Functional linear spatial autoregressive modeling

Sophie Dabo-Niang
LEM CNRS 9221, University of Lille, INRIA-MODAL

7th Statistical Meeting of Avignon-Marseille, 14/06/2019

Outline

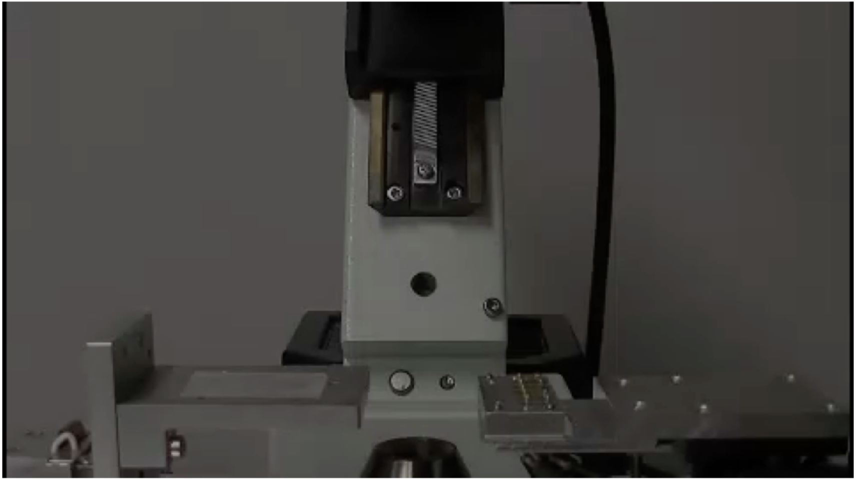
- Introduction : from functional binary discrete models to SAR model
- Functional linear spatial autoregressive (Functional SAR)
- Numerical experiments

Introduction : from functional binary discrete models to SAR model

Motivation

- Developing a model by comprehensive analysis of physical and biological characterization with existing genetic information,
- To predict metastatic potential or migration of a cell by means of physical properties
- Using BioMEMS : MEMS (micro-electro-mechanical systems) applied to biological and/or biomedical.

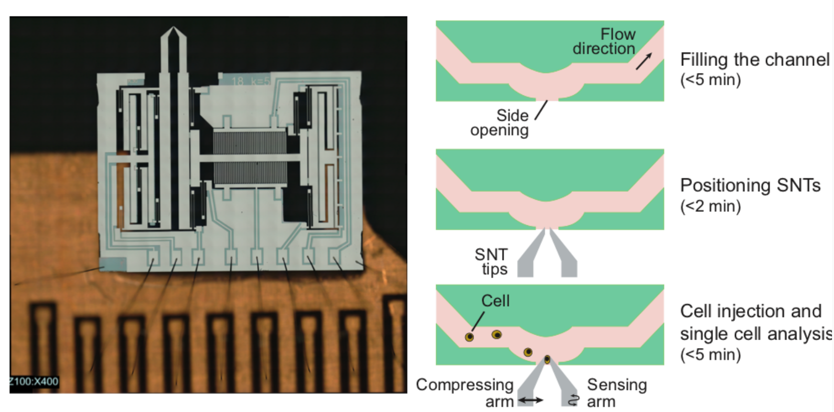
Yesterday and Tomorrow : Physical Single Cell manipulation



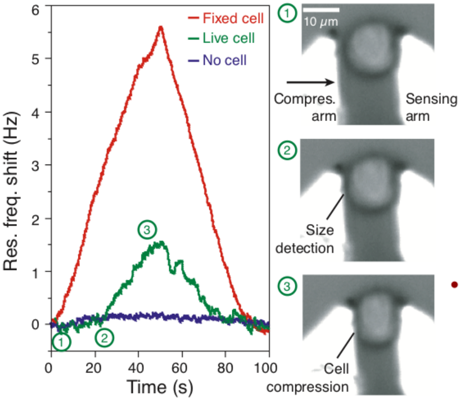
Cell physical/biological properties

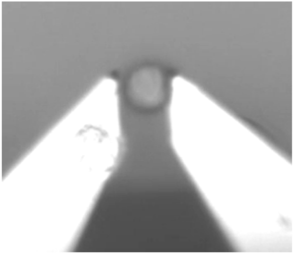
- Cell physical properties : described by her rigidity, viscous losses, membrane capacitance, cytoplasm conductivity, shape recovery and size
- Cell biological properties : migration capability, invasion capacity and metastatic potential (in physiologically relevant conditions),...

