribution standard $\Rightarrow$ l'échantillonnage de π n'est pas simple

Apprentissage Bayésien : les méthodes de Monte Carlo par chaîne de Markov (MCMC) offrent une possibilité simple et universelle d'échantillonner π

Méthode MCMC

Principe : construire une chaîne de Markov qui admette π comme distribution limite

Un exemple de MCMC : l'algorithme de Metropolis-Hastings construit une chaîne de Markov $\{\theta_k,\ k \in \mathbb{N}\}$

Initialisation $\theta_0 \sim p$ et à chaque itération $k = 1, 2, \ldots,$ faire

(i) $\theta' \sim Q(\theta_{k-1}, \cdot)$

(ii) définir $\theta_k = \theta'$ avec proba.:

$$\alpha(\theta_{k-1}, \theta') := 1 \wedge \frac{\pi(\theta' \mid Y) Q(\theta', \theta_{k-1})}{\pi(\theta_{k-1} \mid Y) Q(\theta_{k-1}, \theta')}$$
$$= 1 \wedge \frac{f(Y_1, \ldots, Y_N \mid \theta')}{f(Y_1, \ldots, Y_N \mid \theta_{k-1})} \frac{p(\theta') Q(\theta', \theta_{k-1})}{p(\theta_{k-1}) Q(\theta_{k-1}, \theta')}$$

et $\theta_k = \theta_{k-1}$ avec proba. $1 - \alpha(\theta_{k-1}, \theta')$.

Problème : la complexité de MH est en $\mathcal{O}(N)$ par itération

Exemple 2 (modèle déformable) : apprentissage fréquentiste

- $N = 10,000$ images $Y_1, Y_2, \ldots, Y_N$ et $\mathsf{Y} = \mathcal{M}_{15}\{(0,1)\}$
- pour chaque image Y_n, la classe I_n et le champ de déformation β_n sont inconnues
- le bruit additif est supposé Gaussien de sorte que

$$f(Y_1, \ldots, Y_N, I_n = i, \beta_n = d \,|\, \theta) = \mathcal{N}(\Phi_{\beta_n}\theta_{i_n}, \sigma \mathrm{Id}_{15})$$

- on pose une loi apriori sur les données manquantes $(I_n, \beta_n) \sim g$
- la fonction à maximiser sur Θ est

$$\theta \mapsto f(Y_1, \ldots, Y_N \,|\, \theta) = \sum_i \int_d f(Y_1, \ldots, Y_N, I_n = i, D_n = d \,|\, \theta) \mathrm{d}g(i,d)\,.$$

Apprentissage fréquentiste : l'algorithme Expectation-Maximization (EM) est une façon standard de trouver le maximum de f **en présence de données manquantes**

Algorithme EM

Principe : construire une séquence de paramètres $\{\theta_k,\ k \in \mathbb{N}\}$ qui converge vers un maximum de f

Initialisation avec $\theta_0 \in \Theta$ et à chaque itération $k = 1, 2, \ldots$:

(i) calculer l'espérance conditionnelle sous la loi des données manquantes sachant θ_{k-1}

$$S_k(\theta) = \mathbb{E}\left\{\log f(Y_1, \ldots, Y_N, X_1, \ldots, X_N \,|\, \theta) \,|\, \theta_{k-1}\right\}$$

$$\overset{iid}{=} \sum_{n=1}^{N} \mathbb{E}\left\{\log f(Y_n, X_n \,|\, \theta) \,|\, \theta_{k-1}\right\}$$

(ii) définir θ_k comme un maximiseur de $\theta \mapsto S_k(\theta)$

Problème : la complexité de EM est en $\mathcal{O}(N)$ par itération

Plan

Quelques informations sur mon parcours

Statistique computationelle : modélisation et apprentissage

Apprentissage statistique en présence d'un grand nombre de données : stabilité et efficacité d'approximations de certains algorithmes.

