# 4. Data alignment and clustering

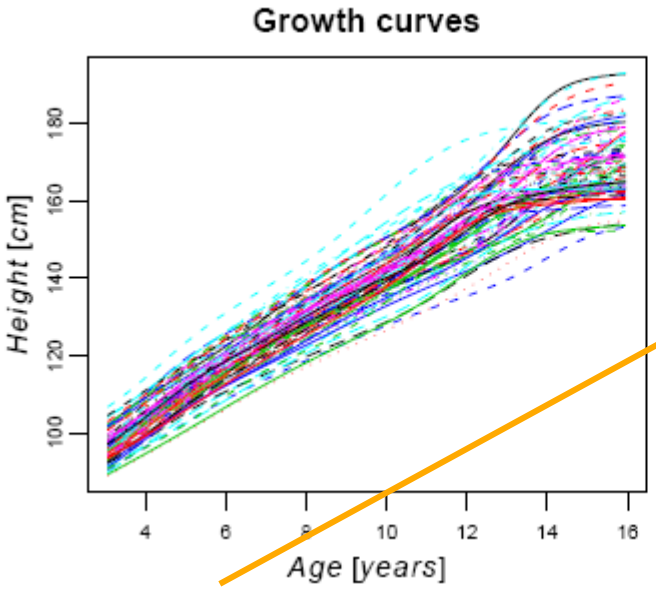
# Course Agenda

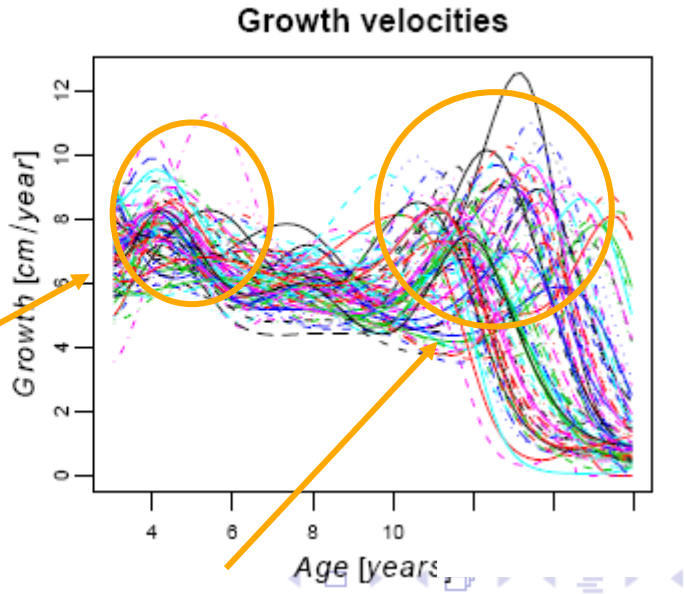
1. **Hilbert space model for functional data**
    - 1.1. Basics notions on Hilbert spaces
    - 1.2. Hilbert space embedding for functional data
    - 1.3. Formal definition of functional data
2. **Smoothing and interpolation of functional data**
    - 2.1. Basis function
    - 2.2. Least square smoothing
    - 2.3. Smoothing with a differential penalization
3. **FDA & Dimensionality reduction in Hilbert spaces**
    - 3.1. Functional Principal Components in Hilbert spaces
    - 3.2. Examples in L2
4. **Data alignment and clustering**
    - 4.1 Phase and amplitude variability
    - 4.2 Landmark and continuous registration
    - 4.3 Decoupling phase and amplitude variability
    - 4.4 K-mean alignment

# 4.1 Misalingned data: the Berkeley growth study data

(Tuddenham and Synder, 1954)

- the data include the heights of 93 children, 54 girls and 39 boys, measured on 31 time instances, not equally spaced;
- the functional form of these data has been reconstructed using monotone smoothing splines.



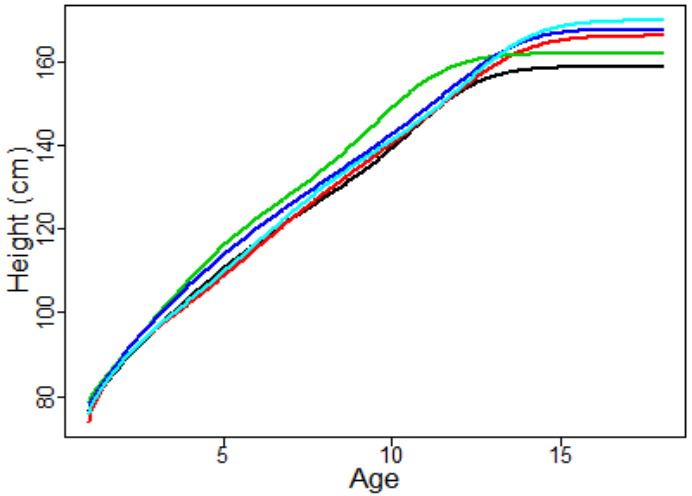

All curves follow a similar course

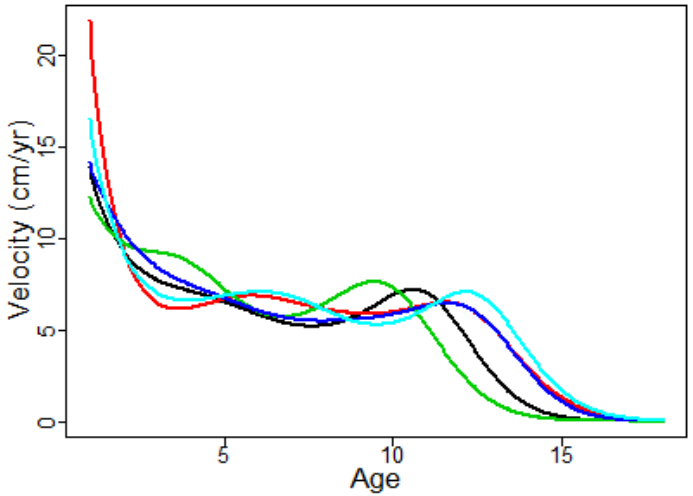
minor peak, 2 -5 years (mid-spurt)

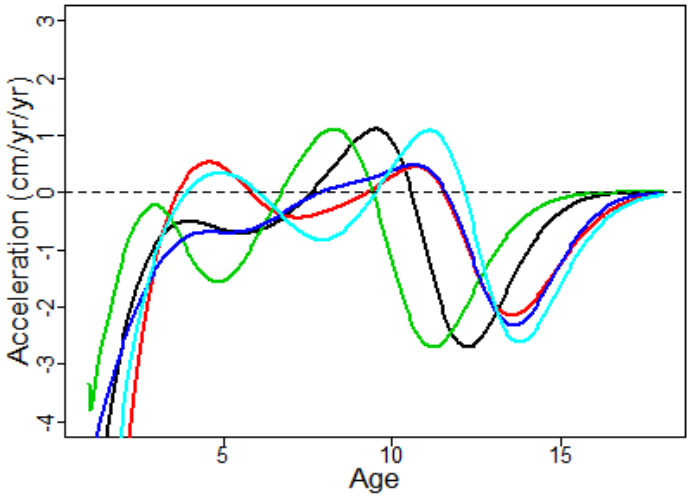
sharp peak, 10 – 16 years (pubertal-spurt)

