参考脚本2：

1. **实验代码**

|  |  |
| --- | --- |
|  | %2021-2022（1）智能仿真实验 实验一 |
|  | %日期：2021-10-20 |
|  | %姓名：谢晔辉 |
|  | %学号：1935031426 |
|  | %作业内容：以中美两国1980年至2016年的GDP历史数据为基础，用多项式拟合和最小二乘拟合二种方法进行曲线拟合，确定其数学模型，并给出拟合过程及分析 |
|  |  |
|  | clear |
|  | clc |
|  |  |
|  | %读入中美两国GDP数据 |
|  | t = (1980 : 2016)'; |
|  | fid = fopen('exp\_1\_CHN.txt'); |
|  | CHN = textscan(fid, '%n', 'commentStyle', '%'); |
|  | CHN = cell2mat(CHN) \* 1e-6; |
|  | fid = fopen('exp\_1\_USA.txt'); |
|  | USA = textscan(fid, '%n', 'commentStyle', '%'); |
|  | USA = cell2mat(USA) \* 1e-6; |
|  | fclose(fid); |
|  |  |
|  | %测试数据 |
|  | real\_CH = [12310409, 13894817, 14279937, 14722730] \* 1e-6; %2017 ~ 2020中国GDP真实值 |
|  | real\_US = [19542979, 20611860, 21433224, 20936600] \* 1e-6; %2017 ~ 2020美国GDP真实值 |
|  | test\_CH = zeros(6, 4); %预测值的残差 |
|  | test\_US = zeros(6, 4); |
|  |  |
|  | %计算拟合多项式的系数 |
|  | tm = (t(end) + t(1)) / 2; %年份的中间值 |
|  | r = (t(end) - t(1)) / 2; %年份区间的一半 |
|  | tt = (t - tm) / r; %统计年度变量变换为拟合自变量 |
|  | N = 6; %多项式的最高阶数 |
|  | Y\_CH = zeros(N, 37); |
|  | Y\_US = zeros(N, 37); |
|  | for n = 1 : N |
|  | Pc = polyfit(tt, CHN, n) %中国GDP拟合多项式的系数 |
|  | Pu = polyfit(tt, USA, n) %美国GDP拟合多项式的系数 |
|  | PCH = vpa(poly2sym(Pc), 5) %中国GDP拟合多项式，保留5位有效数字 |
|  | PUS = vpa(poly2sym(Pu), 5) %美国GDP拟合多项式，保留5位有效数字 |
|  | yCH\_hat = polyval(Pc, tt); |
|  | yUS\_hat = polyval(Pu, tt); |
|  | Y\_CH(n, :) = (CHN - yCH\_hat)'; |
|  | Y\_US(n, :) = (USA - yUS\_hat)'; |
|  |  |
|  | %在指定区间上计算拟合多项式的拟合数据点 |
|  | Ty = (1980 : 2050)'; %拟合曲线扩展的年度区间 |
|  | T = (Ty - tm) / r; %相应的拟合自变量的取值范围 |
|  | Vc = polyval(Pc, T); %扩展区间上中国GDP曲线的数据值 |
|  | Vu = polyval(Pu, T); %扩展区间上美国GDP曲线的数据值 |
|  | ym = max([Vc; Vu]); %区间内最大的曲线值 |
|  |  |
|  | test\_CH(n, :) = Vc(38 : 41)' - real\_CH; %2017 ~ 2020的预测值减去实际值 |
|  | test\_US(n, :) = Vu(38 : 41)' - real\_US; |
|  |  |
|  | %绘制统计数据点图和拟合曲线 |
|  | % figure |
|  | subplot(2, 3, n) |
|  | hold on %允许叠绘 |
|  | tb = (t(end) + 0.5 - tm) / r; %统计区与扩展区分界线位置 |
|  | fh = fill([tb, tb, T(end), T(end)], [0, ym, ym, 0], 'y', 'EdgeColor','none'); %扩展区涂黄色，其区界无色 |
|  | alpha(0.2) %使黄色半透明淡化 |
|  | plot([tb, tb], [0, ym], '--m', 'LineWidth', 3) %画粉色粗分界线 |
|  | scatter(tt, CHN, 'r.'); %中国统计数据点 |
|  | scatter(tt, USA, 'bx'); %美国统计数据点 |
|  | ph = plot(T, Vc, 'r', T, Vu, '--b'); %在扩展区画拟合曲线 |
|  | hold off %不再叠绘 |
|  | box on %使轴系坐标框封闭 |
|  | grid on, grid minor %显示网格并细化 |
|  | axis([T(1), T(end), 0, ym]) %控制坐标轴的范围 |
|  | TL = 1980 : 5 : 2050; %被标识的年份数组 |
|  | tL = (TL - tm) / r; %转换为横坐标的刻度位置 |
|  | xticks(tL); %对横轴刻度 |
|  | xticklabels(string(TL)); %显示横轴刻度的年份标识 |
|  | xtickangle(90); %使标识文字左旋90度 |
|  | legend([ph(1), ph(2)], 'China', 'USA', 'Location', 'NorthWest'); %显示拟合曲线图例 |
|  | text('string', "1980 ~ 2016", 'Units', 'normalized', 'position', [0.16, 0.75]); |
|  | text('string', "统计数据拟合区", 'Units', 'normalized', 'position', [0.14, 0.7]); |
|  | text('string', "2017 ~ 2050", 'Units', 'normalized', 'position', [0.66, 0.75]); |
|  | text('string', "统计数据拟合区", 'Units', 'normalized', 'position', [0.64, 0.7]); |
|  | xlabel('年'), ylabel('万亿美元'); |
|  | title(sprintf("中美两国GDP统计拟合曲线及延伸\n多项式阶数n = %d", n)); |
|  | end |

