Réduction optimale de données

On note Y_t la mesure du CN2 à l'instant t correspondant à la hauteur à l'instant t. On a un échantillon $Y_{t_1}, \cdots, Y_{t_n}$. On suppose que les Y_{t_i} sont décorrélés, gaussiens et de même écart type σ (hypothèses fortes). Elles peuvent être issues d'une fonction non linaire quelconque d'un état caché (issues par exemple d'une modélisation paramétrique du CN2). Le but est de réduire la taille de l'échantillon sans trop de perte d'information et en particulier de préserver les moments. En fait, ici l'échantillon $(\tilde{Y}_1, \cdots, \tilde{Y}_m)$ réduit ne sera pas pris dans l'échantillon original mais sera généré en faisant une combinaison linéaire à partir des échantillons initiaux.

Let $1 \leq m \leq n$ be the new number of observations. The new set of observations denoted by $(\tilde{Y}_1, \cdots, \tilde{Y}_m)$ will be given as a linear combination of the original set of observations $(Y_{t_1}, \cdots, Y_{t_n})$,

$$\tilde{Y}_i = \frac{1}{\|\Phi_i\|^2} \sum_{t=t_1}^{t_n} \Phi_i(t)\, Y_t \ , \tag{1}$$

avec $\|\Phi_i\|^2 = \sum_{t=t_1}^{t_n} \Phi_i^2(t)$ where Φ_i is the Lagrange polynomial of degree $m-1$. For any $1 \leq i \leq m$, let us introduce Φ_i the Lagrange polynomial of degree $m-1$ such that

$$\Phi_i(t) = \begin{cases} 1 & \text{if} \quad t = T_i \\ \\ 0 & \text{if} \quad t = T_j \neq T_i \ . \end{cases} \tag{2}$$

where $T_1 \leq \cdots \leq T_m$ are the roots of the Legendre polynomial Ψ_m of degree m. One can prove that for any $i \neq j$ with $1 \leq i, j \leq m$, T_i is a real such that

$$t_1 \leq T_i \leq t_n \quad \text{and} \quad T_i \neq T_j \ .$$

If $m = n$, then $T_i = t_i$, for any $1 \leq i \leq m = n$.

On a :

$$\langle \Psi_m, t^j \rangle = 0 \ , \tag{3}$$

for any integer $0 \leq j \leq m-1$. Avec le produit scalaire :

$$\langle f, g \rangle = \sum_{t=t_1}^{t_n} f(t)\, g(t) \ , \tag{4}$$

On montre que (cf articles joints) les $\tilde{Y}_i$ sont indépendants gaussiens d'écart-type réduit : $\frac{\sigma}{\|\Phi_i\|_2}$ De plus, l'information contenue dans ces nouvelles

mesures est proche de celle contenue dans les mesures initiales. En particulier, si l'on travaille avec cet échantillon réduit ($m << n$), on préserve les moments. Le polynôme de Legendre et ses racines se calculent rapidement avec des commandes Matlab (https ://fr.mathworks.com/help/symbolic/sym.legendrep.html) sur $[-1, 1]$.

```
syms x
roots = vpasolve(legendreP(7,x) == 0)
```

donne les racines du polynôme de Legendre de degré 7 sur [-1 1]. Pour un intervalle quelconque $[t_1, t_n]$ avec des pas d'échantillonnage quelconque, il suffit de translater et dilater les polynômes et donc les racines. Si le pas d'échantillonage temporel (ou des hauteurs) est toujours le même, les racines sont calculées une fois pour toute.

Si tu codes cette méthode pour tes CN2, vérifie bien (3) et

$$\sum_{i=1}^{m} \|\Phi_i\|_2^2 = n \tag{5}$$

Bien sûr, les hypothèses ne sont pas vérifiées pour ton cas où en particulier tu as des non stationnarités (sauts) et des corrélations (tu peux virer des CN2 proches avant) mais tu peux tester. Il faut un bon compromis : $m = n$ donne l'échantillon initial et à mon avis $m < 50$ suffit largement.