STATISTICAL MODELLING OF THE THICKNESS OF TUNNEL LINING USING RANDOM FUNCTIONS AND PRINCIPAL COMPONENTS ANALYSIS

Daniela JARUŠKOVÁ

Czech Technical University of Prague, Dept. of Mathematics

2011

Change-point detection

At time points $t = 1, 2, \ldots$ we observe $\mathbf{X}_1, \mathbf{X}_2, \ldots$

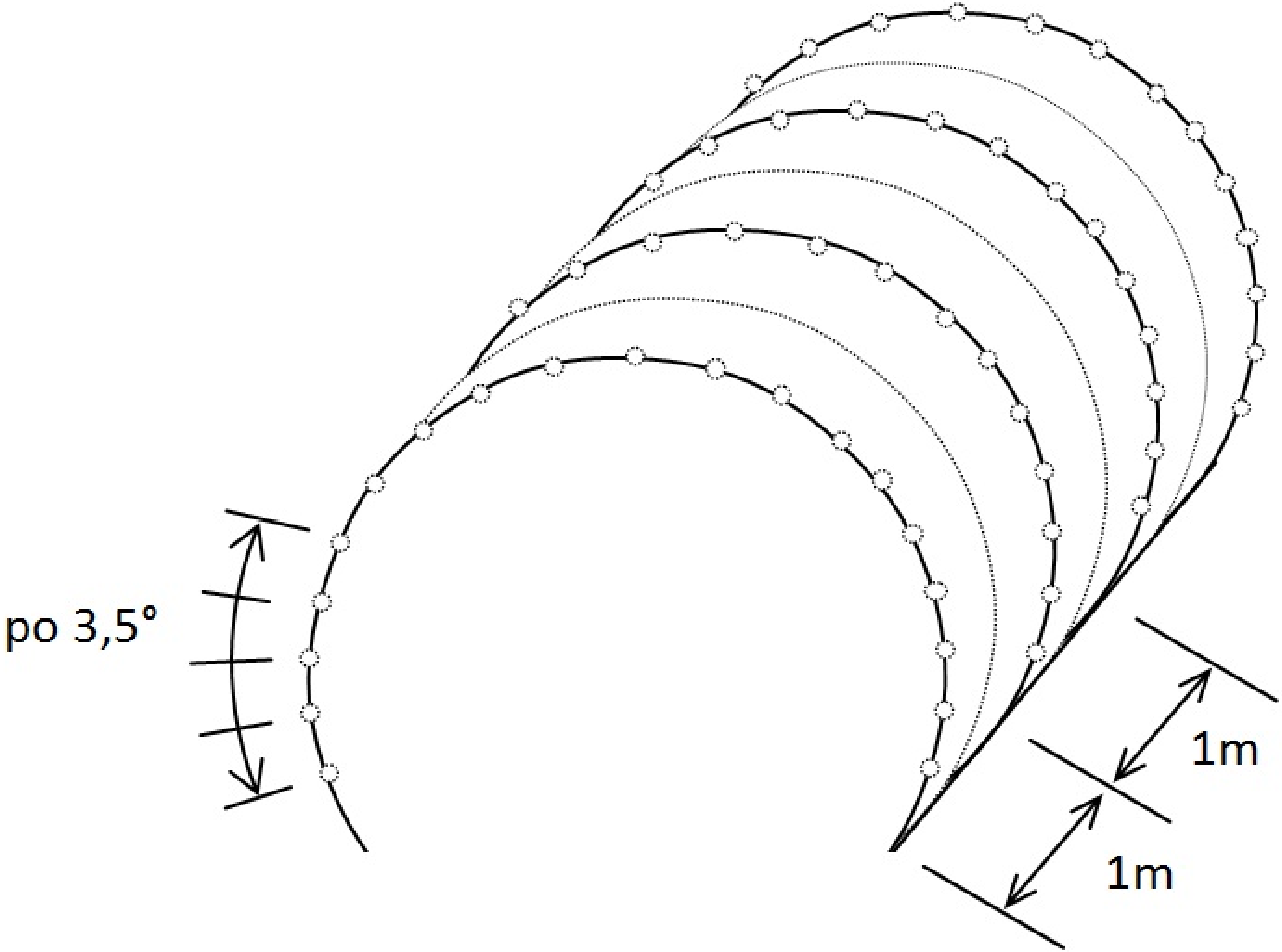
$$H_0: \quad \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \ldots \quad \text{all have the same stochastic behaviour}$$

