

第一次上机实验报告

北京大学化学与分子工程学院 李梓焰 2101110396

摘要 本实验运用岭回归，在最高指数项不超过4的情况下，对数据分别进行关于 x 与 $\cos x$ 的多项式拟合，从而得到相应的四次多项式拟合式。同时，笔者分析了不同多项式次数下最佳的训练结果及相应的测试结果，以及RMS与 $\lg \alpha$ 的关系，并据此选出误差最小的模型。此外，笔者还将原训练集与测试集合并，并采用交叉验证的手段，分别得到关于 x ，关于 $\cos x$ ，以及两者混合的多项式拟合模型，与前一种方法相比，所得模型的误差均有显著减小，表明增大样本数量的同时采用交叉验证，可以进一步提升模型质量。

关键词 岭回归，多项式拟合，正则化系数，均方根误差

1 引言

线性回归模型是回归分析中最常用的模型之一，然而实际运用中，样本数据往往不满足线性关系，因此要引入高次特征项进行拟合，这就是多项式回归。

一般情况下，使用多项式回归可以使模型更加复杂，从而提高模型准确度，但如果模型过于复杂，过于贴近训练数据，有可能导致对真实数据的预测准确度并不高，即欠拟合现象。为了解决模型过拟合的问题，人们引入了正则化的概念，通过在损失函数中添加惩罚项，从而对损失函数的某些参数进行限制，以提高模型的泛化能力。其中最常用的正则化方法便是岭回归，其损失函数可表示为（按scikit-learn的定义）：

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m [h_\theta(x_i) - y_i]^2 + \alpha \sum_{j=0}^n \theta_j^2 \quad (\alpha > 0) \quad h_\theta(x_i) = \sum_{j=0}^n c_j x_i^j$$

α 是正则化系数，它控制正则化项的占比，对模型泛化而言非常重要。 α 过小，正则化的效果弱，得到的模型仍有严重的过拟合现象； α 过大，正则化的效果强，容易导致模型欠拟合，精确度变差。在实际运用中， α 作为超参数，需要结合训练和测试的结果进行多次尝试，寻找最合适的结果，也可以采用交叉验证的方式，获取最合适的数据。

本次上机实验采用scikit-learn自带的Ridge和RidgeCV，结合PolynomialFeatures，对训练样本进行多项式拟合，然后用mean_squared_error计算RMSE，最后用matplotlib对数据进行可视化处理。

2 实验部分

2.1 仪器

2.1.1 硬件

Surface Pro (第5代，处理器参数：Intel® Core™ i5-7300U CPU @ 2.60GHz, 2.71 GHz, 2个内核，4个逻辑处理器；内存容量：8.00 GB)

2.1.2 软件

操作系统：Windows 10家庭版，版本20H2

开发环境：Visual Studio 2019 Community, 64位Anaconda 3 (版本号2021.05，含64位Python 3.8.8、Conda 4.10.1、NumPy 1.20.1、SciPy 1.6.2、Scikit-learn 0.24.1、Matplotlib 3.3.4)

2.2.3 训练和测试数据

训练数据：train.dat，含12组(x, y)数据。

测试数据：test.dat，含300组(x, y)数据。

2.2 实验过程

2.2.1 训练和测试数据的读入

从train.dat读入训练数据，用于训练多项式岭回归模型；从test.dat读入测试数据，以考察不同多项式次数下岭回归模型的最佳训练结果和相应的测试结果。由于训练数据和测试数据的x值均从小到大排序，故无需对两者进行预排序。

2.2.2 计算并绘制RMSE、模型拟合系数和 $\lg \alpha$ 的关系图

使用scikit-learn库中的Ridge（岭回归），结合PolynomialFeatures，在不同 α 值和多项式次数下分别进行关于 x 与 $\cos x$ 的多项式拟合，然后分别计算训练数据与测试数据的预测值，以及两组数据的预测值与实际值的RMSE。随后，给定不同的多项式次数（最高至四次），绘制训练集RMSE与测试集RMSE随 $\lg \alpha$ 变化的关系图，以及模型拟合系数随 $\lg \alpha$ 变化的关系图。

2.2.3 不同多项式次数下最佳超参数 α 的手动选取，相应拟合曲线的绘制与参数保存

按照“训练集误差最小（或与最小值相差极小），且测试集误差尽可能小”的规则，让程序在对不同多项式次数的模型拟合过程中，遍历超参数 α 并拟合，接着记录相应的最佳数值，同时一并记录在最佳超参数下模型拟合系数的值，以及该模型的训练集RMSE和测试集RMSE。然后，将不同多项式次数的拟合模型优化结果进行比较，选出测试集RMSE最佳的模型。

2.2.4 特殊模型 $y = c_0 + x + c_1 \cos x + c_2 \cos^2 x + c_3 \cos^3 x + c_4 \cos^4 x$ 的岭回归拟合及超参数 α 的手动选取

由于scikit-learn库中缺乏固定参数项进行拟合的函数，因此，笔者首先对 $y - x$ 进行多项式拟合，这样，问题便转换为对 $f(x) = \sum_{i=0}^4 c_i \cos^i x$ 的拟合，而这一问题可以按照2.2.3介绍的方法解决。接下来，只需用 $f(x)$ 给出预测值，再与对应的 x 相加，即可得到特殊模型给出的预测值，据此可以给出该模型的训练集RMSE和测试集RMSE。最后，将上述方法所得的预测曲线，与训练数据、测试数据绘制在同一张图上，并输出该模型的系数，完成拟合。

2.2.5（附加实验）合并训练集与测试集，运用交叉验证进行关于 x ，关于 $\cos x$ ，以及两者混合的多项式拟合

由于本次实验的训练集样本数过少，若贸然进行交叉验证，有可能使拟合模型的误差进一步放大，故笔者将原训练集与测试集合并，然后使用scikit-learn库中的RidgeCV（带交叉验证的岭回归），并结合PolynomialFeatures，对训练数据进行多项式拟合，得到不同的多项式次数下，关于 x ，关于 $\cos x$ ，以及两者混合的多项式拟合采用的最优 α 值，同时计算出相应模型的训练集RMSE和测试集RMSE。最后，将优化后的不同的多项式次数下，关于 x ，关于 $\cos x$ ，以及两者混合的多项式拟合模型的预测曲线，与原训练数据、原测试数据绘制在同一张图上，同时将上述模型的优化结果进行比较，选出测试集RMSE最佳的模型。

2.2.6（附加实验）合并训练集与测试集，运用交叉验证，对2.2.4节的特殊模型进行自动化岭回归拟合

按2.2.4-5节所述，笔者将训练集与测试集合并，用scikit-learn库中的RidgeCV，拟合 $f(x) = \sum_{i=0}^4 c_i \cos^i x$ ，并自动完成对超参数 α 的自动选取。之后，用 $f(x)$ 给出预测值，再与对应的 x 相加，即可得到特殊模型给出的预测值，据此可以给出该模型的训练集RMSE和测试集RMSE。最后，将上述方法所得的预测曲线，与训练数据、测试数据绘制在同一张图上，并输出该模型的系数，完成拟合。

3 实验结果与数据

3.1 不同次数下，关于 x 的多项式拟合的RMSE、模型拟合系数与 $\lg \alpha$ 的关系图

读入训练数据与测试数据后，按2.2.2节进行实验，其中多项式次数从0取到4， $\lg \alpha$ 取值范围为 $[-23, 2]$ ，且选取的 $\lg \alpha$ 构成项数为10000的等差数列，由此得到RMSE、模型拟合系数与 $\lg \alpha$ 的关系图，如图1至图5所示。

