Shape Invariant Model: a bayesian point of view

Dominique Bontemps
en collaboration avec Sébastien Gadat

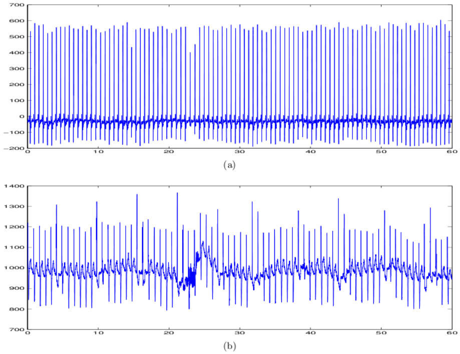
Institut de Mathématiques de Toulouse
Université Paul Sabatier

ANR SPADRO, 9 avril 2014

I - 1 Motivations: biological applications

Problem: We are interested in a situation where some data share a common feature shape.

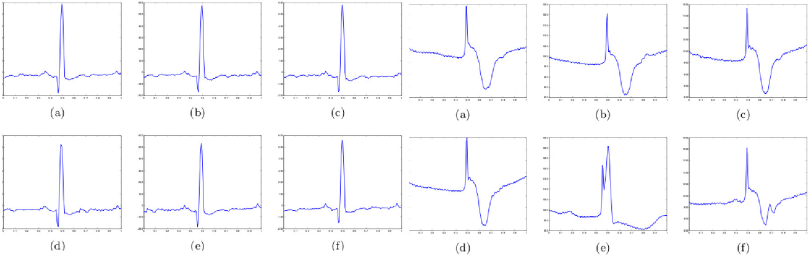
Normal and Arythmic records of an ECG dataset.



Source: *J. Bigot 2013, Fréchet means of curves for signal averaging and application to ECG data analysis, An. App. Stat.*

I - 1 Motivations: biological applications

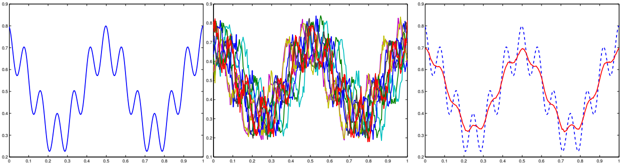
Zoom on the ECG Dataset "



Left: Standard cycles / Right: Arythmic cyles
In each cluster, the signals share a common shape..

I - 1 Motivations: Signal processing

Mean shape estimation ?



Standard approaches fail for deformed signals: $\bar{Y}^n$ is not convincing.

I - 2 Statistical deformable model

Principle: Each signal is a random deformation of a common shape through a group geometrical deformation.

Model: Each observation $Y_m : \Omega \to \mathbb{R},\ m = 1, \dots, n$ is given by

$$Y_m(x) = \mathbf{f} \circ \phi_m(x) + \epsilon W_m(x), \quad x \in \Omega, \text{ where}$$

- $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}$ is the mean shape
- ϕ_m: random deformations (variation around the common shape $\mathbf{f}$).
- W_m is an additive centered noise independent on ϕ_m.
- Noise level ϵ fixed or going to zero.

In this talk: Can we recover $\mathbf{f}$ and $\phi_{\mathbf{m}}$ when $n \to +\infty$?

I - 3 Mathematical statistics

Toy model : Observation of randomly shifted periodic curves $Y_m : [0,1] \to \mathbb{R}$, $m = 1, \ldots, n$ in a white noise model

$$(SIM) \qquad dY_m(x) = f(x - \tau_m)dx + dW_m(x), \text{ where } x \in [0,1] \tag{1}$$

- $f : [0,1] \to \mathbb{R}$ is 1-periodic
- τ_m are i.i.d. random translations whose law is g
- W_m are brownian trajectories independent on the shifts τ_m

Unconsistency of the empirical mean : let $n \to +\infty$

$$\bar{Y}^n(x) = \frac{1}{n}\sum_{m=1}^{n} Y_m(x) \to \mathbb{E}f(x - \tau_1) = \int f(x-\tau)g(\tau)d\tau = f \star g(x) \neq \mathbf{f(x)}$$

Estimation in the SIM when $n \longrightarrow +\infty$ of $f, (\tau_m)_{m=1\ldots n}$ and g.