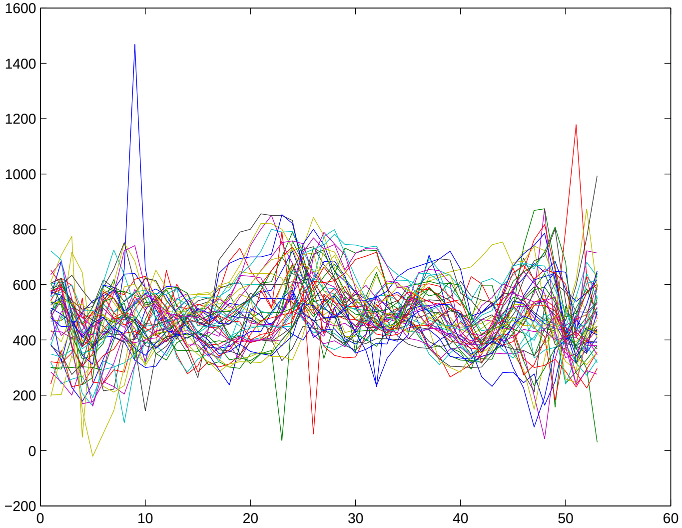
$$H_1: \quad \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k \quad \text{have the different stochastic behaviour from} \quad \mathbf{X}_{k+1}, \mathbf{X}_{k+2}, \ldots$$

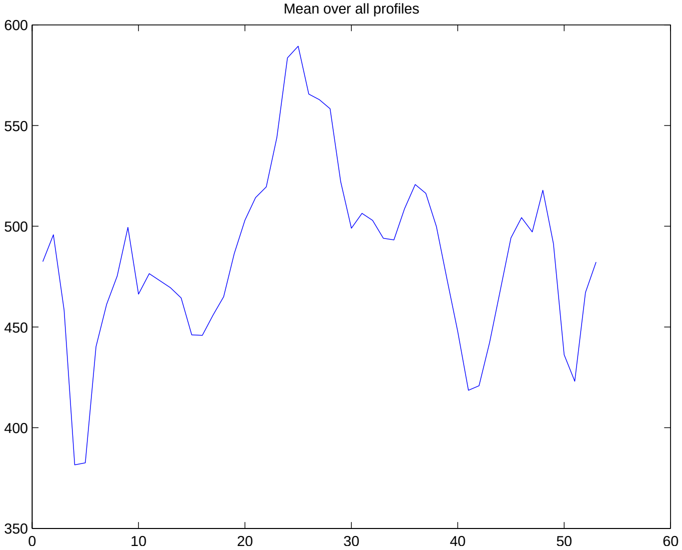
We do not know the value of k.

