Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation.

Estelle Kuhn

INRA, MaIAGE

2015 November 27

[joint work with
Stéphanie Allassonnière (Ecole Polytechnique, CMAP)
and Stanley Durrleman (INRIA, ARAMIS)]

Outline

2D Medical Images : mouse jawbone

Sample of 36 images of mouse jawbone

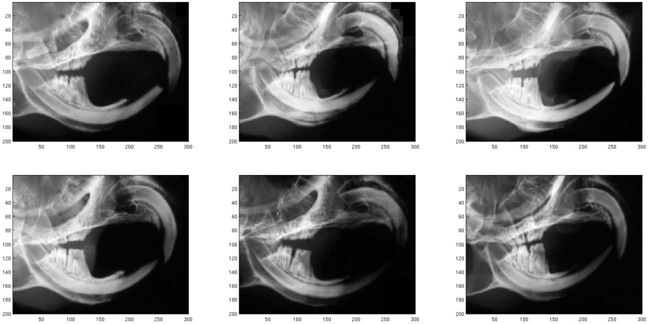


FIGURE : 6 images of the dataset.

Generative Statistical Model for Deformable Template [Allassonnière, Amit, Trouvé (2007)]

Consider n images denoted by $(y_1, \ldots, y_n)$.
For $1 \leq i \leq n$

$$y_i = I_0 \circ \varphi_i + \sigma \varepsilon_i$$

where I_0 is the template, φ_i is the deformation, σ^2 the variance and ε_i the noise.

$\rightarrow$ I_0 and φ_i defined on the whole plane

Parametric model

$$y_i = I_0 \circ \varphi_i + \sigma\varepsilon_i$$

Let $(p_k)_{1 \le k \le k_p}$ be some landmarks on the domain D.
Then for $\alpha \in \mathbb{R}^{k_p}$ we define the template by

$$I_0(x) = (K_p\alpha)(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\alpha(k).$$

Let $(g_k)_{1 \le k \le k_g}$ be some geometrical landmarks on D.
Then for $\beta \in \mathbb{R}^{dk_g}$ we define the field of deformation by

$$\varphi(x) = (K_g\beta)(x) = \sum_{k=1}^{k_g} K_g(x, g_k)\beta(k).$$

Parametric model

$$y_i = I_0 \circ \varphi_i + \sigma\varepsilon_i$$

$$I_0(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\alpha(k).$$

$$\varphi_i(x) = \sum_{k=1}^{k_g} K_g(x, g_k)\beta_i(k).$$

$(\beta_i)_{1\leq i\leq n}$ parameters of the deformations $(\varphi_i)_{1\leq i\leq n}$

$\Longrightarrow$ interest in the global geometrical behaviour

Parametric model

$$y_i = I_0 \circ \varphi_i + \sigma \varepsilon_i$$

$$I_0(x) = \sum_{k=1}^{k_p} K_p(x, p_k)\alpha(k).$$

$$\varphi_i(x) = \sum_{k=1}^{k_g} K_g(x, g_k)\beta_i(k).$$

$(\beta_i)_{1 \leq i \leq n}$ parameters of the deformations $(\varphi_i)_{1 \leq i \leq n}$

$\Longrightarrow$ interest in the global geometrical behaviour

$\Longrightarrow$ consider $(\beta_i)_{1 \leq i \leq n}$ as missing random variables and estimate the parameters of its distribution.

Generative model $(\mathcal{M})$:

$$\begin{cases} \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g) \mid \Gamma_g \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(I_\alpha \circ \varphi_{\beta_i}, \sigma^2 \mathrm{Id}) \mid \beta_1^n, \alpha, \sigma^2 \end{cases}$$

Generative model $(\mathcal{M})$:

$$\begin{cases} \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g) \mid \Gamma_g \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(I_\alpha \circ \varphi_{\beta_i}, \sigma^2 \mathrm{Id}) \mid \beta_1^n, \alpha, \sigma^2 \end{cases}$$

