Estimation in Bayesian mixed effect template model.

Estelle Kuhn

INRA, MaIAGE

18 juin 2015

(joint work with Stéphanie Allassonnière, Alain Trouvé, Stanley Durrleman)

Outline

Context of Computational Anatomy

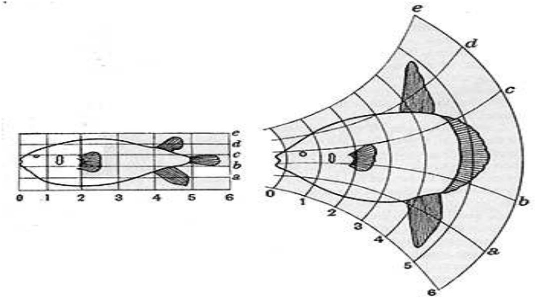
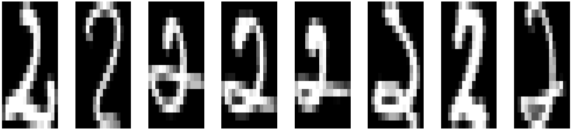
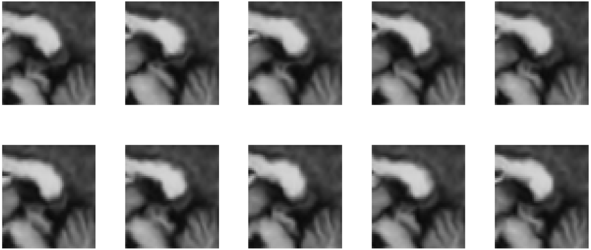


FIGURE: D'Arcy Thompson 1917

* Describing shapes
* Shape matching
* Creating Atlases
* Discrimination/Classification : how to measure differences between shapes

Examples of datasets





Matching problem

- Let I_0 be some template and I_1 be some observed image. $\Longrightarrow$ compute an optimal deformation φ which matches $I_0 \circ \varphi^{-1}$ on I_1 ?
- Variational approach by energy minimization :

$$\underbrace{\frac{1}{2}||\varphi||_H}_{\text{deformation cost}} + \underbrace{\frac{1}{2\sigma^2}||I_1 - I_0 \circ \varphi^{-1}||^2}_{\text{data attachment}}$$

Issue : choice of H, σ^2 et I_0.

State of the art

Arbitrary choice of H and σ^2
Optimized choice of I_0 :

- one of the images of the dataset
- mean image
- mean in some other space (Younes, Joshi, Glaunes, 2005)
- some first statistical approach (Glasbey and Mardia) but necessity of interpolation, no theoretical consistency proved, not adapted to noisy dataset, no generative statistical model.

Objectives

$$\underbrace{\frac{1}{2}||\varphi||_H}_{\text{deformation cost}} + \underbrace{\frac{1}{2\sigma^2}||I_1 - I_0 \circ \varphi^{-1}||^2}_{\text{data attachment}}$$

$\Longrightarrow$ describing the dataset as some sample of some parametric generative coherent statistical model with parameters :

- template I_0
- noise variance σ^2
- global geometric behaviour in the class quantified by H

$\Longrightarrow$ estimating H, σ^2 and I_0 simultaneously

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

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

$$y_i = I_0 \circ \varphi_i^{-1} + \sigma\varepsilon_i$$

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

Parametric models for the template I_0 and for the deformation φ :

$$y_i = I_0 \circ \varphi_i^{-1} + \sigma \varepsilon_i$$

- I_0 and φ_i defined on the whole plane
- fixed regularity depending on the dataset considered

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

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

with $\varphi = I + z_\beta$.

Model parameters

Photometry : $\begin{cases} \alpha & : \text{ parameter of the template } \mathrm{I}_0 \\ \sigma^2 & : \text{ the noise variance} \end{cases}$

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

Interest in the global geometrical behaviour

$\Longrightarrow$ consider $(\beta_i)_{1 \leq i \leq n}$ as missing random variables and introduce a prior law ν on β and estimate its parameters.

