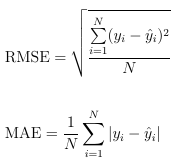
**一．课题要求：预测建模问题**

给定5个时间序列(*y*(*t*), *x*1(*t*), *x*2(*t*), *x*3(*t*), *x*4(*t*))，要求构建预测模型以给出*y*(*t*+1)的值。即构建*y*(*t*+1) = *f*(*y*(*t*), ..., *y*(*t*-*a*), *x*1(*t*), ..., *x*1(*t*-*b*), *x*2(*t*), ..., *x*2(*t*-*c*), *x*3(*t*), ..., *x*3(*t*-*d*), *x*4(*t*), ..., *x*4(*t*-*e*))的函数关系，其中*a*, *b*, *c*, *d*, *e*表示阶次，阶次需要根据数据进行推断。且*x*1(*t*), *x*2(*t*), *x*3(*t*), *x*4(*t*)中有一个与*y*(*t*)无关，请找出。

训练数据（5000）：train\_X = [*x*1(*t*), *x*2(*t*), *x*3(*t*), *x*4(*t*)], train\_y = [*y*(*t*)]

测试数据（500）：test\_X = [*x*1(*t*), *x*2(*t*), *x*3(*t*), *x*4(*t*)], test\_y = [*y*(*t*)]

**性能评估：**RMSE, MAE



**作业要求：**根据训练数据确定阶次，构建预测模型。

**二．具体实现过程**

**1.判断无关的时间序列**

假定5个时间序列(*y*(*t*), *x*1(*t*), *x*2(*t*), *x*3(*t*), *x*4(*t*))的滞后阶数都为5阶，用多元线性回归方法来对模型进行拟合，以及参数的估计。

多元线性回归模型的一般形式为：

，

其中k为解释变量的数目， 称为回归系数。

多元线性回归模型的参数估计，在要求误差平方和为最小的前提下，用最小二乘法或最大似然估计法求解参数。

Matlab实现方法如下：

1. b = regress(Y,X) 确定回归系数的点估计值，其中Y为输出矩阵，X为第一列全为1的输入矩阵
2. [b,bint,r,rint,stats]=regress(Y,X,alpha) 求回归系数的点估计和区间估计、并检验回归模型
3. bint表示回归系数的区间估计
4. r表示残差
5. rint表示置信区间

我们假定模型为：



得出系数矩阵为[-0.2188 0.9212 -0.4335 0.2321 -0.1004 0.0232 0.4485 -0.2881 0.2259 -0.1614 0.0536 -1.0612 0.6392 0.1491 0.0913 0.1218 0.0589 -0.0895 -0.0082 0.0612 -0.0361 0.006 -0.0095 0.0329 -0.0629 0.0239]

RMSE = 1.5487 MAE = 1.08

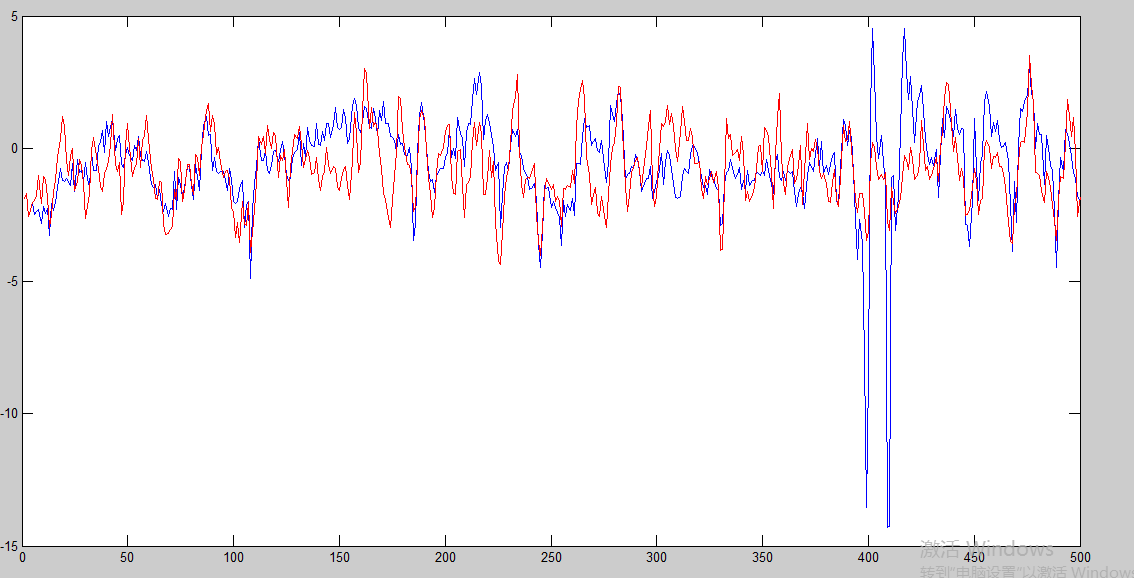
可以看出 的系数普遍比前面三个变量少一个数量级，与输出的相关性很微弱，可以认为是无关变量。接着将输入 去除，得出去除之后的RMSE=1.5481，MAE = 1.06，对输出的影响可以忽略不计，证实 为无关变量。

