高斯过程回归与股票预测

吴同　　无 93　　2019013217

一、问题的引入

股票预测一直是非常热门的研究领域。人们收集股票的收盘价格等一系列的数据，通过回归等处理方法，根据已有的数据拟合出特定的模型，从而预测出未来的股票价格走势。最简单的回归方法是普通的线性回归模型，人们也可以通过改变数据处理方式同样利用线性回归方法实现非线性回归模型。但很重要的一点是它们都是确定性的回归。而近年来人工智能和机器学习领域的学者将更多的目光投向了所谓“随机回归模型”，高斯过程回归（Gaussian Processes Regression，GPR）就是其中的典型代表。我们就此方法进行了探索和研究。

二、GPR 原理

1. 传统回归原理

传统回归方法最初面临的问题是回归模型的选择。研究人员根据不同的数据关系，选择线性、指数、幂次等多种回归模型。模型的选择要由具体的数据关系以及最终的拟合效果决定。

实现了回归模型的选择后，完成相应的数据处理，基本所有回归模型的问题都会归结为线性回归模型的问题。之后进行的是参数的选择，绝大多数情况下采用的是最小二乘法：利用相对于回归模型的误差构建误差函数

$$e = \sum (y - y_{predict})^2$$

当选择的参数使得误差函数最小时，可以认为该回归模型拟合系列数据的效果较好。

构建了完成的回归模型后，就可以利用该模型对未来或者缺失的数据点进行预测。通过上述步骤可以看出，传统回归方法构建的模型，其本身是一个确定性的函数，选择参数最小化误差函数的过程本质上是在寻找拟合已有数据的最好模型。而 GPR 实现的方法和传统回

归方法并没有本质上的区别，但其实现的回归模型是随机的。

2. 高斯过程回归原理

高斯过程是这样一个随机过程：任取任意个时间点，它们上的随机变量均满足高斯/联合高斯分布。这揭示了高斯过程的一个本质，也是可以用于回归的本质：若我已经知道了某些时间点的信息，那么未知时间点的信息分布可以求出，并且不同于完全无信息的分布。因此，如果假设数据满足高斯过程，利用已有的数据拟合出一个高斯过程回归模型，那么对于未来或者缺失的数据，就可以有更加准确的预测。

在假设数据满足高斯过程的前提下，我们假设数据点和数据满足如下关系：

$$y = f(x)$$

其中，x 为数据点，y 为数据，f 为随机函数。那么，单从一个孤立的数据点上看，$f(x)$应该是一个高斯分布。

现在所能了解的仅有这些，因为目前我们所假设的仅有一个大前提——数据满足高斯过程，而我们对于这个高斯高斯过程本身并没有过多的了解。因此下面我们要对具体高斯过程进行分析。

对于一个高斯过程，一旦它的均值函数和协方差函数确定，整个高斯过程也就确定。对于均值函数，我们可以将其简单设定为 0，这是因为我们可以做出进一步假设：这些数据满足的是平稳高斯过程（大多数情况下效果足够好）。这意味着均值函数是一常数，而该常数我们又可以通过数据处理的方法将均值变为 0。因此，我们的主要任务在于选择协方差函数。而确定了协方差函数后，整个高斯过程框架也就确定。

确定了高斯过程框架后，回归模型并没有搭建完成，因为我们只是拥有大体框架，而对于具体的细节——各个参数未曾进行选择。参数的选择利用一般以似然函数作为优化对象进行。假设之前选择的协方差函数（下面称为核）为 $K(x, y)$，那么可以得到似然函数的表达式为：

$$L(\theta) = \frac{1}{2}y^{T}\alpha + \frac{1}{2}\log|K| + \frac{n}{2}\log 2\pi$$

其中θ为参数，包括高斯过程方差、相关长度、噪声方差等，$\alpha = K^{-1}y$。因此，可以通过最大化似然函数的方式寻找最优的参数集合。

最大化似然函数得到最优参数后，便可以进行预测。对于要观测的数据点x_2，我们实际

上求的是

$$f(x_2|f(x_1) = y_1)$$

由已知数据点x_1和选取的核可以求出协方差矩阵

$$K = K(x_1, x_1)$$

由要估计的数据点x_2和已知数据点x_1可以得到

$$K^* = K(x_1, x_2)$$

通过联合高斯分布的边缘分布原理可知：

$$\bar{y_2} = K^{*T} K^{-1} y$$

$$\sigma_{y_2}{}^2 = K(x_2, x_2) - K^{*T} K^{-1} K^*$$

从而得到y_2的分布。

从上面的过程来看，高斯过程回归原理和传统回归方法本质上相差无几：都是经过模型选择、参数选择、预测的步骤实现。传统回归的模型选择对应高斯过程回归的核（协方差函数）的选择（在下一部分会进一步说明）；传统回归的参数选择对应高斯回归过程的参数选择：传统回归采用的是误差函数最小化，本质上也是一种似然函数最大化的过程，而高斯过程采用的正是似然函数最大化取得最优参数；传统回归的预测对应高斯过程的预测：传统回归利用确定性的模型取得预测点的数据，高斯过程回归利用随机模型取得预测点的数据分布。通过以上对比可以看出，作为回归过程，高斯过程回归和传统回归本质相同，唯一不同的是模型的确定性或随机性。这也是两者性能区别所在。

