Non Parametric Modeling of Cyclo-Stationary Markovian Processes
Part II : prediction and dimension reduction

Pierre-François Marteau, Valérie Monbet
University of Bretagne Sud,
Vannes, France
29th June 2006

Abstract
We address in this paper the non parametric modeling of cyclo-stationary multivariate Markovian processes using a continuous state space and discrete time Hidden Markov chain for which all necessary densities are approximated using samples. The model we propose approximates the initial multivariate process (IMP) by decomposing it into a Lower Dimensional Markov Chain (LDMC) for which state transitions are hidden since under the assumption observed through a second stochastic process that from any state of the LDMC generates a multivariate observation. The hidden LDMC state vector coincides with the first coordinates of the state vector of the IMP, while the multivariate observation vector coincides with the last coordinates of the IMP state vector. This decomposition induces a dimension reduction that allows to handle more complex processes, at a computational cost that can be estimated from the data. A Viterbi algorithm is proposed to extract most likely LDMC state trajectories from sequences of observation vectors. The approach can be used to predict state trajectories when the underlying multivariate dynamical process is partially observed. The method is thus of practical use in case of partially missing or noisy sample data. It can also be used as a bootstrap technique when the dimension reduction of the state space is necessary for complexity management. The approach is applied to the modeling of sea state processes for which for instance the mean wind speed is predicted from the observation of the significant wave height and peak period.

Keywords: Nonlinear Time Series, Hidden Markov Models, Resampling, Non Parametric Estimation, Cyclo-Stationarity, Markov Chain, Prediction, Dimension Reduction, Sea State Processes

Introduction
BLABLALBA

1 Continuous Space Hidden Markovian Model (HMM)

1.1 Notations, definitions and hypothesis
Let \( \{ X_t, t > 1 \} \) be a d-dimensional stochastic process in \( \mathbb{R}^d \) where \( \mathcal{B}_d \) is the Borel \( \sigma \) algebra over \( \mathbb{R}^d \). This process can be decomposed in two independent lower dimensional processes: \( \{ S_t, t > 1 \} \) being a u-dimensional stochastic process in \( \mathbb{R}^u \) and \( \{ O_t, t \geq 1 \} \) being a v-dimensional stochastic process in \( \mathbb{R}^v \) with \( u + v = d \).

We suppose that there exists a positive integer \( p < \infty \) such that the stochastic process \( \{ S_t, t > p \} \) with \( Y_t = \{ S_t, S_{t-1}, ..., S_{t-p+1} \} \) forms a cyclo-stationary Markov chain on \( \mathbb{R}^{up}, \mathcal{B}_{up} \), with transition probability function \( P_t(y, A) \) where \( A \subset \mathbb{R}^u \) and \( t \in [0, \text{period}] \). We suppose further that \( P_t \) is slowly varying such that it can be assumed that the process is almost stationary on small time intervals.

We assume that this Markov chain admits a stationary distribution \( \pi_t \) with continuous density \( f_t(y) \) with respect to Lebesgue measure and such
that \( \gamma_t \) is periodic with period \( period \). We also suppose that the transition distribution function \( F_{t,y}(x) = P(X_{t+1} < x | Y_t = y) \) is periodic and admits a continuous probability density function \( \tau(x | y) \).

Furthermore, we suppose that there exists a positive integer \( q < \infty \) such that the stochastic process \( \{Z_t, t \geq p\} \) where \( Z_t = \{O_t, O_{t-1}, \ldots, O_{t-q+1}\} \) is also cyclo stationary and driven according to the conditional distribution \( g_{t,a}(z; - P(Z_t < z) | Y_t = y) \). We assume that this distribution remains almost stationary over a small time interval and admits a continuous probability density function \( g_{i,y} \) that we suppose periodic.

The density probability functions of the cyclo-stationary distribution and the transition probability distribution will be approximated from the data using appropriate kernel estimates. Let \( K_u \) and \( K_v \) be probability density functions on \( \mathbb{R}^u \) and \( \mathbb{R}^v \) respectively, \( K_{up} \) and \( K_{eq} \) probability density functions on \( \mathbb{R}^{up} \) and \( \mathbb{R}^{eq} \) respectively. Let \( \{h_T, T = 0, 1, 2, \ldots\} \) a sequence of positive numbers such that \( h_T \to 0 \) as \( T \to \infty \). We suppose that the density kernels \( K_u, K_v, K_{up} \) and \( K_{eq} \) satisfy the usual conditions.