On-line detection

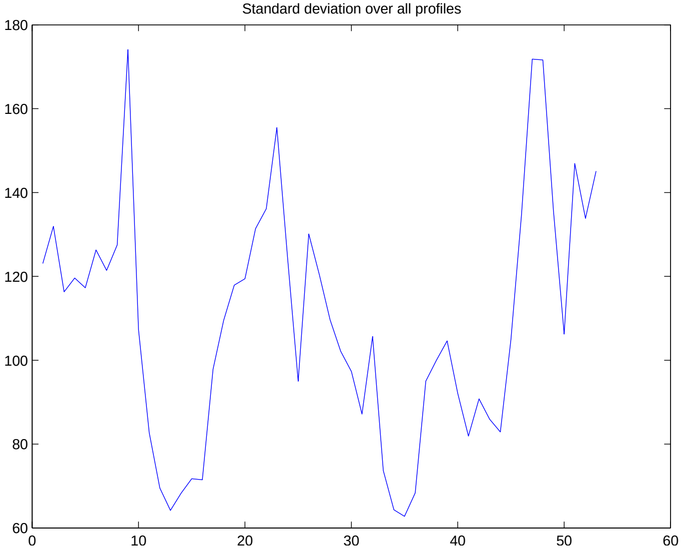
Off-line detection

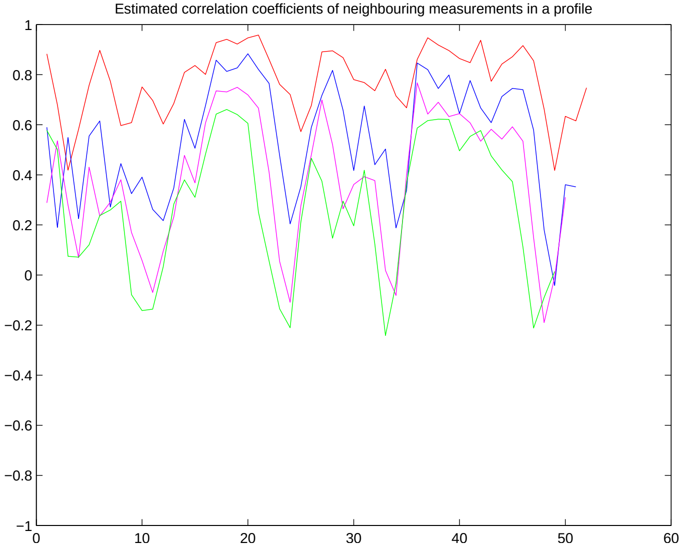

po 3,5°
1m
1m


1600
1400
1200
1000
800
600
400
200
0
−200
0
10
20
30
40
50
60



Mean over all profiles
600
550
500
450
400
350
0
10
20
30
40
50
60


Standard deviation over all profiles
180
160
140
120
100
80
60
0
10
20
30
40
50
60



Estimated correlation coefficients of neighbouring measurements in a profile
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
10
20
30
40
50
60

Typical feature

$\{\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T, \ i = 1, 2, \ldots\}$ p is very large

Basic characteristics:

$E\,\mathbf{X} = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T, \quad Var\,\mathbf{X} = \mathbf{\Sigma}_{pxp}$

The goal of statistical inference:

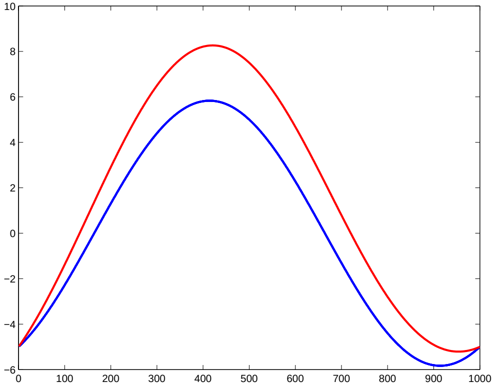
Detect a change (changes) in the mean vector $\boldsymbol{\mu}$ (at least in one component).

There are $2^p - 1$ subsets of indexes where a change can occur.

It would be very nice if we could say before which type of change we expect.

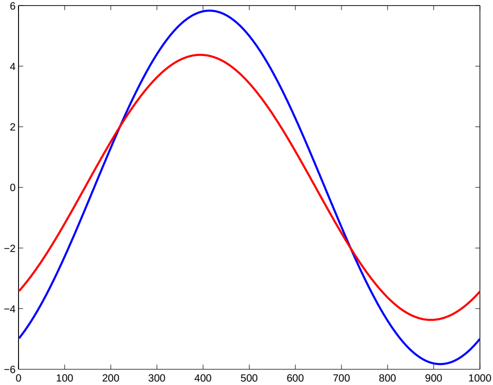
Examples

1. We expect an increase in some components of $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^T$.



Then we can study the sequence of means $\{\bar{X}_i, i = 1, \dots, n\}$ where $\bar{X}_i = \sum_{j=1}^{p} X_{ij}/p$.

2. There is a decrease in the mean of the first p_1 components as well as a decrease in the mean of the last $p - p_2$ components but an increase in the mean of the middle components.