Quelques Illustrations

Approximations d'algorithmes

Thème de recherche actif : conception d'algorithmes qui

1– "imitent" les algorithmes d'apprentissage traditionnels

2– ont une complexité plus faible que $\mathcal{O}(N)$

3– héritent (dans une certaine mesure) des guaranties théoriques de ces algorithmes

Nous présentons deux algorithmes qui approximent MH et EM

- description des algorithmes
- guaranties théoriques
- outils mathématiques pour justifier l'intuition de ces algorithmes

Approximation de MH (avec N. Friel, UCD et P. Alquier, ENSAE)

Simuler une chaîne de Markov $\{\tilde{\theta}_k,\ k \in \mathbb{N}\}$ en utilisant un sous-ensemble des données $Y_U \subset Y$ de taille $n \ll N$ ($U \subset \{1, \ldots, N\}$) pour accepter/rejeter un nouvel état de la chaîne

$$\alpha(\theta, \theta') = 1 \wedge \frac{f(Y \mid \theta')}{f(Y \mid \theta)} \times \frac{p(\theta')Q(\theta', \theta)}{p(\theta)Q(\theta, \theta')}$$

$$\xhookrightarrow{\text{remplacé par}} \hat{\alpha}(\tilde{\theta}, \theta' \mid U) = 1 \wedge \frac{f(Y_U \mid \theta')^{N/n}}{f(Y_U \mid \tilde{\theta})^{N/n}} \times \frac{p(\theta')Q(\theta', \tilde{\theta})}{p(\tilde{\theta})Q(\tilde{\theta}, \theta')}$$

A chaque itération, le sous-ensemble Y_U est tiré suivant une loi

$$\nu_{n,\epsilon}(U) \propto \exp\{-\epsilon \| S(Y) - (N/n)S(Y_U) \|\}, \quad \epsilon > 0\,.$$

Pour un sous-ensemble $Y_U \subseteq Y$, $S(Y_U)$ est un vecteur de statistique résumé de Y_U.

Cette méthode approximant MH est appélée **ISS-MCMC**, acronyme pour

Informed Sub-Sampling MCMC

Complexité : $\mathcal{O}(n)$

Convergence de MH

On definit le niveau de convergence d'une chaîne de Markov par

$$D_k(\theta_0) := \|\Pr(\theta_k \in \cdot \,|\, \theta_0) - \pi\| \,.$$

MH produit une chaîne de Markov ergodique *i.e*

$$\forall\, \theta_0 \in \Theta\,, \qquad \lim_{k\to\infty} D_k(\theta_0) = 0 \qquad \text{p-s}\,.$$

et possiblement la convergence se fait géometriquement,

- simplement
$$\forall\, \theta_0 \in \Theta\,, \qquad D_k(\theta_0) \leq C(\theta_0)\rho^k\,.$$
- uniformément
$$\sup_{\theta_0\in\Theta} D_k(\theta_0) \leq C\rho^k\,.$$

Héritage des propriétés de MH

Il n'y a aucune raison pour que ISS-MCMC soit ergodique :
$\Pr(\tilde{\theta}_k \in A) \not\to_{k\to\infty} \pi(A)$

Hypothèse

Il existe un vecteur de statistiques suffisantes S pour f telle que

$$f(Y \mid \theta) = \psi(\theta) \exp\{\phi(\theta)^T S(Y)\}$$

Proposition

1. *Nous avons*
$$\mathrm{KL}\left(\pi(\cdot \mid Y), \pi(\cdot \mid Y_U)\right) \leq B(Y_U)$$
et $B(Y_U)$ est minimisé pour le(s) sous-ensemble(s) Y_U qui minimise(nt) $\|S(Y) - S(Y_U)\|$.
2. *Asymptotiquement quand $N \to \infty$, nous avons*
$$\mathrm{KL}\left(\pi(\cdot \mid Y), \pi(\cdot \mid Y_U)\right)$$
est minimisé pour le(s) sous-ensemble(s) Y_U qui minimise(nt) $\|S(Y) - S(Y_U)\|$.

