Contents

1. Introduction

1.1 Aim of the Report

Functional data analysis is the analysis of data that are functions. In this report, the analyzed domains are time. The data describes a process that changes continuously over time. In real world, data may be measured more noisy. The purpose of this report is to find the smooth process under the data, while still depicting the trend of variation.

Two data sets are covered in this report: Canadian weather data and fruit fly data. With functional data analysis as a powerful tool, this report is going to show how different types and different number of basis functions fit the data differently. The report seek to output a smooth curve. The number of basis function by using proper methods, or how regression is conducted would be focused on.

1.2 Problems Targeted

The main problem the report focused on is to figure out the functional variable behind the functional data. By least square estimation, fitting the curves would not be a problem. But to evaluate how good is the estimation of the data could be a problem, since there are a number of statistics and skills. For B-spline basis, the selection of knots would also be difficult and worthy of study.

Another way of outputting a smooth curve fitting the trend of the data is to introduce penalty of smoothness. The third problem is occurred when considering the smoothness penalty, the selection of smoothing parameter λ and linear differential operator would affect the simulation results. Thus, proper statistics, or cross-validation would also be necessary in this step.

1.3 Methods Involved

Some important methods in this report involved for both the Canadian weather and the fruit fly data include the following:

1. Choice of basis: Fourier basis, B-spline basis,$\cdots$
2. Fitting: least square estimation (LSE)
3. Choosing no. of basis: integrated mean squared error (IMSE), cross-validation (CV)
4. Using smoothness: second derivative, harmonic acceleration penalty, $\cdots$
5. Find the optimal λ: generalized cross validation (GCV)

2. Models and Inference

In this section, Fourier basis function and B-spline basis is applied. Then, MSE and cross-validation is applied to select the optimal number of basis function. For B-spline basis function, smoothness penalty is introduced, providing smooth curve fitting the data. Procedures and theoretical backgrounds in details would be discussed.

2.1 Problem Reviewed

The models used in this report are mainly Fourier basis function and B-spline basis function. Thus, after creating the number of basis, the problems we face at is to acquire smooth curve fitting the data. To determine the number of basis function, we could use MSE or cross-validation. If we were to introduce the smoothing penalties, it is necessary to find a proper structure of penalty and the value of smoothing parameter λ.

2.2 Methodology

2.2.1 Simulating Different Number of Basis Function

Firstly, create the Fourier basis function with a specific number, then calculated each value a by **create.fourier.basis()** in R evaluated at each point. The Fourier basis are sine and cosine functions of increasing frequency:

$$1, sin(\omega t), sin(\omega t), sin(2\omega t), cos(2\omega t), \cdots, sin(m), cos(m\omega t), \cdots$$

Where $\omega = 2\pi/P$ with P as the period. In R, the **fda** package, the basis functions are scaled, e.g. the second term is indeed:

$$\sqrt{\frac{2}{P}} sin(\frac{n\pi t}{L})$$

Secondly, every Fourier basis function is calculated at each time point from 1 to 365. The second order derivative is also calculated for each basis function through **Lfdobj=2**. (Notice that in this step, the Fourier basis functions are scaled by multiplying $\sqrt{\frac{2}{P}}$).

Thirdly, use specific number of basis function to solve the simulated value $\hat{y}$ by least square estimation, and the second derivative value $\hat{y}''$. Then, plot the original scatter points and simulated curves, as well as the simulated curve of the second order derivative. For 365 time points, attempts could be made for 3, 5, 7, 13, 19, 25, 31, 41, 53, 105, 209, and 365 number of basis.

The splines are monomial segments joined end-to-end. Number of basis functions:

$$\text{order} + \text{number interior knots}$$

Fourthly, a bias-variance simulation is conducted. The simulation started by fitting the data with B-splines, and set the simulated value $x(t_i)$ as the true value (Notice that last few terms of $\hat{y}$ are usually modified to make the curve more smooth near the boundary). Then, error terms are calculated through $\epsilon_i = y_i - x(t_i)$. New data is created by re-arranging the

error terms randomly: $y_i^\star = x(t_i) + \epsilon_{i\star}$. This procedure is repeated for several times (10, 100 or 1000) for each number (if we set it from 3 to 101, we have 50 in total) of Fourier basis. Variance, bias and mean square error(MSE) are calculated. In this case, minimizing integrated mean squared error (IMSE) was used for evaluation:

$$\text{IMSE}[\hat{x}(t)] = \int \text{MSE}[\hat{x}(t)]$$

Where $MSE[\hat{x}(t)] = \text{Bias}^2[\hat{x}(t)] + \text{Var}[\hat{x}(t)]$. The bias of the estimate of $x(t)$: $\text{Bias}[x(t)] = x(t) - E(\hat{x}(t))$. The sampling variance of the estimate: $\text{Var}[\hat{x}(t)] = E[\{\hat{x}(t) - E\hat{x}(t)\}^2]$.

2.2.2 Inference, Estimation and Prediction

For a Single Curve

When fitting the data for a single curve, the procedures in Practical ch2b simply used least square estimation (LSE). For a single curve that we observe:

$$y_i = x(t_i) + \epsilon,$$

and we want to estimate:

$$x(t) \approx \sum_{j=1}^{k} c_j \phi_j(t),$$

k is the number of the basis function. If we assume that the residuals are independent, the least square estimation is calculated when minimizing the sum of squared errors:

$$SSE = \sum_{i=1}^{n} (y_i - x(t_i))^2 = \sum_{i=1}^{n} (y_i - c^T \phi(t_i))^2$$

When using the cross-validation to choose the number of basis, for different number of basis function, firstly calculate the matrix S for linear smooth $\hat{y} = Sy$. For a linear smooth $\hat{y} = Sy$,

$$OCV[\hat{x}] = \sum \frac{(y_i - \hat{x}(t_i))^2}{(1 - s_{ii})^2}$$

Secondly, calculate the ordinary cross-validation score and plot out. Number of basis function was chosen to be the minimal number of basis function that reaches a relatively small variance estimate.

Point-wise confidence bands were calculated after choosing the number of basis function. For each point we would be able to calculate the lower and upper bands for $\hat{y}(t)$ by calculating the variance of $\hat{y}(t)$ first. The estimation of σ^2 is calculated through MSSE (mean sum squared error in validation set). The lower and upper bands for $\hat{y}(t)$ is:

$$\hat{y}(t) \pm \sqrt{Var[\hat{y}(t)]}$$

Where $Var(\hat{y}) = \hat{\sigma}^2 SS^T$, $\hat{\sigma}^2 = \frac{1}{N-K}\text{MSSE}$.

For a Matrix

When we have a N by K matrix as the fruit fly data, we could derive the linear regression of the matrix form. If matrix Φ contains the values $\phi_k(t_j)$, and the real data y is the vector $(y_1, \cdots, y_N)$. We have:

$$SSE(c) = (y - \Phi c)^T (y - \Phi c)$$

The error sum of squares is minimized by the ordinary least square estimate

$$\hat{c} = (\Phi^T \Phi)^{-1} \Phi^T y$$

The we have the estimate:

$$\hat{y} = \phi(t)\hat{c} = \phi(t)(\Phi^T \Phi)^{-1} \Phi^T y$$

Instead of $Var(\hat{y}) = \hat{\sigma}^2 SS^T$ in the previous sessions, the variance of y of the matrix form is:

$$Var[\hat{y}] = \Phi C \Sigma C^T \Phi^T$$

Where $C = (\Phi^T \Phi)^{-1} \Phi^T$, $\hat{y} = c\Phi$, $c = Cy$. Then, confidence band could be calculated as in the previous sessions.

2.2.3 Using Smoothing Penalties

Choosing the number of basis function could be discrete and therefore cause the additional variability. For basis function like B-splines, not only the number, but the location of knots would influence smoothness. Thus, define smoothness explicitly and calculate the penalized fit would be a better alternative.

When fitting the data, we have two desires, to fit the data and maintain smoothness. There are many choices of linear differential operator.

$$D^m x + b_{m-1}(t) D^{m-1} x + \ldots + b_0(t) x$$

The penalized squared error is defined as:

$$\text{PENSSE}_L(x) = (y - x(t))^T (y - x(t)) + \lambda \int [Lx(t)]^2 dt$$

Wwhere λ is a smoothing parameter measuring compromise between fit and smoothness, and $Lx(t)$ has many choices.

For instance, he usual penalized squared error is:

$$\text{PENSSE}_\lambda(x) = [\text{y} - x(\text{t})]^T [\text{y} - x(\text{t})] + \lambda \int \left[D^2 x(t)\right]^2 dt$$

Harmonic acceleration of x is:

$$Lx = \omega^2 Dx + D^3 x$$

With the restriction that $L\cos(\omega t) = 0 = L\sin(\omega t)$.

