Partitionnement de Séries Temporelles Basé sur la Forme des Séries

Brieuc Conan-Guez, Alain Gély, Lydia Boudjeloud-Assala, Alexandre Blansché

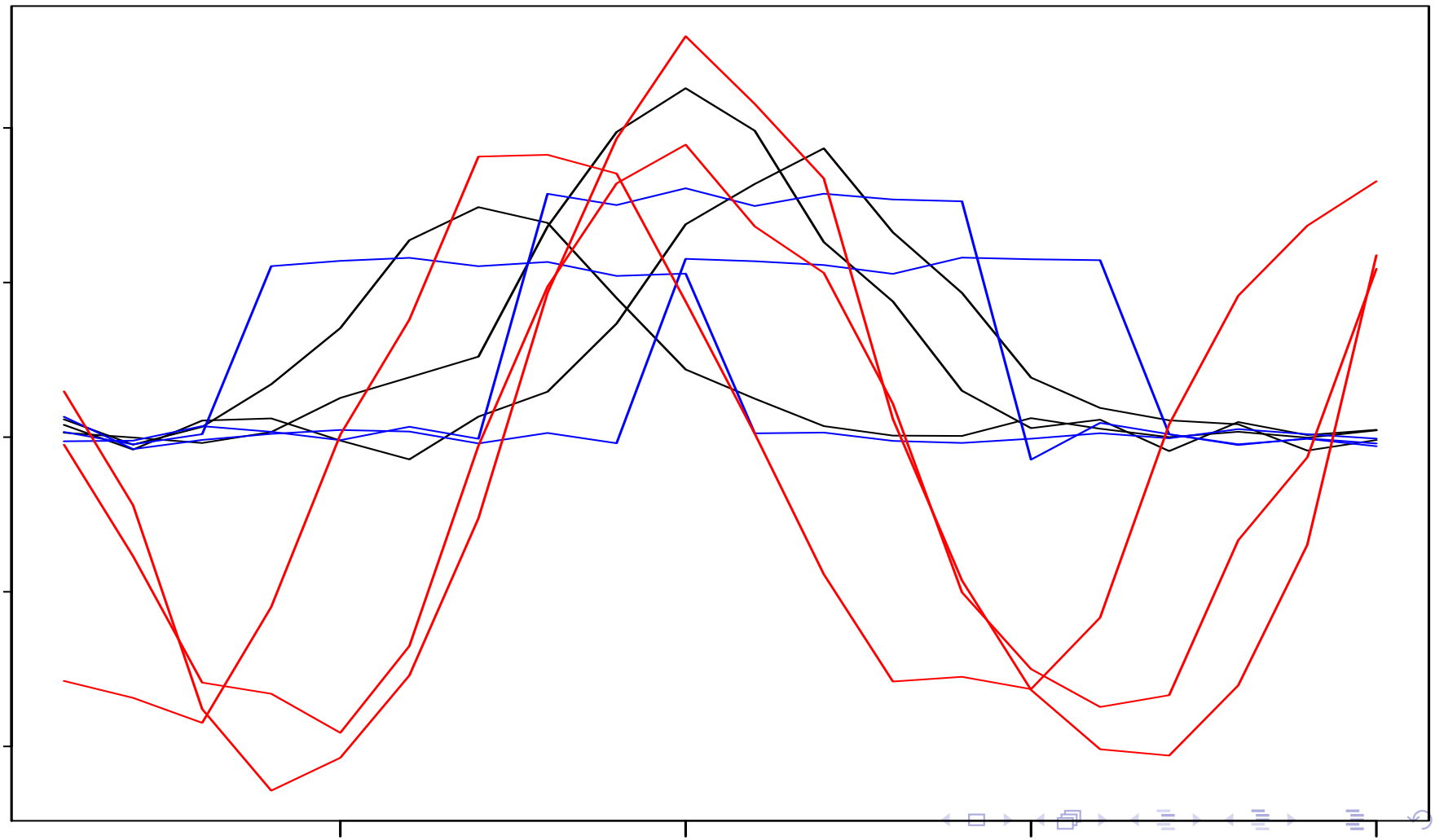
5 septembre 2019

Plan

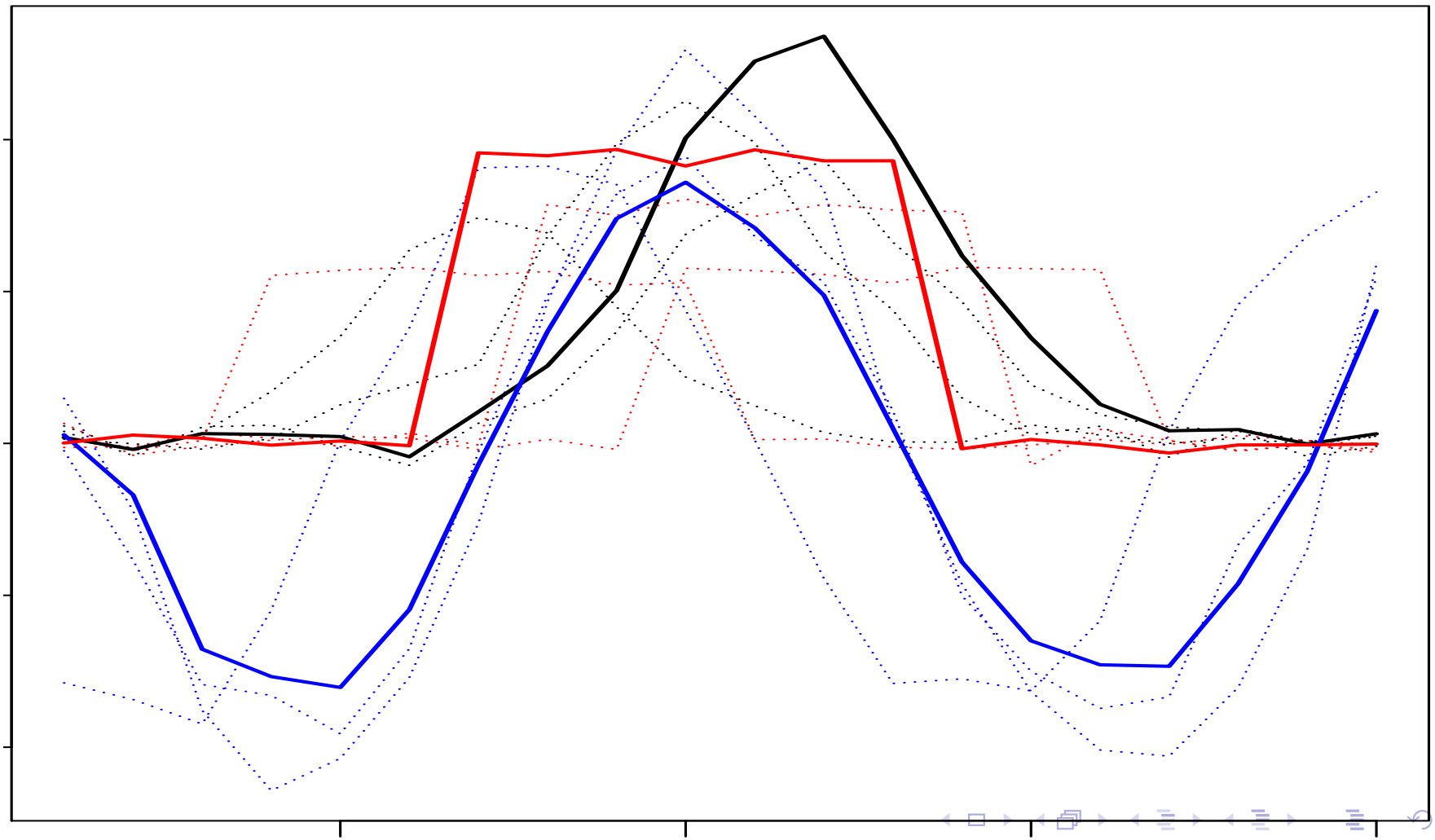
Mesures de ressemblance basées sur la forme

