

Statistics for dependent Data : Problematics, Models and Methods ¹

Samir BEN HARIZ,
Université du Maine, France

ISMAI, Kairouan, Mars 2010

¹A special thanks to Mounir Ben Salah et Hatem Najjar

1. Aim : Use the data for either explanation or prediction.
2. Why Stochastic models?
3. How to model randomness?
4. Methods and techniques.
5. Some results about change point
6. Simulations
7. Perspectives and open questions

Introduction

- ▶ Your Data : X_1, \dots, X_n , n observations. $X_i = X_{t_i}$
 \Rightarrow Extract some information about the law of your experiments.
- ▶ A stochastic model : we assume X_i are random variables defined on some probability space (Ω, \mathcal{A}, P)
 $(X_i)_{i>0}$ is a random sequence or $(X_t)_{t>0}$ a random process.
- ▶ The simplest model : assume independence and identically distributed.
- ▶ The general model : assume stationarity.
 - ▶ A minimal assumption : the law remain unchanged.
 - ▶ If there is a change, first locate the change.
- ▶ How to model and measure dependence over time?
- ▶ Review some basics about inference for Stochastic models.

A simple model: i.i.d

We assume that the X_i are independent identically distributed.

- ▶ The main inference is about the marginal distribution. (The only unknown in the model is the law of X_i).
- ▶ Two possible setting : parametric and nonparametric.
- ▶ Two types of questions : Estimation or Test between hypothesis.
- ▶ An example : " Sondage d'opinions"

Estimation of the mean : SLLN (LFGN) and TLC

- ▶ Assume that X_1, \dots, X_n are i.i.d. with unknown mean m and finite variance.

$$X_i : (\Omega, \mathcal{A}, P) \longrightarrow R$$

We estimate the mean by the empirical mean : $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- ▶ The SLLN ensures the almost sure convergence to the true mean :'

$$\text{For almost all } w \in \Omega, \overline{X(w)} \rightarrow E(X)$$

- ▶ The TLC indicates the speed of the approximation.

$$\sqrt{n}(\bar{X} - E(X)) \implies N(0, \text{var}(X))$$

- ▶ Some examples using Matlab.
- ▶ What about dependance, what about change in the model ?

Memory : What is dependence : LRD and SRD

Let X_1, X_2, \dots, X_n be a stationary sequence

- We would like to model and measure dependence over time.
- We can and do measure dependence by correlations.

$$r(k) = \text{corr}(X_i, X_{i+k}) = \text{Cov}(X_i, X_{i+k})/V(X_i)$$

Assume $\text{Cov}(X_0, X_i) \sim Ci^{-\alpha}$

- ▶ We define LRD by correlations near infinity :
 $\sum_{i=1}^{\infty} |\text{Cov}(X_0, X_i)| = \infty.$
- ▶ if $\alpha < 1 \Leftrightarrow$ Long range dependence (LRD) or Long Memory
- ▶ if $\alpha \geq 1 \Leftrightarrow$ Short range dependence (SRD) or Short Memory
- ▶ LRD is equivalent to a singularity of the spectral density f near zero,

$$r(k) = \int \exp(ikw) f(w) dw$$

Memory : Variance of partial sum

Let X_1, X_2, \dots, X_n be a stationary sequence with correlation

$$r(k) \sim Ck^{-\alpha}, \quad \alpha \geq 0$$

$$f(w) \sim Cw^{-2d}, \quad -1/2 \leq d \leq 1/2$$

Then

$$\text{Var} \left(\sum_{i=1}^n X_i \right) \sim Cn^{2-\max(\alpha,1)} \sim Cn^{2H} \sim Cn^{1+2d}$$

- ▶ α : the decay of the correlation function
- ▶ H The Hurst parameter
- ▶ d the fractional parameter