II Alignment or deconvolution (frequentist)

Results of J. Bigot, S. Gadat, and coauthors.

II-1-a Fréchet empirical mean

$$(SIM) \qquad dY_m(x) = f(x-\tau_m)dx + dW_m(x)$$

- Compute an estimator of τ_j and its inverse to obtain $-\hat{\tau}_j$.
- Align the signals Y_m and take the average:

$$\bar{Y}^n = \frac{1}{n}\sum_{j=1}^{n} Y_j \cdot \hat{g}_j^{-1} = \frac{1}{n}\sum_{j=1}^{n} Y_j(x+\hat{\tau}_j).$$

- Such an estimation is related to a distance associated with G: if $G=[0,1]$ acts on $f \in L^2_{per}([0,1])$: $g_\theta \cdot f(t) = f(t-\theta)$, we define d_G:

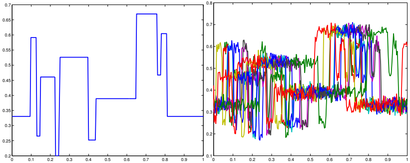
$$d_G^2(Y,h) := \inf_{g\in G} d_E^2(Y, g\cdot h) = \inf_{\tau\in[0,1]} \int_0^1 |Y(x)-h(x-\tau)|^2 dx$$

Fréchet Mean d_G is $\tilde{Y}^n \in \operatorname{argmin}_{h\in\mathcal{H}} \sum_j d_G^2(Y_j, h)$.

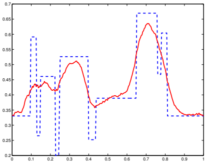
II-1-c 1-Dimension example

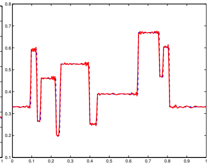
Données :

Signal f / Sample of 10 among $n = 200$



Euclidean mean **Alignment mean**





BG '10

II-1-c 2-Dimension example

Euclidean mean **Alignment mean**

BGL, JMIV, '09

II-2-a Reliability of the Alignment procedures

Model: $dY_j(t) = f(t - \tau_j^*)dt + \epsilon dW_j(t)$, white noise and τ_j^* *i.i.d.* $\sim g$.
Identifiability of the model if we assume that

- (H_g): g is a centered and compactly supported in $\mathbb{T} = [-\frac{1}{4}, \frac{1}{4}]$ distribution.
- (H_f): f satisfies $c_1(f) > 0$

Theorem i) The Frechet empirical mean procedure satisfies

$$\mathbb{P}\left(\frac{1}{n}\sum_{j=2}^{n}(\hat{\tau}_j - \tau_j^*)^2 \geq C(f, \epsilon, n, t, g)\right) \leq 3\exp(-t), \tag{2}$$

BUT $\lim_{n \mapsto +\infty} C(f, \epsilon, n, t, g) > 0$ and $\lim_{n \mapsto +\infty, \epsilon \mapsto 0} C(f, \epsilon, n, t, g) > 0$.
iii) Whatever $(\hat{\tau}_1, \ldots, \hat{\tau}_n)$ estimators of true shifts $(\tau_1^*, \ldots, \tau_n^*)$ are:

$$\mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}(\hat{\tau}_j - \tau_j^*)^2\right) \geq \frac{\epsilon^2}{\sum_{k \in \mathbb{Z}}(2\pi k)^2 |c_k(f)|^2 + \epsilon^2 \int_{\mathbb{T}} \left(\frac{\partial}{\partial \theta}\log g(\tau)\right)^2 g(\tau) d\tau} > 0.$$

iii)

