M7777 Applied Functional Data Analysis
6. Exploratory Data Analysis

Jan Koláček (kolacek@math.muni.cz)

Dept. of Mathematics and Statistics, Faculty of Science, Masaryk University, Brno

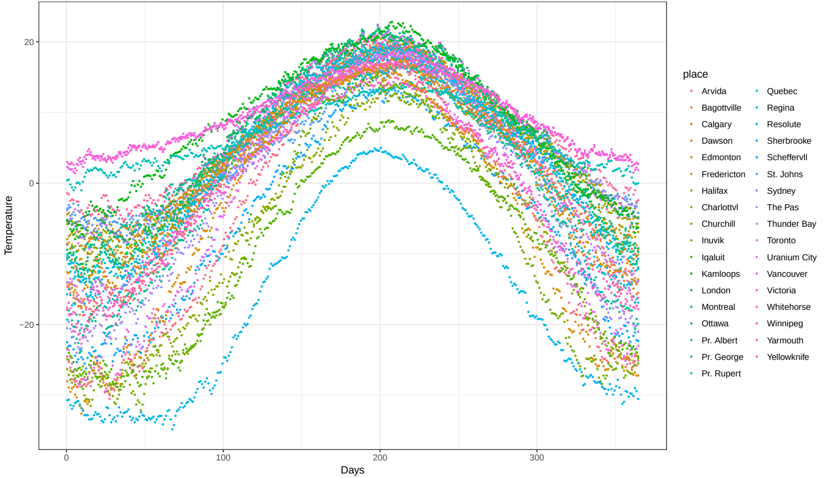
Exploratory Data Analysis

Summary Statistics

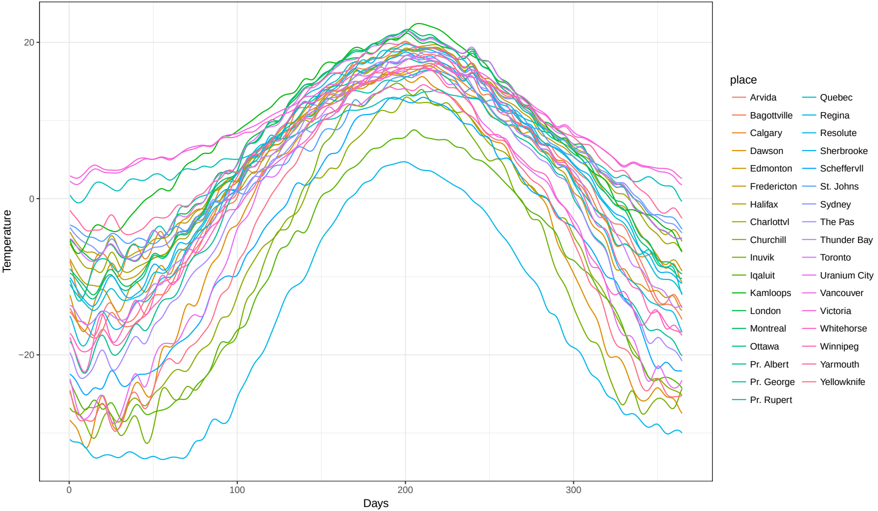
- $X(t)\ldots$ functional random variable $\Rightarrow$ collection of curves $x_1(t), \ldots, x_n(t)$.
- **Expected Value** $\mu(t) = \mathrm{E}X(t)$, **mean** $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$
- **Covariance** $\sigma(s,t) = \mathrm{E}[X(s) - \mu(s)][X(t) - \mu(t)]$, estimate $\hat{\sigma}(s,t) = \frac{1}{n}\sum_{i=1}^{n} [x_i(s) - \bar{x}(s)]\,[x_i(t) - \bar{x}(t)]$
- **Variance** $\mathrm{Var}X(t) = \sigma(t,t)$
- **Correlation** $\rho(s,t) = \frac{\sigma(s,t)}{\sqrt{\sigma(s,s)}\sqrt{\sigma(t,t)}}$

Canadian Weather

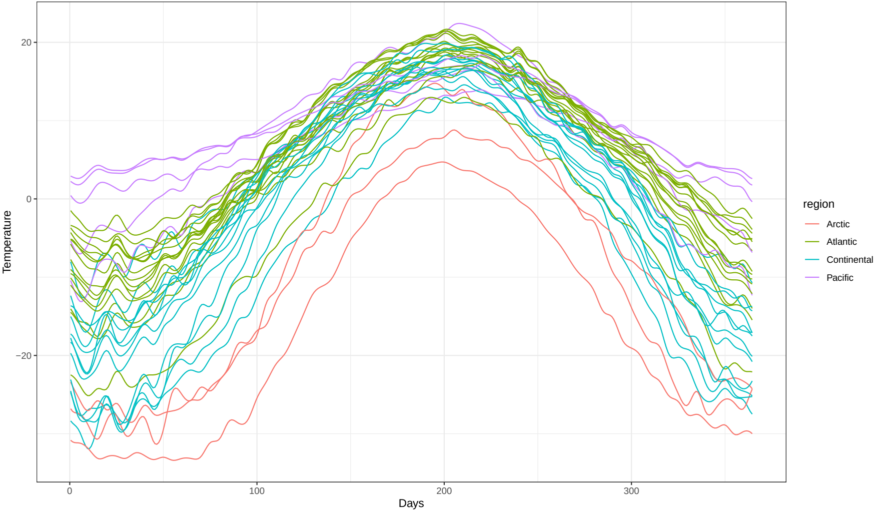
- Daily temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994