3. 核的选择

高斯过程回归中核的选择决定了回归模型的形态。因此，核的选择至关重要。我们了解的核模型有如下几种：

Squared Exponential 核（SE 核）：

$$K = \sigma^2 \exp(-\frac{r^2}{2l^2})$$

Exponential 核（EX 核）：

$$K = \sigma^2 \exp(-\frac{r}{l})$$

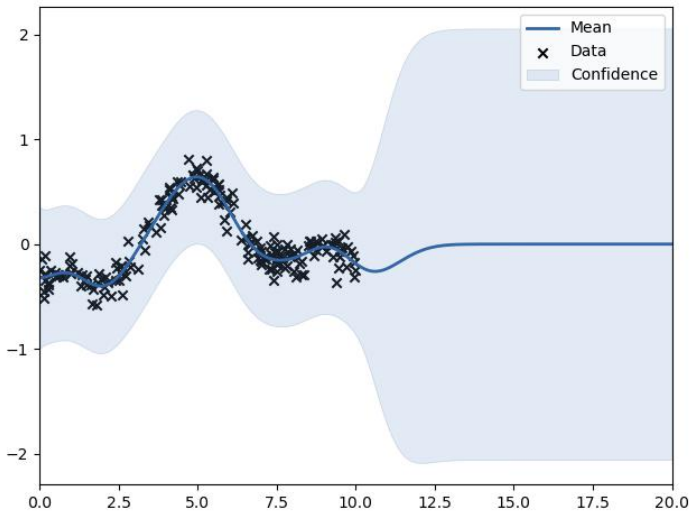
Matern32 核（MT 核）：

$$K = \sigma^2 (1 + \sqrt{3}r) \exp(-\sqrt{3}r)$$

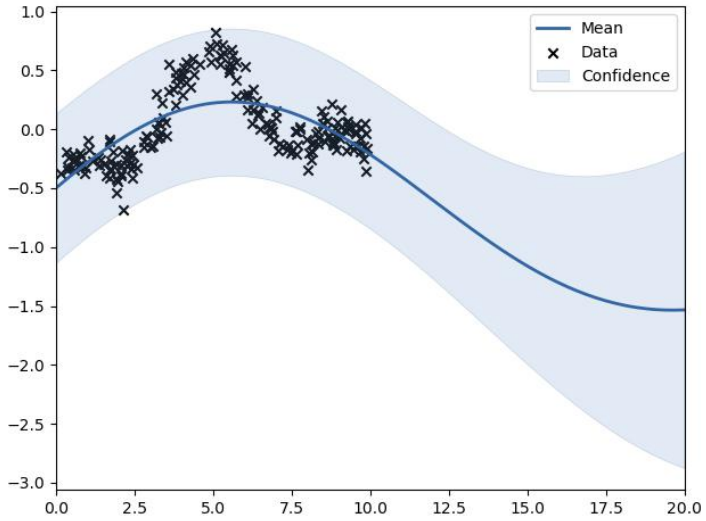
Rational Quadratic 核（RQ 核）：

$$K = \sigma^2 (1 + \frac{r^2}{2})^{-\alpha}$$

不同的核、不同的核参数意味着不同的随机函数模型。例如，对于上述未提到的周期核，其协方差函数为周期函数，那么其预测的数据一般是周期数据，因为每隔一个周期其相关性达到最大。而对于 SE 核，随着间隔增大，其相关性大大降低，因此一般不会用来拟合周期数据。对于不同的参数，拟合效果也不一致。当 SE 核中的 l 参数较小时，较小的间隔就会导致较低的相关，从而导致模型变化较快；当 l 较大时，较大的间隔仍保持较高的相关性，从而模型变化较缓慢。因此，由上可以看出，高斯过程回归中核的选择本质上对应的是传统回归中模型的选择，不同的核有不同的拟合效果；而参数的选择对应着具体模型的细节，例如斜率等。如下两图解释了 SE 核参数 l 的性质（0-10 为数据，10-20 为预测）：



l=1 时均值很快就归于 0



l=10 时很长间隔内仍保持相关

三、高斯过程回归与股票预测

高斯过程回归于股票预测主要包括股票数据选择、数据预处理、评价指标选择、模型建立与预测。

1. 股票数据

股票数据来源于网站：http://www.nasdaq.com/。我选择了 NVIDIA、TESLA、MSFT、AAPL 和 AMDI 五支股票，收集了它们两年内 500 个收盘价格。选取最近的 10 个收盘价格作为测试机，再前面的 400 个收盘价格作为训练集。由于并没有找到文献中 project 内的股票以及数据来源，因此无法和该 project 做比较。

为了方便说明，之后将五只股票数据另命名如下：

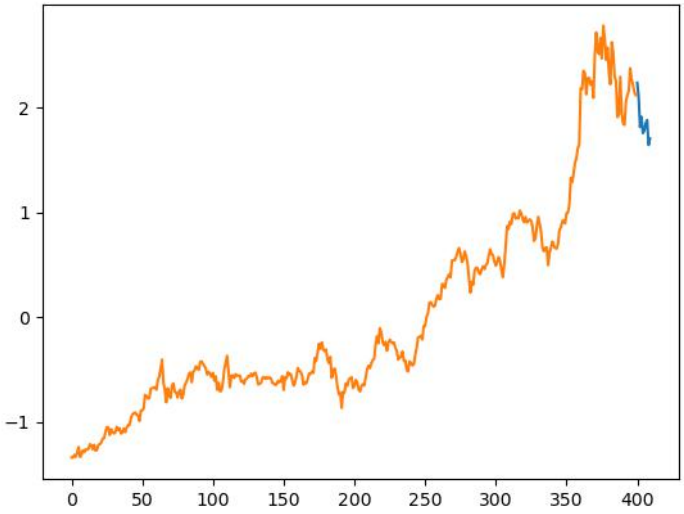
股票	命名
NVIDIA	Data1
TESLA	Data2
MSFT	Data3
AAPL	Data4
AMDI	Data5

