CPdetect : Un package R pour la détection des cassures structurelles par segmentation linéaire

Aristide Houndetoungan[1] Arnaud Dufays[2]

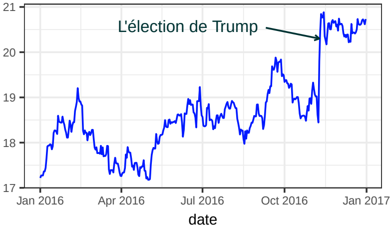
[1] Université Laval

[2] Université de Namur

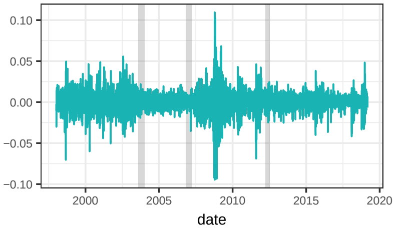
15 Mai 2019

Cassures structurelles

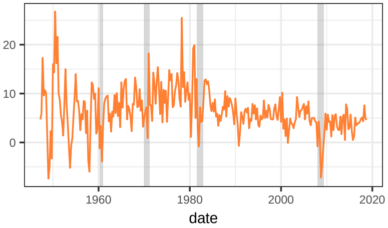
1 $US en Peso Mexicain



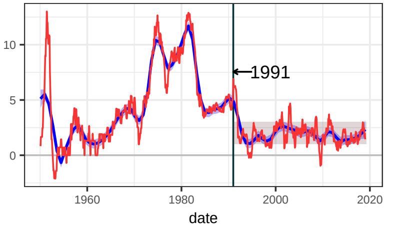
S&P500 log−rendements journaliers



Taux de croissance du PIB Américain



Inflation Canadienne



Plan de la présentation

Spécification du modèle

- Soit $y_{1:T} = \{y_1, \ldots, y_T\}$ une série temporelle observée sur T périodes de m régimes définie par :

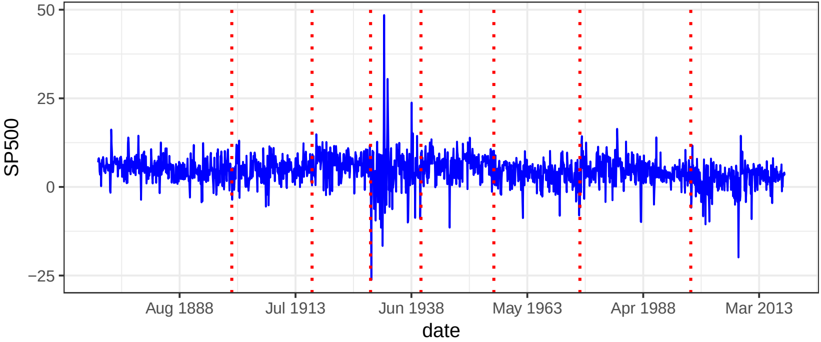
$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_i + \varepsilon_t, \quad \text{si} \quad \tau_{i-1} < t \leq \tau_i$$

- Le régime i est délimité par les dates τ_{i-1} et τ_i, où $\tau_0 = 0$, $\tau_m = T$ et $\tau_i < \tau_{i+1}\ \forall\ i \in [0, m-1]$.
- $\mathbf{x}_t$ contient l'ensemble des variables explicatives de y_t à la période t. Supposons qu'il y a K variables explicatives dans $\mathbf{x}_t$.
- $\boldsymbol{\beta}_i$ est le paramètre associé aux variables explicatives sur le régime i.
- $\varepsilon_t \sim \text{m.d.s}(0, \sigma)$est une séquence de différence martingale.

Exemple : Rendements mensuels du S&P500

- log-rendements du S&P500 de 1871 à 2018 (8 régimes) :

$$y_t = \beta_{0,i} + \beta_{1,i} y_{t-1} + \beta_{2,i} y_{t-2} + \varepsilon_t \quad \text{où} \quad \varepsilon_t \sim N(0, \sigma_i^2),\ t \in [\tau_{i-1}+1, \tau_i]$$



- Yau et Zhao 2016. Inference for multiple change points in time series via likelihood ratio scan statistics. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 78(4), 895-916.

