M2 probabilités et statistiques des nouvelles donnés

Modèle de suite gaussienne

Djamila AZZOUZ
RAZAFIMANDIMBY Henimpitahiana

Professeur référent
Mohamed Hebiri

Année académique 2022 - 2023

I Les différentes de méthodes de seuillage

Objectif : Dans cette partie, nous allons estimer nos coefficients θ_j^* par différentes méthodes de seuillages(fort, faible, non-négative-garrotte). L'idée consiste donc à ne pas tenir compte des coefficients θ_j^* de faible valeur, en considérant qu'ils transcrivent du bruit plus qu'une information sur le signal.
Le principe du seuillage est donc de se donner un seuil τ à partir duquel on peut conserver nos coefficients dans l'estimation.
Supposons que l'on observe $y_1....y_d$ vérifiant le modèle de suite :

$$y_j = a\eta_j + \xi_j$$

où $j = 1...d$, $a \in R$ et $\eta_j \in \{0, 1\}$ des paramètres inconnus tel que : $\sum_{j=1}^{d} \eta_j = [d^{1-\beta}]$

Nous allons passer à l'application numérique : prenons $d = 50$, $\beta = 0.3$ et on fait varier a de 1 à 10. On fixe le seuil $\tau = \sqrt{2 \log d}$. Notons que $\theta^* = a.(\eta_1...\eta_d)^T$.
Pour $j \in \{1, ..., d\}$ on définit les différents estimateurs suivants :

— l'estimateur par seuillage fort qui est défini par : $\hat{\theta}^H = y_j \mathbb{1}_{\{|y_j|>\tau\}}$

— l'estimateur par seuillage faible $\theta^S = y_j.(1 - \frac{\tau}{|y_j|})_+$

— l'estimateur par seuillage non-négative garrotte $\theta^{NG} = y_j.(1 - \frac{\tau^2}{y_j^2})_+$

Traçons d'abord sur le même graphique les fonctions $\hat{\theta}^H$, $\hat{\theta}^S$ et $\hat{\theta}^{NG}$

En calculant la valeur de τ qui est 2.797, on remarque qu'entre $-\tau$ et τ les y_j se retrouvent à 0

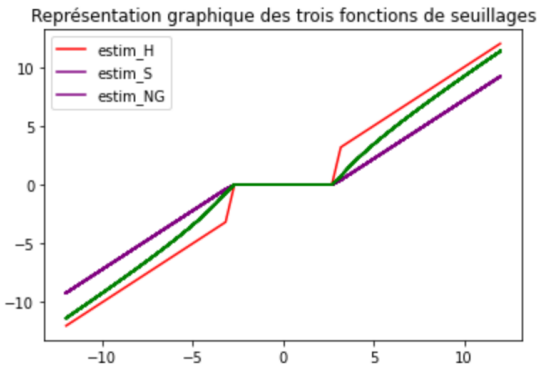

FIGURE 1 – Traçage des fonctions $\hat{\theta}_j^H(y_j), \hat{\theta}_j^S(y_j)$ et $\hat{\theta}_j^{NG}(y_j)$

Maintenant notons $\mathcal{R}(\hat{\theta}, a) = \|\hat{\theta} - \theta^*\|_2^2$ le risque quadratique de l'estimateur $\hat{\theta}$. Nous allons représenter graphiquement pour chacun des estimateurs par seuillage fort(ou dur), faible et non-negative garotte, les quantités $\mathcal{R}(\hat{\theta}, a)$ en variant la valeur de a de 1 à 10

