Procedure of Simulation Study and Smoothing Fruit lies Data

Gong Wenwu, 12031299

March 12, 2021
Emails: 12031299@mail.sustech.edu.cn

Abstract

In this report, there are two main points. Firstly, we will give a bias-variance simulation study to choose the order K of basis expansion model. Also, the criteria order cross-validation has been verified useful in finding the order of basis expansion model. Secondly, we apply many types of basis functions to smooth the fruit flies data set in basis expansion model and the results show that B-spline has a better performance than other basis functions. Thirdly, we smooth the fruit flies data by roughness penalty model and compare with basis expansion model, the results show that the selection of λ may independent of the order K of basis function and roughness penalty may not perform better than basis expansion based on criteria MSE.

Contents

1 Bias-variance Simulation Study

The goal of bias-variance simulation study is choosing the order of expansion, i.e., the number of basis functions K. Based on the lecture notes, we first fit Vancouver precipitation by B-splines and find the best model containing 15 basis functions (norder+knots), further, pretending this is the 'truth' model. We then calculate 'errors' , and create new 'data' by randomly re-arranging these errors. Finally fitted the new data using a Fourier basis and repeat nrep times to calculate bias and variance from samples. In the following code, we just take nrep=1000 and minimize the MSE to choose the optimal number of Fourier basis.

1.1 Step 1: Fitting Vancouver precipitation by B-splines

In this section, we want to fit Vancouver precipitation data by B-splines basis functions. The most important part is to determine the total number of the basis functions (norder+nknots). Firstly, we look at the B-spline basis with 11 interior knots determined by different months, and then choose the order of basis function by minimizing order cross-validation (OCV). Further, we fit the Vancouver precipitation data with 3 interior knots (different seasons), where these basis functions have the different characteristics compared with 11 interior knots. The final results show that 15 (4+11) B-splines basis functional object depict the same essential characteristics in Vancouver precipitation data.
The theory of basis expansion model building as follows. Given the observation $(t_i, y_i), i = 1, 2, \cdots, n$, under the assumption of Standard Model (Gauss-Markov conditions), we model these observations by basis function expansion, shown as in Eq. 1:

$$y_i = x(t_i) + \epsilon_i \tag{1}$$

where $x(t)$ is estimated by $\sum_{j=1}^{K} c_j \phi_j(t)$, where $\phi_j(t)$ is the basis functional object. We determine the coefficients of the expansion c_j by minimizing the least squares criterion and this model form can be called as linear smoother.

$$SSE = \sum_{i=1}^{n} (y_i - x(t_i))^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{K} c_j \phi_j(t))^2$$

Now we give some notations for this linear smoother model. Define a N by K matrix Φ containing the values $\phi_k(t_i)$, a parameter vector $c = (c_1, c_2, \cdots, c_K)$ and a observation vector $y = (y_1, y_2, \cdots, y_n)$. So, the Least-Square errors (SSE) can be written by Eq. 2.

$$SSE(c) = (y - \Phi c)^T (y - \Phi c) \tag{2}$$

Taking the derivative of criterion $SSE(c)$ with respect to c, the coefficients c and fitted values are given by Eq. 3:

$$\widehat{c} = (\Phi^T\Phi)^{-1}\Phi^T y, \qquad \widehat{y} = \widehat{x}(t) = \Phi(\Phi^T\Phi)^{-1}\Phi^T y \hat{=} Sy \tag{3}$$

To achieve the 'best' estimation performance, we can use leave-one-out cross validation, a pure data driven method.

Stepwise variable selection.

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{x}^{-i}(t_i))^2 \tag{4}$$

where $\widehat{x}^{-i}(t_i)$ is the predicted value at t_i computed by using all data except the ith observation. By some inductions [Fan2020], in this linear smoother model, we can minimize the ordinary cross-validation score Eq.5 to choose the best order of B-splines.

$$OCV(\widehat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \widehat{x}(t_i))^2}{(1 - S_{ii})^2} = \frac{1}{n}\sum_{i=1}^{n}(\frac{y_i - \widehat{y}_i}{1 - S_{ii}})^2 \tag{5}$$

The minimization of OCV is 362.4, so the best order is given by 4 and the total number of the basis functions is 15. Therefore, the basis functions are cubic form. Eq. 6 has shown the approximated function with 11 knots:

$$\tau_K = 31, 59, 90, 120, 151, 181, 212, 243, 273, 304, 334$$

and Fig. 1 has plotted the fitted value.