The calculation of penalized fit, for $x(t) = \phi(t)c$, the penalized least squares estimate for c is:

$$\hat{\mathrm{c}} = \left[\mathbf{\Phi}^T W \mathbf{\Phi} + \lambda R\right]^{-1} \mathbf{\Phi}^T W \mathrm{y}$$

It is still a linear smoother:

$$\hat{\mathrm{y}} = \mathbf{\Phi}\left[\mathbf{\Phi}^T W \mathbf{\Phi} + \lambda R\right]^{-1} \mathbf{\Phi}^T W \mathrm{y}$$

The generalized cross validation is discounted for degrees of freedom and λ. It smoothes more than OCV. To choose the smoothing parameter λ, generalized cross validation (GCV) is applied:

$$\mathrm{GCV}(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{\mathrm{SSE}}{n - df(\lambda)}\right)$$

The smoothing spline is also a linear operator

$$\hat{x}(t) = \Phi(t)\left[\mathbf{\Phi}^T W \mathbf{\Phi} + \lambda R\right]^{-1} \mathbf{\Phi}^T W \mathbf{y}$$

Then, linear smooths and linear probes provides a means pf providing the confidence interval estimation for features of x. A confidence interval for linear functionals of x $N[x]$ can be given by:

$$N[\hat{x}] \pm 2\sqrt{N[\Phi] C \Sigma C^T N[\Phi]^T}$$

3. Numerical results

3.1 Practical Ch2b Simulation

In Practical Ch2b, 365 consecutive days' precipitation data of Vancouver is used. After fitting the data with a number of basis function, a true model defined by B-spline is used for calculating the bias, variance and mean squared error. Then, ordinary cross-validation is applied for determining the optimal number of basis. Some results and visualization are provided in this section.

3.1.1 Description of the Data

The precipitation (mm) data of Vancouver has 365 time points, with precipitation ranging from 0 to 8.8 (mm), median is 2.8 (mm). Vancouver is Canada's third most rainy city. As people see it, Vancouver has less precipitation during the summer, but more precipitation (or heavy snow) in the winter.

3.1.2 Creating Basis Function

Firstly, 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 $\sqrt{\frac{2}{P}} \approx 0.07402332$. Then, calculate $\hat{y}$ and $\hat{y}''$ through least square estimation (LSE).

In Figure 1,2,3, The blue lines are the fitted curve, red points are real data. Graphs of 7, 13, and 53 number of Fourier basis functions are shown as follow. Figure 2 has 13 basis functions fitting the trend appropriately, while maintaining smoothness to some degree. Figure 1 fit the points poorly (underfit), and Figure 3 is too oscillating, having large sampling variance. The 2^{nd} derivative plots of the basis function are also shown.

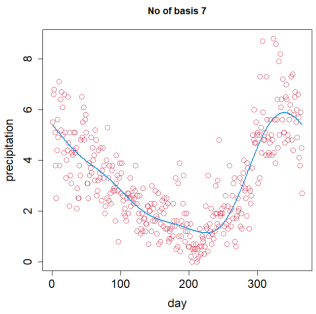


Figure 1: 7 basis fitting

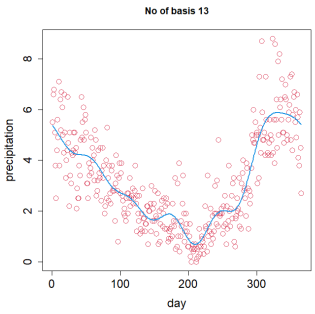


Figure 2: 13 basis fitting

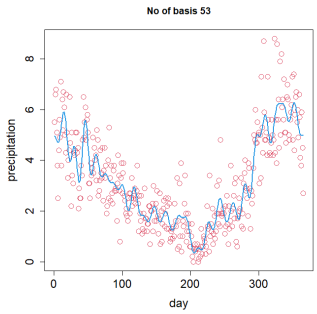


Figure 3: 53 basis fitting

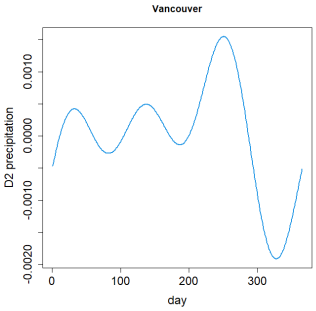


Figure 4: 7 basis D_2

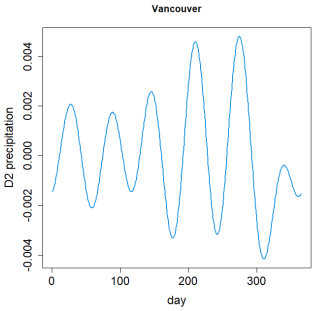


Figure 5: 13 basis D_2

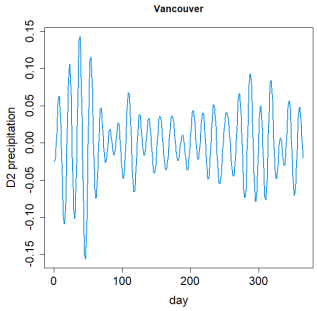


Figure 6: 53 basis D_2

In Figure 4, the second derivative curve oscillates slowly around 0, indicating it is very smooth. Figure 6 oscillates very frequently, possibly that it is not smooth. While Figure 5 is between those two.

3.1.3 Inference Based on True Model

When doing the bias-variance trade-off, in Ch2b the first step is to create a B-spline function. The knots of B-spline is selected to be the first day of every month. There are 11 interior knots, order of splines are set to be 6. Through least square estimation, $\hat{y}$ is calculated.

A special step in Ch2b is that, after simulation of $\hat{y}$, the last 15 terms (from 351 to 365) are modified through $\hat{y}_{350+i} = \hat{y}_{350} + \frac{i}{16}(\hat{y}_1 - \hat{y}_{350})$, $i \in (1, 2, \cdots, 15)$. Values after modification decreases more slowly than original values. This is possibly in consideration of smoothing the curve near the right end of the data.

$\hat{y}_{360:365}$ estimated	$\hat{y}_{360:365}$ modified
5.227 4.995 4.724 4.410 4.047 3.631,	5.692 5.651 5.610 5.568 5.527 5.485

Table 1: Comparison of the last 6 terms of $\hat{y}$ (2 decimals)

Errors are calculated: $\epsilon_i = y - \hat{y}$ for the modified new y_i. Then, randomizing the error terms to create the "true model". This procedure is repeated for 10 times. Then, least square estimation are made for Fourier basis function based on "true model". Each time, variance, bias and mean squared error (MSE) are calculated for 50 different number of basis function (3 to 101). Figure 7 plotted the bias, variance and MSE of all number of basis function (3 to 101). In Figure 8, it only plotted the number of basis from 5 to 21 (5, 7, 9, 11, 13, 17, 19, 21). Variance is always increasing as the number of basis increasing. Bias in decreasing while number of basis functions decreasing. MSE firstly decreases, reach an "elbow" point, then increases. The "elbow" point seemed to be around 13 number of basis functions.

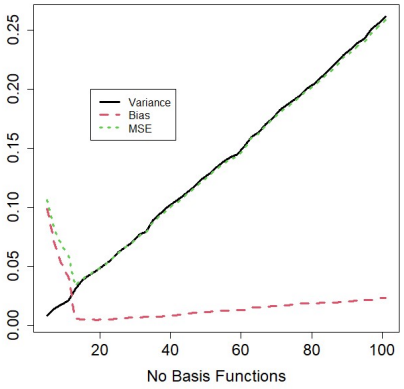


Figure 7: Var, bias and MSE of 3 to 101 number of basis

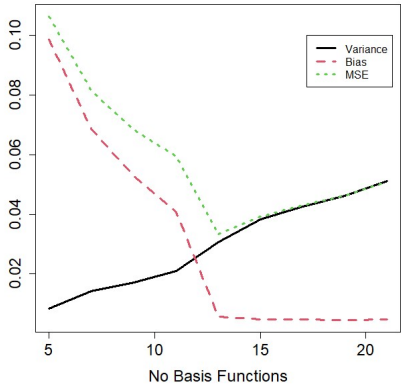


Figure 8: Var, bias and MSE of 3 to 21 number of basis

Then, the next method determining the number of basis function is through cross-validation. By calculating the S matrix of a linear smooth $\hat{y} = Sy$, the ordinary cross validation score (OCV) is then acquired. The cross validation score from 5 to 41 number of basis is plotted in the graph. The choice of number of basis could be selected as 13, the minimal number of basis that variance estimate reaches a turning point (SMSSE = 372.5140).

After 13 number of Fourier basis is chosen, we could be able to estimate the point-wise confidence interval (95%), where $Var(\hat{y}) = 0.3712382$. The 95% confidence interval fit the trend of scatter points well, indicating that our choice of Fourier basis is reasonable.