Héritage des propriétés de MH pour ISS-MCMC

En l'absence de statistiques suffisantes, nous travaillons sous l'hypothèse suivante

Hypothèse

Il existe un vecteur de statistique résumé S et $\gamma > 0$ tels que

$$\forall\, \theta \in \Theta \qquad \left\| \frac{N}{n} \log f(Y_U \mid \theta) - \log f(Y \mid \theta) \right\| \leq \gamma \left\| S(Y) - \frac{N}{n} S(Y_U) \right\|$$

Proposition

- *Si MH converge simplement géométriquement alors ISS-MCMC converge de la même façon vers une approximation de π*
- *Si MH converge uniformément géométriquement alors ISS-MCMC converge de la même façon vers une approximation de π quantifiée par*

$$\begin{aligned} &\|\pi - \tilde{\pi}_k\| \\ &\leq \kappa \sup_{\theta \in \Theta} \mathbb{E}\left\{ \frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}} \right\} \sup_{U \in \mathsf{U}_n} \mathbb{E}\left\{ \frac{p(\tilde{\theta}) Q(\tilde{\theta}, \theta)}{p(\theta) Q(\theta, \tilde{\theta})} \left| \frac{f(Y_U \mid \theta)^{N/n}}{f(Y \mid \theta)} - \frac{f(Y_U \mid \tilde{\theta})^{N/n}}{f(Y \mid \tilde{\theta})} \right| \right\} \end{aligned}$$

Approximation de EM (avec É. Moulines, École Polytechnique)

Hypothèse

Il existe un vecteur de statistique suffisante S i.e une fonction Φ telle que

$$f(Y, X \mid \theta) = \Phi(\theta, S(Y, X)) .$$

Dans notre méthode, les observations $Y_1, Y_2, \ldots$ sont traîtées une par une

Initialisation à $\tilde{\theta}_0 \in \Theta$, la séquence $\{\tilde{\theta}_k,\ k \in \mathbb{N}\}$ est définie par

1. approximation étape E : mise à jour des statistiques suffisantes

$$s_k = s_{k-1} + \varrho_k \left(S(Y_k, X_k) - s_{k-1} \right) , \quad X_k \sim \pi(\cdot \mid Y_k, \tilde{\theta}_{k-1}),$$

   avec $\varrho_k = \varrho_0 k^{-\alpha}$, $\alpha \in (0, 1)$

2. approximation étape M : mise à jour du paramètre

$$\tilde{\theta}_{k+1} = \bar{\theta}(s_{k+1}) := \arg \max_{\theta \in \Theta} \Phi(\theta, s_{k+1}) .$$

Cette méthode approximant EM est appélée **MCoEM**, pour *Monte Carlo online EM*

Complexité : $\mathcal{O}(1)$

Héritage des propriétés théoriques de l'EM

Dans sa forme traditionnelle, la séquence de l'EM converge pour tout $\theta_0 \in \Theta$ vers

$$\Theta^* := \{\theta \in \Theta,\ \nabla_{\theta'} f(Y \mid \theta') \mid_{\theta} = 0\}$$

Proposition

Sous certaines hypothèses techniques (sur la topologie de Θ), la séquence $\{\tilde{\theta}_k,\ k \in \mathbb{N}\}$ produite par MCoEM converge vers

$$\Theta^\dagger := \{\theta \in \Theta,\ \nabla_{\theta'} \mathrm{KL}(\pi^*, f(\cdot \mid \theta')) \mid_{\theta'=\theta}\} = 0\,.$$

presque-sûrement quand $k \to \infty$. π^* est le processus génératif des données qui peut ne pas appartenir à $\mathscr{F}$.

Héritage des propriétés théoriques de l'EM

Monotonicité

- À chaque itération EM vérifie $f(Y \mid \theta_{k+1}) \geq f(Y \mid \theta_k)$
- À chaque itération MCoEM vérifie

$$\mathrm{KL}(\pi^*, f(\cdot \mid \tilde{\theta}_{k+1})) \leq \mathrm{KL}(\pi^*, f(\cdot \mid \tilde{\theta}_k)) + o(\|\tilde{\theta}_{k+1} - \tilde{\theta}_k\|)\,, \quad \text{p-s}$$

Taux de convergence

- Conditionellement à (θ_0, Y), EM est déterministe mais la convergence se fait en $1/\sqrt{n}$
- Conditionellement à $(\tilde{\theta}_0, Y)$, MCoEM est stochastique et la convergence se fait en $1/\sqrt{\varrho_n} \propto 1/\sqrt{n^{-\alpha}}$, $\alpha \in (0, 1)$.
- On peut de plus prouver la normalité asymptotique pour MCoEM:

$$\varrho_n^{-1/2}(\tilde{\theta}_n - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma)\,,$$

où Σ est la matrice de variance covariance d'un problème de Lyapunov défini dépendant de $\mathrm{Var}\{S(Y, X) \mid Y, \theta\}$.

