Structural Components in Functional Data

Juhyun Park*, Theo Gasser and Valentin Rousson
Department of Biostatistics
University of Zürich

July 24, 2007

Abstract

Analyzing functional data often leads to finding common factors, for which functional principal components analysis proves to be a useful tool to summarize and characterize the random variation in a function space. The representation in terms of eigenfunctions is optimal in the sense of L_2 approximation. However, the eigenfuntions are not always directed towards an interesting and interpretable direction in the context of functional data and thus could obscure the underlying structure. This paper proposes an alternative to functional principal component analysis that produces directed components which may be more informative and easier to interpret. These *structural* components are similar to principal components, but are adapted to situations in which the domain of the function may be decomposed into disjoint intervals such that there is effectively independence between intervals and positive correlation within intervals. The approach is demonstrated with examples as well as real data. Properties for special cases are also studied.

Keywords: Functional data analysis, Functional principal component analysis, PCA, Longitudinal data, Smoothing

1 INTRODUCTION

Repeated measurements in the form of curves are increasingly common in the fields of biomedicine and physical sciences. Examples include blood pressure profiles over 24 hours, evoked brain potentials and growth curves. Individual measurements are

*Address for correspondence: Juhyun Park, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, U.K. Email: juhyun.park@lancaster.ac.uk

taken at consecutive time points and repeatedly observed for different subjects. Such measurements are generally called functional data.

Figure 1 about here.

An example of children growth data is shown in Figure 1. These are height measurements (left) and velocity curves (right) of 10 boys at ages between 2 and 21 years (Gasser et al., 1984). Velocity curves are estimated nonparametrically with Gasser-Müller estimator. In order to perform a functional data analysis, these curves will be aligned to eliminate time variability, known as registration. In this paper, we shall assume that registration, if necessary, has been carried out and we shall treat the registered curves as raw data. For details on registration, we refer to Ramsay and Silverman (1997, 2002).

Usually the sample of curves is assumed to have some homogeneous structure in the functional shape, while allowing for individual variability. This variability is commonly charaterized with a few components. Let y_{ij} be the jth observation on the ith individual at time t_j and consider the following regression model

$$y_{ij} \equiv y_i(t_j) = \mu(t_j) + \sum_{k=1}^{K} \xi_{ik}\phi_k(t_j) + \epsilon_{ij}\,, i = 1, \cdots, n; j = 1, \cdots, T\,,$$

where ϵ_{ij} are independent errors. Here the design points are assumed the same for each subject. This is not an essential assumption but it will simplify presentation in Section 2. From now on, such ϕ_ks are referred to as *components*.

When viewing each curve as an independent realization of a stochastic process $X(t)$ in function space, the model has an optimal representation when the common functions $\{\phi_k\}$ and the coefficients $\{\xi_{ik}\}$ are derived from eigenfunctions and eigenvalues of the covariance function of X. These are known as functional principal components. Such models appear in the context of both longitudinal and functional data analysis. In longitudinal data analysis, the random coefficients ξ_{ik} are often associated with random effects with known covariates and may have sparse design points. See also Yao et al. (2005). While the components are prespecified in the longitudinal models, they may be completely determined by the data in functional data analysis. However, situation may arise when some restriction would be useful without fully specifying the components. In particular, when some variability shows certain structure along the time axis, it is beneficial to have components that reflect such a phenomenon. For example, whether the growth process experiences a qualitative change in time would be of interest.

We mainly focus on nonnegative covariance functions, although in practice we would tolerate small negative entries to extract the essential features of a basically positive relationship.

Figure 2 about here.

Consider a simple scenario with two underlying subprocesses that have almost non–overlapping support as shown in Figure 2. Assume that a sample of curves is generated from the sum of these processes when 1) the two processes are independent and when 2) they are positively highly correlated. In the latter case, it can be viewed as being one homogeneous process.

Figure 3 about here.

Figure 3 shows the result of functional PCA for these two cases. The left column corresponds to the nearly uncorrelated case and the right column to the positively correlated case. The first components shown in the top row both suggest an overall level of the process as major source of variability (in our terminology these are *block* components) and the second components, in the second row, suggest the difference between the first and the second processes as the next source (we shall call these *difference* components). Although the percentage of variance differs, functional PCA provides *qualitatively* the same answer whether there are two independent subprocesses or only one process.

Figure 4 about here.

On the other hand, the covariance structure, seen as contour plots in top panels of Figure 4, clearly indicate that the observed process is approximated by two subprocesses in the former case and by one process in the latter case. Below are the corresponding correlation functions. Note that the minimum correlation for the latter is 0.69, compared to 0.2 for the former.

In this paper, we shall define *structural* components to reflect this information both quantitatively and qualitatively. We introduce *block* components and *difference* components to make a distinction between the two cases. In Section 2, we introduce a general framework to obtain *block* components that reflect the underlying structure. We use the simulated examples to illustrate the procedure. Statistical properties are studied in Section 3. Numerical performance is evaluated in Section 4. In particular, comparison to functional PCA and its Varimax rotation is made in simulation studies. Application to real data is also included

2 METHODOLOGY

Consider a stochastic process $X(t)$ with compact support $\mathcal{T} = [0, T]$. Denote the mean function by $\mu(t)$ and the covariance function by $\gamma(s,t) = cov(X(s), X(t))$ and assume that

(A1) $\gamma(s,t) \geq 0$ for all $s,t \in \mathcal{T}$.

Processes with predominantly positive covariances are frequent in functional data and in many cases this property should be satisfied to a good approximation.

We shall decompose a stochastic process into a system of components $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_q$ as in the regression model of Section 1. Similar in spirit to decomposition of analysis of variance, where major variability is captured by main effect and contrasts, we model the process with *block* components and *difference* components. These constitute *structural* components in our model. Structural components were introduced in the multivariate context by Rousson and Gasser (2004). Formal definitions will be introduced below.

2.1 Block and difference components

A component is called a *block* component if $\boldsymbol{\beta}(t) \geq 0$ (or $\boldsymbol{\beta}(t) \leq 0$) for all $t \in \mathcal{T}$, the domain where it is strictly positive (strictly negative) being connected. A *difference* component is an element of nonblock components, i.e. where the $sgn(\boldsymbol{\beta})$ is not constant. A simple example of difference component is seen when the domain where it is strictly positive is connected and when the domain where it is strictly negative is connected. For identifiability we assume that $\int \boldsymbol{\beta}^2(t)\, dt = 1$. We are mainly interested in deriving block components but difference components could be of further interest. For example, we would like to have two block components in the case of a sum of two uncorrelated subprocesses, but to have only one block component in the case of a sum of two highly correlated subprocesses (see Figure 2).

2.2 Correlation between components

Each component function is associated with a random variable $X_k = \int \boldsymbol{\beta}_k(t) X(t)\, dt$. We measure the correlation between two components $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_l$ by the random variables induced by them. Define

$$Corr(\boldsymbol{\beta}_k, \boldsymbol{\beta}_l) = \frac{\int\int \boldsymbol{\beta}_k(s)\gamma(s,t)\boldsymbol{\beta}_l(t)\, ds\, dt}{\sqrt{\int\int \boldsymbol{\beta}_k(s)\gamma(s,t)\boldsymbol{\beta}_k(t)\, ds\, dt}\sqrt{\int\int \boldsymbol{\beta}_l(s)\gamma(s,t)\boldsymbol{\beta}_l(t)\, ds\, dt}}$$

Thus, two components are said to be *uncorrelated* if $\int \boldsymbol{\beta}_k(s)\gamma(s,t)\boldsymbol{\beta}_l(t)\, ds\, dt = 0$. On the other hand, two components are said to be *orthogonal* if $\int \boldsymbol{\beta}_k(t)\boldsymbol{\beta}_l(t)\, dt = 0$. Functional principal components is the only system of components which is orthogonal and uncorrelated. In our approach, components may be non-orthogonal and/or correlated. To avoid that components share too much information, we shall concentrate on systems where the maximal correlation

$$C = \max_{k \neq l} Corr(\boldsymbol{\beta}_k, \boldsymbol{\beta}_l)$$

is smaller than some threshold C_{max}.

