Abstract

In the previous report, the smooth process under functional data has been found. Based on the previous report, Fourier basis with harmonic linear acceleration is used to fit the functional data of fruit flies. Then, the main focus of this report is to carry out dimension reduction technique for interpreting functional data, as well as doing regressions between predictor and response.

Thus, aim of this report firstly seeks interpretability inside the data through PCA, fPCA, PLS and dynamic fPCA. Dynamic fPCA provides the best performance of keeping the most amount of information.

Then, regressions are going to be made in this report to discover the potential relationship between the variation of egg count and life time, through pointwise PCA, functional PCR, low-dimensional functional regression, and introducing smoothing penalty. Both estimate methods of functional regression provide great result.

Four fPCs are found in the fruit flies' data, with great interpretations with respect to the egg counts of different time. Then, the prediction of β indicates that during earlier stage of the time span, large egg counts generally has long life span. This positive effect of egg counts change for the later stage. More numerical results are provided in the numerical part of this research.

Keywords: functional PCA, PLS, Dynamic fPCA, functional PCR

1. Introduction

After functional data is fitted through different basis and smoothed, we would want to explore further into reducing the dimensions of data to seek for better intepretation. If we have response variable (as in the fruit fly data), we would be able to conduct functional regression and principal component regression to study the relationship between independent variable and response.

1.1 Problems Targeted

This report mainly focused on two problems: general and functional principal component analysis, then general and functional linear regression. Problems targeted in this report are:

(a) Conduct a principal components analysis using these smooths. Interpret the PCs that retain 90% of variation.

(b) Try dynamic fPCA. Find if there is any improvements and explain.

(c) Perform a functional linear regression to predict the total lifespan of the fly from their egg laying. Choose a smoothing parameter by cross validation, and plot the coefficient function along with confidence intervals. Calculate the R^2 for regression.

(d) Try a linear regression of lifespan on the principal component scores from your analysis. Test whether the model is significant. Compare with the model obtained by functional linear regression.

1.2 Methods Involved

In this report, methods introduced in order to reach the desired conclusions are provided here. To find the optimal smoothing parameter λ: generalized cross validation (GCV). Summary statistics: the mean curve, covariance matrix and etc. Dimension reduction: principal component analysis (PCA), functional principal component analysis (fPCA), Partial least squares (PLS), dynamic fPCA. Display and visualization: scree plot, cumulative plot, level plot, perspective plot and etc. Regression: linear regression, functional principal component regression, scalar-on-function regression model. Estimation of regression coefficient: low-dimensional regression coefficient function, using a roughness penalty.

2. Models and Inference

Other than trying various basis functions, and retrieving curves with great fitted result and smoothness in report 1. In this report, problems are more focused on finding approaches to provide better intepretability, and doing regressions for multiple curves.

2.1 Problems Reviewed

In this section, Fourier basis function was applied. Then, the generalized cross-validation (GCV) was adopted for determining the optimal smoothing parameter (λ). Some summary statistics are calculated at the same time (mean and variance). The crucial process is to determine and interpret the co-

efficients of principal components, through PCA, fPCA and dynamic fPCA. Then, linear regression and functional linear regression were conducted to depict the relationship between functional attributes and scalar response variable.

2.2 Methodology

In this section, the calculation of several methods based on the idea of PCA are listed. Attempts are made based on PCA, FPCA, dynamic FPCA and additionally partial least square is applied.

2.2.1 Partial Least Square (PLS)

Similar as PCA, PLS is also a dimension reduction method. Idea behind PLS is to find a linear regression model by projecting the predicted variables and the observable variables to a new space. PLS model tries to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.

The general underlying model of multivariate PLS is:

$$X = TP^{\mathrm{T}} + E$$
$$Y = UQ^{\mathrm{T}} + F$$

Where X is matrix of predictors, Y is matrix of responses; T and U are projections of X and Y. P and Q are orthogonal loading matrices, with E and F as error terms.

2.2.2 Functional Principal Component Analysis (FPCA)

Instead of covariance matrix Σ, the functional data used a surface $\sigma(s,t)$. The eigendecomposition of the PCA is written as: $\Sigma = U^T DU = \Sigma d_i u_i u_i^T$. For functions, this is the Karhunen-Loeve decomposition:

$$\sigma(s,t) = \sum_{i=1}^{\infty} d_i \xi_i(s)\xi_i(t)$$

Where d_i represents the amount of variation in direction $\xi_i(t)$. While $d_j / \sum d_j$ is the proportion of variance explained. $\{\xi_1, \ldots\}$ is the basis system.

Thus, the principal component scores are:

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

For the collection of curves $x_i(t)$, $i = 1, \cdots n$. The aim is to find the probe $\xi_i(t)$ that maximizes:

$$Var[\int \xi_i(t)X(t)dt]$$

With the constraint that $\int \xi_1(t)^2 dt = 1$. Then for $\xi_2(t)$, variance is maximized subject to the orthogonality:

$$\int \xi_1(t)\xi_2(t)dt = 0$$

The calculation of fPCA computing is to solve the eigen-equation

$$\int \sigma(s,t)\xi_j(t)dt = \lambda\xi_j(t)$$

In R language, the fda package completes the calculation by transforming the question to calculating the coefficients. When $x_i(t)$ has common basis expansion, it is the same with eigen-functions. Thus, it is possible to re-express eigen-equation in terms of co-efficients. The basis expansion would be therefore apparent for smaller eigenvalues.

2.2.3 Dynamic Functional PCA

Functional principal component analysis (FDA), though a key technique in the field and a benchmark for any competitor, does not provide an adequate dimension reduction in a time series setting. As we are trying to deal with time series data (fruit flies, Canadian weather, knee and hip angles), dynamic fPCA often yields better performance.

When assume $(x_r(t))$ be a stationary process, then the operator $\mathcal{F}_\theta^X$ is called the spectral density operator of $(x_r(t))$ at frequency θ with the kernel

$$f_\theta^X(s,t) = \frac{1}{2\pi}\sum_{h\in\mathbb{Z}} \sigma_h(s,t)e^{-\mathbf{i}h\theta}$$

Then, use dynamic eigendecomposition for $\mathcal{F}_\theta^X(s,t)$ to acquire dynamic eigenvalue $d_j(\theta)$ and dynamic eigenfunction $\psi_j(t|\theta)$. Let

$$\psi_{jl}(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi_j(t \mid \theta)e^{-\boldsymbol{i}l\theta}d\theta$$

Finally, to calulate the j-th dynamic score of $x_r(t)$, which is:

$$f_{jr} = \sum_{l\in\mathbb{Z}} < x_{r-l}, \psi_{jl} > \text{ for } r \in \mathbb{Z}, j \geq 1$$

For dynamic fPCA, the proportion of variance explained by the first p PCs is defined as 1 − NMSE(p), where the normalized mean squared error is

$$NMSE(p) = \frac{\sum_{k=1}^n \|x_k - \hat{x}_k\|^2}{\sum_{k=1}^n \|x_k\|^2}$$

2.2.4 Functional Linear Regression

The functional linear regression examines the predictive relationships between functions based on generalization of linear models. It is possible in three different scenarios: scalar-on-function, function-on-scalar, function-on-function. For the later parts, we usually have scalar response to analyze, thus the scalar-on-function case would be focused on.