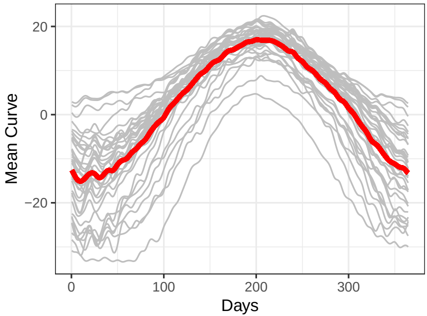
Canadian Weather – smoothed



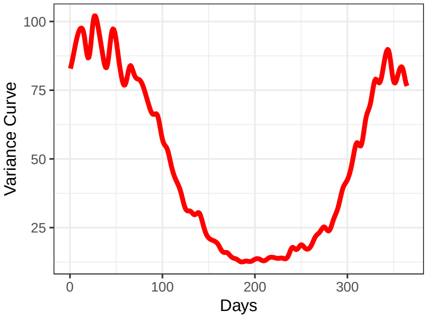
Canadian Weather – smoothed, regions



Exploratory Data Analysis

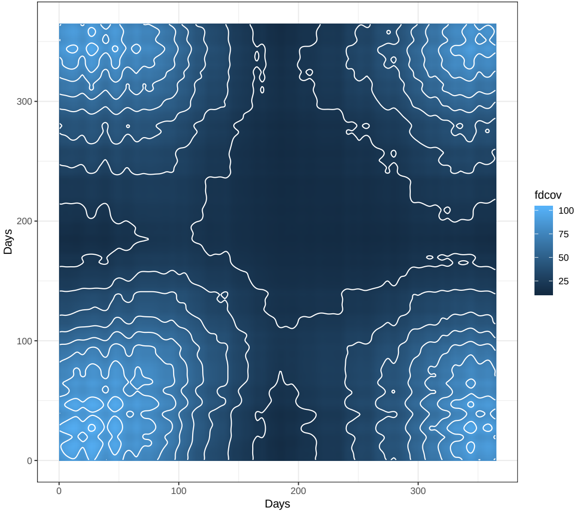


Mean



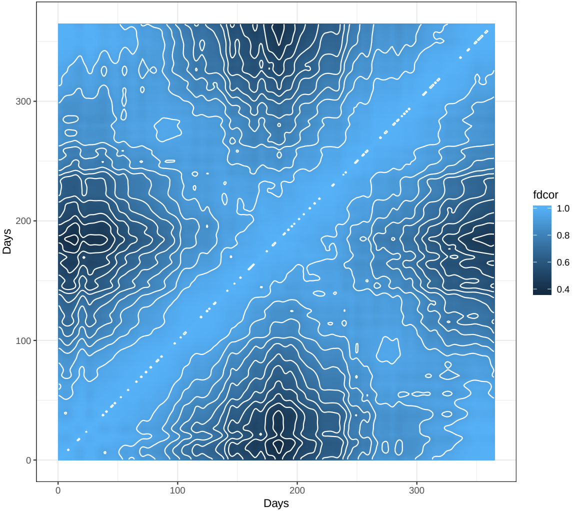
Variance

Exploratory Data Analysis



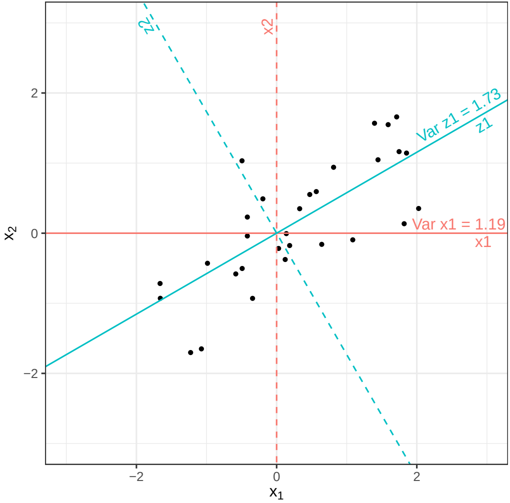
Covariance

Exploratory Data Analysis



Correlation

Multivariate Principal Component Analysis



Directions of greatest variation

Multivariate **P**rincipal **C**omponent **A**nalysis

- Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n),\ \mathbf{x}_i \in \mathbb{R}^d$
- Measure total variation in the data as total squared distance from center
$$\sum_{j=1}^{d} \sum_{i=1}^{n} (x_{ij} - \bar{\mathbf{x}}_j)^2 = \mathrm{tr}\boldsymbol{\Sigma}.$$
- Find $\mathbf{u} \in \mathbb{R}^d,\ ||\mathbf{u}||^2 = \mathbf{u}'\mathbf{u} = 1$ to maximize variance of $\mathbf{u}'\mathbf{X}$
- If $\mathbf{X}$ has a covariance $\boldsymbol{\Sigma}$, the variance of $\mathbf{u}'\mathbf{X}$ is $\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$
- Maximizing $\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}$ with respect to $\mathbf{u}'\mathbf{u} = 1$ tends to solving the eigen-equation
$$\boldsymbol{\Sigma}\mathbf{u} = \lambda\mathbf{u}.$$

Algorithm of PCA

- Estimate covariance matrix

$$\Sigma_{ij} = (\mathbf{x}_i - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}).$$

- Take the eigen-decomposition of $\boldsymbol{\Sigma}$

$$\boldsymbol{\Sigma} = \mathbf{U}'\mathbf{D}\mathbf{U}$$

- Columns of $\mathbf{U}$ are orthonormal; represent a **new basis**.
- $\mathbf{D} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$ is diagonal; entries give variances of data along corresponding directions $\mathbf{U}$.
  $\lambda_j / \sum \lambda_i$ ... "proportion of variance explained".
- Order $\mathbf{D}$, $\mathbf{U}$ in terms of decreasing d_i.
- From original data, $\mathbf{x}_i$, $(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{u}_j$ is the jth principal component score; jth co-ordinate of $\mathbf{x}_i$ in new basis.

Functional Principal Component Analysis – **FPCA**

- Instead of covariance matrix $\mathbf{\Sigma}$, we have a surface $\sigma(s,t)$
- Re-interpret the eigen-decomposition

$$\mathbf{\Sigma}_{ij} = \left(\mathbf{U}'\mathbf{D}\mathbf{U}\right)_{ij} = \sum_{k=1}^{d} \lambda_k \mathbf{x}_i \mathbf{x}_j'$$

- Karhunen – Loève decomposition for functions

$$\sigma(s,t) = \sum_{j=1}^{\infty} \lambda_j \xi_j(s)\xi_j(t),$$

  with $\int \xi_i(t)\xi_j(t)dt = I(i=j)$ (orthonormality).
- The λ_i represents amount of variation in direction $\xi_i(t)$.

Exploratory Data Analysis

- The $\xi_j(t)$ are the **principal components**; successively maximize

$$\lambda_j = \mathrm{Var} \int \xi_j(t)[X(t) - \mu(t)]dt$$

- $\lambda_j / \sum \lambda_i$ ... proportion of variance explained
- Principal component scores are

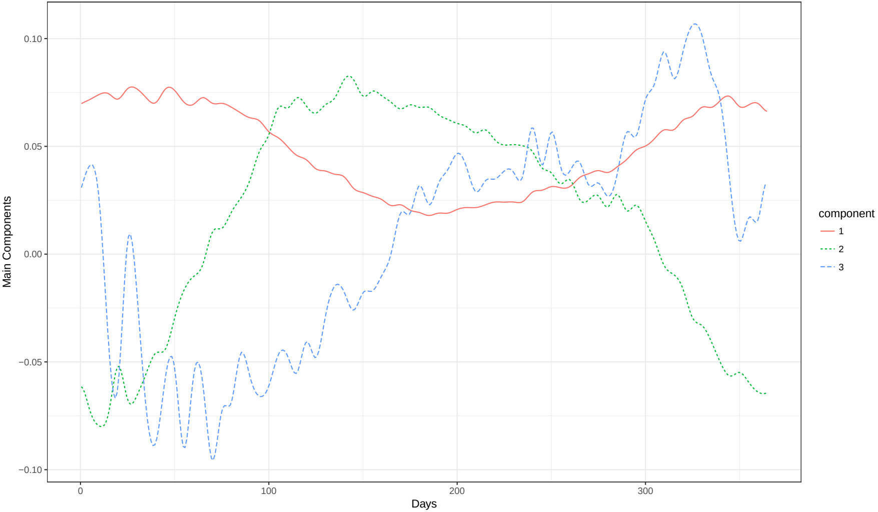
$$c_{ij} = \int \xi_j(t)[x_i(t) - \bar{x}(t)]dt$$

- Backward reconstruction

$$\hat{x}_i(t) = \bar{x}(t) + \sum_{j=1}^{K} c_{ij}\xi_j(t)$$

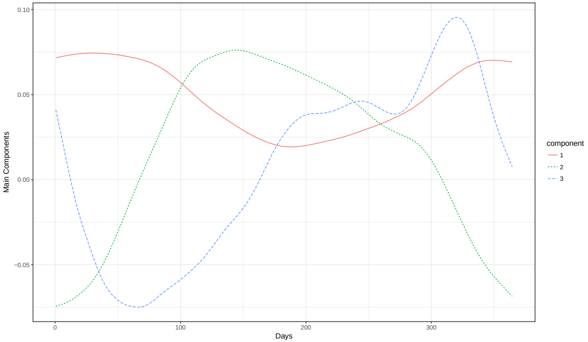
Exploratory Data Analysis

Canadian Temperature Data



First 3 principal components.