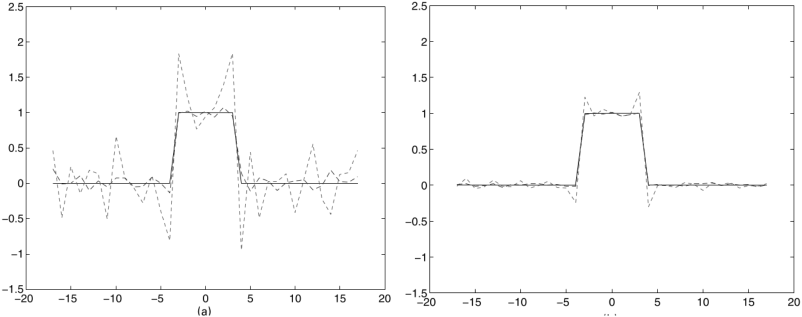
$$\mathbb{E}\|\frac{1}{n}\sum_{j=1}^{n} f(. - \hat{\tau}_j + \tau_j^*) - f\|_2^2 \geq C\epsilon^2 \frac{c_1(f)^2}{\|f'\|_2 + \epsilon^2 I(g)} > 0.$$

Tools: Bernstein's Inequality for i), van Tree's Inequality for ii) and iii).

II-2-b Reliability of the Alignment procedures

Source: [Allassonniere et al, JRSS B'07]

(———): true pattern, (- - - - -) : Warping estimator, (– – –) : other estimator)



Oppositely to the previous Frechet mean experiment, this signal is **slightly more regular**... The Fréchet mean fails.

II-3-a Deconvolution

Hypothesis: We assume that g is known.
Aim : Estimation of f and frequentist minimax risk of the L^2 loss

$$\mathbb{R}_n(\mathbb{F}) = \inf_{\hat{f}_n} \sup_{f \in \mathbb{F}} \mathbb{R}(\hat{f}_n, f), \text{where} \mathbb{R}(\hat{f}_n, f) = \mathbb{E}\|\hat{f}_n - f\|^2$$

Without any deformation: If $\mathbb{F} = H^s(A)$, we have $\mathbb{R}_n(\mathbb{F}) \sim C_A n^{-\frac{2s}{2s+1}}$ **Deconvolution ?** The expectation of each curve is

$$\mathbb{E}\left[f(x - \tau_j)\right] = \int_{\mathbb{R}} f(x - \tau) g(\tau) d\tau = f \star g(x)$$

and the empirical averaging is

$$dY(x) = f \star g(x)dx + \underbrace{\xi(x)dx}_{\text{Non Gaussian}} + \underbrace{\frac{\epsilon}{\sqrt{n}} dW(x)}_{\text{Gaussan}}, \; x \in [0,1].$$

II-3-b Deconvolution estimation

On the Sobolev ball

$$H_s(A) = \left\{ f \in L^2([0,1]) \; ; \sum_{\ell \in \mathbb{Z}} (1 + |\ell|^{2s}) |\theta_\ell|^2 \leq A, \right\} \text{ avec } A > 0, s > 0$$

Theorem (BG, A.O.S. '10)

i) We can build an adaptive (in s) frequentist estimator with Meyer wavelets such that

$$\sup_{f \in H_s(A)} \mathbb{R}(\hat{f}_n, f) = \mathrm{O}(n^{-\frac{2s}{2s+2\nu+1}} \log n)$$

ii) Moreover, if $s > \nu + 1/2 > 1$, one has

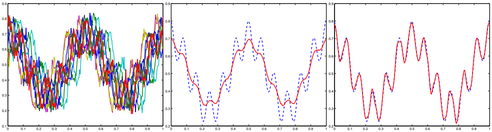
$$\lim_{n \to +\infty} n^{\frac{2s}{2s+2\nu+1}} \mathbb{R}_n(H_s(A)) \geq C(A, s).$$

The L^2 minimal loss is optimal. The least favorable case is the "uniform" law for $g \ldots$. Tools:

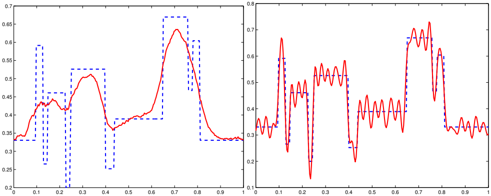
- i) Concentration / Hard Thresholding
- ii) Girsanov's formula & Assouad's Lemma with very annoying computations (!)

II - 3 - c Numerical deconvolution

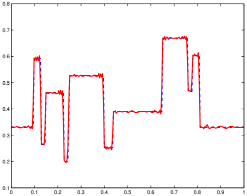
Samples / Empirical mean / Deconvolution



Empirical mean **Deconvolution**



Empirical Fréchet mean



III - 1 - a Bayesian approach: "last chance"?