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

Big structures to learn even with small training set

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

where $\nu_g(d\Gamma_g), \nu_p(d\sigma^2, d\alpha)$ are prior laws on the parameters.
Parameters $\theta = (\alpha, \sigma^2, \Gamma_g)$ are estimated by maximum a posteriori

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

Incomplete data framework

We observe data which are related to unobserved data which are of interest or to missing data.
Some examples

* signal deconvolution
* source separation
* geophysic
* pharmacokinetic
* codant/non codant DNA regions
* images matching

Some statistical models with latent variables

* hidden Markov model
* mixed effects model
* frailty model

General latent variable model

Observed data $y \to$ observed variable
Missing data $\phi \to$ latent variable

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

- Compute the value θ^{ML} that maximises the likelihood $g(y; \theta)$ of the observed data
- Estimate the likelihood of the observations $g(y; \theta^{ML})$.
- Estimate the observed Fisher information matrix $-\partial_\theta^2 \log g(y; \theta^{ML})$.

Heuristics : 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), Wu (1983), Vaida (2005)]

Estimation in missing data model

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

Some existing methods

- Methods based on approximations of the likelihood
  - First order methods (FO, Beal and Sheiner, 1982) Used in NONMEM package (very popular in pharmacokinetics), SAS proc NLMIXED (firo option) for example.
  - First order conditional methods (FOCE, Lindstrom and Bates, 1990)
  - Laplace-EM (Vonesh, 1996) also called mode approximation

  No convergence property or with non realistic assumptions, default of convergence.
- Methods based on the exact likelihood
  - MCEM algorithm (Walker, 1996 ; Fort and Moulines, 2004)
  - PXEM algorithm (Liu, Rubin and Wu, 1998)
  - SPML algorithm (Concordet and Nunez, 2002)
  - SAEM algorithm (Delyon, Lavielle and Moulines, 1999)

  Convergence property for some but high computation times and/or non realistic assumptions.

Heuristics of the stochastic approximation

Quantity of interest in the EM algorithm :

$$Q(\theta|\theta') = E[\log f(y, \phi; \theta)|y; \theta']$$

Sequential approximation of this quantity : at iteration k

- simulate ϕ_k
- compute
  $Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\log f(y, \phi_k; \theta) - Q_{k-1}(\theta)\right]$

Then, we have :

$$\frac{Q_k(\theta) - Q_{k-1}(\theta)}{\gamma_k} = E[\log f(y, \phi; \theta)|y; \theta] - Q_{k-1}(\theta) + \log f(y, \phi_k; \theta) - E[\log f(y, \phi; \theta)|y; \theta]$$

$$\frac{Q_k(\theta) - Q_{k-1}(\theta)}{\gamma_k} \approx E[\log f(y, \phi; \theta)|y; \theta] - Q_{k-1}(\theta) + e_k$$

If $\phi_k \sim p(\cdot|y, \theta)$ then $e_k \approx 0$

Stochastic Approximation of the EM algorithm [Delyon et al (1999), K. et al (2004), Allassonnière et al. (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.
- Maximisation step : $\theta_k = \arg\max Q_k(\theta)$

$+$ converges almost surely toward a stationary point $\widehat{\theta}_g$ of g
– theory in exponential model
– nature of the limit point
– convergence depends on the initial guess

One trajectory of the algorithm

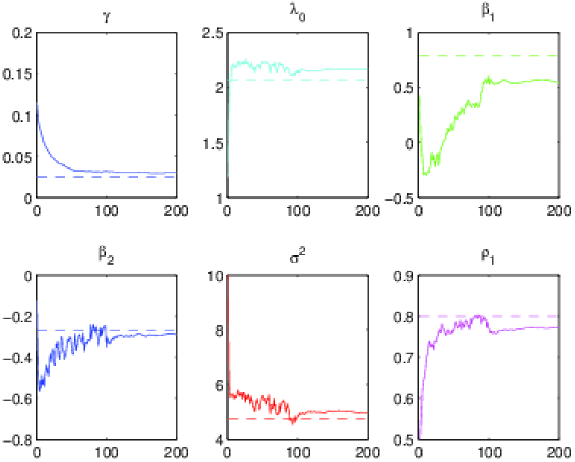


FIGURE: Estimation of the parameters of a correlated frailty model.