The Viterbi Local Grid Bootstrap (VLGB) resampling scheme

The LGB resampling algorithm generates a series \( X_1, X_2, \ldots, X_N \) where the length \( N \) may be chosen independently from the length \( T \) of the observed sequence \( X = (X_i)_{i \in \{1, \ldots, T\}} \). Let us denote \( \hat{Y}_t = \{\hat{X}_t, \hat{X}_{t-1}, \ldots, \hat{X}_{t-p+1}\} \) the state of the generated sequence at time \( t \).

The generation of \( \hat{X}_k \) is obtained by assigning probabilities to a finite subset of convenient states and sampling this subset according to the discrete probability mass. The idea behind the generation of unobserved states is that, if the underlying transition distribution is regular enough (for instance continuous), it is possible through the kernel estimate to assign a probability to unobserved states around the observations \( \{X_k\} \) and to synthesize a new time series that includes observed and unobserved points. The set of "unobserved states" is defined by a discretization of the image, through the Markovian transition operator, of a neighborhood of the current point \( \hat{Y}_t \).

The resampling scheme may be described as follows:

**Initialization step** - Select an initial state \( \hat{Y}_p \), the bandwidth parameter \( h_0 \), the width \( \gamma_T \) of the neighborhood of a given state the time window \( \gamma_0 \) in which the process can be supposed to be stationary and the grid parameters the discretization step \( \Delta_g \) and the edge length \( \gamma_g = N_g + \Delta_g \). \( \Delta_g \) may depend on \( \alpha_T \).

**Step t**

- Let us suppose that the state \( \hat{Y}_t \) is already sampled. The neighborhood \( \hat{V}_Y \) of \( \hat{Y}_t \) is defined by the set of observed \( Y_t \in \{Y_k | d(Y_k, \hat{Y}_t) \leq \sigma_T / 2 \& \in \{T - \gamma_T, T + \gamma_T\} \} \). We note \( I[\hat{V}_Y] \) the set of time index such that for all \( t \in I[\hat{V}_Y], Y_t \in \hat{V}_Y \). Furthermore, we define the image \( \hat{V}_Y(t) \) of \( \hat{V}_YU \) by \( \hat{V}_Y(t) = \{X_{t+1}, t \in I[\hat{V}_Y]\} \subset \mathbb{R}^d \).

- Now, a grid \( G_t = \{g_{t_1}, \ldots, g_{t_j}\} \) is built by discretizing a cube of \( \mathbb{R}^d \) with grid step \( \Delta_g \) and edge length \( \gamma_g \). The cube is centered on the barycentre of \( \hat{V}_Y(t) \) and the edge length \( \gamma_g \) is defined such that the cube includes at least all the elements of \( \hat{V}_Y(t) \). We note \( G_{Y(t)}(t) = \{\hat{V}_Y(t) \cup G_t \}

- Let \( J \) be a discrete random variable taking its values in \( I[G_{Y(t)}(t)] = \{k \in \mathbb{N}, X_k \in G_{Y(t)}(t)\} \) with probability mass function given by:

\[
P(J = k) = \frac{p(Y_t, X_{k+1})}{\sum_{j \in I[G_{Y(t)}(t)]} p(Y_t, X_{j+1})}, \quad \text{1,}
\]

\[
p(Y_t, X_{k+1}) = p_k \sum_{i \in I[\hat{V}_Y]} K_d (\frac{X_{k+1} - X_i}{h_T}) \times K_d (\frac{Y_t - Y_i}{h_T}) \times \exp - \frac{(k-i) \mod \text{period}}{\sigma_{\text{period}}}, \quad \text{2,}
\]

where

\[
h_T = \left( \frac{4}{(2d + 1)n} \right)^{1/(d+1)} h_0, \quad \text{3,}
\]
The sampled state at time $t+1$ is such that $X_{t+1} = X_J$.

