Analyse fonctionnelle de données
Jour 3 : recalage temporel

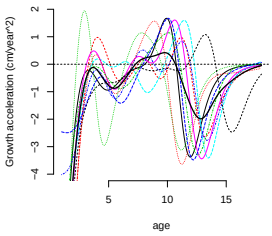
Jean-Marc Lasgouttes, Inria Paris

jean-marc.lasgouttes@inria.fr

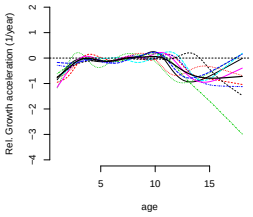
Partie I. Pourquoi recaler les données

Courbes de croissance (rappel)

Données la taille de 39 garçons et 54 filles et l'âge auquel elle a été mesurée



Courbe d'accélération de croissance pour 10 filles (avec la moyenne).

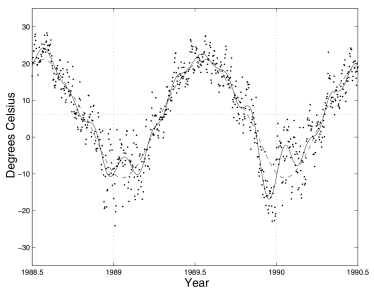


Courbe d'accélération relative pour 10 filles (avec la moyenne).

Problème les courbes sont décalées en temps. Les moyennes n'ont pas les mêmes caractéristiques que les courbes.

Climat Canadien

Données on a la température moyenne relevée chaque jour de 1961 à 1994 à Montréal. On sait que certaines années l'hiver peut arriver plus tôt ou plus tard. On cherche à quantifier ce phénomène.

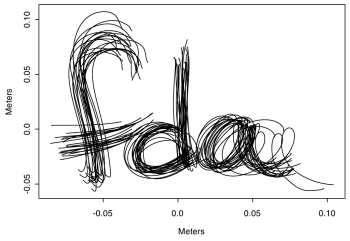


Température journalière à Montréal (points) de mi 1989 à mi 1990, avec courbe lissée (pleine) et tendance périodique (pointillé)

Type de données une seule courbe sur 33 ans
Lissage des données $y(t)$ 500 B-splines d'ordre 6
Tendance périodique $y_0(t)$ 9 bases de Fourier de période de base 365.25 pour faire apparaître les phénomènes saisonniers
Écart type entre les deux courbes : 2.15°C
Observation Vague de froid en 1990 : serait moins exceptionnelle si plus tardive

Écriture manuscrite

Données on fait écrire 20 fois le mot « fda » à plusieurs sujets. On mesure les coordonnées (x, y, z) de la pointe du stylo 200 fois par seconde.



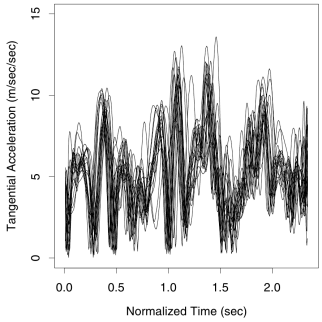
Les caractères « fda » écrits 20 fois.

Propriétés
- données en 3 dimensions (on peut mesurer la vitesse et l'accélération tangentielle)
- forme des caractères variable
- vitesse d'écriture variable

Accélération tangentielle on regarde la norme du vecteur d'accélération tangentielle

$$A_T(t) = \sqrt{X''(t)^2 + Y''(t)^2}$$

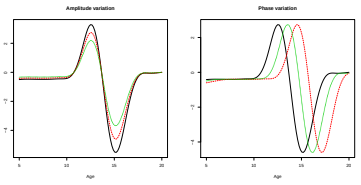
Représentation



Échelle toutes les courbes sont ramenées à 2.3 s
Problème les courbes ne sont pas alignées entre elles

Partie II. Principe du recalage de données

Principe



Gauche (resp. droite) : 3 courbes qui ne diffèrent que par l'amplitude (resp. la phase)

2 types de variabilité
- *En amplitude* : variations plus importantes à un instant donné selon les courbes
- *En phase* : variations pas aux mêmes instants

2 temps par exemple pour la croissance
- âge chronologique
- âge physiologique

Conséquence la moyenne ne ressemble pas aux courbes individuelles

Spécificité seule la variabilité en amplitude existe sur les données multivariées.

