Random forest for dependant data: simulation results

26/06/2019

We propose to simulate a first data set corresponding to time series data with seasonnalities. Let:

$$y_t = \cos(\omega_1 t) + \cos(\omega_2 t) + \varepsilon_t$$

where $\omega_1 = 2\pi/40$ corresponds to a low frequency term and $\omega_2 = 2\pi/20$ a (relatively) high frequency term. ε_t is iid gaussian, mean 0 and variance σ^2.

We have the intuition that dependant random forest methods could work well in the case where the iid forest has an error wich contains time dependant patterns. W illustrate that bellow.

To model a seasonnality of period T with random forest, we build canonical seasonnal variables as $x_t^T = (1, 2, ...T, 1, 2, ..., T....)$. This is the way we do it to model the yearly seasonnality of the load consumption. Here we choose as input of the forest x_t^{20} which means that the forest doesn't include the covariat that model the low frequencies.

We generate a data set of size $n = 200$ and split the data into two sets of size 100, one for learning the forests, one for forecasting. We set $\sigma = 0.5$. An example is plotted on Figure 1.

We compute the forecasting mse and present the results on Figure 2 ($\sigma = 0.5$) and 3 ($\sigma = 1$). In both case the best performances are achieved with moving block boostrap with blocks of sizes arround 20. This corresponds to the seasonnality of the residuals of the iid forest.