Physical properties : Experimental steps



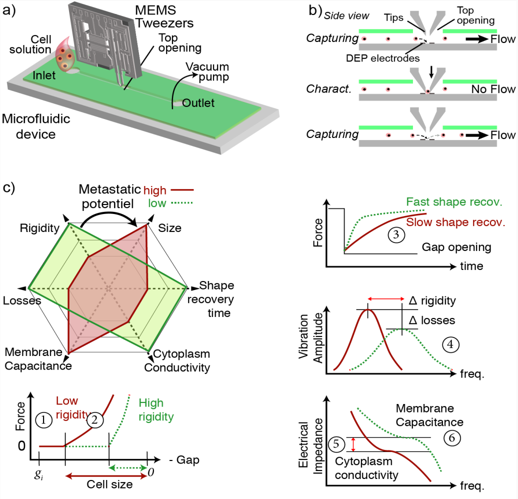
Cell compression



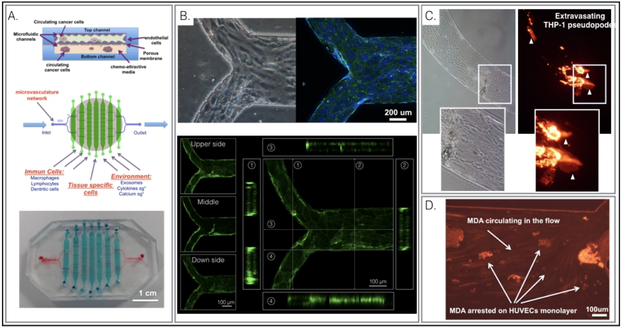


- Difference between live and fixed cells
  - 100 s period
  - 0-50 V cycles (corresponding to ~4.5 μm change in the gap)

Genetic cell data : physical characterization



Genetic cell data : biological characterization



Choice-Based or Case-Control Sampling

- Prospective studies are disadvantageous in terms of time and cost
- Use existing records to identify people with a certain health problem (cases) and a similar group without the problem (controls) and observe there characteristics to some risk factors.
- Oversample a rare disease of interest to increase the accuracy of an analysis.

Binary Functional Choice Model for sampling data

- In a given population, assume that one observe $Y \in \{0, 1\}$ and $X \in \mathcal{X} \subset L^2(\mathcal{T})$, with $\mathcal{T} \subset \mathbb{R}$.
- Let the parameters of interest $\alpha^* \in \mathbb{R}$ and $\theta^*(\cdot) \in L^2(\mathcal{T})$:

$$P\left(Y = 1 | X, \alpha^*, \theta^*(\cdot)\right) \equiv E\left(Y | X\right) = \Phi\left(\alpha^* + \int_{\mathcal{T}} X(t)\theta^*(t)dt\right)$$

  with $\Phi(\cdot)$ is some cumulative distribution function strictly increasing and $E\left(X(t)\right) = 0, \ \forall t \in \mathcal{T}$.
- The population is divided according to the values of Y into two stratum $\mathcal{J}(0) = \{(0, X), \ X \in \mathcal{X}\}$ and $\mathcal{J}(1) = \{(1, X), \ X \in \mathcal{X}\}$.
- Let $Q(i) = P\left(Y = i\right)$ be the population share of choice i, $Q^* = Q(1)$.

Existing works on functional sampling data

- Functional PCA for sampling data : Cardot et al. (2010),...
- Functional mean and variance estimates for sampling data : Cardot et al. (2011),...
- Functional logit model applied to case-control to test association between a dichotomous trait and multiple genetic variants in a region : Fan et al. (2014).
- Functional logit model applied to case-control head circumference growth data : Petrovich et al. (2018)
- Non-functional Binary Choice Models under sampling data : Manski and McFadden (1981), Cosslett (1981, 2013), Imbens (1992),...

- Let $H(i)$ be the probability according to which the stratum i is drawn and let $H^* = H(1)$.
- First draw a stratum i with a probability $H(i)$. Then, select an observation (i, X) randomly from $\mathcal{J}(i)$.
  The conditional density of Y given $X = x$ is

$$g(i|x) = \frac{P\left(i|x, \alpha^*, \theta^*(\cdot)\right) H(i)/Q(i)}{\sum_{j=0}^{1} P\left(j|x, \alpha^*, \theta^*(\cdot)\right) H(j)/Q(j)}, \quad x \in \mathcal{X},\ i \in \{0, 1\}$$

- Let E_s denote the expectation with respect to the CBS (choice-based or case-control)

$$E_s(\cdot) = H(0)E(\cdot|Y=0) + H(1)E(\cdot|Y=1).$$

  When $H^* = Q^*$ we have $E(\cdot) = E_s(\cdot)$.
- Assume that we have a prior information, allowing knowledge on Q^* and H^*.