Exemple : Rendements mensuels du S&P500

- log-rendements du S&P500 de 1871 à 2018
- Estimation du modèle par segment (fonction `lm` en boucle), 32 coefficients

```
# Résultat
```

```
##                         Intercept   AR1    AR2
## 1871-02 à 1899-11      3.845 0.412 -0.155
## 1899-12 à 1917-02      3.393 0.202  0.092
## 1917-03 à 1929-09      5.774 0.223 -0.104
## 1929-10 à 1940-07      3.707 0.354 -0.181
## 1940-08 à 1956-03      4.709 0.211  0.044
## 1956-04 à 1974-09      2.499 0.251  0.012
## 1974-10 à 1998-07      3.717 0.302 -0.101
## 1998-08 à 2018-09      1.808 0.252 -0.072
```

```
# Problème de surparamétrisation
# Estimations potentiellement imprécises
```

Notre approche

Critique de l'approche séquentielle

- Toutes les K composantes de $\boldsymbol{\beta}_i$ changent entre deux régimes consécutifs.
- Problème de surparamétrisation.
- Estimations potentiellement imprécises.

Question d'intérêt

- Etant donné $\boldsymbol{\tau} = \{\tau_0, \ldots, \tau_m\}$, quelles sont les composantes de $\boldsymbol{\beta}_i$ qui changent réellement lorsqu'on passe du premier régime au deuxième, du deuxième au troisième, ...?

Notre approche

1. Contrôle le problème de surparamétrisation.
   - Pour chaque nouveau régime, seuls les paramètres qui ont subi une variation significative vont réellement varier.
2. Utilise un ensemble de cassures structurelles potentielles comme input.
   - N'importe quelle méthode de détection de cassures dans la littérature peut être utilisée (Ex. Bai et Perron, 1998 ; Yau et Zhao, 2016 ; ...).
   - L'ensemble peut contenir outre les vraies cassures, quelques cassures fallacieuses.

- **Nouvelle approche implémentée dans un package R, CPdetect, que nous développons** (exécution très rapide, code optimisé en C++).

Notre approche

```
## Method:  PSELO
## Variance:  constant
## dependent variable: SP500
## Sample size N: 1772
## AR order: 2
## Intercept term: yes
## Exogenous variables: (0)
##
## ******
## Parameter a
##   Intercept         AR1         AR2
## 0.042514036 0.005564119 0.006653331
##
## Parameter lambda
## [1] 65
##
##
## Number of regimes
## Intercept       AR1       AR2        se
##         7         1         1         1
##
## Coefficients in difference:
##                           Intercept       AR1         AR2
## 1871-02-02 | 1899-11-02 |  3.9886456 0.2978332 -0.08929543
## 1899-12-02 | 1917-02-02 |  0.0000000 0.0000000  0.00000000
## 1917-03-02 | 1929-09-02 |  1.2064181 0.0000000  0.00000000
## 1929-10-02 | 1940-07-02 | -1.6519697 0.0000000  0.00000000
## 1940-08-02 | 1956-03-02 |  1.4504610 0.0000000  0.00000000
## 1956-04-02 | 1974-09-02 | -2.3015416 0.0000000  0.00000000
## 1974-10-02 | 1998-07-02 |  0.9864619 0.0000000  0.00000000
## 1998-08-02 | 2018-09-02 | -1.9329986 0.0000000  0.00000000
##
## Coefficients in level:
##                          Intercept       AR1         AR2
## 1871-02-02 | 1899-11-02 |  3.988646 0.2978332 -0.08929543
## 1899-12-02 | 1917-02-02 |  3.988646 0.2978332 -0.08929543
## 1917-03-02 | 1929-09-02 |  5.195064 0.2978332 -0.08929543
## 1929-10-02 | 1940-07-02 |  3.543094 0.2978332 -0.08929543
## 1940-08-02 | 1956-03-02 |  4.993555 0.2978332 -0.08929543
## 1956-04-02 | 1974-09-02 |  2.692013 0.2978332 -0.08929543
## 1974-10-02 | 1998-07-02 |  3.678475 0.2978332 -0.08929543
## 1998-08-02 | 2018-09-02 |  1.745477 0.2978332 -0.08929543
##
## Residual standard error (se.):  3.75
## log-likelihood:  -4848.3285
## Penalized log-likelihood:  -5253.32
## Marginal likelihood (%):  35.364
```