2.3 Variance extracted by a system of components

We assess optimality of components by corrected sum of variances explained by components. Following Gervini and Rousson (2004), we consider

$$Var(\boldsymbol{\beta}_1) + \sum_{k=2}^{q} Var(\boldsymbol{\beta}_k|\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{k-1}),$$

where

$$Var(\boldsymbol{\beta}_k|\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{k-1}) = \int\int \boldsymbol{\beta}_k(s)\gamma_{(k-1)}(s,t)\boldsymbol{\beta}_k(t)\, ds\, dt\,.$$

Here $\gamma_{(k-1)}$ is the residual covariance function subtracting linear prediction in terms of $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{k-1})$. Normalization with respect to principal components $\phi_1, \cdots, \phi_q$ measures relative loss of optimality:

$$O = \frac{Var(\boldsymbol{\beta}_1) + \sum_{k=2}^{q} Var(\boldsymbol{\beta}_k|\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{k-1})}{Var(\phi_1) + \sum_{k=2}^{q} Var(\phi_k|\phi_1, \cdots, \phi_{k-1})}\,.$$

This criterion has been introduced in the multivariate context and it has been shown that $O \leq 1$, equality holding if and only if the $\boldsymbol{\beta}_k$ are principal components. It is designed for penalizing systems of components which are correlated. The "price to pay" for replacing principal components with a suboptimal system which is better interpretable, for example, a system with more than one block component, can then be quantified. This criterion hence allows to compare different candidate component models. Note that this criterion is not symmetric with respect to the order of components. For structural components, we order first block components by decreasing variance, and then difference components by decreasing variance.

2.4 Problem statement

Now we can state our problem formally. We are interested in a system $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_q$ which maximizes

(1) the number of block components N_b

(2) the optimality criterion O

under the constraint that the maximum correlation $C \leq C_{max}$. Thus, if there is no system with $N_b > 1$ and $C \leq C_{max}$, then the solution to this problem is given by

principal components. Otherwise the solution is in general different from principal components.

In subsequent sections, a stepwise approach for estimating the components from a sample of curves is proposed. These are based on the covariance function estimator.

2.5 Estimation of covariance function

Estimating the covariance function γ is closely related to estimating functional principal components. Diggle and Verbyla (1998), Staniswalis and Lee (1998), Wu and Zhang (2002) and Yao et al. (2003) among many others have used kernel type smoothing estimator. Alternatively, a prespecified basis function can be used as in Silverman (1996). The differences center around the different smoothing techniques adopted.

Our approach can encompass different versions of covariance function estimators and the subtle differences would not be a major issue for the purpose of our analysis. The common mean function is estimated simply by taking the average of the data at design points without smoothing followed by subtraction from data points. The remaining curve corresponds to subject specific variation, subject to a measurement error. At this stage individual (kernel) smoothing is applied before computing the covariance function. To avoid any systematic bias caused by different smoothing parameters, the same bandwidth is used. Our estimator of the covariance function is the covariance function of the smoothed residual process.

2.6 Estimating block components

In this section, we explain how to estimate the block components, as well as their number N_b. For this, we first consider all possible partitions of $[0, T]$ into two disjoint intervals $[0, t]$ and $[t, T]$. Let $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ be the leading eigenfunctions of the covariance functions restricted on these intervals. In general, these functions are positive, otherwise, we approximate $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ to be positive by setting zero where they are negative so that we extract a basically positive relationship. We extend $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ to the whole interval $[0, T]$ by adding zero on the other interval. These functions may have discontinuity at t, which is of measure zero. For each separation point t, we evaluate the criteria C and O, obtaining functions $C(t)$ and $O(t)$. If the correlation $C(t)$ is larger than C_{max} throughout $\mathcal{T}$, the solution to our problem stated in Section 2.4 is given by the principal components. Otherwise, we consider the two block components $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ that maximize $O(t)$ under the domain where $C(t) \leq C_{max}$. Alternatively, we may choose the separation point t using specific knowledge from the field of application (some values may make more "sense" than others). We then try to split the block components obtained into two further block components using a similar algorithm. This sequential approach ends when it is no longer possible to

split the block components without surpassing the correlation threshold. In practice, the algorithm often stops with $N_b = 1$ or $N_b = 2$ block components. In these cases, the system obtained is considered the solution to our problem. If there are more than two block components, however, this sequential algorithm may miss it. In order to be sure to find the solution, one would need to use a global algorithm investigating all possible partitions of $[0, T]$ into three or more disjoint intervals.

Remark 1: In our analysis, we allow discontinuity at t for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. It is possible, though, to slightly modify the function so that it becomes continuous or differentiable. For example, to obtain a continuous function, linear interpolation at the border may be implemented. To obtain a smooth function, a constrained smoothing may be applied with boundary condition. However, because an additional refinement procedure can be arbitrary and does not influence the conclusion, we do not pursue this topic further.

Figure 5 about here.

The procedure is illustrated with the examples shown in Figure 3. Again, panels from the left column correspond to two subprocesses and those to the right to one homogeneous process. The functions $C(t)$ and $O(t)$ are seen in the top panels of Figure 5. For two subprocesses, the correlation $C(t)$ is found below the cut-off value $C_{max} = 0.3$ on a whole interval around 0, and the maximum of the optimality criterion $O(t)$ within that interval is found at $t_0 = -0.04$. This leads to two block-components (lower left panel). For one homogeneous process, the correlation function $C(t)$ stays well above the cut-off value, and we hence define only one block component (lower right panel).

2.7 Adding difference components

Once block components are defined, we may add difference components to the system. This is done in order to increase the percentage of variance extracted and to obtain further structural information. For each i, define a residual process $r_i(t)$ by subtracting its linear prediction in terms of $\boldsymbol{\beta}_k, k = 1, \cdots, N_b$ as

$$r_i(t_j) = y_i(t_j) - \mu(t_j) - \sum_{k=1}^{N_b} \theta_{ik} \boldsymbol{\beta}_k(t_j)\,.$$

Denote the covariance function of the residual process by γ_r and apply principal components analysis. Alternatively, the residual dispersion matrix subtracting its best linear prediction in terms of the $\boldsymbol{\beta}_k$ can be directly obtained from

$$\Gamma_r = \Gamma - \Gamma B (B'\Gamma B)^{-1} B'\Gamma\,,$$

where $\Gamma = \{\gamma(t_i, t_j)\}$ and $B = \{\boldsymbol{\beta}_i(t_j)\}$ are matrices evaluated at design points. See Rao (1968).

3 Properties

To establish statistical properties of the proposed procedure, we shall restrict our attention to two important situations. If there are two subprocesses, the procedure should be able to detect it by defining two block components. If there is only one process, we should come up with one block component. The following assumptions will be used.

(A2) The covariance function γ is continuous, strictly positive-definite, and the trace of γ, $\int \gamma(u, u)\, du$, is finite.

(A3) All eigenvalues λ_j have multiplicity 1, so that $\lambda_1 > \lambda_2 > \cdots > 0$.