I.1 Estimateur par seuillage fort(dur) $\hat{\theta}^H$

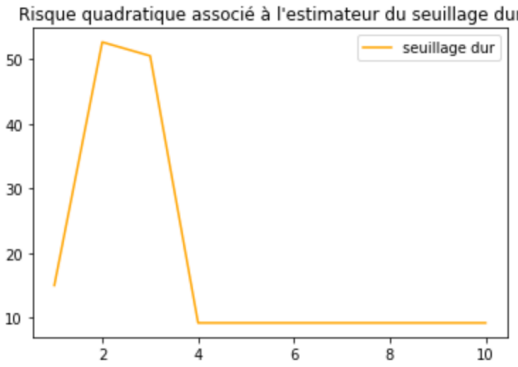

FIGURE 2 – Quantités de $\mathcal{R}(\hat{\theta}, a)$ pour l'estimateur par seuillage fort

La méthode par seuillage fort nous permet de voir que le risque commence à augmenter pour $a = 1$ et même on arrive à avoir un pic quand a est entre 2 et 4. Puis dès qu'il atteint le pic il diminue directement et reste stable pour $a = 4$ ce qui signifie qu'on arrive vite à écraser le bruit.

I.2 Estimateur par seuillage faible $\hat{\theta}^S$

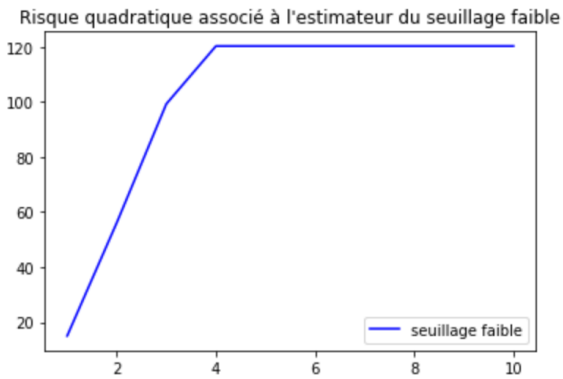

FIGURE 3 – Quantités de $\mathcal{R}(\hat{\theta}, a)$ pour l'estimateur par seuillage faible

Pour l'estimateur par seuillage faible, pour $a = 4$ on voit qu'il y a un pic de risque et ce-ci reste stable à $a = 6$ en restant élevé qui est du aux données qui sont bruitées.

I.3 Estimateur par seuillage non-negative garotte $\hat{\theta}^{NG}$

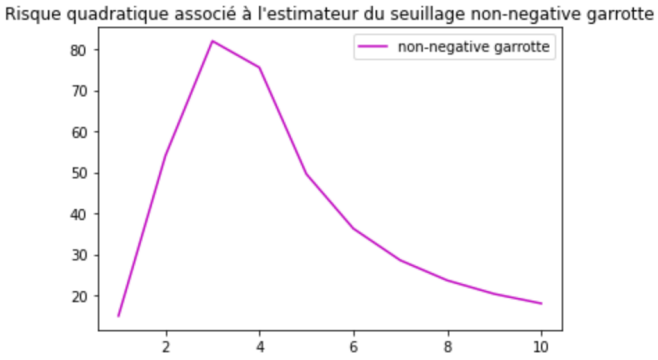

FIGURE 4 – Quantités de $\mathcal{R}(\hat{\theta}, a)$ pour l'estimateur par seuillage non-negative garotte

On remarque que pour l'estimateur par seuillage non-negative garrotte, pour $a = 4$ il y a toujours un pic. Puis le risque diminue quand la valeur de a augmente i.e plus grand que 4 mais moins rapide que dans le risque du seuillage dur, ceci signifie aussi qu'on arrive aussi à écraser le bruit mais de façon lente.

Remarque : En effet, nous constatons que le risque quadratique quadratique $\mathcal{R}(\hat{\theta}, a)$ de l'estimateur $\hat{\theta}$ dépend de la valeur de a. Mais parmi les estimateurs de seuillages qu'on a vu précédemment, on voit que la méthode par seuillage fort est plus fiable et efficace que les deux autres méthodes car le risque atteint son pic puis diminue avec une stabilité quand a est supérieur ou égale à 6.