Canadian Temperature Data



First 3 principal components – smoothed.

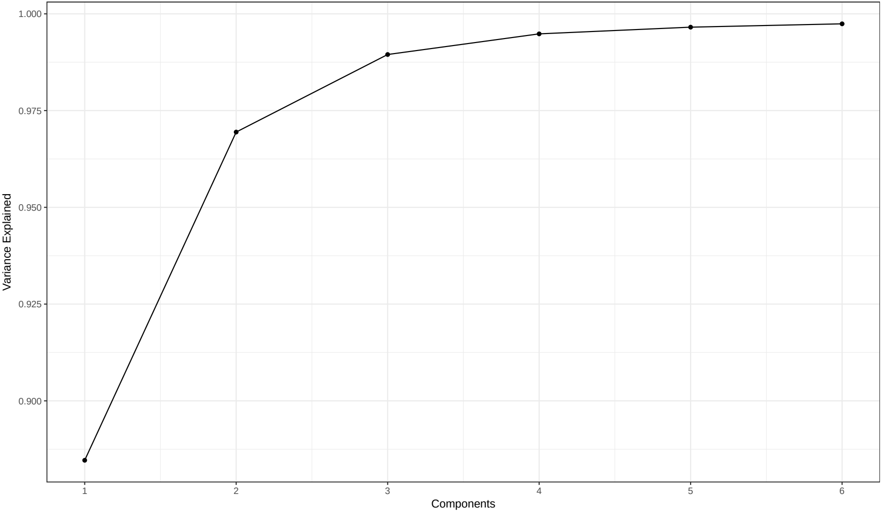
Interpretation:

PC 1 over-all temperature

PC 2 Summer vs Winter

PC 3 Spring vs Fall

Canadian Temperature Data

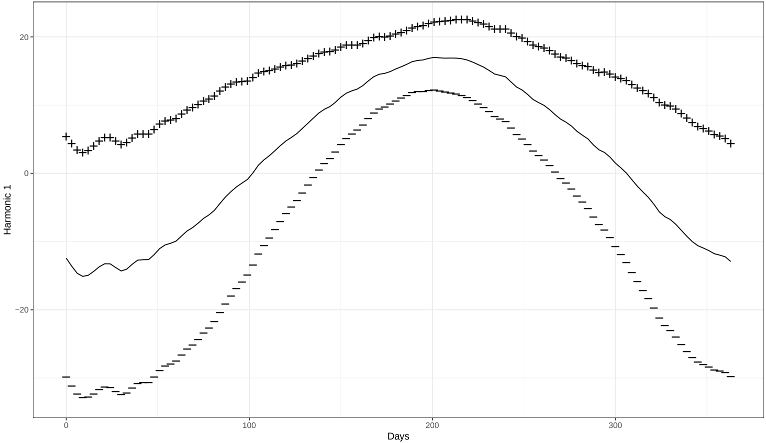


Cumulative variance explained.

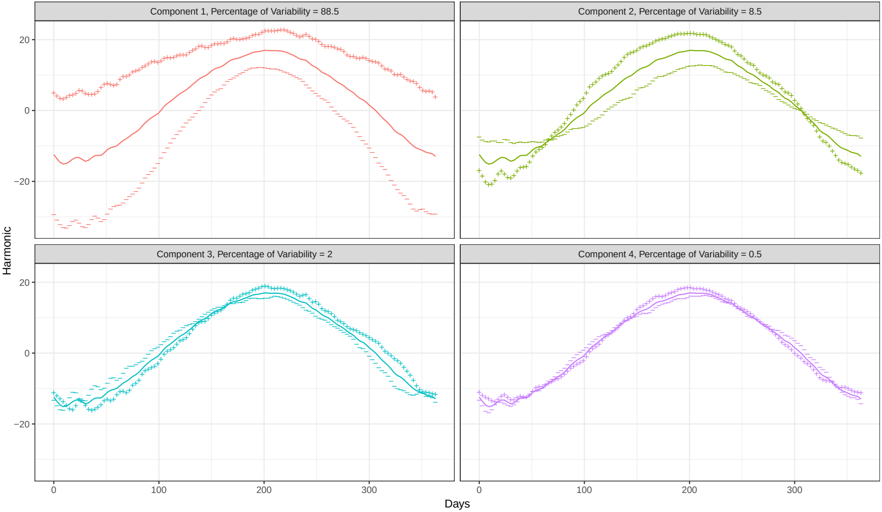
Display of Principal Components

Best way to obtain an idea of variation for each component is to plot

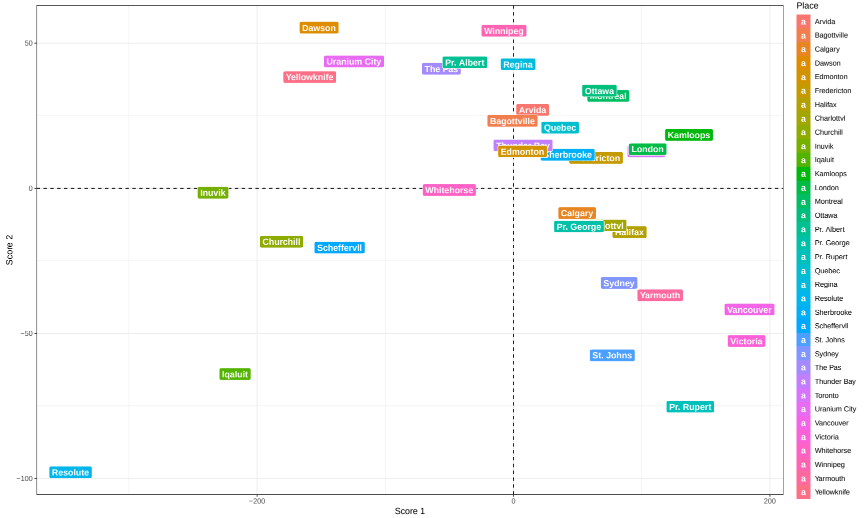
$$\bar{x}(t) \pm 2\sqrt{\lambda_i}\xi_i(t)$$