Infeasible maximum likelihood

Let $\left(Y_n = i_n, \{X_n(t),\ t \in \mathcal{T}\}\right),\ n = 1, \ldots, N$ be independent observations drawn through the CBS process.

The conditional log likelihood :

$$L(\alpha, \theta(\cdot)) = \sum_{n=1}^{N} \log \left(\frac{P\left(Y = i_n | X_n, \alpha, \theta(\cdot)\right) H(i_n)/Q(i_n)}{\sum_{j=0}^{1} P\left(Y = j | X_n, \alpha, \theta(\cdot)\right) H(j)/Q(j)} \right) \quad (1)$$

Objective : estimate α^* and $\theta^*(\cdot)$ by maximizing $L(.,.)$.

- Truncated Likelihood Method (Müller and Stadtmüller (2005)).

Truncated Likelihood Method

- Let p_N be a positive sequence of integers, increasing as $N \to \infty$.
- Let Γ_s denotes the covariance operator of the $\mathcal{X}-$valued random function under the CBS, defined as

$$\Gamma_s x(t) = \int_{\mathcal{T}} K(t, v) x(v) dv, \; x \in \mathcal{X}, \; t \in \mathcal{T}.$$

  with the kernel $K(t, v) = E_s(X(t)X(v))$.
  Γ_s is a compact self-adjoint Hilbert-schmidt operator.
- Let $\{\varphi_j, j = 1, 2, \dots\}$ be the orthonormal eigen-basis of Γ_s and λ_j the associated eigenvalues.

Truncated Likelihood Method

X and $\theta^*(\cdot)$ can be expanded as follow

$$X(t) = \sum_{j \geq 1} \varepsilon_j \varphi_j(t), \qquad \theta^*(t) = \sum_{j \geq 1} \theta_j^* \varphi_j(t)$$

with

$$\varepsilon_j = \int_{\mathcal{T}} X(t)\varphi_j(t)dt, \quad \text{and} \quad \theta_j^* = \int_{\mathcal{T}} \theta^*(t)\varphi_j(t)dt$$

then

$$\int_{\mathcal{T}} X(t)\theta^*(t)dt = \sum_{j \geq 1} \theta_j^* \varepsilon_j.$$

Consider the decomposition

$$U_{p_N} = \alpha^* + \sum_{j=1}^{p_N} \theta_j^* \varepsilon_j, \qquad V_{p_N} = \sum_{j=p_N+1}^{\infty} \theta_j^* \varepsilon_j.$$

Under the CBS :

$$E_s\left(Y|X\right) = g\left(1|X\right) \equiv \mu\left(U_{p_N} + V_{p_N}\right)$$

where

$$\mu(t) = \frac{\Phi(t)H(1)/Q(1)}{\Phi(t)(H(1)/Q(1)) + (1-\Phi(t))H(0)/Q(0)}.$$

- Let $\tilde{\theta} = (\theta_0^*, \theta_1^*, \ldots, \theta_{p_N}^*)^T$ with $\theta_0^* = \alpha^*$ and $\theta_1^*, \ldots, \theta_{p_N}^*$ the p_N first coefficients of $\theta^*(\cdot)$ and let $\varepsilon_0 = 1$.

Estimation Procedure

The truncated conditional likelihood function is defined for $\theta \in \mathbb{R}^{p_N+1}$:

$$\tilde{L}_{p_N}(\theta) \quad = \quad \sum_{n=1}^{N} Y_n \log \mu\left(\eta_n\right) + \left(1 - Y_n\right) \log\left(1 - \mu\left(\eta_n\right)\right) \tag{2}$$

where $\eta_n = \sum_{j=0}^{p_N} \theta_j \varepsilon_j^{(n)}$ with $\varepsilon_j^{(n)} = \int X_n(t)\varphi_j(t)dt,\ \varepsilon_0^{(n)} = 1$.

Then $\tilde{\theta}$ is estimated by

$$\hat{\theta} = \operatorname{argmax}\left\{\tilde{L}_{p_N}(\theta),\ \theta \in \mathbb{R}^{p_N+1}\right\}$$

So $\hat{\alpha}^* = \hat{\theta}_0$ and the estimate of $\theta^*(\cdot)$ is

$$\hat{\theta}(t) = \sum_{j=1}^{p_N} \hat{\theta}_j \varphi_j(t).$$

Asymptotic Results

Under appropriate assumptions (identification,...) :

$$\frac{N(\hat{\theta} - \tilde{\theta})^T \Delta_{p_N} (\hat{\theta} - \tilde{\theta}) - (p_N + 1)}{\sqrt{2(p_N + 1)}} \to \mathcal{N}(0, 1)$$

$p_N \to \infty$, $N^{-1/4} p_N \to 0$.