However, each child follows his/her own biological clock

# 4.1 The growth curves of 5 girls



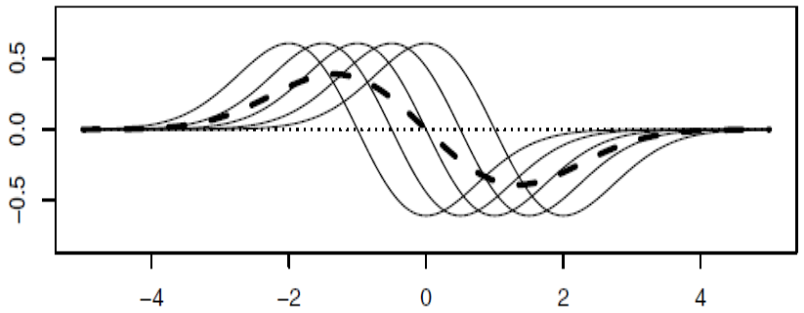




Growth, growth velocities and accelerations for 5 girls of the Study

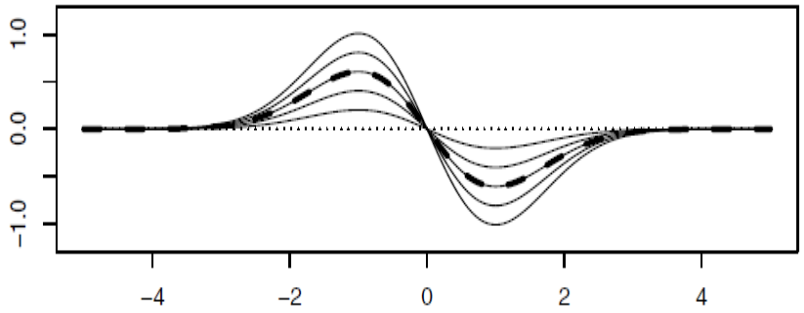
# 4.1 Phase and amplitude variability

Ramsay Silverman 2005 Springer

Phase variability




Amplitude variability



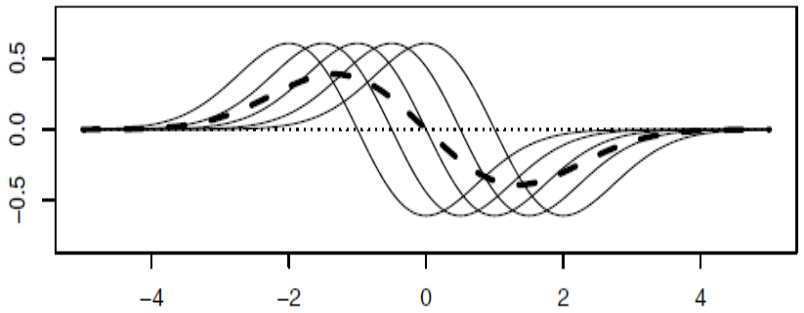

Phase variability: different curves exhibit more or less the same features but these features occur at different times or space locations for different statistical units.

If not taken properly into account, the misalignment acts as a confounding factor and may blur subsequent analyses.

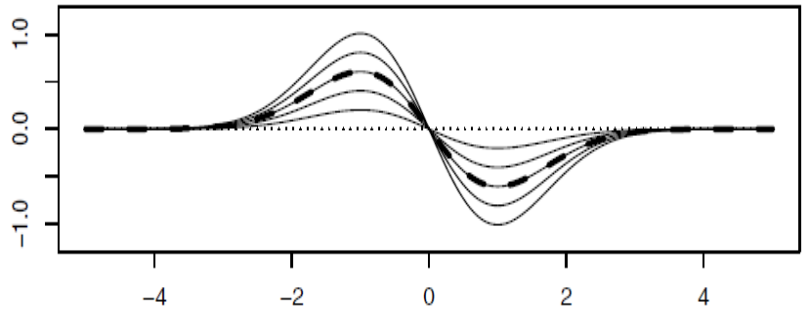
# 4.1 Phase and amplitude variability

Ramsay Silverman 2005 Springer

Phase variability




Amplitude variability




## Registration of a set of functions

Given $n$ curves $c_1(t)\ldots, c_n(t)$, find suitable warping functions $h_1(t),\ldots, h_n(t)$ such that $c_1(h_1(t))\ldots, c_n(h_n(t))$ are the most similar.

The functions $h_i$ should be increasing; they capture the phase variability. Amplitude variability is the remaining variability among the aligned curves in the vertical direction.

In some cases, time or location is merely shifted from curve to curve, for example, because the measurements are started at random time points. For these situations, it is natural to use $h_i(t)=t+delta_i$. In other situations, phase variation is a matter of dilation, in which case $h_i(t)=alpha_i t$ is a natural choice of warping function. In yet other situations, the time or space deformation is more complex.

# 4.1 Phase and amplitude variability

Ramsay Silverman 2005 Springer

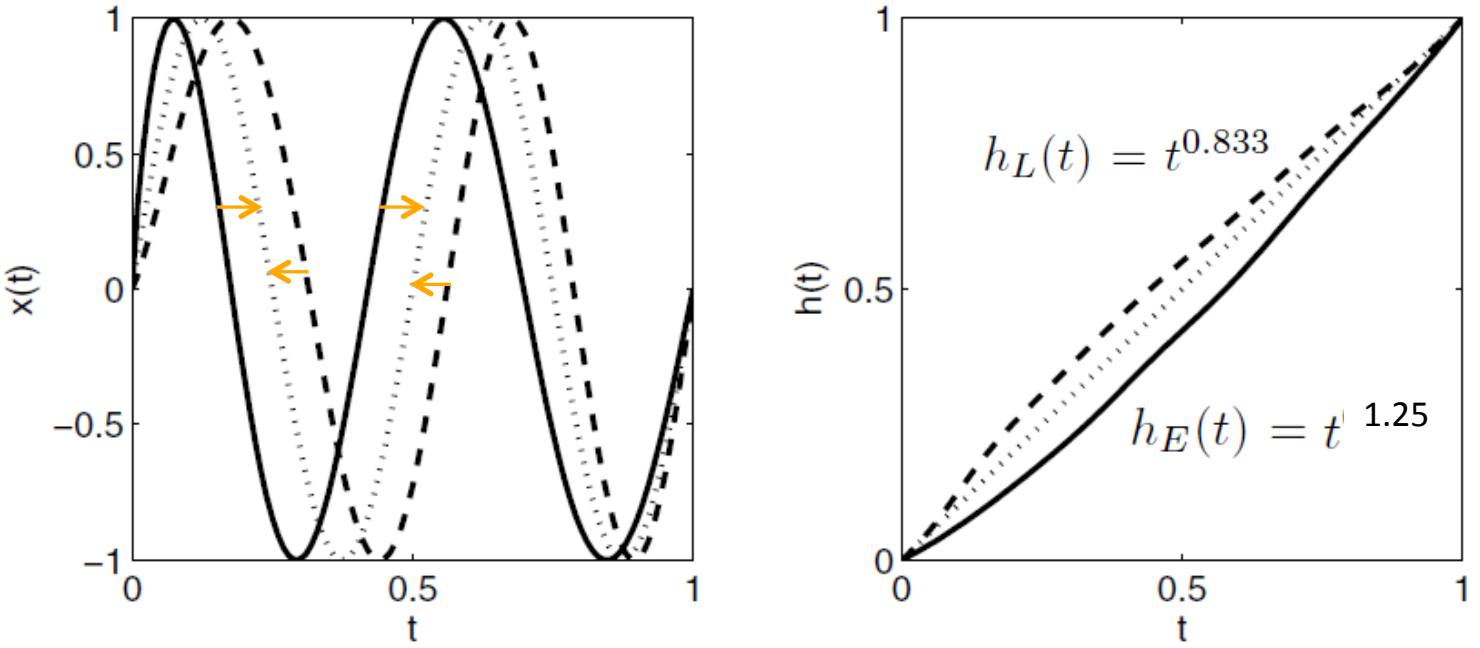
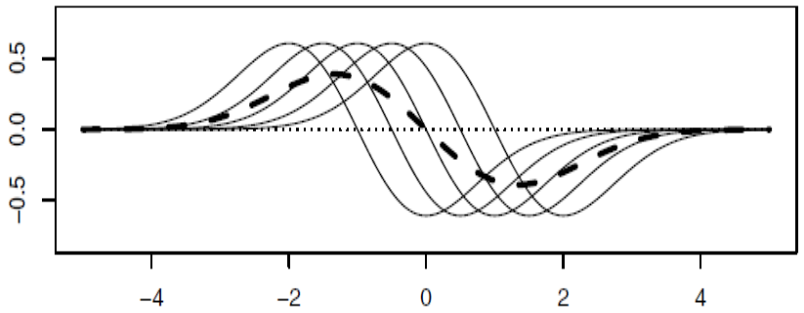