$\Longrightarrow$ big structures to learn even with small training set

Bayesian generative model $(\mathcal{M})$:

$$\begin{cases} (\Gamma_g, \alpha, \sigma^2) \sim \nu_g \otimes \nu_p \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g) \mid \Gamma_g \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(I_\alpha \circ \varphi_{\beta_i}, \sigma^2 \mathrm{Id}) \mid \beta_1^n, \alpha, \sigma^2 \end{cases}$$

where $\nu_g(d\Gamma_g), \nu_p(d\sigma^2, d\alpha)$ are prior laws

Bayesian generative model $(\mathcal{M})$:

$$\begin{cases} (\Gamma_g, \alpha, \sigma^2) \sim \nu_g \otimes \nu_p \\ \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}(0, \Gamma_g) \mid \Gamma_g \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}(I_\alpha \circ \varphi_{\beta_i}, \sigma^2 \mathrm{Id}) \mid \beta_1^n, \alpha, \sigma^2 \end{cases}$$

where $\nu_g(d\Gamma_g), \nu_p(d\sigma^2, d\alpha)$ are prior laws

Parameters $\theta = (\alpha, \sigma^2, \Gamma_g)$ are estimated by maximum a posteriori

$$\hat{\theta}_n = \arg\max h(\theta | y_1^n)$$

General latent variable model

Denote by y the observed variable
and by ϕ the latent variable

Assume the complete likelihood f of (y, ϕ) belongs to a parametric family $\{f(y, \phi; \theta), \theta \in \Theta\}$.

General latent variable model

Denote by y the observed variable
and by ϕ the latent variable

Assume the complete likelihood f of (y, ϕ) belongs to a parametric family $\{f(y, \phi; \theta), \theta \in \Theta\}$.

$\Longrightarrow$ Compute the value θ^{ML} that maximises the observed likelihood given by :

$$g(y; \theta) = \int f(y, \phi; \theta) d\phi$$

The EM algorithm [Dempster et al. (1977)]

Heuristic : if ϕ were observed, then consider $\log f(y, \phi; \theta)$
$\Longrightarrow$ consider $E[\log f(y, \phi; \theta) | y; \theta]$.

The EM algorithm [Dempster et al. (1977)]

Heuristic : if ϕ were observed, then consider $\log f(y,\phi;\theta)$
$\Longrightarrow$ consider $E[\log f(y,\phi;\theta)|y;\theta]$.

Iteration k of the algorithm :

- Expectation step :

$$Q(\theta|\theta_{k-1}) = E[\log f(y,\phi;\theta)|y;\theta_{k-1}]$$

- Maximization step :

$$\theta_k = Argmax \;\; Q(\theta|\theta_{k-1})$$

The EM algorithm [Dempster et al. (1977)]

Heuristic : if ϕ were observed, then consider $\log f(y,\phi;\theta)$
$\Longrightarrow$ consider $E[\log f(y,\phi;\theta)|y;\theta]$.

Iteration k of the algorithm :

- Expectation step :

$$Q(\theta|\theta_{k-1}) = E[\log f(y,\phi;\theta)|y;\theta_{k-1}]$$

- Maximization step :

$$\theta_k = Argmax \;\; Q(\theta|\theta_{k-1})$$

+ increase of $Q \Longrightarrow$ increase of the observed likelihood g
+ converges toward a stationary point $\widehat{\theta_g}$ of g
– theory in exponential model
– nature of the limit point
– convergence depends on the initial guess
– expression of $Q(\theta|\theta')$ often analytically intractable

Stochastic Approximation of the EM algorithm [Delyon et al (1999), K., Lavielle (2004), Allassonnière, K., Trouvé (2010)]

Iteration k of the algorithm :

- Simulation step : $\phi^k \sim \Pi_{\theta_{k-1}}(\phi^{k-1}, \cdot)$
  where Π_θ is a transition probability of an ergodic Markov Chain having the posterior distribution $p(\cdot|y, \theta)$ as stationary distribution,
- Stochastic approximation :
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k[\log f(y, \phi^k, \theta) - Q_{k-1}(\theta)]$ where (γ_k) is a decreasing sequence of positive step-sizes such that $\sum \gamma_k = +\infty$ and $\sum \gamma_k^2 < +\infty$.
- Maximisation step : $\theta_k = \arg\max Q_k(\theta)$

+ converges almost surely toward a stationary point $\widehat{\theta_g}$ of g

Anisotropic Metropolis Adjusted Langevin Algorithm [Allassonnière, K. (2015)]

$\Longrightarrow$ Metropolis Hastings algorithm with optimized proposal
Let π be the density of the target distribution and for $b > 0$ the drift $D(x) = \nabla \log \pi(x) 1_{|\nabla \log \pi(x)| < b} + b 1_{|\nabla \log \pi(x)| > b}$
Iteration k of the algorithm :