Scalar-on-function

In general linear regression, there are fewer covariates than observations. But if y_i and $x_i(t)$ are used, there are infinitely many covariates for

$$y_i = \int \beta(t)x_i(t)dt + \epsilon_i$$

To improve the identifiability, if the smoothness of $\beta(t)$ is agreed, then it is possible to fit by penalized squared error:

$$\text{PENSSE}_\lambda(\beta) = \sum_{i=1}^{n}\left(y_i - \alpha - \int \beta(t)x_i(t)dt\right)^2 + \lambda\int [L\beta(t)]^2 dt$$

Where $\beta(t) = \sum c_i\phi_i(t)$. And when $x_i(t)$ are represented by the same basis. By calculation,

$$\hat{\text{y}} = \int \hat{\beta}(t)x_i(t)dt = Z\begin{bmatrix}\hat{\alpha}\\ \hat{\text{c}}\end{bmatrix} = S\text{y}$$

Choosing parameter is always conducted through cross-validation.

It is also possible to derive the confidence interval of $\beta(t)$:

$$\Phi(t)\hat{\text{c}} \pm 2\sqrt{\Phi(t)^T\ \text{Var}[\hat{\text{c}}]\Phi(t)}$$

Multivariate and Mixed Functional Linear Regression

In addition to having only functional covariates, it is also possible to habe scalar covariates z and multiple functional covariates $x_i(t), \cdots x_K(t)$ at the same time.

$$y_i = \alpha + z_i\gamma + \sum_{j=1}^{k}\int \beta_j(t)x_{ij}(t)dt + \epsilon_i$$

The penalized sum of squares is

$$\sum_{i=1}^{n}\left(y_i - \alpha - z_i\gamma + \sum_{j=1}^{k}\int \beta_j(t)x_{ij}(t)dt\right)^2 + \sum_{j=1}^{K}\lambda_j\int \left[L_j\beta_j(t)\right]^2 dt$$

When $\zeta = \begin{bmatrix}\alpha & \gamma^T & c_1 & \cdots & c_k\end{bmatrix}^T$, and $R_j = \int L_j\Phi_j(t)L_j\Phi_j(t)^T dt$,

$$R = \begin{bmatrix}0 & 0 & \cdots & 0\\ 0 & \lambda_1 R_1 & \cdots & \vdots\\ \vdots & \vdots & \ddots & 0\\ 0 & \cdots & 0 & \lambda_k R_K\end{bmatrix}$$

Then ζ is possible to be calculated through:

$$\hat{\zeta} = \left(Z^T Z + R\right)^{-1} Z^T\text{y}$$

Principal Components Regression

For $y_i = \beta_0 + \sum\beta_j x_{ij} + \epsilon_i$. When the goal is to mainly focus n future prediction, with covariates x having high dimensionality and correlation, it is possible to take the principal component ξ_j to $x_i = \sum_{j=1}^{p}\alpha_{ij}\xi_j$. Then the model becomes:

$$y_i = \beta_0' + \sum_{j=1}^{p'}\beta_j'\alpha_{ij} + \epsilon_i$$

The model kept α_{ij} uncorrelated, and reduce the dimension through choosing a specific number of PCs.

Functional PCR

For functional data analysis, the principal components regression could also be generalized. For each functional factor score f_{ij}, it is possible to use the model:

$$y_i = \beta_0 + \sum\beta_j f_{ij} + \epsilon_i$$

Recall that $f_{ij} = \int x_i(t)\xi_j(t)dt$, in calculation, we always use

$$y_i = \beta_0 + \int \beta(t)x_i(t)dt + \epsilon_i$$

For the functional PCR, the confidence interval of $\hat{\beta}(t) = \sum\hat{\beta}_j\xi_j(t)$ is able to be calculated with

$$\text{Var}[\hat{\beta}(t)] = \sum \text{Var}\left[\hat{\beta}_j\right]\xi_i^2(t)$$

3. Numerical results

In this report mainly focus on the fruit flies data, similar with the data provided in practical ch3and ch4a, the egg count and lifetime of fruit files are functional data. The different part is that the data itself contains the predictor variable X (egg count) and response variable Y (life time). Thus, it is possible to make regression between life time and egg count of fruit files inside within the information in this data set.

In the sessions below, the fruit files data are smoothed at first with smoothing penalties. Some summary statistics are calculated (mean, variance), then for dimension reduction fPCA, Dynamic fPCA, fPLS are applied and compared with visualization. After dimension reduction, Scalar on functions regressions were made for original data points, functional curves and principal components. Interpretations of components and regressions were made as much as possible.

3.1 Description of the Data

The fruit files' data contains 50 files with egg count at 26 time points and each fly's life time. The independent variable in this data is egg count(X), the response variable is life time(Y). Egg count ranges between 0 to 112, with mean of 37.58, life time ranges between 124 and 739, with the mean of 466.26. The scalar on function regression is possible for this data set/ Other data sets used in practical ch3 and ch4a are also listed for comparison in Table 1.

The temperature data (Celsius) summarizes data collected at 35 different weather stations in Canada on the average daily temperature for each day of the year. It ranges from -34.8 to 22.8, with mean of 1.88.

The precipitation (mm) data in Canadian weather data has collected from 35 different weather stations in Canada average daily rainfall for each day of the year rounded to 0.1 mm. It is a very skewed distribution ranges from 0 to 16.4, with mean of 2.17, and median of 1.7.

The knee and hip angle data (gait) collected hip and knee angle in degrees through a 20 point walking movement cycle for 39 boys. The data contains components of standardized gait (from=0.025, to=0.975, by=0.05), subject ID, and gait variable ("Hip Angle" or "Knee Angle").

The Australia fertility data has 86 years of fertility (count of live birth per 1,000), for ages between 15 and 49. Count of live birth ranges between 0.01 to 260.88, with the mean of 70.50.

3.2 Exploratory Data Analysis

Firstly, the temperature data introduced are used. I created 365 Fourier basis function, calculating the value and its 2^{nd} derivative at each time point. The amplitude of sine and cosine function is $\sqrt{\frac{2}{P}} \approx 0.277 = c^{\star}$. Then, for the saturated model, introduce the harmonic acceleration penalty $Lx = D^3x + w^2Dx$ and $\lambda = 100$ to smooth the curves. λ is chosen as report 1 suggests that most 39 of 50 flies achieve the lowest GCV scores below $\lambda = 100$. For this data, $\omega = 2\pi/26$.

Then some summary statistics would be able to calculate. A mean curve is calculated and plotted in Figure 1.1 (see coefficients of the mean curve in Appendix A.1), as well as the variance between curves. The mean egg count of fruit flies initially rises, then decreases with time. Curves are more dispersed around the 5^{th} to 10^{th} time points. Since We cannot plot the covariance matrix directly (could

Table 1: Data Set Description.
Including data sets used in Practical Ch3, Practical ch4b and in project 2. The data sets include temperature data of Canadian weather, precipitation, knee and hip data, Australia fertility data. The fruit flies data used in the next section is also listed here.