The above procedure assigns sampling probabilities to a set of points that includes the successors of the observed neighbors of the last sampled state and the points of a local grid positioned around these successors. This discrete probability mass may be considered as transition probabilities between $Y_t$ and $X_J$. It depends both on the density of the original sequence around $X_J$ and on the density of the observed state around $Y_t$. As the kernels $K_d$ and $K_{dp}$ are continuous on $\mathbb{R}^d$ and $\mathbb{R}^{dp}$ respectively, we are able to assign probabilities to unobserved points of the grid and consequently to sample unobserved states.

Properties of the bootstrap

According to the above sampling scheme for any fixed length $T$ of the observed serie, the generated series $\{X_t, t = 1, 2, \cdots\}$ evolves following a dependence structure to be characterized. Theorems ?? and ?? describe the asymptotic validity of the LGB procedure. Asumptions are given in appendix.

Theorem 1 For given $T$ and $X_1, \cdots, X_T$, for every kernel satisfying (K1) and every fixed kernel bandwidth $h_T$, there exists probability one $\nu_0 \in \mathbb{N}$ such that the generated Markov serie $Y_T = \{Y_t, t > 0\}$ is positive recurrent, irreducible and aperiodic Markov chain.

The probabilistic properties of the bootstrap Markov chain $\hat{Y}$ depends on the chosen kernel and on its bandwidth parameter $h_T$. Indeed, if $h_T$ is sufficiently close to $0$, the generated time series will be an exact reproduction of the reference serie. On the contrary, if $h_T$ is too large, the bootstrap procedure will be unable to restore the statistic properties of the reference Markov chain. The grid parameters have no influence on the estimation of the transition probabilities itself. It is shown in Monbet (2004) that the grid parameters impact directly on the reproduction ratios of subsequences of the reference time series inside the sampled one.

Theorem 2 Under assumptions (K1)-(K3), (A1)-(A3), if $h_T \to 0$ and $T h_T^{dp} \to \infty$ as $T \to \infty$ and $\sigma_T \to c h_T$ with $c$ a positive constant. For all $x \in S'$ and $y \in S$, we have for the one step transition and the stationary distributions.

1. $F_{\hat{Y},T}(x) \to F_{\hat{Y},T}(x)$ weakly in Prohorov as $T \to \infty$

2. $f_{\hat{Y},T}(y) \to f_{\hat{Y},T}(y)$ weakly in Prohorov as $T \to \infty$.

where $S'$ is a compact subset of $\mathbb{R}^d$ such that $f_{\hat{Y},T}(x') > 1$ for all $x \in S'$ and $S = S' \times \cdots \times S' \subset \mathbb{R}^{dp}$. $f_{\hat{Y},T}$ denotes the stationary probability density function of the random variable $X_T$.

The first part of theorem ?? ensures the weak convergence of the transition distribution function of the simulated series to the transition distribution of observed time series when the observation time becomes long enough. This result is important because a Markov process may be entirely specified by its transition distribution function when the stationary mode is achieved.

The second part of the theorem gives the weak convergence of the stationary distribution function of the LGB Markov chain to the stationary distribution function of the reference time serie.

Practical choice of the resampling parameters

In the LGB procedure, the transition probabilities are estimated non parametrically. One of the difficulties of kernel based estimation lies in the choice of the initial bandwidth parameter $h_0$. Some solutions have been proposed in the litterature to help in the choice this parameter, but non it really efficient for multivariate data. So that one of the best solution remain to try several values until the simulated series restore well some chosen statistics.

The grid parameters length of edge $\sigma_T$ and step $\Delta_\sigma$ are also sensitive. They determine the number $N_\sigma$ of the non observed states included in the neighborhood of the current point for the sampling $N_\sigma = \left( \frac{2 \Delta_\sigma}{\sigma_T} \right)^d$. When the ratio $\sigma_\sigma/\Delta_\sigma$ tends to zero, LGB algorithm tends to become a standard Local Bootstrap procedure (Paparoditis (2002)) as none unobserved state is added to the observations. If $\sigma_\sigma$ is large compared to $h_T$, non observed states will be far from the center of the grid and they will have a very low probability to be sampled. Ideally, $\sigma_\sigma$ has to be chosen so that each non observed state has a too large probability to be reached. Now we can observe that for fixed $\sigma_\sigma$, the smaller $\Delta_\sigma$, the larger the number of accessible states. Unfortunately the
complexity of the algorithm increases exponentially with $\Delta_g$, so that the grid step should not be too small.