Outils mathématiques pour établir des résultats en présence d'approximations (1/2)

Pour une chaîne de Markov, on utilise la théorie de perturbation des opérateurs linéaires:

- caractérisation de la chaîne par l'opérateur de transition

$$K(\theta, A) := \Pr(\theta_{n+1} \in A \,|\, \theta_n = \theta)$$

- la loi de probabilité de la chaîne est simplement l'itération de K

Comment une approximation locale se propage-t-elle asymptotiquement?

Exemple de résultat:

Proposition (Mitrophanov (AAP, 2005))

Soit une chaîne $\{\tilde{\theta}_n,\ n \in \mathbb{N}\}$ *ayant pour transition* $\tilde{K}$*. Si* K *est uniformément ergodique alors pour n suffisament grand, nous avons :*

$$\| \Pr\{\theta_n \in \cdot\} - \Pr\{\tilde{\theta}_n \in \cdot\} \| \leq \kappa_n \| K - \tilde{K} \|$$

$\Rightarrow \tilde{K}$ joue le rôle d'une version bruitée de l'opérateur de référence (le noyau de transition de MH dans notre cas)

Outils mathématiques pour établir des résultats en présence d'approximations (2/2)

L'EM peut également être réecrit en terme d'opérateur

$$\theta_{n+1} = K(\theta_n), \qquad K(\theta) = \bar{\theta}\left(\mathbb{E}\{S(Y, X) \,|\, \theta\}\right)$$

K est un opérateur non linéaire
$\Rightarrow$ les points stationnaires de l'EM sont dans $\ker h$, $h := K - \mathrm{Id}$

Approximation stochastique (Robbins et Monro, 1951) : trouver les zéros d'une fonction h ne pouvant être évaluée que de façon bruitée $\hat{h}(\theta) \approx h(\theta)$

Sous certaines hypothèses sur l'approximation $\hat{h}$, la séquence $\{\tilde{\theta}_n\,, n \in \mathbb{N}\}$

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n + \varrho_n \hat{h}(\tilde{\theta}_n)\,,$$

où $\varrho_n \searrow 0$, $\sum_n \varrho_n = \infty$, $\sum_n \varrho_n^2 < \infty$ alors :

$$\tilde{\theta}_n \to \{\theta \in \Theta\,, h(\theta) = 0\} \text{ p-s}\,.$$

MCoEM : $h \xrightarrow{\textit{remplacé par}} \hat{h} = \bar{\theta}\left(S(Y_n, X_n)\right) - \mathrm{Id}, \qquad (Y_n, X_n) \sim \delta_{y_n} \otimes \pi(\,\cdot \,|\, \tilde{\theta}_n, Y_n)$