Names	Size	Predictor (X)	Response (Y)	Range	Mean
Temperature (°C)	365×35	daily temperature	Null	$[-34.8, 22.8]$	1.88
Precipitation (mm)	365×35	daily rainfall	Null	$[0, 16.4]$	2.17
Knee and Hip (°)	20×30×2	hip angle	Null	$[-12, 64]$	26.69
		knee angle		$[0, 82]$	29.97
Australia fertility	86×35	count of live birth	Null	$[0.01, 260.88]$	70.50
Fruit Flies	$26 \times (50+1)$	egg count	life time	$x \in [0, 112]$, $y \in [124, 739]$	$\bar{X} = 37.58$, $\bar{Y} = 466.26$

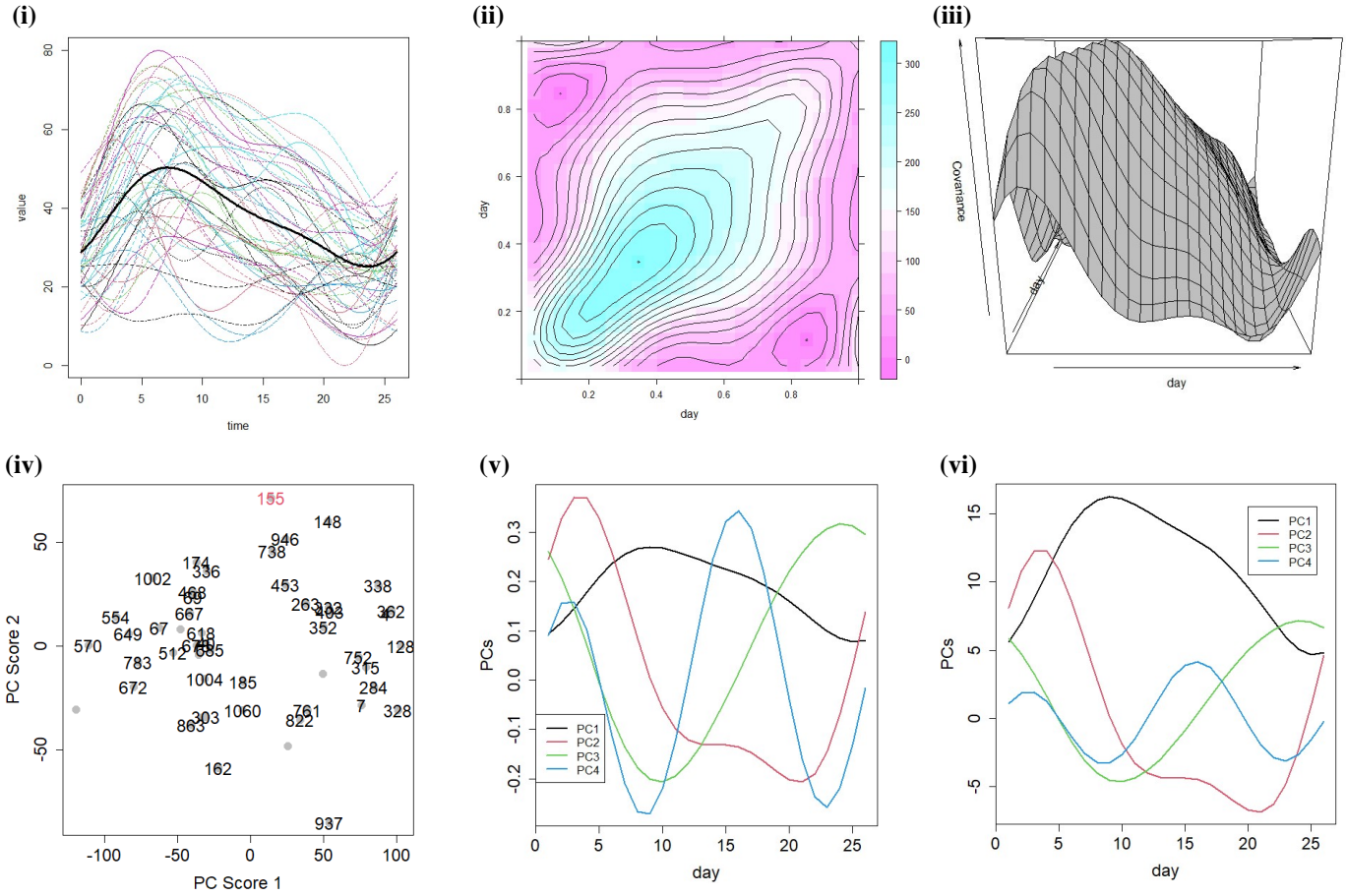


Figure 1: Visualization of Fourier fitted curves of fruit flies data. **(i)** plots the 50 curves of fruit flies, with a mean curve in bold. **(ii)** is the level plot with contour, with the covariance range of $[1.03, 300.99]$. Between 5 and 10 days, curves have larger variance. **(iii)** plots the perspective plots, with the height has covariance. Both figures imply the large variance happen between the time point of 5 to 10 days. **(iii)** is the projection on the first two fPCs, with each fruit fly labeled. **(v)** shows the variation of first 4 fPC of time points, possible interpretations are provided. **(vi)** shows the variation of scaled 4 fPC of time points, multiplied by the correspond eigenvalue.

create infinitely many covariances), thus evaluate it at 26×26 fine grid and then plot the level plot with contour, colored by variance, and heat map. Regions with more cyan means larger variance, pinky areas have less variance (Figure 1.2, Appendix B.1). Figure 1.2 shows that larger covariance within each day approximately from 5 to 10 days, with the 9^{th} day as the highest. The perspective plot (Figure 1.3) of the variance surface between days would also be able to visualize through fda package, with the peaks and saddle points clearly shown in a 3-dimensional space. Both figures provide clear visualization of a large covariance between 5 to 10 days within a few days around.

3.3 Functional Principal Components (FPCA)

The next step is to evaluate the fPCA for the fruit flies' egg count. In this report, the uncentered data is used. Assume each $x_i(t)$ has a common Fourier basis expansion, then must the eigen-functions. Then, we use the scree plot of eigenvalue, as well as the cumulative plot of proportion of variance to choose the number of PCs (Appendix B.2). In the scree plot, the subjective understanding of the 'elbow point' is subtle. In the cumulative plot, 3 PCs are enough to exceed the 90% of variance (95.84%). While 4 PCs account for 98.54% of variance. For better display, only the basis with coefficients larger than 0.05 are displayed in the table of 4 PCs (fPC is ξ_j in Karhunen-Loeve), for a detailed version, please refer to Appendix A.2. In fda package, it can re-express eigen-equation in terms of coefficients.

In Table 2, the 1^{st} PC is mainly constructed from combination of constant and $\sin(\omega t)$. The coefficients for constant is very high (0.94), which implies that it is very likely a weighted average of curves. For the 2^{nd} fPC, it heavily relies more on the $\sin(\omega t)$, $\cos(\omega t)$, $\sin(2\omega t)$ basis. The 3^{rd} relies more on combination of constant and $\cos(\omega t)$. The first 4 fPCs all have very different combination of

Table 2: Summary of Functional Principal Components. **Panel A** provides the display of the first 4 fPCs (coefficients larger than 0.05), the coefficients re-express eigen-equation. **Panel B** shows the variance of each fPC. **Panel C** provides the coefficients after varimax rotation, coefficients are more aggregated towards particular trend. The constant term decreases, factors have less intersections.