Maintenant, on va procéder à la méthode de sélection de variables de notre modèle(MSG) qui revient à déterminer les coordonnées non nulles du vecteur θ^* et étudier le risque de sélection de ces variables.
On définit donc le risque de selection de variables par :

$$\mathcal{R}^{MS}(\hat{\theta}, a) = \sum_{j=1}^{M} |\eta_j - \hat{\eta}_j|$$

La valeur du risque de sélection de variables est

$$[15.0, 11.0, 5.0, 4.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]$$

Les quatre premières variables sont non nulles. Autrement dit, le risque est élevé pour la première variable. Elle diminue pour la deuxième, la troisième et la quatrième et à partir de la cinquième variable il est nul. L'écart-type du risque MS est 5.287721626560915 ; et le risque moyen de la sélection de variables est 3.2
La figure 5 nous confirme qu'à partir de la quatrième variable le risque est nul.

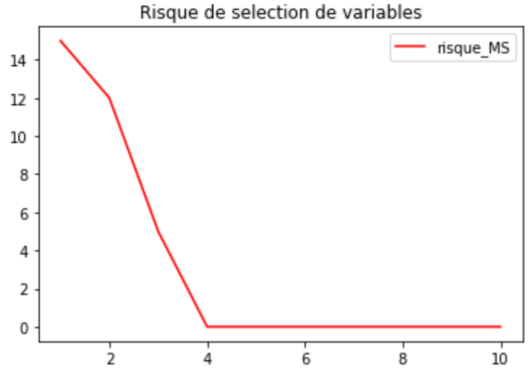

FIGURE 5 – Risque pour la sélection de variables

Remarque : En regardant la quantité du risque pour l'estimateur par seuillage fort $\mathcal{R}(\hat{\theta}, a)$ et la quantité du risque de sélection de variables $\mathcal{R}^{MS}(\hat{\theta}, a)$, les méthodes en effet sont efficaces et donnent les mêmes résultats.

II Méthode de détection de ruptures

Dans cette partie, nous allons étudier le problème de détection de ruptures, autrement dit nous allons estimer les instants où le signal présente un changement dans sa distribution.
Nous disposons donc des observations $y_1, ...y_d$, vérifiant le modèle MSG suivant :

$$y_j = \theta_j^* + \varepsilon.\xi_j$$

où ξ_j est le bruit qui suit une loi $N(0,1)$, $\theta^* = (\theta_1^*, ...\theta_d^*)^T$ et $\varepsilon = \frac{1}{\sqrt{d}}$
Comme nous l'avons déjà mentionnés, on doit détecter les instances de ruptures de notre θ_j^*. Regardant alors la présentation graphique de ce vecteur :

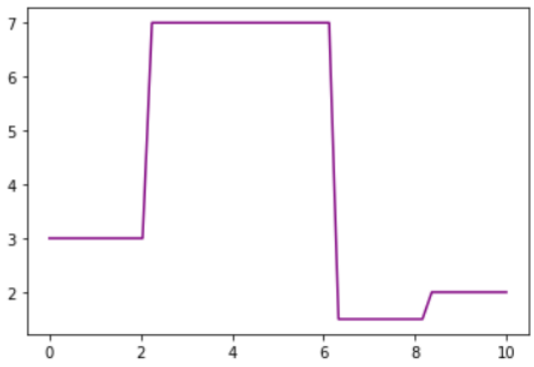

FIGURE 6 – Représentation graphique des θ_j^*

En observant le graphe associé à θ_j^*, on détecte trois ruptures :

— la première : en 2

— la deuxième entre 6 et 7

— la troisième entre 8 et 9

Nous allons introduire dans un premier temps l'ensemble de sparsité :

$$J^* = \{j \in \{1, ..., d-1\}, \Delta_j^* \neq 0\}$$

où $\Delta_j^* = y_{j+1} - y_j$, il dépend que de nos données observées.

Vérifions que le cardinal de J^* égal au moins 2.