Figure 7.9. The left panel shows the target function, $x_0(t) = \sin(4\pi t)$, as a dotted line; an early function, $x_E(t) = \sin(4\pi t^{0.8})$, as a solid line; and a late function, $x_L(t) = \sin(4\pi t^{1.2})$, as a dashed line. The corresponding warping functions that register the early and late curves to the target are shown in the right panel.

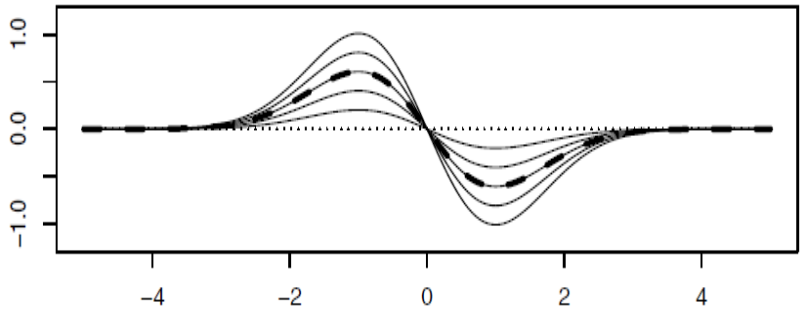
# 4.1 Phase and amplitude variability

Ramsay Silverman 2005 Springer

Phase variability




Amplitude variability




**Registration of a set of functions**

Find **suitable** warping functions $h_1, \ldots, h_n$ such that $c_1 \circ h_1, \ldots, c_n \circ h_n$ are the most **similar**.

**Landmark Approach**: known **landmarks** along the curves that are aligned so that landmarks occurs at the same abscissa points.

**Continuous Approach**: define a measure of similarity/dissimilarity between curves, that are aligned in order to maximize/minimize their similarity/dissimilarity.

# 4.2 Landmark registration

Ramsay Silverman 2005 Springer

Landmarks: significant (univocally identifiable) shape-events in a curve, e.g. crossings of zero, peaks, valleys, points of inflection.

$c_1, \ldots, c_n$, where $c_i : [0, T] \to \mathbb{R}^d$

Suppose

- $L$ landmarks; for the $i-$th curve, located at $t_{i1}, \ldots, t_{iL}$
- a template curve $c_0$ is available with landmark locations $t_{01}, \ldots, t_{0L}$
  If not, we can define $t_{0j}$ as the average of the $t_{ij}$'s

Warping function for the $i-th$ curve: any strictly increasing function $h_i$ s.t.

- $h_i(0) = 0$
- $h_i(t_{0j}) = t_{ij}$, for $j = 1, \ldots, L$
- $h_i(T) = T$

Notation: the warping functions will be the inverse of these $h_i$

▶ $(0, 0), (t_{01}, t_{i1}), .0\ .\ , (t_{1L}, t_{iL}), (T, T)$ : interpolated by a piece-wise line, a polygon or higher order monotone splines (strictly increasing)

# 4.2 Continuous registration

- Landmark-based registration may require significant user input and can be sensitive to the accuracy of the landmark identification.

- In some applications it is not possible to identify well-defined features that can be taken as landmarks

**Alternative strategy**: Continuous registration

Main idea:

- definition of a suitable distance (or closeness) measure between curves, which measures dissimilarity (or similarity) between curves.

- the curves are thus aligned by warping their time or space abscissa parameters choosing the optimal warping function in some class of admissible warping functions in order to minimize the final distance among the curves or, equivalently, maximize their final similarity.

# 4.2 Continuous registration

The problem of decoupling amplitude and phase variability is not univocally defined as different measures of distance or similarity between curves can be considered, as well as different classes of admissible warping functions (e.g., simple translations or dilations, increasing linear transformations or more complex increasing transformations), leading to different registration results.

The choice of the couple formed by dissimilarity/similarity measure and admissible warping functions defines the distinction between phase variability and amplitude variability in the specific problem under analysis.

This choice must thus be problem specific.

# 4.3 Decoupling phase and amplitude variabilities

Sangalli, Secchi, Vantini, Veneziani 2009 JASA

$(\rho, W)$ must satisfy properties that ensure that the aligning problem is well-posed and the corresponding procedure is coherent

- $\rho$
  - Bounded
  - Reflexive
  - Symmetric
  - Transitive
- $W$
  - Convex vector space
  - Group structure with respect to function composition
- $(\rho, W)$ Properties of coherence

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \rho(\mathbf{c}_1 \circ h, \mathbf{c}_2 \circ h), \quad \forall\, h \in W$$

*W*-invariance of the index

(Isometry of the group, parallel orbits)

→ $\rho(\mathbf{c}_1 \circ h_1, \mathbf{c}_2 \circ h_2) = \rho\left(\mathbf{c}_1 \circ h_1 \circ h_2^{-1}, \mathbf{c}_2\right) = \rho\left(\mathbf{c}_1, \mathbf{c}_2 \circ h_2 \circ h_1^{-1}\right)$

$(\rho, W)$ defines on the considered set of functions $\mathcal{C}$ a partition in equivalence classes