Plan

Quelques informations sur mon parcours

Statistique computationelle : modélisation et apprentissage

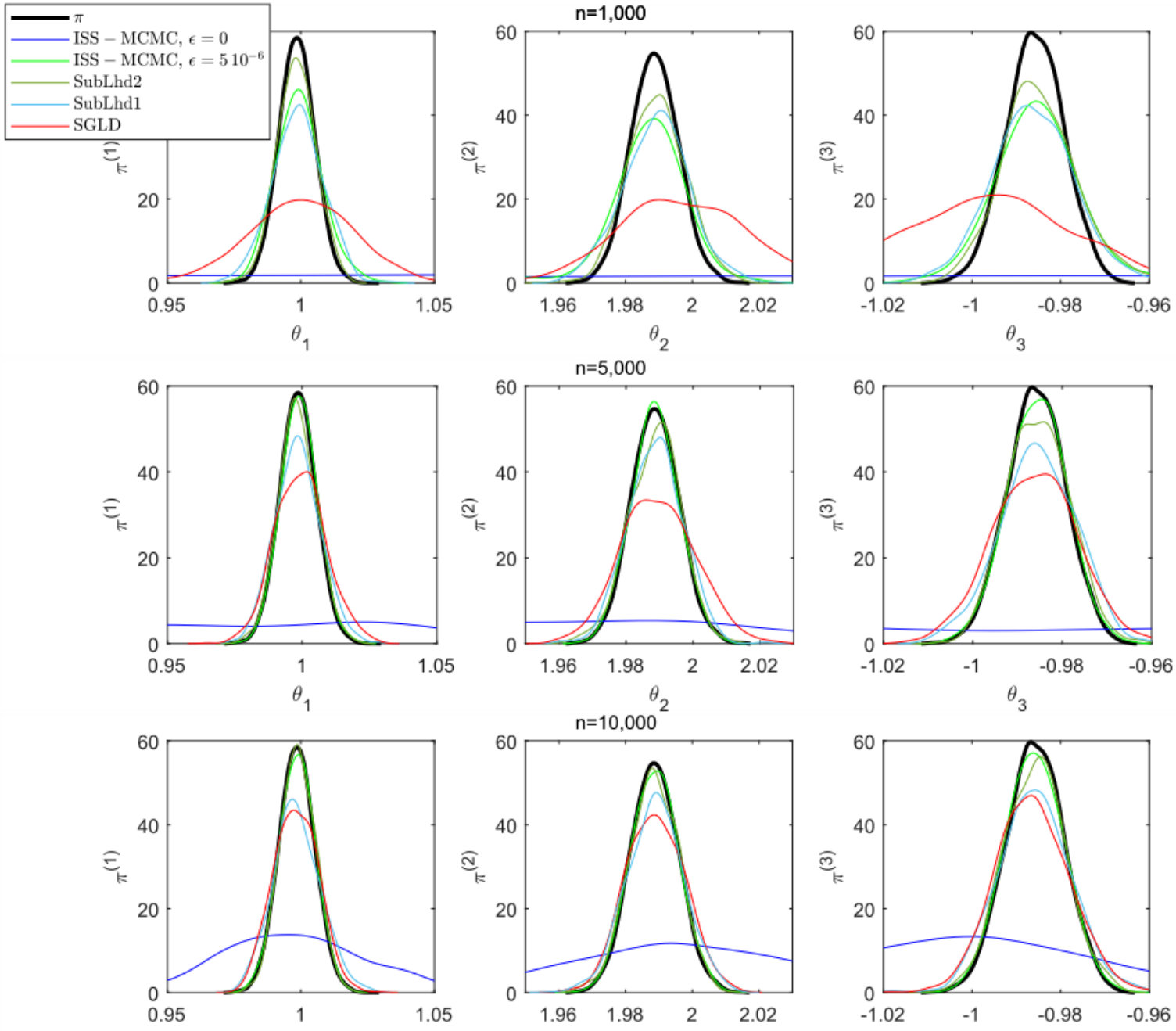
Apprentissage statistique en présence d'un grand nombre de données : stabilité et efficacité d'approximations de certains algorithmes.

Quelques Illustrations

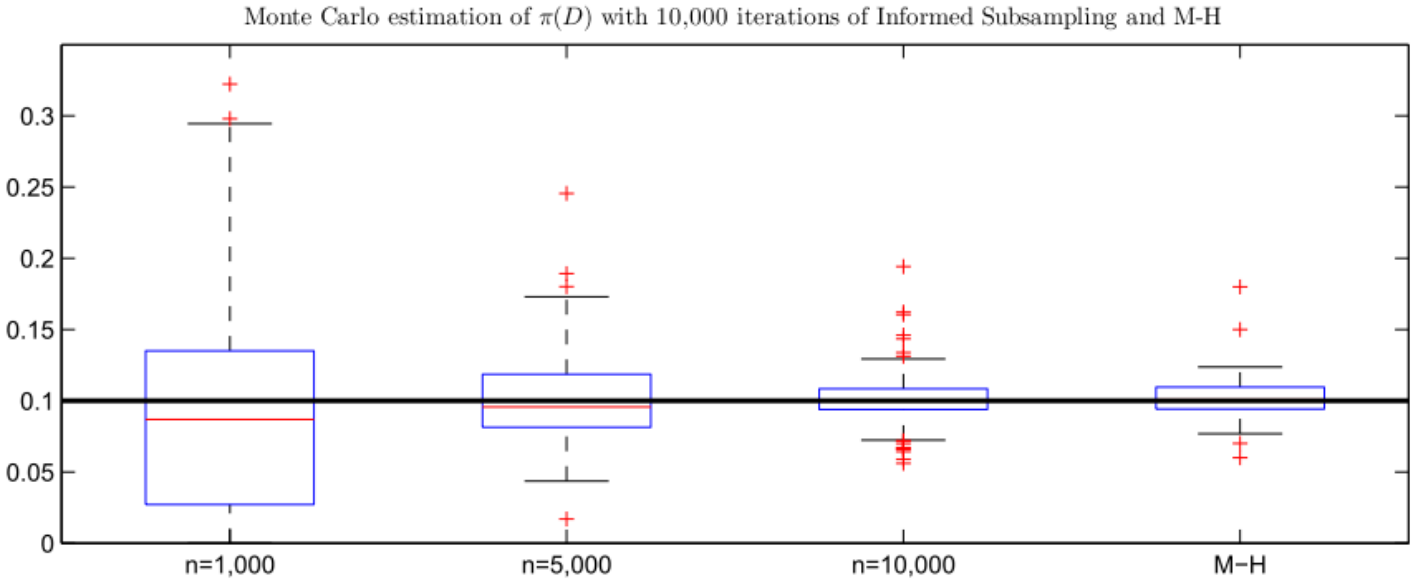
Régression logistique (1/3)