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

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

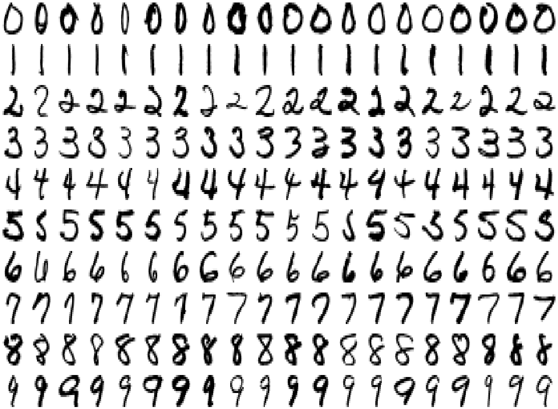
- $X_c | X_k \sim \mathcal{N}\left(X_k + \delta D(X_k), \delta \Sigma(X_k)\right)$ where $\Sigma(x) = \varepsilon Id + D(x) D(x)^T$ with $\varepsilon > 0$.
- compute the acceptance ratio

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

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

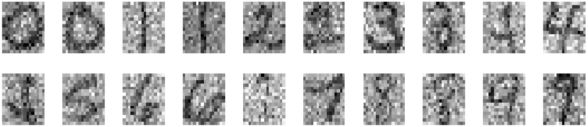
Results : ergodicity of the AMALA chain
convergence and CLT of AMALA-SAEM algorithm

Some images of the US postal database :



In noisy setting :

Examples of data :



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

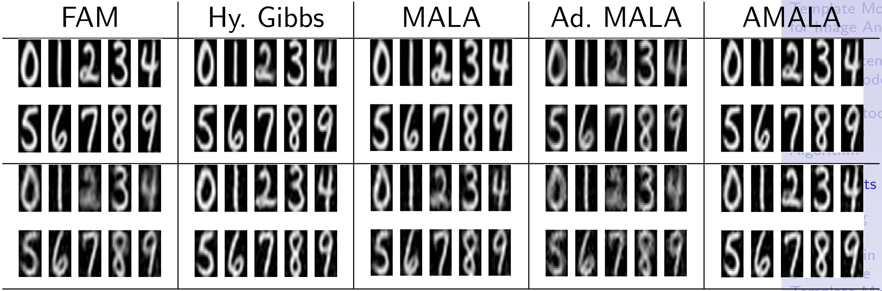


FIGURE: Estimated templates using five algorithms with original data (first line) and noisy data (second line). The training set includes 20 images per digit. The hidden variable are of size $2k_g = 72$.

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

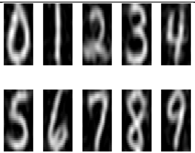
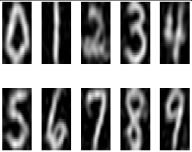
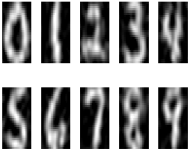
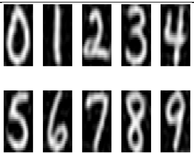
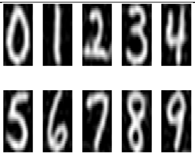
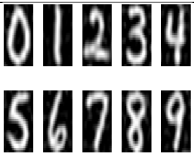
Defor. size/ Samp.	$2k_g = 72$	$2k_g = 128$	$2k_g = 200$
MALA			
AMALA			

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.

Noise variance estimation

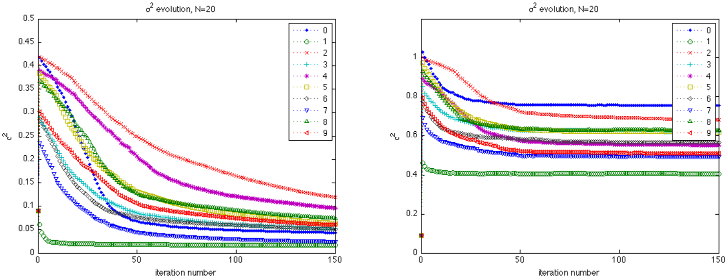


FIGURE: Evolution of the estimation of the noise variance along the AMALA-SAEM iterations. Left : original data. Right : noisy data.