# 4.3 Decoupling phase and amplitude variabilities

Vantini 2012 TEST

Sangalli Secchi Vantini 2014b EJS

| dissimilarity $d$ | warpings $W$ |
|---|---|
| $\|\|c_1 - c_2\|\|$ | $W_{shift}$ |
| $\|\|c_1' - c_2'\|\|$ | $W_{shift}$ |
| $\|\|(c_1 - \bar{c_1}) - (c_2 - \bar{c_2})\|\|$ | $W_{shift}$ |
| $\left\|\left\|(c_1' - \bar{c_1'}) - (c_2' - \bar{c_2'})\right\|\right\|$ | $W_{shift}$ |
| $\left\|\left\|\frac{c_1}{\|\|c_1\|\|} - \frac{c_2}{\|\|c_2\|\|}\right\|\right\|$ | $W_{affinity}$ |
| $\left\|\left\|\frac{c_1'}{\|\|c_1'\|\|} - \frac{c_2'}{\|\|c_2'\|\|}\right\|\right\|$ | $W_{affinity}$ |
| $\left\|\left\|\text{sign}(c_1')\sqrt{\|c_1'\|} - \text{sign}(c_2')\sqrt{\|c_2'\|}\right\|\right\|$ | $W_{diffeomorphism}$ |

# 4.3 Curve alignment: iterative procedure



If a template (prototype) curve $\varphi$ is known, then it is enough to align each curve to this template

If the template is unknown then it must be estimated from the data, leading to a complex optimization problem

find $\varphi \in \mathcal{C}$ and $\underline{\mathbf{h}} = \{h_1, \ldots, h_N\} \subset W$ such that

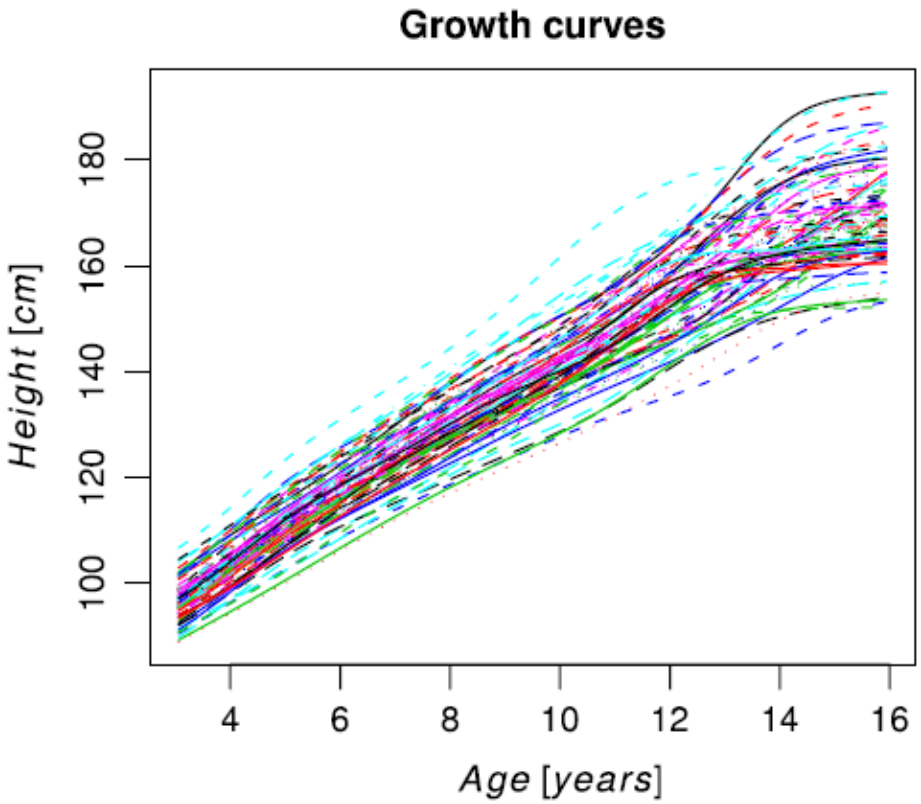
$$\frac{1}{N}\sum_{i=1}^{N} \rho(\varphi, \mathbf{c}_i \circ h_i) \geq \frac{1}{N}\sum_{i=1}^{N} \rho(\psi, \mathbf{c}_i \circ g_i)$$

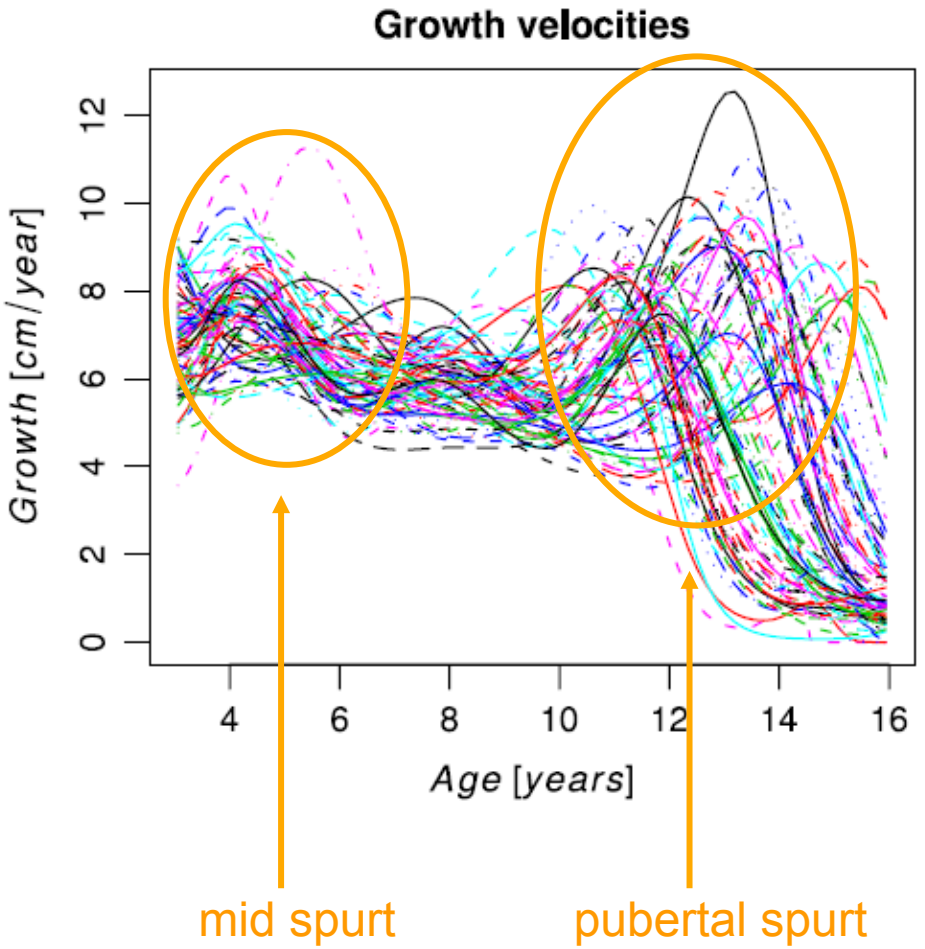
for any other $\psi \in \mathcal{C}$ and $\underline{\mathbf{g}} = \{g_1, \ldots, g_N\} \subset W$

# 4.3 Example: Berkeley Growth Study data

(Tuddenham and Synder, 1954)