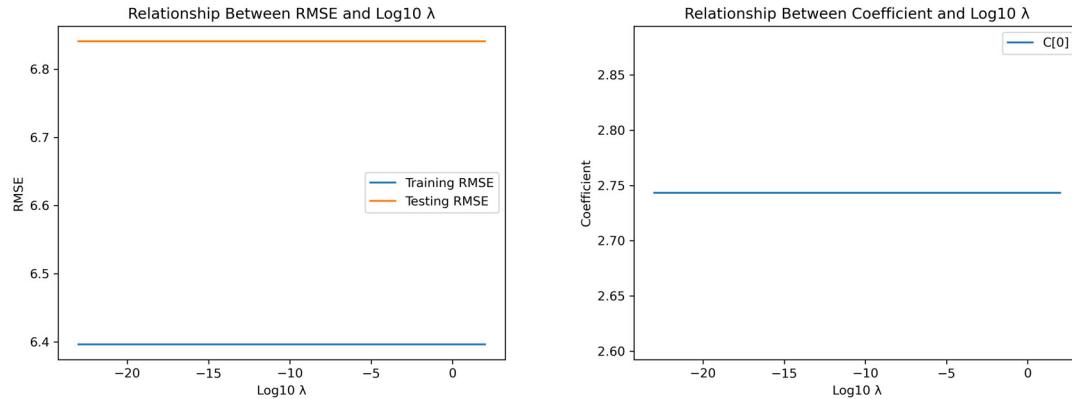


图1 最高次为0的关于x的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

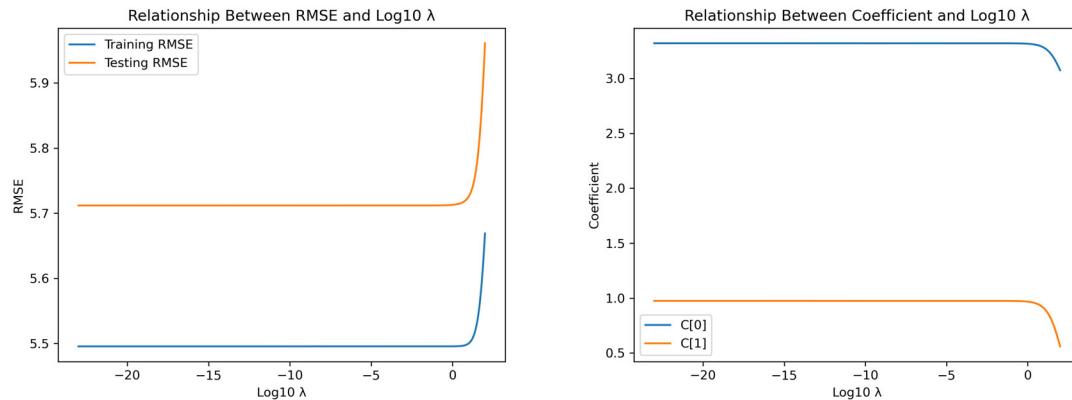


图2 最高次为1的关于x的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

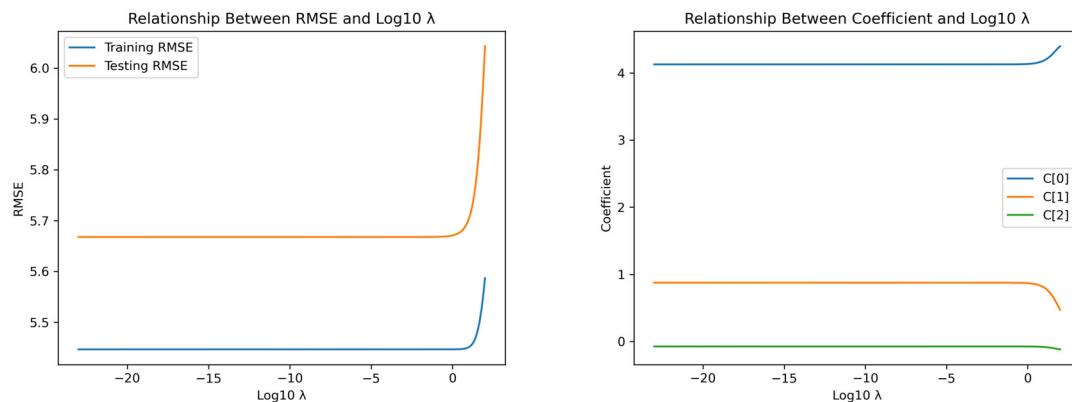


图3 最高次为2的关于x的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

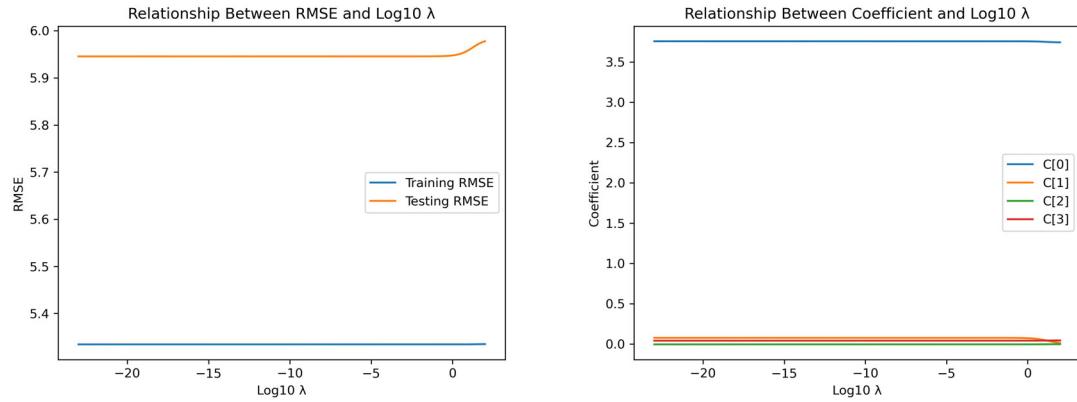


图4 最高次为3的关于x的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

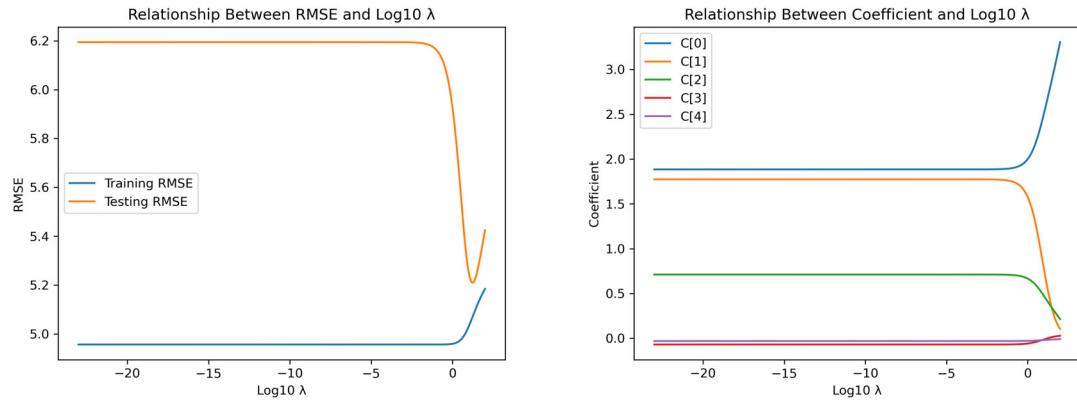


图5 最高次为4的关于x的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

3.2 不同次数下，关于x的多项式拟合的模型最佳超参数的选取，及对应模型训练集与测试集RMSE

显然，为使测试集误差最小而忽视训练集误差，相当于预知实验结果去修改模型，这种做法有作弊的嫌疑。然而，如果不采纳训练集的预测结果，就无法评判模型优劣。为兼顾训练集与测试集，笔者决定采用“训练集误差最小（或与最小值相差极小），且测试集误差尽可能小”的超参数选取原则，这样，模型既能与训练集契合，又不至于过拟合而损失泛化性（尽管可能不如使测试集误差最小的模型）。

根据这一原则，笔者进行了2.2.3节的实验，所得的不同次数下，关于x的多项式拟合模型最佳超参数及系数如表1所示，相应的训练集与测试集RMSE如表2所示，模型拟合曲线如图6所示。

表1 不同次数下，关于x的多项式拟合的模型最佳超参数及各项系数

最高次数	α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
0	1.000×10^{-23}	2.743				
1	1.000×10^{-23}	3.320	9.750×10^{-1}			
2	1.000×10^{-23}	4.129	8.755×10^{-1}	-7.464×10^{-2}		
3	1.000×10^{-23}	3.755	7.743×10^{-2}	-3.578×10^{-3}	4.221×10^{-2}	
4	1.380	2.037	1.513	6.517×10^{-1}	-5.530×10^{-2}	-2.858×10^{-2}

表2 不同次数下，关于x的多项式拟合的模型最佳超参数、训练集RMSE与测试集RMSE

最高次数	α	训练集RMSE	测试集RMSE
0	1.000×10^{-23}	6.396	6.841
1	1.000×10^{-23}	5.495	5.712
2	1.000×10^{-23}	5.446	5.668
3	1.000×10^{-23}	5.334	5.946
4	1.380	4.962	5.849

