Wygładzanie i klasyfikacja krzywych funkcjonalnych

Karolina Sopata

6 czerwca 2022

Wygładzanie splajnu przy użyciu "kary na szorstkość"

- rozszerzenia bazy mogą stanowić dobre przybliżenie danych funkcjonalnych
- popularną i efektywną metodą jest metoda najmniejszych kwadratów
- podejście oparte na karze na „szorstkość” jest jedną z metod aproksymacji danych funkcjonalnych
- metoda ta daje lepsze wyniki zwłaszcza przy estymacji pochodnych
- metoda oparta na optymalizacji kryterium dopasowania

Określanie szorstkości

$[D^2x(t)]^2$ - krzywizna funkcji

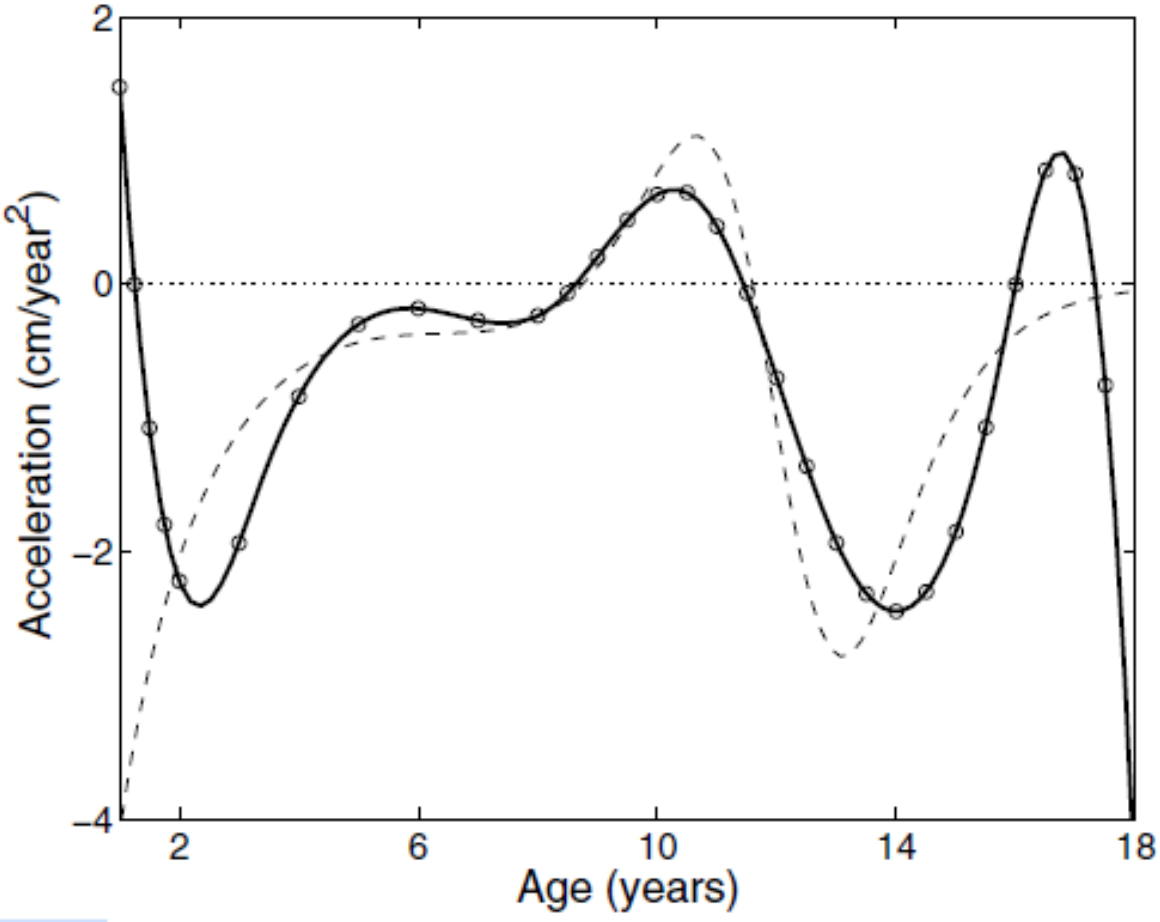
Naturalna miara szorstkości:

$$\mathtt{PEN}_2(x) = \int [D^2x(s)]^2\, ds$$

Generalnie:

$$\mathtt{PEN}_m(x) = \int [D^mx(s)]^2\, ds$$

Przykład



Rysunek: Krzywa przyspieszenia wzrostu, żródło: Ramsay, Functional Data Analysis, Fig. 4.3.