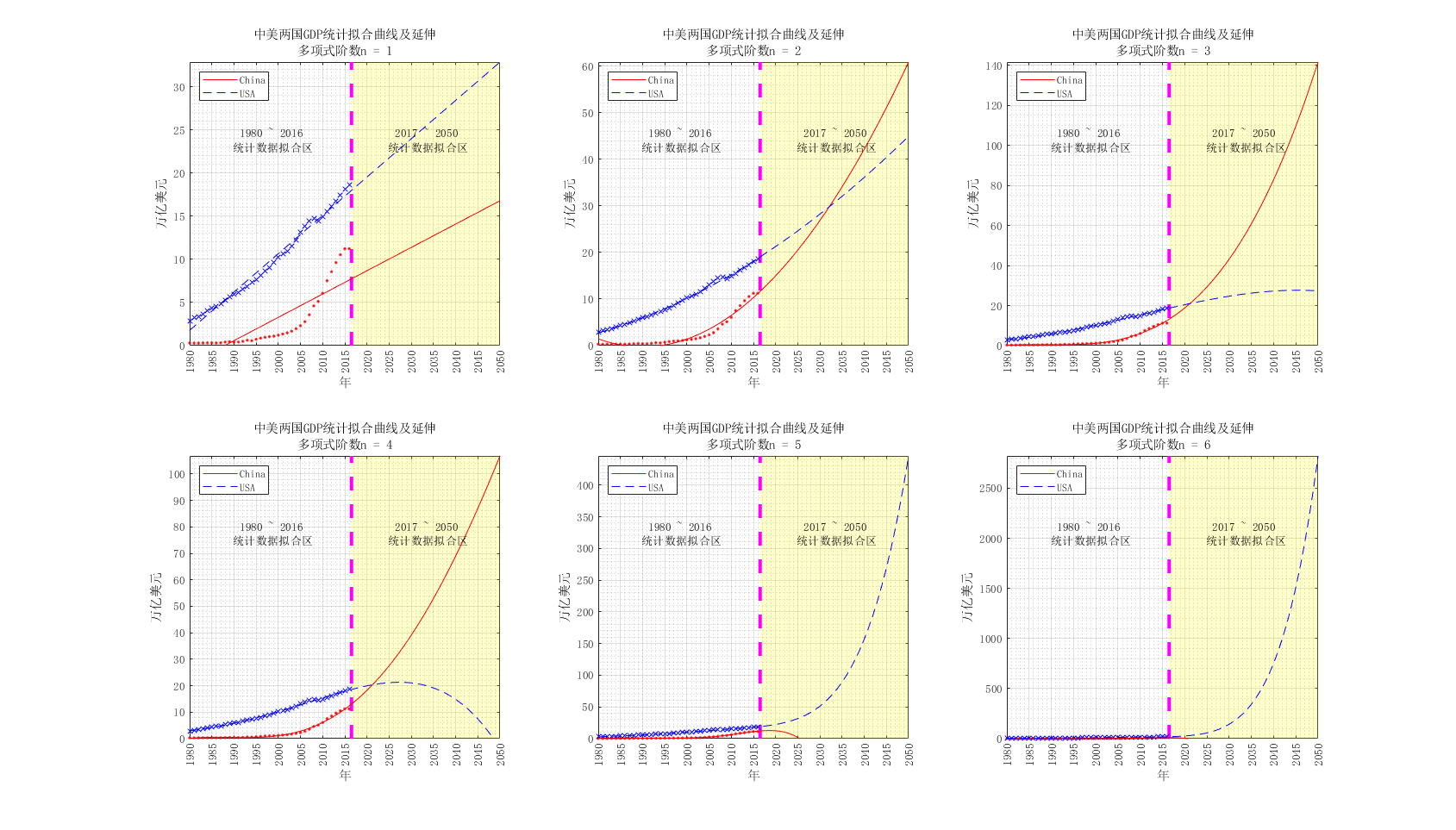