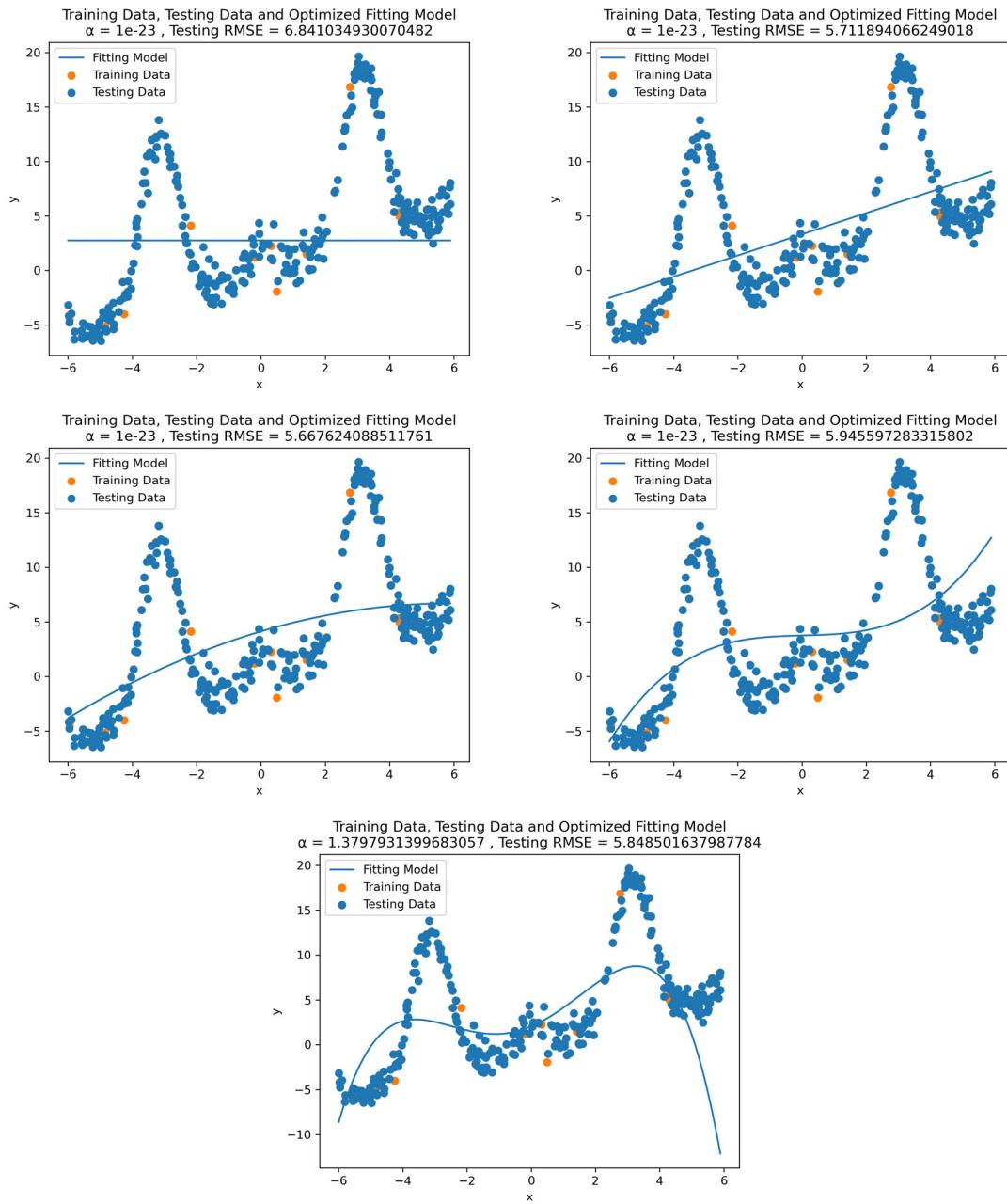


图6 不同次数下，关于x的多项式拟合的模型曲线（从左到右，从上到下，次数依次为0、1、2、3、4）

3.3 不同次数下，关于 $\cos x$ 的多项式拟合的RMSE、模型拟合系数与 $\lg \alpha$ 的关系图

类似于3.1节，仍设多项式次数从0取到4， $\lg \alpha$ 取值范围为 $[-23, 2]$ ，且选取的 $\lg \alpha$ 构成项数为10000的等差数列，按2.2.2节进行实验，可得RMSE、模型拟合系数与 $\lg \alpha$ 的关系图，如图7至图11所示。

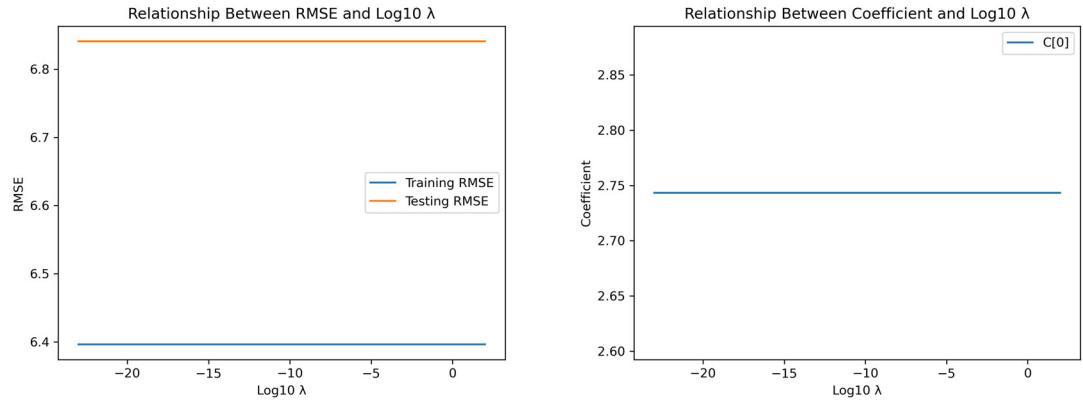


图7 最高次为0的关于 $\cos x$ 的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

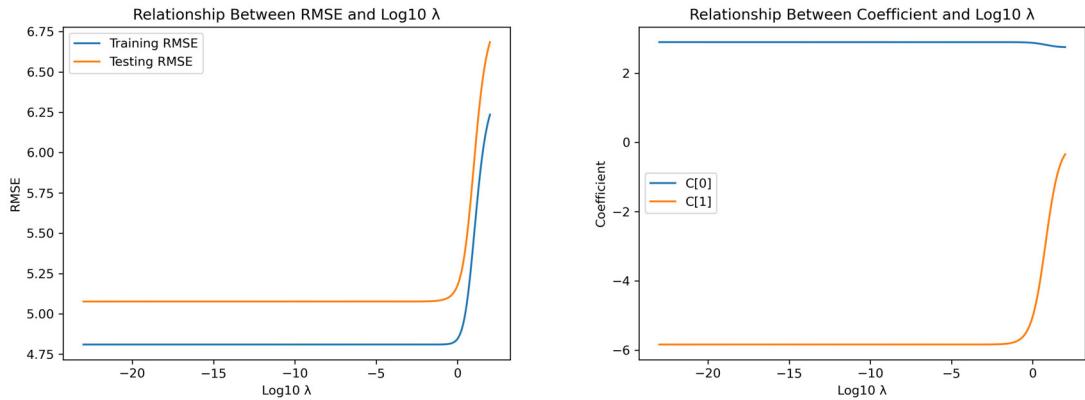


图8 最高次为1的关于 $\cos x$ 的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

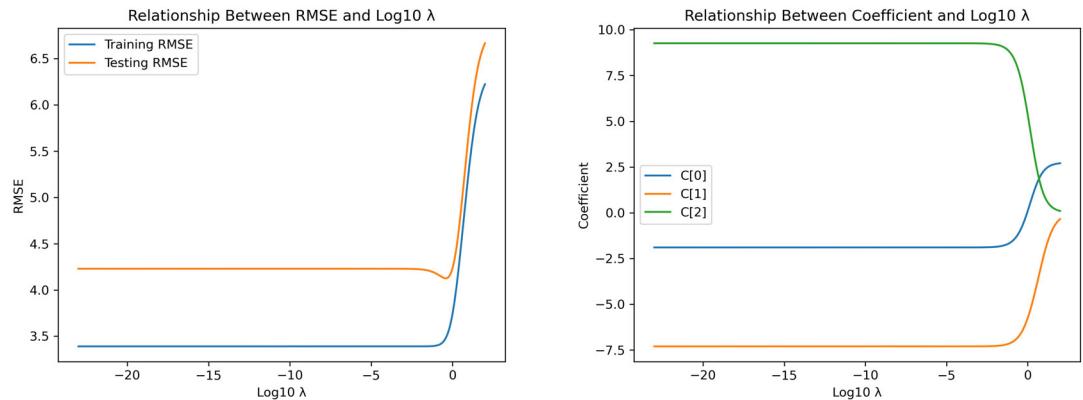


图9 最高次为2的关于 $\cos x$ 的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

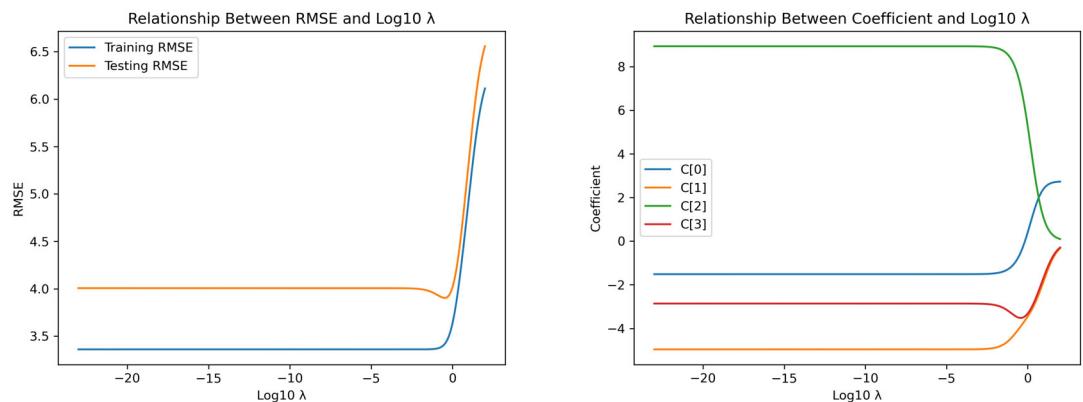


图10 最高次为3的关于 $\cos x$ 的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

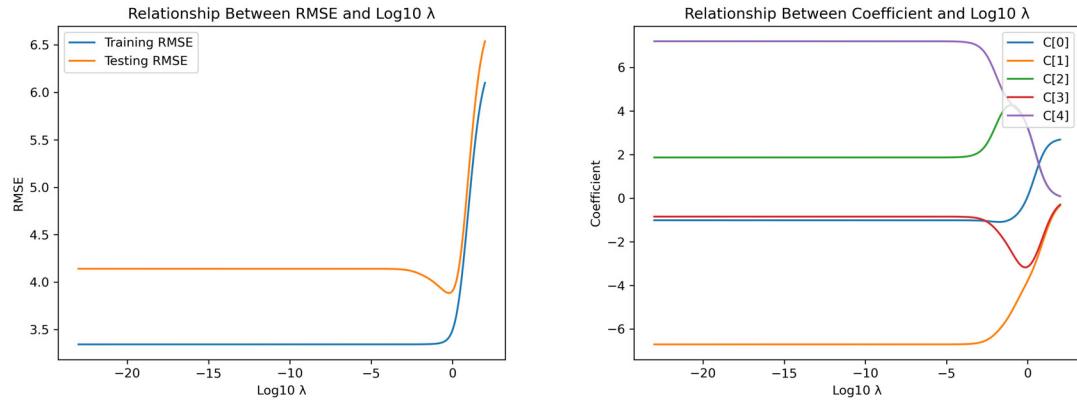


图11 最高次为4的关于 $\cos x$ 的多项式拟合模型关系图 (左: RMSE- $\lg \alpha$ 关系图; 右: 系数- $\lg \alpha$ 关系图)