Kryterium wyboru parametru wygładzającego

x(t) - wektor wartości otrzymanych przy użyciu funkcji x na wektorze argumentów t

Kompromis pomiędzy gładkością a dopasowaniem do danych:

$$\mathrm{PENSSE}_{\lambda}(x|\mathbf{y}) = [\mathbf{y} - x(\mathbf{t})]'\mathbf{W}[\mathbf{y} - x(\mathbf{t})]^2 + \lambda \times \mathrm{PEN}_2(x)$$

de Boor (2002)
Krzywa x, która daje najmniejszą wartość powyższego kryterium jest splajnem sześciennym z węzłami w punktach t_j, dla j=1,...,n.

Wektor współczynników c

1. Przypadek bez parametru wygładzającego.
Postać funkcji:

$$x(t) = \sum_{k}^{K} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t) = \boldsymbol{\phi}'(t)\mathbf{c},$$

Estymacja współczynników:

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}'\mathbf{y}$$

Predykcja zmiennej objaśnianej:

$$\hat{\mathbf{y}} = \boldsymbol{\Phi}(\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y} = \mathbf{S}_{\phi}\mathbf{y},$$

Wektor współczynników c

2. Przypadek z parametrem wygładzającym.

$$\mathtt{PEN}_m(x) = \int [D^m x(s)]^2 \, ds$$

$$\begin{aligned} &= \int [D^m \mathbf{c}' \boldsymbol{\phi}(s)]^2 \, ds \\ &= \int \mathbf{c}' D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) \mathbf{c} \, ds \\ &= \mathbf{c}' [\int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) \, ds] \mathbf{c} \\ &= \mathbf{c}' \mathbf{R} c \ , \end{aligned}$$

gdzie:

$$\mathbf{R} = \int D^m \boldsymbol{\phi}(s) D^m \boldsymbol{\phi}'(s) \, ds \ .$$

Wektor współczynników c

2. Przypadek z parametrem wygładzającym, c.d.

$$\texttt{PENSSE}_m(\mathbf{y}|\mathbf{c}) = (\mathbf{y} - \mathbf{\Phi c})'\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}) + \lambda \mathbf{c}'\mathbf{R}\mathbf{c}$$

$$-2\mathbf{\Phi}'\mathbf{W}\mathbf{y} + \mathbf{\Phi}'\mathbf{W}\mathbf{\Phi c} + \lambda \mathbf{R c} = 0,$$

Estymator wektora współczynników c:

$$\hat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \lambda \mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}\mathbf{y}$$