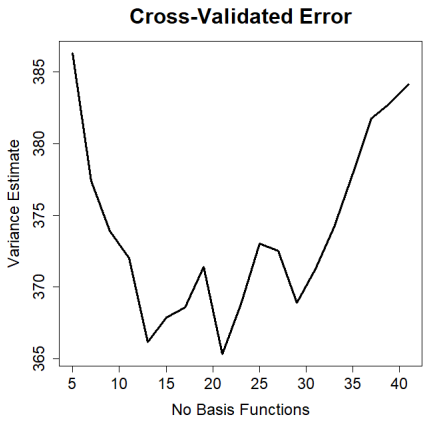


Figure 9: SMSSE of cross-validation

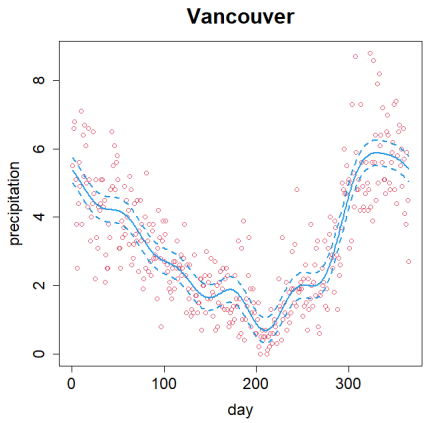


Figure 10: Point-wise confidence interval

3.1.4 USING SMOOTHING PENALTIES

The another way of acquiring a smooth curve, other than choosing the optimal number of basis function, is to introduce smoothing penalty for estimation. After applying the B-spline basis function with 11 interior points (one for each month) of order 6. In CH2b, it started with least square smooth, then plotted the curve of 2^{nd} order differential operator object. Additionally, the derivative of the harmonic acceleration is calculated.

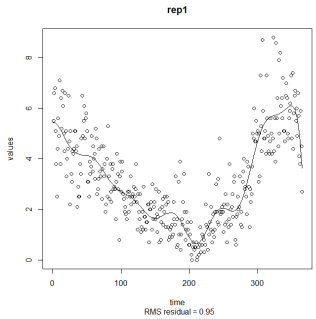


Figure 11: Least square smooth

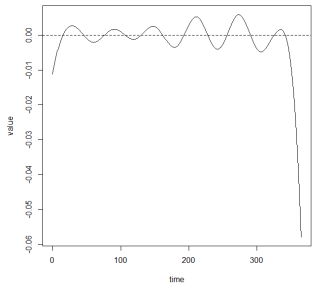


Figure 12: 2^{nd} order derivative

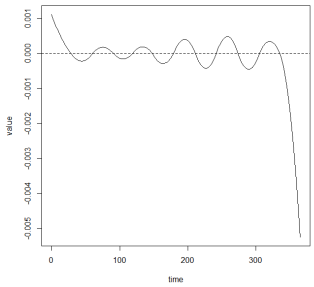


Figure 13: Harmonic acceleration

The harmonic acceleration penalty is given by:

$$Lx = D^3x + \omega^2 Dx$$

For the Canadian Weather data, $\omega = 2\pi/365$. The plot is shown in Figure 13.

Question: Can you represent D2lfd using the vec2Lfd command?

Answer: Yes, through D2veclfd= vec2Lfd(c(0,0), c(0, 365)).

Evaluating the Lfd object helps us to identify how 'smoothness' is being violated. Near the boundary, both derivatives are far from the value = 0 line. This indicates that near the

boundaries, curves are not smooth. Calculating the Lfd object is more useful in examining collections of functions. Suppose we create the functional data of average daily Canadian weather of B-spline basis function, with no smoothness penalty at first. In Figure 14, the graph is the daily average weather of 35 cities fitted by B-spline (11 interior knots and order 6). The second derivative of the lines are plotted in Figure 15, indicating the roughness near boundaries. Then, looking at Figure 16, the derivative of harmonic acceleration is oscillating around 0 for points in the middle 50 to 300, indicating the smoothness of the B-spline basis fitting in the middle. This is the smoothness effect of basis expansion, the next step is to see how introducing smoothness penalty improve the smoothness of fitting.

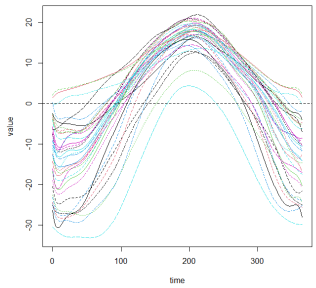


Figure 14: Least square smooth

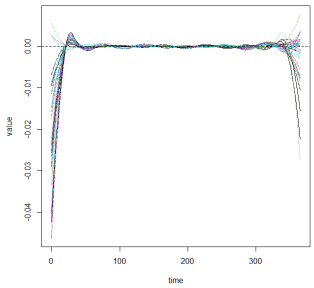


Figure 15: 2^{nd} order derivative

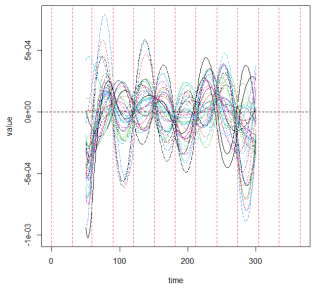


Figure 16: Harmonic acceleration

Firstly, create a saturated B-spline basis ($k = 367$), with order of 4. The fdPar function in package fda defines a functional parameter object, including saturated model, linear differential operator and smoothing parameter λ. When fdPar objects are put in the smooth.basis function, it introduces the smoothing penalty to calculate PENSSE_λ for estimation.

Question: does the harmonic acceleration penalty work for the B-spline basis above? How can you rectify this if not?

Answer: The harmonic acceleration penalty matrix cannot be evaluated for derivative of order 3 for B-splines of order 4.

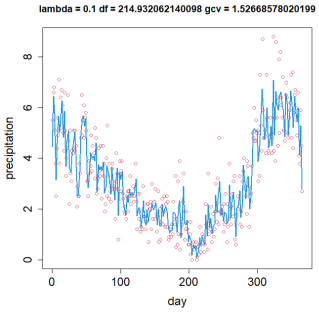


Figure 17: λ=0.1

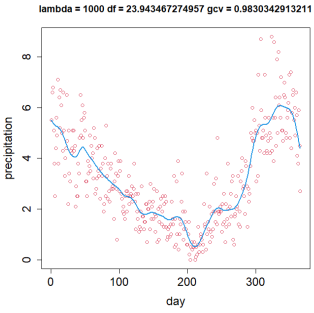


Figure 18: λ=1000

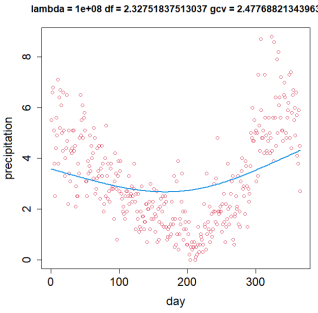


Figure 19: λ=10^8

Then, we would be able to examine how generalized cross validation (GCV), degrees of freedom (DF) and squared estimate of errors (SSE) varies with λ. This is shown in Figure 17, Figure 18 and Figure 19. The values of λ varies on logarithmic scale. In least square fitting, the degrees of freedom used to smooth the data is exactly the number of basis function k. As λ increases, the degrees of freedom decreases, while generalized cross validation score firstly decreases, then increases as λ increases. When $\lambda = 1000$, the generalized cross validation (GCV) score reaches a relatively low value.

3.1.5 FINDINGS

1. Not only finding optimal basis function would return a smooth fitting, introducing smoothness penalty would also be a useful method.

2. Compared with bias and variance, MSE would be more appropriately to determine the optimal number of basis function.

3. Introducing the smoothness penalty would be an effective method, especially for basis function B-spline, only determining the number of basis is not enough.

3.2 Fruit Fly Data

The fruit fly's data contains 50 flies and 26 time points, having several curves rather than just one. Thus, in this section, similar methods are applied when creating the basis function, choosing the number of basis function, and using smoothness penalty. The minor difference is that linear regression and inference could be done to an N by K matrix.

3.2.1 DESCRIPTION OF THE DATA SET

This is a data set of 50 flies' egg counts at 26 time points, the response variable is the eggcount of each fly. The time points vary from V13 to V38, in R language realization, was denoted as the 1^{st} day to the 26^{th} day. The response variable egg count ranges from 0 to 112, with the average of 37.58, median of 37, and 81 zeros.

3.2.2 PROBLEMS TO ADDRESS

In this section, the first step is still to create the basis function. But for the fruit fly data, I'm going to use monomial basis, Fourier basis, B-spline. After that, some inferences are made to determine optimal number of basis function through MSE and OCV, and confidence interval are calculated. After that, introducing smoothing penalties for estimation was also conducted. For summary, the methods in this sections are later compared with each other for fitting result, and some important findings are listed.