Examples : ARMA and ARIMA

- ▶ AR processes : $X_t = a_0 + a_1 X_{t-1} + \dots + X_{t-p} + \epsilon_t$ $(\epsilon_t) i.i.d$
- ▶ ARMA processes : B is the back shift operator, Φ, Θ two polynomial functions of order p and q .

$$\Phi(B)X_t = \Theta(B)\epsilon_t \quad (\epsilon_t) i.i.d$$

- ▶ ARIMA processes

$$(1 - B)^{-d} X_t = \epsilon_t$$

- ▶ An example : Regression Analysis

$$Y_t = f(X_t^1, \dots, X_t^p) + \epsilon_t$$

- ▶ Y_t the output variables or the answer
- ▶ f is the trend or the link function.
- ▶ X^1, \dots, X^p input variables.
- ▶ (ϵ_t) the error term.

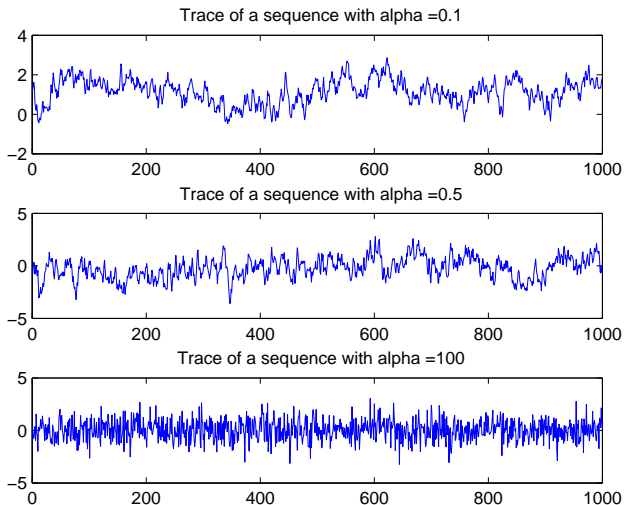


Figure: Traces of a stationary sequence for different values of α

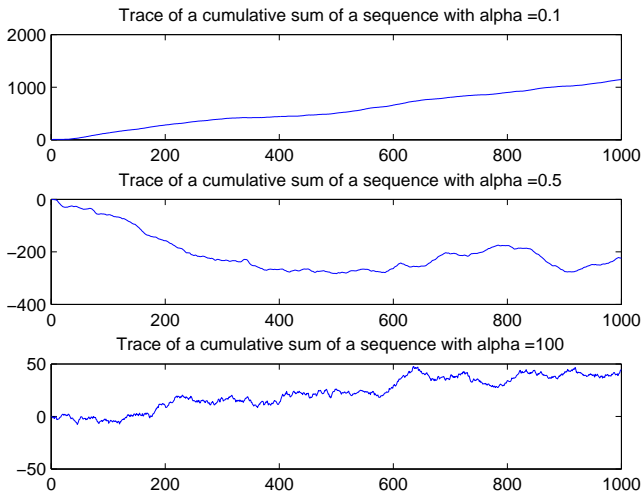


Figure: The cumulative sum process for different values of α

The change point problem : the setting

- ▶ Let $(X_i)_{i=1..n}$ be a sequence in a measurable space E .

$$\mathcal{L}(X_i) = \begin{cases} P_n & \text{if } i \leq n\theta \\ Q_n & \text{if } i > n\theta, \end{cases}$$

where $0 < \theta < 1$ is the location of the change-point.

- ▶ We aim to estimate the location of the change-point θ using an estimator of the following general type:

$$\hat{\theta}_n = \frac{1}{n} \min (\operatorname{argmax}_{1 \leq k < n} \{N(D_k)\}), \quad (1)$$

- ▶ N is a (possibly random) semi-norm on the space \mathcal{M} of signed finite measures on E ,

$$D_k = \left[\frac{k}{n} \left(1 - \frac{k}{n} \right) \right]^{1-\gamma} \left(\frac{1}{k} \sum_{i=1}^k \delta_{X_i} - \frac{1}{n-k} \sum_{i=k+1}^n \delta_{X_i} \right), \quad (2)$$

and γ is a parameter satisfying $0 \leq \gamma < 1$.