3.4 不同次数下, 关于 $\cos x$ 的多项式拟合的模型最佳超参数的选取, 及对应模型训练集与测试集RMSE

类似于3.2节, 笔者进行了2.2.3节的实验, 所得的不同次数下, 关于 $\cos x$ 的多项式拟合模型最佳超参数及系数如表3所示, 相应的训练集与测试集RMSE如表4所示, 模型拟合曲线如图12所示。

表3 不同次数下, 关于 $\cos x$ 的多项式拟合的模型最佳超参数及各项系数

最高次数	α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
0	1.000×10^{-23}	2.743				
1	1.000×10^{-23}	2.897	-5.840			
2	5.952×10^{-2}	-1.711	-7.181	8.897		
3	6.305×10^{-2}	-1.283	-4.579	8.534	-3.196	
4	7.712×10^{-2}	-9.858×10^{-1}	-5.360	4.267	-2.307	4.429

表4 不同次数下, 关于 $\cos x$ 的多项式拟合的模型最佳超参数、训练集RMSE与测试集RMSE

最高次数	α	训练集RMSE	测试集RMSE
0	1.000×10^{-23}	6.396	6.841
1	1.000×10^{-23}	4.810	5.076
2	5.952×10^{-2}	3.393	4.195
3	6.305×10^{-2}	3.364	3.960
4	7.712×10^{-2}	3.350	3.995

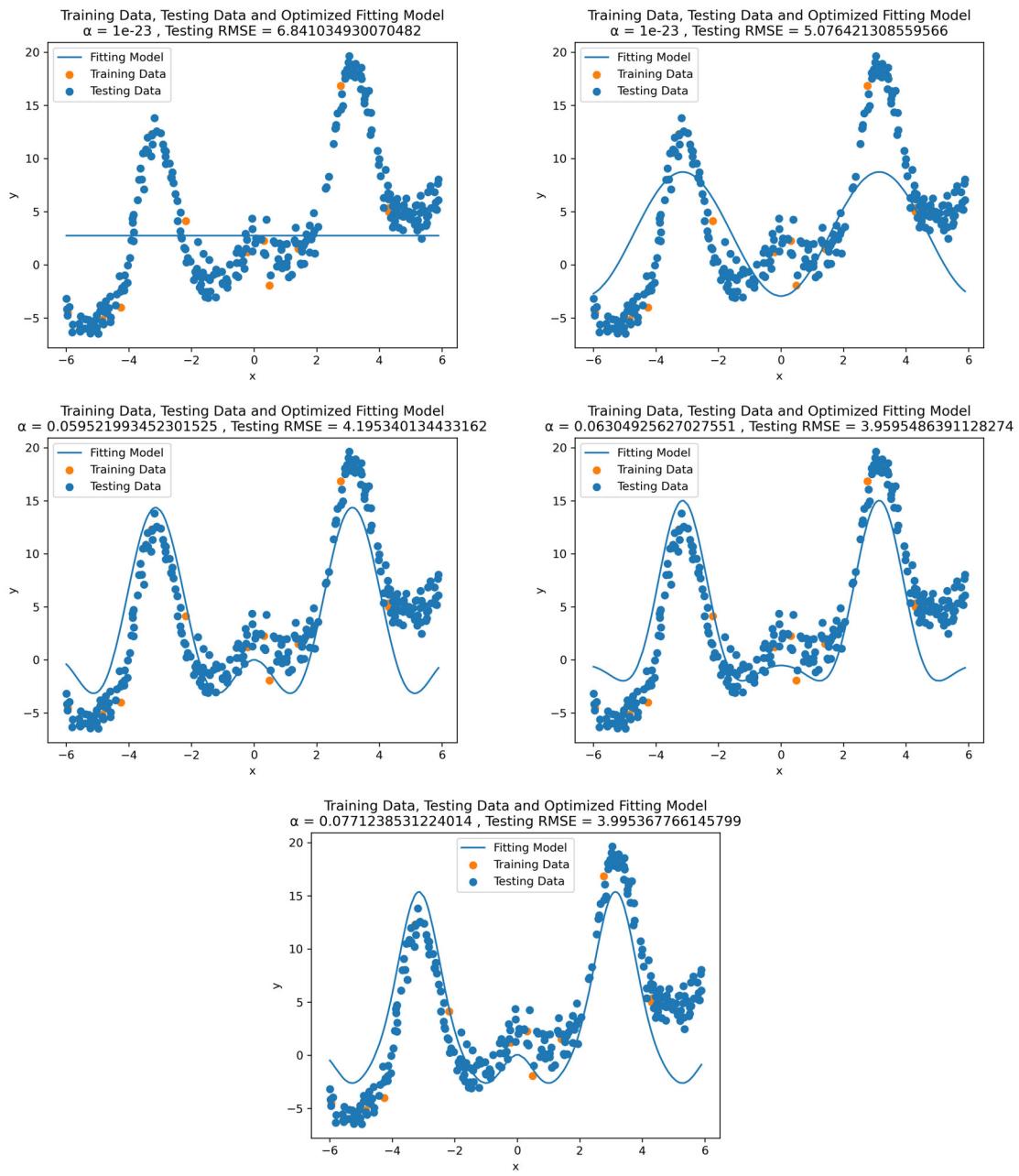


图12 不同次数下，关于 $\cos x$ 的多项式拟合的模型曲线（从左到右，从上到下，次数依次为0、1、2、3、4）

3.5 特殊模型 $y = c_0 + x + c_1 \cos x + c_2 \cos^2 x + c_3 \cos^3 x + c_4 \cos^4 x$ 的岭回归拟合及超参数 α 的手动选取

按照2.2.4节所述流程，笔者对特殊模型进行岭回归拟合，并手动选取了最佳超参数 α ，其各项系数如表5所示，训练集及测试集RMSE如表6所示，最终的拟合曲线如图13所示。

表5 特殊模型最佳超参数及各项系数

α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
1.150×10^{-1}	-9.122×10^{-1}	-5.133	4.271	-2.525	4.285

表6 特殊模型最佳超参数、训练集RMSE与测试集RMSE

α	训练集RMSE	测试集RMSE
1.150×10^{-1}	1.675	1.627

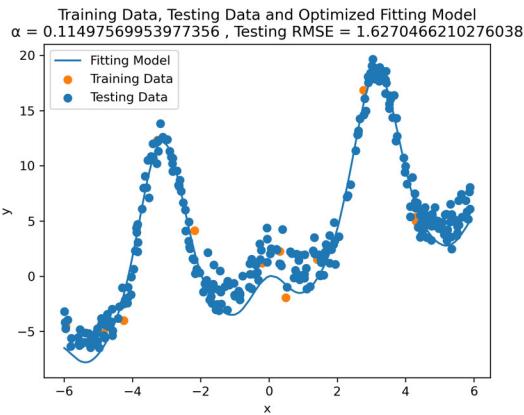


图13 特殊模型的拟合曲线

3.6 (附加实验) 合并训练集与测试集, 运用交叉验证进行关于 x 与关于 $\cos x$ 的多项式拟合

按2.2.5节所述流程, 笔者首先进行关于 x 的多项式拟合, 其中交叉验证方式为5折验证 (5-fold cross validation), 最终得不同次数下, 关于 x 的多项式拟合模型的各项系数, 如表7所示, 相应模型的原训练集、原测试集RMSE如表8所示, 最终的拟合曲线如图14所示。

表7 采用交叉验证后, 不同次数下, 关于 x 的多项式拟合的模型最佳超参数及各项系数

最高次数	α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
0	1.000×10^{-23}	4.225				
1	1.000×10^2	3.974	9.619×10^{-1}			
2	1.000×10^2	5.159	9.795×10^{-1}	-9.384×10^{-2}		
3	1.000×10^2	5.157	1.095	-9.351×10^{-2}	-5.561×10^{-3}	
4	1.000×10^2	1.635	1.045	8.300×10^{-1}	-4.839×10^{-3}	-3.016×10^{-2}