Panel A: fPC Coefficients

Basis Function	PC1	PC2	PC3	PC4
const	0.9435535		0.32172928	
$\sin(\omega t)$	0.1361661	0.7721753		
$\cos(\omega t)$		0.4944415	0.77378978	
$\sin(2\omega t)$		0.3954977	0.05234972	0.88946393
$\cos(2\omega t)$			0.06453471	0.22891670
$\sin(3\omega t)$				0.06658826

Panel B: fPCA Variance

	PC1	PC2	PC3	PC4
Proportion of Variance	0.66359470	0.20120939	0.09362974	0.02698746
Intuition	weighted average	earlier v.s. later	last+earliest v.s. earlier	3^{rd} v.s. 2^{nd} and 4^{th} quartile

Panel C: Factor Coefficients

Basis Function	PC1	PC2	PC3	PC4
const	0.4943229		0.5057947	
$\sin(\omega t)$	0.5819104	0.4219617		
$\cos(\omega t)$		0.5454999	0.4692213	
$\sin(2\omega t)$		0.5458775		0.4447997
$\cos(2\omega t)$				0.1147436

basis function. Visualizations of each fPCs are made to provide intuitions of each fPC. Interpretations are also provided based on the varation of curves and its variance. Figure 1.5 shows the variation of the 4 fPCs with time. The intuition is that FPC1 is mainly a weighted average of all curves. FPC2 could be interpreted as the large egg counts at the earlier stage versus the less in later sessions. FPC3 could be interpreted as the large egg count at the end and beginning versus less before. FPC4 could be depicted as large egg count near the 15^{th} day versus less between the time near 9^{th} and 23^{th} days. Figure 1.5 is the projection of the data onto the first 2 fPCs plane. Take fly 155 (in red) for illustration, it has the highest PC2, while it has large egg counts in the first half of time than the later half (highest egg count happens at the 2^{nd} and 4^{th} time points with 75 egg count). Figure 1.6 shows the scaled fPCs, which is multiplied by the correspond eigenvalue λ_i to combine the importance of information.

The variation of each component is plotted by $\bar{x}(t) \pm 2\sqrt{d_i}\xi_i(t)$, with d_i as the variance explained, thus the wider the variation band means larger information. Thus, in Figure 2.1, FPC1 variation band is wide along the x-axis, simply implies that it contains information of all time. FPC2 has wide variation band for the first ten days, then intersects and changes direction for illustrating the negative information of later egg count. FPC3 mainly has information of positive effect at the margins versus in the middle. FPC4 contains more information of positive effect between the 3^{rd} quartile versus 2^{nd} and 4^{th} quartiles of time points (Figure 2.2,3,4). The interpretations based on the variation coincides with previous understanding of fPC curves.

3.4 Partial Least Square Regression (PLS)

Compared with PCA, PLS is also a dimension reduction technique, with components selected to minimize the covariance between the distance of response variable Y. Since in the fruit flies' data, we also have a response variable (life time), it is possible to use PLS. By using cross-validation, PLS selects 4 components accounting to 99.52% of variance, which is a huge increase compared to PCA. It is only less than the performance of Dynamic fPCA

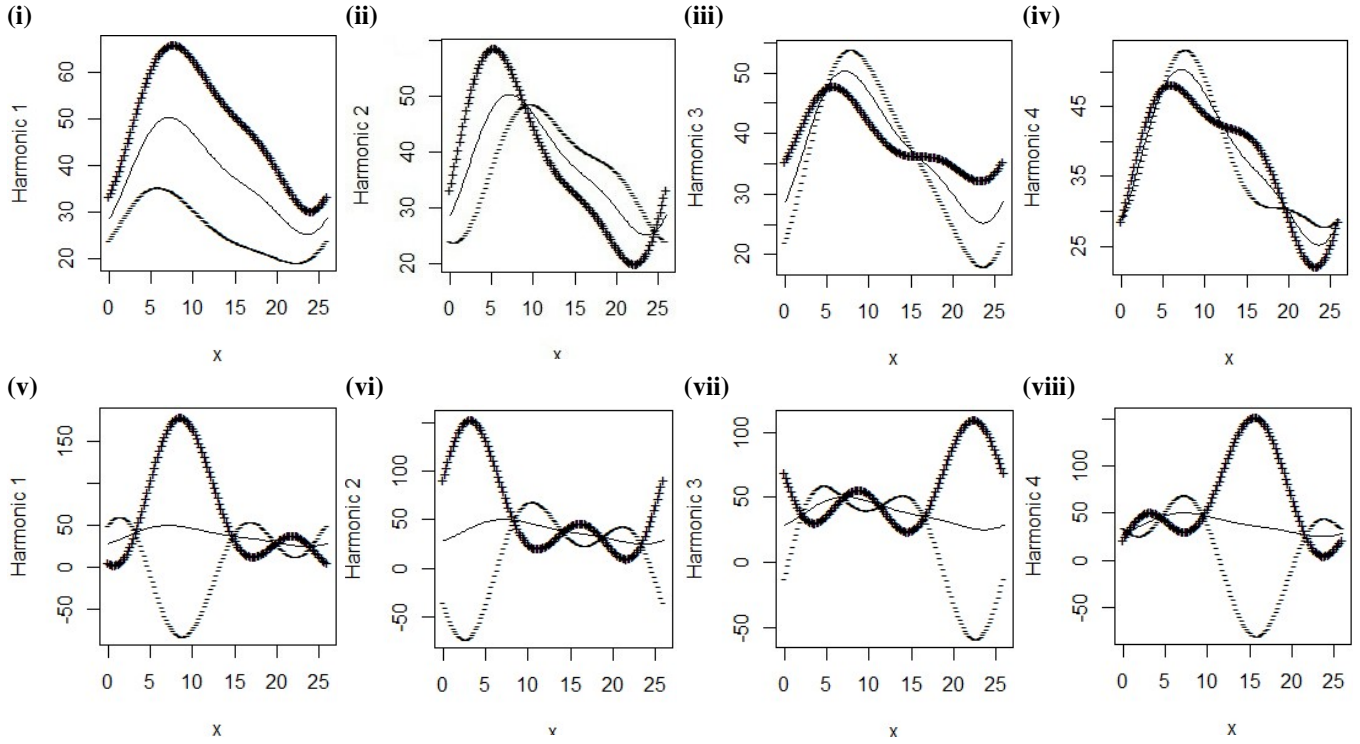


Figure 2: the first 4 fPCA and after Varimax rotation. Smooth curves are fPC, curves with + is the upper variation curve, and − is the lower variation curve. **(i)** is fPC1, contains information (66.4% of variability) of all time. **(ii)** is fPC2 (20.1% of variability), positive information in earlier versus negative in later time. **(iii)** is fPC3 (9.4% of variability), positive information at earlier times versus the beginning and the end. **(iv)** is fPC4 (2.7% of variability), positive information in the 3^{rd} quartile versus 1^{st} and 4^{th} quartiles of times. **(v)** is the 1^{st} factor (34.2% of variability), with information aggregated between 5 to 15 days. **(vi)** is the 2^{nd} factor (23% of variability), information aggregated between 0 to 7 days. **(vii)** is the 3^{rd} factor (13.7% of variability), information aggregated between 17 to 25 days. **(viii)** is the 4^{th} factor (27.6% of variability), information aggregated between 10 to 20 days.

(Table 3). The detailed coefficients of each components are in (Appendix A.5)

3.5 Varimax Rotation

Similar with PCA, after fPCA, it is also possible to carry out varimax rotation to find more interpretable basis as for factor analysis. The idea behind is try to find co-ordinate system where PC loadings are either very large or very small, through Varimax criterion of maximizing $Var(u_i^2)$.

Figure 1.5,6,7,8 display the components after varimax rotation, In fda, Varimax tends to emphasize particular regions. variability of each component becomes more balanced. The 1^{st} component emphasizes the information aggregated between 5 to 15 days. The 2^{nd} factor contains most the information between 0 to 7 days. The 3^{rd} component has more information between 17 to 25 days. While the 4^{th} factor mainly shows the information between 10 to 20 days. The four factors merely overlap, still accounting for 98.54% of variance, but with nicer display property.

3.6 Dynamic fPCA

The Dynamic fPCA requires to fit $x(t)$ by a set of basis function at first, then it would calcualte filters and provide dynamic fPCs. The direct use of fruit files' orginal data into dynamic fPCA would provide the result of DPCA. Thus, use the data already smoothed at the beginning with Fourier basis and penalty to compare with result provided by fPCA.