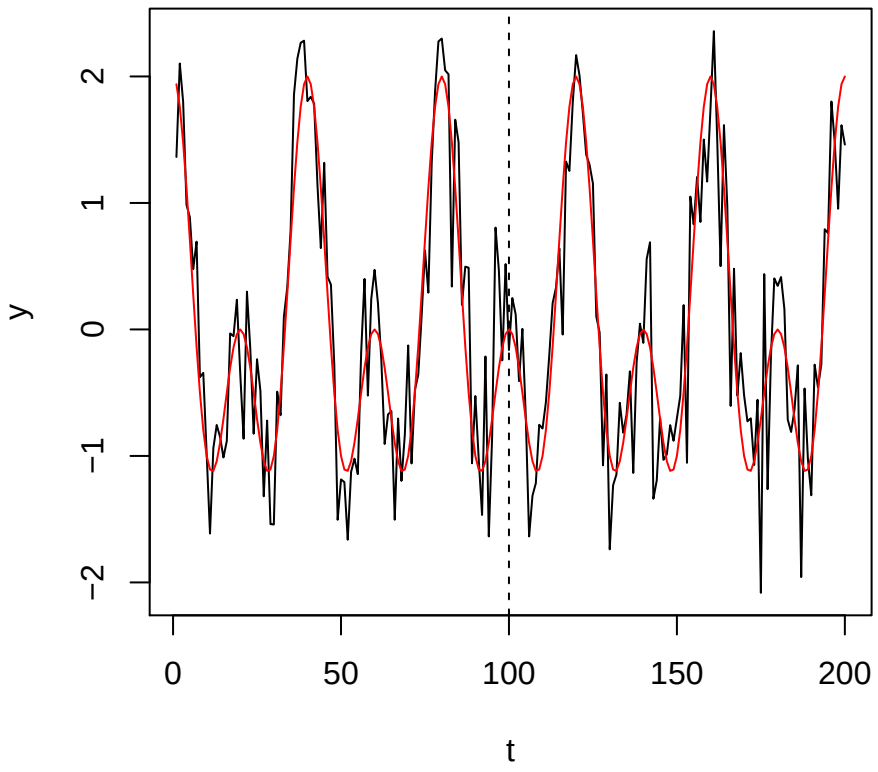

Figure 1: simulated seasonnal data

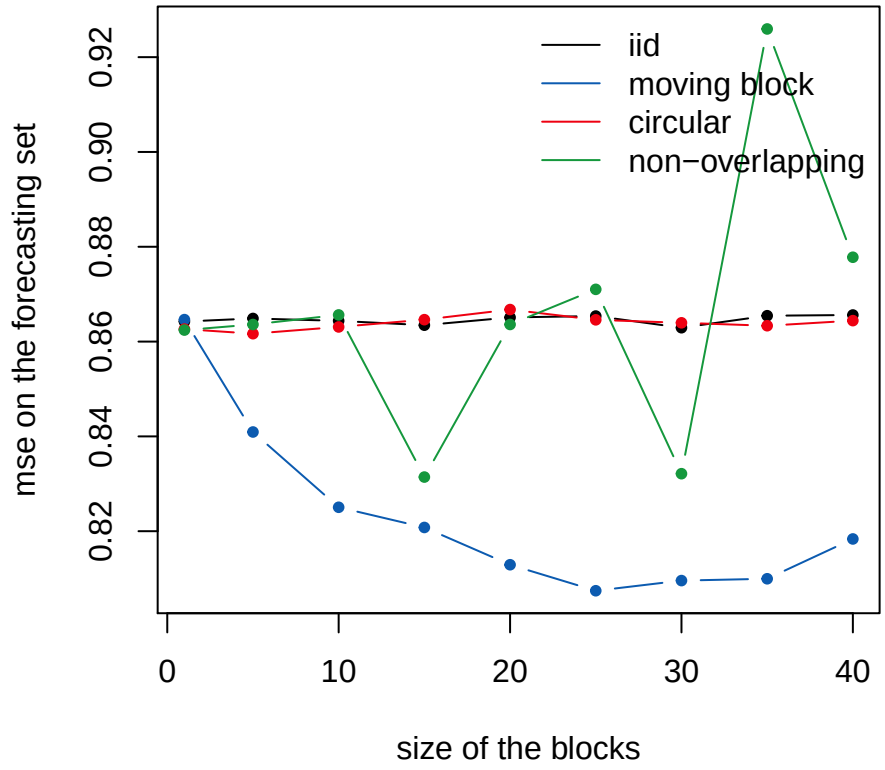

Figure 2: mse on the forecasting set in function of the size of the blocks, sigma=1/2

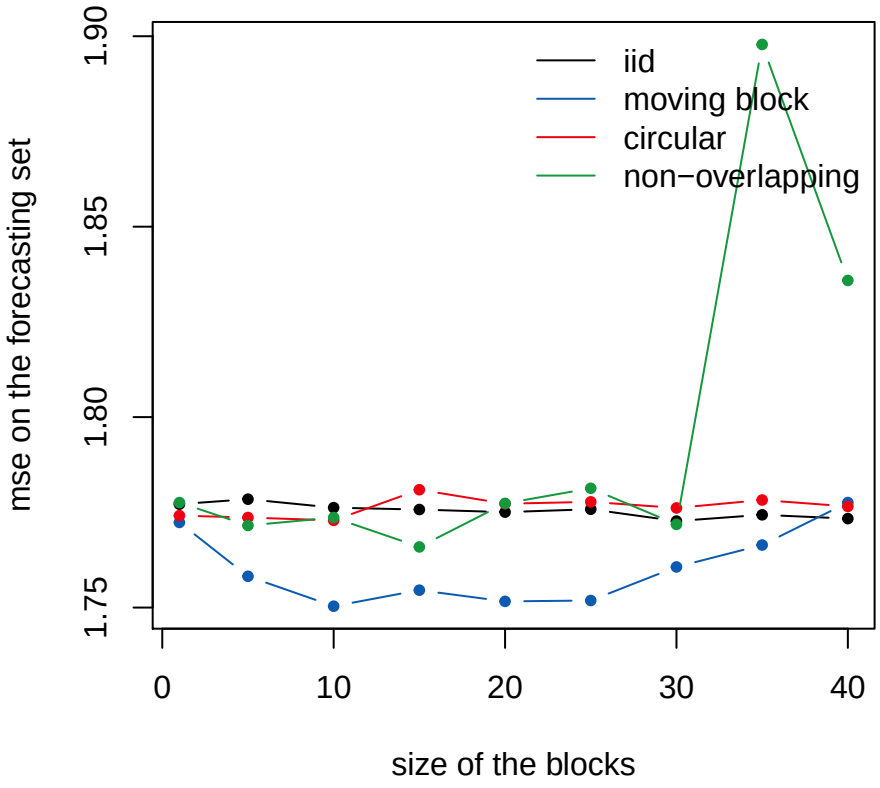

Figure 3: mse on the forecasting set in function of the size of the blocks, sigma=1