(A4) $\sup_{u,v \in \mathcal{T}} |\hat{\gamma}(u, v) - \gamma(u, v)| \to 0$ in probability.

Assumption (A2) is standard. When some eigenvalues have multiplicity greater than 1, the corresponding eigenfunctions are not uniquely defined and thus one needs to deal with the subspace generated by the eigenvectors as in Dauxois et al. (1982) and Boente and Fraiman (2000). Because our criteria only require the leading eigenfunction for each partition, we assume that it is uniquely defined as in (A3). Our focus is not so much on the covariance function estimator as on the behaviour of criteria functions based on it, as long as it is (uniformly) consistent. For example, kernel-based smoothing estimators used in Staniswalis and Lee (1998) and Yao et al. (2005), similar to our estimator, satisfy (A4). For roughness penalty approach, see Silverman (1996) and Cardot (2000). Thus, any reasonable covariance function estimator could be incorporated in the procedure and the properties stated below would be equally applicable. Proofs are found in the Appendix.

Below we write $\hat{C}(t)$ and $\hat{O}(t)$ for the respective estimators of $C(t)$ and $O(t)$ calculated with $\hat{\gamma}$. Theorem 1 shows that these estimators are consistent. In particular, one can estimate consistently the minimum value of $C(t)$. This will be needed to show the consistency of our procedure in subsequent sections.

Theorem 1 *Assume (A2)-(A4). Then, we have*

$$\sup_t |\hat{C}(t) - C(t)| \to 0 \quad \textit{in probability},$$

$$\sup_t |\hat{O}(t) - O(t)| \to 0 \quad \textit{in probability},$$

as $n \to \infty$.

3.1 Case of two subprocesses

Consider the case of two subprocesses. Equivalently, suppose that the covariance function has an ideal partition with two blocks.

$(A1')$ For a fixed t_0, $\gamma(u,v) = 0$ for $(u-t_0)(v-t_0) < 0$. Otherwise, $\gamma(u,v) \geq 0$, where

$$0 < c_1 \leq \frac{\int_{t_0}^{T} \gamma(u,u)\, du}{\int_0^{t_0} \gamma(u,u)\, du} \leq c_2 < \infty\,,$$

for some positive constants c_1 and c_2.

It turns out that one cannot detect this cut-point t_0 based on the correlation criterion $C(t)$ alone. The reason for this is that this function does not have a unique minimum, but is equal to zero on an interval. Thus we look for the point where the optimality criterion $O(t)$ reaches maximum on that interval.

Lemma 1 *Assume* $(A1'), (A2)-(A3)$ *Then, there exists a neighborhood of* t_0, $\mathcal{N}(t_0)$ *such that*

$$C(t) = 0\,, \quad \textit{for all } t \in \mathcal{N}(t_0)\,,$$

and

$$O(t_0) \geq O(t)\,, \quad \textit{for all } t \in \mathcal{T}\,.$$

Moreover, the maximizer is unique.

Theorem 2 *Assume* $(A1'), (A2)-(A4)$. *Define for given* $\tau > 0$

$$\hat{t}_0 = \arg \max_{t:|\hat{C}(t)| \leq \tau} \hat{O}(t)\,.$$

Suppose that for large enough N, $\hat{t}_0$ *is uniquely defined for* $n \geq N$ *as* $n \to \infty$ *with probability tending to one. Then*

$$\hat{t}_0 \to t_0 \quad \textit{in probability}\ .$$

In summary, in such an ideal case as assumed in $(A1')$, our procedure will consistently define block components which correspond to the blocks in the covariance function.

3.2 Case of one process

Now we consider the case of one homogeneous process. In terms of the covariance function, this means that

$(A1'')$ $\frac{\gamma(u,v)}{(\gamma(u,u)\gamma(v,v))^{1/2}} \geq c > 0$, for all $u, v \in \mathcal{T}$.

Here we are mainly concerned with c relatively *large*, for example $c = 0.5$ or larger.

The following lemma shows that the correlation criterion $C(t)$ will in turn be larger than this threshold c. Its corollary states that this will also happen consistently in the sample.

Lemma 2 *Assume* $(A1'')$, $(A2) - (A3)$. *Then, we have*

$$C(t) \geq c\,, \quad \textit{for all } t \in \mathcal{T}\,.$$

Corollary 1 *Assume* $(A1'')$, $(A2) - (A4)$. *Then, there exists a sequence* c_n *that converges to* c *as* $n \to \infty$ *such that*

$$\hat{C}(t) \geq c_n \quad \textit{in probability}\ .$$

Therefore, in case $(A1'')$ holds, our procedure will consistently define only one block-component as soon as threshold c is larger than the pre-specified cut-off value C_{max}.

4 Numerical performance

Numerical performance was studied through simulation and application to three real data sets including growth data shown in Figure 1. Further examples are weather data and gait data from Ramsay and Silverman (1997). Motivating examples shown in Figure 2 serve as the basis of simulation studies.

4.1 Simulation studies

We compare our method (SCA) to functional PCA and its varimax rotation (implemented in Matlab) to assess differences with respect to the three criteria proposed in Section 2: number of block components (N_b), correlation(C) and optimality(O). We consider the following model

$$y_i(t_j) = \alpha_{i1}\frac{1}{\sqrt{2\pi}0.3}\exp\left(-\frac{(t_j+0.5)^2}{0.3}\right) + \alpha_{i2}\frac{1}{\sqrt{2\pi}0.3}\exp\left(-\frac{(t_j-0.5)^2}{0.3}\right)\,,$$

Table 1: Comparison of methods for approximate block structure (left) and nonblock structure (right).

	$\sigma_{12} = 0.0$				$\sigma_{12} = 0.5$			
	N_b	$N_b(10\%)$	Opt	Corr	N_b	$N_b(10\%)$	Opt	Corr
SCA	1.95	1.95	0.98	0.04	1.0	1.0	1.0	0.0
FPCA	0.97	1.41	1.0	0.0	1.0	1.0	1.0	0.0
Varimax	0	2.0	0.99	0.06	0	2.0	0.78	0.70

where

$$(\alpha_{i1}, \alpha_{i2}) \sim iid \text{ Normal } \left(0, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right).$$

We are mainly interested in the dependence of N_b on the underlying covariance structure. We set $\sigma_{11} = 0.8$, $\sigma_{22} = 0.7$ and σ_{12} between 0 and 0.5. Results below are based on 1000 simulations, where 50 random curves are observed at 101 equally spaced design points on [-1,1]. For SCA, we use $C_{max} = 0.3$.

Table 1 presents average results for $\sigma_{12} = 0.0$ and $\sigma_{12} = 0.5$. The former case corresponds to two independent subprocesses (with almost non-overlapping supports), whereas the latter case corresponds to one homogeneous process. Thus, a good solution according to our criteria should have $N_b = 2$ and $N_b = 1$, respectively. While this is mostly achieved by SCA (N_b being on average 1.95 and 1), FPCA has only one block component in either cases (N_b being on average 0.97 and 1). Strictly speaking, Varimax had no block component ($N_b = 0$). However, it often produces components which resemble block components. If we relax a bit our criterion to allow small perturbation, say 10% of squared norm, in counting the number of block components as

$$N_b(10\%) = \sum_{k=1}^{q} I\Big(min\Big\{\int_{\boldsymbol{\beta}_k>0} \boldsymbol{\beta}_k^2(t)\,dt, \int_{\boldsymbol{\beta}_k<0} \boldsymbol{\beta}_k^2(t)\,dt\Big\} < .1\Big)$$

then Varimax has 2 block components in both cases ($N_b(10\%) = 2$). Thus, neither FPCA nor Varimax could distinguish between one and two subprocesses, whereas SCA mostly does. Moreover, the average optimalty of SCA is larger than 0.98 in both cases, the average correlation between components remaining smaller than 0.04. By way of contrast, the average optimality of Varimax is only 0.78 in the one process case, the average correlation between components being as high as 0.7.