Improvements are made compared with fPCA. A table of comparison between orginal data with PCA, Dynamic PCA; as well as FPCA and Dynamic FPCA with smoothed Fourier basis function are provided (Table 3). Notice that the variance of dynamic fPCA is defined as 1 − NMSE(p)normalized mean squared error. Compared with fPCA, dynamic fPCA outperforms in both proportion of variance and cumulative proportion, with 99.94% of 4 fPCs than 98.54%. While in Table 3, Panel C, D, dynamic PCA still outperforms general PCA with a total of 71.65% variance than 67.62%. Which further states that dynamic fPCA outperforms other methods discussed above.

Table 3: Summary of Proportion of Variance explained of PCs. **Panel A** provides the proportion of the first 4fPCs with smoothed Fourier basis function, account to 98.54%, which is a high ratio (in bold). **Panel B** is variance proportion of first 4 dynamic fPCs after smoothed Fourier basis function, account to 99.94%, which even higher. **Panel C**, **D** and **E** provide the proportion by PCA and Dynamic PCA, PLS using the orgininal data, which has obviously lower variance.

Panel A: fPCA (Fourier basis)

	PC1	PC2	PC3	PC4
Proportion of Variance (%)	0.6636	0.2012	0.0936	0.0270
Cumulative (%)	0.6636	0.8648	0.9584	0.9854

Panel B: Dynamic fPCA (Fourier basis)

	PC1	PC2	PC3	PC4
Proportion of Variance	0.6863	0.2526	0.0424	0.0180
Cumulative	**0.6863**	**0.9389**	**0.9814**	**0.9994**

Panel C: PCA (original data)

	PC1	PC2	PC3	PC4
Proportion of Variance	0.3445	0.2112	0.0641	0.0563
Cumulative	0.3445	0.5558	0.6199	0.6762

Panel D: Dynamic PCA (original data)

	PC1	PC2	PC3	PC4
Proportion of Variance	0.3618	0.2146	0.0814	0.0587
Cumulative	0.3618	0.5764	0.6578	0.7165

Panel E: PLS (original data)

	PC1	PC2	PC3	PC4
Proportion of Variance	0.3169	0.5618	0.6165	0.6469
Cumulative	0.9230	0.9743	0.9909	0.9952

The reason behind this improvement would likely to be that Dymanic fPCA considers the property of data as time-series and assumed the weakly stationarity of $x_r(t)$, including the lag-h covariance kernel. It is very likely the historical egg counts would affect the future egg counts, especially for days that are close on the timeline.

3.7 Functional Linear Regression

For the fruit flies' data, it has response Y which is the flies' lifetime. Thus, we would want to relate lifetime to the shape of egg count. Thus, regressions are going to be conducted. However, different from the general linear regression, the regression of scalar-on-functions is through $y_i = \alpha + \sum \beta_j x_{ij} + \epsilon_i$. Where β_t could usually be represented through a basis expansion $\beta(t) = \sum c_i \phi_i(t)$. In later sessions, we would conduct several attempts and find optimal ways of doing functional data regression.

3.7.1 First Idea of Linear Regression

Notice that in linear regression, we must have fewer covariates than observations. Thus, the first idea is to use observations y_i, $x_i(t)$, and choose $t_1, \cdots, t_k$. I initially chose 13 basis functions and 10 points with equal distance. Then do linear regression for $y_i = \alpha + \mathbf{x}_i\beta + \epsilon$. It provides nice fit with $R^2 = 0.9873$, and adjusted $R^2 = 0.9844$. The regression result could be written as:

$$\begin{aligned} y_i = {} & 1.5665x_i(t1) + 3.8118x_i(t2) + 2.8401x_i(t3) \\ & + 2.3593x_i(t4) + 0.8838x_i(t5) + 0.4873x_i(t6) \\ & + 0.1081x_i(t7) - 0.9637x_i(t9) - 14.5537 \end{aligned}$$

However, some of the regression coefficients are not significant, for details please refer to Appendix A.3. For this consideration, we would going to look for more robust regressions.

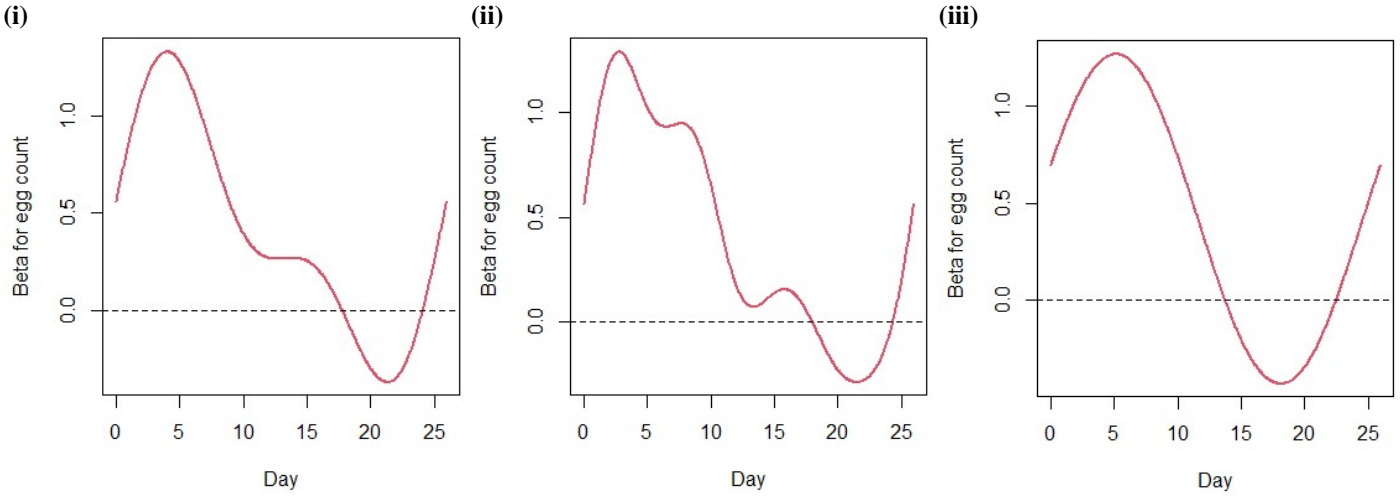


Figure 3: Estimate of $\beta(t)$. **(i)** is from the first estimate of beta with 5 Fourier basis function and constant intercept α. **(ii)** is from the second estimate of smoothing parameter $\lambda = 100$. **(iii)** is from the second estimate of smoothing parameter $\lambda = 10^{12.5}$, it is very smooth compared to the others. But we still need to figure out a *beta* both retaining the trend and keeping smoothness.

3.7.2 Functional PCR

Then, we notice that the coefficients of linear regression are not significant. Usual least squares may be inappropriate when there could be multicollinearity inside the data. Thus, functional PCR could be conducted by modelling $y_i = \beta_0 + \sum \beta_j f_{ij} + \epsilon_i$. By calculation, the model could be written as

$$y_i = 2.11597 f_{i1} + 2.94422 f_{i1} - 0.05287 f_{i2} + 0.99858 f_{i3} + 466.26000$$

Compared with introducing the penalty terms, this method avoids the need for cross-validation. Three of four coefficients are significant (<0.01), see Appendix A.4. The result is more robust and avoids multicollinearity. It still keeps a relatively high $R^2 = 0.9844$ and adjusted $R^2 = 0.983$. Combined with the overall F − statistics = 709.1 with a p − value $< 2.2 * 10^{-16}$, the model is significant.