$$y = \widehat{c_0} + \widehat{c_1}x + \widehat{c_2}x^2 + \widehat{c_3}x^3 + \widehat{c_4}(x - \tau_1)_+^2 + \cdot + \widehat{c_{14}}(x - \tau_{11})_+^2 + \epsilon \tag{6}$$

1.2 Step 2: Create new data by randomly re-arranging the errors

Based on the B-spline basis expansion model, we have the 'truth' value of $x(t) = \Phi\widehat{c}$, then the assumption model errors can be given by ϵ (Eq. 7) and we can create 'new' data by sampling from errors guided by Eq. 8

$$\epsilon_i = y_i - x(t_i) \tag{7}$$

$$y_i{}^* = x(t_i) + \epsilon_i{}^* \tag{8}$$

To get a more accurate result, we repeat this process 1000 times.

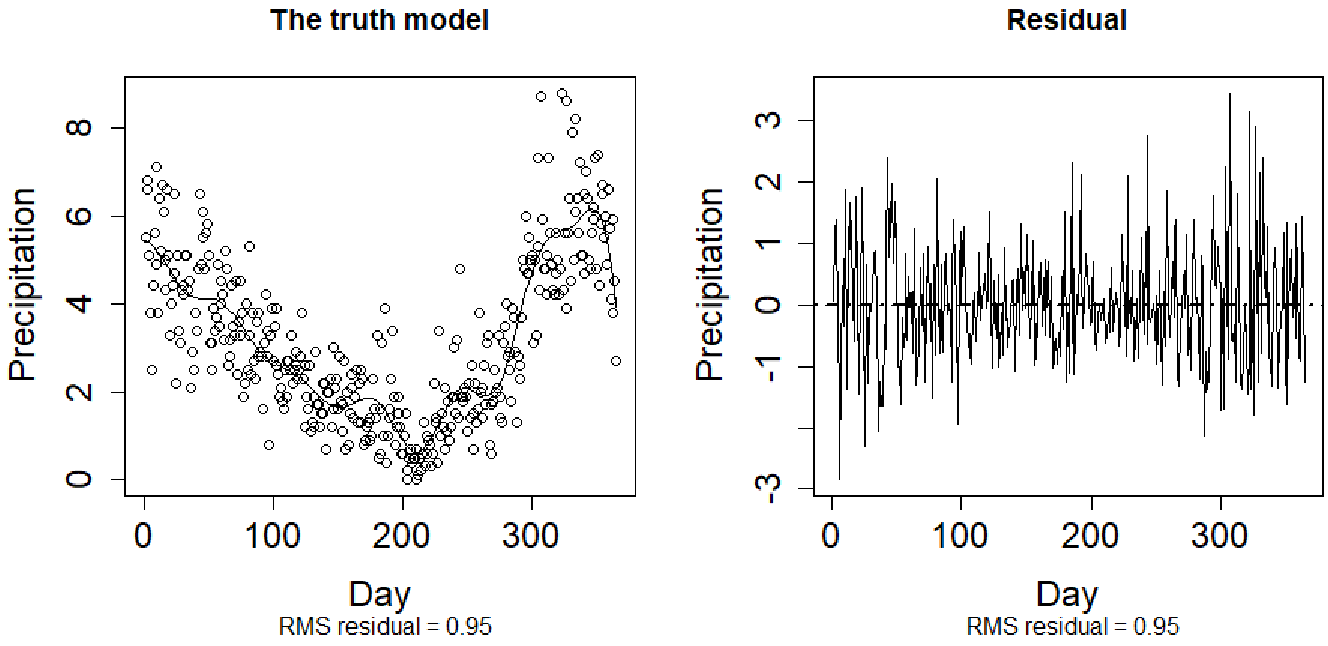


Fig. 1: The truth model and residual of precipitation in Vancouver.