Définition

Données on dispose de n fonctions $y_i(t)$ définies sur $[0, T_i]$. On se donne un intervalle $[0, T_0]$ de référence.

Torsion du temps une fonction h_i de torsion du temps de $[0, T_0] \mapsto [0, T_i]$ a les propriétés suivantes
- $h_i(0) = 0$ et $h_i(T_0) = T_i$
- h_i est strictement croissante (le temps reste ordonné)
- donc h est bijective : $h_i^{-1}(h_i(t)) = t$

Principe soit $y_0(t)$ une fonction de référence. On cherche une fonction de torsion du temps $h_i(t)$ telle que

$$y_i(h_i(t)) = y_0(t) + \epsilon_i(t)$$

où $\epsilon_i(t)$ est petit par rapport à $y_i(t)$.

Condition on suppose donc que les fonctions y_i et y_0 ont plus ou moins la même forme.

Méthode de Procuste

Problème en général on ne connaît pas y_0, il faut le déterminer.

Détermination de y_0 on suit les pas suivants

1. on aligne linéairement les fonctions sur $[0, T_0]$
2. on pose $y_0^{(1)}(t)$ la moyenne empirique des courbes y_i.
3. on recale les fonctions y_i sur $y_0^{(1)}$ en déterminant les fonctions $h_i^{(1)}$
4. on pose $y_0^{(2)}(t)$ la moyenne empirique des courbes recalées $y_i \circ h_i^{(1)}$.
5. on recale les fonctions y_i sur $y_0^{(2)}$ en déterminant les nouvelles fonctions $h_i^{(2)}$
6. On pourrait continuer, mais en général ça suffit

Remarque on préfère souvent recaler les dérivées des fonctions : leurs variations sont souvent plus visibles.

Décomposition de l'erreur quadratique moyenne

Notations on a des fonctions $y_i(t)$ et des fonctions de torsion $h_i(t)$. On note $\bar{y}(t)$ la moyenne des y_i et $\bar{y}^*(t)$ celle des $y_i \circ h_i$

Erreur quadratique moyenne totale on peut la décomposer en amplitude

$$E^2_{\text{tot}} = \frac{1}{n}\sum_{i=1}^{n} \int [y_i(t) - \bar{y}(t)]^2 dt = E^2_{\text{amp}} + E^2_{\text{phas}}$$

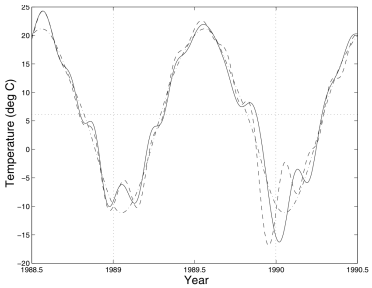
$$E^2_{\text{amp}} = C_R \frac{1}{n}\sum_{i=1}^{n} \int [y_i(h_i(t)) - \bar{y}^*(t)]^2 dt$$

$$E^2_{\text{phas}} = C_R \int \bar{y}^*(t)^2 dt - \int \bar{y}(t)^2 dt$$

$C_R - 1$ est lié à la covariance des dérivées h_i' et des fonctions recalées au carré $y_i(h_i(t))^2$

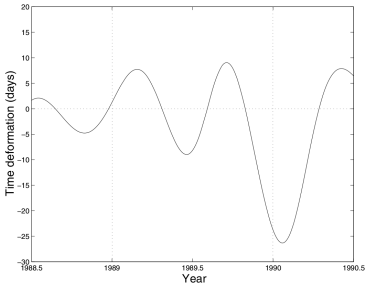
Proportion de variation due à la phase : $R^2 = E^2_{\text{phas}}/E^2_{\text{tot}}$

Exemple : températures



Température lissée recalée (pleine) de mi-1989 à mi-1990, courbe lissée brute (pointillé) et tendance périodique (pointillé+point)

Méthode on recale la courbe lissée $y(t)$ sur la partie périodique $y_0(t)$
Détails on utilise une base de 140 B-splines d'ordre 5 (tous les 3 mois).
Écart type entre les deux courbes : 1.73°C
Observation l'hiver 1990 est plus normal



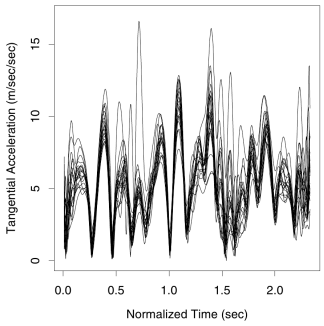
Température lissée recalée (pleine) de mi-1989 à mi-1990, courbe lissée brute (pointillé) et tendance périodique (pointillé+point)