Metoda uogólnionej walidacji krzyżowej

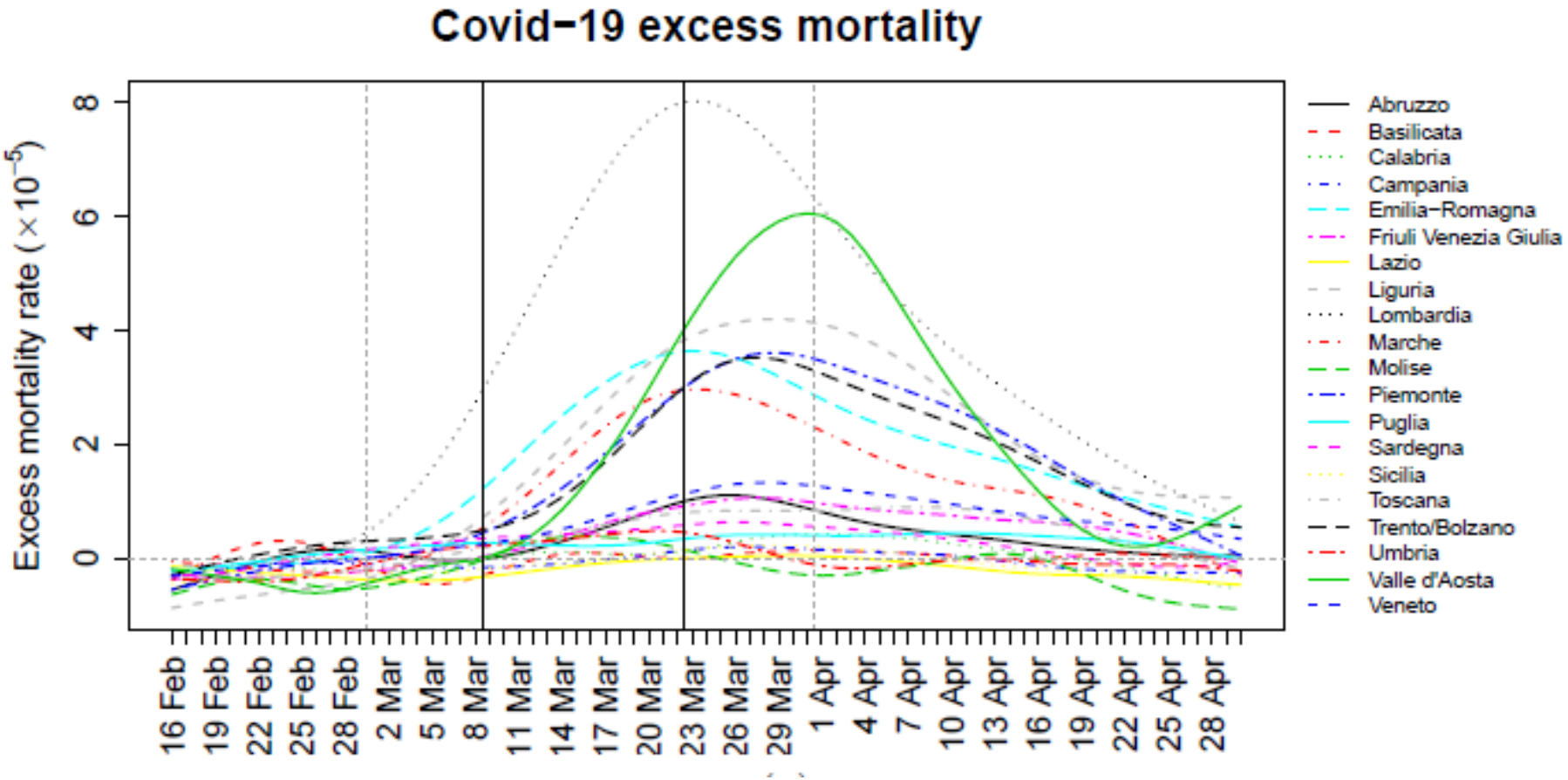
Kryterium:

$$\mathtt{GCV}(\lambda) = \frac{n^{-1}\,\mathbf{SSE}}{[n^{-1}\mathrm{trace}\,(\mathbf{I} - \mathbf{S}_{\phi,\lambda})]^2}$$

Równoważna forma:

$$\mathtt{GCV}(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{\mathbf{SSE}}{n - df(\lambda)}\right)$$

Wygładzanie krzywych - przykład



Rysunek: Wygładzone krzywe śmiertelności na Covid-19 we Włoszech, źródło: M. Cremona, Probabilistic K-mean with local alignment for clustering and discovery in functional data

Klasyfikacja metodą K-means

- Ustalenie liczby klas.
- Inicjalizacja centroidów: tasujemy zbiór danych, a następnie wybieramy losowo K punktów.
- Iteracje: do momentu, aż nie będzie żadnych zmian w centroidach, tzn. przydział punktów do klastrów nie będzie się zmieniał
  - obliczenie sumy kwadratów odległości między punktami a wszystkimi centroidami
  - przypisanie każdego punktu danych do najbliższego klastra (centroidu)
  - obliczenie centroidów klastrów, jest to średnia wartość wszystkich punktów należących do każdego klastra

K-means algorytm

Funkcja celu:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

K-means algorytm

Expectation step:

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$

$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

Maximization step:

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}}$$

Klasyfikacja metodą fuzzy C-means

- ustalenie c - liczby klastrów, wybór parametru rozmycia m (zwykle m $1.25 < m < 2$) i inicjalizacja macierzy partycji $U^{(0)}$
- wyliczenie centroidów $C^{(k)}$ przy ustalonym $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

- aktualizacja macierzy partycji $U^{(k)}, U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Klasyfikacja metodą probKMA

- Rozważamy zbiór N krzywych (d-wymiarowych) $x_i : R \to R^d$, i = 1,...,N.
- Celem jest zidentyfikowanie K (d-wymiarowych) centrów klas $v_k : (0, c_k) \to R^d$.
- Dla każdej krzywej znajdujemy przesunięcie, tak aby minimalizować odległość każdej z nich od centrum klasy (v_k), gdzie funkcja przesunięcia jest klasy $W = \{h : t \to t + s, s \in \mathbb{R}\}$.
- Każdej rozważanej krzywej przypisujemy prawdopodobieństwo należenia do danej klasy. Funkcja prawdopodobieństwa zdefiniowana jest w następujący sposób:

$$p_k : \{x_1, ..., x_N\} \to [0, 1]$$

gdzie $p_k(x_i) = p_{k,i}$.

Probabilistic K-means Algorithm

Spotykamy się z następującym problemem optymalizacji: chcemy znaleźć K centrów klas $(v_1, ..., v_k)$, macierz prawdopodobieństw P oraz macierz przesunięć S, tak aby zminimalizować uogólnione kryterium najmniejszych kwadratów zdefiniowane w kontekście analizy danych funkcjonalnych. Kryterium zdefiniowane jest następująco:

$$G_m(P, S, v_1, ..., v_K) = \sum_{i=1}^{N} \sum_{k=1}^{K} (p_{k,i})^m d^2(\tilde{x}_{i,s_{k,i}}, v_k), \qquad (1)$$

gdzie $m > 1$ jest pewnym ustalonym parametrem kontrolującym stopień rozmycia, natomiast $\tilde{x}_{i,s_{k,i}} = x \circ h_{k,i}$ przesuniętą krzywą; $d(\cdot, \cdot)$ jest odległością krzywej od centrum klasy.

Probabilistic K-means Algorithm

Inicjalizacja:
Zadanie liczby klas K oraz długości centroidów, wylosowanie początkowych prawdopodobieństw należenia do danej klasy oraz przesunięć.
Iteracje:

- wyliczenie centroidów $\hat{v}_k^{(it)}$
- wybór przesunięcia $\hat{s}_{k,i}^{(it)}$ oraz przypisanie krzywych do klas
- obliczenie prawdopodobieństwa $\hat{p}_{k,i}^{(it)}$ należenia krzywych do danej klasy

Kryterium stopu:

$$BC_k = -log\left(\sum_{i=1}^{N}\sqrt{p_{k,i}^{(it)}\,p_{k,i}^{(it-1)}}\right)$$

Probabilistic K-means Algorithm

Rozważmy zbiór $B = \{i \in \{1, ..., N\} | d(\tilde{x}_{i,\hat{s}_{k,i}}, \hat{v}_k) > 0, \forall k\}$ - zbiór indeksów krzywych, których odległość od każdego centrum klasy jest większa od 0. Przy ustalonych wartościach $\hat{v}_1, ..., \hat{v}_K$ oraz $\hat{S}$ estymator $\hat{P} = [\hat{p}_{k,i}]$, który globalnie minimalizuje kryterium (1) przyjmuje następującą postać:

$$\hat{p}_{k,i} = \left(\sum_{l=1}^{K} \left(\frac{d^2\left(\tilde{x}_{i,\hat{s}_{k,i}}, \hat{v}_k\right)}{d^2\left(\tilde{x}_{i,\hat{s}_{l,i}}, \hat{v}_l\right)} \right)^{\frac{1}{m-1}} \right)^{-1}, \tag{2}$$

dla $i \in B$ oraz:

$$\hat{p}_{k,i} = \begin{cases} 0, & \text{dla } k : d^2(\tilde{x}_{i,\hat{s}_{k,i}}, \hat{v}_k) > 0 \\ \in [0,1], & \text{dla } k : d^2(\tilde{x}_{i,\hat{s}_{k,i}}, \hat{v}_k) = 0 \end{cases}, \tag{3}$$