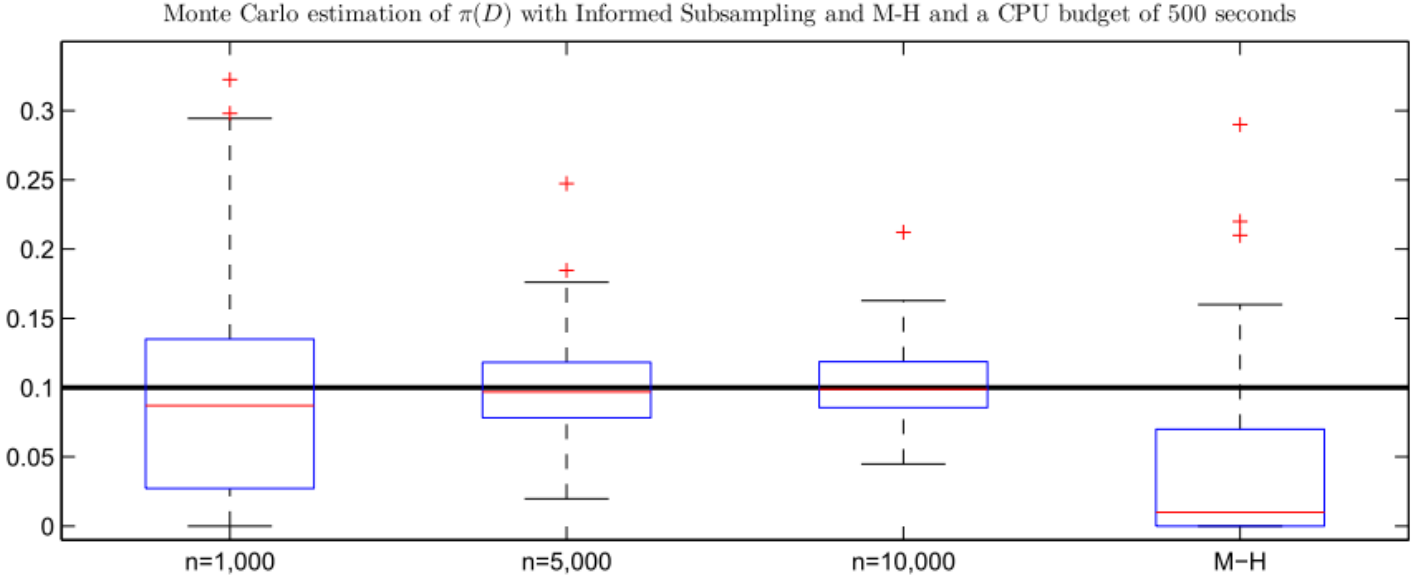
algorithm	time/iter.(s)	iter. completed	RMSE	$\widehat{\text{var}\{\pi(D)\}}$
M–H	10	50	0.1417	0.004
ISS-MCMC, $n = 1,000$	0.05	$10,000$	0.1016	0.0104
ISS-MCMC, $n = 5,000$	0.08	$6,250$	0.0351	0.0012
ISS-MCMC, $n = 10,000$	0.13	$3,840$	0.0267	0.0007
SGLD, $n = 1,000$	0.08	$6,000$	0.1370	0.0157
SGLD, $n = 5,000$	0.11	$5,250$	0.0996	0.0100
SGLD, $n = 10,000$	0.12	$4,500$	0.0326	0.0011