表8 采用交叉验证后, 不同次数下, 关于 x 的多项式拟合的模型最佳超参数、原训练集RMSE与原测试集RMSE

最高次数	α	原训练集RMSE	原测试集RMSE
0	1.000×10^{-23}	6.566	6.666
1	1.000×10^2	5.535	5.672
2	1.000×10^2	5.518	5.582
3	1.000×10^2	5.553	5.570
4	1.000×10^2	5.435	4.639

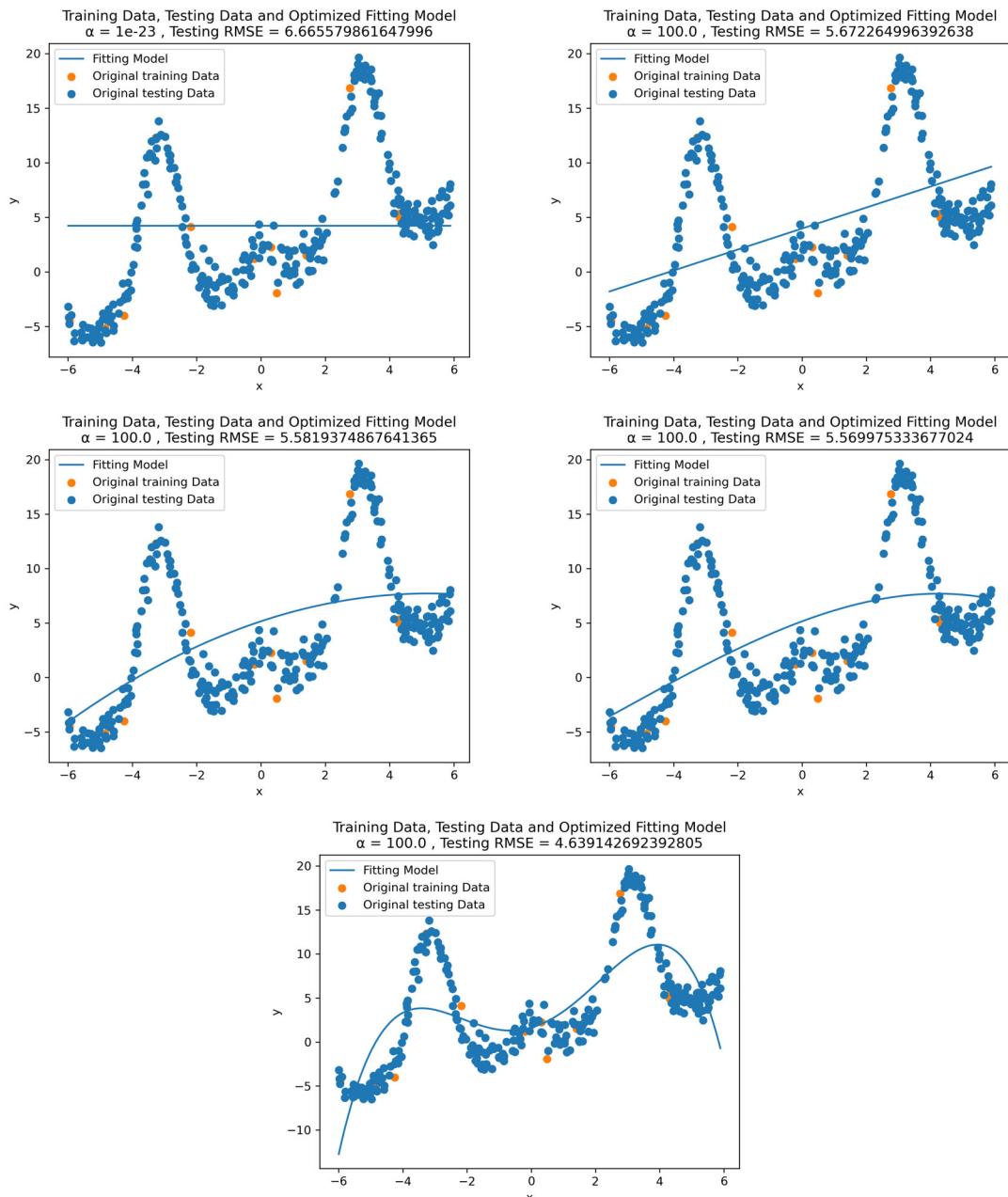


图14 采用交叉验证后，不同次数下，关于x的多项式拟合的模型曲线（从左到右，从上到下，次数依次为0、1、2、3、4）

接下来是关于 $\cos x$ 的多项式拟合，其中交叉验证方式仍为5折验证，最终得不同次数下，关于 $\cos x$ 的多项式拟合模型的各项系数，如表9所示，相应模型的原训练集、原测试集RMSE如表10所示，最终的拟合曲线如图15所示。

表9 采用交叉验证后，不同次数下，关于 $\cos x$ 的多项式拟合的模型最佳超参数及各项系数

最高次数	α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
0	1.000×10^{-23}	4.225				
1	2.251×10^1	3.748	-5.680			
2	1.325	-1.943×10^{-1}	-5.684	8.557		
3	1.782×10^{-13}	-2.154×10^{-1}	-1.084	8.296	-6.311	
4	1.000×10^{-23}	6.939×10^{-1}	-1.237	-2.371×10^{-1}	-5.829	8.881

表10 采用交叉验证后，不同次数下，关于 $\cos x$ 的多项式拟合的模型最佳超参数、原训练集RMSE与原测试集RMSE

最高次数	α	原训练集RMSE	原测试集RMSE
0	1.000×10^{-23}	6.566	6.666
1	2.251×10^1	4.887	5.013
2	1.325	3.828	3.856
3	1.782×10^{-13}	3.557	3.709
4	1.000×10^{-23}	3.729	3.619

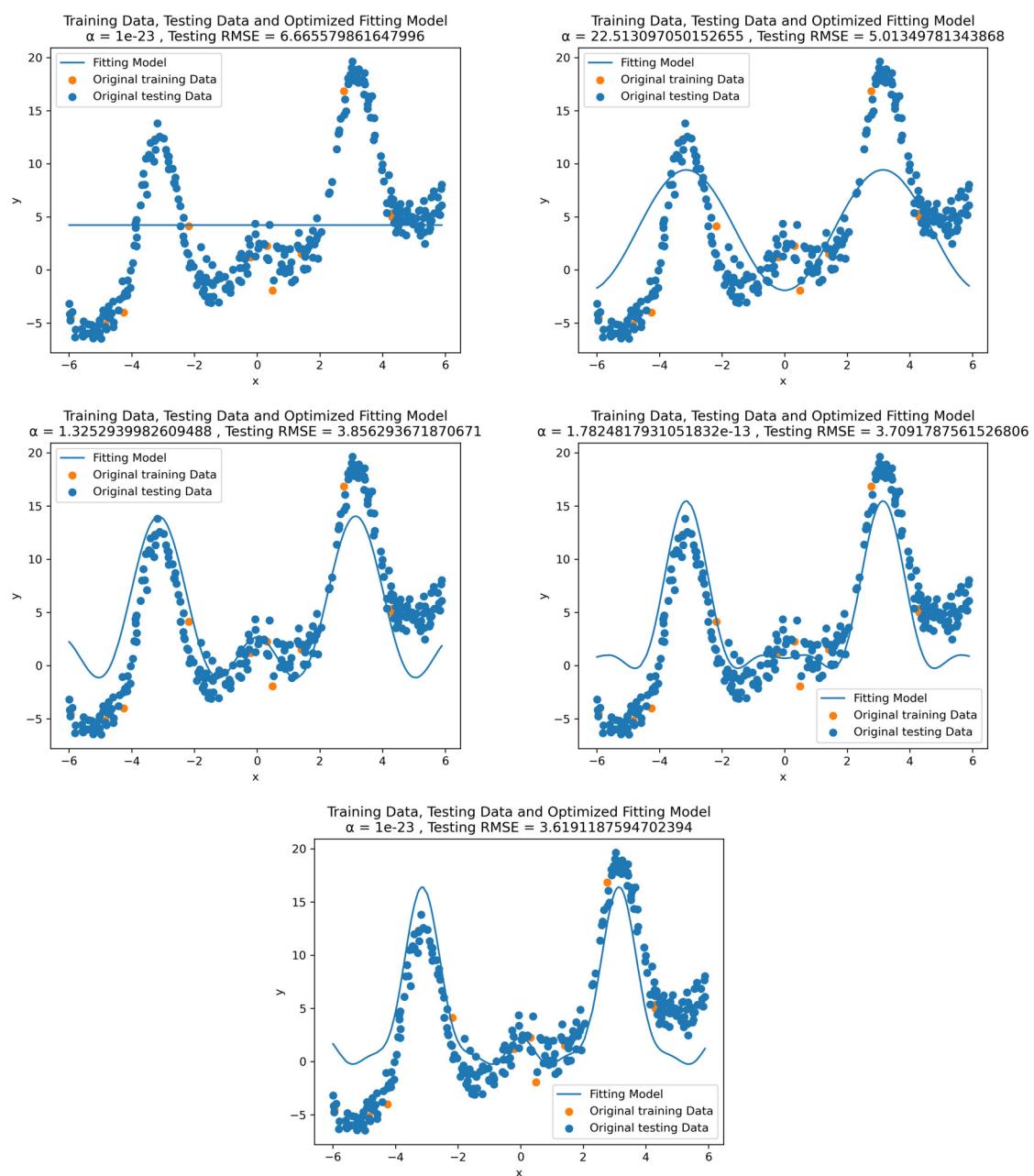


图15 采用交叉验证后，不同次数下，关于 $\cos x$ 的多项式拟合的模型曲线（从左到右，从上到下，次数依次为0、1、2、3、4）

3.7 (附加实验) 合并训练集与测试集, 运用交叉验证进行任意次方的 x 与任意次方的 $\cos x$ 混合的多项式拟合

按2.2.5节所述流程, 笔者还进行了任意次方的 x 与任意次方的 $\cos x$ 混合的多项式拟合, 其中交叉验证方式仍为5折验证, 最终得任意次方的 x 与任意次方的 $\cos x$ 混合的多项式拟合模型的各项系数, 如表11所示 (记 x 与 $\cos x$ 的最高次数为 $(p_{\text{lin}}, p_{\cos})$, 下同), 相应模型的原训练集、原测试集RMSE如表12所示, 最终的拟合曲线如图16至图19所示。