In later sessions, examples are mostly using the fly 512. For comparison, other flies may be used. For details, please refer to Appendix.

Fly Label	Summary of Statistics
512	Min=0, Max=55, Median=21, no. of zeros=2

Table 2: Summary of Egg Counts of Fruit Fly 512 (2 Decimals)

3.2.3 Creating Different Basis Function

In the functional data lectures, monomial basis, Fourier basis and B-spline basis are mainly discussed. Thus, fitting results of the above basis functions are provided.

Monomial Basis

For the basis function $\phi(t)$ in this sessions, the order of monomial is set to be 0 to 8. For a monomial of order 5, the basis function is based on the form of $\phi(t) = (1, t, t^2, t^3, t^4, t^5)$. In realization, when calculating each term, to prevent from the complexity of calculation, the values are all using the form $t^\star = \frac{t-13}{5}$.

It is also possible to output the form of the monomial: $\hat{y} = 158.79 - 9.29t^\star + 1.40t^{\star 2} - 2.0t^{\star 3} - 1.19t^{\star 4} + 0.70t^{\star 5}$. Larger terms would easily over-run smaller terms.

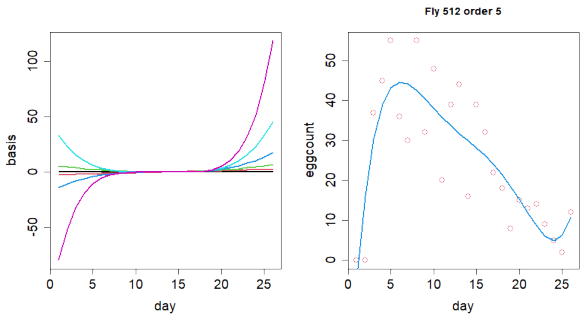


Figure 20: Basis functions and monomial fitting curves of order 5

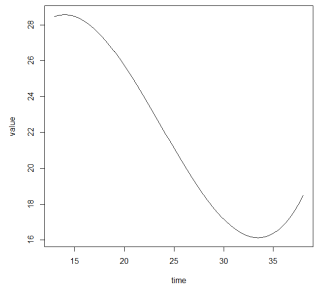


Figure 21: 2^{nd} derivative of order 5 Polynomial fitting

In Figure 20, the monomial fitting curve of order 5 is shown. For more orders from 1 to 9, see Appendix. However, the problem with the monomial basis is the poor performance in rates of change, the derivatives of the function does not fit well.

Fourier Basis

Considering the rates of change should also fit the data as well, then Fourier basis functions are generated. Basis functions of sines and cosine functions are created with increasing frequency. The constant L in front of basis function is $L = \frac{2}{P} \approx 0.277$, and $\omega = \frac{2\pi}{26} = 0.24$. Only 3 and 5 number of basis functions are shown in Figure 22 and Figure 23. For more details of Fourier basis function, please refer to the Appendix.

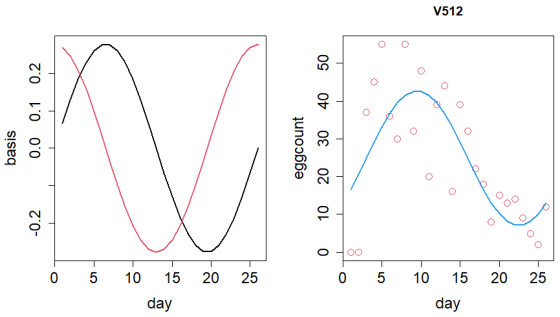


Figure 22: $\Phi(t) = (1, \sin(\omega t), \cos(\omega t))$

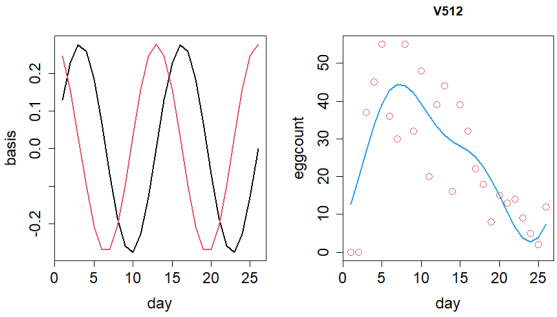
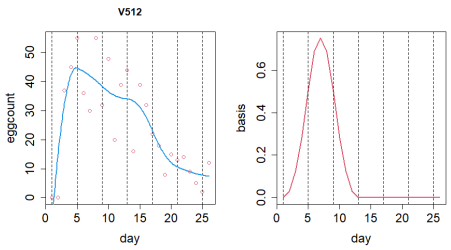


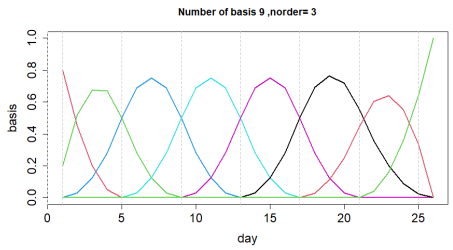
Figure 23: $\Phi(t) = (1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t))$

B-SPLINE BASIS

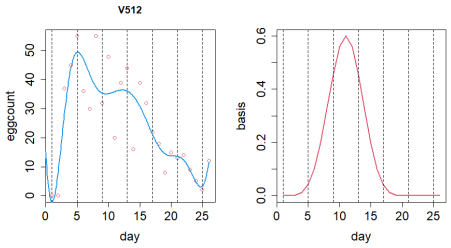
Now, we look at the spline basis. The choice of the interior knots are subjective at this moment. After examining the graph carefully, 6 interior points are chosen, since the first two terms are very far away from other points, the intervals of B-spline is chosen to be (0,1,5,9,13,17,21,26). The total number of basis function is norder + 6. For B-spline function, order 1 to 6 fitting result and one of the basis function are calculated and plotted.



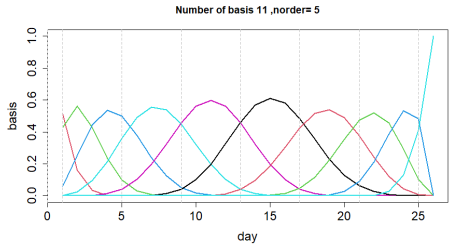
(i) Cubic Spline



(ii) All basis of Cubic Spline



(iii) Spline of order 5



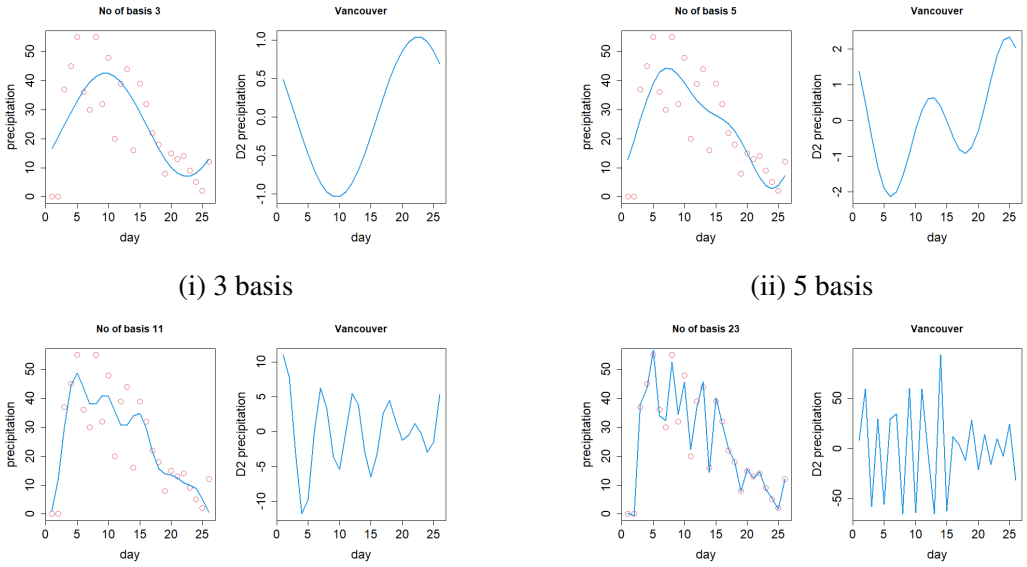
(iv) All basis of spline of order 5

Figure 24: Spline fitting and all of their basis

The cubic spline fits the egg count data well, with the smooth curve fitting the trend. While as comparison, the spline of 5 order already has the trend to overfit near the left boundary. For more please refer to the Appendix.