Plan de la présentation

Approche exhaustive

- Tester toutes les possibilités.
- Si par exemple on a un modèle avec 2 paramètres et une cassure, on aura 4 possibilités à tester.
  1. Deux paramètres sans cassure (peu probable) ;
  2. Cassure dans le premier paramètre seul ;
  3. Cassure dans le second paramètre seul ;
  4. Cassure dans les deux paramètres.
- Au lieu de choisir une seule possibilité, nous proposons un critère convergent qui permet de comparer les modèles sous forme de probabilité.
- Tous les modèles peuvent être utilisés avec leur probabilité pour la prévision.

Approche exhaustive avec **CPdetect**

Soit le processus y_t suivant :

$$y_t = \begin{cases} 1.35 + \varepsilon_t & \text{si } 1 \leq t \leq 205 \\ 1.35 + 0.7 y_{t-1} + \varepsilon_t & \text{si } 206 \leq t \leq 350 \end{cases}$$

où $\varepsilon_t \sim N(0,1)$.

```r
# Simulation du processus
set.seed(2019)
library(CPdetect)
N         <- 350
y1        <- numeric(N)
y1[1:205] <- 1.35 + rnorm(205)
for (t in 206:N) {
  y1[t]   <- 1.35 + 0.7*y1[t - 1]  + rnorm(1)
}
```

Approche exhaustive avec **CPdetect**

```r
# Estimation de l'ensemble de cassures potentielles
cas.pot1 <- detectcp(formula = y1 ~ 1, pmax = 3)
summary(cas.pot1)

##
## method: Yau and Zhao (2016)
## dependent variable: y1
## Sample size N: 350
## Optimal AR order*: 1
## Intercept term: yes
## Exogenous variables: (0)
##
## Optimal window radius h: 43
## Number of regimes: 4
## Method: Yau and Zhao (2016)
## Variance: dynamic
##
##
##  Start End
##      1  43
##     44 207
##    208 306
##    307 350
##
## (*) The optimal AR order p is determined by minimizing the MDL statistic such that p<4
```

Approche exhaustive avec **CPdetect**

```
exh1     <- cplm(cas.pot1) # Estimation du modèle par l'approche exhaustive

## 64 homoscedactic models

summary(exh1) # Affiche le meilleur modèle mais il y a des options pour demander d'autres

## Bayesian estimation
## dependent variable: y1
## Sample size N: 350
## AR order: 1
## Intercept term: yes
## Exogenous variables: (0)
##
## ******
## Number of regimes
## Intercept       AR1        se
##         1         2         1
##
## Coefficients in difference:
##             Intercept         AR1
##   1 |  43 |  1.315587 -0.04555363
##  44 | 207 |  0.000000  0.00000000
## 208 | 306 |  0.000000  0.70645142
## 307 | 350 |  0.000000  0.00000000
##
## Coefficients in level:
##                    Intercept         AR1
##   1 |  43 | coef | 1.3155871 -0.04555363
##           |   sd | 0.1021573  0.06847603
##  44 | 207 | coef | 1.3155871 -0.04555363
##           |   sd | 0.1021573  0.06847603
## 208 | 306 | coef | 1.3155871  0.66089779
##           |   sd | 0.1021573  0.03062641
## 307 | 350 | coef | 1.3155871  0.66089779
##           |   sd | 0.1021573  0.03062641
##
## Residual standard error (se.):  0.97
## Marginal likelihood (%):  63.749
```