3.7.3 Regress with Basis Coefficient Expansion

Two estimates of the regression coefficients will be calculated. Both redefine the problem using a basis coefficient expansion of beta.

First Estimate

The first estimate is through low-dimensional regression coefficient function: use a constant function for alpha, and 5 Fourier basis functions for beta. By calculation, for one time the intercept is approximately 3.464844 and for $\beta(t)$,

$$\beta(t) = 0.0006c^\star - 0.0014c^\star \sin(\omega t) + 0.0042c^\star \cos(\omega t) - 0.0161c^\star \sin(2\omega t) + 0.0028c^\star \cos(2\omega t)$$

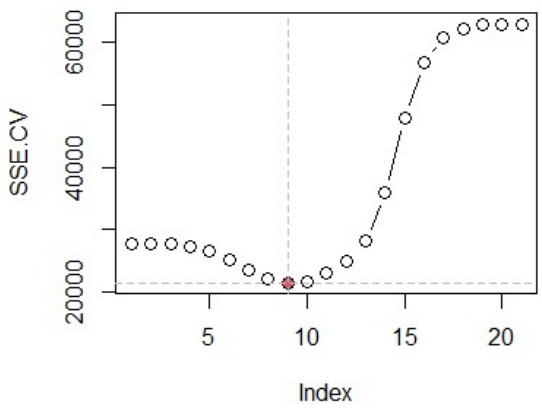
Where P is the period and $\omega = 2\pi/P$, in fda package it creates scaled basis function with $c^\star = \sqrt{\frac{2}{P}}$. And the fitted $\beta(t)$ is in Figure 3.1, it is Since $R^2 = 1 - \frac{SSE}{SST}$, by calculation, $R^2 \approx 0.999227$, which is a very high ratio.

Second Estimate

The second is to estimate using a roughness penalty. This is very much like smoothing. If $x_i(t)$ are represented by a basis, using the same basis often works well. The first step is to set up a harmonic acceleration operator $Lx = \omega^2 Dx + D^3 x$, then replace our previous choice of basis for beta estimate by a functional parameter, with Fourier basis function of 27 basis functions. Figure 3.2 is the $\beta(t)$ estimated with smoothing parameter $\lambda = 100$. Figure 3.3 has $\lambda = 10^{12.5}$, which returns a very smooth curve of β. The optimal choice of smoothing parameter λ could be derived through cross-validation.

As in Figure 4.1, when $\lambda = 10^4$, it achieves the minimal SSE, which equals to 21331.21. If we further taken a fine grid for λ for logarithmic scale between 3.5 to 4.5, a more precise $\lambda^\star = 11748.98$, for this plot please refer to Appendix B.3,4.

(i)



(ii)

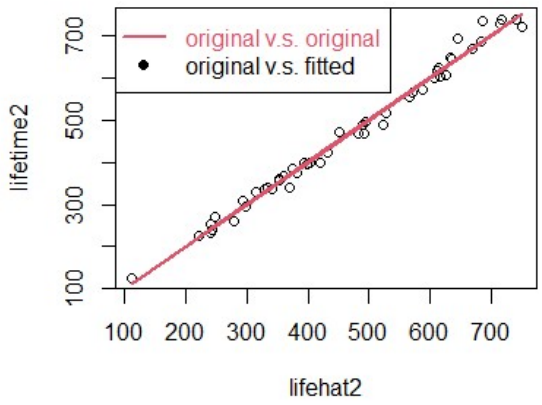


Figure 4: **(i)** Cross-validation to find optimal λ from 1 to 10^10, with logarithmic interval of 0.5. The optimal lambda is at $\lambda = 10^4$, SSE = 21331.21. **(ii)** has x-axis of fitted lifetime and y-axis of original lifetime, with the red line as original data points.

In Figure 4.2, it plotted the scatter plot of points with x-axis as fitted lifetime, and y-axis as original lifetime. It is very similar to the line of original data v.s. original data. Which coincides with the high R^2 achieved.

With the constant term of -0.4507, the output of the result is :

$$\begin{aligned}\beta(t) &= 2.0865c^\star + 2.3143c^\star \sin(\omega t) \\ &+ 0.5355^\star \cos(\omega t) + 0.0028c^\star \sin(2\omega t) \\ &+ 0.0209c^\star \cos(2\omega t)\end{aligned}$$

INFERENCES AND CONFIDENCE INTERVAL

After the optimal number of λ is chosen for regressions with smoothing penalties, we are able to calculate the R^2 and confidence interval for regression. In this model, $R^2 = 0.989453$, and SSE = 21320.23. We can also calculate confidence interval and also obtain an estimate of σ^2, where $\hat{\sigma}_e^2 = \text{SSE}/(n-df) = 313.894$, and the intervals are $\Phi(t)\hat{c} \pm 2\sqrt{\Phi(t)^T \text{Var}[\hat{c}]\Phi(t)}$, are plotted in dashed lines (Figure 6).

The intuitive provided by the estimated curve is that, before 17 days, egg counts have positive effect to lifetime. Especially for time between the 0 to 10 days. After 17 days, having more eggs has negative effect on the lifetime of fruit flies, especially between 20 and 23 days. E.g., for fruit fly 155 as I previously mentioned, it has very high PC2, which means it has high egg counts at beginning versus at

(i) **(ii)** **(iii)**

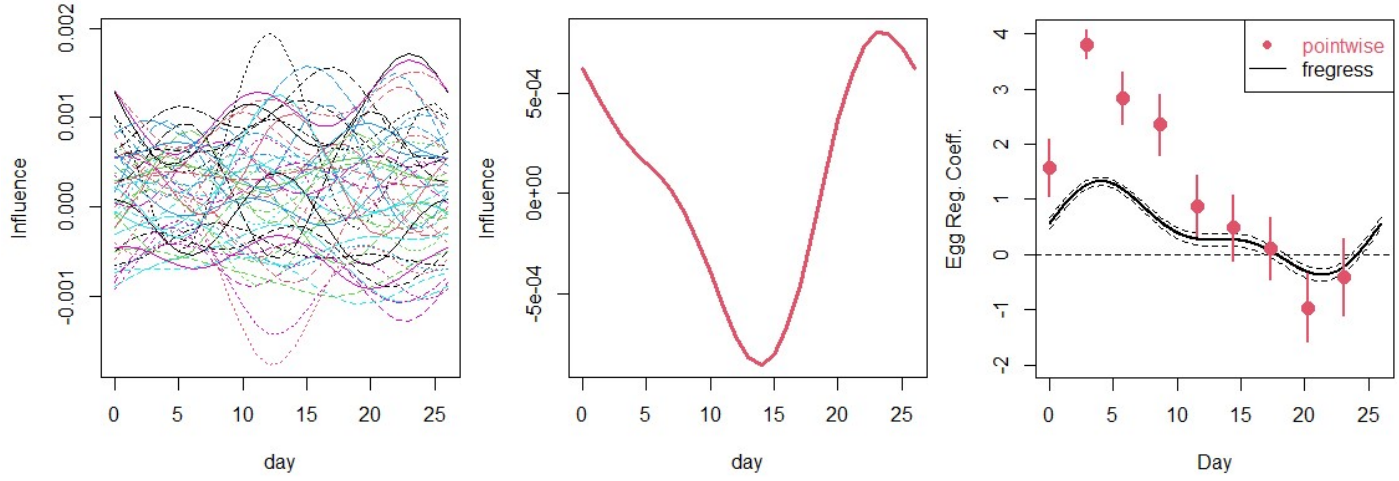


Figure 5: **(i)** is the influence of each observation on β. **(ii)** is the influence of fly 155 on the estimation of β, it has positive influence to β except for approximately 10 to 20 days. **(iii)** plotted the result by fRegress with smoothing parameter in black ($R^2 = 0.9895$), and pointwise estimation by red ($R^2 = 0.9873$).