2. 数据预处理

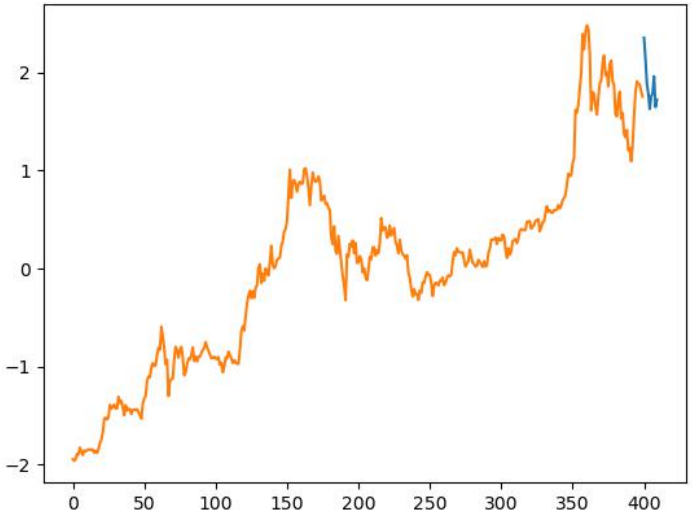
为了方便后续处理，我们对股票原始数据进行了一定的预处理。

首先我们为了满足模型中 0 均值的假设，将原始训练集数据进行了减去均值的处理，从而使得到的训练集数据均值满足假设。另外，我们还将数据进行了归一化处理：将均值归零后的数据除以其标准差，便于后续的比较和作图。

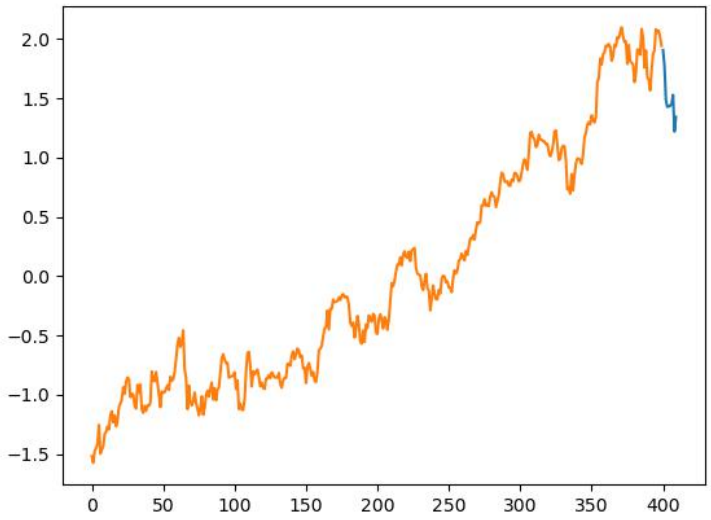
经过处理后得到的数据作图如下：



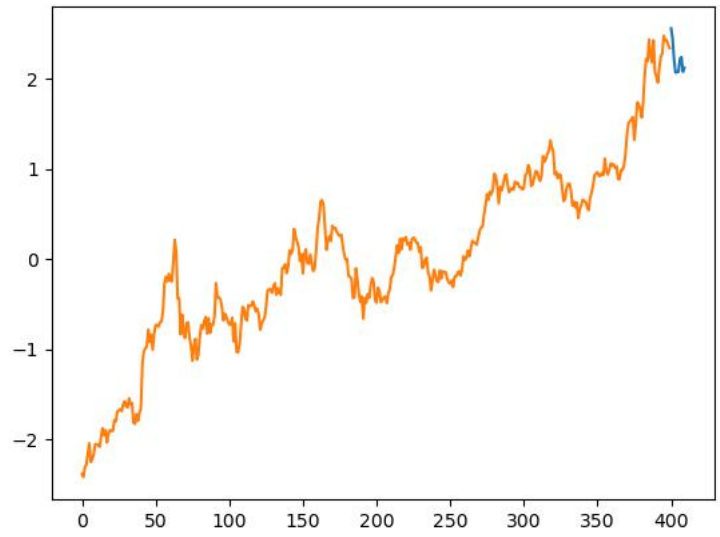
Data1



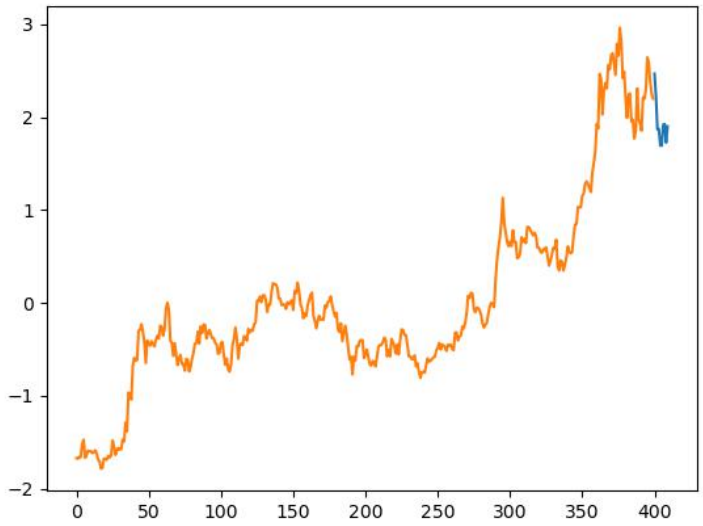
Data2


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
50
100
150
200
250
300
350
400