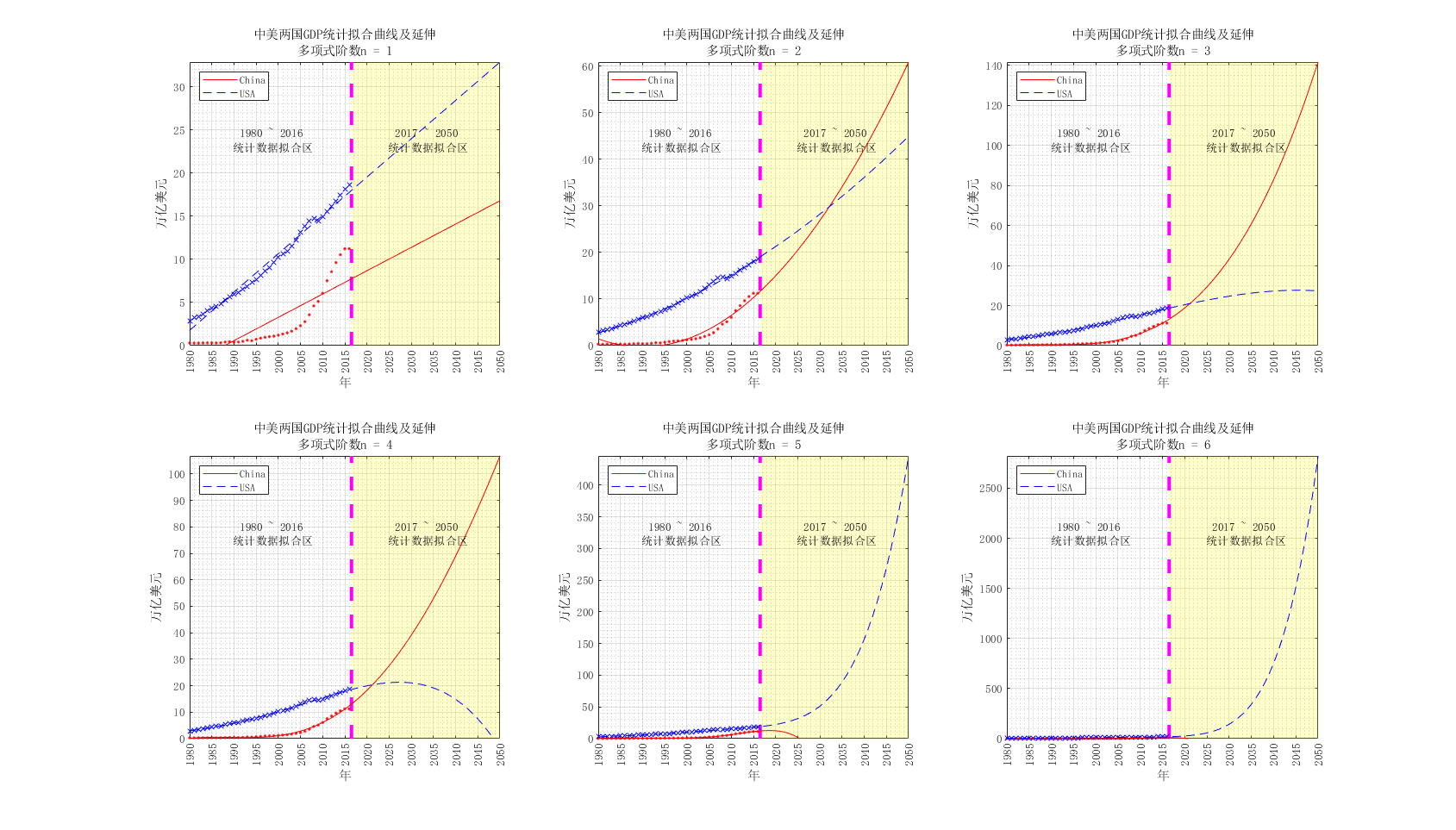