1.3 Step 3: Simulation of bias-variance

The order of basis expansion K is very important between the fitting noise and information loss, i.e., the value of bias and variance with respect to the estimator. However, too many basis functions means small bias but large sampling variance and too few basis functions means small sampling variance but large bias. So, the trade-off between bias and variance can be used to choose the number of basis functions (Order of expansion K). Based on the expression of bias and variance, the best number of basis functions can be given by minimizing mean squared error (MSE) (Shown as in Eq.9).

$$\begin{gathered} Bias(\hat{x}(t)) = x(t) - E\hat{x}(t) \\ Var(\hat{x}(t)) = E(\hat{x}(t) - E\hat{x}(t))^2 \\ MSE(\hat{x}(t)) = E(\hat{x}(t) - x(t))^2 = (Bias(\hat{x}(t)))^2 + Var(\hat{x}(t)) \end{gathered} \tag{9}$$

By the 'created' data sets, we would like to find out the order of basis expansion by minimizing the integrated mean squared error (IMSE, shown as in Eq. 10). The simulation results (see Fig. 2) shown that the best order K existed when bias not decrease and choosing 13 basis functional object is the best choice for Fourier basis expansion model.

$$IMSE(\hat{x}(t)) = \int MSE(\hat{x}(t))dt \tag{10}$$

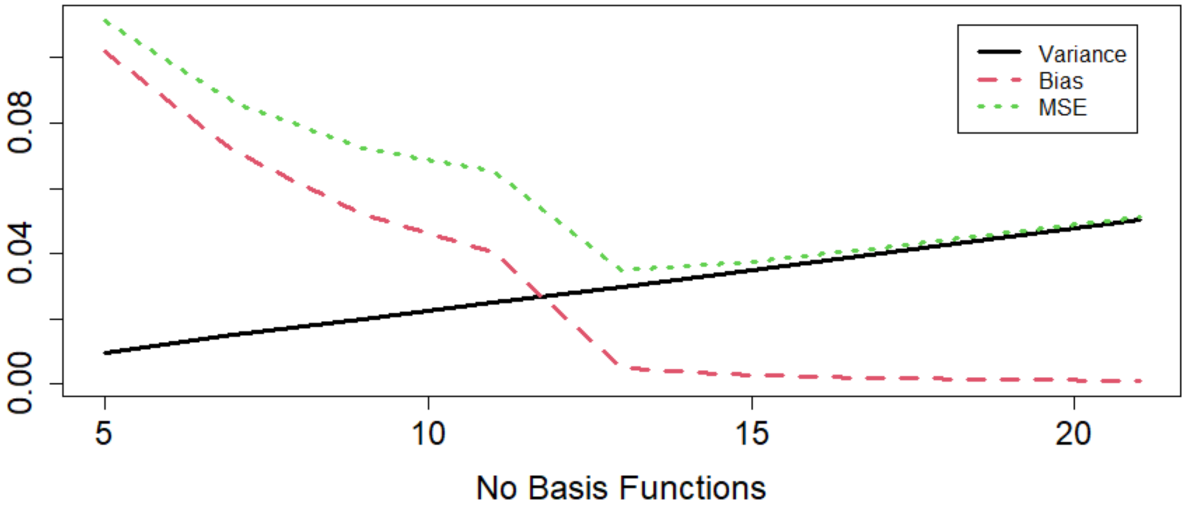


Fig. 2: The simulation results when repeats 1000 times. Sampling variance increases rapidly when we use too many basis functions, but squared bias tends to decay more gently to zero at the same time. However, we see there that the best results for mean squared error are obtained with 13 basis functions.

1.4 Step 4: Estimation and point-wise confidence bands

From Step 3, we have found a best Fourier basis expansion model to fit Vancouver precipitation data. According to the LSE procedures, we can estimate the parameters and Fig.3 has shown the fitted model which is same as the 'truth' model.

Confidence bands can provide an impression of how well the curve is estimated and are typically computed by adding and subtracting a multiple of the standard errors. According to Eq. 3, the variance-covariance matrix of the fitted values is given by:

$$Var(\hat{y}) = S^T Var(y) S, \qquad Var(y) = \sigma^2 \quad unknown.$$

Under the standard model assumption, we can use SSE given by least-square estimate of c to approximate σ^2 [Ramsay2005]. Then the variance-covariance matrix can be estimated by Eq. 11 and we can calculate lower and upper bands for $\hat{y}$ at each point (Eq.12). Fig. 4 have shown the point-wise confidence bands.

$$\widehat{Var}(y) = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{356 - 13} \tag{11}$$

$$\hat{y} \pm 2 \cdot (\frac{S^T(y - \Phi\hat{c})^T(y - \Phi\hat{c})S}{356 - 13})^{1/2} \tag{12}$$