A Norm on the space \mathcal{M}

- ▶ For each semi-norm, we require a family of functions \mathcal{F} .
- ▶ For a measure ν on E , and $f : E \rightarrow \mathbb{R}$, we define $N(\nu)$ as $N(\nu) \equiv \int f(x)\nu(dx)$.
- ▶ EXAMPLE 1. $\mathcal{F} = \{\mathbb{1}_{\cdot < x_i}, i = 1, \dots, n\}$, we define norms of a measure ν via the quantities $d_i = \nu(\mathbb{1}_{\cdot < x_i})$. For example, Kolmogorov-Smirnov and L^p norm

$$N(\nu) = \sup_{1 \leq i \leq n} |d_i|, \quad N_p(\nu) = \left(\frac{1}{n} \sum_{i=1}^n |d_i|^p \right)^{1/p}, \quad (3)$$

- ▶ EXAMPLE 2. For $\mathcal{F} = \{f^p : x \rightarrow x^p, p = 1, \dots, +\infty\}$ we define the semi-norm by

$$N(\nu) \equiv \sum_{f \in \mathcal{F}} d(f) |\nu(f)|,$$

where $d(f)$ is a sequence of positive weights.

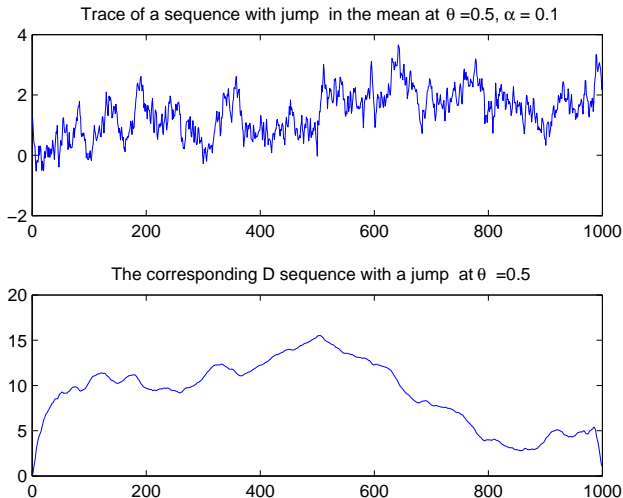


Figure: A sequence with a change in the mean

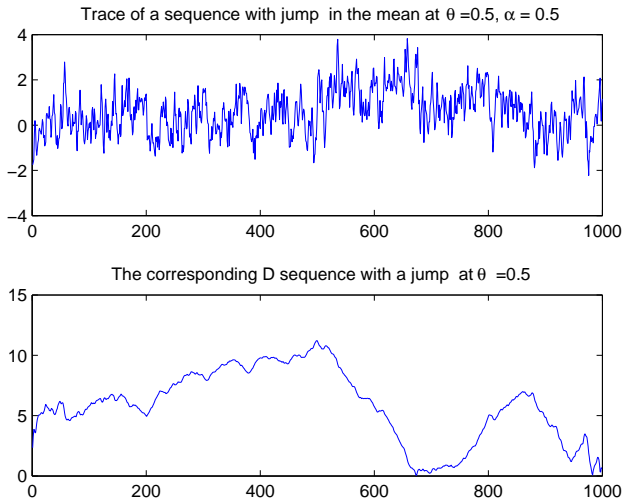


Figure: A sequence with a change in the mean

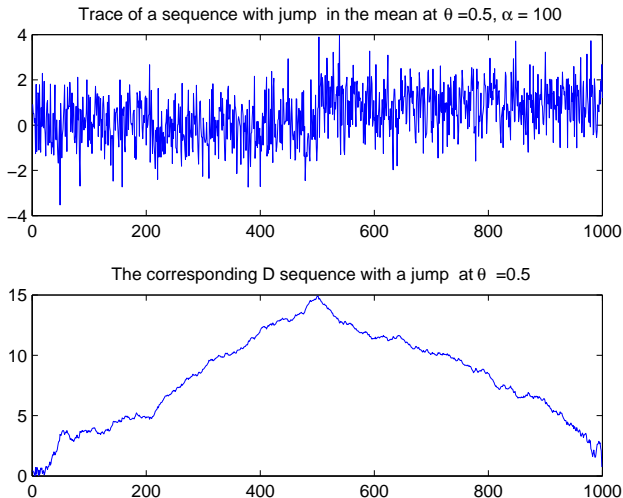


Figure: A sequence with a change in the mean

Assumption 1 : The dependence structure of the sequence

We now turn our attention to the dependence structure of the sequence.

There exist constants $C > 0$ and $0 < \rho < 1$, such that for any m

$$\sup_{f \in \mathcal{F}} \sup_{k, m, k+m \leq n} \mathbb{E} \left(\sum_{i=k}^{k+m} [f(X_i) - \mathbb{E}(f(X_i))] \right)^2 \leq Cm^{2-\rho}. \quad (4)$$

In particular it is the case if there exists $C > 0$

$$\sup_{f \in \mathcal{F}} \sup_{1 \leq i \leq n-m} |\text{corr}(f(X_i), f(X_{i+m}))| \leq Cm^{-\rho}. \quad (5)$$