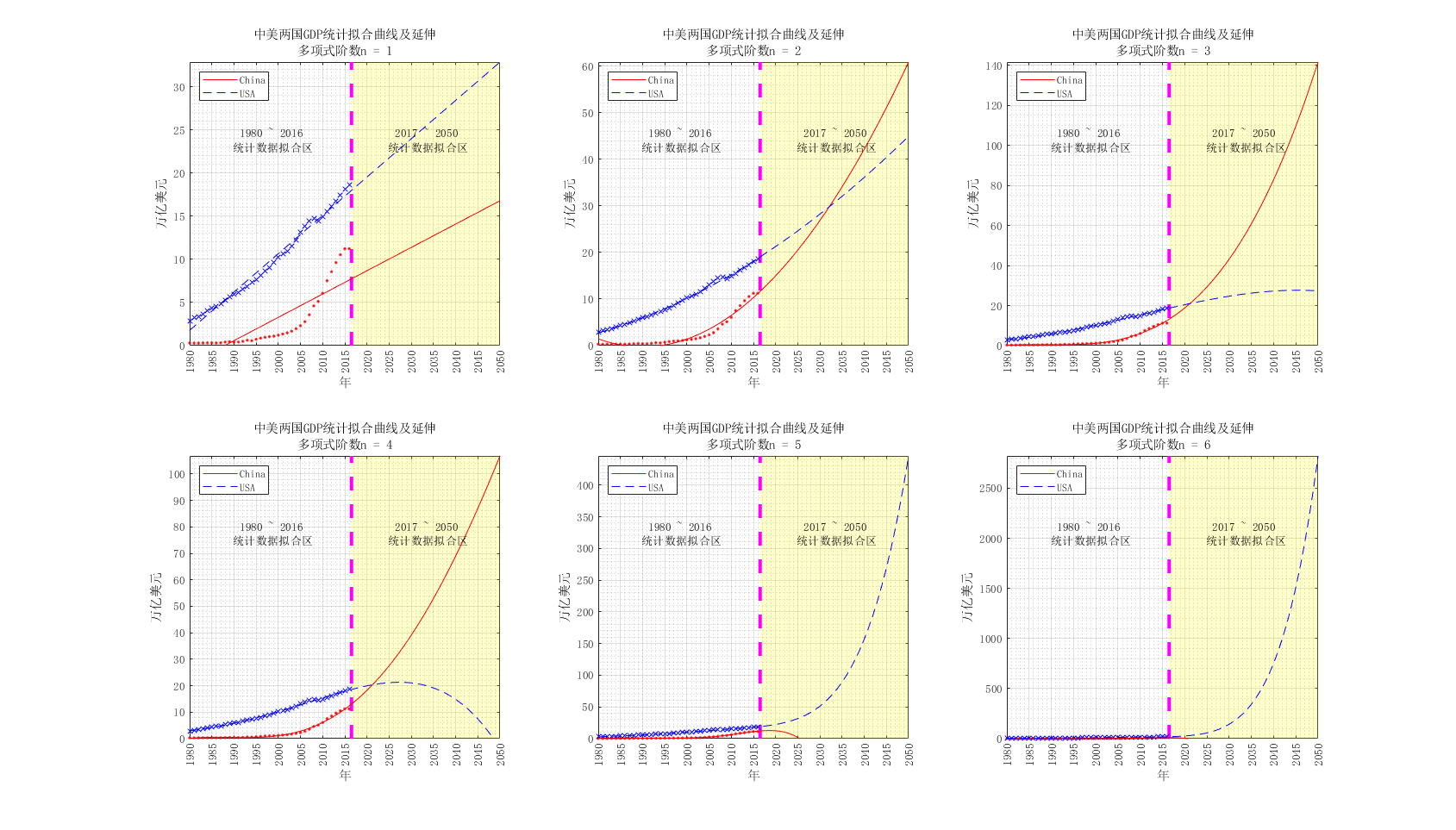
1. **实验结果**