Further simulations point into the same direction; details can be obtained from the first author.

4.2 Real data examples

Now we present application to three real data sets.

Growth data: The procedure is applied to the growth data shown in Figure 1. It has been conjectured (Karlberg, 1987; Karlberg et al., 1987) that during growth there are at least two (or three if including infancy) nearly non–overlapping periods in which onset of secretion of new hormones promotes different phase of growth. This prompted the use of different parametric forms for different periods, however, phase changes are rather manually identified based on individual growth curves. Stützle et al. (1980) adopted a semiparametric approach, where two almost non-overlapping additive components are estimated by nonlinear regression via shape–invariant modeling. It is an interesting question whether our method can identify the separation between the different phases.

For 120 boys and 112 girls, measurements were taken from age 0 to age 20, half-yearly around puberty and yearly otherwise. Because of the huge variability in infancy, the analysis is restricted to age 2 to 20, for which two subprocesses could be postulated. The analysis is based on the velocity trajectories, estimated directly from the raw data using Gasser-Müller estimator. A registration step is implemented based on landmarks (Gasser et al., 1984; Kneip and Gasser, 1992). The pattern is similar for both sexes but the timing of the pubertal growth spurt, corresponding to a peak in the velocity curves, occurs earlier and is smaller for girls than for boys.

Figure 6 about here.

The top left panel of Figure 6 shows 20 samples of smoothed velocity curves for boys after registration, together with the mean function (thick line). The middle left panel shows the correlation criterion (solid line) and optimality criterion (dashed line). The curves are evaluated at the original data points and linearly interpolated. As the correlation becomes negative after age 9.9, the absolute correlation is plotted. Observe that the correlation stays low for a wide range of values, suggesting the presence of independent subprocesses. Putting the correlation threshold to $C_{max} = 0.3$, the optimality criterion is maximized at age 11.8. Thus, our procedure define a first block component between ages 2 and 11.8, and a second block component between ages 11.8 and 21, representing two roughly independent growth subprocesses. Alternatively, since optimality is almost constant in the age range for which correlation is low, one may select another age to separate subprocesses (in agreement with an expert in the field) with practically no loss in optimality.

In particular, since the correlation function has a clear minimum at age 9.9, and since the optimality criterion is almost as high at 9.9 as at 11.8, age 9.9 in this example is also a natural estimate to separate blocks. In the bottom panel are added block

components based on this separation ($t_0 = 9.9$). These two block components cannot be splitted into further blocks without surpassing the correlation threshold. Thus, we conclude that two subprocesses, but not more, are present in the data. Similar conclusions can be drawn for girls (not shown).

Weather data: Our second example is the monthly mean temperature of 35 Canadian weather stations from Ramsay and Silverman (1997) shown in the top right panel of Figure 6. The mean is estimated by kernel smoothing with a bandwidth 0.4, set by visual inspection. Unlike the growth process, the climate process is expected to be homogeneous across time and regions. The correlation criterion shown in the middle right panel confirms this hypothesis, as the function remains very high, close to 1, on the whole range. Thus, our procedure selects in this example only a single block component to indicate that no structural change occurs and that the underlying process is homogeneous. In that case, components suggested by our procedure are the same as those produced by functional PCA, and the first two of them are shown in the bottom right panel. The same conclusions hold for the daily temperature measurements.

Gait data: The procedure is now applied to gait data, concentrating on the knee. These consist of the angles formed by the hip and knee of 39 children over a gait cycle (Ramsay and Silverman 1997). This is an interesting example because it is not clear in advance whether the underlying process consists of one homogeneous process or not. As a registration step is required, landmark registration with maxima is used.

Figure 7 about here.

The smoothed curves of registered data are shown in the upper left panel of Figure 7, with the mean curve in thick line. The corresponding correlation and optimality criteria are shown in the upper right panel. In contrast to the first two examples, the correlation function is neither negligible, nor very high, but is around 0.5 on a large domain. When applied with a cut-off value of $C_{max} = 0.3$, our procedure selects a single block component, as is shown in the bottom right panel. However, using a more liberal value of C_{max}, a solution with two block components is conceivable for which optimality is still above 80%, compared to functional principal components. The two block components obtained are shown in the lower left panel, which separate the earlier preparation movement from the later major movement. It would be interesting to discuss with specialists of the field the merits of the two solutions.

5 Appendix

Proof of Theorem 1 We use the following lemma.

Lemma 3 *(p.147, Rudin (1976)) The sequence of functions $\{f_n\}$ defined on E, a subset of metric space, converges uniformly on E if and only if for every $\epsilon > 0$, there exists an integer N such that $m \geq N, n \geq N$ implies*

$$|f_n(t) - f_m(t)| < \epsilon$$

for each $t \in E$.

We first introduce some notations. Define the bounded linear operator associated with the covariance function $\mathbf{\Gamma} : L^2[0,T] \to L^2[0,T]$ as

$$(\mathbf{\Gamma} f)(u) = \int_0^T \gamma(u,v) f(v)\, dv \qquad \text{for all } f \in L^2([0,T])\,,$$

with norm $||\mathbf{\Gamma}||_{\mathcal{L}} = \sup_{||f||\leq 1} ||\mathbf{\Gamma} f||$, where $||\cdot||$ is the usual norm in the space $L^2[0,T]$, distinguished from $||\cdot||_2$ for the norm in $L^2([0,T]\times[0,T])$. Similarly, the empirical covariance operator $\hat{\mathbf{\Gamma}}$ will be defined through the estimated covariance function $\hat{\gamma}$. Then it holds that

$$||\mathbf{\Gamma}||_{\mathcal{L}} \leq ||\gamma||_2\,. \tag{1}$$

Note that, instead of $\hat{C}$, we write C_n for the estimator of C based on n observations. For each t, denote by $\boldsymbol{\alpha}_{n,t}$ and $\boldsymbol{\beta}_{n,t}$ the extended leading eigenfunctions for each partial covariance function, which are defined on the whole interval. Then C_n can be written as

$$C_n(t) = \frac{(\boldsymbol{\alpha}_{n,t}, \mathbf{\Gamma}_n \boldsymbol{\beta}_{n,t})}{\sqrt{(\boldsymbol{\alpha}_{n,t}, \mathbf{\Gamma}_n \boldsymbol{\alpha}_{n,t})(\boldsymbol{\beta}_{n,t}, \mathbf{\Gamma}_n \boldsymbol{\beta}_{n,t})}}\,. \tag{2}$$