- problématique de *clustering* de séries temporelles
- mesures de ressemblance basées sur la **forme des séries**
  - *shape based measure*
  - invariante par changement d'échelle (normalisation des valeurs)
  - invariante par décalage dans le temps (alignement temporel)
- mesure DTW : invariance plus générale

Exemple : 9 séries en 3 classes



Exemple : barycentres calculés



Pertinence des mesures basées sur la forme

- origine temporelle mal définie ou inconnue
  - signaux périodiques
  - outil de mesure induit un décalage (spectromètre)
  - distribution temporelle de tweets associés à un hashtag

Méthodes existantes

- K-Spectral Centroïd (KSC)
  - J. Yang et J. Leskovec (2011)
  - séries à valeurs positives
  - distributions temporelles de tweets
- K-Shape (KS)
  - Paparrizos et Gravano (2017)
  - séries à valeurs quelconques
  - validation expérimentale conséquente
    - distance euclidienne, DTW, DTW contraint
    - partitionnement plat, hiérarchique, méthodes spectrales,...
- KSC et K-Shape ont une complexité en $O(L^3)$

Notations

- série temporelle : x
- longueur des séries : L
- opérateur de décalage temporel d'offset o : $\tau_o(x)$
- produit scalaire : $x \cdot \tau_o(y)$
  - calculé sur l'intersection des supports de x et de $\tau_o(y)$
  - $o \in \mathcal{O} = \{-L+1, \ldots, L-1\}$
- corrélation croisée normalisée

$$CC_N(x,y)(o) = \frac{x \cdot \tau_o(y)}{\|x\|\|y\|}$$

Mesures de ressemblance existantes

- mesures basées sur une transformation de CC_N

$$d(x,y) = \min_{o \in \mathcal{O}} \; g\left(\; CC_N(x,y)(o) \; \right)$$

- KSC :

$$g_{KSC}(cc) = \sqrt{1 - cc^2}$$

- K-Shape :

$$g_{KS}(cc) = 1 - cc$$

Mesures proposées

- proposition pour la méthode K-Forme (KF)

$$g_E(cc) = \sqrt{1 - cc}$$

- proposition pour la méthode K-SoftForme (KSF)
  - identique à la précédente (même g_E)
  - le calcul du $\min_{o \in \mathcal{O}}$ remplacé par un softMin
  - ne fournit pas un alignement
  - motivé par l'approche softDTW (Cuturi et Blondel 2017)

$$d_{KSF}(x, y) = \text{softMin}_{o \in \mathcal{O}}\ g_E\ (\ CC_N(x, y)(o)\ )$$

Rappel sur la fonction *softMin*

- fonction softMin pour un vecteur $v \in \mathcal{R}^n$

$$\text{softMin}^{\gamma}(v_1, \ldots, v_n) = -\gamma \log \left(\frac{\sum_o e^{-\frac{v_o}{\gamma}}}{n} \right)$$

- dépend du paramètre de régularisation γ
- propriétés :
  - f-moyenne généralisée
  - fonction régulière (calcul de gradient)
  - converge vers min par valeurs supérieures

$$\lim_{\gamma \to 0} \text{softMin}^{\gamma}(v) = \min(v)$$

Calcul des mesures (et du gradient)

- calcul de tous les $CC_N(x, y)(o)$ semble en $O(L^2)$
- mais classiquement la FFT permet un calcul efficace

$$CC(x, y) \equiv \mathcal{F}^{-1}\left(\mathcal{F}(x^0)\ \mathcal{F}^*({}^0y)\right)$$

- évaluation de toutes les mesures en $O(L\log(L))$
- précalcul des transformées
- gradient de $d_{KSF}(x, y)$
  - la complexité semble encore une fois quadratique
  - mais le calcul fait apparaitre un produit de convolution
  - grâce à la FFT, complexité en $O(L\log(L))$