Then we can try to detect this type of change by studying the characteristics:

$$\{\bar{X}_i(p_1, p_2) - \bar{X}_i(1, p_1)\}$$
$$\{\bar{X}_i(p, p_2) - \bar{X}_i(p_1, p_2)\}$$

Often, we are looking for a change in the mean in one, two, ..., m characteristics that have usually a form of linear combinations of the components of the observed vectors:

$$A_i^{(1)} = \mathbf{a}_1^T \mathbf{X}_i, \quad A_i^{(2)} = \mathbf{a}_2^T \mathbf{X}_i, \quad \dots, \quad A_i^{(m)} = \mathbf{a}_m^T \mathbf{X}_i, \quad i = 1, \dots, n.$$

Statistical inference - hypotheses testing

The alternative has to be set before the statistical inference is performed.

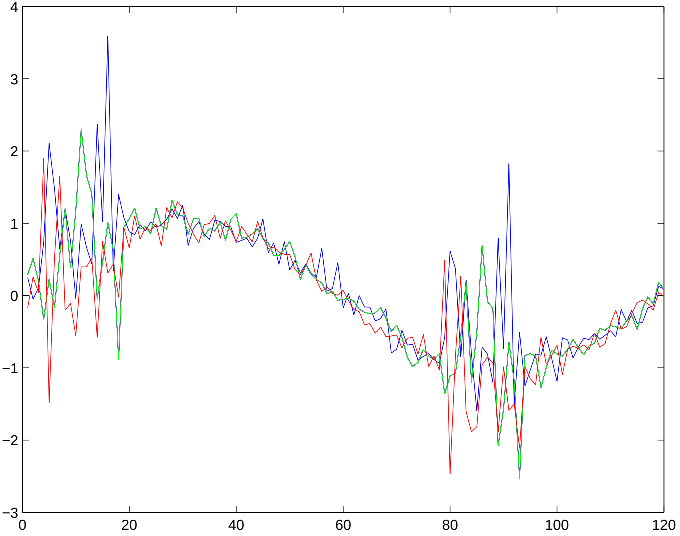
It can happen that we do not know which linear combinations to choose. Is it possible that the data speak for themselves?

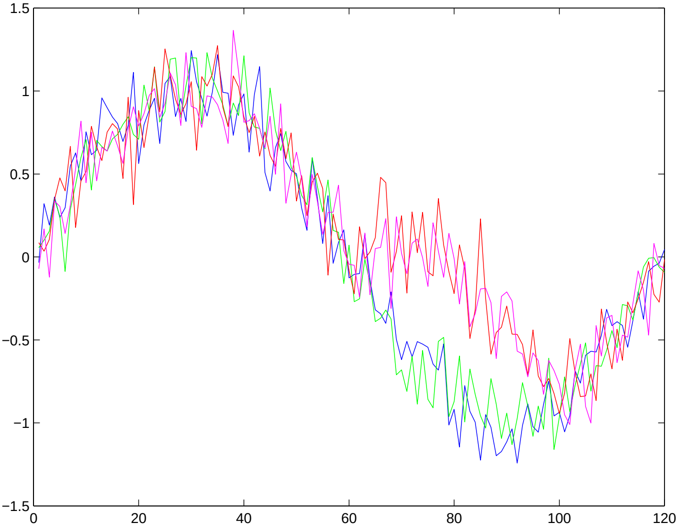
Solution: Choose such linear combinations $\mathbf{a}^T\mathbf{X}$, $\mathbf{b}^T\mathbf{X}, \dots$ that have the largest variance (supposing $||\mathbf{a}|| = 1, ||\mathbf{b}|| = 1, \dots$)

Why? Two possibilities:

1. The large variance of the chosen linear combination is natural - we have chosen a wrong linear combination.

2. The large variance of the chosen linear combination is caused by a shift in the mean of this linear combination - we have chosen a right linear combination.


4
3
2
1
0
−1
−2
−3
0
20
40
60
80
100
120



1.5
1
0.5
0
−0.5
−1
−1.5
0
20
40
60
80
100
120

Method of choosing the linear combinations with the largest variances is well-known - method of principal components.