In view of Lemma 3, we will study the behaviour of $|C_n(t) - C_m(t)|$ for each t and suppress the dependence on t in $\boldsymbol{\alpha}, \boldsymbol{\alpha}_n, \boldsymbol{\beta}$ and $\boldsymbol{\beta}_n$ from now on. First observe that

$$\begin{aligned}
&|(\boldsymbol{\alpha}_n, \mathbf{\Gamma}_n \boldsymbol{\beta}_n) - (\boldsymbol{\alpha}_m, \mathbf{\Gamma}_m \boldsymbol{\beta}_m)| \\
&\quad\leq\ |(\boldsymbol{\alpha}_n, (\mathbf{\Gamma}_n - \mathbf{\Gamma}_m)\boldsymbol{\beta}_n)| + |(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_m, \mathbf{\Gamma}_m \boldsymbol{\beta}_n)| + |(\boldsymbol{\alpha}_m, \mathbf{\Gamma}_m(\boldsymbol{\beta}_n - \boldsymbol{\beta}_m))| \\
&\quad\leq\ ||\boldsymbol{\alpha}_n||||(\mathbf{\Gamma}_n - \mathbf{\Gamma}_m)\boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_m||||\mathbf{\Gamma}_m \boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_m||||\mathbf{\Gamma}_m(\boldsymbol{\beta}_n - \boldsymbol{\beta}_m)|| \\
&\quad\leq\ ||\boldsymbol{\alpha}_n||||\mathbf{\Gamma}_n - \mathbf{\Gamma}_m||_{\mathcal{L}}||\boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_m||||\mathbf{\Gamma}_m||_{\mathcal{L}}||\boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_m||||\mathbf{\Gamma}_m||_{\mathcal{L}}||\boldsymbol{\beta}_n - \boldsymbol{\beta}_m|| \\
&\quad\leq\ ||\boldsymbol{\alpha}_n||||\gamma_n - \gamma_m||_2||\boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_m||||\gamma_m||_2||\boldsymbol{\beta}_n|| + ||\boldsymbol{\alpha}_m||||\gamma_m||_2||\boldsymbol{\beta}_n - \boldsymbol{\beta}_m||\,,
\end{aligned}$$

where the last inequality follows from (1). Note that (A4) implies that $||\gamma_n - \gamma||_2 \to 0$ in probability. Thus, by Lemma 3, it is enough to show that $||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}||$ and $||\boldsymbol{\beta}_n - \boldsymbol{\beta}||$ converge. In fact, the uniform convergence of the covariance function implies the uniform convergence of the corresponding eigenfunctions. According to lemma 3.1 of Bosq (1991), the jth eigenfunction estimator of ϕ_j satisfies

$$||\hat{\phi}_j - \phi_j|| \leq a_j ||\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}||_{\mathcal{L}}\,,$$

where

$$\begin{aligned} a_1 &= 2\sqrt{2}(\lambda_1 - \lambda_2)^{-1} \\ a_j &= 2\sqrt{2}\max\{(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1}\} \quad \text{if } j \geq 2\,. \end{aligned}$$

(Remark: The original theorem is proved under non-smoothed empirical covariance operator and its extension to a kernel-smoothed covariance operator is referred to Boente and Fraiman (2000).) From (1), we may write it as

$$||\hat{\phi}_j - \phi_j|| \leq a_j||\hat{\gamma} - \gamma||_2\,.$$

This property still can be applied to our situation where for each t, $\boldsymbol{\alpha}_n$ is not an eigenfunction of $\boldsymbol{\Gamma}_n$ but an eigenfunction of $\boldsymbol{\Gamma}_n$ restricted to $[0, t]$, for if the full covariance function converges uniformly, so does its restriction on the subinterval. Write $\gamma_{n,t}$ for the corresponding covariance function associated with $\boldsymbol{\alpha}_n$ and $\gamma_{n,-t}$ for the covariance function associated with $\boldsymbol{\beta}_n$.

$$\begin{aligned} ||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}|| &\leq a_{1,t}||\gamma_{n,t} - \gamma_t||_2 = a_{1,t}\sqrt{\int_0^t\int_0^t |\gamma_n(u,v) - \gamma(u,v)|^2\,du\,dv} \\ &\leq a_{1,t}\sqrt{\int_0^T\int_0^T |\gamma_n(u,v) - \gamma(u,v)|^2\,du\,dv} \\ &= a_{1,t}||\gamma_n - \gamma||_2\,, \end{aligned}$$

where $a_{1,t}$ is the constant a_1 calculated from $\gamma_{n,t}$. Similar result can be derived for $||\boldsymbol{\beta}_n - \boldsymbol{\beta}||$ with an appropriate constant $b_{1,-t}$. This leads to the convergence of $(\boldsymbol{\alpha}_n, \boldsymbol{\Gamma}_n\boldsymbol{\beta}_n)$, in particular, that of $(\boldsymbol{\alpha}_n, \boldsymbol{\Gamma}_n\boldsymbol{\alpha}_n)$ and $(\boldsymbol{\beta}_n, \boldsymbol{\Gamma}_n\boldsymbol{\beta}_n)$. It follows from (2) and Lemma 3 that C_n converges uniformly. On the other hand, it can be seen that $C_n(t)$ converges to $C(t)$ for each t. Therefore, we conclude that C_n converges to C uniformly. Because O is also a functional of the covariance function and its eigenfunctions, similar argument applies to O.

Proof of Lemma 1 For a given t, the leading eigenfunctions for each partition satisfy:

$$\begin{aligned} (\Gamma_t\boldsymbol{\alpha})(u) &= \int_0^t \gamma(u,v)\boldsymbol{\alpha}(v)\,dv = \lambda_1(t)\boldsymbol{\alpha}(u)\,, \\ (\Gamma_t\boldsymbol{\beta})(u) &= \int_t^T \gamma(u,v)\boldsymbol{\beta}(v)\,dv = \lambda_2(t)\boldsymbol{\beta}(u)\,, \end{aligned}$$

Here, for simplicity of notation, we suppress the dependence on t in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Suppose that $t \leq t_0$ and consider $\boldsymbol{\beta}$. Because of condition $(A1')$, only one of the following equations will be used to produce the leading eigenfunction $\boldsymbol{\beta}$.

$$\begin{aligned}\int_t^{t_0} \gamma(u,v)\boldsymbol{\beta}_1(v)\,dv &= \lambda_{2,1}(t)\boldsymbol{\beta}_1(u) \qquad t \leq u \leq t_0 \\ \int_{t_0}^{T} \gamma(u,v)\boldsymbol{\beta}_2(v)\,dv &= \lambda_{2,2}(t)\boldsymbol{\beta}_2(u) \qquad t_0 \leq u \leq T\end{aligned}$$

Observe that $\lambda_{2,2}(t)$ does not depend on t, while

$$\begin{aligned}\lambda_{2,1}(t) &= \int_t^{t_0}\int_t^{t_0} \boldsymbol{\beta}_1(u)\gamma(u,v)\boldsymbol{\beta}_1(v)\,du\,dv \\ &\leq \sqrt{\int_t^{t_0}\int_t^{t_0} \gamma(u,v)^2\,du\,dv} \\ &\leq \int_t^{t_0} \gamma(u,u)\,du \leq \sup_{u\in[t,t_0]} |\gamma(u,u)||t-t_0|\,.\end{aligned}$$

As a consequence, $\lambda_{2,1}(t) \to 0$ and $\lambda_{2,2}(t) = \lambda_{2,2}(t_0)$ as $t \to t_0$. So, there exist a neighborhood $\mathcal{N}(t_0)$ such that for all $t \in \mathcal{N}(t_0), \boldsymbol{\beta}(u) = \boldsymbol{\beta}_2(u), t_0 \leq u \leq T$ and 0 otherwise. Therefore,

$$\begin{aligned}&\int_0^T\int_0^T \boldsymbol{\alpha}(u)\gamma(u,v)\boldsymbol{\beta}(v)\,du\,dv \\ &= \int_0^{t_0}\int_0^T \boldsymbol{\alpha}(u)\gamma(u,v)\boldsymbol{\beta}(v)\,du\,dv + \int_{t_0}^T\int_0^T \boldsymbol{\alpha}(u)\gamma(u,v)\boldsymbol{\beta}(v)\,du\,dv \\ &= \int_0^{t_0}\int_0^t \boldsymbol{\alpha}(u)\gamma(u,v)\boldsymbol{\beta}(v)\,du\,dv + \int_{t_0}^T\int_0^t \boldsymbol{\alpha}(u)\gamma(u,v)\boldsymbol{\beta}(v)\,du\,dv\,,\end{aligned}$$

where the first term is zero because $\boldsymbol{\beta}(u) = 0$ for $0 \leq u \leq t_0$ and the second term is zero because of $(A1')$. Symmetric arguments apply to $\boldsymbol{\alpha}$ when $t > t_0$. Therefore, this implies that $C(t) = 0$ for all $t \in \mathcal{N}(t_0)$. Regarding O, note that under the assumption $(A1')$, the two block components separated at t_0 are indeed eigenfunctions of the covariance function γ, thus the weight functions of principal components. Because O measures the sum of variability and the principal components uniquely maximize the sum of variability, it follows that $O(t_0)$ is the unique maximum.