Display of Principal Components



Display of Scores



Display of Scores – regions

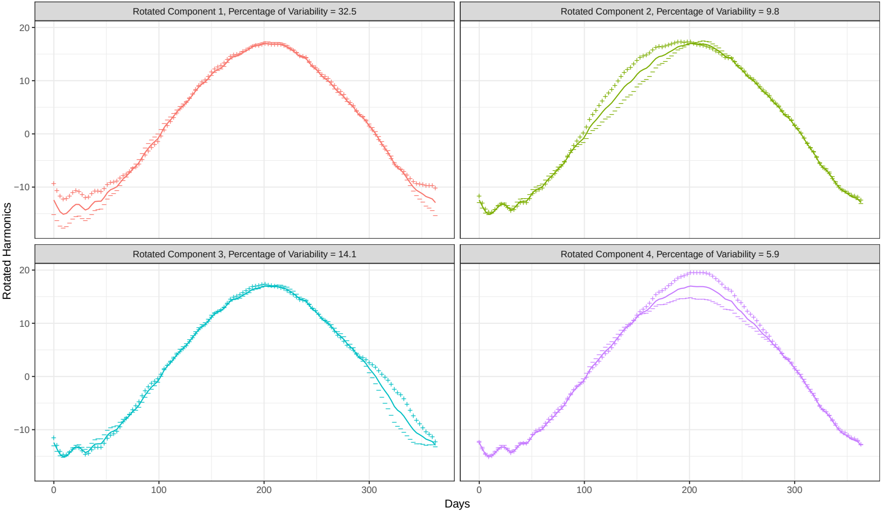
Rotating Principal Components

- find new orthonormal components by rotating the original PC's

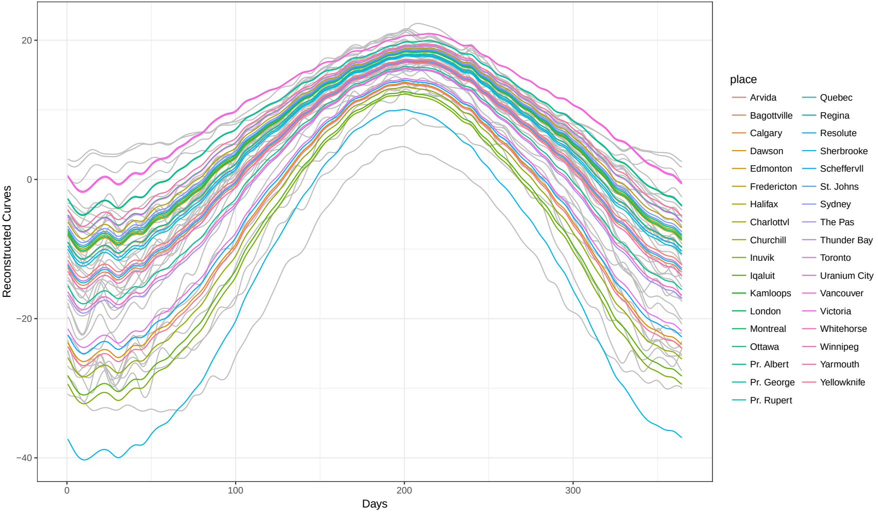
$$\psi = \mathbf{T}\xi$$

  that will be a little easier to interpret
- use the VARIMAX algorithm from multivariate PCA
  - let $\mathbf{B} = (\xi_i(t_j))_{ij}$ be matrix $K \times n$, let $\mathbf{A} = \mathbf{TB}$
  - find $\mathbf{T}_{VARIMAX}$ maximizing $\mathrm{Var}(a_{11}^2, \ldots, a_{Kn}^2)$ over all orthonormal matrices $\mathbf{T}$
  - it happens if a_{ij} are strongly positive, strongly negative or tend to 0
- collectively, all rotated functions ψ still account for the same part of variation as functions ξ (but they divide this variation in different proportions)
- the rotated component scores are no longer uncorrelated! (however ψ may have better interpretation)