Data3


2
1
0
−1
−2
0
50
100
150
200
250
300
350
400

Data4



Data5

3. 评价指标选择

我们采用了 4 种评价指标综合对预测的结果进行评估，它们分别是：

均方误差 MSE：

$$\mathrm{MSE} = \frac{1}{n}\sum (y_{test} - y_{predict})^2$$

它可以很好的评估预测向量和评估向量的距离，但是受异常值影响较大。

平均绝对误差 MAE：

$$\mathrm{MAE} = \frac{1}{n}\sum |y_{test} - y_{predict}|$$

它可以减少异常值的影响，和 MSE 一起评估结果会更加真实。

平均绝对百分比误差 MAPE：

$$\mathrm{MAPE} = \frac{1}{n}\sum \frac{|y_{test} - y_{predict}|}{y_{test}}$$

它用于评估平均偏离的百分比。

异常值数量 NOO（Number of Outliers）：

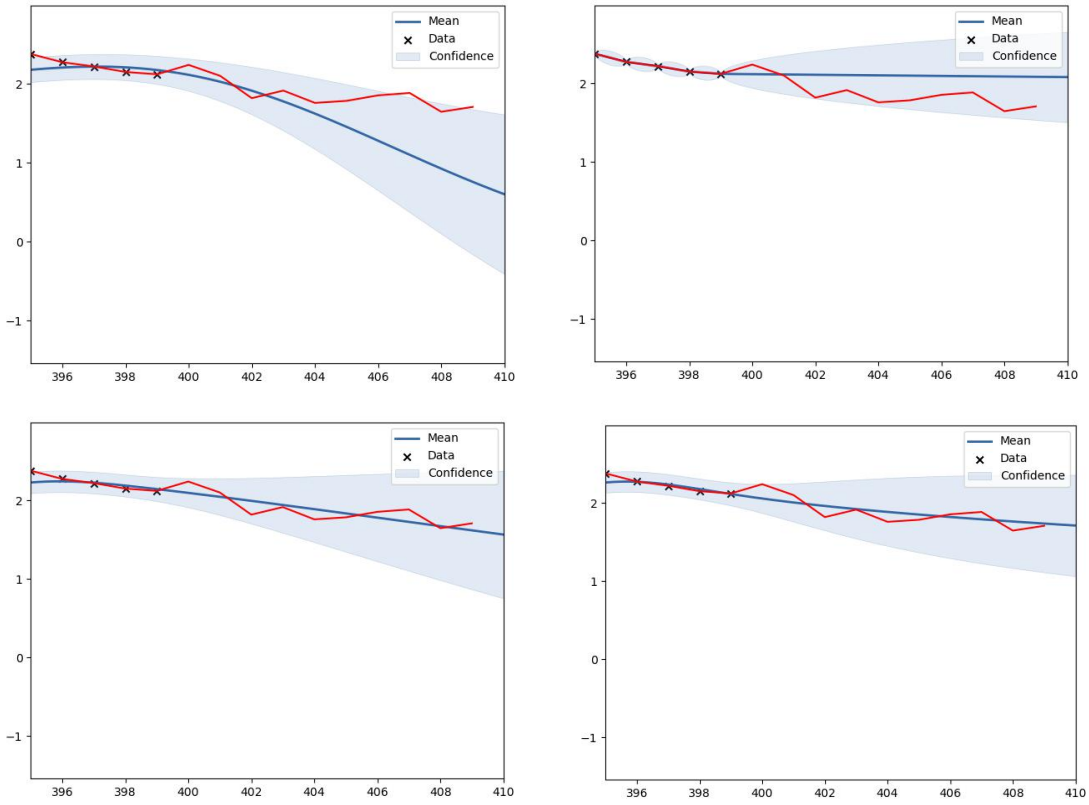
$$\mathrm{NOO} = \sum (|y_{test} - y_{predict}| > 2\sigma)$$