Proof of Theorem 2 We prove that the negation of the claim contradicts to the argument below. To make the dependence on n of the estimator clear, write the

nth criteria functions as C_n and O_n. All arguments below hold in probability without making explicit references. Observe that t_0 is uniquely defined for two block structure $(A1')$. Then for every $\delta > 0$, there exists some $\epsilon = \epsilon(\delta) > 0$ such that

$$O(t) + \epsilon < O(t_0) - \epsilon\,, \quad \text{if } |t - t_0| > \delta\,.$$

Because of uniform convergence of O_n, we have for every $\epsilon_1 > 0$ and for all t, there exists $n_0 = n_0(\epsilon_1) \geq N$ such that for all $n \geq n_0$,

$$|O_n(t) - O(t)| < \epsilon\,. \tag{3}$$

Then, for every $\delta > 0$, there exists $n_0 = n_0(\epsilon(\delta)) \geq N$ such that for all $n \geq n_0$,

$$O_n(t) < O(t_0) - \epsilon(\delta)\,, \quad \text{if } |t - t_0| > \delta\,. \tag{4}$$

This will contradict to the negation of the argument. To see why, suppose that the claim is false. Define for n

$$t_{0,n} = \arg\max_t O_n(t)\,.$$

Then for some $\delta_0 > 0$, there exists $n \geq \tilde{n}, n \geq N$ for all $\tilde{n}$ such that

$$|t_{0,n} - t_0| > \delta_0\,, \quad \text{and } O_n(t_{0,n}) \geq O_n(t_0)\,.$$

Given $\epsilon_0 = \epsilon(\delta_0)$, choose $\tilde{n} = n_0(\epsilon_0)$ that satisfies (3). Then, there exists $n \geq n_0 \geq N$ such that

$$O_n(t_{0,n}) > O(t_0) - \epsilon(\delta_0)\,, \quad \text{if } |t_{0,n} - t_0| > \delta_0\,,$$

which is contradictory to (4).

Now suppose that C_n is also used and $t_{0,n}$ is defined as

$$t_{0,n} = \arg\max_{t:|C_n(t)|\leq\tau} O_n(t)\,,$$

for some $\tau > 0$. Note that $C(t_0) = 0$. Then given τ and for all $t \in \mathcal{N}(t_0)$, there exists n_1 such that for all $n \geq n_1 \geq N$,

$$|C_n(t)| \leq \tau\,.$$

If we replace ϵ by $min(\epsilon, \tau)$ and n_0 by $max(n_0, n_1)$ in the above argument, the same holds true.

Proof of Lemma 2 It follows from $(A1'')$ combined with Cauchy-Schwarz inquality.

Proof of Corollary 1 Because $\sup_t |\hat{C}(t) - C(t)| \to 0$ in probability, with probability tending to 1, for every ϵ and for every t, there exists $n_0 \geq N$ such that for all $n \geq n_0$,

$$|C_n(t) - C(t)| < \epsilon \,.$$

Because $C(t) \geq c$, this implies that

$$C_n(t) \geq c - \epsilon \,.$$

For $\epsilon_n \to 0$, take $c_n = c - \epsilon_n$.

References

Boente, G. and R. Fraiman (2000). Kernel-based functional principal components. *Statistics & Probability Letters 48*, 335–345.

Bosq, D. (1991). Modelization, nonparametric estimation and prediction for continuous time processes. In G. Roussas (Ed.), *Nonparametric Function Estimation and Related Topics*, Nato Asi series, pp. 509–529. Klewer Academic: Dordrecht.

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics 12*, 503–538.

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of multivariate analysis 12*, 136–154.

Diggle, P. J. and A. P. Verbyla (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics 54*, 401–415.

Gasser, T., H. G. Müller, W. Köhler, L. Molinari, and A. Prader (1984). Nonparametric regression analysis of growth curves. *Ann. Statist. 12*, 210–229.

Gervini, D. and V. Rousson (2004). Criteria for evaluating dimension-reducing components for multivariate data. *The American Statistician 58*, 72–76.

Karlberg, J. (1987). On the modeling of human growth. *Statist. Med. 6*, 185–192.

Karlberg, J., J. G. Fryer, I. Engström, and P. Karlberg (1987). Analysis of linear growth using a mathematical model ii. from 3 to 21 years of age. *Acta Paediatrica Scandinavica 337*, 12–29.

Kneip, A. and T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist. 20*, 1266–1305.

Ramsay, J. and B. Silverman (1997). *Functional Data Analysis*. Springer–Verlag: New York.

Ramsay, J. and B. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer: New York.

Rao, C. R. (1968). *Linear Statistical Inference and its Applications*. Wiley: New York.

Rousson, V. and T. Gasser (2004). Simple component analysis. *Appl. Statist. 53*, 539–555.

Rudin, W. (1976). *Principles of Mathematical Analysis* (3rd ed.). McGraw–Hill.

Silverman, B. W. (1996). Smoothed functional principal component analysis by choice of norm. *Ann. Statist. 24*, 1–24.

Staniswalis, J. and J. Lee (1998). Nonparametric regression analysis of longitudinal data. *J. Am. Statist. Ass. 93*, 1403–1418.

Stützle, W., T. Gasser, L. Molinari, R. H. Largo, A. Prader, and P. J. Huber (1980). Shape-invariant modeling of human growth. *Annals of Human Biology 7*, 507–528.

Wu, H. and J. Zhang (2002). Local polynomial mixed-effects models for longitudinal data. *J. Am. Statist. Ass. 97*(459), 883–897.

Yao, F., H. G. Müller, A. J. Clifford, S. R. Dueker, J. Follett, Y. Lin, B. A. Buchholz, and J. S. Vogel (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics 59*, 676–685.

Yao, F., H. G. Müller, and J. L. Wang (2005). Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass. 100*, 577–590.