表11 采用交叉验证后, 任意次方的 x 与任意次方的 $\cos x$ 混合的多项式拟合模型最佳超参数及各项系数

$(p_{\text{lin}}, p_{\cos})$	α	$C[0]$	$C_{\text{lin}}[1]$	$C_{\text{lin}}[2]$	$C_{\text{lin}}[3]$	$C_{\text{lin}}[4]$
(1,1)	1.000×10^{-23}	3.414	9.884×10^{-1}			
(1,2)	1.000×10^{-23}	-6.813×10^{-1}	1.008			
(1,3)	1.000×10^{-23}	-5.926×10^{-1}	9.933×10^{-1}			
(1,4)	8.958×10^{-2}	8.270×10^{-2}	9.848×10^{-1}			
(2,1)	1.202×10^1	3.948	9.925×10^{-1}	-3.825×10^{-2}		
(2,2)	1.000×10^{-23}	-6.200×10^{-1}	1.009	-4.366×10^{-3}		
(2,3)	5.125×10^{-2}	-4.264×10^{-1}	9.953×10^{-1}	-1.142×10^{-2}		
(2,4)	3.526×10^{-1}	1.223×10^{-1}	9.864×10^{-1}	-5.386×10^{-3}		
(3,1)	1.000×10^{-23}	3.841	1.182	-3.314×10^{-2}	-8.773×10^{-3}	
(3,2)	3.555×10^{-15}	-6.120×10^{-1}	1.096	-4.284×10^{-3}	-4.085×10^{-3}	
(3,3)	3.555×10^{-15}	-4.270×10^{-1}	1.054	-1.130×10^{-2}	-2.739×10^{-3}	
(3,4)	1.171×10^{-12}	1.597×10^{-1}	9.863×10^{-1}	-4.920×10^{-3}	-3.662×10^{-5}	
(4,1)	1.138×10^{-12}	6.187	1.184	-7.823×10^{-1}	-7.642×10^{-3}	2.537×10^{-2}
(4,2)	4.288×10^{-1}	-7.101×10^{-1}	1.096	3.114×10^{-2}	-4.129×10^{-3}	-1.207×10^{-3}
(4,3)	2.112×10^{-1}	-5.111×10^{-1}	1.054	1.406×10^{-2}	-2.779×10^{-3}	-8.559×10^{-4}
(4,4)	5.718×10^{-1}	4.237×10^{-2}	9.946×10^{-1}	1.055×10^{-2}	-3.839×10^{-4}	-5.433×10^{-4}

$(p_{\text{lin}}, p_{\cos})$	$C_{\cos}[1]$	$C_{\cos}[2]$	$C_{\cos}[3]$	$C_{\cos}[4]$
(1,1)	-6.592			
(1,2)	-5.698	9.041		
(1,3)	-2.004	8.616	-5.039	
(1,4)	-2.130	2.321	-4.665	6.546
(2,1)	-5.996			
(2,2)	-5.689	9.029		
(2,3)	-1.951	8.570	-5.081	
(2,4)	-2.151	2.699	-4.632	6.124
(3,1)	-6.503			
(3,2)	-5.685	9.012		
(3,3)	-1.955	8.571	-5.071	
(3,4)	-2.082	2.223	-4.715	6.634
(4,1)	-9.692			
(4,2)	-5.518	8.991		
(4,3)	-1.896	8.583	-5.001	
(4,4)	-2.124	2.945	-4.575	5.888

表12 采用交叉验证后，任意次方的x与任意次方的 $\cos x$ 混合的多项式拟合模型最佳超参数、原训练集RMSE与原测试集RMSE

$(p_{\text{lin}}, p_{\cos})$	α	原训练集RMSE	原测试集RMSE
(1,1)	1.000×10^{-23}	4.253	3.488
(1,2)	1.000×10^{-23}	1.948	1.405
(1,3)	1.000×10^{-23}	1.480	1.135
(1,4)	8.958×10^{-2}	1.445	9.729×10^{-1}
(2,1)	1.202×10^1	4.075	3.497
(2,2)	1.000×10^{-23}	1.949	1.405
(2,3)	5.125×10^{-2}	1.478	1.129
(2,4)	3.526×10^{-1}	1.444	9.727×10^{-1}
(3,1)	1.000×10^{-23}	4.118	3.465
(3,2)	3.555×10^{-15}	1.965	1.397
(3,3)	3.555×10^{-15}	1.480	1.125
(3,4)	1.171×10^{-12}	1.442	9.716×10^{-1}
(4,1)	1.138×10^{-12}	3.496	3.235
(4,2)	4.288×10^{-1}	1.999	1.395
(4,3)	2.112×10^{-1}	1.509	1.123
(4,4)	5.718×10^{-1}	1.465	9.727×10^{-1}

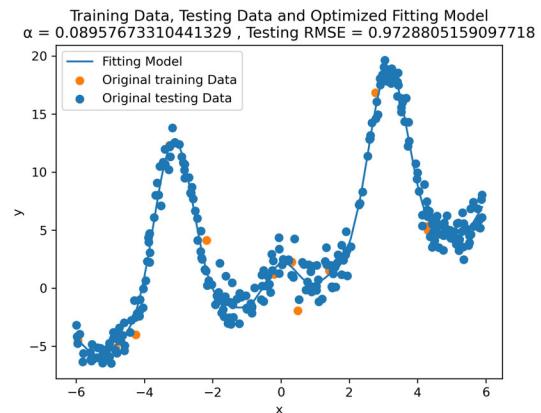
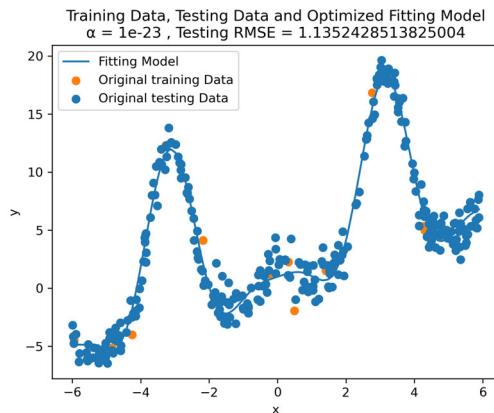
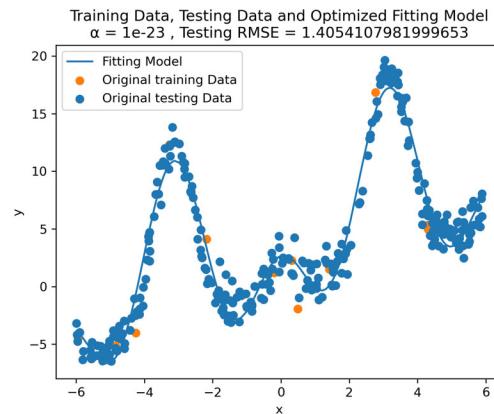
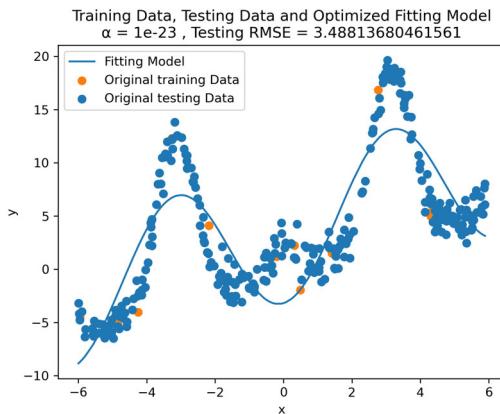


图16 采用交叉验证后, x 最高次数为1时, 混合多项式模型的拟合曲线 (从左到右, 从上到下, $\cos x$ 次数依次为1、2、3、4)