它统计偏离置信区间的数据点个数来判定预测异常值的数量。

4. 模型的建立与预测

在五支股票数据的回归建模和预测的过程中，我们均采用了 SE 核、EX 核、MT 核以及 RQ 核，并对 4 个核的结果进行了对比。

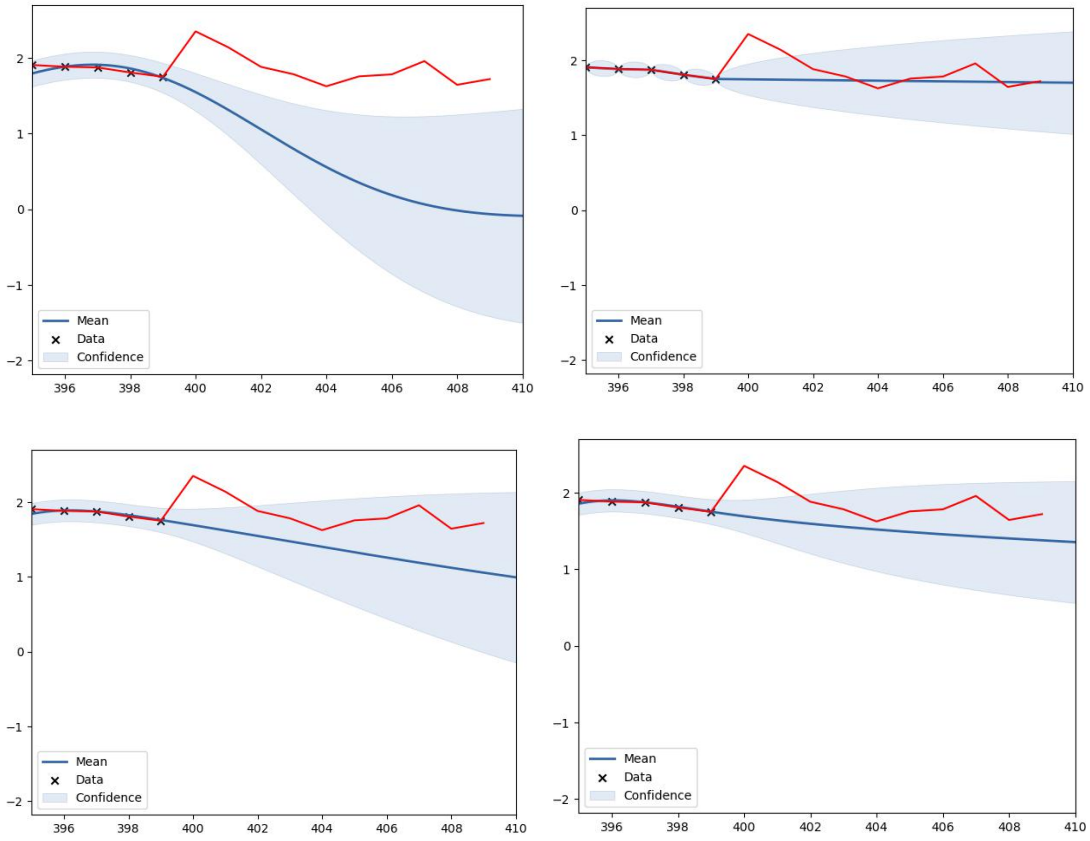
Data1：



由上到下、左到右依次是 SE 核、EX 核、MT 核以及 RQ 核的预测结果

核	MSE	MAE	MAPE	NOO
SE	0.2541	0.3924	21.98%	2
EX	0.0791	0.2542	14.22%	0
MT	0.0114	0.0931	4.96%	0
RQ	0.0108	0.0895	4.75%	0

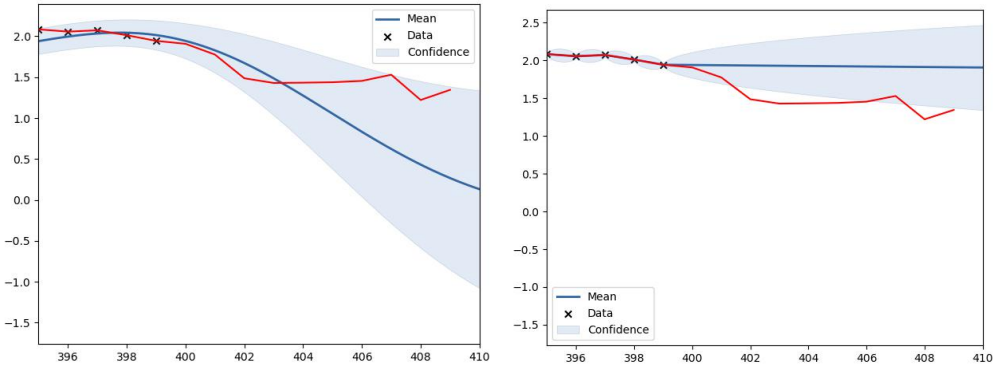
Data2:

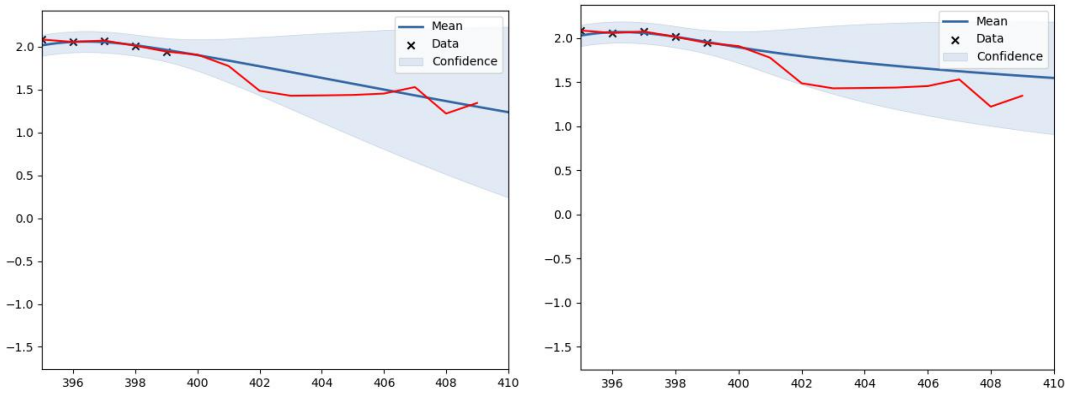


由上到下、左到右依次是 SE 核、EX 核、MT 核以及 RQ 核的预测结果

核	MSE	MAE	MAPE	NOO
SE	407561	638.4	34631%	10
EX	0.0629	0.1725	8.40%	2
MT	0.2724	0.4952	26.42%	2
RQ	0.1467	0.3491	18.12%	2

Data3:

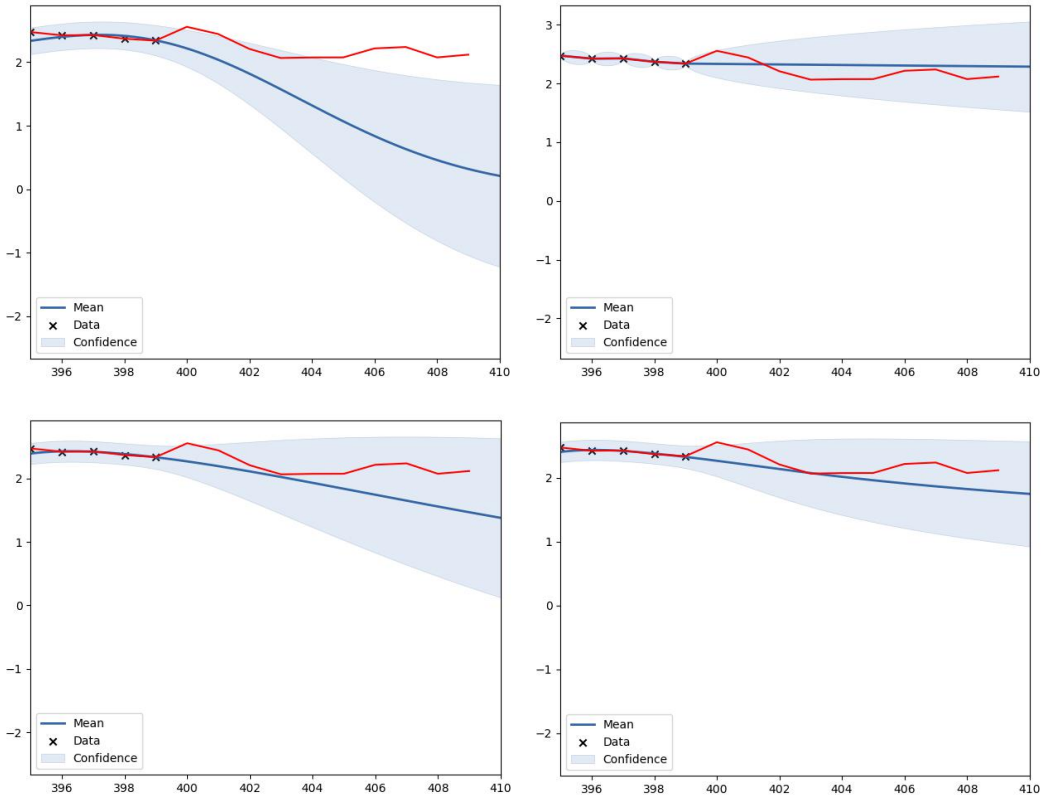




由上到下、左到右依次是 SE 核、EX 核、MT 核以及 RQ 核的预测结果

核	MSE	MAE	MAPE	NOO
SE	0.3209	0.4268	30.58%	0
EX	0.2124	0.4232	30.02%	7
MT	0.0255	0.1297	9.03%	0
RQ	0.0581	0.2124	15.13%	0

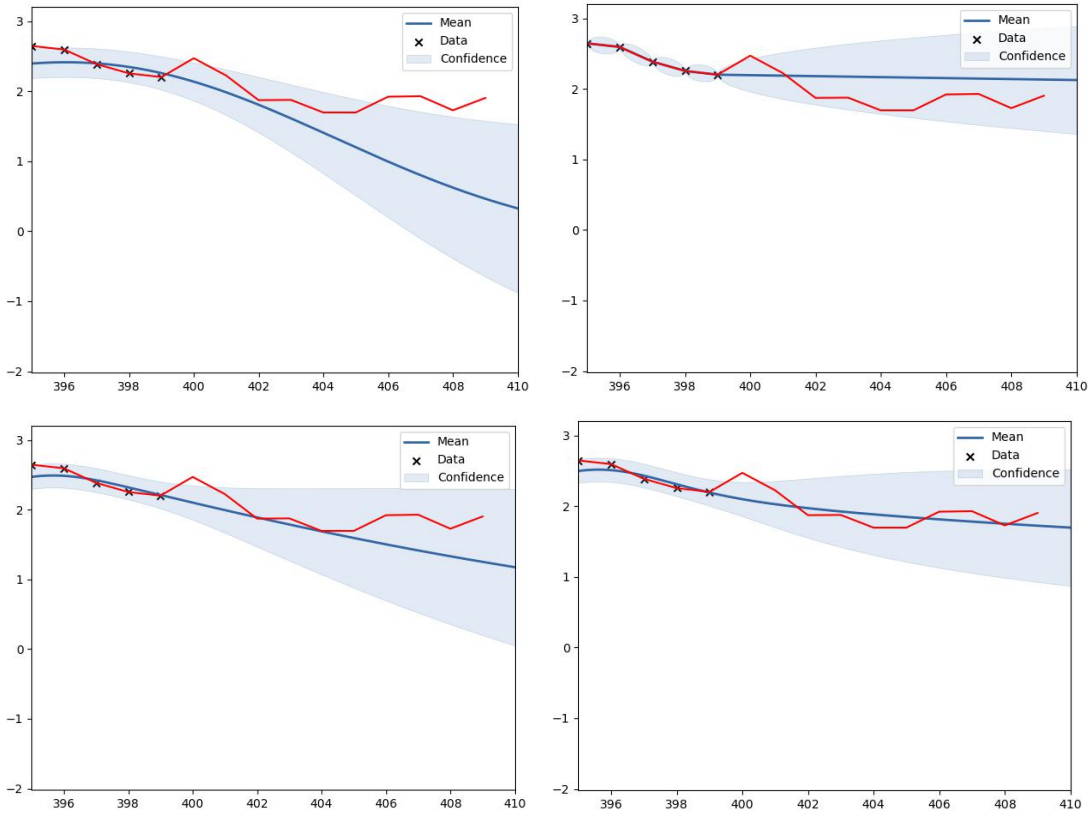
Data4:



由上到下、左到右依次是 SE 核、EX 核、MT 核以及 RQ 核的预测结果

核	MSE	MAE	MAPE	NOO
SE	1.2581	0.9784	45.28%	7
EX	0.0345	0.1730	7.96%	0
MT	0.1488	0.3278	14.91%	1
RQ	0.0571	0.2048	9.15%	1

Data5:



由上到下、左到右依次是 SE 核、EX 核、MT 核以及 RQ 核的预测结果

核	MSE	MAE	MAPE	NOO
SE	0.5993	0.6286	33.45%	5
EX	0.1020	0.2942	15.94%	1
MT	0.1226	0.2790	14.27%	1
RQ	0.0314	0.1515	7.59%	1

综合五支股票的数据预测结果以及评价指标可以看出：MT 核、RQ 核相比于 SE 核、EX 核构建的模型的预测效果更好。其中，SE 核在 data2-data5 的模型构建中都出现了较大偏离的现象，在 data2 中甚至出现了病态的情况。这很大原因是参数相关间隔 l 过小，从图中模型均值很快回到 0 可以看出。EX 核在大部分表现良好，但是在 data3 的模型构建中偏离较大。MT 核、RQ 核构建的模型在所有数据中都表现优异，因此对于股票预测，MT 核、以及 RQ 核是较好的选择。另外，四个核都存在的问题是：如果恰好在第一个预测点上有较大的变化，模型是难以追踪的，这体现在 data2、data4、data5 的预测结果上。但这类问题我们目前仍没有想到缓解或者是解决的方法。

四、思考与探究

对于上述出现的问题，我们进行了一些思考与探究。

在某些核预测不准，导致在未知未来数据、没有评估指标的情况下，选核成了一个问题。对于这种情况，我们采用了自动选核的算法，使得预测结果的稳定性和性能都得到提高。

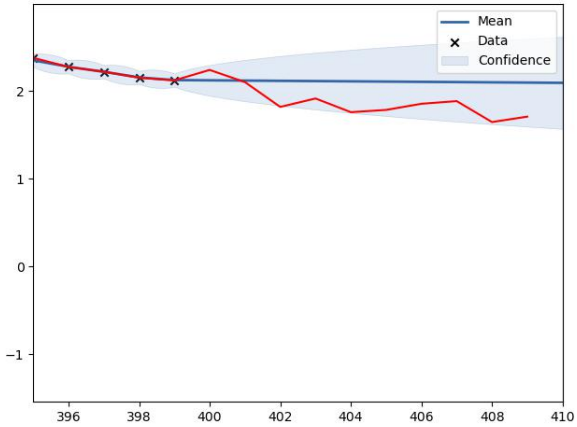
对于预测不准（例如 SE 出现的巨大偏离）的问题，我们认为可以设置一定模型参数优化限制，从而减少过拟合的出现。

1. 自动选核

自动选核采用的是贪心算法。首先创建一个核集合，对每个核都进行模型构建并计算似然值，选出似然值最大的核（记为核 0）。之后再从集合中选出核，依次与核 0 进行加或者乘的组合，在这些组合中选出似然值最大的核（记为核 1）。如此反复循环，直到得到核 n 停止。

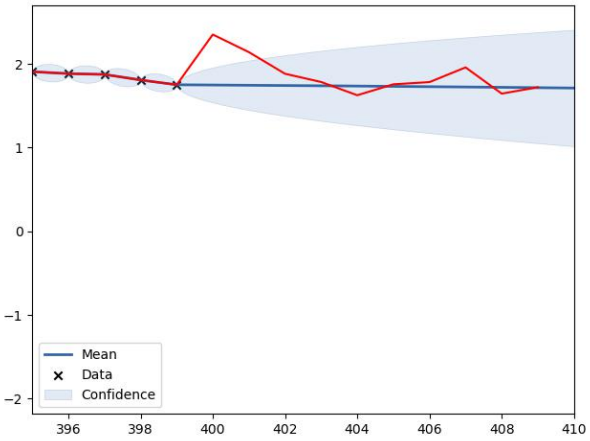
由于自动选核的过程中每一次核的组合都需要进行模型的构建，因此计算复杂度很大。由于算力有限，我们仅执行了一次组合，一窥自动选核的效果。结果如下。

Data1：



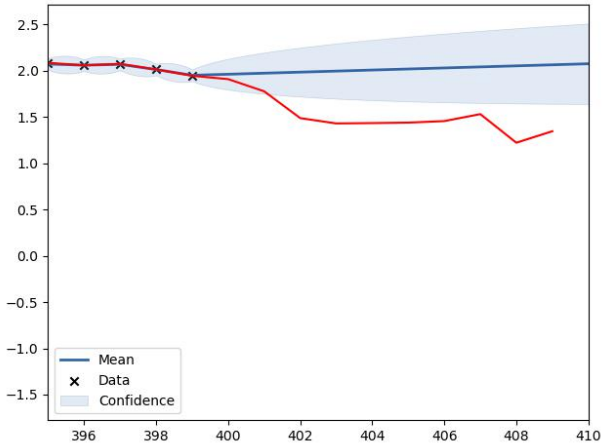
核	MSE	MAE	MAPE	NOO
RQ+EX	0.0835	0.2615	14.63%	1