- ▶ This assumption simply states that for each of the functions f in \mathcal{F} the correlation between $f(X_i)$ and $f(X_{i+m})$ must decay algebraically or faster with m as $m \rightarrow \infty$.

The results : Theorem 1 and 2

- ▶ In Theorem 1 we develop conditions that can deal with countable families of functions and norms that are bounded by weighted moments.
- ▶ In Theorem 2 we consider the case of uncountable families of functions. In this case we need to control the size of the family.
- ▶ For f in \mathcal{F} we set

$$\nu(f) = \int f(x)\nu(dx)$$

Theorem 1 : The weighted moments

Theorem (1)

Assume that

$$N(\nu) \leq \sum_{f \in \mathcal{F}} d(f) |\nu(f)|, \quad (6)$$

There exists a sequence b_n such that

$$\mathbb{P}(N(P_n - Q_n) > b_n) \rightarrow 1 \text{ as } n \rightarrow \infty, \quad (7)$$

with $b_n \geq b > 0$ then

$$\hat{\theta}_n - \theta = O_p(n^{-1}). \quad (8)$$

- ▶ Equation (7) controls the rate at which the semi-norm of the difference between the two distributions decays to zero.

Assumption 2 : Controlling the size of the family

- ▶ Given two functions l and u , the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. Given a norm $\|\cdot\|$ on a space containing \mathcal{F} , an ε -bracket for $\|\cdot\|$ is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$.
- ▶ The bracketing number $N_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|_X)$ is the minimal number of ε -brackets needed to cover \mathcal{F} .
- ▶ A family \mathcal{F} is said to satisfy Assumption 2 if $\forall \varepsilon > 0, N_{[\cdot]}(\varepsilon, \mathcal{F}, \|\cdot\|_X) < \infty$ where $\|\cdot\|_X$ is a norm satisfying $\sup_{n \in \mathbb{N}} |P_n(|f|)| + |Q_n(|f|)| \leq \|f\|_X$.