En observant le graphe si dessus qui représente les différentes valeurs prises par Δ_j^*, on voit bien qu'on a réussit à détecter au moins deux ruptures. Cependant, on voit toujours la présence du bruit ξ_j, que nous devons éliminer. Pour cela nous allons estimer Δ_j^* par le seuillage dur défini précédemment dans l'exercice 1 et regardant ce qui se passe.

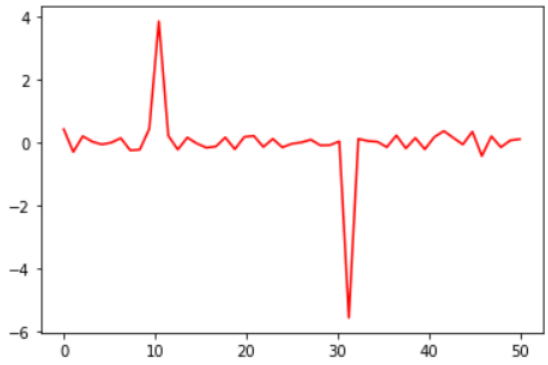

FIGURE 7 – Représentation graphique des ruptures

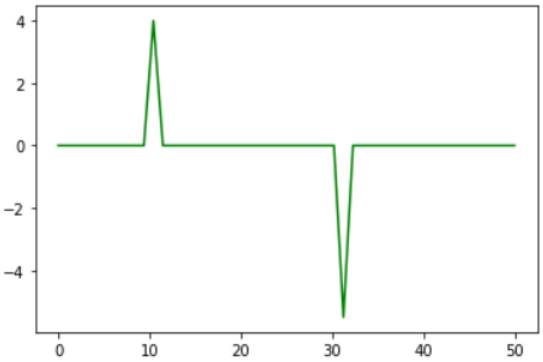

FIGURE 8 – Détection de ruptures

Observation : Dans la figure 7, on a deux ruptures avec les bruits. Puis la figure 8, en estimant notre Δ_j^*, on arrive à avoir seulement les lieux de ruptures de signal et grâce à la méthode par seuillage dur le bruit a été éliminé.

III Traitement de donnée réelle pour la détection de rupture

Dans cette partie, nous allons appliquer les résultats obtenus dans l'exercice 2 sur le jeu de données **"Émissions de CO2 et de polluants des véhicules commercialisés en France pour l'année 2014"**, que nous avons récupérés sur le site **Data.gouv.fr**.

Nous avons choisit de prendre la variable explicative "$conso - mixte$" qui représente la consommation extra urbaine de carburant(en l/100km). Notre but est de détecter les lieux de ruptures du signal.

Rappelons que la consommation normale du carburant en 2014 pour le **Diesel** était de $4,4\ l/km$ et pour **l'essence** $5,10\ l/km$.

Prenons 50 observations pour montrer que notre sparsité vaut bien au moins 2 ou 3 comme l'exo 2.

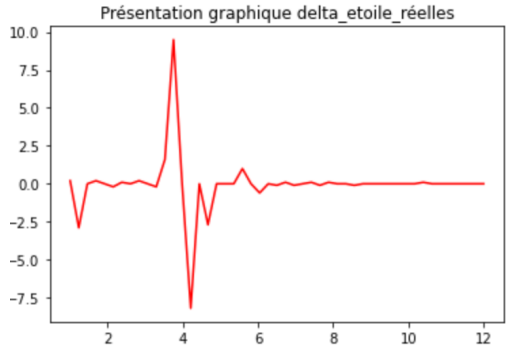

FIGURE 9 – Représentation graphique des ruptures

FIGURE 10 – Détection de ruptures

Observations : La figure 9 nous montre qu'on a au moins trois ruptures avec les bruits. Après avoir utiliser la méthode par seuillage dur pour notre estimateur, le bruit a été supprimé(figure 10). On voit maintenant qu'on a que trois ruptures, la première est en mois de janvier, la seconde entre mars et avril et la dernière entre le avril et mai. Autrement dit, il y a une baisse de consommation du carburant pour 100 km en janvier puis une surconsommation entre mars et avril , après encore une baisse et enfin la consommation devient stable pour le reste jusqu'à la fin d'année.
En effet, on a trois sparsités, donc le cardinal de $\hat{J}^*$ est 3 et les valeurs non nuls de notre ensemble sont :