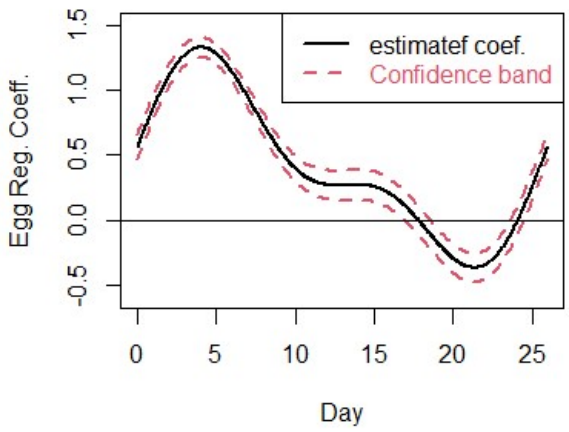


Figure 6: 95% Confidence interval. The variation of coefficient of egg counts is in black, with red dashed lines as upper and lower interval.

the end. Its' life time is 726, where the mean life time is 466.26, it has life time larger than 92% flies. The β curve has larger variance when coefficients change mildly.

Influence

We would also be able to study the influence of the i-th observation on the estimation of β since $\widetilde{\beta}(t) = \left[0\Phi(t)^T\right]\left(Z^TZ+\lambda R\right)^{-1}Z^T\mathrm{y} = \sum_{i=1}^{n} b_i(t)y_i$. Then, tries to plot the effect of each of the 50 observations. Take the influence of fly 155 as example, it has a very chaotic dispersion. But for each fly, say fly 155 in this case (Figure 5.2), it has has positive influence to $\beta(t)$, except for for approximately 10 to 20 days.

3.7.4 Comparison of Regression Results

We've already applied point-wise linear regression (M_1), functional PCR (M_2), low-dimensional functional regression (M_3), functional regression with smoothing parameter (M_4). In Figure 5.3, the fitted coefficients for functional regression with smoothing parameter and point-wise regression is compared. The functional regression curve is more flat and has smaller variance, while point-wise curve is fluctuated and has larger variance and wider interval.

Table 4: R^2 of 4 Regression Methods. Point-wise linear regression (M_1), functional PCR (M_2), low-dimensional functional regression (M_3), functional regression with smoothing parameter (M_4) are summarized. With highest R^2 in bold.

	M_1	M_1	M_1	M_1
R^2 (%)	0.9873	0.9844	**0.9992**	0.9895

From the R^2 statistics, the low-dimensional functional regression performs the best. However, functional regression introducing smoothing parameter also performs well, it is likely if choose the value with fine grids carefully, it will have a higher R^2. Thus, both estimates of functional regression have good performance (Table 4).

4. Conclusion

In this report, two main parts about the fruit flies' data are done: to increase better interpretation, PCA, fPCA, PLS, dynamic FPCA are done, dynamic fPCA achieves the best result by having more assumptions on the property of time-series. 4 PCs accounting for 99.94% of variance are selected, provides information of different interval of days. Then, regressions are made between lifetime and egg count of fruit flies, both of the regressions provide very high R^2, especially the first estimate by functional regression function with constant α and β with 5 Fourier basis. The second estimate through introducing smoothing also performs well, it's performance might be restricted to the selection of λ by logarithmic scale of 0.01. Interpretations are made as much as possible as in the table, but there could be better explanation or any possible ones. It would still be better to have test and train datas to test the validility of the model.

Appendix

Appendix A

(1) The Mean Curve of 50 flies with Fourier Basis

Day	Value	Day	Value	Day	Value	Day	Value
1	191.60154	9	-0.0662	17	-2e-05	25	-1e-04
2	35.06892	10	0.00459	18	0.00046	26	-0.00015
3	-20.53979	11	-0.00847	19	0.00022	27	-0.00032
4	4.12801	12	-0.00084	20	0.00016		
5	-10.61063	13	0.00224	21	-0.00025		
6	0.42746	14	0.00147	22	1e-05		
7	-0.64576	15	-0.00027	23	4e-05		
8	-0.00557	16	0.00031	24	7e-05		

(2) First 4 fPCs (accounting for 98.54% **of variation)**

Basis Function	PC1	PC2	PC3	PC4
const	0.9471099	0.0387337	0.3185592	-0.000325
sin 1	0.1363011	0.8501058	-0.5086207	-0.0063916
cos 1	-0.2905185	0.5251146	0.7998888	-0.0056076
sin 2	-9.56e-05	0.0086178	0.0011255	0.9822954
cos 2	-0.0019297	-0.0004611	0.0012384	0.1840667
sin 3	-3.01e-05	0.0003027	0.0001357	0.0334252
cos 3	-3.35e-05	-9.09e-05	-2.45e-05	-0.0010854
sin 4	-2.6e-06	3.86e-05	3e-06	0.0039304
cos 4	-5.1e-06	-1.16e-05	2.9e-06	-0.0028377
sin 5	1e-06	3.1e-06	-1.7e-06	0.0005716
cos 5	-8e-07	-7.2e-06	-2e-07	-0.0007592
sin 6	2e-07	9e-07	4e-07	7.24e-05
cos 6	-1e-07	-1.9e-06	7e-07	-0.0001063
sin 7	-1e-07	4e-07	-1e-07	6.65e-05
cos 7	2e-07	-4e-07	1e-07	-3.63e-05
sin 8	1e-07	1e-07	1e-07	-2.56e-05
cos 8	0	-2e-07	1e-07	-3.74e-05
sin 9	0	2e-07	1e-07	1.73e-05
cos 9	1e-07	-1e-07	1e-07	2e-06
sin 10	0	0	1e-07	9e-07
cos 10	0	-1e-07	1e-07	-4.2e-06
sin 11	0	0	0	-4e-07
cos 11	0	-1e-07	0	-4.6e-06
sin 12	0	-1e-07	0	-4.7e-06
cos 12	0	0	0	-4.9e-06
sin 13	0	-1e-07	0	-1.15e-05
cos 13	0	0	0	-7e-07

(3) Point-wise Linear Regression Coefficients

Basis Function	Estimate Std.	Error	t value	Pr(>|t|)
Intercept	-14.5537	11.1398	-1.306	0.19886
t(eggvals)1	1.5665	0.2569	6.098	3.45e-07 ***
t(eggvals)2	3.8118	0.1305	29.202	<2e-16 ***
t(eggvals)3	2.8401	0.2360	12.037	7.10e-15 ***
t(eggvals)4	2.3593	0.2752	8.572	1.35e-10 ***
t(eggvals)5	0.8838	0.2741	3.225	0.00251 **
t(eggvals)6	0.4873	0.2987	1.631	0.11066
t(eggvals)7	0.1081	0.2793	0.387	0.70082
t(eggvals)8	-0.9637	0.3086	-3.122	0.00333 **
t(eggvals)9	-0.4069	0.3519	-1.156	0.25448
t(eggvals)10	NA	NA	NA	NA

(4) Functional PCR Coefficients

Basis Function	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	466.26000	3.01494	154.650	<2e-16 ***
as.matrix(eggpc)V1	2.11597	0.05020	42.150	<2e-16 ***
as.matrix(eggpc)V2	2.94422	0.09116	32.298	<2e-16 ***
as.matrix(eggpc)V3	-0.05287	0.13364	-0.396	0.694251
as.matrix(eggpc)V4	0.99858	0.24885	4.013	0.000224 ***