Déformation du temps représentée par la fonction

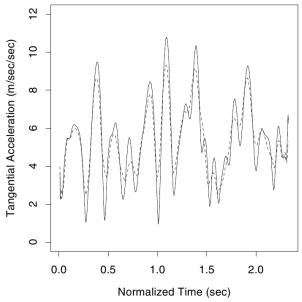
$$h(t) - t$$

Observation l'hiver 1990 avait 25 jours d'avance

Exemple : écriture manuscrite



Courbes d'accélération tangentielle pour les courbes recalées



Courbes d'accélération moyenne pour les courbes brutes (pointillé) et recalées (trait plein)

Propriétés

- les courbes d'accélération sont maintenant comparables
- les moyennes sont plus précises et détaillées

Partie III. Les méthodes de recalage

Recalage par décalage uniforme

Données on dispose de fonctions $y_i(t)$ que l'on veut recaler sur une fonction $y_0(t)$ uniquement en utilisant un décalage δ_i.

Critère à minimiser on cherche le δ_i qui minimise

$$\frac{1}{T_0 - |\delta_i|} \int [y_i(t + \delta_i) - y_0(t)]^2 dt$$

Limitations

- on est limité à de simple translations de courbes
- l'intervalle de définition de la fonction et de la moyenne est réduit de $|\delta_i|$

Recalage par points de repère

Point de repère c'est un point reconnaissable d'une courbe

- passage par zéro ou à une valeur particulière
- maximum/minimum

Données

- pour chaque courbe $y_i(t)$ (dont $y_0(t)$), on identifie les valeurs des R points de repère t_{ir} pour $r = 1, \ldots, R$.
- on pose $t_{i0} = 0$ et $t_{i,R+1} = T_i$

Objectif trouver des fonctions $h_i(t)$ telles que $h_i(t_{0r}) = t_{ir}$.

Méthodes

- construire une fonction $h_i(t)$ linéaire par morceaux (le plus rapide)
- construire une fonction lisse $h_i(t)$ qui minimise

$$\sum_{r=0}^{R+1} [h_i(t_{0r}) - t_{ir}]^2 + \lambda \int h_i''(t)^2 dt$$

- faire de même avec des fonction h_i monotones (le plus lent)

Recalage fonctionnel par moindres carrés

Données on a juste les fonctions $y_0(t)$ et $y_i(t)$

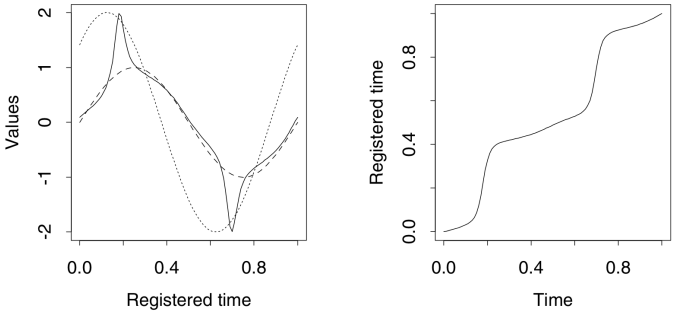
Critère on cherche le $h_i(t)$ qui minimise

$$\int [y_i(h_i(t)) - y_0(t)]^2 dt + \lambda \int h_i''(t)^2 dt$$

Problèmes

- le critère des moindres carrés est adapté à comparer les amplitudes, pas les phases
  exemple : la moyenne est un estimateur des moindres carrés
- quand 2 fonctions ont des différences d'amplitude et de phase, la torsion de temps va servir à corriger les problèmes d'amplitude

Exemple sur une translation pure



Gauche : courbe à recaler (points), cible (tirets) et résultat (pleine). Droite : fonction de torsion