Sangalli Secchi Vantini Vitelli 2010 CSDA




93 children, 39 boys and 54 girls

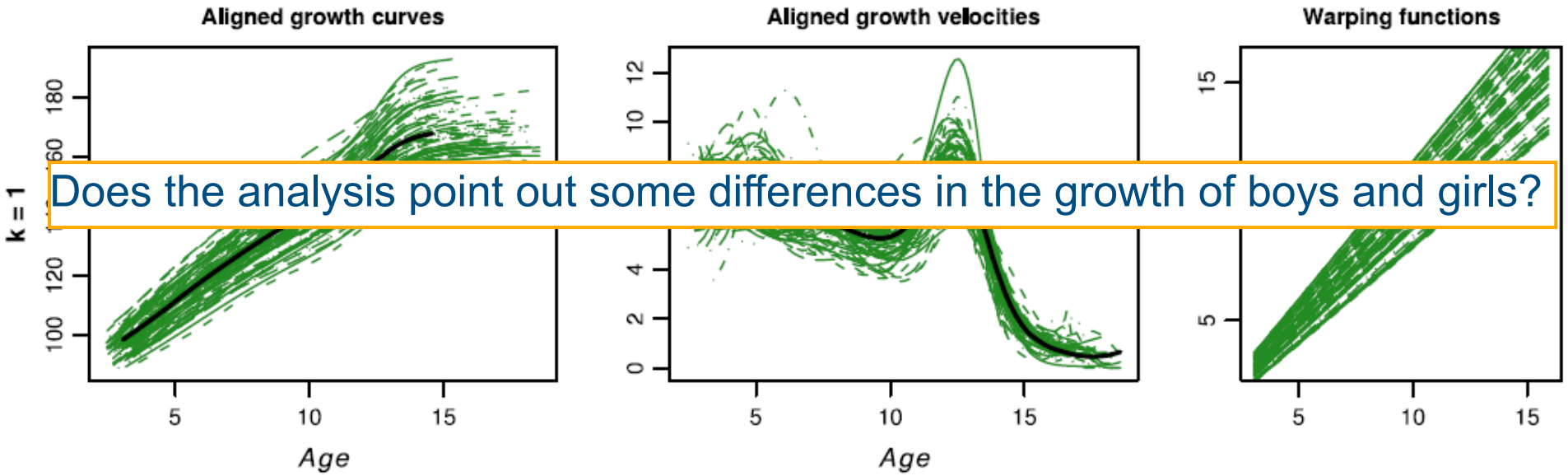
mid spurt

pubertal spurt

Curves estimated by monotonic cubic regression splines, implemented using the R package *fda*

# 4.3 Berkeley Growth Study data

Sangalli Secchi Vantini Vitelli 2010 CSDA




Does the analysis point out some differences in the growth of boys and girls?

Results of continuous alignment using the following similarity index and class of warping functions $(\rho, W)$:

$$\rho(c_i, c_j) = \frac{\int_{S_{ij}} c_i'(s) c_j'(s) ds}{\sqrt{\int_{S_{ij}} c_i'(s)^2 ds} \sqrt{\int_{S_{ij}} c_j'(s)^2 ds}}$$

← the focus is on growth patterns, rather than on the absolute heights of the children or on their more or less pronounced growths

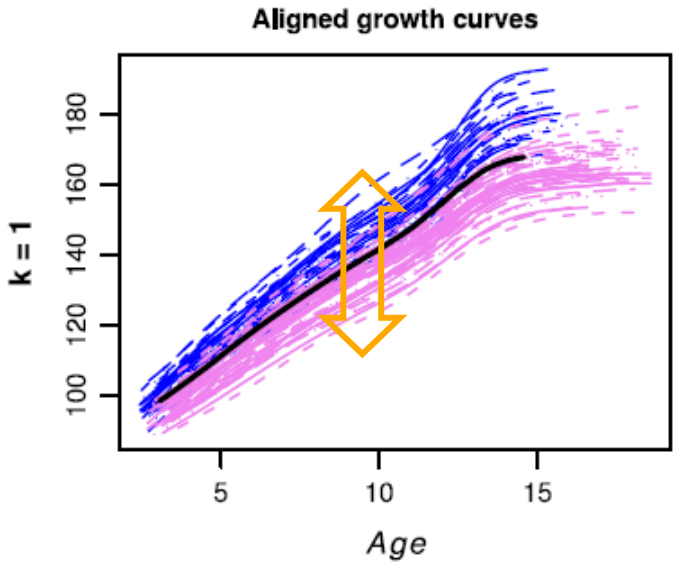
$$\rho(c_i, c_j) = 1 \;\Leftrightarrow\; \exists a \in \mathbb{R}, b \in \mathbb{R}^+ : c_i(t) = a + b\, c_j(t)$$

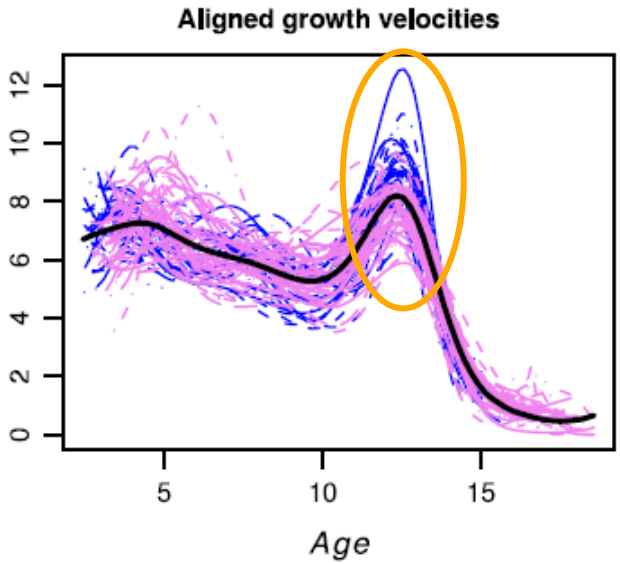
$$W = \{h : h(t) = mt + q \text{ with } m \in \mathbb{R}^+, q \in \mathbb{R}\}$$

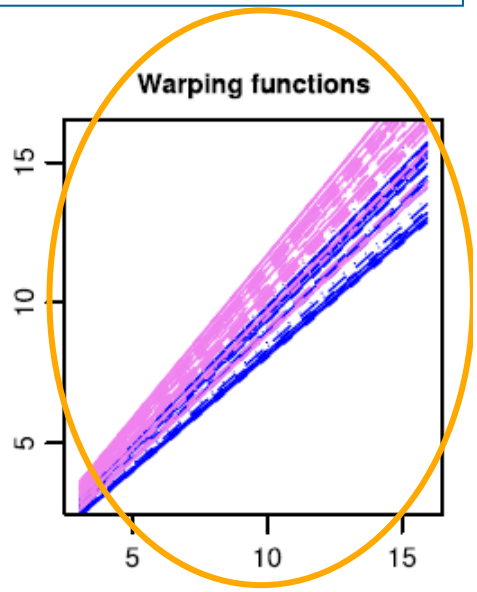
← constant modifications of the running speeds of the children biological clocks

# 4.3 Berkeley Growth Study data

Sangalli Secchi Vantini Vitelli 2010 CSDA




Once the biological clocks are aligned

the height of boys stochastically dominates the one of girls for any registered biological age