(5) PLS Coefficients (account for 99.54% **of variability)**

Lifetime	1^{st} Comp	2^{nd} Comp	3^{rd} Comp	4^{th} Comp
V13	0.8377	1.1268	1.0109	1.0304
V14	0.7797	1.1044	1.1441	1.2862
V15	1.0550	1.2056	1.0147	1.1634
V16	0.6960	0.9494	1.1737	0.9826
V17	0.7542	0.7337	0.9251	0.9879
V18	0.7083	0.8111	1.0905	1.0815
V19	0.7115	0.7208	0.9956	0.9520
V20	0.6866	0.6743	0.8569	0.8346
V21	0.5670	0.5729	0.7913	1.0636
V22	0.4747	0.4053	0.6466	0.8128
V23	0.4261	0.2498	0.1573	-0.1316
V24	0.5041	0.4180	0.2350	0.2150
V25	0.3025	0.2013	0.2185	0.2292
V26	0.4589	0.3372	0.3490	0.2638
V27	0.4397	0.3385	0.0275	-0.1178
V28	0.3728	0.2614	0.0648	-0.0147
V29	0.3317	0.1656	-0.0151	-0.0139
V30	0.4252	0.3264	0.0719	0.1073
V31	0.3085	0.1757	0.0575	0.0870
V32	0.0964	-0.0172	-0.1888	-0.3120
V33	0.2835	0.1410	-0.0234	-0.0704
V34	-0.0191	-0.1681	-0.1924	-0.1262
V35	0.1871	0.0427	-0.0027	-0.0677
V36	-0.0614	-0.1785	-0.0502	0.2070
V37	-0.0635	-0.1019	0.1385	0.2830
V38	-0.2277	-0.3816	-0.2446	-0.0524

Appendix B

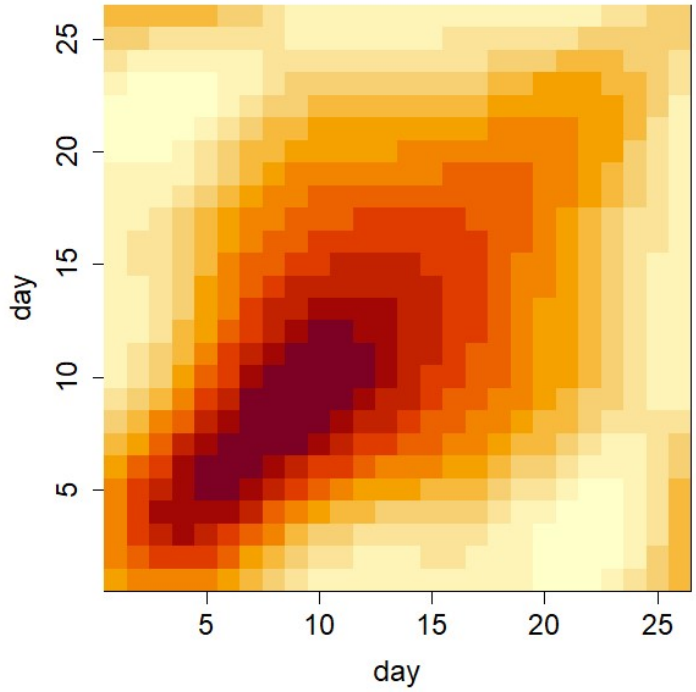


Figure 1: Image Plot with heated regions indicating high covariance and cold regions with less covariance.

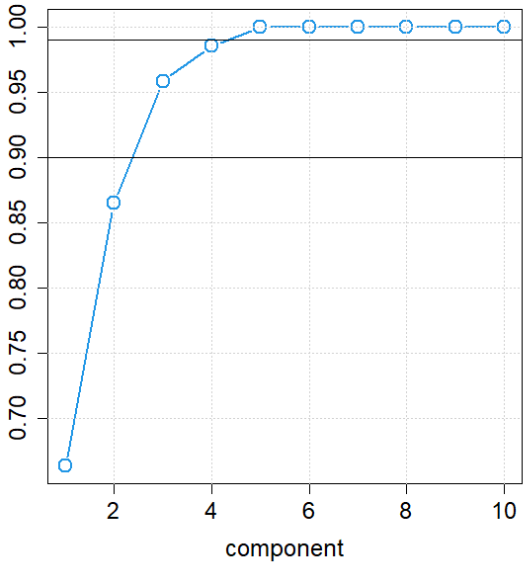


Figure 2: Cumulative Plot of of proportion of variance for fPCs. The variance at 90% requires 2 components, while 99% requires 4 fPCs.

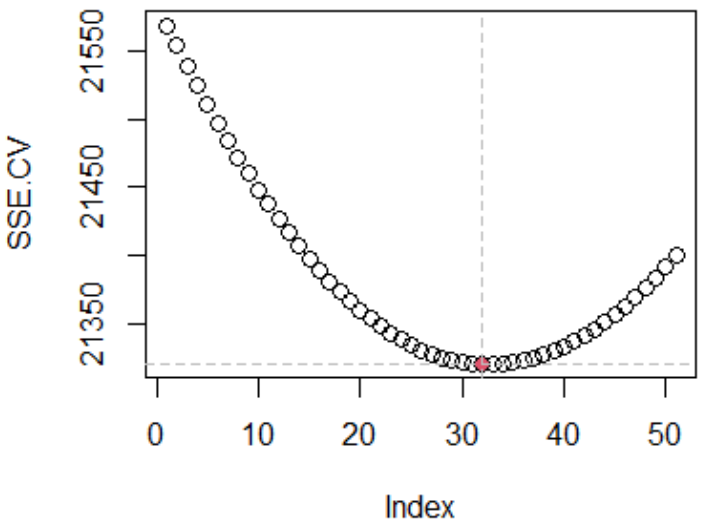


Figure 3: Fine grid of cross validation for smoothing parameter λ. From 3.75 to 4.25, use the logarithmic scale of 0.01, when $\lambda^\star = 11748.98$, the cross validation has smallest scores.

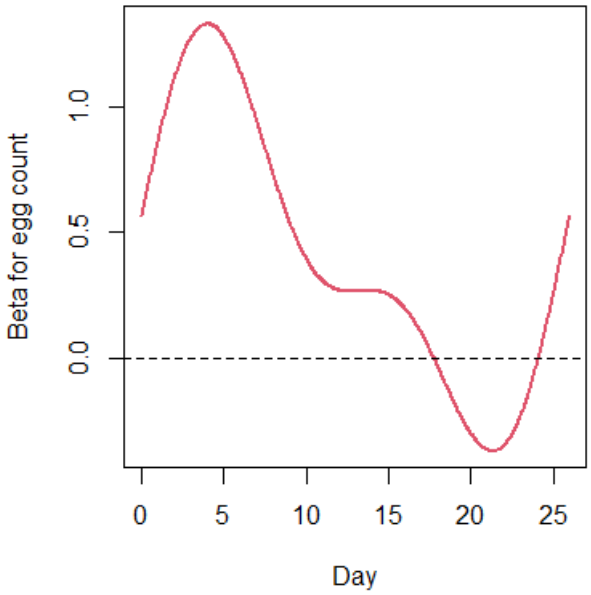


Figure 4: Beta estimated with the optimal λ selected as $\lambda^\star$. It is very similar to β estimated through the low-dimensional functional regression.