Plan

K-Spectral Centroïd et K-Shape : barycentres

- calcul du barycentre μ des $\{x_1, \ldots, x_n\}$
- alignement optimal de x_i par rapport à μ : o_i^*
- o_i^* fournis lors de la phase d'affectation
- o_i^* maintenus fixes pour la phase de représentation
- KSC et K-Shape optimisent le "critère d'inertie" de d_{KSC}

$$\mu = \arg\max_{\mu} \sum_i CCN^2(\mu, x_i)(o_i^*)$$

K-Spectral Centroïd et K-Shape : barycentres

- équivalent à la maximisation d'un quotient de Rayleigh

$$\mu = \arg\max_{\mu} \frac{\mu^t S \mu}{\mu^t \mu}$$

- μ vecteur propre de la plus grande valeur propre λ_{max} de S
- signe de μ à déterminer
- diagonalisation $O(L^3)$
- optimisations possibles (ESANN 2018)
  - calcul incrémental de la matrice S
  - méthode de la puissance itérée pour le calcul μ
- complexité observée de KSC et K-Shape : $O(L^2)$

K-Forme : barycentres

- K-Forme : critère d'inertie équivalent à

$$\mu = \arg\max_{\mu} \sum_{i} CCN(\mu, x_i)(o_i^*)$$

- calcul du barycentre μ
  - somme sur le support de μ des $\tau_{o_i^*}(x_i)/||x_i||$
  - normalisation unitaire
- complexité linéaire en L
- complexité de K-Forme : $O(L \log(L))$

K-SoftForme : barycentres

- K-SoftForme : critère d'inertie

$$\mu = \arg\min_{\mu} \sum_{i} d_{KSF}^{\gamma}(\mu, x_i)^2$$

- les "alignements" ne sont pas fixés
- nécessité d'une méthode d'optimisation non linéaire (BFGS, gradient conjugué,...)
- peu de descentes de gradient si barycentre quasi-stabilisé
- complexité de K-SoftForme : $O(L \log(L))$

Centrage des barycentres

- soit la matrice de centrage

$$Q = Id - \frac{1}{L}\mathbf{1}$$

- K-Shape
  - μ est à présent vecteur propre de $Q^t SQ$
- K-Forme
  - centrage à chaque calcul de μ
- les complexités des méthodes ne sont pas modifiées

Plan

Données et implémentation

- UCR Time Series Classification Archive
- 85 jeux de données avec une partition réelle
- juste la partie TRAINING
- implémentation en R
- optimisations : FFT et méthode de la puissance itérée
- unique station de travail (20 jours de calcul)

Protocole de test

- méthodes testées
  - K-Shape et K-Forme avec ou sans centrage
  - K-SoftForme avec 3 γ différents
  - K-Means, K-DTW (avec ou sans fenêtre)
- 10 initialisations différentes
- même initialisation pour toutes les méthodes
- comparaison avec la partition réelle : Rand-Index
- résultats présentés
  - moyenne des 10 Rand-Index
  - temps cumulé des 10 exécutions