Asymptotic Results

Let $\{\varphi_j^G, j = 1, 2 \ldots\}$ denotes the eigen-basis of the operator associated to the kernel G defined by

$$G(t, v) = E_s\left(\frac{\mu'^2(\eta)}{\sigma^2(\mu(\eta))}X(t)X(v)\right),\ t, v \in \mathcal{T}$$

$\sigma^2(t) = t(1-t)$.

Considering the metric in $L^2(\mathcal{T})$ defined by

$$d_G^2\left(f, g\right) = \int\int\left(f(t) - g(t)\right)G(t, v)\left(f(v) - g(v)\right)dtdv,\ f, g \in L^2(\mathcal{T}).$$

As $N \to \infty$ we have,

$$\frac{Nd_G^2\left(\hat{\theta}(\cdot), \theta^*(\cdot)\right) - p_N}{\sqrt{2p_N}} \to \mathcal{N}(0, 1).$$

Biased OML intercept estimate : logit case

Let the conditional log-likelihood function when the sampling scheme is ignored

$$\mathcal{L}(\alpha, \theta(\cdot); Y, x) = Y \log(\Phi(\eta(\alpha, \theta))) + (1 - Y) \log(1 - \Phi(\eta(\alpha, \theta))).$$

Let $\mathcal{F}(\alpha, \theta(\cdot); x)$ be (under the true parameters α^* and $\theta^*(\cdot)$)) :

$$\mathcal{F}(\alpha, \theta(\cdot); x) = E_s\left(\mathcal{L}(\alpha, \theta; Y, x) | \alpha^*, \theta^*(\cdot)\right).$$

Let $\delta = \log(H^*/Q^*)$ and $\bar{\delta} = \log((1 - H^*)/(1 - Q^*))$; by use of the definition of $E_s(\cdot)$,

$$\begin{aligned} \mathcal{F}(\alpha, \theta(\cdot); x) &= \frac{1 + e^{\eta(\alpha^* + \delta, \theta^*)}}{1 + e^{\eta(\alpha^* + \delta - \bar{\delta}, \theta^*)}} \times \\ & E\left(\mathcal{L}(\alpha, \theta(\cdot); Y, x) | \alpha^* + (\delta - \bar{\delta}), \theta^*(\cdot)\right). \end{aligned}$$

Biased OML intercept estimate : logit case

- Consistent OML estimate $\hat{\theta}_{\mathrm{RS}}(\cdot)$ of $\theta(\cdot)$ and a biased estimate of the intercept, with (in the logit case) bias

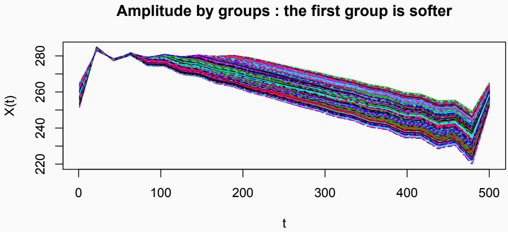
$$\log((1 - Q^*)H^*/(1 - H^*)Q^*).$$

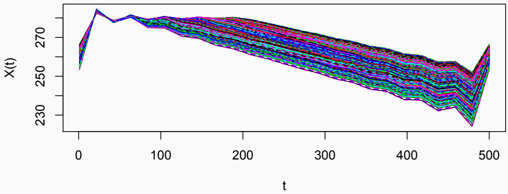
- $\hat{\theta}_{\mathrm{RS}}$ is not asymptotic normal if $Q^* \neq H^*$.

Functional clustering of physical fake data by BioMEMS technology

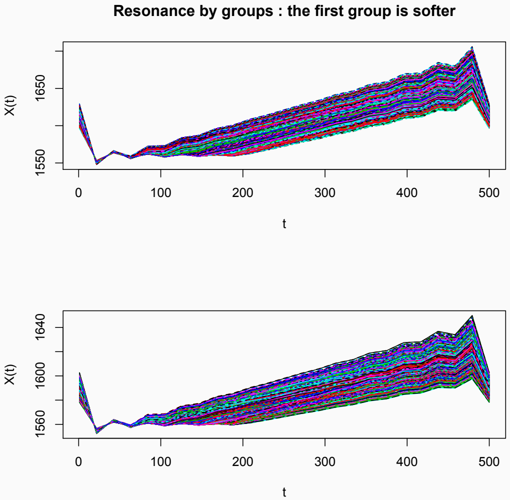
- We have both softer (agressive/metastatic) and less-soft (non-metastatic) cell lines using continuous monitoring (500 monitoring points) of cells' physical properties (size and stiffness from frequency resonance, viscosity from the amplitude data) under different mechanical cell stimulation
- Namely, resonance frequency and amplitude monitoring data of two groups (metastatic, non-metastatic) of cells (1000 cells per group) are analyzed.
- Functional clustering methods to recognize cell status (aggressiveness).