- **Model:**
$$dY_j(x) = f^0(x - \tau_j)dx + \epsilon dW_j(x),$$
where (τ_j) $i.i.d. \sim g^0$ and f^0, g^0 are unknown and belong to non parametric space.
- **Bayesian approach:** Define a prior law π_1 on f^0 and π_2 on g^0, use the posterior to build estimators of f^0 and g^0.
$$\pi := (\pi_1 \otimes \pi_2) \text{ et } \pi_n := \pi[\quad . \quad |Y_1^n]$$
- **Frequentists questions:**
Question 1: When $n \longmapsto +\infty$, $\pi_n \longmapsto \mathbb{P}_{f^0,g^0}$?
Question 2: What is the contraction rate around $\mathbb{P}_{f^0,g^0}$?
Question 3: Results related to the functional objects themselves (L^2 metric)?

III - 1 - b Bayesian approach: Mixture description

Few notations:

- $\mathcal{H}^{\ell}(A)$ truncated Sobolev ball of maximal frequency ℓ and radius $\|.\|_2 \leq A$.
- $\mathfrak{M}$: set of probability measures on $[0,1]$.
- **Sieve:** $\mathcal{P}_{l,w} := \left\{\mathbb{P}_{f,g} \quad \text{s.t.} \quad f \in \mathcal{H}^{l}(w),\, g \in \mathfrak{M}\right\}$

In the Fourier domain: if $\theta^0 = \theta^0(f) := (\theta^0_{-\infty}, \ldots \theta^0_{-1}, \theta^0_0, \theta^0_1, \ldots, \theta^0_{-\infty})$ and

$$c_{\ell} \quad = \quad \theta^0_{\ell} e^{-i2\pi\ell\tau} + z_{\ell} \text{ where } z_{\ell} \sim_{i.i.d.} N_{\mathbb{C}}(0,1)\,, \qquad \tau \sim g^0,$$

We have the multivariate mixture model

$$\mathbb{P}_{\theta^0, g^0} = \int_0^1 \gamma_{\theta^0 \bullet \tau} d g^0(\tau).$$

$\gamma_{\theta \bullet \tau}$ is an infinite complex standards gaussian law whose mean is

$$\forall \ell \in \mathbb{Z} \qquad (\theta \bullet \tau)_{\ell} = \theta e^{-i2\pi\tau\ell}.$$

Proposition: For any ϵ small enough, and $w_{\epsilon} \leq \sqrt{\ell_{\epsilon}}$

$$\log N(\epsilon, \mathcal{P}_{\ell_{\epsilon}, w_{\epsilon}}, d_{VT}) \lesssim \ell_{\epsilon}^2 \left[\log \frac{1}{\epsilon} + \log \ell_{\epsilon}\right].$$

We loose a term ℓ_{ϵ} (unknown f) $\ell_{\epsilon}^2 \log \frac{1}{\epsilon} \geq \ell_{\epsilon} \log \frac{1}{\epsilon}$ (covering of mixture models dim l_{ϵ}).
Tool: Follow the covering strategy of [GvdV,01] on gaussian mixture models and use χ^2_k concentration, $k \mapsto +\infty$.

III - 2 - a Building the prior (general mixtures)

Prior built through a tensorial product of a prior on f and a prior on g.

- Pick ℓ_{max} such that $p(\ell_{max} = k) \propto e^{-k^2 \log k}$ as a threshold frequency [RR,12]
- Each active coefficient follows a $\mathcal{N}_{\mathbb{C}}(0, \xi_n^2)$. [RR,12]
- Dirichlet process $D(\alpha)$ as a prior on g.