Régression logistique (2/3)



Régression logistique (3/3)





Modèle déformable

- Performance computationelle

	MCoEM	EM (N=50)	EM (N=300)
temps CPU / itération (sec.)	20	225	1,570

- Performance en classification de chiffre manuscrit

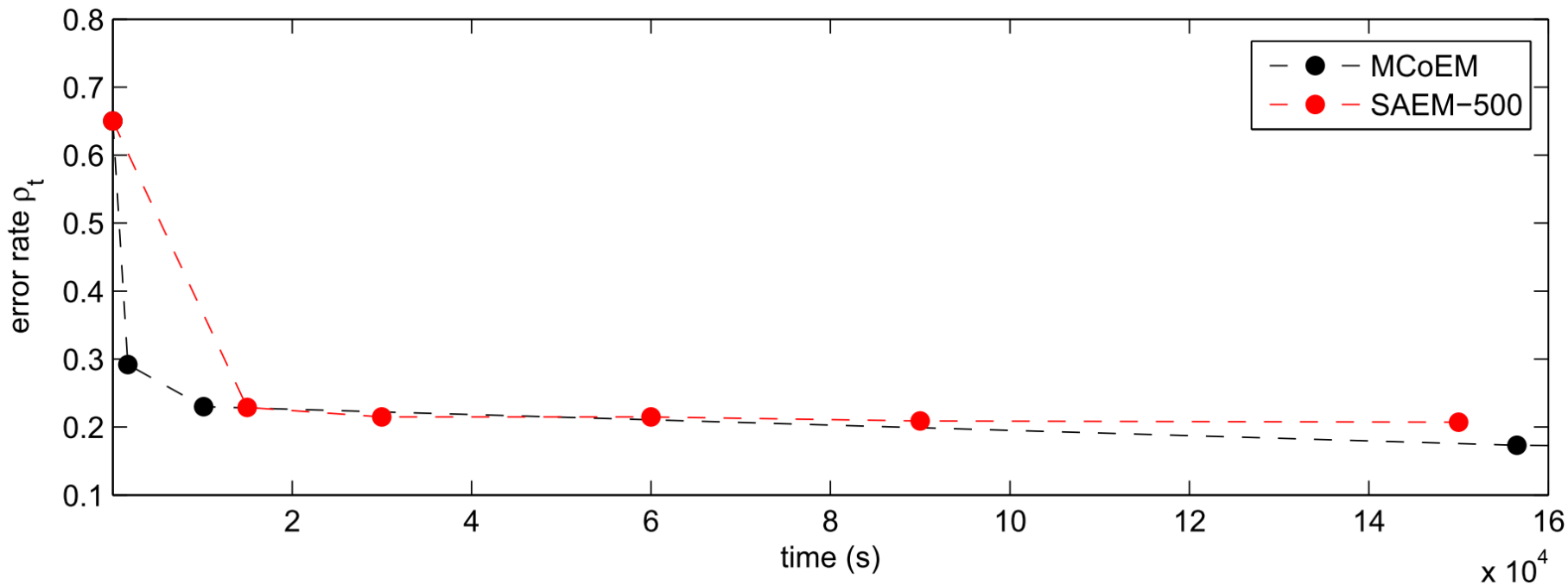


Figure: Erreur de classification en fonction du temps CPU

Discussion

- Il est crucial d'établir des **approximations scalables** d'algorithmes bien connus afin d'exploiter au mieux les progrés des systèmes de mesure et de stockage
- Nous avons proposé de telles approximations pour **Metropolis-Hastings** et **Expectation-Maximization**
- Trouver des guaranties théoriques pour ce genre de méthode représente un enjeu de taille, allant au delà de la communauté statistique
- Les résultats énoncés ici reposent sur **des hypothèses pas toujours réalistes** (existence de statistiques suffisantes ou "bonnes" statistiques résumées, convergence uniforme, Θ compact...)
- D'autres outils théoriques ont été ou pourraient être utilisès comme le **couplage** de processus stochastiques ou la **théorie spectrale** des opérateurs