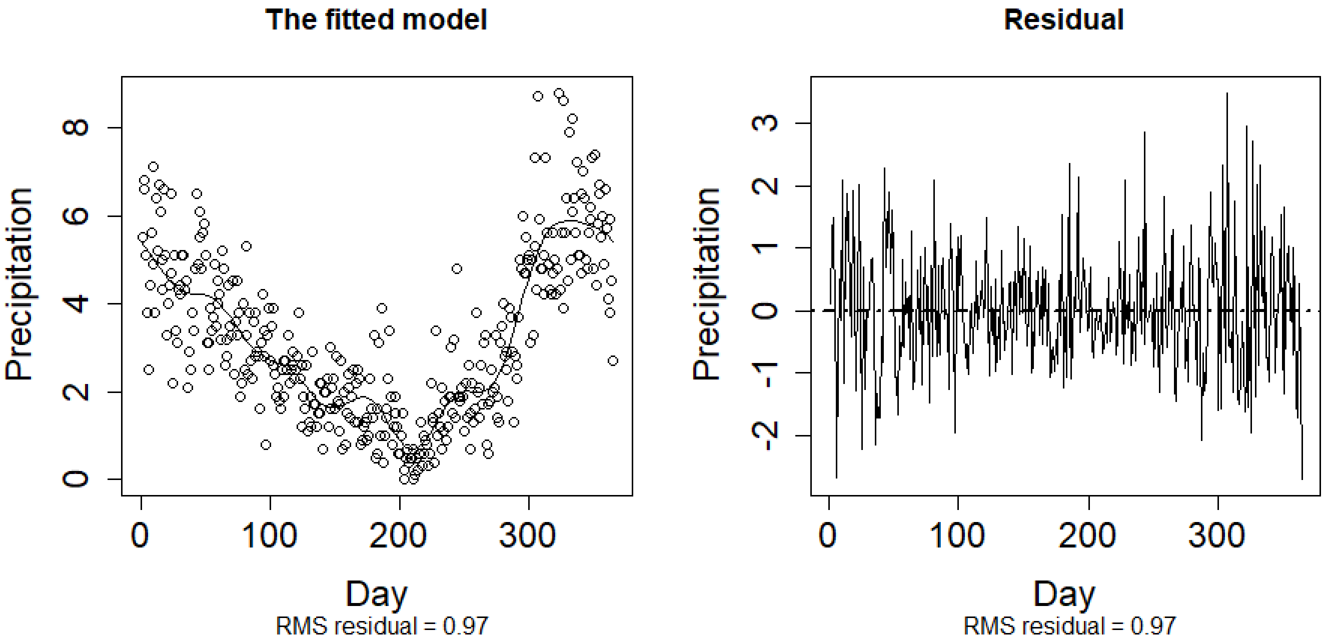


Fig. 3: The model fitted and residual of precipitation in Vancouver.