$$\hat{J}^* = \{-2.900000095, 9.499999049, -8.199998858999999\}$$

Remarque : La cardinal de notre ensemble change en augmentant l'échantillon, il faudra donc calculer un nouveau seuil τ selon la taille du nouveau échantillon. Nous allons mettre deux figures de détection de ruptures pour $d = 300$.

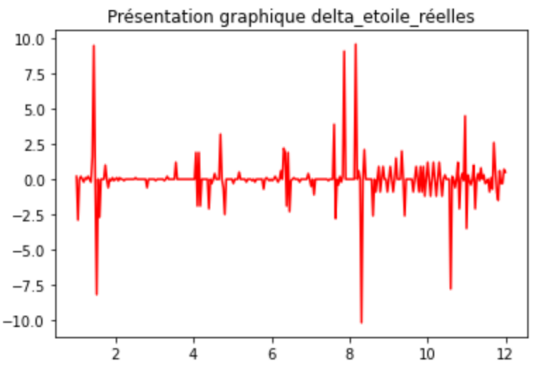

FIGURE 11 – Représentation graphique des ruptures

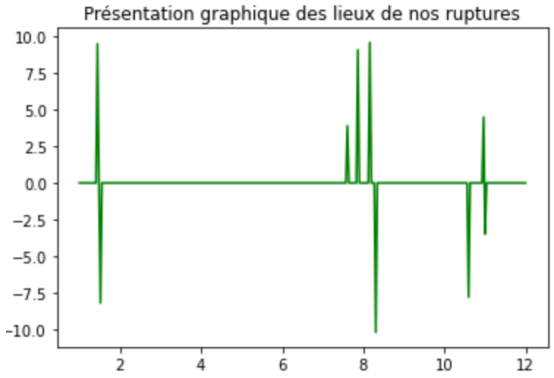

FIGURE 12 – Détection de ruptures

Observation : On a dans la figure 12 neuf ruptures, i.e le nombre de sparsité est neuf : il y a eu une surconsommation du carburant en janvier, entre juillet et août, en août exactement puis une baisse de consommations entre janvier et février, entre août et septembre, en novembre 2014.

Conclusion : Pour étudier la prédiction d'un modèle de suite gaussienne, il nous faut un bon estimateur . Pour cela la méthode par seuillage dur ou fort est la plus efficace. Et pour minimiser le risque, soit on utilise le risque quadratique $\mathcal{R}(\hat{\theta}, a) = \|\hat{\theta} - \theta^*\|_2^2$ soit le risque $\mathcal{R}^{MS}(\hat{\theta}, a) = \sum_{j=1}^{M} |\eta_j - \hat{\eta}_j|$. On constate aussi que notre meilleur estimateur nous aide à sélectionner les θ_j^*.
Puis on a vu dans l'exercice 2 la méthode de détection de ruptures qui nous a permis d'estimer le nombre de sparsité. Et grâce à l'estimateur par seuillage dur on pourrait supprimer les bruits. En connaissant les variables explicatives incluses dans le modèle on a introduit l'ensemble de sparsité qu'on a utilisé pour la sélection des"θ_j^*" non nuls. Ceci nous permet de connaître les variables qui sont dans le modèle.
Enfin, pour appliquer cette méthode de détection de ruptures nous avons choisit les données concernant les "Émissions de CO2 et de polluants de véhicules commercialisés en France pour l'année 2014". On remarque que plus on augmente le nombre d'observations plus on aura de sparsité. Nous pouvons dire que le problème de détection de ruptures est un problème de régression ayant pour but d'estimer les instants où un signal présente des changements dans la distribution.