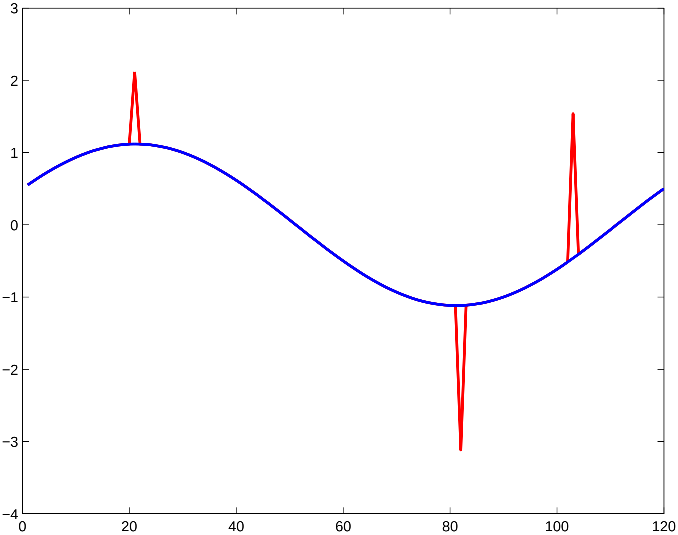
We compute

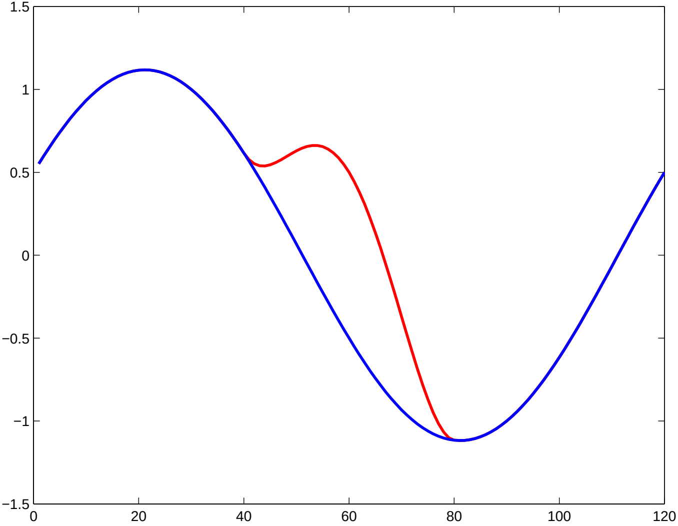
$$\begin{array}{ll} \text{eigenvalues} & \lambda_1 > \cdots > \lambda_p \\ \text{eigenvectors} & \mathbf{w}_1, \ldots, \mathbf{w}_p \end{array}$$

of the variance-covariance matrix $Var\mathbf{X}$ and consider the linear combinations:

$$\mathbf{w}_1^T\mathbf{X}, \ldots, \mathbf{w}_m^T\mathbf{X}, \quad m \text{ is small.}$$

(Then, we test for a change in mean of these linear combinations.)


3
2
1
0
−1
−2
−3
−4
0
20
40
60
80
100
120

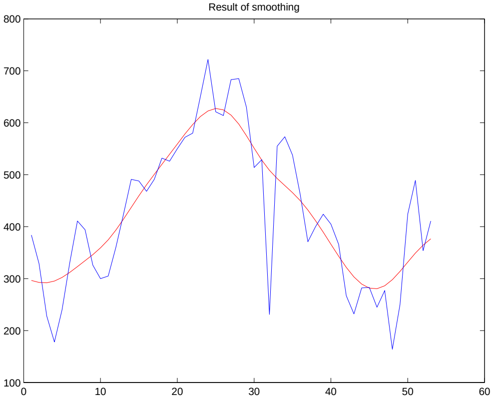


1.5
1
0.5
0
−0.5
−1
−1.5
0
20
40
60
80
100
120

We prefer "smooth" type of changes

We replace the observed vectors with many components by "smooth" functions.