3.2.4 Inference Based on True Model

For the Fourier basis function, the goal in this section is to find out the optimal number of basis function. We would simulate all basis function, and by MSE or CV, make the final decision. In this section, examples are made for fly 512 first, but it is hard to get the optimal number through OCV. Thus, another fly labeled 453 was added for comparison.



(i) 3 basis

(ii) 5 basis

(iii) 11 basis

(iv) 23 basis

Figure 25: Fourier basis fitting and basis function

By examining the Fourier fitting and the 2^{nd} derivative of basis function, when the number of basis functions increases, the curve tend to be very rough, so is the second order derivative. More is provided in Appendix.

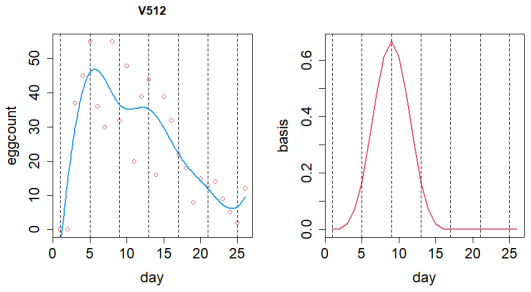


Figure 26: B-spline of order 4 previous interior knots

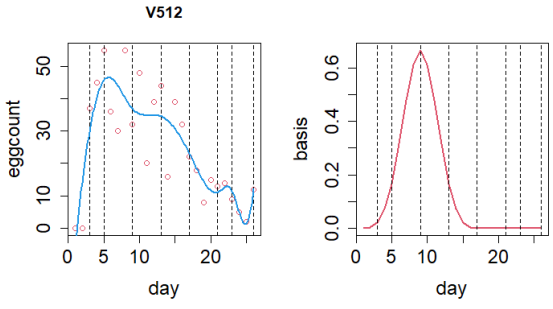


Figure 27: B-spline of order 4 adjusted interior knots

Now we do a bias-variance trade-off. The first step is to generate the "true model" that would be used for later variance and bias calculation. I used the B-spline basis function that generates in the previous sessions, with norder = 4 and rearranging 7 interior knots (3,5,9,13,17,21,23), adding more knots to the middle and subtracting the ones near the left boundary. It fitted better for the variation during the 18 to 23 days. The next special step is to fix the values near the right boundary to prevent from overfitting (24, 25 and 26 days). Errors are the real data subtracted by $\hat{y}^\star$. Then, the new data $y^\star$ is generated by rearranging the error terms to the real value. Finally, fit the new data using a Fourier basis and repeat 10 times to calculate bias and variance from the sample.

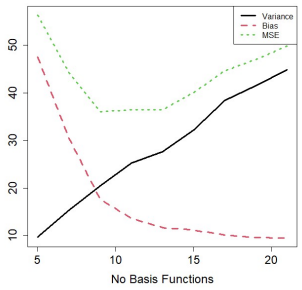


Figure 28: Var, Bias and MSE based on True Model

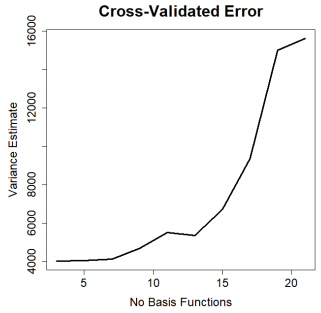


Figure 29: OCV score. It is always increasing

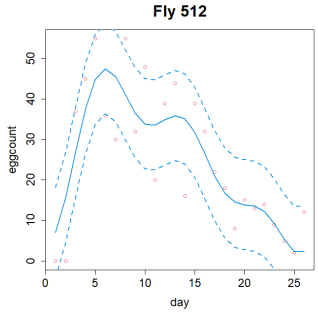


Figure 30: Pointwise Confidence Interval

The variance keeps increasing as the number of basis function increases, while bias keeps decreasing. The MSE firstly decreases, then increases. When the number of basis function equals to 7, MSE=36.04 reaches the minimum. Thus, the number of Fourier function equals to 7 is a possible appropriate decision.

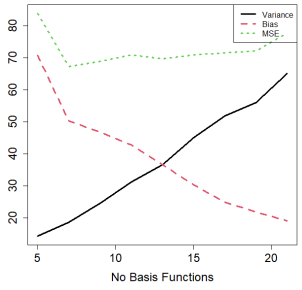


Figure 31: Fly 453 MSE

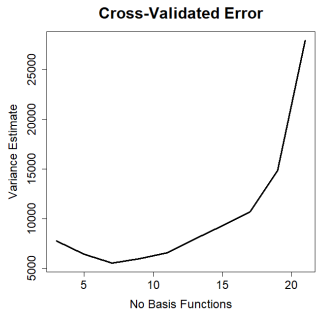


Figure 32: Fly 453 OCV

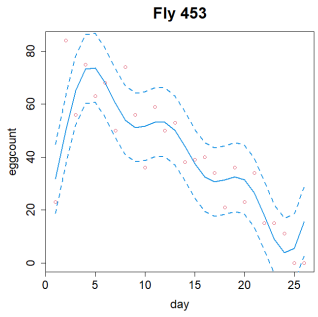


Figure 33: Fly 453 CI

Further using ordinary cross validation (OCV) for this problem, fly 512 has a cross validation score that keeps rising, which is hard to tell the result. However, if we change a fly, say fly 453, it's cross validation score has a very obvious "elbow" point as in Figure 31. It would be better to combine the MSE and OCV method of deciding the optimal number of basis functions.

For fly 453, the MSE suggests the optimal number of Fourier basis functions to be 7, and the ordinary cross validation (OCV) score also suggests 7 basis functions. Thus, it uses $\Phi(t) = (1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \sin(3\omega t), \cos(3\omega t))$ as basis functions. The 95% point-wise confidence interval of fly 453 is plotted in the dashed line in Figure 33, it contains the majority of points.

3.2.5 Using Smoothing Penalties

We will continue on producing the result for Fly 512, since it seems to be a tough question of deciding the number of basis function (through CV it is not possible), For the reason that selection of knots as well as the number of order would both effect the result. Thus, another alternative would be introducing appropriate linear differential operator for estimation. It is where the penalized method becomes useful.

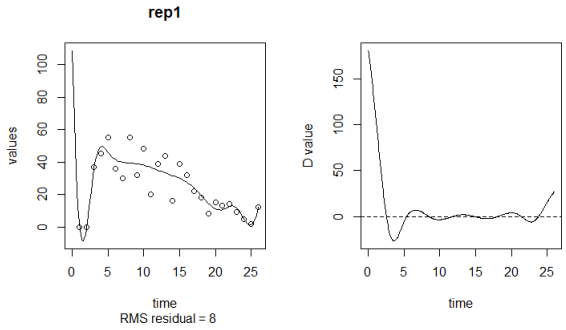


Figure 34: Fly 512 fit and 2^{nd} derivative

The procedure starts with creating a B-spline basis function as before. The 7 interior knots and order remain the same as in previous sessions. In Figure 34, it firstly plotted the fitted curve, and then its second derivative. The second derivative is approximately oscillating around 0. Figure 36 is the derivative of the harmonic acceleration, still oscillating around 0 (especially in the middle).

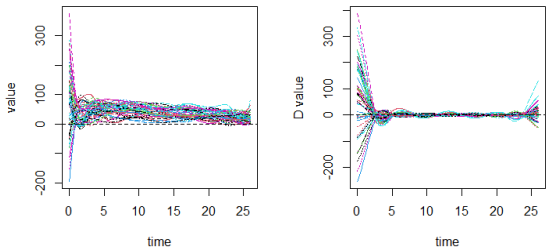


Figure 35: Fitted Curve and D2 of All Flies

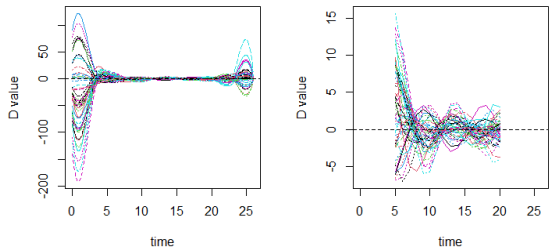


Figure 36: Harmonic Acceleration of

Then calculate the derivatives for all of the 50 flies. Firstly, all fitted curves are shown, then the 2^{nd} derivative and harmonic acceleration ($Lx = D^3x + w^2Dx$). Both derivatives are oscillating around 0, and more mild in the middle. If we look at the harmonic acceleration of 50 flies between 5 to 20 time points, its' absolute value is bounded by a smaller constant.