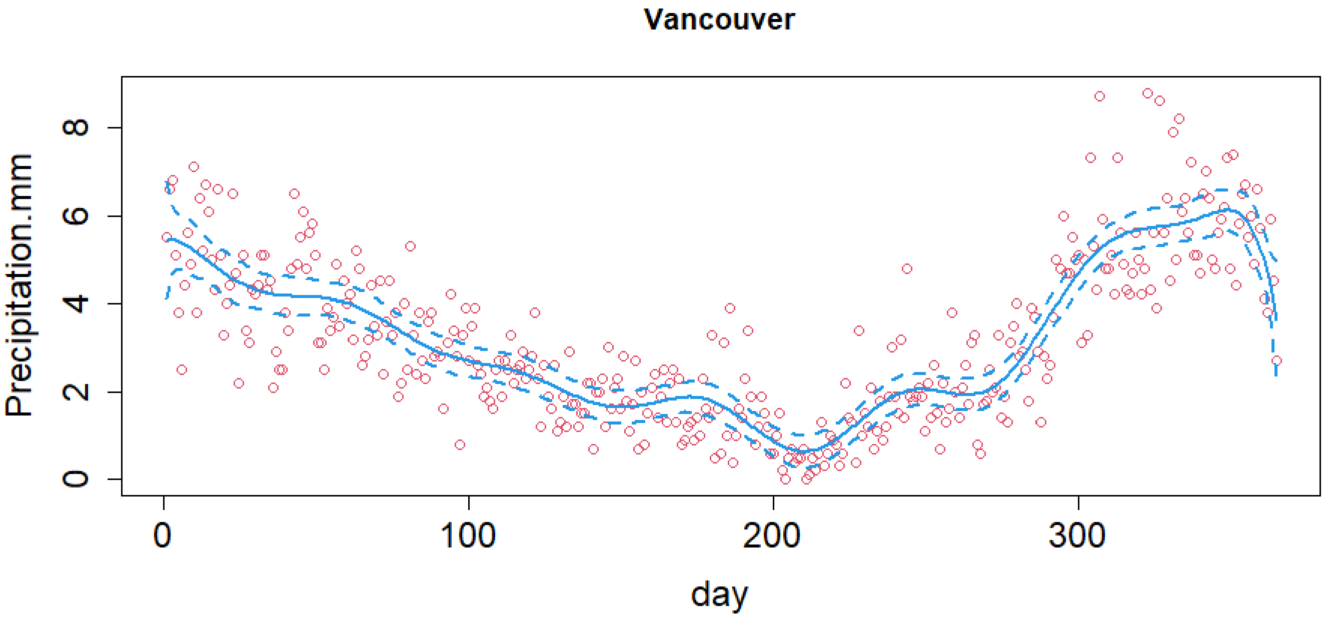


Fig. 4: Confidence bands for $\hat{y}$ at each point.

2 Smooth the fruit flies data

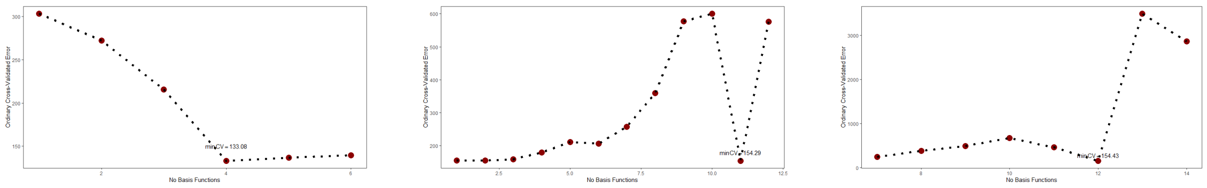
In this section, the focus moves on to smooth the fruit flies data compared with basis expansion model and roughness penalty model. The results will be guided by the following steps and we will repeat theses procedures on each fruit fly.

- Step 1: Minimize the cross-validation score to find out the best order basis expansion K;
- Step 2: Smooth the data by basis expansion model and find out the most suitable basis function;
- Step 3: Smooth the data by roughness penalty model based on the most suitable basis function given by Step 2 and select the tuning parameter λ by GCV;
- Step 4: Compare the performance of basis expansion and roughness penalty model.

In the following contents, we will give the example (512) to show the whole procedures of model building and discuss the results of the final choice model.

2.1 Finding out the best order of different basis expansion

Just as shown in Eq. 4, we can calculate the OCV for different basis functions. For polynomial and Fourier basis, we have picked 4 and 11 parameters of basis function with the minimal CV equal to 133.08 and 154.29, respectively. For B-spline basis function, we determine that a cubic basis and pick the best number of knots by minimizing OCV. The result shows that the order of B-spline basis functions $K = 12$ is the best choice and the CV equals to 154.43. The following Fig.5 has shown the best order K results.

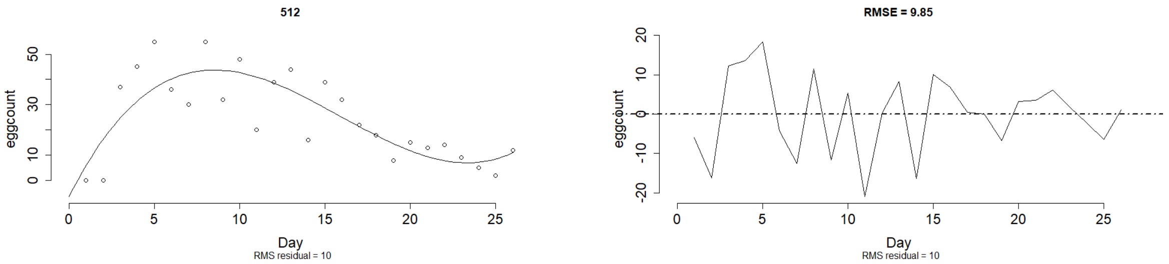


(a) Polynomial basis (b) Fourier basis (c) B-spline basis

Fig. 5: Minimal OCV based on different order K for different basis expansion. There is a local minimum in different type basis functions which is corresponding to the basis expansion order K.