Recalage fonctionnel par minimisation de valeur propre

Idée supposons que $y_0(t)$ et $y_i(h_i(t))$ ne diffèrent que par l'amplitude, par exemple

$$y_i(h_i(t)) = A y_0(t)$$

Alors la seconde valeur propre de la matrice suivante sera 0

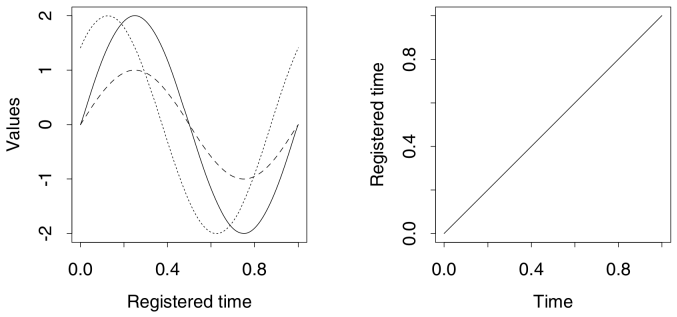
$$\mathbf{T}(h_i) = \begin{pmatrix} \int y_0(t)^2 dt & \int y_0(t) y_i(h_i(t))\, dt \\ \int y_0(t) y_i(h_i(t))\, dt & \int y_i(h_i(t))^2 dt \end{pmatrix}$$

Critère d'optimisation en utilisant un lissage on obtient le critère ($\mu_2[\cdot]$ est la seconde valeur propre)

$$\mu_2[\mathbf{T}(h_i)] + \lambda \int h_i''(t)^2 dt$$

Fonction monotone on peut exiger une fonction monotone (et remplacer h_i'' par W_i''), mais le calcul est beaucoup plus long

Exemple sur une translation pure



Gauche : courbe à recaler (points), cible (tirets) et résultat (pleine). Droite : fonction de torsion

Comparatif des méthodes

Décalage uniforme rapide et simple à mettre en œuvre, mais qualité limitée et réduction de l'étendue de t

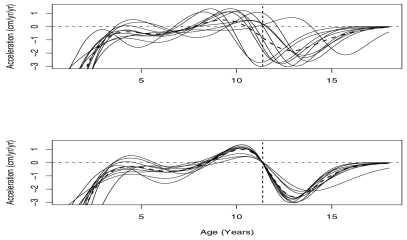
Point de repère assez rapide, et résultats assez bons ; par contre le positionnement des points de repère peut être assez difficile et devoir être fait à la main

Moindres carrés plutôt bon et facile à mettre en œuvre ; par contre peut être très lent

Valeur propre en général la meilleure méthode, facile aussi d'utilisation ; par contre peut être très lent

Exemple : croissance

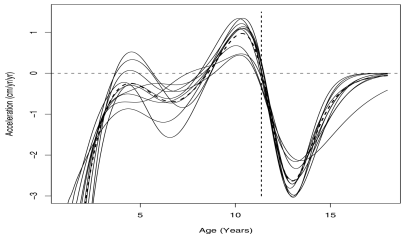
Utilisation d'un point de repère le plus facile est de noter quand une courbe passe par 0



Haut : courbe d'accélération de croissance pour 10 filles ; bas : mêmes courbes recalées sur le point de repère de l'annulation de l'accélération de la croissance à la puberté

Base 4 B-splines d'ordre 3, $\lambda = 10^{-12}$
Proportion de la variation due à la phase : $R^2 = 0.70$
Observation la moyenne (tirets) est maintenant plus représentative des courbes, sauf aux âges prépubères

Recalage fonctionnel on recale les courbes précédentes sur leur moyenne

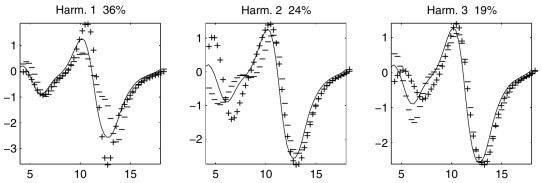


Courbe d'accélération de croissance après recalage par point de repère et recalage fonctionnel.

Base 15 B-splines d'ordre 5, $\lambda = 1$
Observation la poussée de croissance prépubère est maintenant correctement alignée

ACP sur les amplitudes

Méthode on décompose les fonctions $y_i(h_i(t))$ (et on fait une rotation varimax)



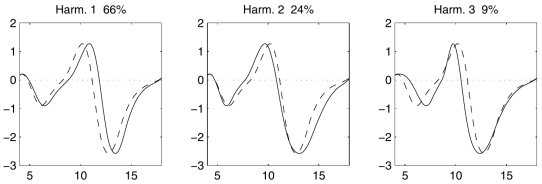
Décomposition des axes recalés pour 10 filles.

Interprétation

- axe 1 : variation pendant la poussée de croissance de la puberté,
- axes 2 et 3 : variation pendant la période prépubère

ACP sur les phases

Méthode on décompose les fonctions $h_i(t)$.