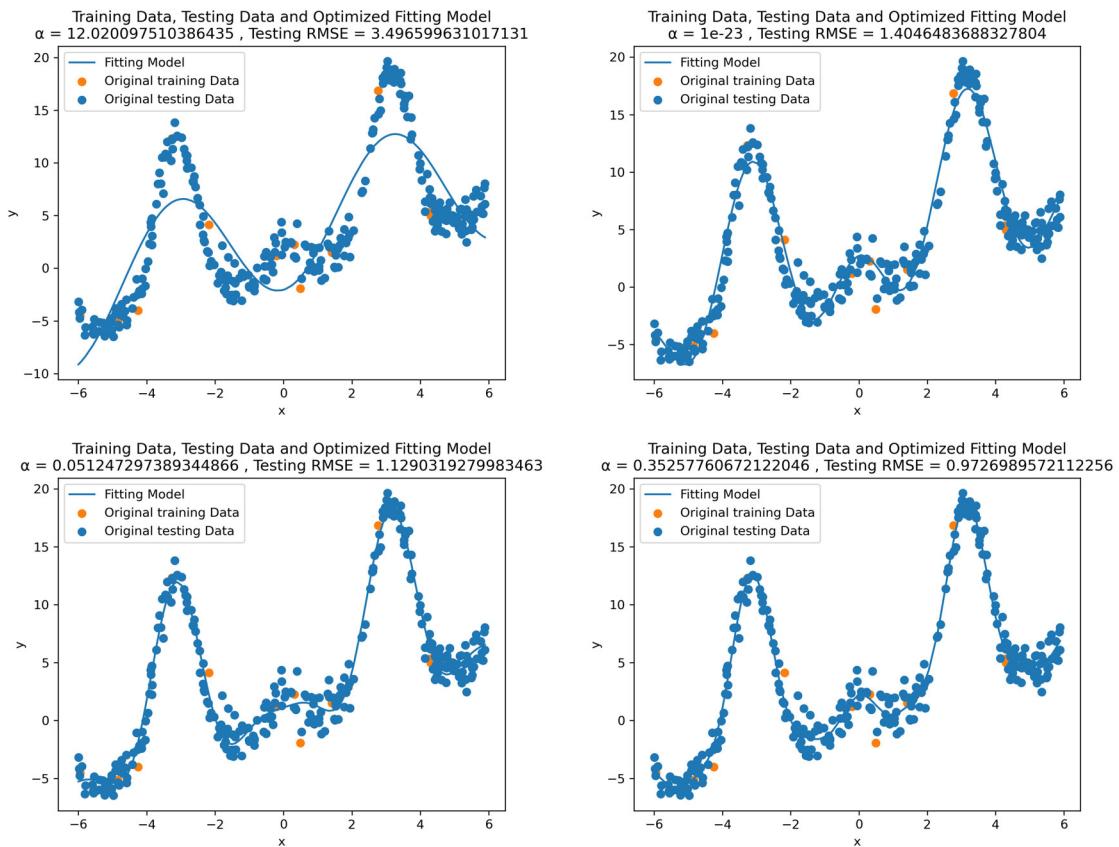


图17 采用交叉验证后, x 最高次数为2时, 混合多项式模型的拟合曲线 (从左到右, 从上到下, $\cos x$ 次数依次为1、2、3、4)

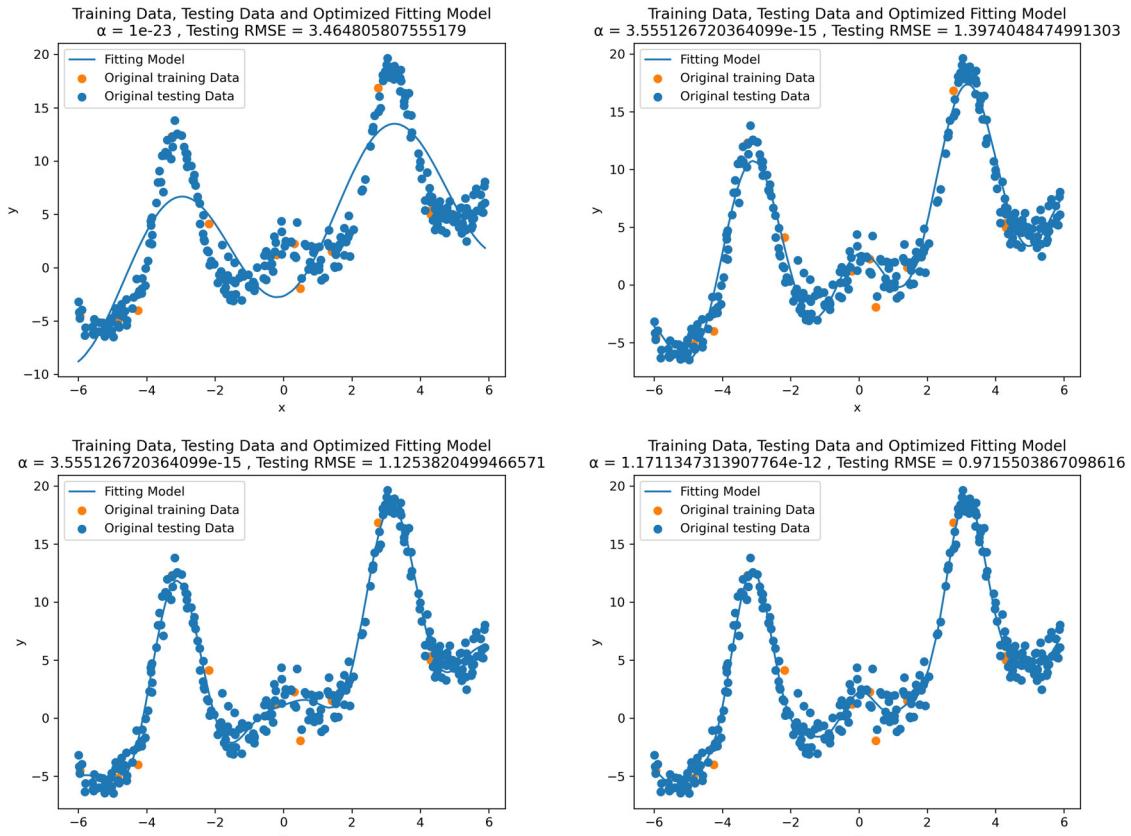


图18 采用交叉验证后, x 最高次数为3时, 混合多项式模型的拟合曲线 (从左到右, 从上到下, $\cos x$ 次数依次为1、2、3、4)

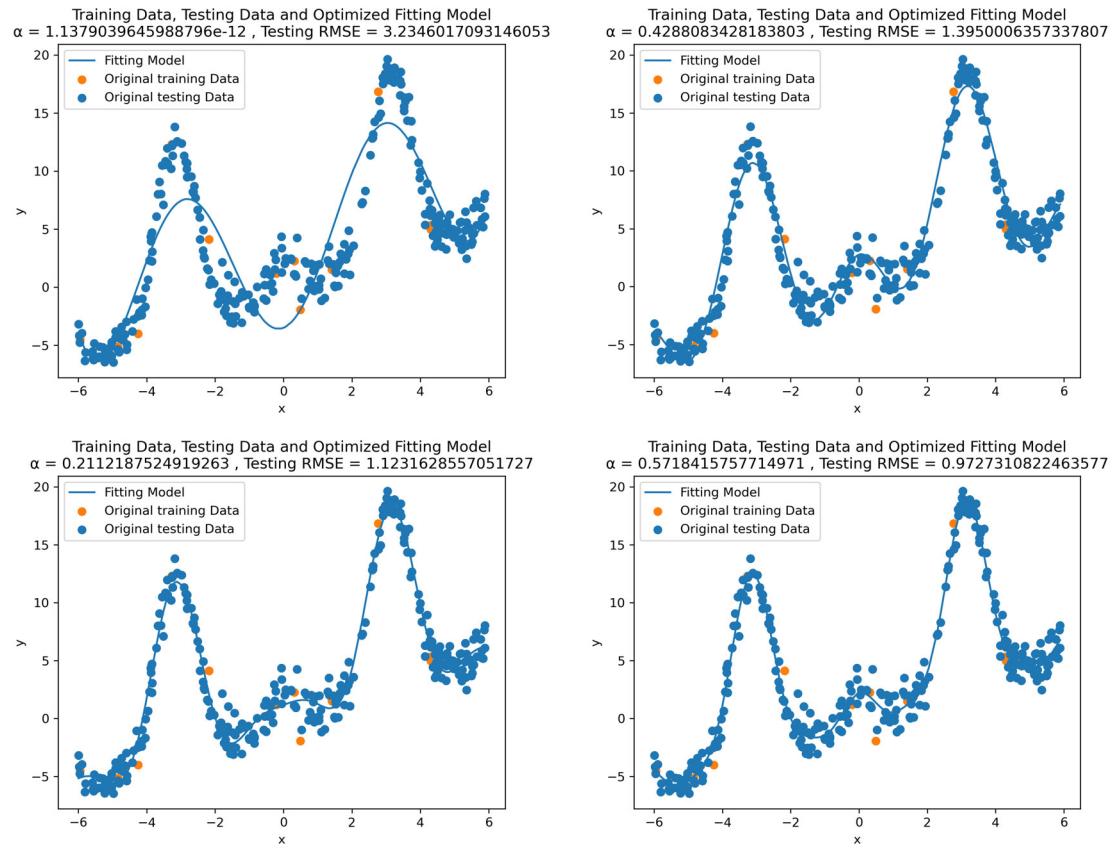


图19 采用交叉验证后, x 最高次数为4时, 混合多项式模型的拟合曲线 (从左到右, 从上到下, $\cos x$ 次
数依次为1、2、3、4)