Approche sélective

- L'approche exhaustive est seulement possible en petite dimension.
  - Nombre de modèles à considérer : $2^{(m-1)K}$, où K est le nombre de variables explicatives et m le nombre de régimes.
- En grande dimension, nous réécrivons le modèle en différence première des paramètres :

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_1 + \mathbf{x}_t'\left(\sum_{j=2}^{m} \Delta\boldsymbol{\beta}_j \mathbb{I}(t \geq \tau_{j-1})\right) + \sigma\varepsilon_t, \quad \text{pour} \quad \tau_{i-1} < t \leq \tau_i$$

$$\mathbf{y} = \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$$

- Les paramètres en niveau sont obtenus par $\boldsymbol{\beta}_k = \boldsymbol{\beta}_1 + \sum_{j=2}^{k} \Delta\boldsymbol{\beta}_j$.
- Pénaliser la différence première de chaque paramètre

Approche sélective avec **CPdetect**

Soit le processus y_t suivant :

$$y_t = \begin{cases} 1.35 + 2.5x_{1t} - 1.9x_{2t} - 3x_{3t} + \varepsilon_t & \text{si } 1 \leq t \leq 205 \\ 1.35 + 0.7y_{t-1} + 0.5x_{1t} - 1.9x_{2t} + \varepsilon_t & \text{si } 206 \leq t \leq 350 \end{cases}$$

où $\varepsilon_t \sim N(0, 1)$.

```
# Simulation du processus
y2          <- numeric(N)
X           <- cbind(rnorm(N, 0, 2), rpois(N, 3), runif(N))
y2[1:205] <- 1.35 + X[1:205,] %*% c(2.5, -1.9, -3) + rnorm(205)
for (t in 206:N) {
  y2[t]    <- 1.35 + 0.7*y2[t - 1]  + sum(X[t,]*c(0.5, -1.9, 0)) + rnorm(1)
}
# Estimation de l'ensemble de cassures potentielles
cas.pot2  <- detectcp(formula = y2 ~ X, pmax = 3)
# Sélection
sel2      <- selectcp(cas.pot2)
# Estimation Bayésienne de l'approche sélective
exh2      <- cplm(sel2)

## 8 homoscedactic models
```

Approche sélective avec **CPdetect**

```
summary(sel2)

## Method:  PSELO
## Variance:  constant
## dependent variable: y2
## Sample size N: 350
## AR order: 1
## Intercept term: yes
## Exogenous variables: (3)  X1 X2 X3
##
## ******
## Parameter a
##   Intercept         AR1          X1          X2          X3
## 0.030854970 0.001679121 0.005418199 0.005652651 0.038209999
##
## Parameter lambda
## [1] 20.121
##
##
## Number of regimes
## Intercept       AR1        X1        X2        X3        se
##         1         2         2         1         2         1
##
## Coefficients in difference:
##             Intercept          AR1        X1        X2        X3
##   1 | 103 |  1.039939 -0.002832586  2.556199 -1.844277 -2.792047
## 104 | 205 |  0.000000  0.000000000  0.000000  0.000000  0.000000
## 206 | 350 |  0.000000  0.688702761 -2.035260  0.000000  2.751192
##
## Coefficients in level:
##             Intercept          AR1        X1        X2          X3
##   1 | 103 |  1.039939 -0.002832586 2.5561988 -1.844277 -2.79204679
## 104 | 205 |  1.039939 -0.002832586 2.5561988 -1.844277 -2.79204679
## 206 | 350 |  1.039939  0.685870175 0.5209387 -1.844277 -0.04085445
##
## Residual standard error (se.):  0.97
## log-likelihood:  -482.7646
## Penalized log-likelihood:  -560.58
## Marginal likelihood (%):  80.185
```

Plan de la présentation

Stratégies de Hedge Funds

- Base de données Credit Suisse.
- Données mensuelles, 14 stratégies et plusieurs facteurs de risque.
- Déterminer les facteurs de risque explicatifs de l'indice *Fixed-income arbitrage* (FIA).
- Etude déjà réalisée par Fung et Hsieh (2004) où un modéle linéaire sans cassure structurelle est estimé.
- Méthode Yau et Zhao (2016) pour détecter les cassures potentielles : 4 regimes
  - Modèle AR(1) avec facteurs de risque exogénes, au total 9 variables explicatives.
    - Nombre de modèles à considérer : $2^{3\times 9} = 134\ 217\ 728$.
  - Approche sélective, très rapide (30 secondes).