Amplitude


Amplitude by groups : the first group is softer
X(t)
280
260
240
220
0
100
200
300
400
500
t

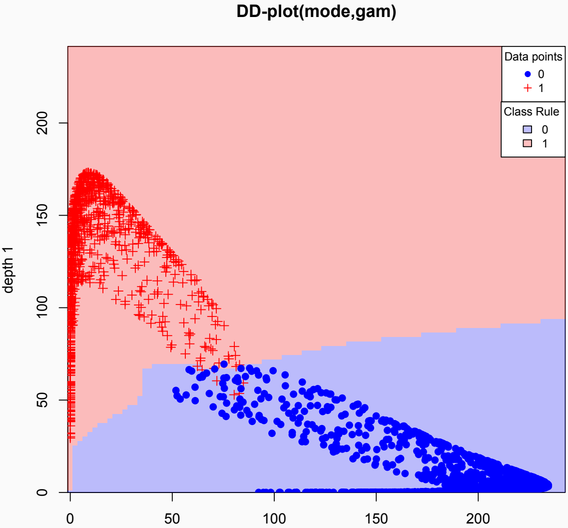

X(t)
270
250
230
0
100
200
300
400
500
t

Resonance



Resonance by groups : the first group is softer
X(t)
1550
1650
0
100
200
300
400
500
t
X(t)
1560
1600
1640
0
100
200
300
400
500
t

Plot classification/prediction analyses using FPCA and GAM showing very good classification of metastatic characteristics (blue for metastatic cell, red for non-metastatic)



Functional linear spatial autoregressive models

Spatial lattice data

"*Economic activities and epidemiological data are often located in space*",

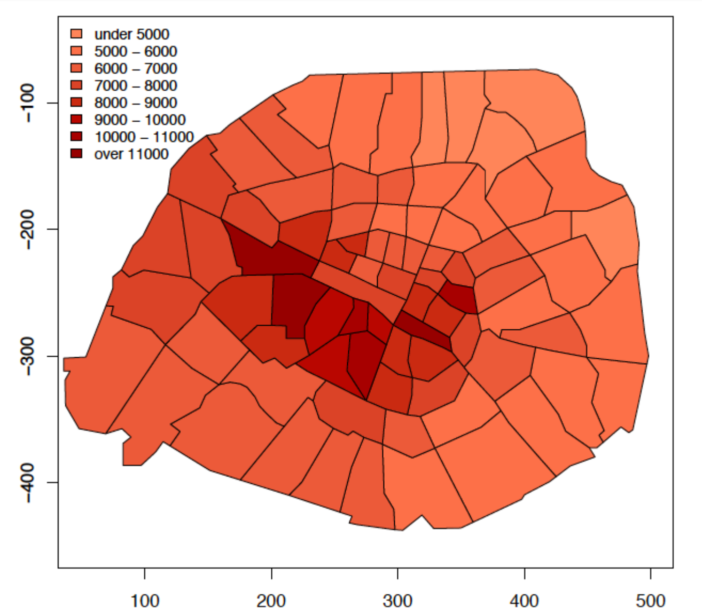
Pinkse and Slade [1998]

Let a spatial set $\mathcal{S} \subset \mathbb{R}^2$. Locations are in general geographical units (zip codes, regions,...) or structured sites in a neighbor network.

They are linked to a large number of models :

- auto-regressive models : spatial auto-correlation : a location i depends on some neighbor locations j.
- Markov random fields used in images analyses
- ...

Price of houses (m2) in Paris (2009)



Neighbors

Let $\mathcal{S} = (s_i)_{i=1,n}$.