图3.1 多项式阶数为1的拟合曲线 图3.2 多项式阶数为2的拟合曲线



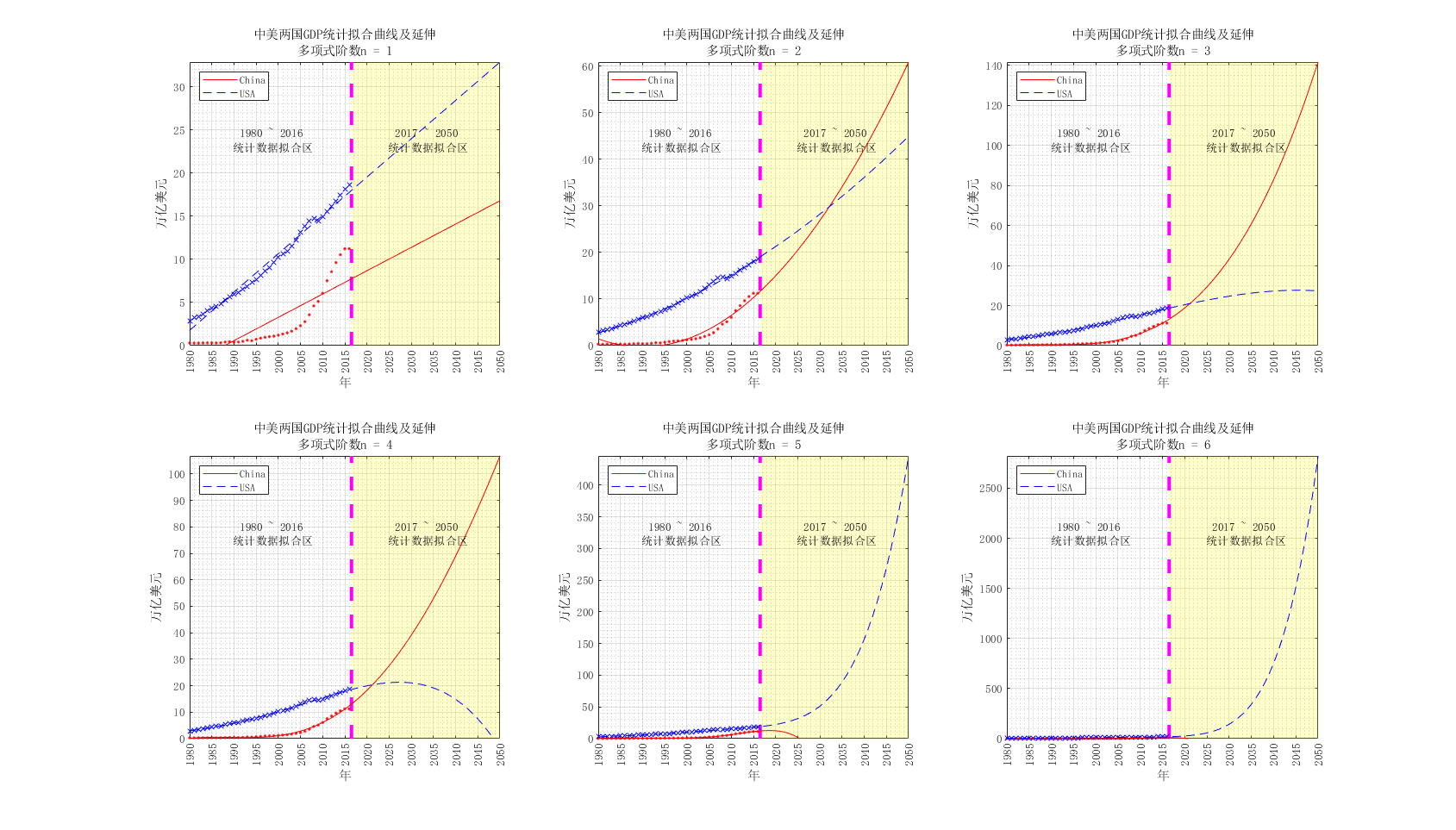
图3.3 多项式阶数为3的拟合曲线 图3.4 多项式阶数为4的拟合曲线

图3.5 多项式阶数为5的拟合曲线 图3.6 多项式阶数为6的拟合曲线

1. **拟合评价指标**

其中评价指标的定义如下：

拟合优度为将1减去剩余平方和占总离差平方和的比例定义为相关指数，记为

总离(偏)差平方和

Q为残差平方和(Sum of Squared Residuals)

均方误差(Mean Squared Error，MSE)

表3.1 残差平方和

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **多项式阶数** | **中国** | **美国** |
| **残差平方和** |  | 129.2074 | 11.19805 |
|  | 17.9929 | 2.854453 |
|  | 5.033087 | 2.236973 |
|  | 4.951182 | 2.171671 |
|  | 2.185224 | 1.713067 |
|  | 0.414128 | 1.288912 |

表3.2 均方误差（MSE）

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **多项式阶数** | **中国** | **美国** |
| **均方误差（MSE）** |  | 3.492091 | 0.30265 |
|  | 0.486295 | 0.077147 |
|  | 0.136029 | 0.060459 |
|  | 0.133816 | 0.058694 |
|  | 0.05905 | 0.046299 |
|  | 0.011193 | 0.034835 |

表3.3 拟合优度

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **多项式阶数** | **中国** | **美国** |
| **拟合优度** |  | 0.704591 | 0.986731 |