Rotated Principal Components

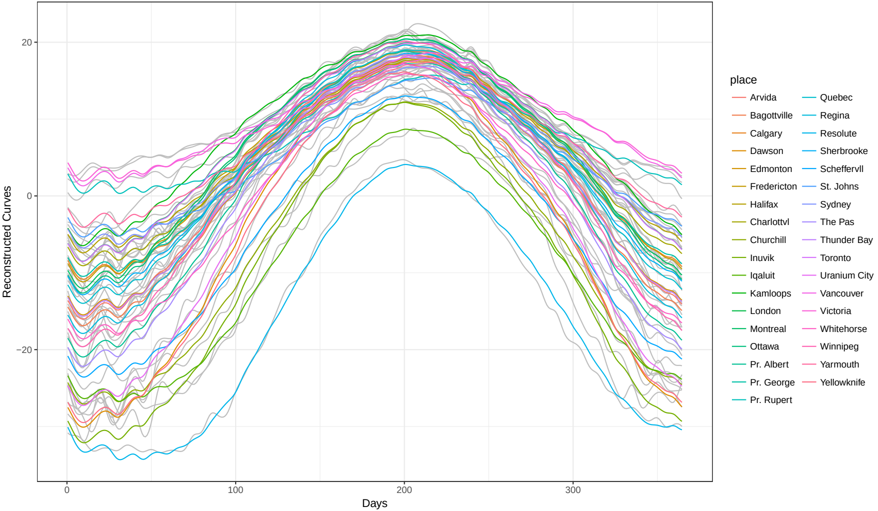


Reconstruction



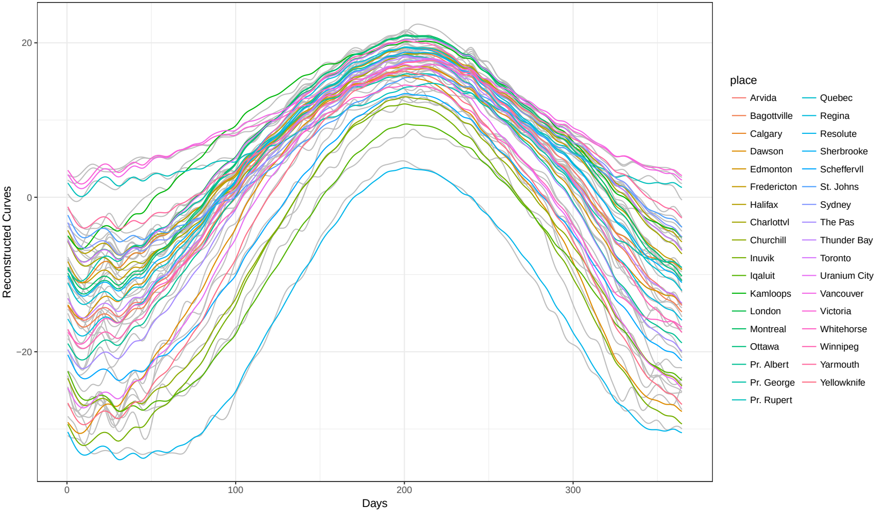
Reconstruction – 1 principal component

Exploratory Data Analysis



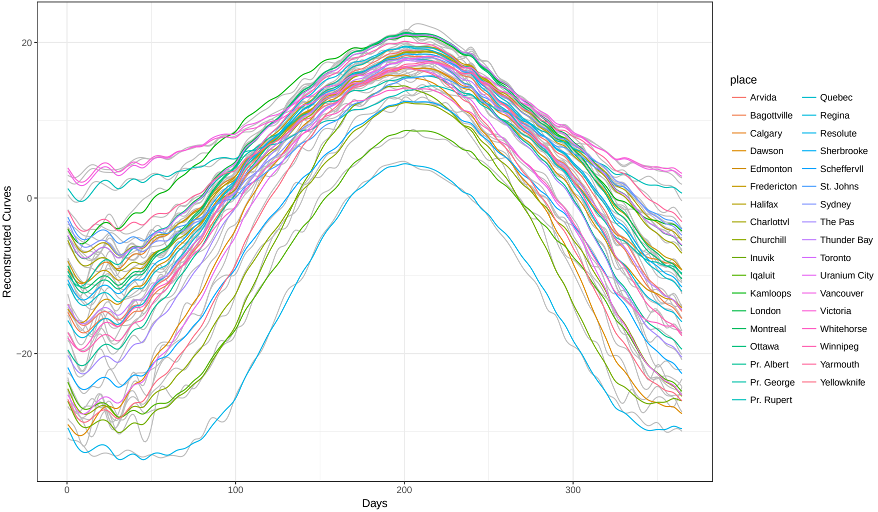
Reconstruction – 2 principal components

Exploratory Data Analysis



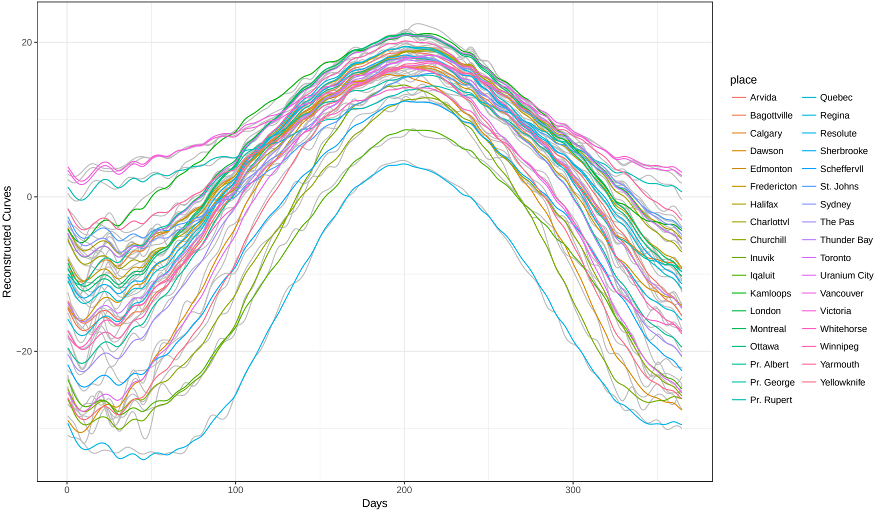
Reconstruction – 3 principal components

Exploratory Data Analysis



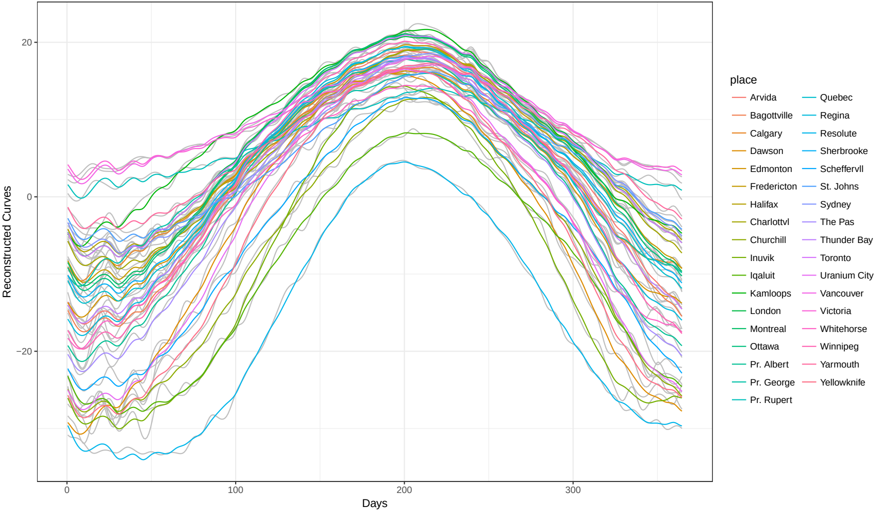
Reconstruction – 4 principal components

Exploratory Data Analysis



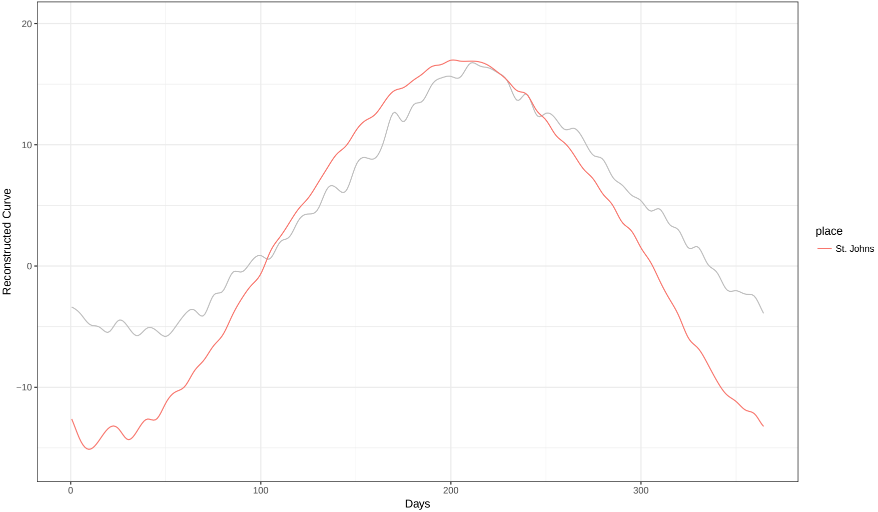
Reconstruction – 5 principal components

Exploratory Data Analysis

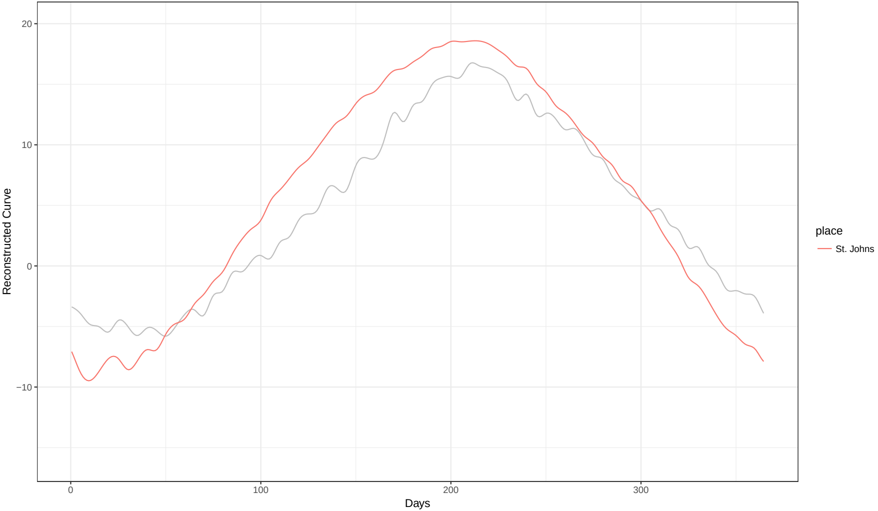


Reconstruction – 6 principal components

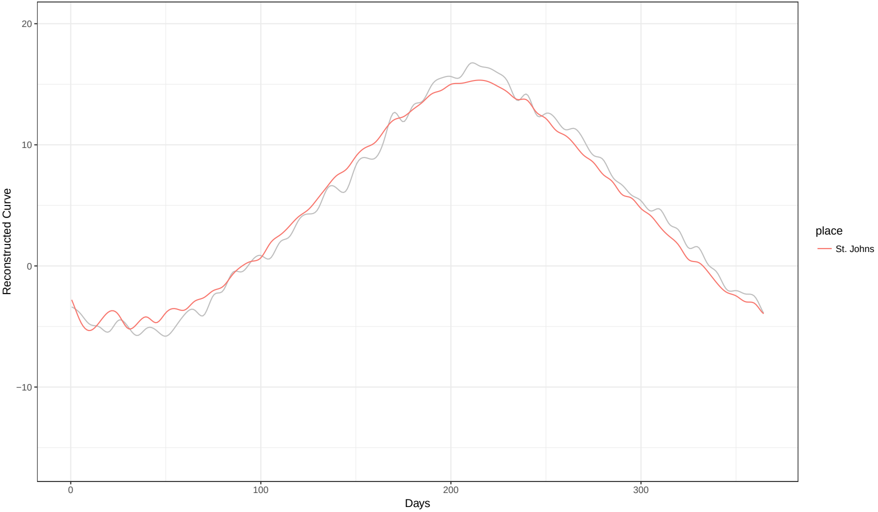
Exploratory Data Analysis



Mean Curve

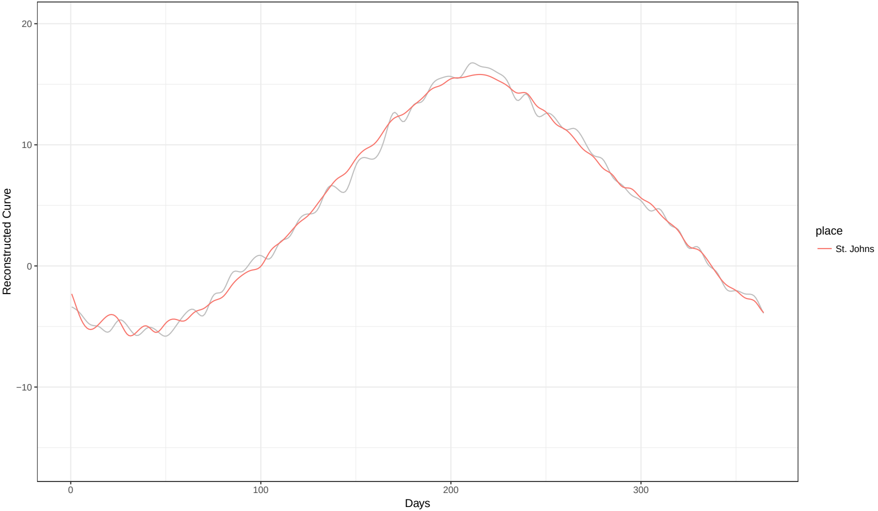


Reconstruction – 1 principal component



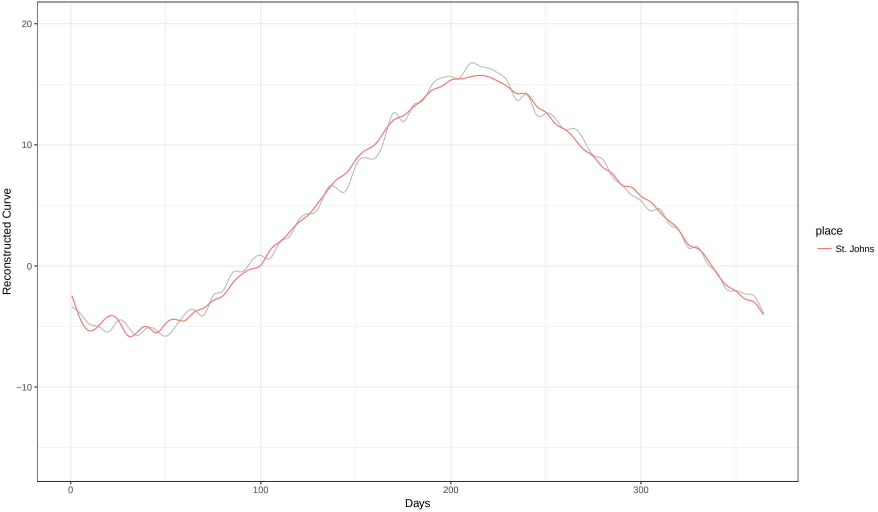
Reconstruction – 2 principal components

Exploratory Data Analysis

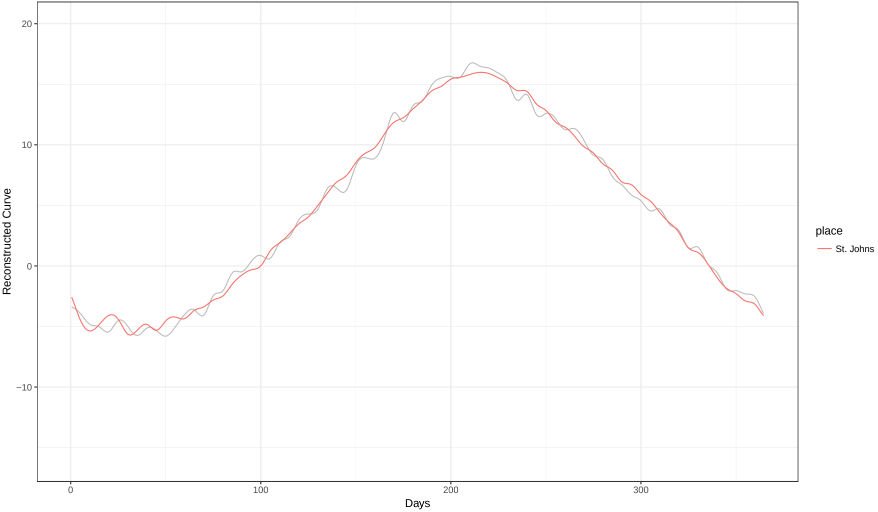


Reconstruction – 3 principal components

Exploratory Data Analysis

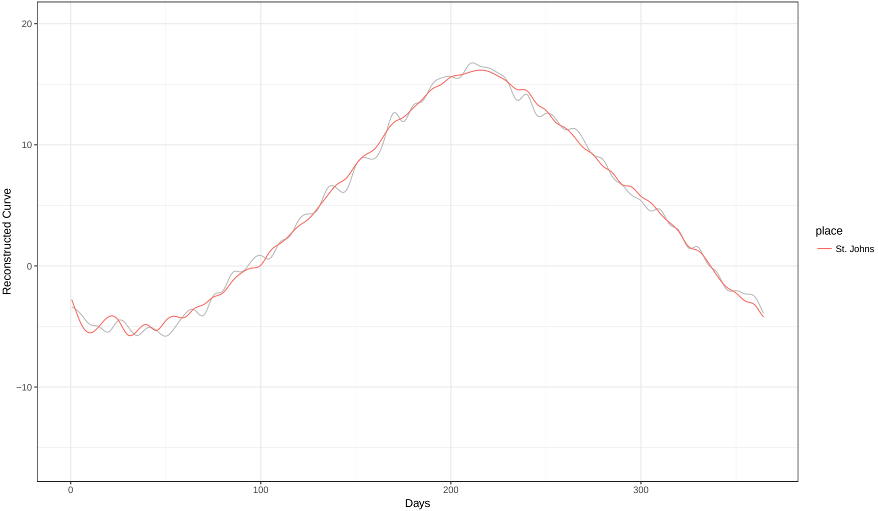


Reconstruction – 4 principal components



Reconstruction – 5 principal components

Exploratory Data Analysis



Reconstruction – 6 principal components

Problems to solve

1. Pinch-Force Data