Neighbor graph : binary relation on $\mathcal{S} \times \mathcal{S}$: s_i is neighbor of s_j.
Neighbor matrix : weigth matrix

$$w_{ij} = \begin{cases} 1 & \text{if } s_i \text{ and } s_j \text{ are neighbors} \\ 0 & \text{if } s_i \text{ and } s_j \text{ are not neighbors or } s_i = s_j \end{cases}$$

or

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{si } s_i \text{ and } s_j \text{ are neighbors } d_{ij} = \mathrm{dist}(s_i, s_j) \\ 0 & \text{if } s_i \text{ and } s_j \text{ are not neighbors or } s_i = s_j \end{cases}$$

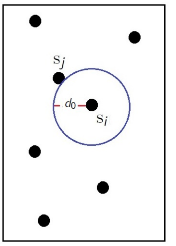
Spatial econometric models

Three different types of spatial interaction effects :

- Endogenous interaction effects among response $\mathbf{Y}$.
- Exogenous interaction effects among predictors $\mathbf{X}$.
- Interaction effects among the disturbances terms ϵ.

Spatial econometric models

- Sample of n observations collected from points or regions in sites $s_i \in \mathbb{R}^k;\ i = 1, \ldots, n$ on $\mathcal{S} \subset \mathbb{R}^k,\ k \geq 2$
- $\forall s_i, s_j \in \mathcal{S} : ||s_i - s_j|| \geq d_0$
- We observe $(X_{s_i}, Y_{s_i});\ i = 1, \ldots, n$
- $\mathbf{Y}$ be the vector of responses.



Spatial econometric models

The full model (Manski model) :

$$\boxed{\begin{array}{c} \mathbf{Y} = \rho W \mathbf{Y} + \mathbf{X}\beta + W\mathbf{X}\eta + u, \\ u = \lambda M u + \epsilon \;\; \texttt{;} \;\; \epsilon \;\texttt{i.i.d.} \sim (0, \sigma_\epsilon^2 I_n), \end{array}}$$

Special cases :

- SLM ($\lambda = 0$ and $\eta = 0$) :
  $\mathbf{Y} = (I_n - \rho W)^{-1}(\mathbf{X}\beta + \epsilon)$
- SEM ($\rho = 0$ and $\eta = 0$) :
  $\mathbf{Y} = \mathbf{X}\beta + (I_n - \lambda M)^{-1}\epsilon$
- Kelejian-Prucha (general) model ($\eta = 0$) :
  $\mathbf{Y} = (I_n - \rho W)^{-1}\mathbf{X}\beta + (I_n - \rho W)^{-1}(I_n - \lambda M)^{-1}\epsilon$
- Error terms no longer independent, violation of Gauss-Markov assumption, then OLS does not works

Functional lattice data

- In many applied domains, spatially correlated functional data are available : economic, environmental, hydrology, ...

  Example : curves of daily concentration of ozone at near municipalities

- Some works are developed to deal with spatially correlated functional data
  - Functional geostatistical data :

    Kriging methods : Bohorquez et al. (2016), Giraldo et al. (2010),...

    Nonparametric regression : Dabo et al. (2011), Ternynck (2014),...
  - Lattice functional data : less developed

    Ruiz-Medina (2012) : prediction of SAR hilbertian processes

    Pineda-Ríos and Giraldo (2016) : FLMs with SAR disturbance process

$\hookrightarrow$ SAR models with functional covariates :
Ahmed et. al. (2017), Huang et al. (2018)

Model

- At n spatial units, we observe a covariate function $\{X(t), t \in \mathcal{T}\}$ and a real-valued variable Y.
- $W_n = (w_{ij})_{1 \le i,j \le n} = o(h_n^{-1})$ is the spatial weights matrix associated to these units

$\hookrightarrow$Relation between Y_i and X_i is modeled by the FSAR :

$$Y_i = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_j + \int_{\mathcal{T}} X_i(t)\theta^*(t)dt + U_i, \qquad i = 1, \ldots, n, \tag{3}$$

- λ_0 is the autoregressive parameter
- $\{X_i(t)\, t \in \mathcal{T}\},\ i = 1, \ldots, n$ are i.i.d
- The disturbances U_i are i.i.d and such that $E(U_i^2) = \sigma_0^2$.

QML (Lee, 2004) is one of the popular estimation methods (ML ; Ord (1975), 2SLS ; Kelejian and Prucha (1998), GMM ; Smirnov and Anselin (2001)) when $X \in \mathbb{R}^p$.

Quasi-Likelihood function

Model (3) can be rewritten as

$$S_n \mathbf{Y}_n = \mathbf{X}_n(\theta^*(\cdot)) + \mathbf{U}_n, \tag{4}$$

where

- $\mathbf{X}_n(\theta^*(.))$ is the vector of i-th element $\int_{\mathcal{T}} X_i(t)\theta^*(t)dt$.
- $S_n = I_n - \lambda_0 W_n$, let $S_n(\lambda) = I_n - \lambda W_n$.