Data2：



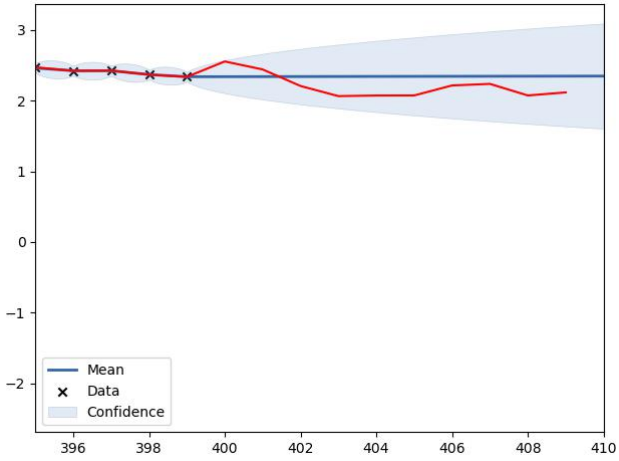
核	MSE	MAE	MAPE	NOO
RQ+EX	0.0619	0.1688	8.21%	2

Data3:



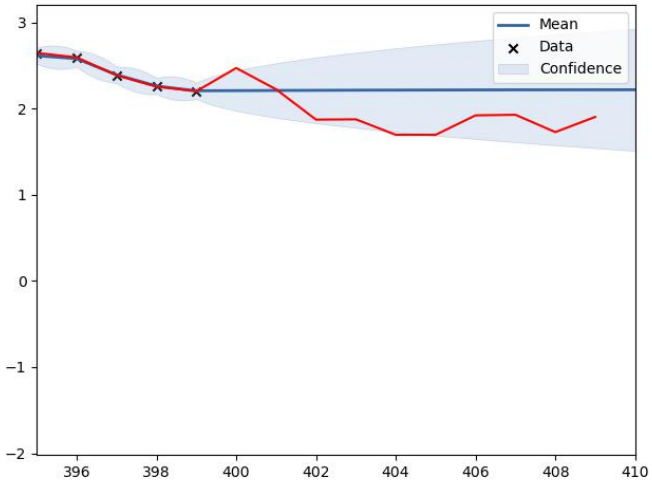
核	MSE	MAE	MAPE	NOO
RQ+EX	0.3068	0.5098	36.14%	8

Data4:



核	MSE	MAE	MAPE	NOO
RQ+EX	0.0454	0.2011	9.29%	0

Data5:



核	MSE	MAE	MAPE	NOO
SE+EX	0.1350	0.3383	18.41%	2

通过上述结果可以看出，自动选核在绝大多数情况下是优于最差结果的，因此如果不断选核，结果大概率是越来越好的。但实际上也可以看出，自动选核的结果除了在 data4 中比单核最优结果好，其余都是不如最优结果的。这很可能是由于自动选核算法在选核时仅考虑似然值大小，而似然值大小仅决定已知数据的拟合效果，并没有关系到未来数据的预测效果，因此很容易造成过拟合。由上述结果可以知道，EX 核的拟合效果总是最好的，但预测结果总不是最好的，因此存在一定的过拟合。

2. 设置优化限制

设置优化限制是指模型在根据似然函数值进行最优化参数时，设置一定的参数限制，以防止过拟合造成的预测偏离。这不仅可以缓解 SE 模型出现的偏离现象，也可以提升自动选核的结果。

设置优化限制有两种方法。一是根据经验设定限制，这需要大量的先验知识。二是利用功率谱分析，将大部分功率限制在奈奎斯特采样频率内，从而得到参数 l 的限制。（也可以利用正则项限制似然函数的优化，比如将参数优化的目标函数改为 $J(\theta) = L(\theta) + l$）

五、总结

高斯过程回归主要可以总结为以下几个步骤：

1. 数据的预处理
2. 核的选择和模型的建立
3. 参数优化
4. 预测

其中，数据的预处理主要是将数据均值变为 0 以使其满足 0 均值平稳高斯过程；核的选择至关重要，它涉及到拟合模型的精准和预测结果的准确；参数优化利用拟合评价函数——似然函数的最大化进行；预测利用了高斯过程任意点满足联合高斯分布的性质以及联合高斯边缘分布特性。

在普通高斯过程回归中，我们遇到两类问题：

1. 核的选择。由于不同的核选择会导致预测结果不同，我们通过五组数据测试了四种核，并经过对比得出 MT 核和 RQ 核的拟合、预测效果最优。另外，我们也采用了自动选核的方法进行了实验，以提高结果的稳定性。
2. 过拟合问题。由于参数优化的目标函数仅仅是似然函数，并没有考虑到和预测有关的参数限制，因此常常会导致过拟合的出现。我们通过设置经验限制值、限定功率范围、修改目标函数等方法将过拟合的现象减少。

虽然可能不同的核会导致预测结果相去甚远，但是股票价格走势还是能在一定误差范围预测的，这也体现出了高斯过程回归在股票预测中的巨大潜力。

六、致谢

首先感谢李刚老师本学期的教学工作，您循循善诱、形象生动的讲解让我更深刻地理解了随机过程的知识，为本次大作业打下基础，并让我产生了浓厚的兴趣，乐意投入时间研究。

另外感谢 gpy 包的编写团队，是他们编写的包让我省去了造轮子的麻烦，可以花更多时间在对高斯过程回归的理解和探索上。

七、附件与说明

附件中共有 5 份 python 代码，GP1-GP5 分别对应 data1-data5 的代码。另外有从网站上下载的股票数据 data1-data5。

本次高斯过程回归和股票预测主要是基于 gpy 包实现的，以省去很多造轮子的麻烦。因此代码运行需要下载 gpy 包。