The second Theorem

Theorem (2)

Assume that the semi-norm satisfies:

$$N(\nu) \leq \sup \{ |\nu(f)|, f \in \mathcal{F} \}, \quad (9)$$

where $\sup \{ \|f\|, f \in \mathcal{F} \} < \infty$. If

- ▶ Assumptions 1 and 2 are satisfied.
- ▶ There exists a sequence b_n such that

$$\mathbb{P}(N(P_n - Q_n) > b_n) \rightarrow 1 \text{ as } n \rightarrow \infty, \quad (10)$$

- ▶ $b_n^{-1} N_{[\cdot]}(b_n \varepsilon_n, \mathcal{F}, \|\cdot\|_X) [n^{-\rho/2} + n^{\gamma-1}] \rightarrow 0$, where ε_n is any sequence that tends to zero as $n \rightarrow \infty$. Then we have

$$\hat{\theta}_n - \theta = O_p \left(n^{-1} [b_n^{-1} N_{[\cdot]}(b_n \varepsilon_n, \mathcal{F}, \|\cdot\|_X)]^{2/\bar{\rho}} \right). \quad (11)$$

Corollary

Assume that the semi-norm satisfies (9) and (10) with $b_n \geq b > 0$ then

$$\hat{\theta}_n - \theta = O_p(n^{-1}). \quad (12)$$

- ▶ In the case $b_n > b > 0$ Theorems 1 and 2 both give the same $O_p(n^{-1})$ rate for both $\rho < 1$ and $\rho \geq 1$
- ▶ For Theorem 1 in the case $b_n \rightarrow 0$ with $\rho \geq 1$ it is possible to obtain the rate $O_p(n^{-1}b_n^{-2} \ln^2(nb_n^2))$ which can represent a marginally better result.

Some ideas behind the results

- ▶ We define $t_k \equiv k/n$, then $D_k \equiv D_n(t_k)$ where

$$D_n(t) = t^{1-\gamma}(1-t)^{1-\gamma} \left(\frac{1}{nt} \sum_{i=1}^{[nt]} \delta_{X_i} - \frac{1}{n(1-t)} \sum_{i=[nt]+1}^n \delta_{X_i} \right)$$

and $w(t) = t^\gamma(1-t)^\gamma$.

- ▶ We rewrite $D_n(t)$ as the sum of its mean and a centered random component, $B_n(t)$,

$$D_n(t) = \frac{1}{w(t)} [(P_n - Q_n)g(t) + B_n(t)], \quad (13)$$

where $g(t) = t(1-\theta)\mathbb{1}_{t \leq \theta} + \theta(1-t)\mathbb{1}_{t > \theta}$ is a piecewise linear function that takes its maximum at the change-point ($t = \theta$) and B_n is the empirical bridge measure given by

$$B_n(t) = W_n(t) - tW_n(1), \quad (14)$$

$$W_n(t) = \frac{1}{n} \sum_{i=1}^{[nt]} [\delta_{X_i} - \mathcal{L}(X_i)], \quad (15)$$

Simulations

- ▶ We generated a sequence by taking

$$X_i = Y_i^2 - 1 \text{ if } i \leq n\theta \text{ and } X_i = 1 - Y_i^2 \text{ if } i > n\theta.$$

- $(Y_i)_{i=1..n}$ with zero mean and unit variance with a covariance given by $r(n) = (1 + n^2)^{-\alpha/4} \sim n^{-\alpha/2}$.

- The marginal distributions before and after the jump have the same mean and variance, but have different skewness.

- ▶ We show results for the estimator that uses the Kolmogorov-Smirnov norm (KS) and the L^1 norm . The parameter γ is equal 0.5.
- ▶ We considered independent sequences, SRD sequences with $\alpha = 1.5$ and LRD sequences with values of $\alpha = 1.0, 0.8, 0.6$ and 0.4. We present simulations in which the sequence length, n , varies between 1000 and 6000.
- ▶ The Mean Absolute Error $MAE \equiv \mathbb{E}(|\hat{\theta}_n - \theta|)$ for each value of α was estimated using 10000 different sequences.