Theorem:[BGI] If $f^0 \in \mathcal{H}_s$ with $s \geq 1$, then for $\epsilon_n = n^{-[s/(2s+2) \wedge 3/8]} \log n$:

$$\exists M > 0 \qquad \pi_n \left\{ \mathbb{P}_{f,g} \quad t.q.\, d_H(\mathbb{P}_{f,g}, \mathbb{P}_{f^0,g^0}) \leq M\epsilon_n \right\} = 1 + o_p(1)$$

Comments:

- Polynomial rate.
- The supplementary $\ell_\epsilon^2 \log \frac{1}{\epsilon}$ impacts the rate $2s/(2s+2)$ instead of $2s/(2s+1)$, we must estimate g^0...
- Adaptive prior on s

III - 2 - b Building the prior (smooth mixtures)

Same prior on f, Gaussian process for g [vdVvZ,08]:

- $\mathfrak{M}_\nu([0,1])(B) := \left\{ g \in \mathfrak{M} \,|\, \exists\forall k \in \mathbb{Z} \quad c|k|^{-\nu} < |c_k(g)| < C|k|^{-\nu} \right\}$ with $\|g\|_\nu \leq B$
- For any continuous trajectory f on $[0,1]$ define J and $J_k = J_{k-1} \circ J$ as

$$J(f)(t) := \int_0^t f(s)ds - t \int_0^t f(u)du$$

- B a brownian bridge on $[0,1]$, build

$$w = \underbrace{J_{k_\nu}(B)}_{k_\nu \text{ regularization of } \textit{periodized} \text{ B}} + \overbrace{\sum_{j=1}^{k_\nu} Z_j \psi_j}^{\text{Start and end at random on } \{0,1\}},$$

where $\psi_j(t) = \cos 2\pi jt + \sin 2\pi jt$ and $(Z_j)_{1 \leq j \leq k_\nu} \sim \mathcal{N}_{\mathbb{R}}(0,1)$.

- The prior samples the distributions with density g_w: $g_w = e^w \left(\int_0^1 e^w \right)^{-1}$.
- Following arguments of [vdVvZ,08] & [LS01] to obtain the small ball property.

Théorème:[BGII] If $f^0 \in \mathcal{H}_s$, $g^0 \in \mathfrak{M}_\nu([0,1])$ and $\epsilon_n = n^{-[s/(2s+2) \wedge \nu/(2\nu+2) \wedge 3/8]} \log n$:

$$\pi_n \left\{ \mathbb{P}_{f,g} \quad t.q.\ d_H(\mathbb{P}_{f,g}, \mathbb{P}_{f^0,g^0}) \leq M\epsilon_n \right\} = 1 + o_p(1)$$

Comments:

- Polynomial rate but depends on ν.
- Adaptive with s **but now with** ν.

III - 3 Bayesian consistency around f^0 and g^0 (smooth mixtures)

- **Identifiability condition:** $\mathcal{F}_s = \{f \in \mathcal{H}_s \text{ s.t. } c_1(f) > 0\}$
- Theorem: If $(f, g) \in \mathcal{F}_s(A) \times \mathfrak{M}_\nu([0,1])$, the model is identifiable.
  Tool: Show $d_{VT}(\mathbb{P}^1_{f,g}, \mathbb{P}^1_{\tilde{f},\tilde{g}}) > 0 \Rightarrow c_1(f) = c_1(\tilde{f})$ and use a Laplace transform argument.
- **Result on elements in $\mathcal{F}_s(A)$ or $\mathfrak{M}_\nu([0,1])$?**
  Theorem: If $(f^0, g^0) \in \mathcal{F}_s(A) \times \mathfrak{M}_\nu([0,1])$, then (similar result for g^0)

$$\Pi_n\left(f : \|f - f^0\|_2^2 > M(\log n)^{-\frac{4s\nu}{2s+2\nu+1}} \,|Y_1, \dots, Y_n\right) = o_p(1)$$

- Theorem: From a Minimax point of view, we can show

$$\liminf_{n \longrightarrow +\infty} (\log n)^{2s+2} \inf_{\hat{f} \in \mathcal{F}_s(A)} \sup_{(f^0, g^0) \in \mathcal{F}_s(A) \times \mathfrak{M}_\nu([0,1])} \|\hat{f} - f^0\|_2^2 \geq c,$$

and

$$\liminf_{n \longrightarrow +\infty} (\log n)^{2\nu+1} \inf_{\hat{g} \in \mathcal{F}_s(A)} \sup_{(f^0, g^0) \in \mathcal{F}_s(A) \times \mathfrak{M}_\nu([0,1])} \|\hat{g} - g^0\|_2^2 \geq c.$$

Tool: Fano's Lemma applied on a very very anoying networks of $(f_j, g_j)_j$ such that

$$\forall j \neq j' \qquad f_j \star g_j \simeq f'_j \star g'_j.$$

III - 4 - a - Bayesian end of the story?