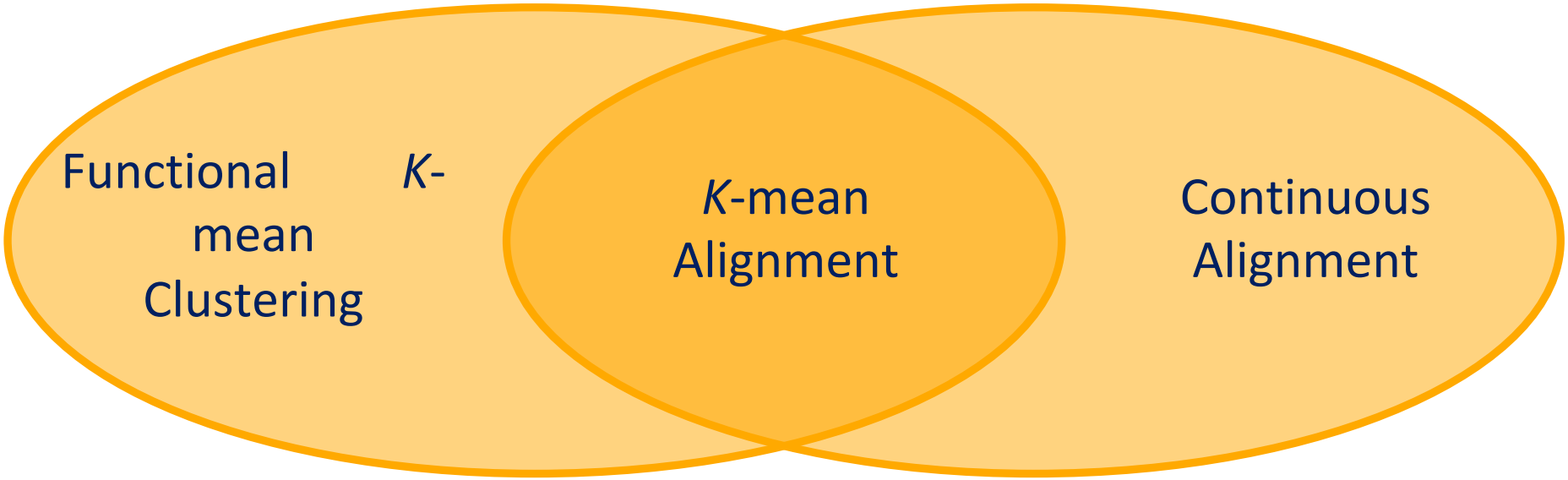
boys have a more pronounced growth during puberty (more prominent growth velocity peak)

Neat separation of boys and girls in the phase.

The biological clocks of boys and girls run at different speeds

# 4.4 Simultaneous registration and classification: the K-mean Alignment algorithm

Sangalli Secchi Vantini Vitelli 2010 CSDA



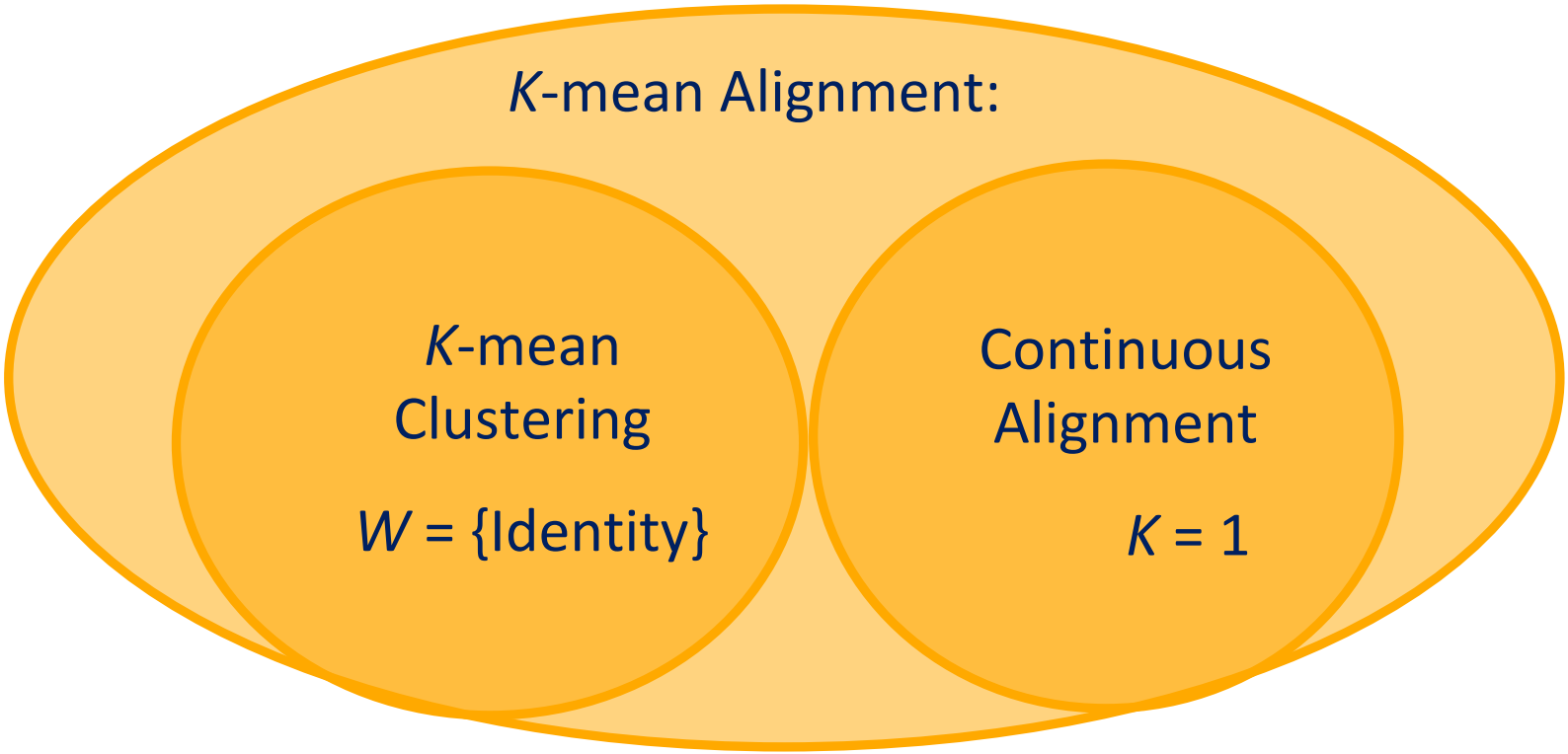

→ *K*-mean Clustering
with warping allowed

→ Continuous Alignment
with *K* templates

Code for *K*-mean alignment: R package `fdakma`, available from CRAN

# 4.4 K-mean Alignment

Sangalli Secchi Vantini Vitelli 2010 CSDA




Code for *K*-mean alignment: R package `fdakma`, available from CRAN

# 4.4 K-mean Alignment

Sangalli Secchi Vantini Vitelli 2010 CSDA

Goal of **Alignment**:
**Decoupling Phase and Amplitude Variability**

+

Goal of ***K*-mean** Clustering:
**Decoupling Within and Between-cluster (Amplitude) Variability**