3.8 (附加实验) 合并训练集与测试集, 运用交叉验证, 对2.2.4节的特殊模型进行自动化岭回归拟合

按2.2.6节所述流程, 笔者对特殊模型进行自动化岭回归拟合, 笔者对特殊模型进行岭回归拟合, 并自动选取最佳超参数 α , 其中交叉验证方式仍为5折验证。最终各项系数如表13所示, 训练集及测试集RMSE如表14所示, 拟合曲线如图20所示。

表13 采用交叉验证后, 特殊模型最佳超参数及各项系数

α	$C[0]$	$C[1]$	$C[2]$	$C[3]$	$C[4]$
1.000×10^{-23}	6.939×10^{-1}	-1.237	-2.371×10^{-1}	-5.829	8.881

表14 特殊模型最佳超参数、训练集RMSE与测试集RMSE

α	训练集RMSE	测试集RMSE
1.000×10^{-23}	1.489	1.052

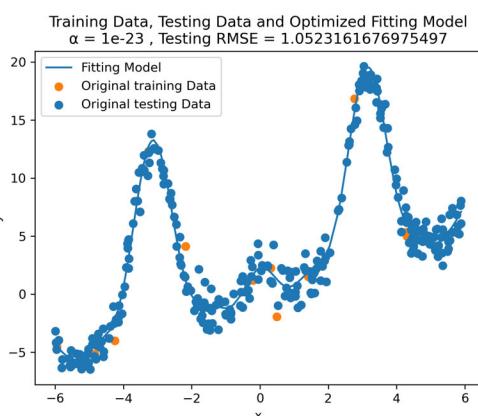


图20 特殊模型的拟合曲线

4 分析与讨论

4.1 不同次数的多项式模型随超参数 α 的变化趋势

从3.1、3.3节的数据可知，对于关于 x 的多项式拟合模型而言，若最高次数不超过3，则提高超参数 α 的值非但不能减小拟合误差，反而导致训练时与测试时的误差皆有上升，考虑到 α 的作用是抑制系数过大的拟合项，消除过拟合，而模型本身参数较少时，加入 α 反而会进一步抹消参数，加剧模型的欠拟合，因此当 α 增大时，便有模型误差增大的可能。而最高次数为4时，随着 α 的增大，训练集RMSE先保持不变，至 $\alpha = 1$ 逐渐上升，而测试集RMSE先保持不变，至 $\alpha = 0.001$ 开始迅速下降，随后在 $\alpha = 10$ 附近反弹，表明适当增大 α 对该模型有减小过拟合的作用，另一方面也表明关于 x 的多项式模型至少要四次方才能刻画出数据的性质。

对于关于 $\cos x$ 的多项式拟合模型而言，情况略有不同。若最高次数不超过1，则提高超参数 α 的值同样导致训练时与测试时的误差皆有上升，而最高次数不小于2时，随着 α 的增大，训练集RMSE逐渐上升，而测试集RMSE先保持不变，至 $\alpha = 0.001 \sim 0.01$ 时开始略有下降，随后迅速上升。上述现象同样表明，适当增大 α 对该模型有减小过拟合的作用，另一方面也表明，关于 $\cos x$ 的多项式模型至少要二次方才能刻画出数据的性质，比关于 x 的多项式模型最少参数量还要少。

4.2 手动选取最佳超参数 α 时，不同次数的多项式模型比较

观察3.2节的数据，笔者发现，尽管关于 x 的四次多项式模型体现了数据的分布趋势（测试集随 x 的增大形成两个凸峰），但其测试集RMSE却大于关于 x 的二次多项式模型。查看图6发现，二次多项式模型虽然没有显示出两个凸包，但整条曲线却与数据点“基线”保持较好的一致，而四次多项式模型虽然展现出两个凸峰，但是两侧曲线却是向下延伸，与数据点的距离愈发增大，这一趋势与数据点是相悖的（测试集“基线”随 x 的增大应逐渐增大），因此导致了测试集RMSE的增大。当然，如果进一步增大惩罚项，测试集RMSE还有减小的空间，但这样的举措也隐含偏离训练集的风险，两者仍需稍加权衡。

相较而言，3.4节的数据也有类似的现象，不过差异不明显。其中，最高次数为2的关于 $\cos x$ 的多项式模型，已经较为准确地刻画出测试集的两个凸峰，而最高次数为3、4的关于 $\cos x$ 的多项式模型，则仅仅在两侧曲线的处理上略有差异，虽然最高次数为3时的模型，具有最小的测试集RMSE，但与最高次数为4时的模型相比，差距并不算非常显著。

4.3 合并数据集结合交叉验证对模型拟合的影响

比较3.2、3.4、3.6节的数据，笔者还发现，在合并训练集与测试集，并结合交叉验证，重新划分数据与训练后，得到的关于 x 和关于 $\cos x$ 的多项式拟合模型，其对（原）训练集和（原）测试集的偏差均有显著下降，这与对数据的利用，以及模型评价的方法有关：采用多折交叉验证，意味着随机用一部分数据训练，另一部分数据测试，因此消除了训练时采样的偏差；而增大样本数，则保证随机划分后，训练集与测试集仍有足够多的数据，避免了多折交叉验证训练及评价的失真。

4.4 实验结论

(1) 在最高指数项不超过4的情况下，若采用“训练集误差最小，且测试集误差尽可能小”的规则，则优化后，关于 x 的四次多项式模型为

$y = 2.037 + 1.513x + 0.6517x^2 - 0.05530x^3 - 0.02858x^4$ ，规范化系数为 $\alpha = 1.380$ ，训练集、测试集RMSE分别为4.962、5.849。另一方面，仅从测试集RMSE来看，关于 x 的多项式拟合的最佳模型为二次多项式模型，不过从曲线形状看，四次多项式模型的曲线形状更接近样本数据点的变化趋势；

(2) 在最高指数项不超过4的情况下，按照(1)所述的原则，关于 $\cos x$ 的四次多项式模型为 $y = -0.9858 - 5.360 \cos x + 4.267 \cos^2 x - 2.307 \cos^3 x - 4.429 \cos^4 x$ ，规范化系数为 $\alpha = 0.07712$ ，训练集、测试集RMSE分别为3.350、3.995。同时可知，仅从测试集RMSE来看，关于 $\cos x$ 的多项式拟合的最佳模型为三次多项式模型，而从曲线形状看，四次多项式模型的曲线形状与更接近数据点分布的形状；

(3) 仍按照 (1) 所述的原则, 对特殊模型进行岭回归拟合, 则最佳模型应为

$$y = -0.9122 + x - 5.133 \cos x + 4.271 \cos^2 x - 2.525 \cos^3 x + 4.285 \cos^4 x, \text{ 规范化系数为 } \alpha = 1.150 \times 10^{-1}, \text{ 训练集、测试集RMSE分别为} 1.675, 1.627;$$

(4) 在最高指数项不超过4的情况下, 如果将训练集和测试集合并, 再进行交叉验证, 然后去测试原测试集数据, 则关于 x 的四次多项式模型为

$y = 1.635 + 1.045x + 0.8300x^2 - 0.004839x^3 - 0.03016x^4$, 规范化系数为 $\alpha = 100.0$, 原训练集、原测试集RMSE分别为5.435、4.639。上述模型也是关于 x 的多项式拟合的最佳模型, 且该模型优于未合并数据且未交叉验证的模型, 表明增加训练数据量, 并结合交叉验证, 可以改善模型质量;

(5) 在最高指数项不超过4的情况下, 按照 (4) 所述的原则, 关于 $\cos x$ 的四次多项式模型为
 $y = 0.6939 - 1.237 \cos x - 0.2371 \cos^2 x - 5.829 \cos^3 x + 8.881 \cos^4 x$, 规范化系数为
 $\alpha = 1.000 \times 10^{-23}$, 原训练集、原测试集RMSE分别为3.729、3.619, 它也是关于 $\cos x$ 的多项式拟合的最佳模型, 且该模型优于未合并数据且未交叉验证的模型;

(6) 在最高指数项不超过4的情况下, 按照 (4) 所述的原则, 可得任意次方的 x 与任意次方的 $\cos x$ 混合的多项式拟合最佳模型, 其 x 最高次数为3, $\cos x$ 最高次数为4, 相应的表达式为

$$\begin{aligned} y = & 0.1597 + 0.9863x - 4.920 \times 10^{-3}x^2 - 3.662 \times 10^{-5}x^3 \\ & - 2.082 \cos x + 2.223 \cos^2 x - 4.715 \cos^3 x + 6.634 \cos^4 x \end{aligned}$$

该模型规范化系数为 $\alpha = 1.171 \times 10^{-12}$, 对应的原训练集、原测试集RMSE分别为1.442、0.9716;

(7) 仍按照 (4) 所述的原则, 对特殊模型进行岭回归拟合, 则最佳模型应为

$$y = 0.6939 + x - 1.237 \cos x - 0.2371 \cos^2 x - 5.829 \cos^3 x + 8.881 \cos^4 x, \text{ 规范化系数为 } \alpha = 1.000 \times 10^{-23}, \text{ 原训练集、原测试集RMSE分别为} 1.490, 1.052;$$

(8) 改变 α 值对欠拟合和较好拟合的模型的优化并不明显, 而对过拟合模型改善显著。