- The logarithm loss is very disapointing ...

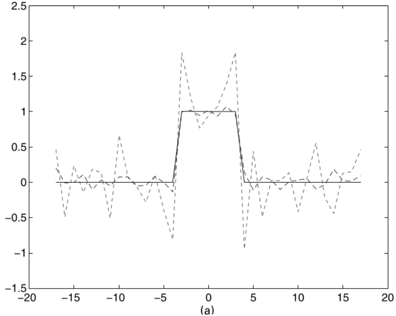
BUT

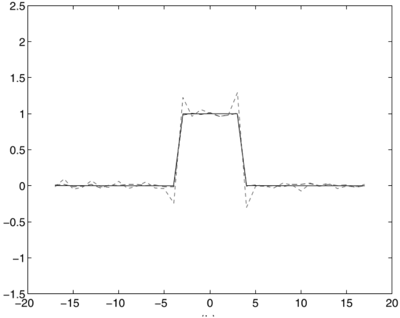
- **Did we use the right distance to measure the performance of the estimators?**
- Recall the Fréchet distance (on orbits defined through the action of translations)

$$d_F(f_1, f_2)^2 = \inf_\tau \|f_1^{-\tau} - f_2\|_2^2.$$

- Source: [Allassonniere et al, JRSS B'07]

(——): true pattern, (- - - - -) : Warping estimator, (– – –) : Posterior realisation)





IV Conclusion

Important remarks:

- Inverse problem with unknown (or not) operator (depending on the knowledge on g).
- The smoothness of g is important.

Criticisms:

- Optimality of NP Bayesian rates with respect to the Fréchet distance?
- Efficient algorithms to compute the posterior distribution?

Extensions:

- More realistic model: σ unknown, n signals, J points (BG'14?).

Bibliography

[1] Allassonnière, S. & Kuhn, E. & Trouvé, A. (2010) Construction of Bayesian deformable models via stochastic approximation algorithm: A convergence study. Bernoulli Journal, 16 (3), p.641-678.

[2] Bhattacharya, R. & Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds (i). Annals of Statistics, 31(1):1-29.

[3] Bigot, J., Gadat, S. , Klein, T. & Marteau, C. (2013) Intensity estimation of non-homogeneous Poisson processes from shifted trajectories, Electronic Journal of Statistics.

[4] Bigot, J. & Gadat, S. (2010) A deconvolution approach to estimation of a common shape in a shifted curves model, Annals of Statistics, 38(4), 2422-2464.

[5] Bigot, J. & Charlier B. (2011) On the consistency of Fréhet means in deformable models for curve and image analysis, Electronic Journal of Statistics, 5, 1054-1089.

[6] Bigot, J., Christophe C. & Gadat, S. (2012) Random action of compact Lie groups and minimax estimation of a mean pattern, IEEE Transactions on Information Theory, 58(5) .

[7] Bigot, J. & Gendre, X. (2013) Minimax properties of Fréchet means of discretely sampled curves, Annals of Statistics.

[8] Bontemps, D. & Gadat, S. (2012) Bayesian consistency in the non parametric randomly shifted curves model (I), (II), preprint.

[9] Bontemps, D. & Gadat, S. (2013) Bayesian consistency in the non parametric randomly shifted curves model (III), in preparation.

[10] Ghosal, S. & Ghosh, J. & van der Vaart, A. Convergence rates of posterior distributions. Ann. Statist. 28 (2000), no. 2, 500-531.

[11] Ghosal, S. & van der Vaart, A. Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann. Statist. 35 (2007), no. 2, 697-723.

[12] van der Vaart, A. & van Zanten, H. Rates of contraction of posterior distributions based on Gaussian process priors. Ann. Statist. 36 (2008), no. 3, 1435-1463.