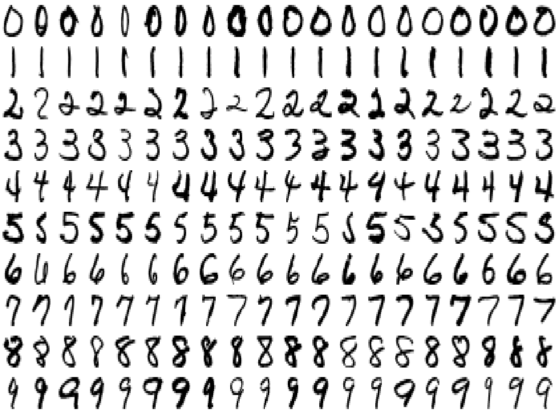
- $X_c | X_k \sim \mathcal{N}\left(X_k + \delta D(X_k), \delta(\varepsilon Id + D(x)D(x)^T)\right)$ with $\varepsilon > 0$.
- compute the acceptance ratio

$$\rho(X_k, X_c) = \min\left(1, \frac{\pi(X_c) q_c(X_c, X_k)}{q_c(X_k, X_c)\pi(X_k)}\right).$$

- update $X_{k+1} = X_c$ with probability $\rho(X_k, X_c)$ and $X_{k+1} = X_k$ with probability $1 - \rho(X_k, X_c)$

Results : ergodicity of the AMALA chain
convergence and CLT of AMALA-SAEM algorithm

Some images of the US postal database :



Template estimation in $(\mathcal{M})$:

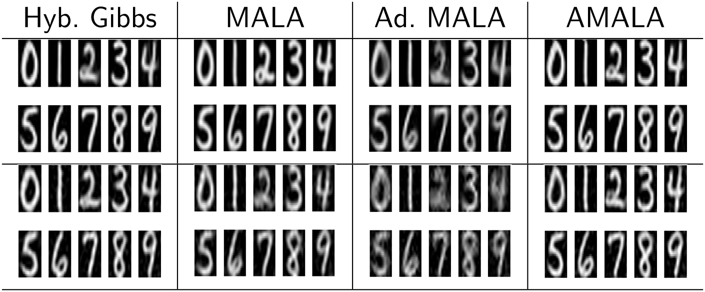


FIGURE : Estimated templates using four samplers with original data (first line) and noisy data (second line). The training set includes 20 images per digit. The hidden variable are of size $2k_g = 72$.

Template estimation in $(\mathcal{M})$:

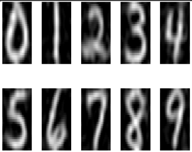
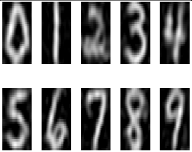
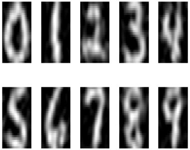
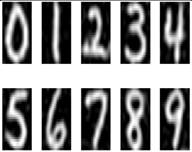
Defor. size/ Samp.	$2k_g = 72$	$2k_g = 128$	$2k_g = 200$
MALA			
AMALA			

FIGURE : Estimated templates using MALA and AMALA samplers in the stochastic EM algorithm on noisy training data. The training set includes 20 images per digit. The dimension of the hidden variable increases from 72 to 200.

Optimizing geometrical landmark locations in deformable template model ($\mathcal{M}$)

$\Longrightarrow$ considering geometrical landmarks as global variables of the model ($\mathcal{M}$) and estimating their locations

$$\varphi_\beta(x) = \sum_{k=1}^{k_g} K_g(x, r_{g,k})\beta(k).$$

Optimizing geometrical landmark locations in deformable template model ($\mathcal{M}$)

$\Longrightarrow$ considering geometrical landmarks as global variables of the model ($\mathcal{M}$) and estimating their locations

$$\varphi_\beta(x) = \sum_{k=1}^{k_g} K_g(x, r_{g,k})\beta(k).$$

$$\begin{cases} \theta = (\alpha, \sigma^2, \Gamma_g, \bar{r}) \sim (\nu_p \otimes \nu_g) \\ r_g \sim \mathcal{N}(\bar{r}, \sigma_r^2 Id) |\ \theta, \\ \beta_i \sim \mathcal{N}(0, \Gamma_g) |\ \theta,\ \forall 1 \le i \le n, \\ y_i | \beta_i, r_g \sim \mathcal{N}(I_\alpha \circ (\varphi_{\beta_i}^{r_g})), \sigma^2 Id) \mid \beta_i,\ \theta,\ r_g,\ \forall 1 \le i \le n. \end{cases}$$

with the prior distribution as before and Gaussian for $\bar{r}$.