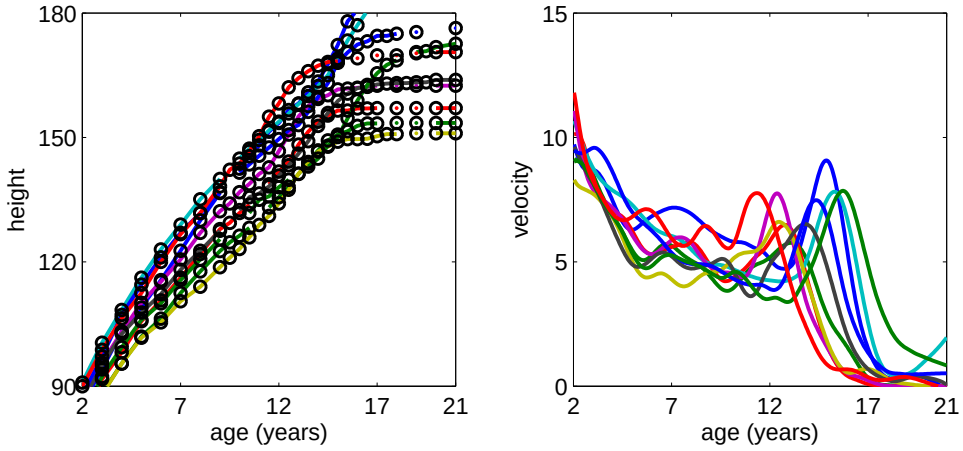

Figure 1: Height growth curves (left) and velocity curves (right) for 10 boys. Velocity curves are estimated nonparametrically with Gasser-Müller estimator. These curves can be aligned to eliminate time variability.

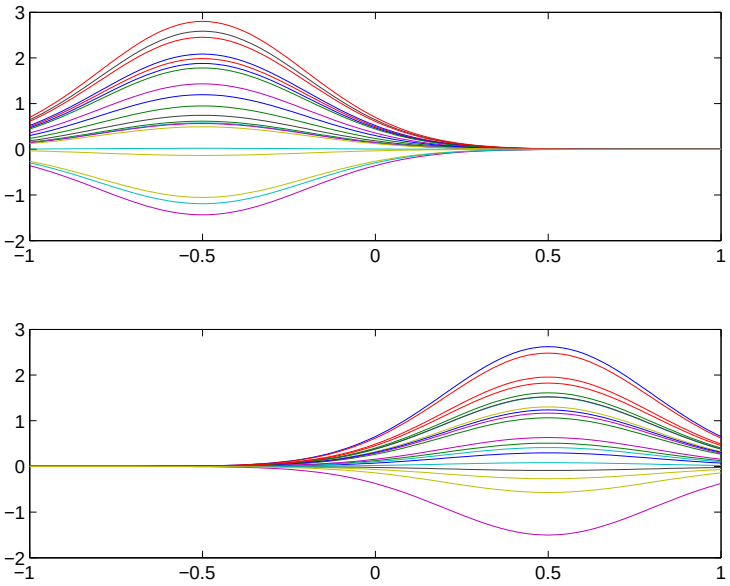

Figure 2: Example of curves that have almost non–overlapping support. Data are constructed by adding up these two curves.

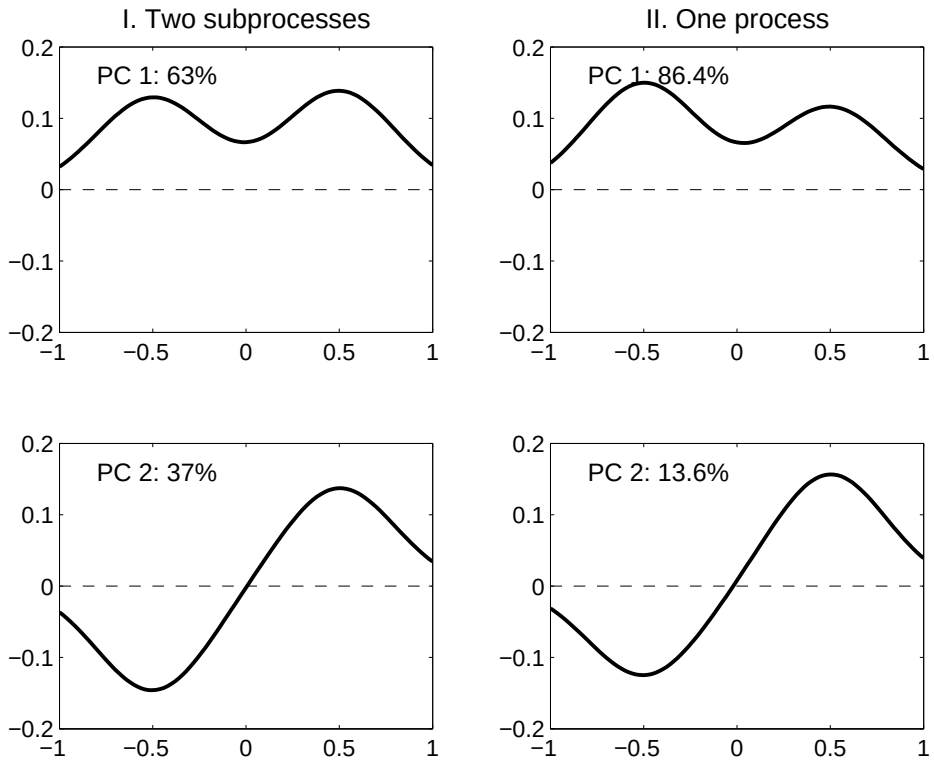

Figure 3: Functional PCA for two subprocesses (left, shown in Figure 2) and one process (right). Both provide *qualitatively* the same answer.

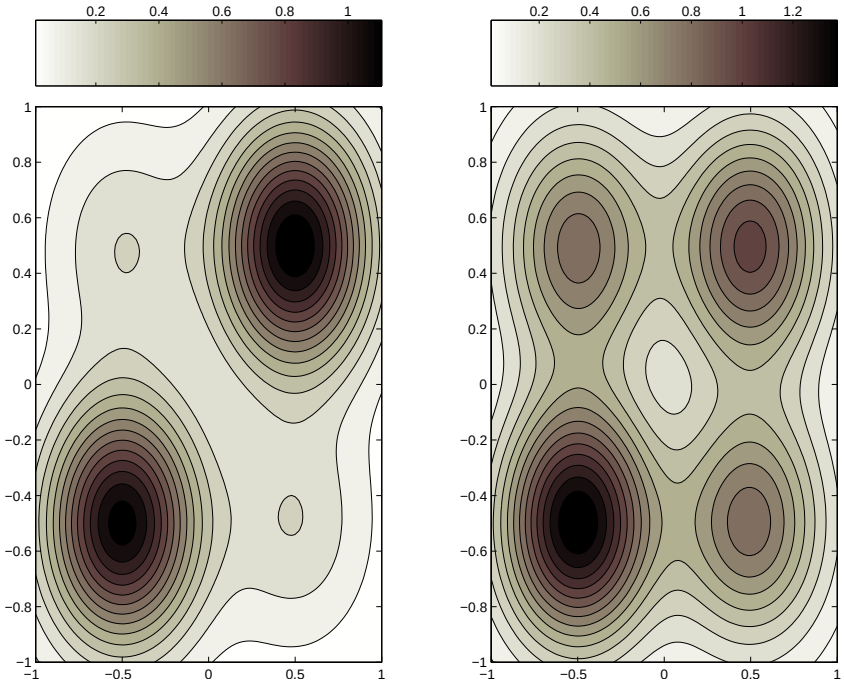

Figure 4: Contour plots of covariance (top) and correlation (bottom) function. Distinct structures between two subprocesses (left) and one process (right) are visualized. The higher, the darker. Ranges in numbers are added in the bottom of each plot.