Stratégie FIA

				Approche par segment					
Période	Int.	AR1	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM
1994.03 à 1999.04	0.40	0.31	0.06	0.05	-1.93	-10.90	-0.00	-0.01	0.00
	(0.14)	(0.09)	(0.03)	(0.04)	(0.74)	(1.69)	(0.01)	(0.01)	(0.01)
1999.05 à 2007.06	0.39	0.24	-0.01	-0.01	-0.58	-2.04	-0.00	0.01	0.01
	(0.08)	(0.09)	(0.02)	(0.02)	(0.36)	(0.69)	(0.00)	(0.00)	(0.00)
2007.07 à 2010.06	0.29	0.22	0.22	-0.24	-2.48	-4.08	-0.00	-0.03	-0.01
	(0.29)	(0.11)	(0.05)	(0.10)	(1.22)	(1.02)	(0.03)	(0.02)	(0.02)
2010.07 à 2016.03	0.15	0.43	0.06	-0.05	-0.39	-1.64	-0.00	-0.00	0.00
	(0.06)	(0.08)	(0.02)	(0.03)	(0.33)	(0.52)	(0.00)	(0.00)	(0.00)
				Approche sélective (prob = 92%)					
Période	Int.	AR1	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM
1994.03 à 1999.04	0.32	0.27	0.08	-0.00	-1.30	-10.16	-0.00	-0.01	0.00
	(0.06)	(0.04)	(0.03)	(0.02)	(0.29)	(1.21)	(0.00)	(0.01)	(0.00)
1999.05 à 2007.06			-0.01			-3.14		0.00	
			(0.02)			(0.41)		(0.01)	
2007.07 à 2010.06			0.23	-0.23				-0.03	
			(0.03)	(0.06)				(0.01)	
2010.07 à 2016.03			0.04	-0.01				-0.00	
			(0.03)	(0.05)				(0.00)	

Prédiction des indices

- Au total 6 modèles considérés.
- Backtest et calcul des RMSEs pour les 14 stratégies.
- Notre approche a 5 fois la plus petite valeur du RMSE.

Strat.	HFI	CNV	DSB	EME	EMN	EDR	EDD
Linéaire	1.33	**1.84**	2.71	2.31	4.16	1.31	1.16
CP	**1.18**	1.92	2.71	2.22	4.57	1.31	1.12
Linéaire - JBF	1.34	1.84	2.70	2.22	4.18	1.31	1.16
CP - JBF	1.27	2.31	2.77	**2.09**	4.87	1.36	1.09
TVP	1.45	1.91	2.90	2.48	**4.11**	1.48	1.47
Sel. seg.	1.27	1.89	**2.64**	2.14	28.90	**1.31**	**1.08**
Strat.	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Linéaire	**1.53**	0.97	1.32	1.92	1.72	3.19	**1.21**
CP	1.53	0.97	1.32	**1.68**	1.67	3.19	1.21
Linéaire - JBF	1.55	0.97	1.43	1.95	1.64	3.19	1.23
CP - JBF	1.56	0.97	1.60	1.73	1.71	3.19	1.42
TVP	1.85	1.19	**1.24**	2.12	1.71	3.84	1.61
Sel. seg.	1.55	**0.95**	1.36	1.71	**1.50**	3.19	1.25

Plan de la présentation

Incertitude sur les cassures

Soit le processus y_t suivant :