2D Medical Images : Splenium of the Corpus Callossum

Sample of 47 images of the corpus callosum and part of the cerebellum

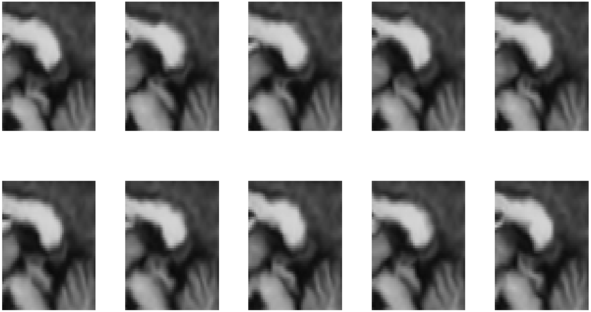


FIGURE: 10 images of the dataset.

Templates estimation

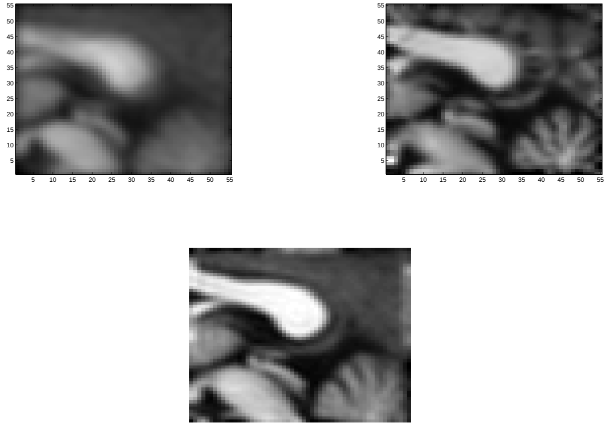


FIGURE: Medical image template estimation. Top row left : mean image. Top row right : Gibbs-hybrid SAEM estimated template. Bottom row : AMALA-SAEM estimated template.

Optimizing geometrical landmark locations in deformable template model ($\mathcal{M}$) [Allassonnière, Durrleman et K. (2015)]

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

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

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

Optimizing the geometrical landmarks in deformable template model

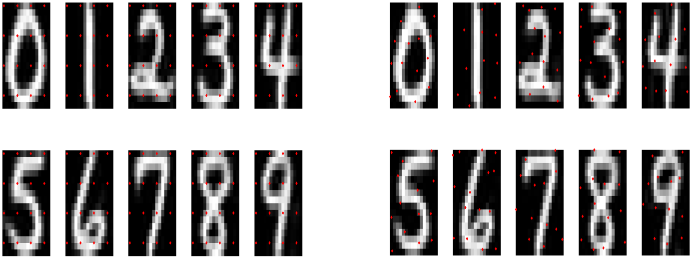


FIGURE: Estimated templates with 16 control points with either fixed (left) or estimated (right) control points positions.

Optimizing the geometrical landmarks in deformable template model

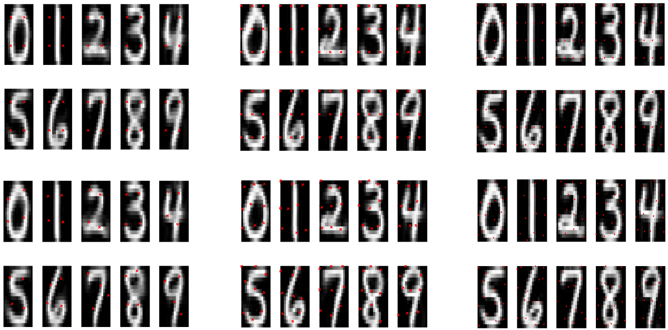


FIGURE: Estimated templates with varying numbers of control points : 4 (left), 9 (middle) and 16 (right), with either fixed (top) or estimated (bottom) control points positions.