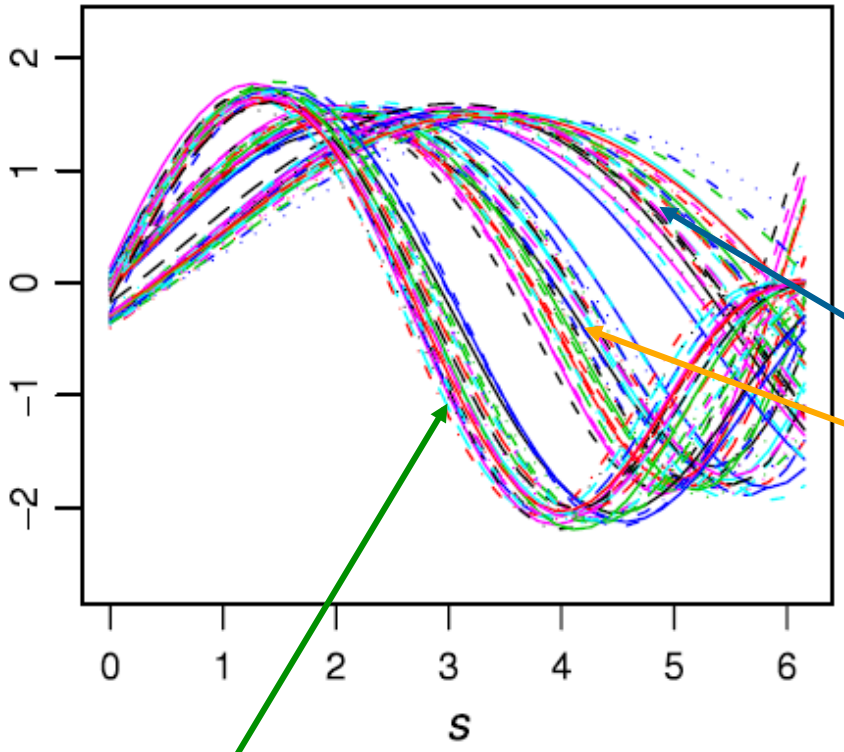
=

Goal of ***K*-mean Alignment**:
**Identifying Phase Variability, Within-cluster Amplitude Variability and Between-cluster Amplitude Variability**

(disclosing clustering in the phase)

# 4.4 A small part of a larger simulation study…

Sangalli Secchi Vantini Vitelli 2010 CSDA




2 AMPLITUDE CLUSTERS
(2 template curves)

How many different
prototype curves have
generated these data?

ONE has associated a
further CLUSTERING IN
THE PHASE

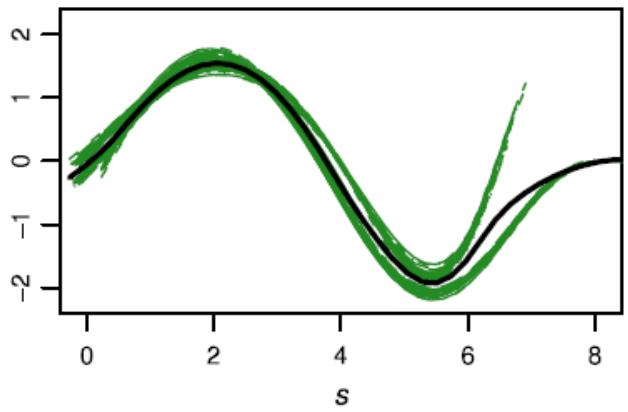
$$1 * \sin(s) + 1 * \sin\left(\frac{s^2}{2\pi}\right)$$

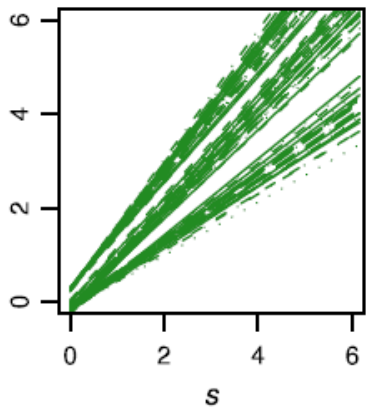
$$2 * \sin(s) - 1 * \sin\left(\frac{s^2}{2\pi}\right) + (1 + \varepsilon_{4i})s) + (1 + \varepsilon_{2i}) * \sin\left(\frac{(\varepsilon_{3i} + (1 + \varepsilon_{4i})s)^2}{2\pi}\right)$$