So the conditional log quasi-likelihood function associated to (4) is :

$$\begin{aligned} L_n(\lambda, \theta(\cdot), \sigma^2) = & -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) + \ln|S_n(\lambda)| \\ & - \frac{1}{2\sigma^2}\left[S_n(\lambda)\mathbf{Y}_n - \mathbf{X}_n(\theta(\cdot))\right]'\left[S_n(\lambda)\mathbf{Y}_n - \mathbf{X}_n(\theta(\cdot))\right] \end{aligned} \tag{5}$$

$\hookrightarrow$ Maximize truncated version of (5).

Truncated Quasi-Likelihood function

Let the truncated log quasi-likelihood :

$$\begin{aligned}\tilde{L}_n(\lambda, \theta, \sigma^2) &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) + \ln|S_n(\lambda)| \\ &\quad -\frac{1}{2\sigma^2}\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\theta\right]^{'}\left[S_n(\lambda)\mathbf{Y}_n - \xi_{p_n}\theta\right],\end{aligned}$$

with

- ξ_{p_n} is an $n \times p_n$ matrix of elements given by

$$\varepsilon_j^{(i)} = \int_{\mathcal{T}} X_i(t)\varphi_j(t)dt\,, \qquad i = 1, \ldots, n \;\; j = 1, \ldots, p_n.$$

- θ is the $1 \times p_n$ vector of parameters given by

$$\theta_j = \int_{\mathcal{T}} \theta(t)\varphi_j(t)dt\,, \qquad j = 1, \ldots, p_n.$$

- Let $\theta^* = (\theta_1^*, \ldots, \theta_{p_n}^*)^{'}$

QML Estimators

- For a fixed λ, the truncated log quasi-likelihood is maximized at

$$\hat{\theta}_{n,\lambda} = (\xi_{p_n}^{'}\xi_{p_n})^{-1}\xi_{p_n}^{'}S_n(\lambda)\mathbf{Y}_n = (\hat{\theta}_{nj,\lambda})_{j=1,\dots,p_n}$$

and

$$\hat{\sigma}^2_{n,\lambda} = \frac{1}{n}\mathbf{Y}_n^{'}S_n^{'}(\lambda)M_nS_n(\lambda)\mathbf{Y}_n,$$

where $M_n = I_n - \xi_{p_n}(\xi_{p_n}^{'}\xi_{p_n})^{-1}\xi_{p_n}^{'}$.

- The concentrated truncated log quasi-likelihood function of λ is :

$$\tilde{L}_n(\lambda) = \tilde{L}_n(\lambda, \hat{\theta}_{n,\lambda}, \hat{\sigma}^2_{n,\lambda}) = -\frac{n}{2}(\ln(2\pi)+1) - \frac{n}{2}\ln\hat{\sigma}^2_{n,\lambda} + \ln|S_n(\lambda)|.$$

$\hookrightarrow$ Then the estimator of λ_0 which maximizes $\tilde{L}_n(\lambda)$ is $\hat{\lambda}_n$, and

- Those of θ^* and σ_0^2 are, respectively, $\hat{\theta}_{n,\hat{\lambda}_n}$, $\hat{\sigma}^2_{n,\hat{\lambda}_n}$
- The corresponding estimator of $\theta^*(\cdot)$ is : $\hat{\theta}_n(t) = \sum_{j=1}^{p_n}\hat{\theta}_{nj,\hat{\lambda}_n}\varphi_j(t)$.

Asymptotic Results

The QMLE $\hat{\lambda}_n$ is consistent and satisfies

$$\sqrt{\frac{n}{h_n}}(\hat{\lambda}_n - \lambda_0) \to \mathcal{N}(0, s^2_{\lambda_0}),$$

$\hat{\sigma}^2_n$ is consistent and satisfies

$$\sqrt{n}(\hat{\sigma}^2_{n,\hat{\lambda}_n} - \sigma^2_0) \to \mathcal{N}(0, s^2_{\sigma_0})$$

Asymptotic Results

We have

$$\frac{n\left(\hat{\theta}_{n,\hat{\lambda}_n}-\theta^*\right)^{'}\Gamma_{p_n}\left(\hat{\theta}_{n,\hat{\lambda}_n}-\theta^*\right)-p_n}{\sqrt{2p_n}}\to\mathcal{N}(0,\sigma_0^4).$$

Moreover,

$$\frac{nd^2\left(\hat{\theta}_n(\cdot),\theta^*(\cdot)\right)-p_n}{\sqrt{2p_n}}\to\mathcal{N}(0,\sigma_0^4),$$

where $d^2(\cdot,\cdot)$ denotes the metric defined in $L^2(\mathcal{T})$ by

$$d^2(f,g)=\int_{\mathcal{T}}\int_{\mathcal{T}}(f(t)-g(t))\,E(X(t)X(s))\,(f(s)-g(s))\,dtds,$$

for all $f,g\in L^2(\mathcal{T})$.