Optimizing the number of geometrical landmarks

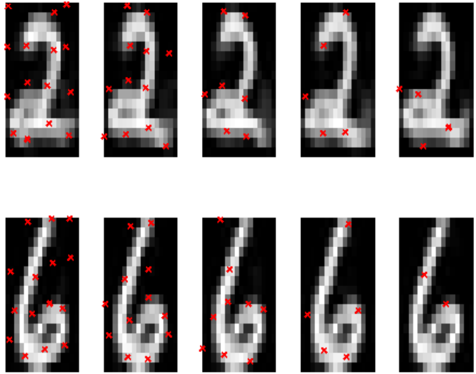


FIGURE: Evolution of the estimated templates and of their number of active control points with respect to the threshold parameter. From left to right : λ equals to $0.3, 0.45, 0.6, 0.75$ and 0.8.

Optimizing the number of geometrical landmarks

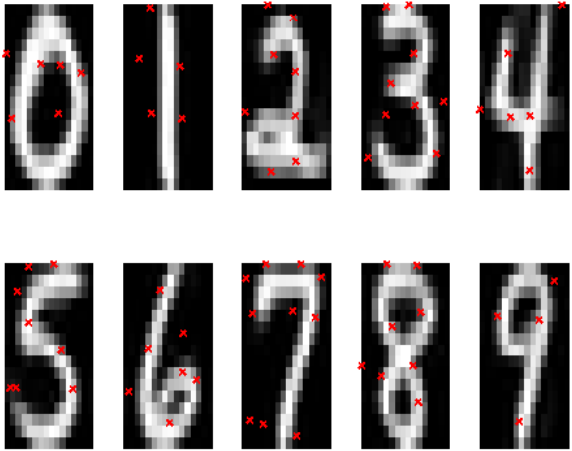


Figure: Estimated templates with their optimal numbers and positions of control points.

2D Medical Images : mouse jawbone

Sample of 36 images of mouse jawbone

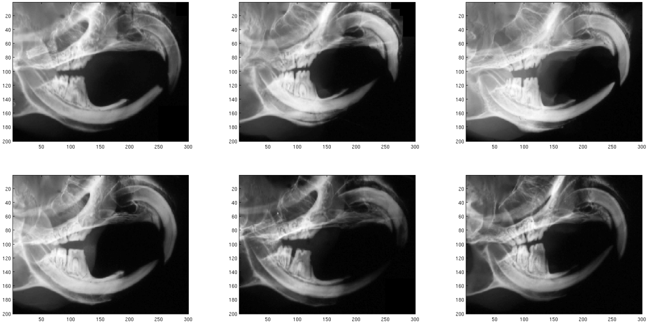


FIGURE: 6 images of the dataset.

Templates estimation

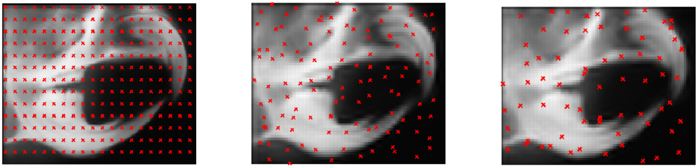


FIGURE: Estimated templates of the mouse mandible images obtained with 260 fixed control points (left), with 117 (middle) and 70 (right) estimated control points.

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)

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

Optimizing geometrical landmarks locations in large deformation model

In large deformation model, initial momenta of the deformation and geometrical landmarks are solution of some coupled Hamiltonian system.
Calculating the gradient of the posterior with respect to the momenta leads for free the one with respect to the geometrical landmarks.
$\Longrightarrow$ considering geometrical landmark as global variables of the model $(\mathcal{M})$ and estimating their optimal location

$$\begin{cases} \dfrac{dr_{g,k}(t)}{dt} = \mathbf{K_g}(r_{g,k}(t))\alpha_k(t) \\ \\ \dfrac{d\alpha_k(t)}{dt} = -\dfrac{1}{2}\nabla_{r_{g,k}(t)}\mathbf{K_g}(\alpha_k(t), \alpha_k(t)) . \end{cases} \tag{1}$$