## Aligned and clustered curves

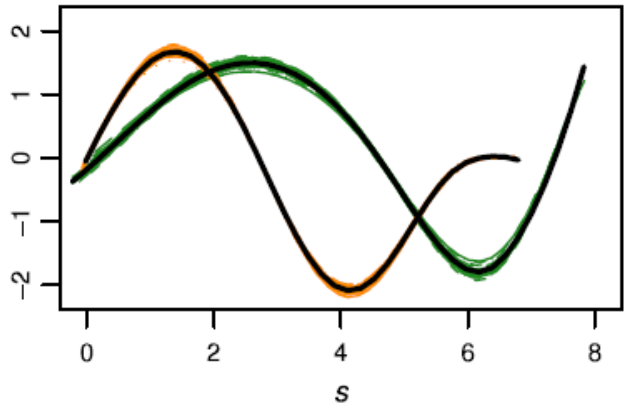
## Warping functions

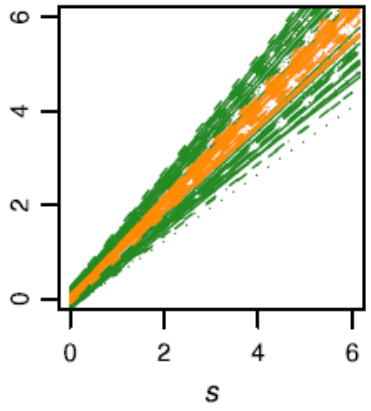
K = 1



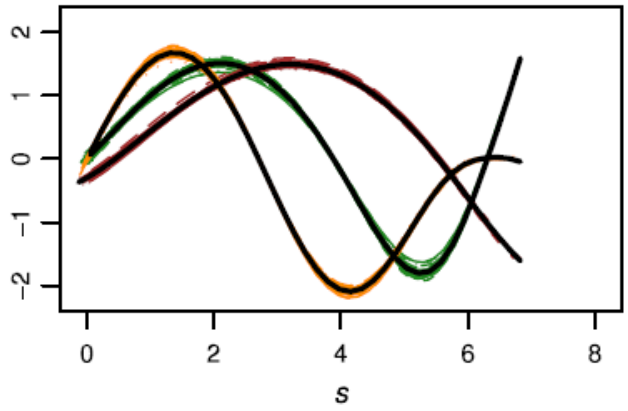



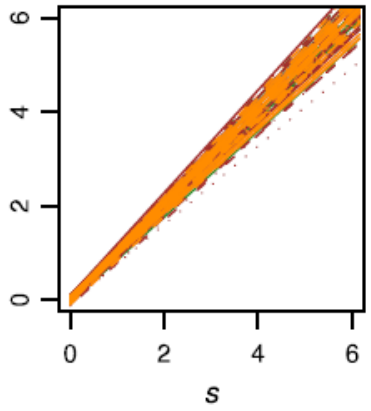

K = 2



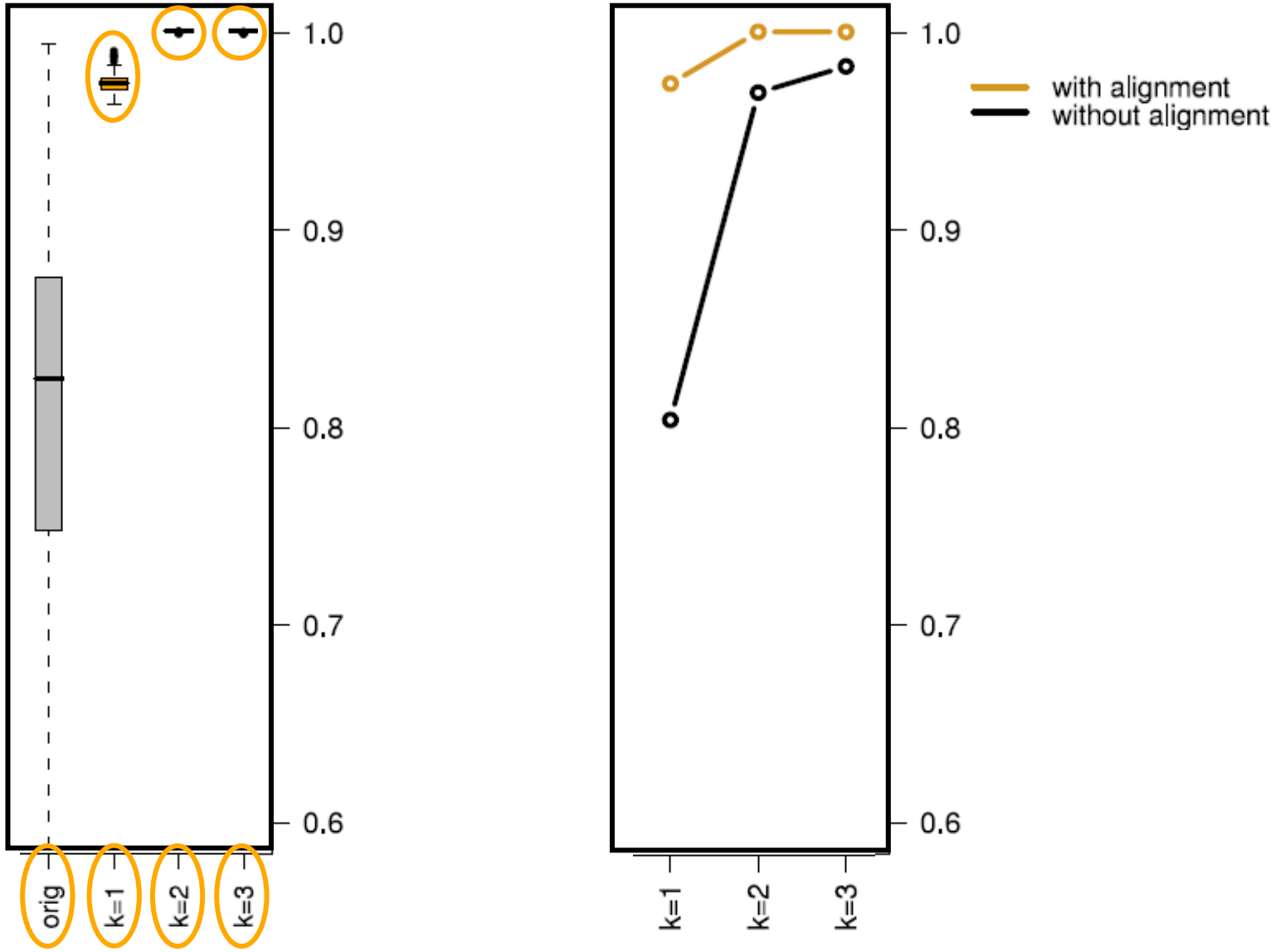




K = 3



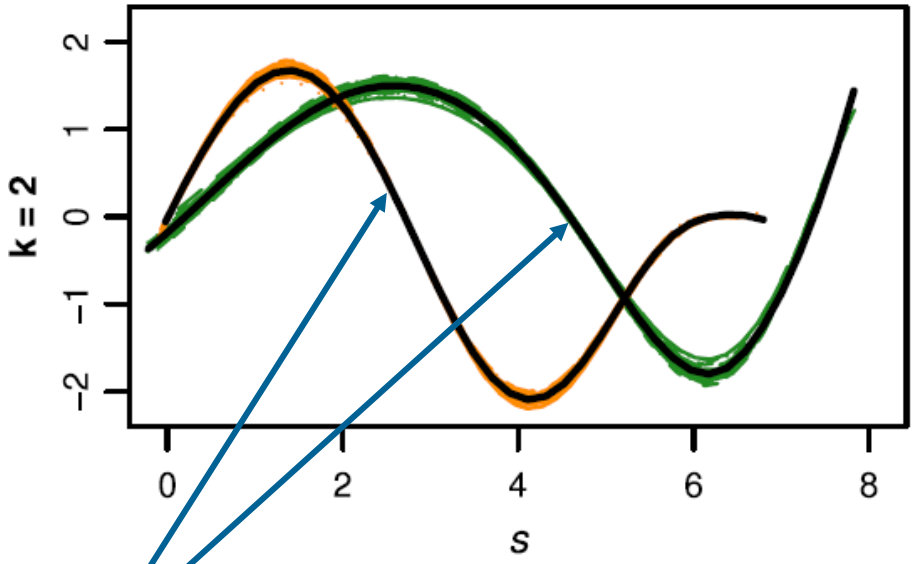



# 4.4 A small part of a larger simulation study…

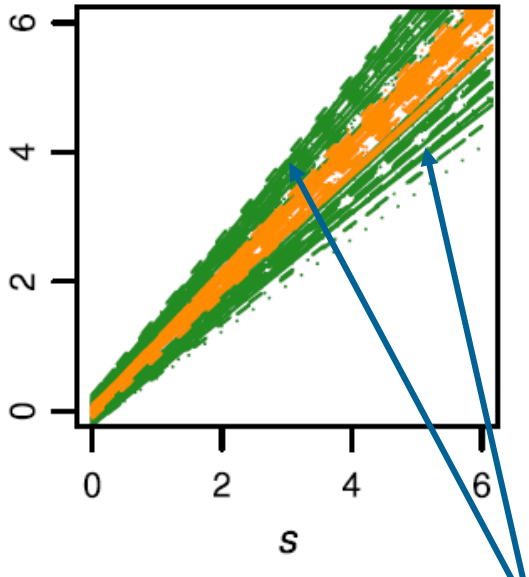


# 4.4 A small part of a larger simulation study…

Aligned and clustered curves




Warping functions




Curve aligned in 2 AMPLITUDE CLUSTERS

CLUSTERING IN THE PHASE

*k*-mean alignment is able to efficiently detect true amplitude clusters and also to disclose clustering structures in the phase