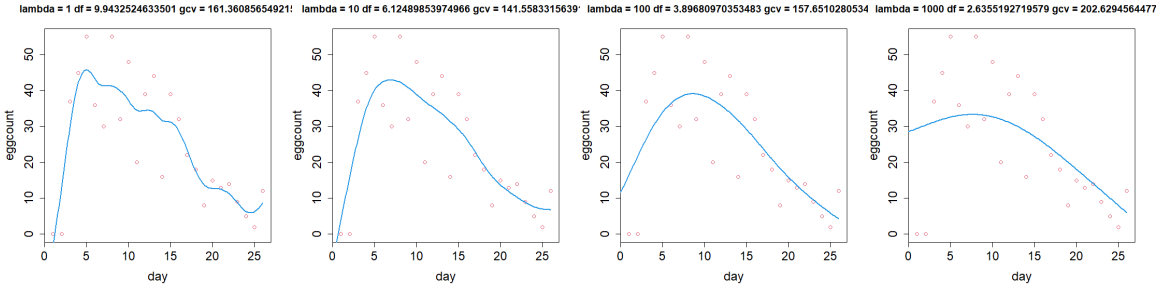


Figure 37: From left to right: $\lambda = 1$, $\lambda = 10$

Figure 38: From left to right: $\lambda = 100$, $\lambda = 1000$

Then, we would want to use the penalized squared error instead of least squared error for regression. For the first step, a saturated B-spline basis function is created, with number of basis function equals to 27. Then, the generalized cross validation (GCV), degrees of freedom (DF), and SSE were calculated. When using the logarithmic scale of λ, the plot shows the minimimal GCV score at $\lambda = 10$, when $GCV = 141.56$. If we set λ to be between 8 and 12, the minimum is also reached at $\lambda = 10$. As lambda increases, the fitted curve tends to be more close to a straight line (since the penalized term used here is $J_2[x] = \int \left[D^2x(t)\right]^2 dt$). Then, draw the variation of degrees of freedom (DF), SSE and GCV score. The GCV plot shows a dip near $\lambda = 10$. For more plots, refer to Appendix.

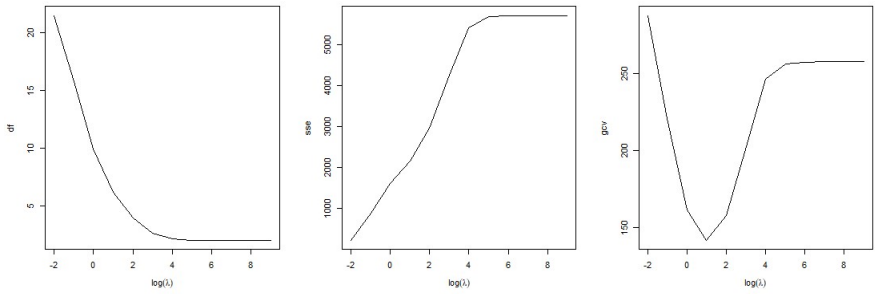


Figure 39: From left to right: DF, SSE, GCV

Other than only using the $J_2[x]$ penalizing term, further attempts could be made through using $J_1[x]$ and $J_3[x]$. $J_1[x]$ penalizes the square of the slope or velocity, $J_2[x]$ penalizes the squared acceleration, while $J_3[x]$ penalizes the squared rate of change of acceleration. But the B-spline that I used previously has norder = 4, which is not possible for minimizing the derivative of order 3. Thus, this time, interior points are the same as previous, while the order of the B-spline basis is set to be 6. In Figure 40, 41 and 42, as the order of derivative

increases, the value of optimal λ increases, and the curve tends to be more smooth, with a lower GCV score. This is possibly true that when the higher order of derivatives are penalized, the lower order derivative as well as the fitted curve tend to be more smooth.

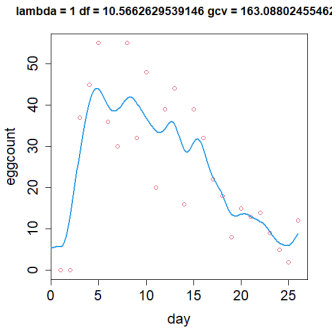


Figure 40: $J_1[x]$ selected $\lambda = 1$

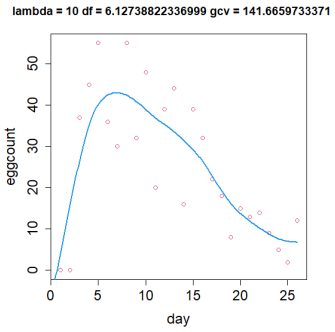


Figure 41: $J_2[x]$ selected $\lambda = 10$

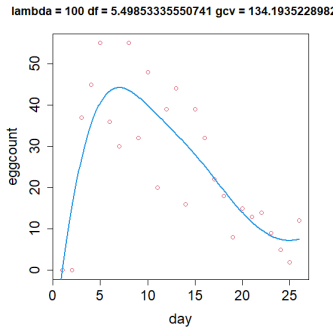


Figure 42: $J_3[x]$ selected $\lambda = 100$

For all of the 50 flies, the GCV scores could be calculated, with the optimal value of λ determined on the logarithmic scale. The table of the optimal λ value with the smallest GCV score is provided. 30 flies of 50, 60% of egg counts data selected λ approximately equals to 10 or 100.

λ Value	0.01	0.1	1	10	100	1000	10000	10^9
Counts	5	2	2	17	13	3	1	7

Table 3: Distribution of λ value when GCV is at maximum

There are about 9 flies that have GCV score consistently decreasing, and 5 flies that have GCV score that has small λ value, where the GCV method failed to provide an optimal solution. Thus, further method could be studied in future.

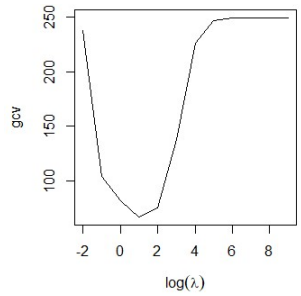


Figure 43: Most cases

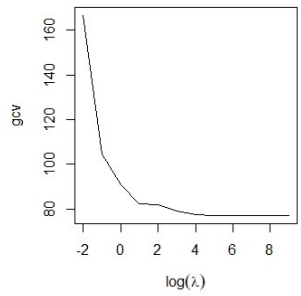


Figure 44: Increasing GCV

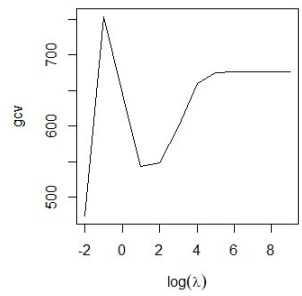


Figure 45: Low GCV at small λ

3.2.6 MODEL COMPARISON

When fitting a curve, several basis functions could be used: monomial, Fourier, and B-spline. Monomial basis has poor fitting performance to derivatives, thus B-spline and Fourier basis were mainly used.

When doing basis expansion, other than variance or bias, the number of basis function was firstly determined through MSE. Some flies could also use the ordinary cross validation score. However, it would be better to combine them together.

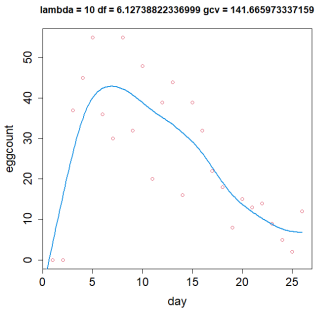


Figure 46: $J_2[x]$, $\lambda = 10$

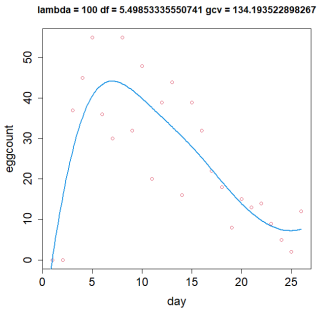


Figure 47: $J_3[x]$, $\lambda = 100$

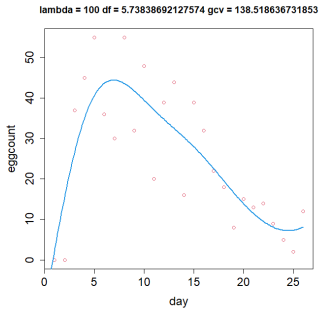


Figure 48: Harmonic Acceleration, $\lambda = 100$

Other then basis expansion, using smoothness penalty is more flexible, especially useful for basis like B-spline. Thus, with different penalizing term $J_m[x]$, the same choice of λ returns different $\hat{y}$ value. Figure 47 used the 3^{rd} derivative as penalized term, Figure 48 used the harmonic acceleration. Though having the same optimized λ at logarithmic scale, $GCV \approx 134.19$ for $J_3[x]$, and $GCV \approx 134.52$ for harmonic acceleration. This further illustrates that using $J_3[x]$ to penalize has a better estimation. Optimal λ were determined in the logarithmic scale.

3.2.7 IMPORTANT FINDINGS

1. The monomial basis is not an ideal approach in this question, for its poor performance on fitting the derivatives.

2. When choosing the number of basis function for Fourier basis, it would be better to combine the MSE and CV method (some flies have a monotonically increasing CV score).