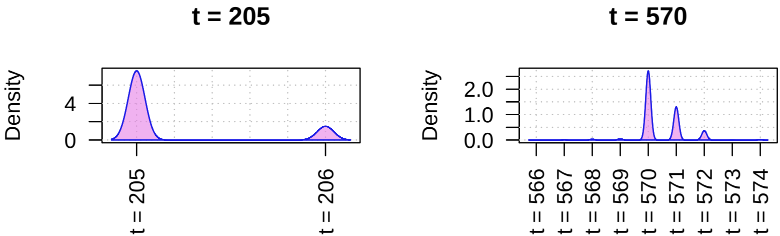
$$y_t = \begin{cases} 1.35 + 2.5v_{1t} - 3v_{2t} + \varepsilon_{1t} & \text{si } 1 \leq t \leq 205 \\ 1.35 + 0.7y_{t-1} + 0.5v_{1t} + \varepsilon_{1t} & \text{si } 206 \leq t \leq 350 \\ 1.35 + 0.7y_{t-1} + 0.5v_{1t} + \varepsilon_{2t} & \text{si } 351 \leq t \leq 570 \\ 0.7y_{t-1} + 1.5v_{1t} + v_{2t} + \varepsilon_{2t} & \text{si } 571 \leq t \leq 700 \end{cases}$$

où $\varepsilon_{1t} \sim N(0,1)$ et $\varepsilon_{2t} \sim N(0,4)$.

```
# Simulation du processus
y3        <- numeric(700)
V         <- cbind(rnorm(700, 0, 2), rpois(700, 3))
y3[1:205] <- 1.35 + V[1:205,] %*% c(2.5, -3) + rnorm(205)
for (t in 206:350) {
  y3[t]   <- 1.35 + 0.7*y3[t - 1]  + sum(V[t,]*c(0.5, 0)) + rnorm(1)
}
for (t in 351:570) {
  y3[t]   <- 1.35 + 0.7*y3[t - 1]  + sum(V[t,]*c(0.5, 0)) + rnorm(1, 0, 2)
}
for (t in 571:700) {
  y3[t]   <- 0.7*y3[t - 1]  + sum(V[t,]*c(1.5, 1)) + rnorm(1, 0, 2)
}
# Estimation Bayésienne de l'approche sélective
exh3      <- cplm(formula = y3 ~ V, pmax = 3, selection = TRUE)

## Potential change detection
## Change points selection
## Change points linear model estimation
## 10 homoscedactic models
```

Incertitude sur les cassures

```
inc3      <- rcp(exh3, R = 10)# Distribution postérieure des cassures
summary(inc3)

## Uncertainty of change points
##
## Method                   : D-DREAM
## Gelman and Rubin's R     : [1.002 1.007]
##
##               Mean Median Inf.CI.95% Sup.CI.95%
## t = 205 205.1658    205        205        206
## t = 570 570.4436    570        570        572

plot(inc3, type = "density")
```

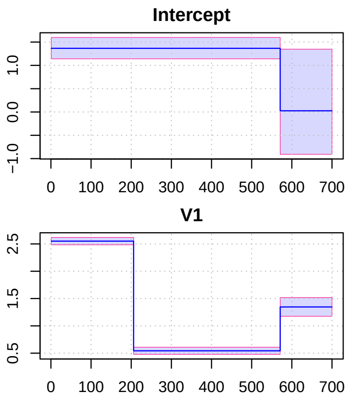


Hétéroscédasticité

```r
# Directement l'estimation
exh4     <- cplm(formula = y3 ~ V, pmax = 3, selection = TRUE, variance = "dynamic")

## Potential change detection
## Change points selection
## Change points linear model estimation
## 10 heteroscedactic models

# Affiche seulement la variance
summary(exh4)$models[[1]]$se

##                   se.
##   1 | 205 | 1.0492544
## 206 | 350 | 0.9329631
## 351 | 570 | 2.0627739
## 571 | 700 | 1.9223516

# Affiche les coefficients en première différence
summary(exh4)$models[[1]]$coefficients.in.diff

##              Intercept         AR1         V1        V2
##   1 | 205 |   1.368575 0.006927179  2.5507977 -3.014106
## 206 | 350 |   0.000000 0.695352600 -2.0073102  3.010723
## 351 | 570 |   0.000000 0.000000000  0.0000000  0.000000
## 571 | 700 |  -1.419790 0.000000000  0.7991528  1.042968
```