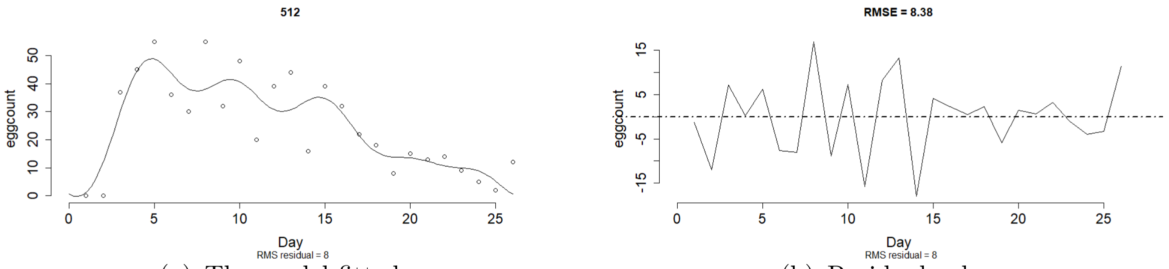
2.2 Smooth the data by basis expansion

Based on the procedures of basis expansion (Just as noted in Sec. 1.1), we can estimate the parameters for different functional object. Fig. 6, Fig.7 and Fig. 8 have shown the different results of basis expansion model fitted and the residual.



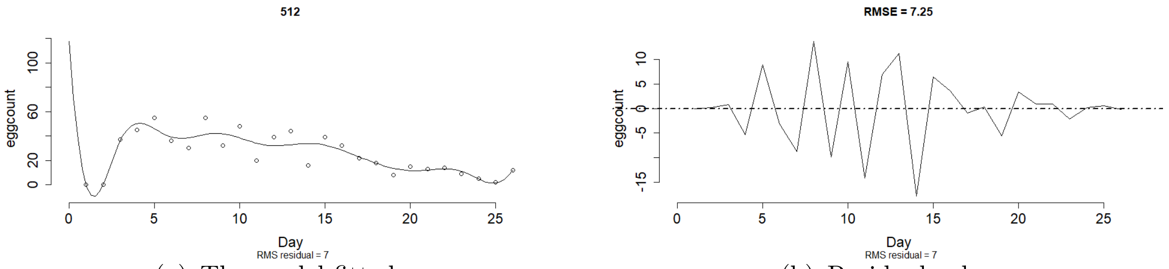
(a) The model fitted (b) Residual value

Fig. 6: Polynomial basis expansion model



(a) The model fitted (b) Residual value

Fig. 7: Fourier basis expansion model



(a) The model fitted (b) Residual value

Fig. 8: B-spline basis expansion model

At same time, we can compute the root-mean-squared-error (RMSE) of the model fitted values relative to the truth value so as to evaluate the model performance.

$$RMSE(t) = \sqrt{(E\hat{y}(t) - y)^2} \tag{13}$$

We have the RMSE value of different basis functions are 9.85, 8.38 and 7.25 respectively. For this data set, maybe we can say the B-spline basis expansion model has a better performance than other types of basis expansion. So, in the following roughness penalty model, the B-spline basis function will be used to smooth the number 512 fruit fly data.

2.3 Smooth the data by roughness penalty

We have saw in LSE estimator that basis expansions can provide good approximations to functional data provided that the basis functions have the same essential characteristics in the data set. However, the basis expansion model implies clumsy discontinuous control over the degree of smoothing, and we wonder if it is not possible to get better results with other methods. So, in this section, the roughness penalty or regularization approach has been discussed. The roughness penalty method is based on optimizing a fitting criterion, same as LSE, which defines what a smooth of the data is trying to achieve. Moreover, roughness penalty approaches can be applied to a much wider range of smoothing problems than simply estimating a curve [Ramsay2005].