3. For fly 512, spline of order 4 provides a plausible trend, while with higher order, the curves tend to overfit. After appropriate adjustment of interior knots, fitting performance would increase.

4. Using smoothing penalties is also an effective method. But different choice of penalty terms would yields different result, the $J_3[x]$ penalty when $\lambda = 100$, as well as $J_2[x]$

penalty when $\lambda = 10$ has a relatively small GCV score. Thus, both result could be considered.

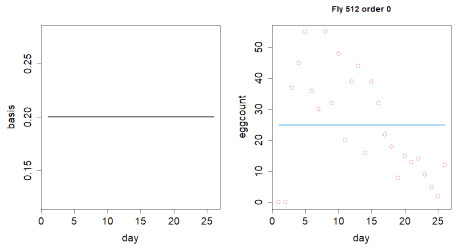
5. Among all of the fruit fly data, 60% them they have similar selection of optimal λ of 10 to 100. But some have abnormal GCV, and should be checked carefully.

4. Conclusion

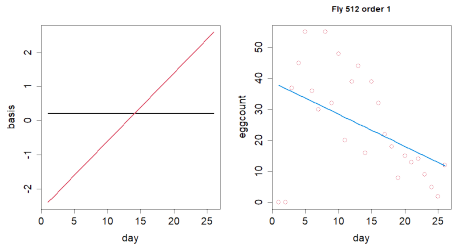
In this report, two data sets were examined and processed. The Canadian weather data (precipitation of Vancouver, and daily average temperature) and fruit fly data (egg counts). Both of them went through the process of creating different basis function, statistical inference and confidence interval, selecting the number of basis function and smoothing data with penalties. After selection of optimal λ, we would be able to output a smooth curve with minimum GCV score. For 50 flies, the distribution of optimal λ could be calculated. It could be further used for predictions given a future time point. To enhance the study, more choices of basis function and penalized terms could be further introduced, and more validation method could be applied to optimize the fitted result.

Appendix

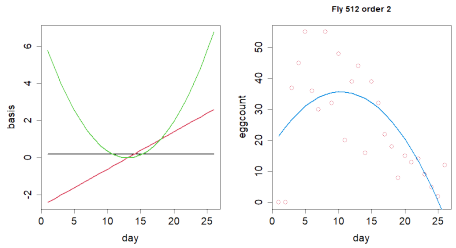
Monomial Fitting of Fly 512



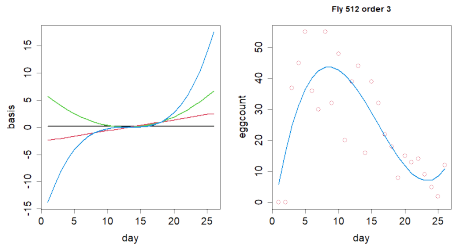
(i) order 0



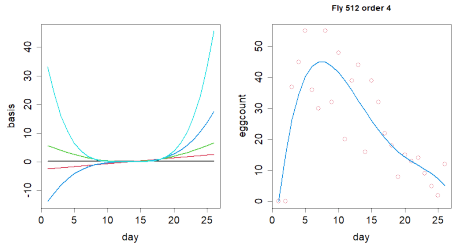
(ii) order 1



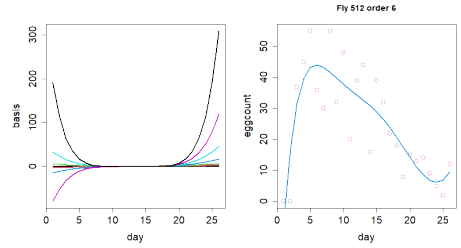
(iii) order 2



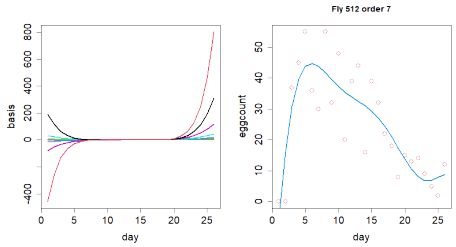
(iv) order 3



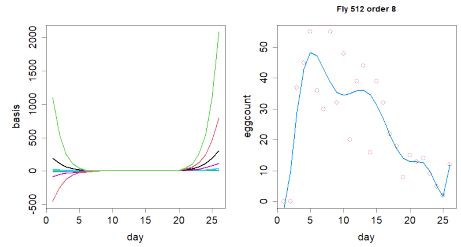
(v) order 4



(vi) order 6



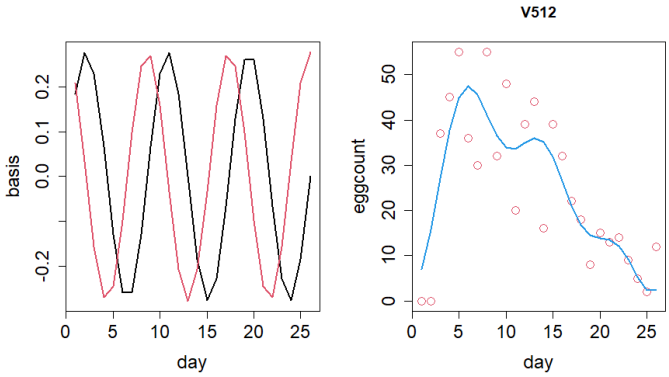
(vii) order 7



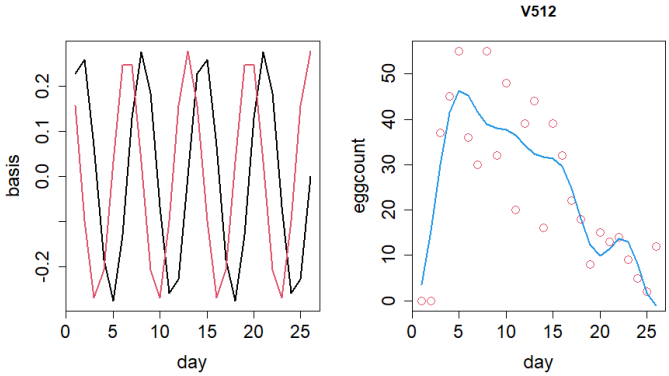
(viii) order 8

Figure 49: Monomial fitting of Fly 512

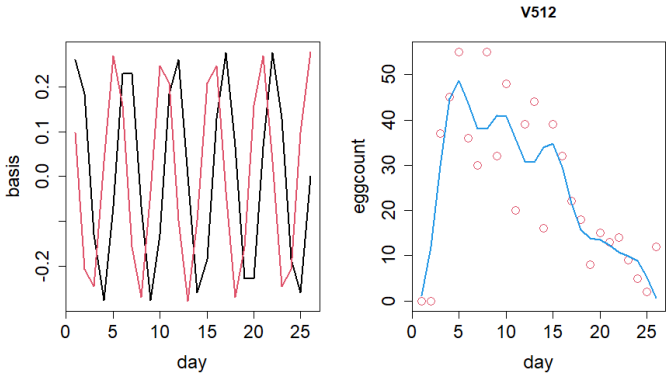
Fourier Basis Function of Fly 512



(i) 7 Fourier basis function



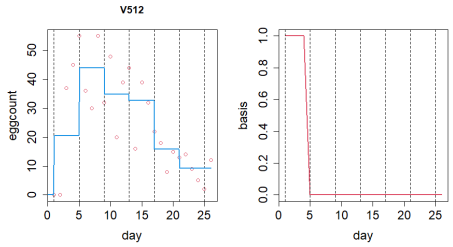
(ii) 9 Fourier basis function



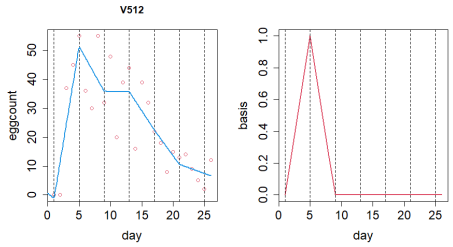
(iii) 11 Fourier basis function

Figure 50: Fourier basis functions. As the number of Fourier basis function increases, the curves failed to be smooth.