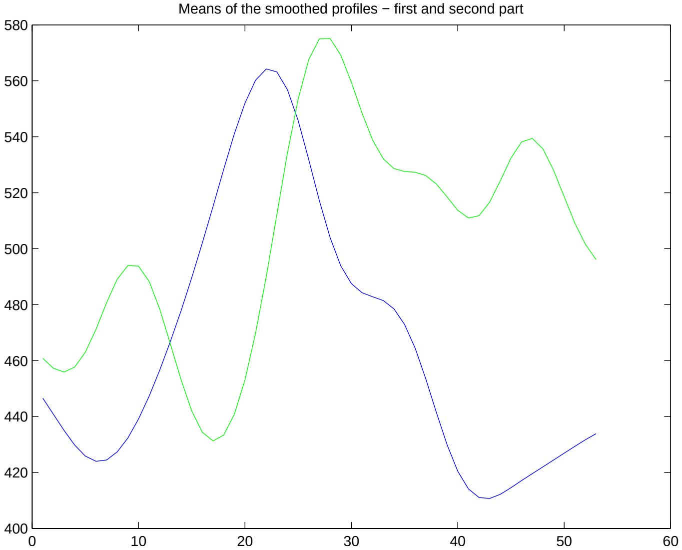
The data are not anymore vectors but they are smooth functions.



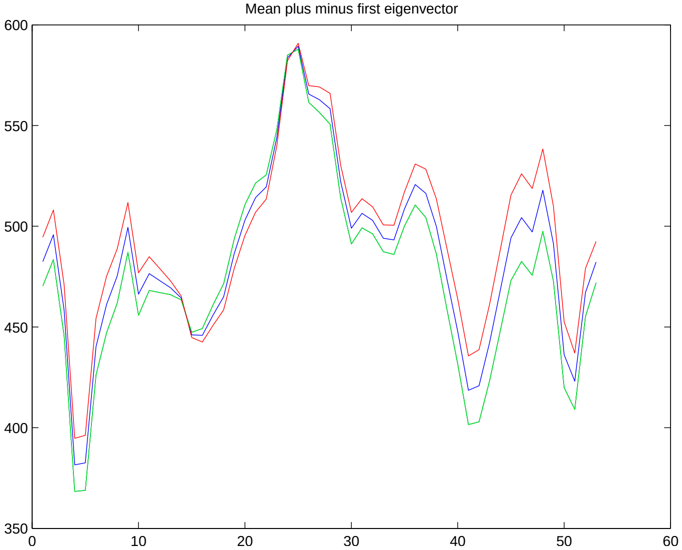
Functional data analysis

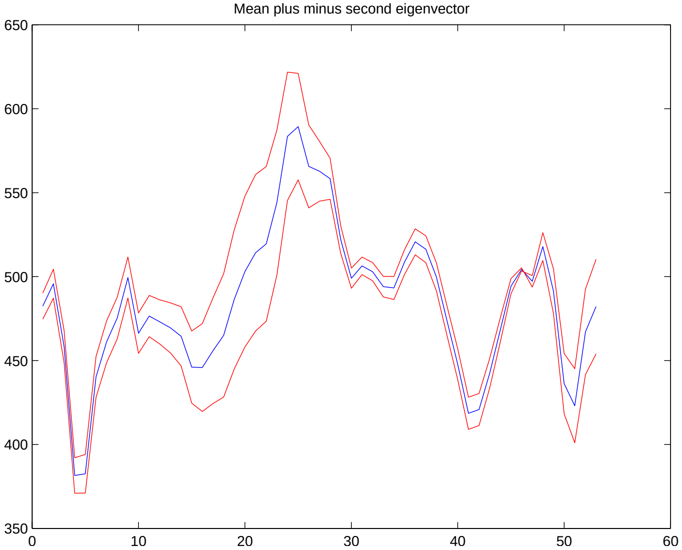
Instead of dealing with a space R^p we are dealing with a space of integrable functions $L^2([0, T])$. We can use again the method of principal components replacing

$$
\begin{aligned}
\mathbf{w}_1^T \mathbf{X}_i \quad &\text{by} \quad \int_0^T \mathbf{w}_1(t)\mathbf{X}_i(t)\,dt \\
\mathbf{w}_2^T \mathbf{X}_i \quad &\text{by} \quad \int_0^T \mathbf{w}_2(t)\mathbf{X}_i(t)\,dt \\
&\vdots
\end{aligned}
$$



Means of the smoothed profiles − first and second part
580
560
540
520
500
480
460
440
420
400
0
10
20
30
40
50
60


Mean plus minus first eigenvector
600
550
500
450
400
350
0
10
20
30
40
50
60


Mean plus minus second eigenvector
650
600
550
500
450
400
350
0
10
20
30
40
50
60