**Application: profitability of an offshore barge**

In order to demonstrate how the LGB algorithm run for real data, we propose two numerical examples: hindcast data of OCEANWEATHER database at point . 22 years of data are available with 6 hours time step.

Let us suppose that a barge is positioned at the point of the data. Off shore operations can be performed only if the significant wave height $H_s$ is low enough, if the propagation direction of the sea states $\Theta$ does not induce too much roll and if the period of the sea states $T_p$ does not correspond to the resonance frequency of the structure and its confidence interval. Simulations will be used to estimate the mean necessary time to perform an operation from a barge given the date of the beginning of the operation. In the same way, we could study the profitability of the barge at the given point.

We simulate 10 times 22 years of data. In order to well estimate the mean necessary time to perform an operation it is important that some statistics observed in the data are well restored, such as:

- the marginal and joint instantaneous distributions of $H_s$, $T_p$ and $\Theta$
- the persistence duration of storms
- the persistence duration of calm weather

Results are presented in two parts

**First part**

First the marginal and joint statistics of couple $(H_s, \Theta)$ are shown in figures ?? to ?? for generated data learned on 22 years of observed data. The distributions functions (fig. ??) permit to observe that the LGB algorithm allows to restore quite well the statistical features of the data. We have to remark that the generator tends to simulate a little bit to much low values (or values that are frequent in the data) and not enough extrem high values (or values that are not frequent in the data). This is due to the choice of the bandwidth parameter of the non parametric estimators of the transition probabilities. This drawback well known of non parametrical estimators is also observed on the joint density probability functions of the cartesian coordinate corresponding to the polar coordinates $(H_s, \Theta)$ . The persistence statistics (fig. ?? to ??) show that the time dependence structure of the sea state processes is restored by the generated data. Figures ?? and ?? give the cumulative distribution function of persistence at storms and calm between storms for respectively January and August months. It show that the LGB algorithm achieve to take into account the non stationnarity of the studied data.

**Second part**

Let us now discuss figure ?? This figure represents the mean time necessary to perform an offshore operation under next constraints: $H_s < 2$, $T_p \notin [9.5, 13]$ and $\Theta \notin [150, 300]$, given the month of the beginning of the operation. The empirical confidence intervals are added to the punctual estimate. The generated data used to calculated the necessary mean time are learned on only 5 years of observed data. On the figure ??, generated statistics are compared to observed statistics calculated with the complete observed time series (22 years of data). The mean time estimated with the simulated data is very close to the mean time compute with the observed data. This example shows that the LGB algorithm is an interesting tool to make richer or longer a set of observed data in order, for instance, to empirically study the profitability of
Figure 2: Joint probability density function of couple \(U = H_s \cdot \sin(\Theta), V = H_s \cdot \cos(\Theta)\) Top: observed data, Bottom: generated data.

Figure 4: January month - Top: persistence of storms (left: higher as 2/3max(H_s^{max}), right higher as 1/3max(H_s^{max})) : persistence of calm between storms. Blue line: observed data, red line: generated data. Dotted red line: confidence interval.

Figure 3: Persistence of storms higher as 2/3max(H_s) (left: H_s; Right: \(\Theta\)). Blue line: observed data, red line: generated data. Dotted red line: confidence interval.

Figure 5: August month - Top: persistence of storms (left: higher as 2/3max(H_s^{max}), right higher as 1/3max(H_s^{max})). Blue line: observed data, red line: generated data. Dotted red line: confidence interval.
an offshore structure.

Conclusion

In this paper we have presented a nonparametric resampling procedure referred as LGB that can be used to model and synthesize multivariate Markov processes of general order $p$. This procedure extends the one proposed by Paparoditis and Politis (2002) in two ways: firstly, it cope with multivariate processes and secondly unobserved states can be generated by the proposed resampling algorithm. The last feature is obtained by means of a grid used to discretize locally the image through the transition operator of the neighborhood of the current state. Theoretical results show that the proposed resampling procedure is asymptotically correct.

Application results show that the proposed algorithm permits to generate long sequences of data given observed time series. A large set of statistical feature of the data are restored in the generation such as instantaneous distribution functions and persistance statistics. The main drawback of the proposed simulation scheme is that it tends to overestimate the number of occurrence of very large data. This point could be improve in implementing a best procedure to chose the bandwidth parameter of the non parametrical function estimators.