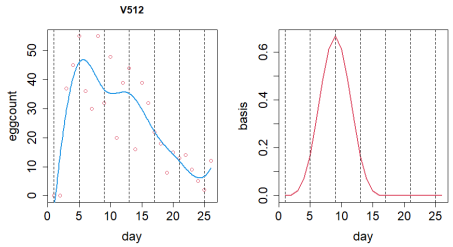
B-spline of Fly 512



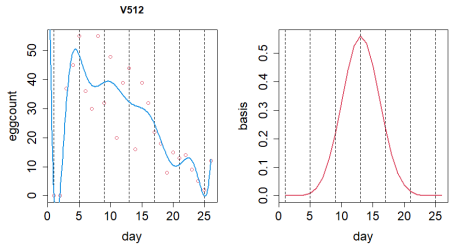
(i) order 1



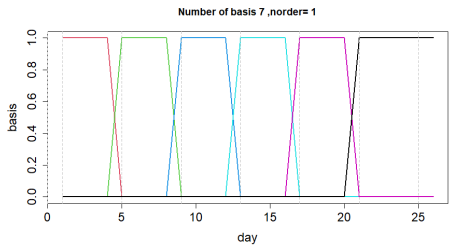
(ii) order 2



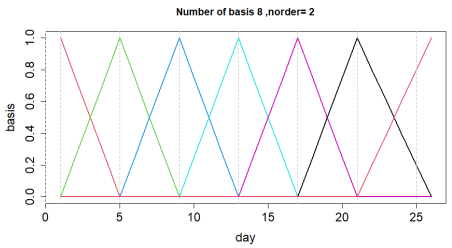
(iii) order 4



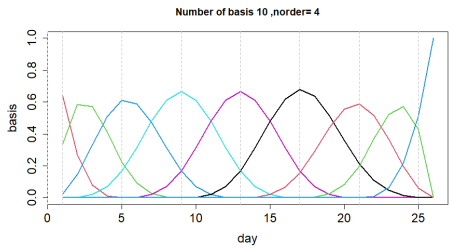
(iv) order 6



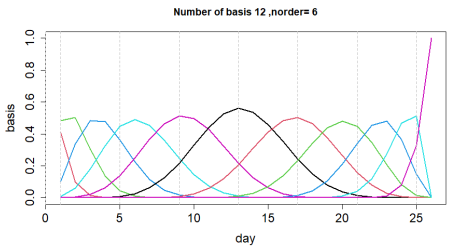
(v) order 1



(vi) order 2



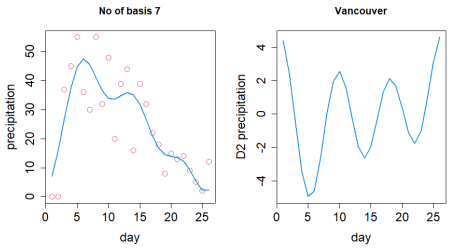
(vii) order 4



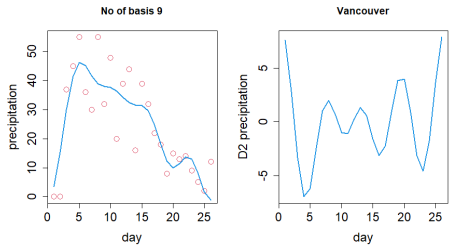
(viii) order 6

Figure 51: B-spline fitting and all basis functions

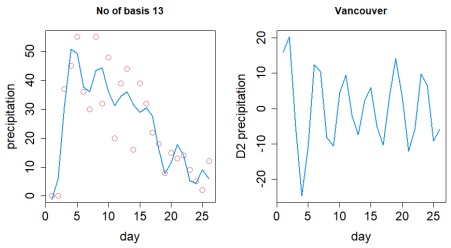
Fourier Basis Function and 2^{nd} Derivative



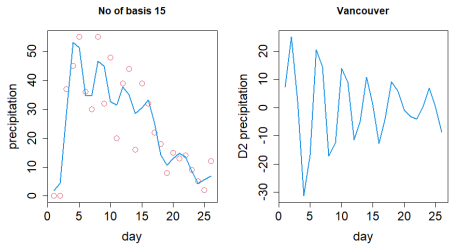
(i) 7 basis



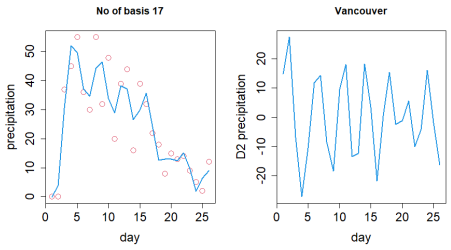
(ii) 9 basis



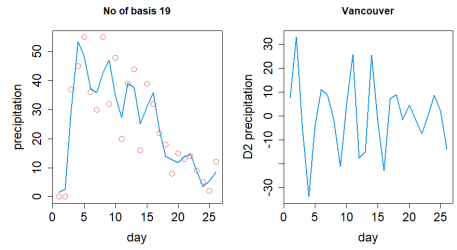
(iii) 13 basis



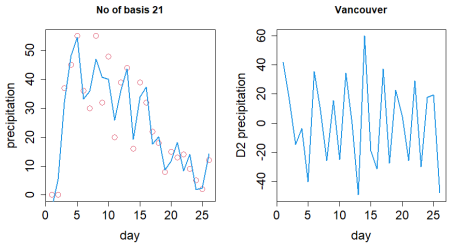
(iv) 15 basis



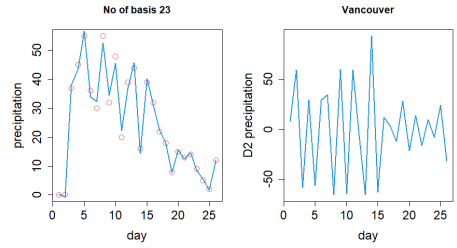
(v) 17 basis



(vi) 19 basis



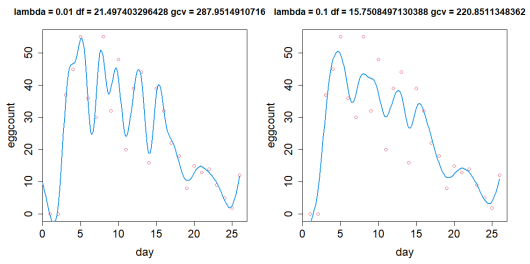
(vii) 21 basis



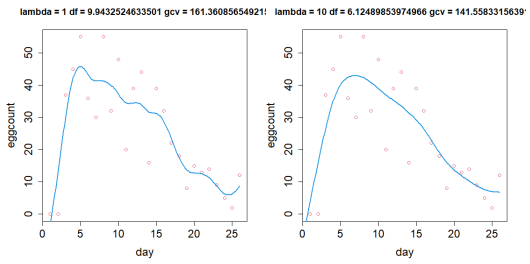
(viii) 23 basis

Figure 52: Fourier basis fitting and 2^{nd} derivative

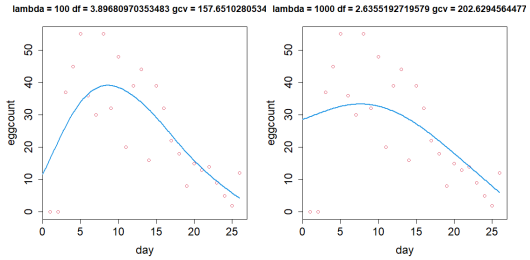
Fourier basis fitted by penalized smoothness



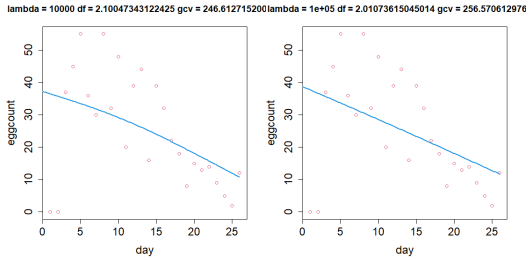
(i) $\lambda = 0.01$ and $\lambda = 0.1$



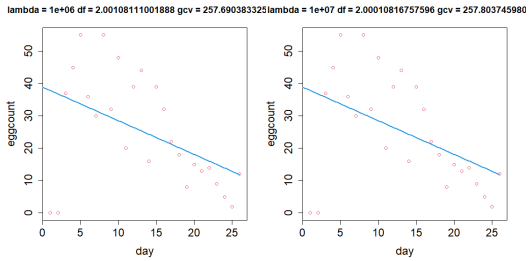
(ii) $\lambda = 1$ and $\lambda = 10$



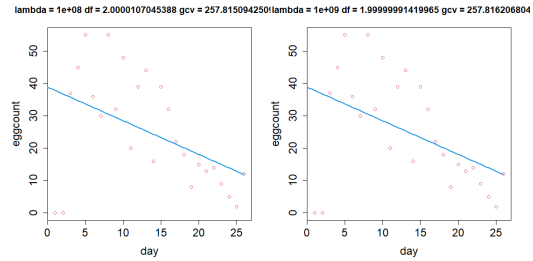
(iii) $\lambda = 100$ and $\lambda = 1000$



(iv) $\lambda = 10000$ and $\lambda = 10^5$



(v) $\lambda = 10^6$ and $\lambda = 10^7$



(vi) $\lambda = 10^8$ and $\lambda = 10^9$

Figure 53: Fourier basis fitting with penalized smoothness