Hétéroscédasticité

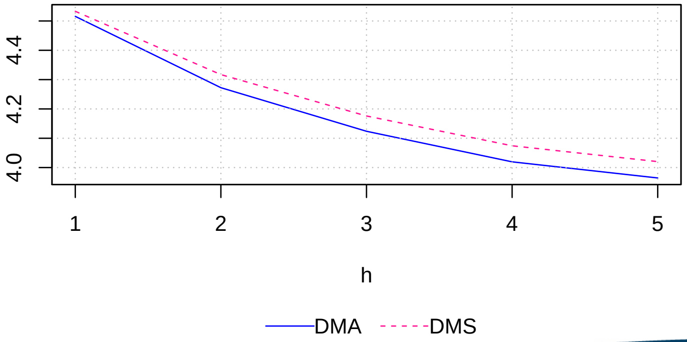
```r
# Afficher la dynamiaue des coefficients
par(mar = c(2, 4.1, 2, 2.1))
plot(exh4)
```



Prédiction

- Utiliser le premier modèle AR pour prédire à un horizon $h = 5$.
- Utiliser tous les 64 modèles (Avantage par rapport aux méthodes standard).

```
pred1    <- predict(exh1, h = 5)
plot(pred1, separate = FALSE)
```



Backtest

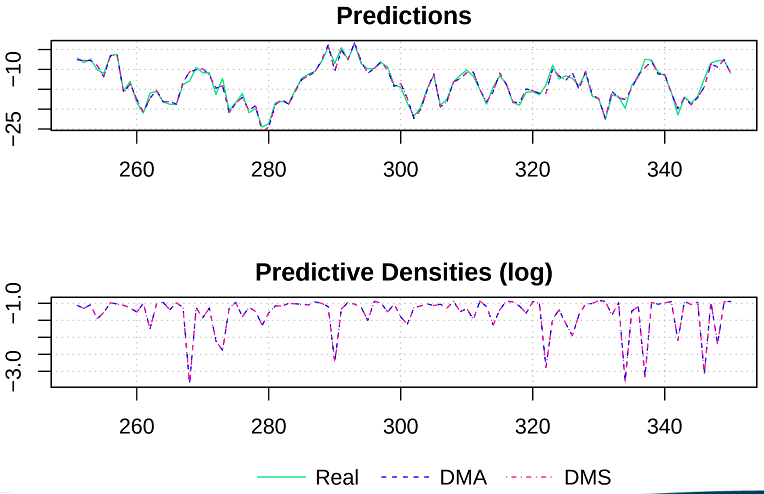
- Prédiction à chaque période pour un horizon $h = 1$.
- Comparaison avec la valeur réalisée et calcul du *Root Mean Square Error* (RMSE).

```
back2 <- backtest(formula = y2 ~ X, pmax = 3, selection = TRUE, nsample = 250)
```

```
print(back2)

## Sample size               : 350
## Training sample size      : 250
## Predicting Sample size    : 100
## Updating time             : 1
##
## Predictive density (log)
##                       DMA: -1.299
##                       DMS: -1.299
##
## Root-Mean-Square Error
##                       DMA: 0.854
##                       DMS: 0.854
```

Backtest

```r
plot(back2)
```



Conclusion

- Méthode de Segmentation Linéaire Sélective.
    1. Détecte les paramètres qui ont réellement changé lorsqu'on change de régime.
    2. Réduit les variations non importantes des paramètres à zéro.
- Contribution Empirique.
    1. Facilite l'interprétation des résultats.
    2. Meilleure performance de prévision.
- Extensions vers les modèles multivariés.

MERCI