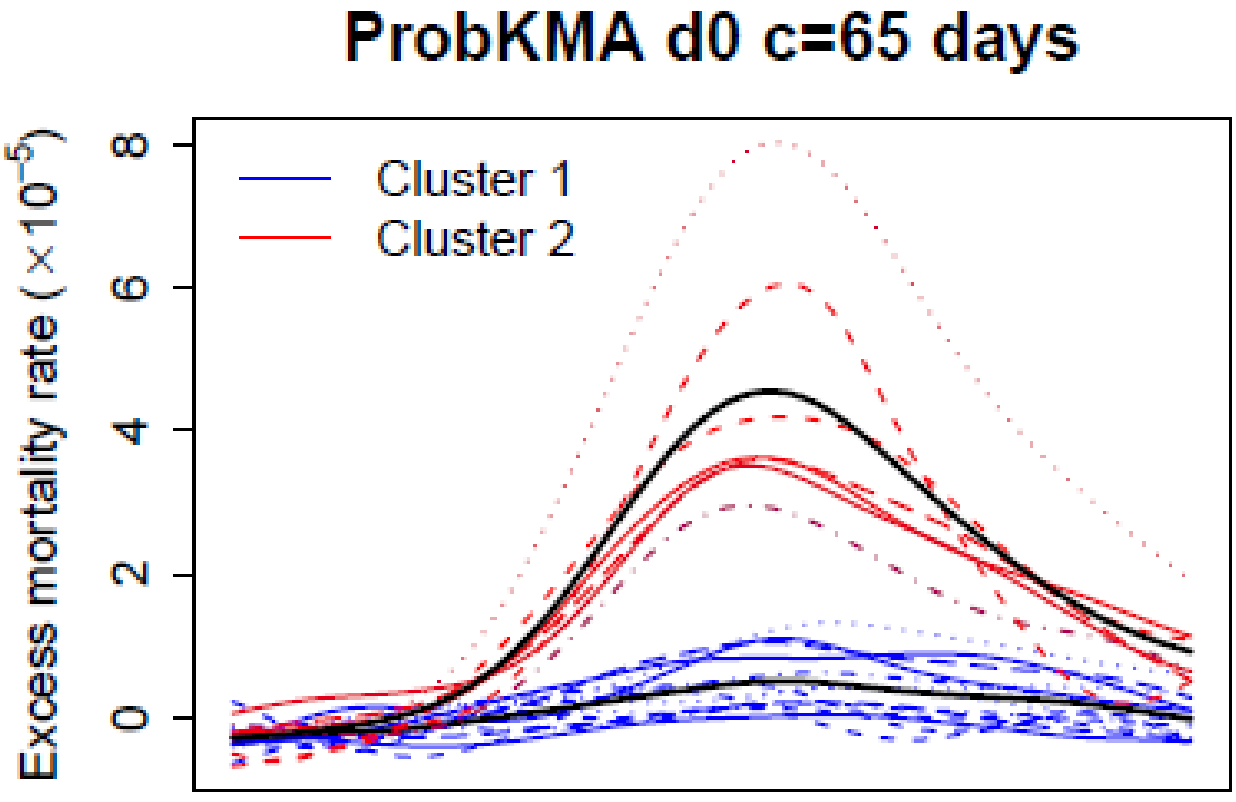
dla $i \notin B$.

Probabilistic K-means Algorithm

Przy ustalonych wartościach prawdopodobieństw należenia do danej klasy w macierzy $\hat{P}$ oraz przesunięć w macierzy $\hat{S}$, unikalny estymator centrum klasy, który minimalizuje funkcję $G_m(\cdot, \hat{P}, \hat{S})$ ma postać:

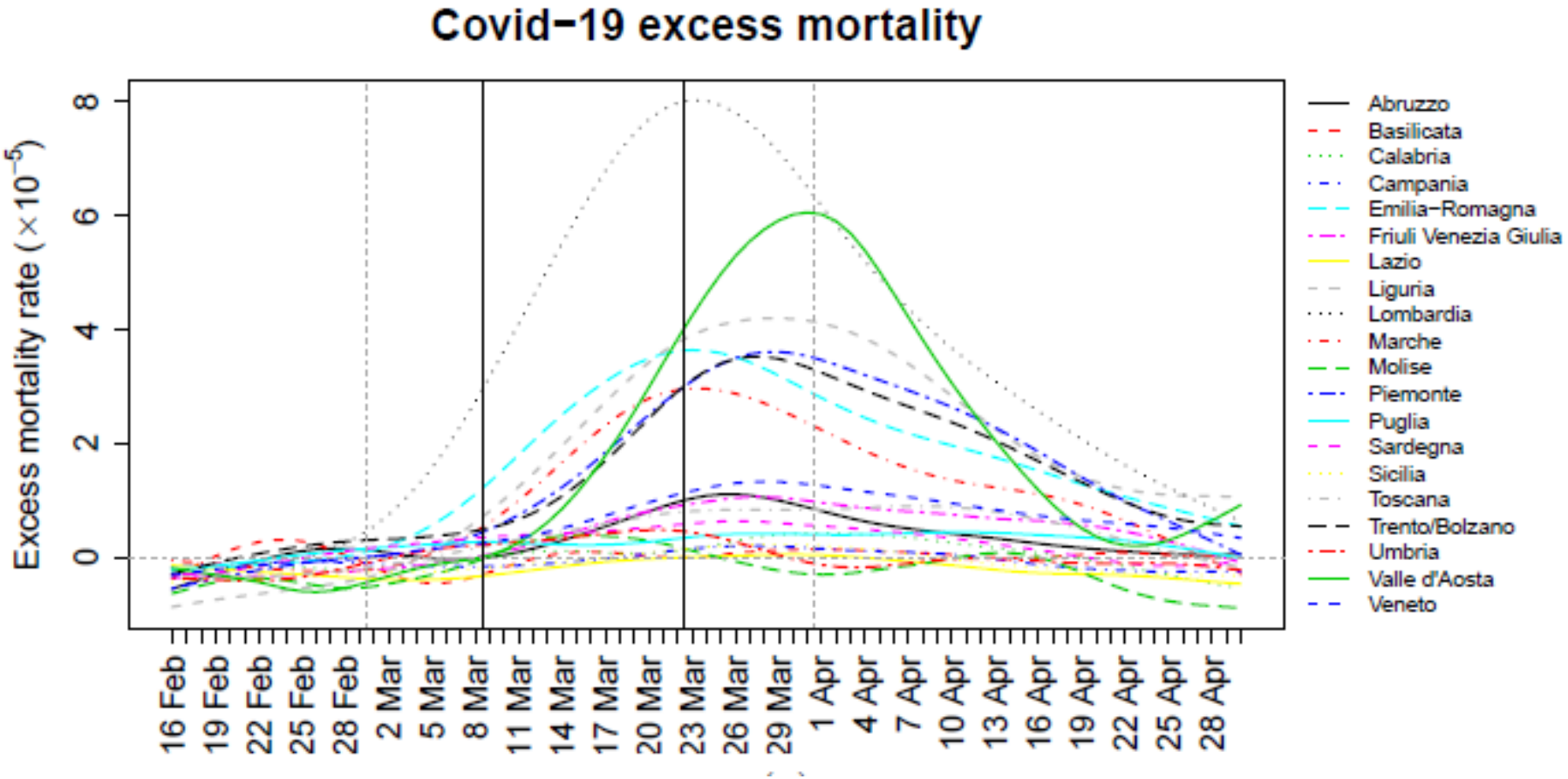
$$\hat{v}_k = \frac{\sum_{i=1}^{N} (\hat{p}_{k,i})^m \tilde{x}_{i,\hat{s}_{k,i}}}{\sum_{i=1}^{N} (\hat{p}_{k,i})^m}, \tag{4}$$