We will show in a second paper a Vitterbi version of LGB algorithm which allows to predict a component of a sea state parameter vector given the observation of the other components. The Vitterbi algorithm will be tested in a spatial context.

Acknowledgment

This work has been performed in collaboration with Metocean group of Ifremer.

References


Appendix

Usual conditions for the density kernels

(K1) $K_i(\cdot)$ is continuous, symmetric, bounded, i.e. $\sup\{K_i(x) : x \in \mathbb{R}^d\} < \infty$ and $K_i(x) > 0 \ \forall x \in (\mathbb{R}^d)^i$ for all $i \in \{d, dp\}$.

(K2) $\|x\|K_i(x) \to 0$ as $\|x\|$ tends to infinity for all $i \in \{d, dp\}$.

Assumptions for the theorems

Let us first introduce some assumptions. $F_0$ defines the mapping $y \in S \mapsto F_y(\cdot)$ where $S$ is a compact set of $\mathbb{R}^{dp}$ and $F_y$ belongs to the space of distributions defined on a compact set $S' \subset \mathbb{R}^d$. We suppose that $F_0$ is continuous for a suitable metric.

Let us define, for any time $t$, the transition distribution $F_\tau$ (eq. ??) and the stationary distribution $F$ (eq. ??) of the observed (or reference) time series $(Y_t)$:

$$\begin{align*}
F_\tau(x) &= P(X_{t+1} < x | Y_t = y), \\
X_{t+1} &\sim \mathbb{R}^d, \ Y_t \sim \mathbb{R}^{dp} \quad (4) \\
F(y) &= P(Y_t < y), \ Y_t \sim \mathbb{R}^{dp} \quad (5)
\end{align*}$$

We need to introduce some assumptions on the distribution functions $F_\tau$ and $F$ to ensure the convergence of their estimates.

(A1) We suppose that there exists a positive integer $p < \infty$ such that the stochastic process $(Y_t, t > p)$ with $Y_t = \{X_t, X_{t-1}, ..., X_{t-p+1}\}$ forms a cyclostationary Markov chain on $(\mathbb{R}^{dp}, B_{dp})$ with transition probability function $P_t(y, z)$ where $A \subset \mathbb{R}^d$ and $t \in [0, \text{period}]$. We suppose further that $P_t$ is slowly varying such that it can be assumed that the process is almost stationary on small time intervals.

We assume that this Markov chain admits a stationary distribution $\pi_t$ with continuous density $f_{t,y}$ with respect to Lebesgue measure and such that $\pi_t$ is periodic with period $\text{period}$. We also suppose that the transition distribution function $P_{t,y}(x) = P(X_{t+1} < x | Y_t = y)$ is periodic and admits a continuous probability density function $\pi_t(x, y)$.

(A2) $F$ and $F_y$ are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{dp}$ and $\mathbb{R}^d$ respectively and have bounded densities.

(A3) $F_0$ is Prohorov continuous.

(A4) For every time $t$, the stationary probability density function $f_t$ of $F$ is positive on a compact set $S \subset \mathbb{R}^{dp}$. $S$ is defined for a given reference process such that $Y_t \in S$ a.s. for all time $t$.

We remark that the compactness of the support of $f_t$ is a technical assumption which is non restrictive for the applications. For instance, every physical system with finite energy has states that take their values on a compact subset, most of the biological parameters are bounded, etc.

The Prohorov continuity is defined in a topological space equipped with a Prohorov measure. If $M$ denotes the space of all probability measures on $(\Omega, \mathcal{F})$, the weak convergence topology on $M$ is metrizable by the Prohorov distance $d_{pr}(\cdot, \cdot)$. The distance $d_{pr}(G_1, G_2)$ of two measures $G_1$ and $G_2$ is given by

$$d_{pr}(G_1, G_2) = \inf \{\varepsilon | G_1(A') \leq G_2(A') + \varepsilon \ \forall A' \subset \Omega \} \quad (6)$$

where $A' = \{y | d(A, y) < \varepsilon \}$ and $d$ is a distance in $\mathbb{R}^d$. 

7