相关性分析

一、分析思路

对数据进行相关性分析，经常会计算相关系数 r 。在本项目中，我们想要使用多入多出的方法对时间序列进行趋势预测。使用多入多出的前提是多个输入变量和一个或者多个输出变量相关。如果不相关，则没有使用多入多出的必要。

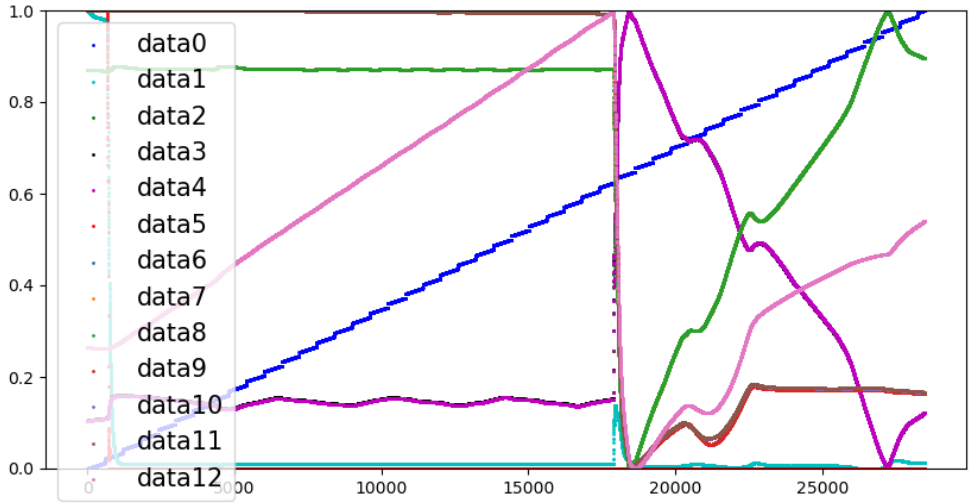
经过观察，需要进行预测的数据呈现近似线性关系，所以可以对数据进行低通滤波，计算输入输出之间的线性相关系数。

可以先绘制散点图和折线图，初步判断各变量之间是否相关，再计算相关系数。

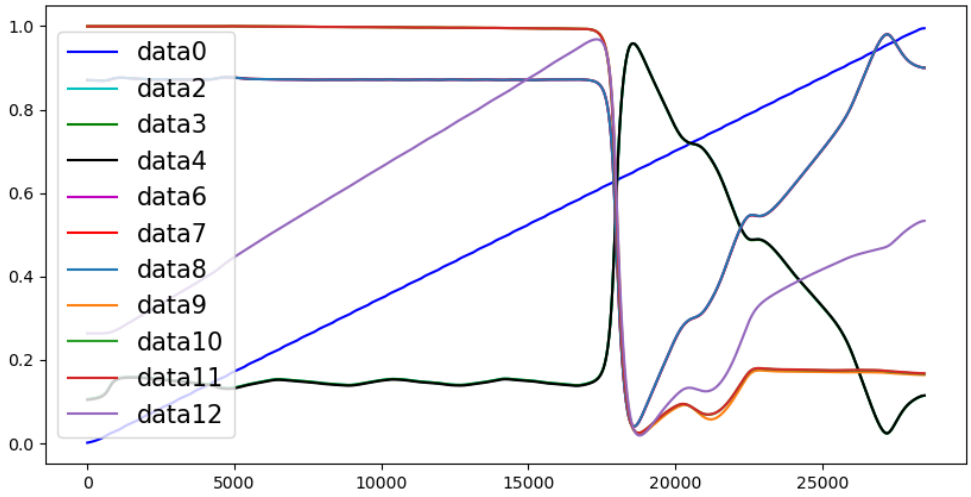
二、数据归一化以及图像绘制

由于各数据的单位不同，数量级不同，直接绘制图形无法判断是否相关，可以先进行归一化，再绘制图像。

归一化以后直接绘制的散点图如下：



data 12 是需要预测的数据，筛去与 data 12 明显无关的变量（data 1 5），对数据进行低通滤波，得到如下折线图：



三、线性相关系数计算

通过计算皮尔逊相关系数判断相关程度。对于变量X和Y，通过以下公式计算：

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

由于协方差总是小于各个标准偏差的乘积，因此*ρ*的值在*-1*和*+1*之间变化。

协方差用于衡量两个变量的总体误差。如果两个变量的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。 如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。所以可以通过协方差来衡量变量之间的相关性。并且通过除以标准差对相关性进行归一化。

以下是相关系数计算结果（从data 0到12依次和data 12 的相关系数）：

> 0 2 3
>
> 11 12

$$\begin{bmatrix} 1 & -0.07765352 \\ -0.07765352 & 1 \end{bmatrix} \begin{bmatrix} 1 & -0.66240778 \\ -0.66240778 & 1 \end{bmatrix} \begin{bmatrix} 1 & -0.66196483 \\ -0.66196483 & 1 \end{bmatrix}$$

> 4 6 7

$$\begin{bmatrix} 1 & -0.66261607 \\ -0.66261607 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.66798844 \\ 0.66798844 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.66798844 \\ 0.66798844 & 1 \end{bmatrix}$$

> 8 9 10

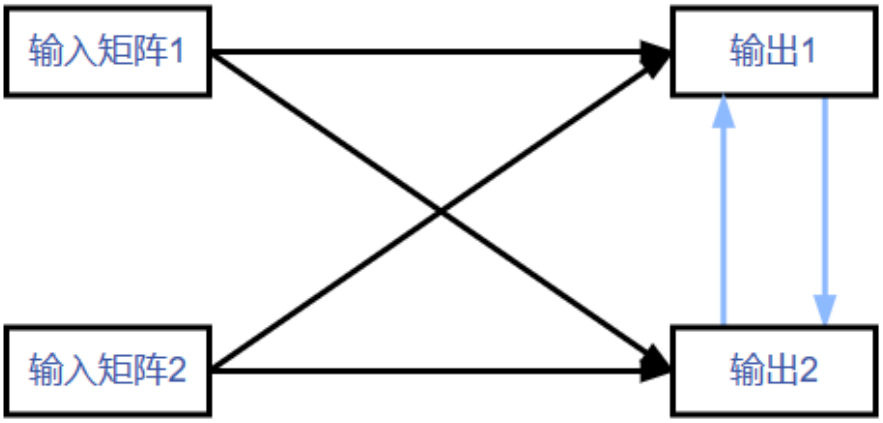
$$\begin{bmatrix} 1 & 0.66771781 \\ 0.66771781 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.622095 \\ 0.622095 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0.62137908 \\ 0.62137908 & 1 \end{bmatrix}$$

> 11 12

$$\begin{bmatrix} 1 & 0.62153983 \\ 0.62153983 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

四、进一步工作

根据相关性计算的结果，大部分输入和输出中度相关(≈0.6)，可以使用多入单出策略。但是由于没有相关资料，只知道单个输出量，无法分析输出量之间的相关性。需要通过查阅文档，得到输出量有哪些，以及它们对应的输入量，进一步分析输出之间的相关性，以及不同输出所对应的输入之间的交叉关系。



目前只分析了输入矩阵1到输出1之间的相关性，也就是对应的输入输出之间的相关性。还需要分析不同输出之间的相关性，以及交叉的输入输出的相关性。