In B-spline basis expansion smoothing, the mean squared error (MSE) shown as in Eq. 10, is one way of capturing the quality of estimator. The roughness penalty makes it more explicit by sacrificing the bias and then achieve an improvement on MSE. (We can prove that by doing derivative based on MSE) [Fan2020]. There are many notations should be discussed before illustrating the roughness penalty model. Firstly, we need to quantify the 'roughness' of a function which is determined by data set. The most useful 'roughness' is the square of second derivative just as defined in Eq. 14:

$$[D^2x(t)]^2 = [D^2\Phi c(t)]^2 \tag{14}$$

where the second derivative function at t can be integrated by the expression defined in Eq. 15, which is called the curvature of function $x(t)$.

$$PEN_2(x) = \int [D^2x(s)]^2 ds = \int c^T D^2\Phi(t) D^2\Phi^T(t) c dt = c^T Rc \tag{15}$$

For the fruit flies data, we assure that the curvature 'roughness' is more reliable compared with others penalty (The data is not periodicity). Secondly, the next step is to modify the least squares fitting criterion (Eq. 2), so as to allow the roughness penalty $PEN_2(x)$ to play a important role in defining $x(s)$. We define a compromise that trades off smoothness against data by defining the penalized residual sum of squares as:

$$PENSSE_\lambda(x) = (y - x(t))^T(y - x(t)) + \lambda PEN_2(x) \tag{16}$$

where parameter λ is a smoothing parameter that measures the rate of exchange between data fitting, as measured by the residual sum of squares in the first term, and variability of the function $x(t)$, as quantified by $PEN_2(x)$ in the second term. Estimating this roughness penalty smooth model also can be considered as the linear smoother which the estimated parameters satisfy:

$$\hat{c}_\lambda = (\Phi^T\Phi + \lambda R)^{-1}\Phi^T y, \qquad \hat{y} = \hat{x}(t) = \Phi(\Phi^T\Phi + \lambda R)^{-1}\Phi^T y = S_\lambda y \tag{17}$$

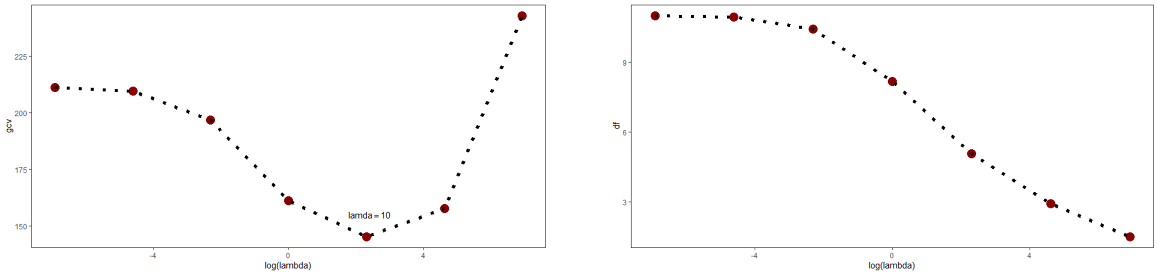
Just as shown in Eq. 17, where we smooth data using a roughness penalty instead of basis expansion, we switch from defining the smooth in terms of degrees of freedom K (basis expansion matrix S) to defining the tuning parameter λ (matrix S_λ). So, the most important part of data smoothing is to choose the tuning parameter in a reasonable way. Fortunately, we have a common sense method for the tuning parameter which is defined as generalized cross-validation measure (GCV) [Gu2002]. The criterion is usually expressed as:

$$GCV(\lambda) = \frac{n^{-1}SSE}{[n^{-1}trace(I - S_\lambda)]^2} = (\frac{n}{n - df(S_\lambda)})(\frac{SSE}{n - df(S_\lambda)}) \tag{18}$$

So, the best tuning parameter λ is given by:

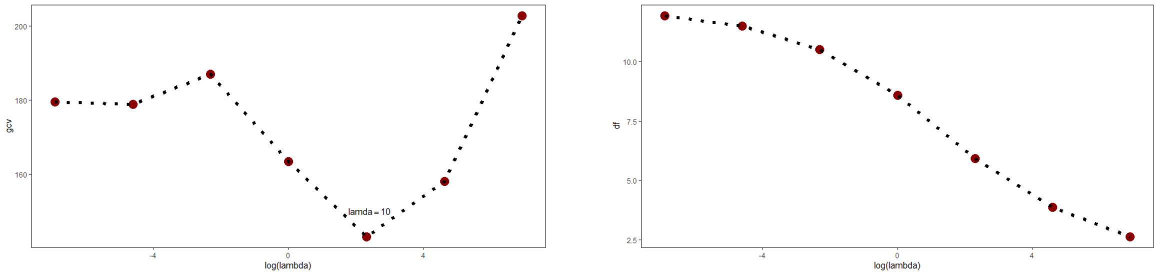
$$\lambda^{GCV} = \arg\min_\lambda(\frac{n}{n - df(S_\lambda)})(\frac{SSE}{n - df(S_\lambda)}) \tag{19}$$

Based on the above induction and the best tuning parameter λ given by Eq. 19, the roughness penalty model fitted of sample data (512) have the following results: we have the RMSE value of different basis functions are 9.64 and 9.09 respectively, which are both bigger than the value of basis expansion model. For this data set, maybe we can also say the B-spline have a better performance than Fourier basis function under same tuning parameter λ = 10. Further, the penalized method cannot improve the smooth performance (based on criteria RMSE) for the 512 data set.