|  | 0.958862 | 0.996618 |
|  | 0.988493 | 0.997349 |
|  | 0.98868 | 0.997427 |
|  | 0.995004 | 0.99797 |
|  | 0.999053 | 0.998473 |

为了评估该拟合模型的预测能力（即泛化能力），我通过资料查询找到了中国与美国在2017 ~ 2020年的GDP数据（表3.4）来进行测试。

表3.4 2017 ~ 2020中美GDP数据

|  |  |  |
| --- | --- | --- |
| **年份** | **中国（单位：万亿美元）** | **美国（单位：万亿美元）** |
| **2017** | 12.310409 | 19.542979 |
| **2018** | 13.894817 | 20.611860 |
| **2019** | 14.279937 | 21.433214 |
| **2020** | 14.722730 | 20.936600 |

根据现实的数据得到在多项式阶数不同的情况下数学模型的预测能力，如表3.5所示。

表3.5 2017 ~ 2020年中美GDP预测评价

|  |  |  |  |
| --- | --- | --- | --- |
| **评价指标** | **多项式阶数** | **中国** | **美国** |
| **均方误差（MSE）** |  | 30.8873 | 3.240379 |
|  | 0.246607 | 0.295277 |
|  | 7.332244 | 1.024341 |
|  | 5.433973 | 1.719643 |
|  | 2.241563 | 0.545967 |
|  | 46.06533 | 9.546296 |

从表3.1、3.2和3.3可以看出，当多项式阶数越高，拟合的准确度是在不断提高的，但是这并不代表该数学模型具有很好的预测和泛化能力。如表3.5所示，对于中国和美国在2017 ~ 2020年的GDP数据，阶数为2阶的多项式在预测数据上的均方误差相对较小。而随着多项式的阶数不断升高，其均方误差也有明显的升高，明显表现出了过拟合。