Optimizing the geometrical landmarks locations

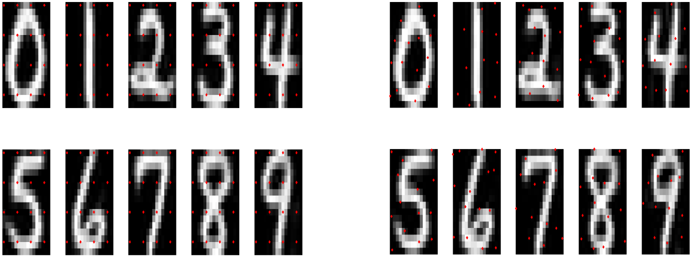


FIGURE : Estimated templates with 16 control points with either fixed (left) or estimated (right) control points positions.

Optimizing the number of geometrical landmarks

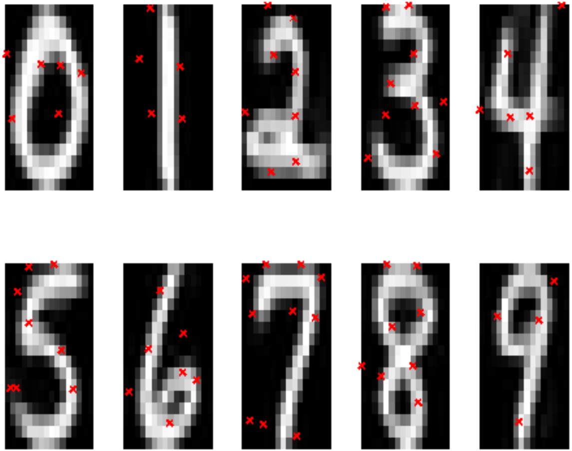


FIGURE : Estimated templates with their optimal numbers and positions of control points.

Templates estimation with mouse jawbone images

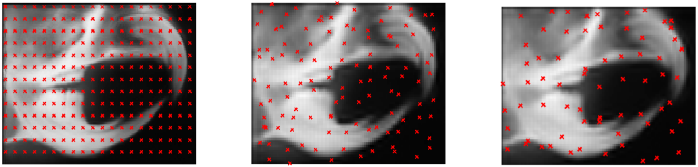


FIGURE : Estimated templates of the mouse mandible images obtained with 260 fixed control points (left), with 117 (middle) and 70 (right) estimated control points.

Application to maize leaf morphology analysis

with A. Ressayre and C. Dillmann

FIGURE : One image of maize leaf.

Bibliography

- [$A.A.T.$] Allassonnière, Amit, Trouvé, Toward a coherent statistical framework for dense deformable template estimation, JRSS B (2007)
- [$A.K.T.$] Allassonnière, Kuhn, Trouvé, Construction of Bayesian deformable models via a stochastic approximation algorithm : A convergence study, Bernoulli (2010)
- [$A.K.$] Allassonnière, Kuhn, Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation, CSDA (2015)
- [$A.D.K.$] Allassonnière, Durrleman, Kuhn, Bayesian Mixed Effect Atlas Estimation with a Diffeomorphic Deformation Model, SIAM Journal of Imaging Sciences (2015)

Thank you for your attention !

Large Deformation Framework [Trouvé and Younès (1995) among others]

Idea : Build a diffeomorphic map as successive instantaneous steps of time dependent local small deformations
$\varphi_t = Id + z_t$ where (z_t) is called the velocity field.
The motion of a point r_0 describes a curve satisfaying the Flow Equation for $t \in [0, 1]$

$$\begin{cases} \frac{dr(t)}{dt} = z_t(r(t)) \\ r(0) = r_0 \end{cases}.$$

The deformation φ_1 is defined as follows :

$$\forall r_0 \in D, \quad \varphi_1(r_0) = r(1).$$

$+$ under some conditions if $\forall t \; z_t \in H$ Hilbert space
then existence and unicity of the solution φ_1
which is a C^1 diffeomorphic map.
$-$ expensive in computational cost