**2.判断模型滞后阶数**

根据系数矩阵，将系数偏小的项去除，新的模型为：



测试拟合结果如下图：

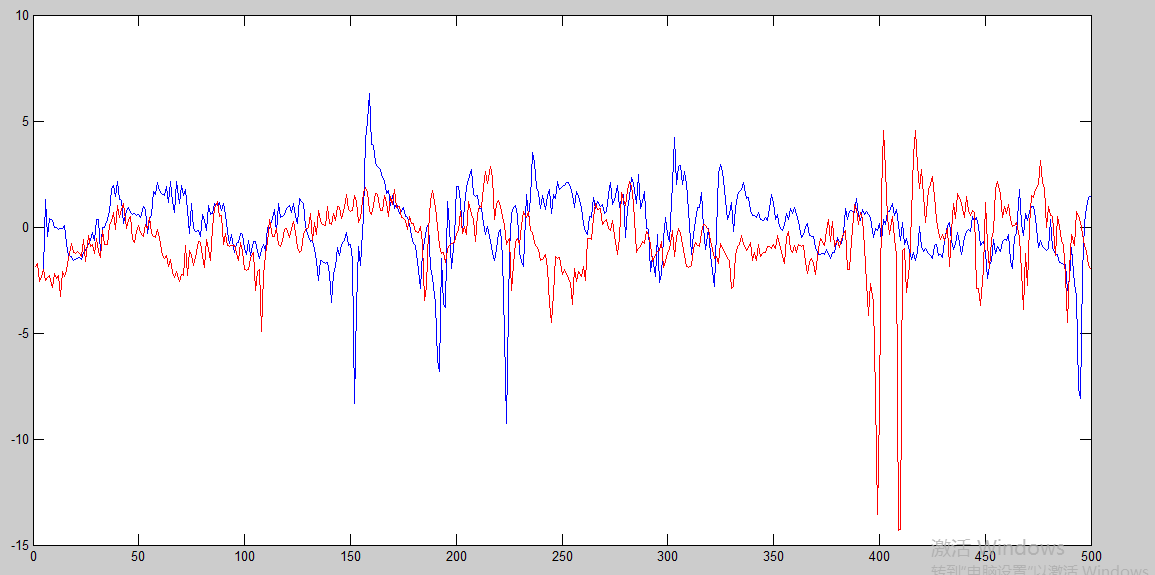


多元线性回归的思想是基于概率论和数理统计的基础上的，希望通过样本所反映的总体信息来揭示总体的规律性，但是对于偶然发生的干扰并不能很好的进行跟踪，这就造成了平整的部分，多元线性回归拟合的很好，但是对于尖峰信号，多元线性回归并不能识别。同时造成RMSE与MAE性能指标并不好。所以仅用它来分析一下无关向量，改用神经网络来进行学习并预测时间序列。

**3.神经网络建模预测**

1、我们首先使用了BP算法对时间序列模型进行训练，实验结果如下：

RMSE = 2.53 MAE=1.85



拟合效果很差，分析原因在于挑选合适的初始权值，学习率，隐节点数目非常困难。初始权值的选择会严重影响神经网络的泛化能力，极易陷入局部最小。同时学习率的选择也同样会对神经网络造成很大影响，太小会导致收敛太慢，大了又会导致不收敛。

2、随后我们考虑使用RBF-OLS算法来对该模型进行训练，RBF网络可以任意精度逼近任意的非线性函数，且具有全局逼近能力，从根本上解决了BP网络的局部最优问题，而且拓扑结果紧凑，结构参数可以实现分离学习，收敛速度快。

**三．总结**

针对本课题，我们采用了多元线性回归和神经网络两种方法进行解决，其中，在神经网络部分我们尝试了采用BP网和RBF网。通过多元线性回归方法，我们较快地得到了实验结果，并根据实验结果确定了系统阶次和无关变量。但是通过得到的拟合曲线可以看出，实际拟合效果并不是太好，于是我们又采用了神经网络方法。然而训练时间却大大增长，即使将训练过程放在一台服务器上运行，实际的训练时间也超过了十个小时，而且最后得到的性能指标RMSE和MAE反而变差了。造成这种情况的原因可能是初始权值、学习率等参数设置的不合适。

本课题比较接近实际工程应用。通过完成本课题，我们进一步提高了利用神经网络解决实际问题的能力。但同时也反映出我们对神经网络的一些理论理解得不是很全面，对于一些问题，虽然有解决的思路，但是却不知道该如何进行实际操作。这些理论和实践方面所欠缺的能力，有待于通过日后的学习进一步提高。