ProbKMA na danych Covid-19



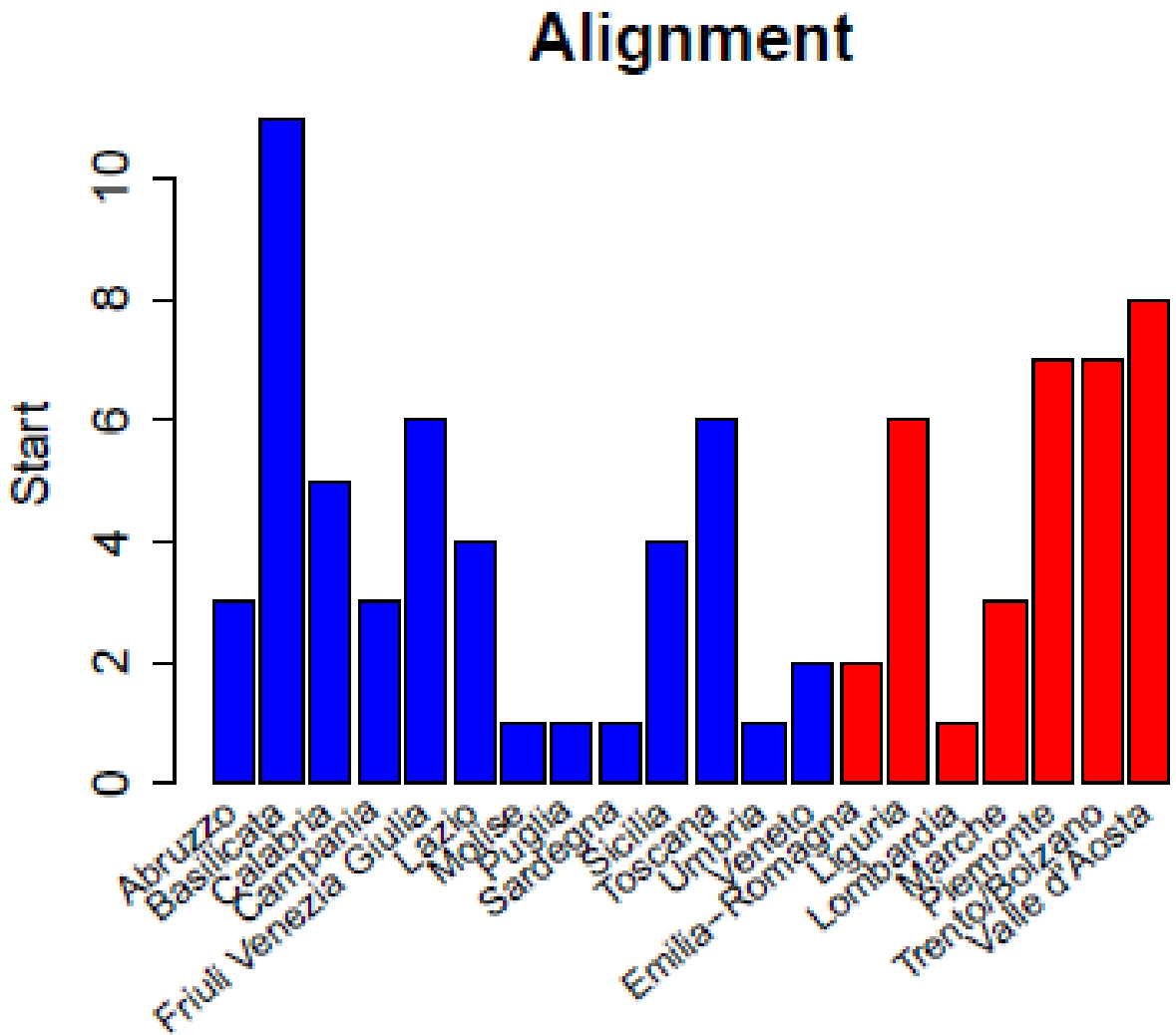
Rysunek: Klasyfikacja krzywych algorytmem probKMA, źródło: M. Cremona, Probabilistic K-mean with local alignment for clustering and discovery in functional data.

Wygładzanie krzywych



Rysunek: źródło: M. Cremona, Probabilistic K-mean with local alignment for clustering and discovery in functional data.

ProbKMA na danych Covid-19



Rysunek: Przesunięcia (liczba dni) krzywych, źródło: M. Cremona, Probabilistic K-mean with local alignment for clustering and discovery in functional data.

Dziękuję za uwagę!