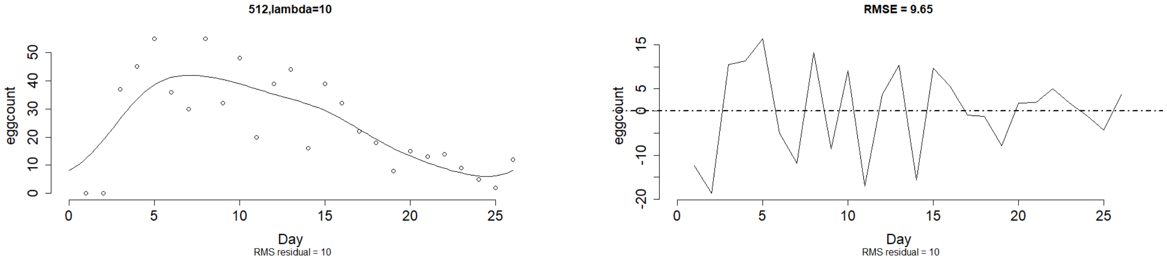
(a) GCV (b) df

Fig. 9: Choosing the tuning parameter λ for Fourier basis.



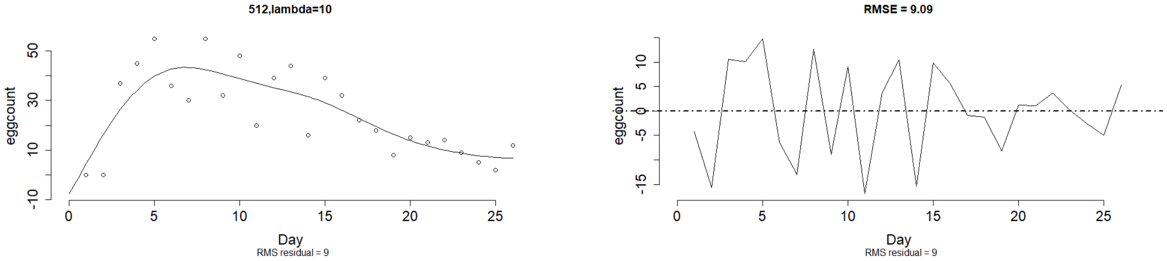
(a) GCV (b) df

Fig. 10: Choosing the tuning parameter λ for B-spline basis.



(a) The fitted model (b) Residual

Fig. 11: Penalized model for Fourier basis when $\lambda = 10$.



(a) The fitted model (b) Residual

Fig. 12: Penalized model for B-spline basis when $\lambda = 10$.

3 Conclusions

- The goal of bias-variance simulation study is to choose the order K of basis expansion model and most important part of simulation study is to understand the ‘truth’ model and ‘created’ data.
- We can apply different basis functions to smooth data and the choice is determined by the characteristic of data set.
- OCV is the one of criteria to find the best order K of basis expansion model.
- We should not fix the number of knots in B-spline basis expansion model when we smoothing the real data set. That is, the statement ‘The number of basis functions is $(n-2)+4=n+2$ where n is the number of sampling points’ may not applicable in basis expansion model.
- The selection of λ may be independent of the number of nbasis (Order K) in roughness penalty model and the tuning parameter λ can be determined by minimizing GCV.
- Honestly, we can choose tune parameter λ to make sure that the MSE of roughness penalty method is smaller than basis expansion. However, the MSE of most fitted roughness penalty model (determined by GCV) may bigger than basis expansion model.

References

[1] Fan, J. and Li, R. and Zhang, C.-H.and Zou, H.. Statistical Foundations of Data Science. (2nd ed.). CRC Press.

[2] Ramsay, J.O. and Silverman, B.W.. Functional Data Analysis. (2005) Springer, New York.

[3] Gu, C.. Smoothing Spline ANOVA Models. (2002) New York: Springer, New York.