   Load the variable `pinch` from the `fda` package. The variable `pinch` contains 20 replications of a subject pinching between their thumb and forefinger. For each replicate, the force of the pinch was recorded at 151 time points.
   - Smooth the data by B-spline bases with second-derivative penalties and plot the result (see Figure 1).
   - Conduct a principal components analysis of these data. How many components do you need to recover 90% of the variation? Do the components appear satisfactory? Plot the principal components (see Figue 2).
   - Try a smoothed PCA analysis. Choose the smoothing parameter by cross-validation. Plot the cross-validation curve (see Figure 3). Plot the new smoothed principal components (see Figure 4). Does this appear to be more satisfactory? Can you interpret the principle components?
   - Apply a varimax rotation to the smoothed principle components and plot them (see Figure 5). Does this rotation change your interpretation?

Problems to solve

2. Handwriting Data
   Load the variable `handwrit` from the `fda` package.
   - Smooth the data by B-spline bases with second-derivative penalties. Select a reasonable number of knots given the nature of the data and the number of observations. You may choose a smoothing parameter as any reasonable value. You should expect this to be small. Plot the smoothed curves and plot the mean (see Figure 6).
   - Conduct a principle components analysis on the bivariate data. How many components are necessary to explain 90 % of the variation? Interpret the two leading components, including a plot of the mean writing with variation in this components around it (see Figure 7).

3. Medfly Data
   Load the variable `medfly` from the `medfly.RData` file. The data consist of records of the number of eggs laid by 50 fruit flies on each of 31 days, along with each individual's total lifespan.
   - Smooth the data for the number of eggs, choosing the smoothing parameter by GCV. Plot the smooths (see Figure 8).
   - Conduct a principal components analysis using these smooths. Are the components interpretable? How many do you need to retain to recover 90 % of the variation? Plot the components (see Figure 9). If you believe that smoothing the PCA will help, do so.
   - Divide the population to 2 groups by the lifespan level: flies with "low" level (lifetime less then a half) and flies with "high" level (the others). Plot the PCA scores of the first principal component against the PCA scores of the second principal component for all samples. For each point set the color by its group (see Figure 10). What can we conclude?

Problems to solve

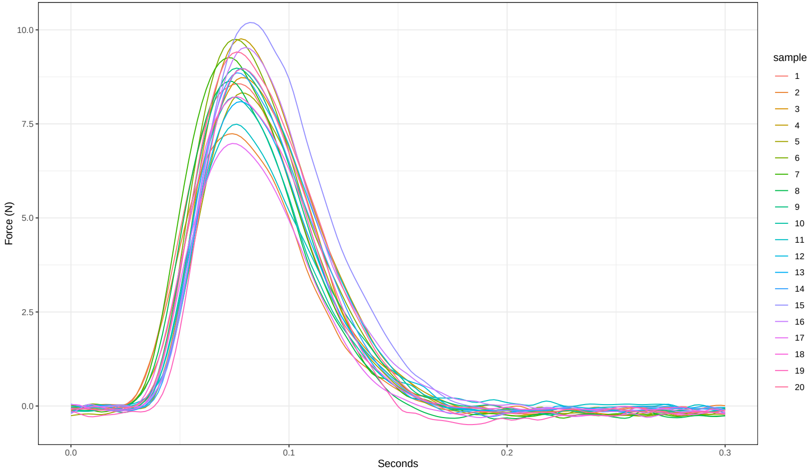


Figure 1.

Problems to solve

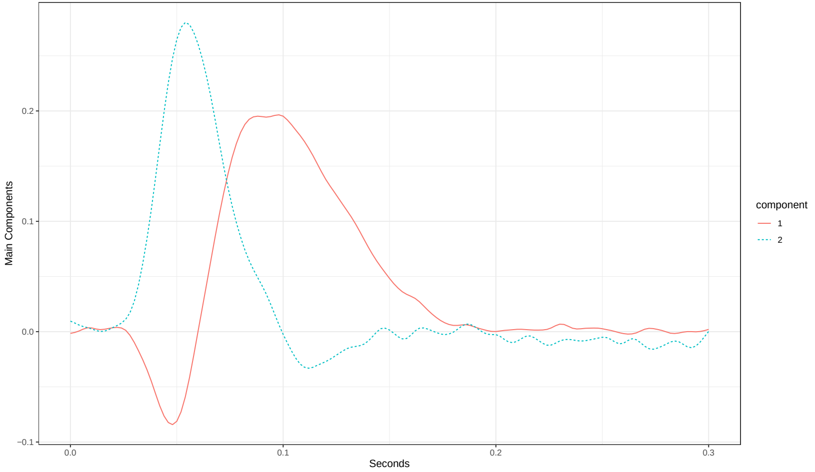


Figure 2.

Problems to solve

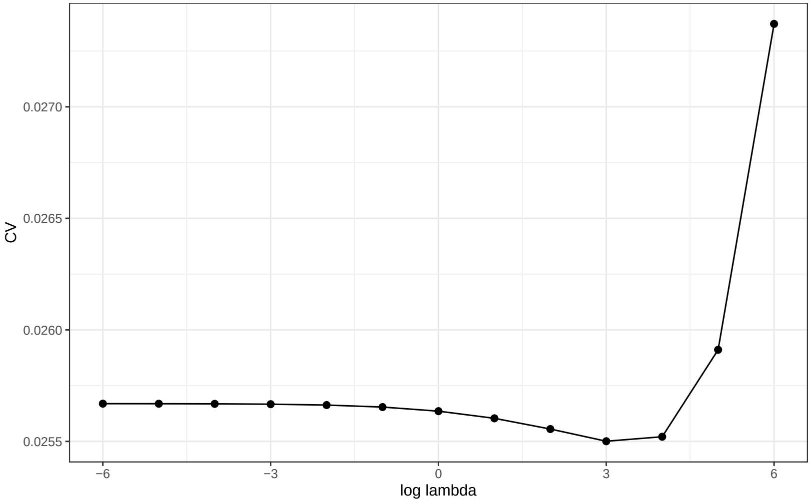


Figure 3.