Qualité du partitionnement

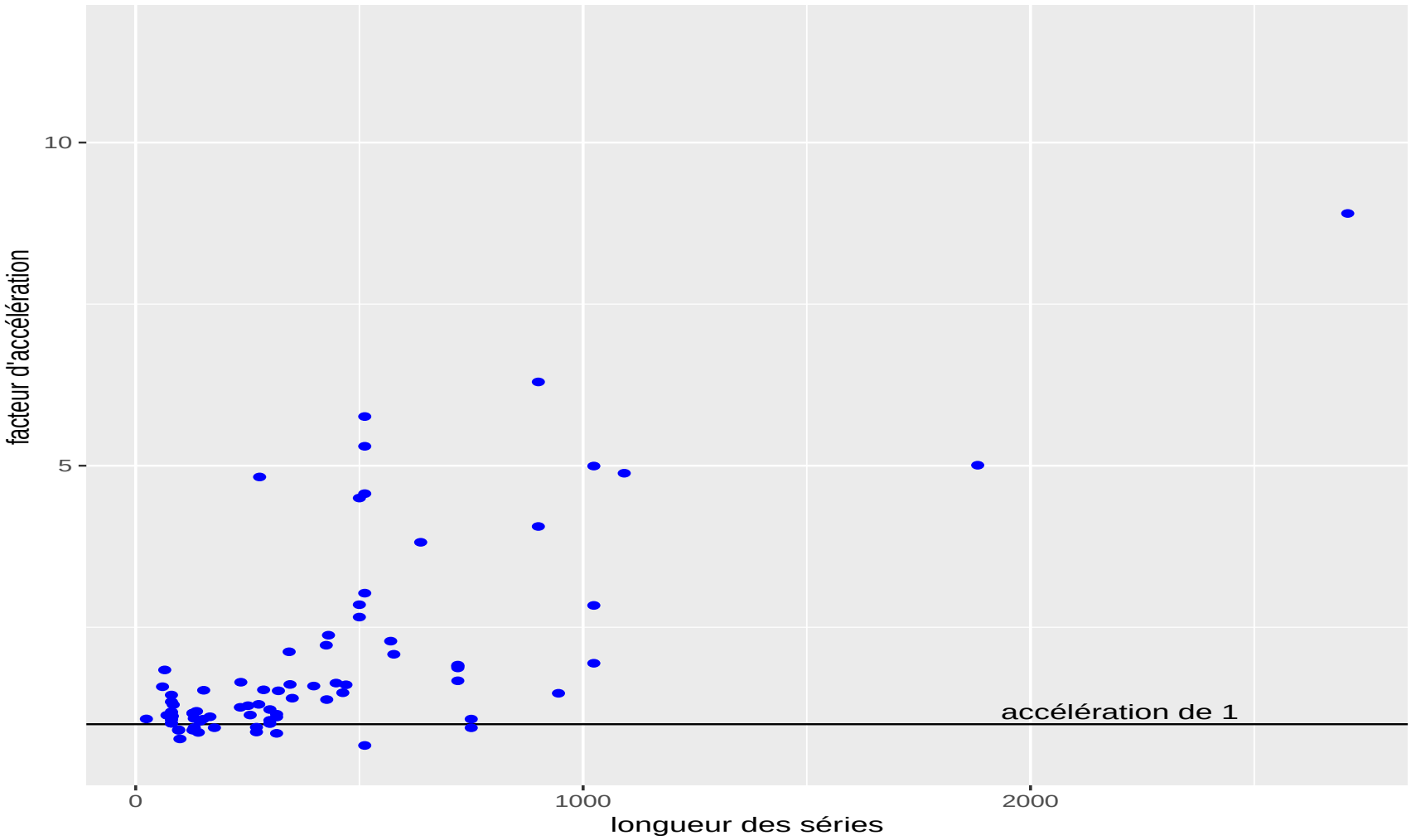
méthode	paramètre	moyenne des 85 RI	classé devant K-Shape
K-Shape	centré	0.6995	
K-Forme	centré	0.6989	41
K-SoftForme	0.001	0.6988	40
K-DTW	sans fenêtre	0.6927	39
K-Means		0.6873	38