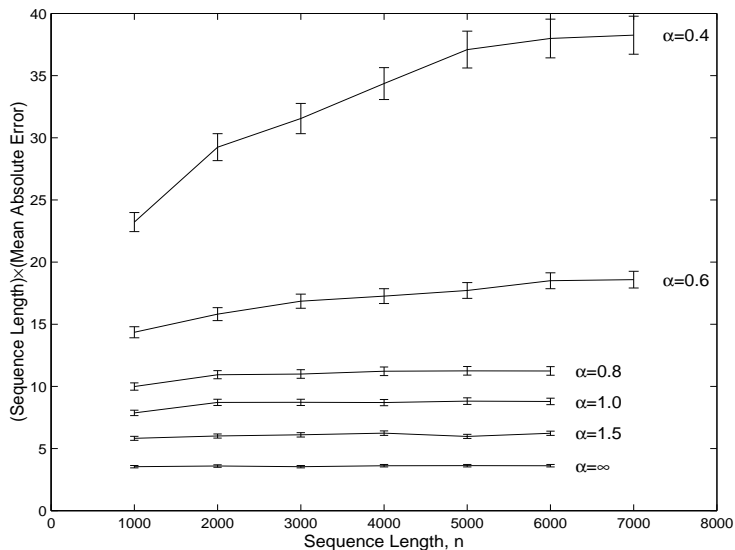


Figure: The MAE of $n(\hat{\theta}_n - \theta)$ for different values of α under the L1 norm

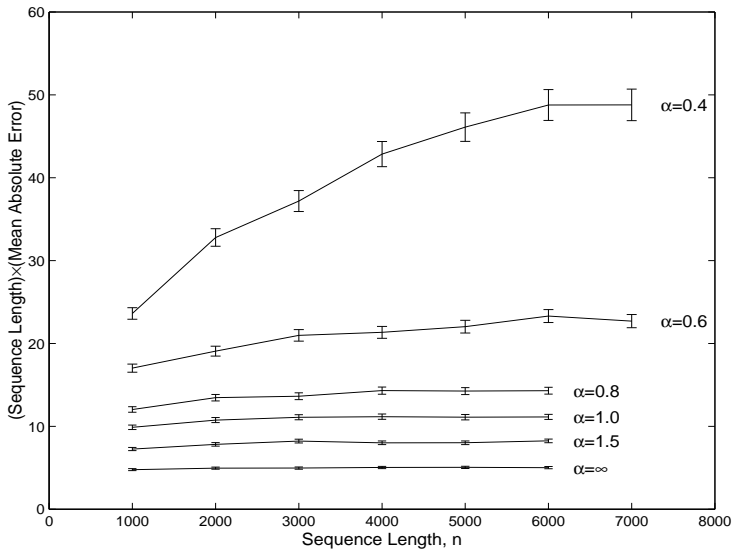







Figure: The MAE of $n(\hat{\theta}_n - \theta)$ for different values of α under the KS norm

- ▶ How to estimate the memory parameter d .
- ▶ How to test and locate possible change in the parameter d .
- ▶ A challenging problem : High frequency Data
- ▶ Model and estimate changing of variance over time : GARCH and augmented GARCH.

THANKS and welcome for questions or comments

-  Samir Ben Hariz, et al., Change-point detection for long-range dependent sequences in a general setting, *Nonlinear Analysis: Theory, Methods and Applications*, December 2009.
-  Ben Hariz, Samir; Wylie, Jonathan J.; Zhang, Qiang. Optimal rate of convergence for nonparametric change-point estimators for nonstationary sequences. *Ann. Stat.* 35, No. 4, (2007).
-  Ben Hariz, S. and Wylie, J.J. Rates of convergence for the change-point estimator for long-range dependent sequences . *Statistics and Probability Letters* (2005).
-  Carlstein, E. (1988), Nonparametric change-point estimation. *Ann. Statist.* 16, 188–197.
-  Dumbgen, L. (1991), The asymptotic behavior of some nonparametric change-point estimators, *Ann. Statist.* 19, 1471–1495.