Problems to solve

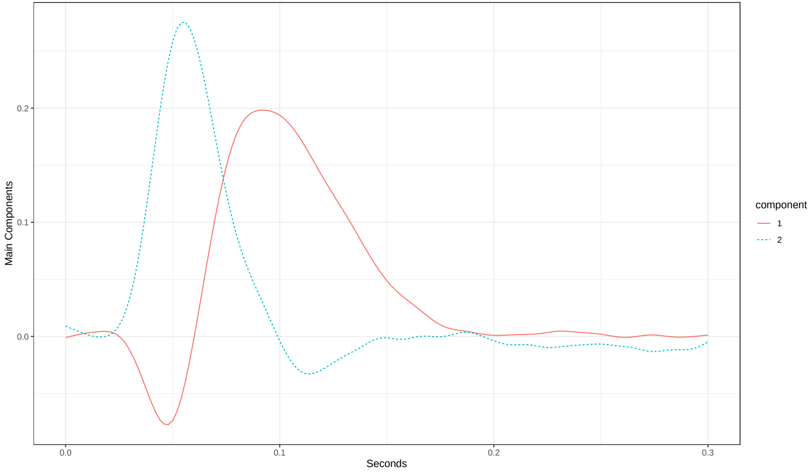


Figure 4.

Problems to solve

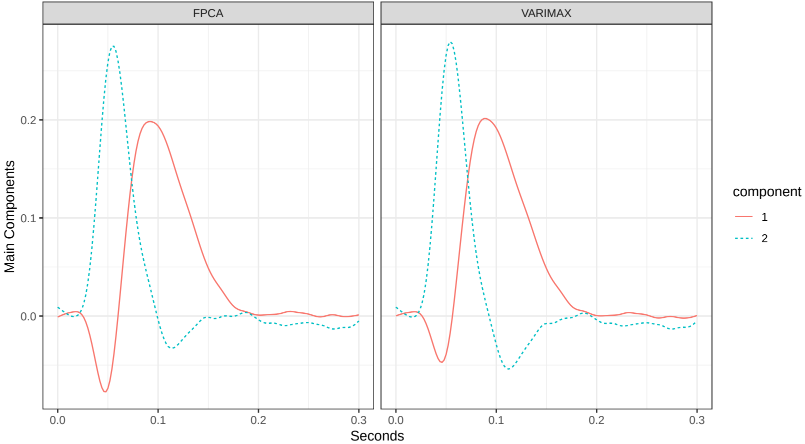


Figure 5.

Problems to solve

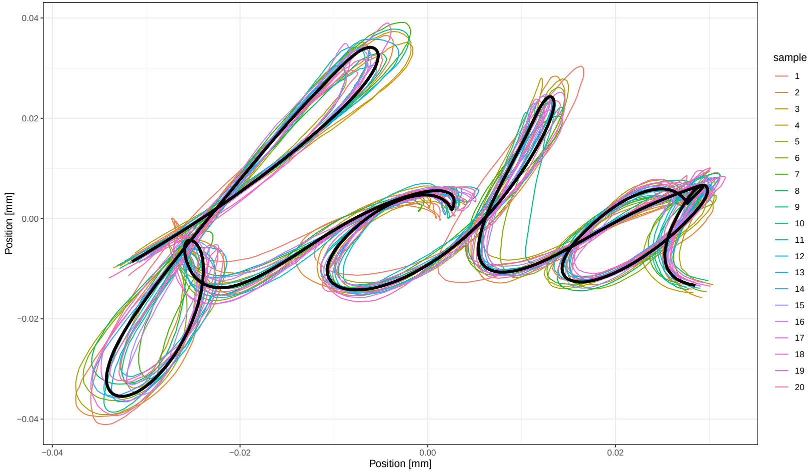


Figure 6.

Problems to solve

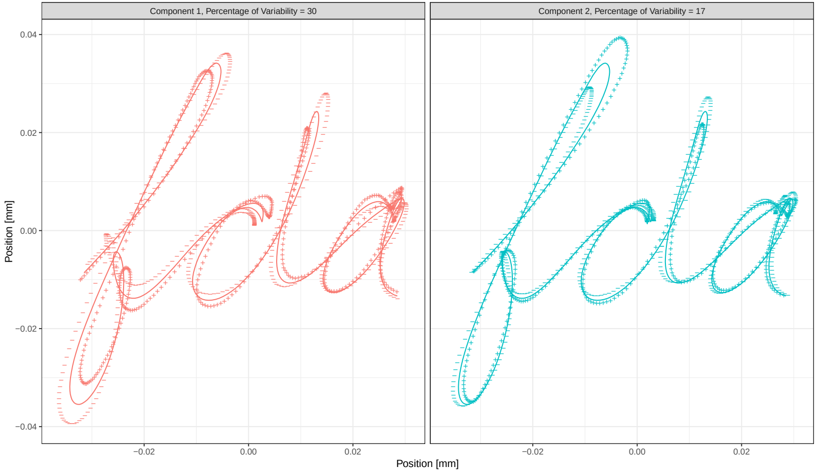


Figure 7.

Problems to solve

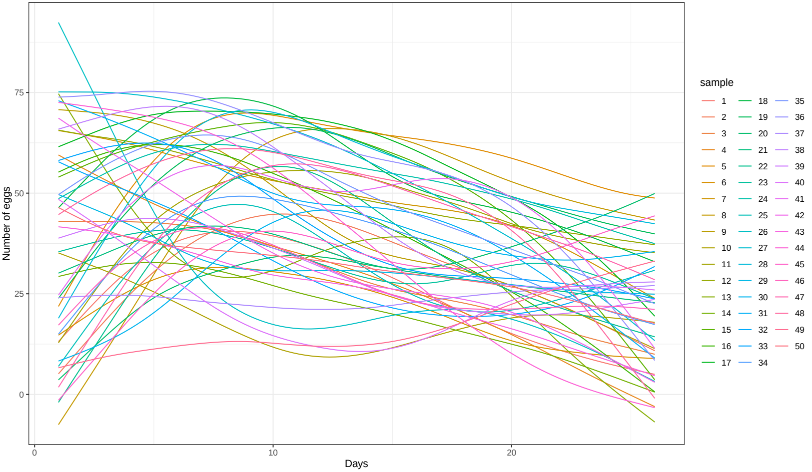


Figure 8.

Problems to solve

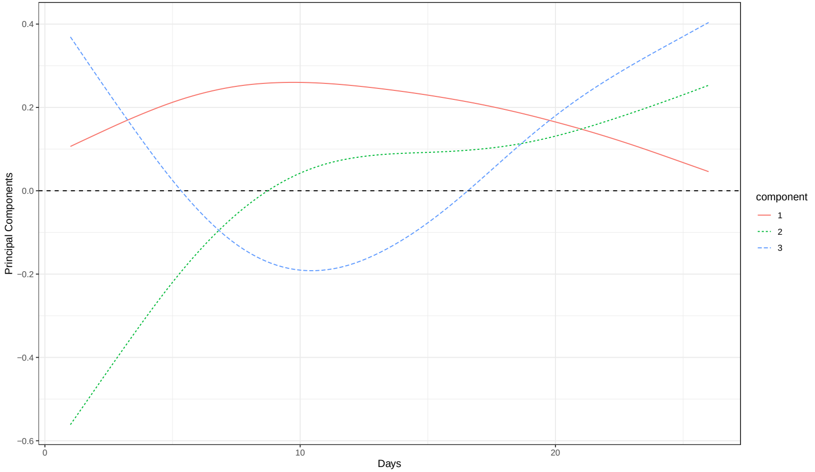


Figure 9.

Problems to solve