Durée des 10 exécutions

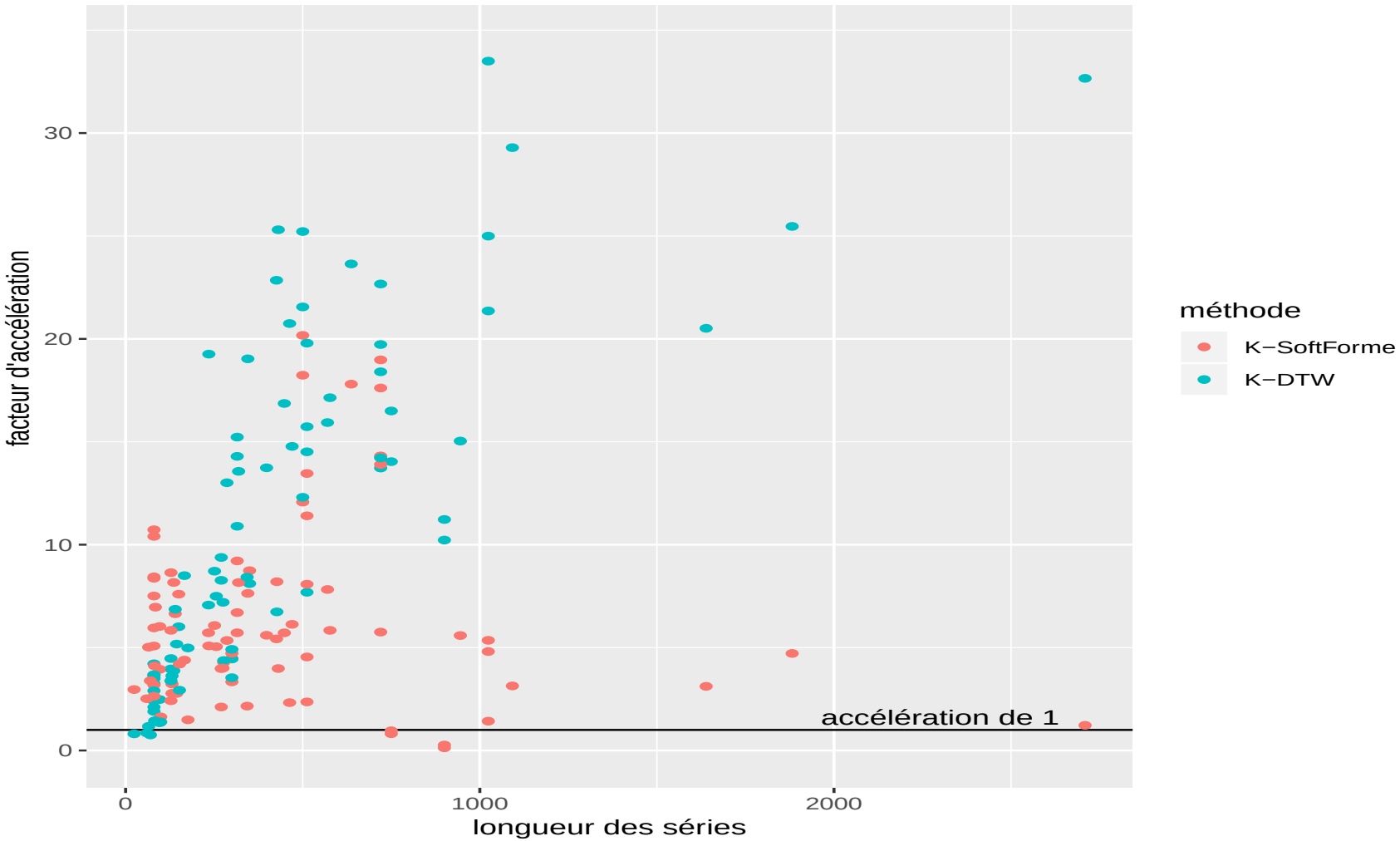
- pour HandOutlines ($L = 2709$, $n = 370$ et 2 classes)

méthode	paramètre	durée
K-Means		5 sec
K-Forme	non centré	5 min
K-Shape	non centré	45 min
K-SoftForme	0.001	55 min
K-DTW	sans fenêtre	24 h

Accélération de K-Forme par rapport à K-Shape



Accélération de K-Shape % à K-SoftForme et K-DTW



Conclusions

- confirmation expérimentale de la qualité de K-Shape
- K-Forme a des résultats très similaires
- implémentation simple, exécution rapide
- à implémenter dans les librairies de séries temporelles

- K-SoftForme a des résultats satisfaisants
- la mesure est plus universelle grâce à sa régularité
- adaptable à des problématiques de discrimination