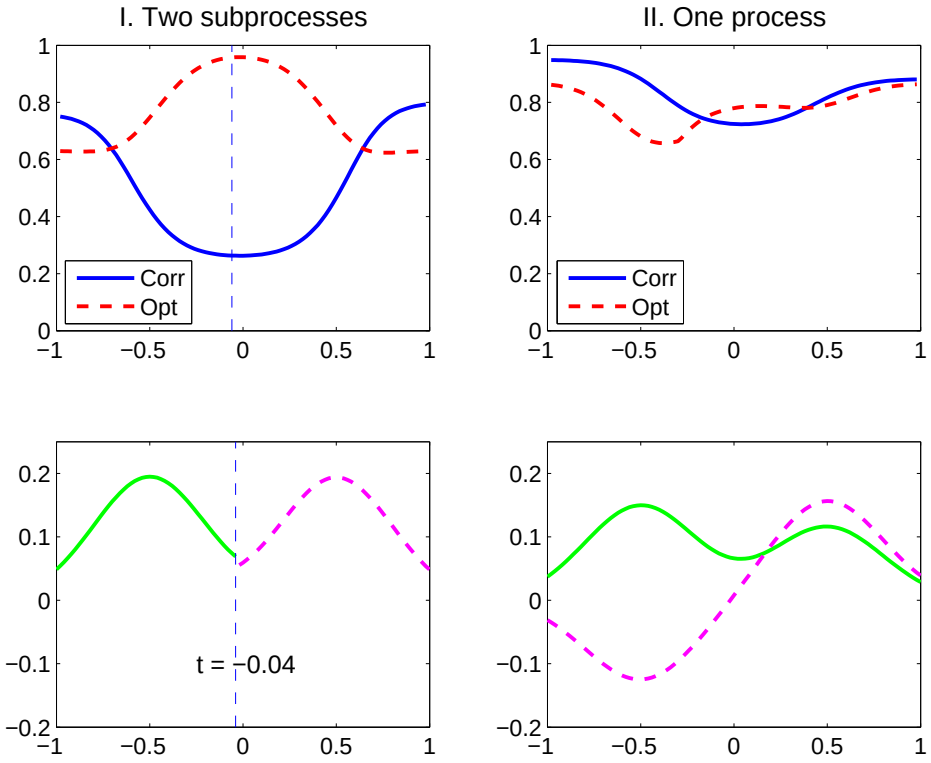

Figure 5: Top row shows diagnostic plots for two subprocesses (left) and one process (right) cases, shown in Figure 4. For the left are suggested two block components, separated at $t = -.04$ and one block component for the right. Bottom row shows selected first two components.

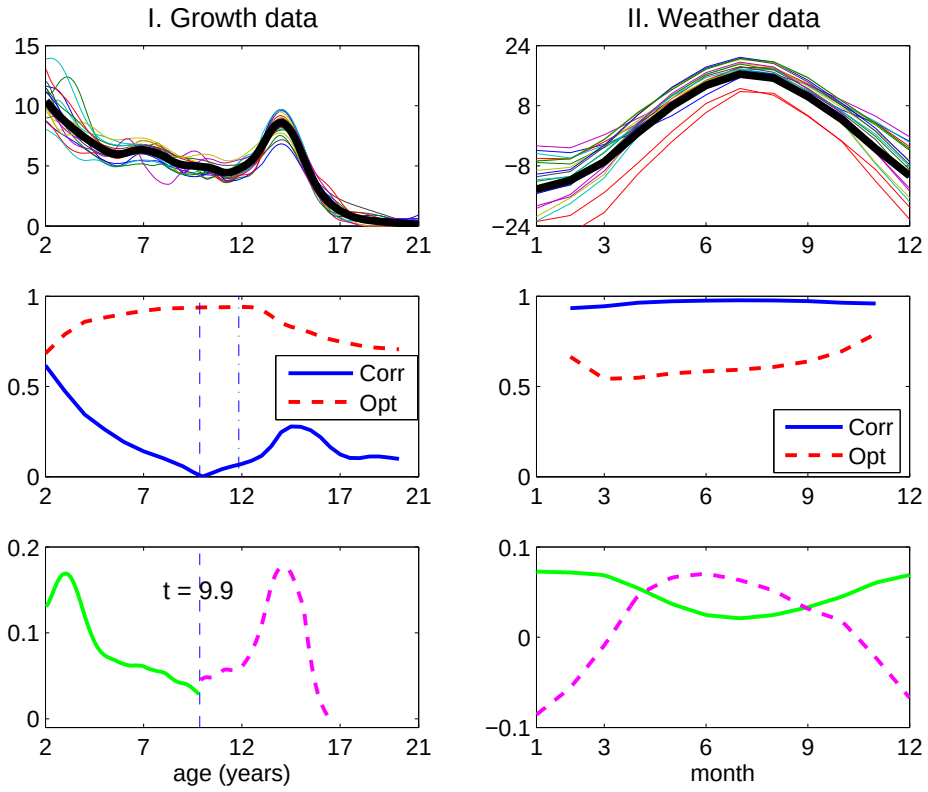

Figure 6: Analysis of structural components for growth curves (left) and weather data (right). Criteria curves are shown in the middle panel, indicating two subprocesses in the left and one process in the right. Suggested two components are drawn in the bottom.

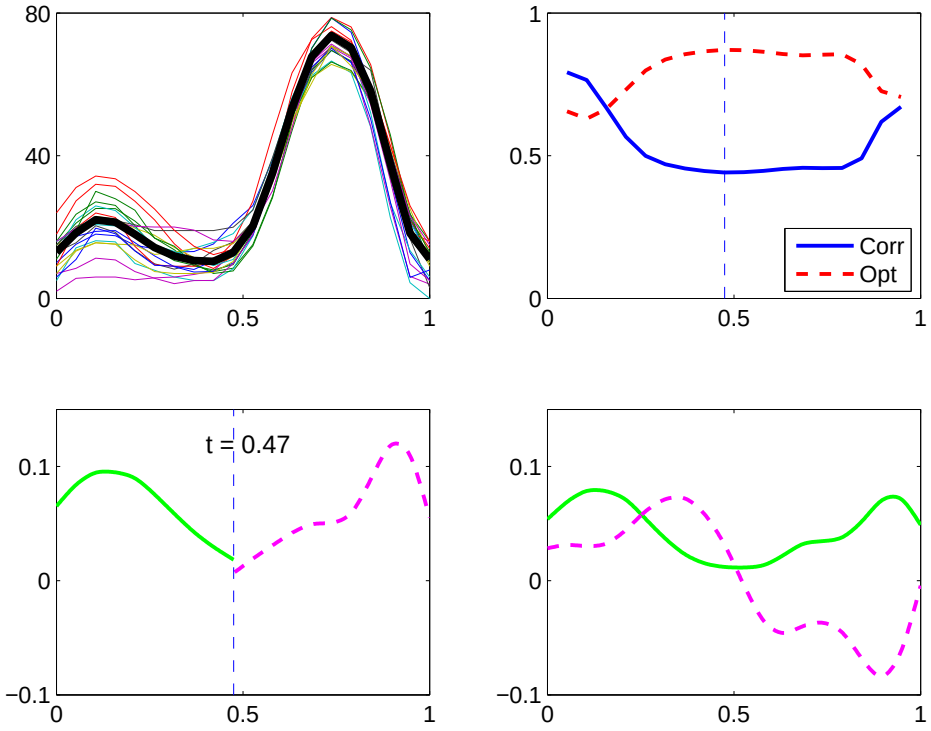

Figure 7: Registered gait data (knee) in the upper left panel with criteria curves in the right. Vertical line indicates possible separation of block components with relatively high correlation (0.5). Two block components are shown in lower left, compared to two principal components in lower right panel. Block components are approximately 85% optimal against principal components, as is indicated by dashed curve in top right panel.