Décomposition des fonctions de recalage pour 10 filles.

Interprétation

- axe 1 : retard de croissance,
- axe 2 : avance de croissance jusqu'au ralentissement de la poussée de croissance de la puberté,
- axe 3 : retard sur la poussée prépubère, mais avance sur la poussée de puberté.

Partie IV. Utilisation du paquet « fda »

Recalage uniforme

Fonction calcule des décalages pour un ensemble de fonctions

```
regfd0obj <- register.fd0(y0fd, yfd)
```

Paramètres
- `y0fd` : données fonctionnelles décrivant la fonction cible
- `yfd` : données fonctionnelles multivariées à recaler

Valeur retournée une liste contenant
- `regfd` : données fonctionnelles décrivant les fonctions recalées
- `dregfd` : données fonctionnelles décrivant la différence entre `regfd` et `y0fd`
- `offset` : vecteur contenant le décalage pour chaque courbe de `yfd`

Recalage par points de repère

Fonction calcule des décalages pour un ensemble de fonctions

```
landmarkobj <- landmarkreg(fdobj, ximarks, x0marks,
    WfdPar, monwrd=FALSE, ylambda=1e-10)
```

Paramètres utiles
- `fdobj` : données fonctionnelles multivariées à recaler
- `ximarks` : une matrice contenant les positions des points de repère (un par ligne) pour les courbes de `fdobj` (une par colonne) à recaler
- `x0marks` : un vecteur contenant les positions des points de repère pour la fonction cible
- `Wfdparobj` : un objet produit par `fdPar()`

Paramètres peut-être utiles
- `monwrd` : la valeur `TRUE` demande l'utilisation de fonction monotone (plus lent) ; à essayer en cas de problème
- `ylambda` : en grande dimension, si des oscillations apparaissent, on peut essayer d'augmenter cette valeur

Valeur retournée une liste contenant
- `regfd` : données fonctionnelles décrivant les fonctions recalées
- `warpfd` : données fonctionnelles décrivant les fonctions de torsion du temps

Utilisation on peut marquer les points de repère à la main sur chaque courbe grâce à la fonction `locator()`. Exemple de code

```
par(mfrow=c(1,1), ask=TRUE)
for (i in 1:10) {
  plot(fdobj[i], main=paste("Case",i))
  ximarks[i] = locator(1)$x
}
```

Sinon, il faut trouver une méthode automatique de marquage

Recalage fonctionnel

Fonction calcule des décalages pour un ensemble de fonctions

```
regfdobj <- register.fd(y0fd, yfd, WfdParobj,
    conv=1e-04, iterlim=20, dbglev=1,
    periodic=FALSE, crit=2)
```

Paramètres
- `y0fd` : données fonctionnelles décrivant la fonction cible
- `yfd` : données fonctionnelles multivariées à recaler
- `Wfdparobj` : un objet produit par `fdPar()`
- `crit` : critère à minimiser (1 pour moindres carrés, 2 pour seconde valeur propre)
- `periodic` : si `TRUE`, on considère que les fonctions sont périodiques et un décalage peut être ajouté à chaque fonction
- `conv`, `iterlim`, `dbglev` : voir `density.fd`

Valeur retournée une liste contenant
- `regfd` : données fonctionnelles décrivant les fonctions recalées
- `warpfd` : données fonctionnelles décrivant les fonctions de torsion du temps
- `Wfd` : données fonctionnelles décrivant les $W(x)$ des fonctions de torsion
- `shift` : vecteur de décalages (si `periodic == TRUE`)

Décomposition phase/amplitude

Fonction calcule la proportion de l'erreur quadratique due à la phase

```
apdecomp <- AmpPhaseDecomp(xfd, yfd, hfd, rng=xrng)
```

Paramètres
- `xfd` : données fonctionnelles brutes
- `yfd` : données fonctionnelles recalées
- `hfd` : données fonctionnelles pour les fonction de torsion du temps
- `rng` : intervalle sur lequel on fait la comparaison (optionnel)

Valeur retournée une liste contenant
- `MS.amp` : erreur quadratique moyenne en amplitude E^2_{amp}
- `MS.phas` : erreur quadratique moyenne en phase E^2_{phas}
- `RSQR` : proportion d'erreur quadratique due à la phase R^2
- `C` : constante C_R