Numerical results : simulations

We consider the model

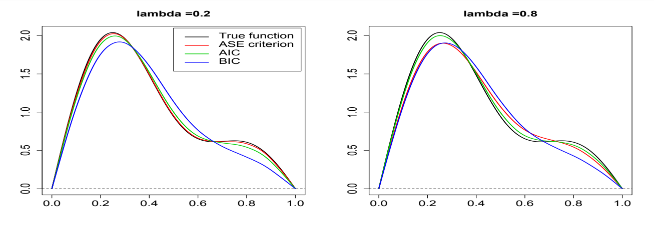
$$Y_i = \lambda_0 \sum_{j=1}^{n} w_{ij} Y_j + \int_0^1 X_i(t)\theta(t)dt + U_i, \quad X(t) = \sum_{j=1}^{20} \varepsilon_j \varphi_j(t),$$

where $\theta^*(t) = \sum_{j=1}^{20} \theta_j \varphi_j(t)$, $U_i \sim \mathcal{N}(0, \sigma_0^2 = 1)$.

- $\varepsilon_j \sim \mathcal{N}(0, 1/j)$ for $j \geq 1$, $\theta_j = 1/j$, for $j = 1, 2, 3$ and $\theta_j = 0$ for $j > 3$
- W_n is constructed by taking the 8 neighbors of each unit using kNN method.
- Use of the eigen-basis, FPCA associated to the kernel $\hat{K}(t, v) = \dfrac{1}{n-1} \sum_{i=1}^{n} X_i(t) X_i(v)$.
- Using AIC, BIC and *ASE* criterion to chose the number of PC (p_n).
- Comparing IMSE $= \int (\theta(t) - \hat{\theta}(t))^2 dt$

Simulations

	$\lambda_0 = 0.2$			$\lambda_0 = 0.8$		
	ASE	AIC	BIC	ASE	AIC	BIC
λ	.1912	.1912	.1916	.7941	.7950	.7950
	(.0800)	(.0799)	(.0801)	(.0321)	(.0318)	(.0321)
σ^2	.9833	.9839	.9871	.9957	.9889	.9925
	(.0687)	(.0688)	(.0690)	(.0745)	(.0720)	(.0536)
IMSE	.0394	.0548	.0881	.0811	.0492	.0880
	(.0409)	(.0484)	(.0476)	(.0970)	(.0476)	(.0536)
PCs	2.980	2.725	2.395	2.615	2.775	2.405
	(.1404)	(.4476)	(.4901)	(.6315)	(.4186)	(.5022)



Problem of Spatial dependence + sampling models

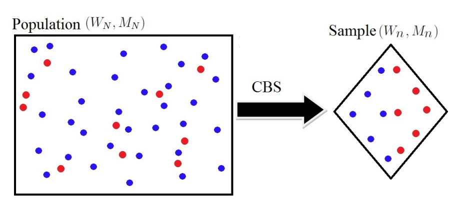
Suppose that in the population

- N observations
- $\tilde{\mathbf{X}}_N$ be a N random variables from a functional exogenous variable
- each couple $(\tilde{X}_i, \tilde{Y}_i)_{;i=1,\ldots,N}$ is based on $\tilde{Y}_i = \mathbb{I}\left(\tilde{Y}_i^* \geq 0\right)$ via

$$\tilde{\mathbf{Y}}_N^* = \rho W_N \tilde{\mathbf{Y}}_N^* + \tilde{\mathbf{X}}_N \theta + (I_N - \lambda M_N)^{-1} \nu_N,$$
$$\nu_N \sim (0, I_N)$$

Problem of Spatial dependence + case-control sampling

Stratifying the population (using fine stratifying sampling) :



- sample $(X_i, Y_i)_{i=1,\ldots,n}$ of size n
- difficult to have the right spatial connection (W_N, M_N) in the population
- (W_n, M_n) are not able to identify the autoregressive parameters ρ and/or λ.
- not possible to estimate the parameters using the existed estimation methods.

Conclusion and Perspectives

- Exogenous weight matrices
- Extension of the full econometric model
- Relation between economic growth and emissions from carbon dioxide, sulfur dioxide, carbon monoxide,...,
- More flexible models

Thank You for your time