1. **数学模型**

表3.6 数学模型

|  |  |  |
| --- | --- | --- |
| **多项式阶数** | **数学模型** | |
| **中国** | **美国** |
|  | 4.8654\*x + 2.7396 | 7.9977\*x + 9.7708 |
|  | 5.5151\*x^2 + 4.8654\*x + 0.79913 | 1.5106\*x^2 + 7.9977\*x + 9.2393 |
|  | 3.6242\*x^3 + 5.5151\*x^2 + 2.5723\*x + 0.79913 | - 0.79109\*x^3 + 1.5106\*x^2 + 8.4983\*x + 9.2393 |
|  | - 0.55955\*x^4 + 3.6242\*x^3 + 6.0201\*x^2 + 2.5723\*x + 0.746 | - 0.49963\*x^4 - 0.79109\*x^3 + 1.9615\*x^2 + 8.4983\*x + 9.1918 |
|  | - 6.3541\*x^5 - 0.55955\*x^4 + 11.044\*x^3 + 6.0201\*x^2 + 0.90306\*x + 0.746 | 2.5873\*x^5 - 0.49963\*x^4 - 3.8123\*x^3 + 1.9615\*x^2 + 9.178\*x + 9.1918 |
|  | - 9.9921\*x^6 - 6.3541\*x^5 + 13.725\*x^4 + 11.044\*x^3 + 1.0334\*x^2 + 0.90306\*x + 0.99444 | 4.8899\*x^6 + 2.5873\*x^5 - 7.4901\*x^4 - 3.8123\*x^3 + 4.4019\*x^2 + 9.178\*x + 9.0703 |

1. **拟合过程及结果分析**

为了使得程序更简洁，我将输入数据放在了exp\_1\_CHN.txt和exp\_1\_USA.txt，好处时修改文件内容较为方便，而且在txt文档中可以加入注释，在MATLAB读取数据的过程中可以屏蔽掉注释。

为了验证多项式阶数为多少时，拟合模型的性能较好，所以分别选取了多项式阶数为1 ~ 6进行实验拟合，通过for循环来实现。每次在循环体内 先在当前多项式阶数的情况下，根据输入数据的值分别得到中美GDP多项式的系数，接着再将输入年份带入多项式，把得到的值与原始GDP数值作差，即得到了残差，将此残差进行保存用作之后的评价指标计算，如表3.1、3.2和3.3所示。

作图上，使用了subplot函数就，将6张图显示在同一个窗口上，这样可以清晰看出多项式阶数不同时拟合曲线的对比。同时，因为不同图的坐标轴范围差异很大，所以图例的放置使用了比例来实现。

1. **最小二乘法拟合**

|  |  |
| --- | --- |
|  | %2021-2022（1）智能仿真实验 实验一 |
|  | %日期：2021-10-20 |
|  | %姓名：谢晔辉 |
|  | %学号：1935031426 |
|  | %作业内容：最小二乘拟合 |
|  |  |
|  | clear |
|  | clc |
|  |  |
|  | t = (1980 : 2016)'; |
|  | fid = fopen('exp\_1\_CHN.txt'); |
|  | CHN = textscan(fid, '%n', 'commentStyle', '%'); |
|  | CHN = cell2mat(CHN) \* 1e-6; |
|  | fid = fopen('exp\_1\_USA.txt'); |
|  | USA = textscan(fid, '%n', 'commentStyle', '%'); |
|  | USA = cell2mat(USA) \* 1e-6; |
|  | fclose(fid); |
|  |  |
|  | tm = (t(end) + t(1)) / 2; %年份的中间值 |
|  | r = (t(end) - t(1)) / 2; %年份区间的一半 |
|  | tt = (t - tm) / r; %统计年度变量变换为拟合自变量 |
|  |  |
|  | y\_1 = polyfit(tt, CHN, 3) |
|  | y\_hat\_1 = polyval(y\_1, t); |
|  | f1 = y\_1(1) \* tt.^3 + y\_1(2) \* tt.^2 + y\_1(3) \* tt + y\_1(4); |
|  | f = @(a, tt)a(1)\*tt.^3 + a(2)\*tt.^2 + a(3)\*tt + a(4); |
|  | [a, e] = lsqcurvefit(f, [1, 1, 1, 1], tt, f1) %最小二乘拟合 |
|  |  |
|  | y\_2 = polyfit(tt, USA, 3) |
|  | y\_hat\_2 = polyval(y\_2, t); |
|  | f2 = y\_2(1) \* tt.^3 + y\_2(2) \* tt.^2 + y\_2(3) \* tt + y\_2(4); |
|  | f = @(a, tt)a(1)\*tt.^3 + a(2)\*tt.^2 + a(3)\*tt + a(4); |
|  | [a, e] = lsqcurvefit(f, [1, 1, 1, 1], tt, f2) %最小二乘拟合 |