## CoFlux波动相关性

**背景**

在互联网公司里，通常会监控成千上万条时间序列，用于保障整个系统或平台的稳定性，在这种情况下，如果能够对多条时间序列判断其是否相关，对于监控是十分有效的。

**CoFlux整体介绍**

输入：两条时间序列

输出：

1. 波动相关性：两条时间序列是否存在波动相关性
2. 前后顺序：如果两条时间序列相关，输出其先后波动顺序，同时进行或者前后顺序
3. 方向性：如果两条时间序列相关，输出其波动方向。

CoFlux没有对时间序列进行异常检测，直接从时间序列的历史数据出发，判断两条时间序列之间的波动相关性，进一步分析先后顺序和波动方向。

**CoFlux计算流程**

**提炼时间序列的波动曲线特征**

已知一个长度为的时间序列，对于任意一个特征提取工具（detector），可以得到一条关于的预测值曲线。于是针对这一特征提取工具（detector）可以得到一个波动特征序列，其中，其中。因此一个特征提取工具可以对应一个波动序列特征，即一个时间序列。对于个特征图提取工具，可以对应条波动特征序列，并且其长度都是。

原文次此处共使用了86种特征提取工具。

文章认为：

1、对于任何一条时间序列，总有一个特征提取工具可以相对准确的提炼其波动特征

2、如果两条时间序列和波动相关，那么的一个波动特征序列与的一个波动特征序列也应该是相关的

在提炼时间序列的波动曲线特征之后，还需要进行以下几个步骤：

特征工程的放大（Amplify）：对波动特征序列进行放大，让某些波动序列特征更加明显；

修正：当两条时间序列之间存在时间前后的漂移，两条时间序列之间存在空隙，需要对其中一条时间序列进行平移操作；

**波动特征曲线处理**

1. 对波动特征曲线做的归一化
2. 对归一化后的波动特征曲线进行特征放大处理，定义函数

则放大之后的波动特征曲线为。

1. 对于两条放大后的特征曲线和，计算其相关性、先后顺序和方向。令

其中为个0，。特别地，当时，，那么可以定义和的内积

这里的内积指的是向量之间的内积，同时定义相关性为

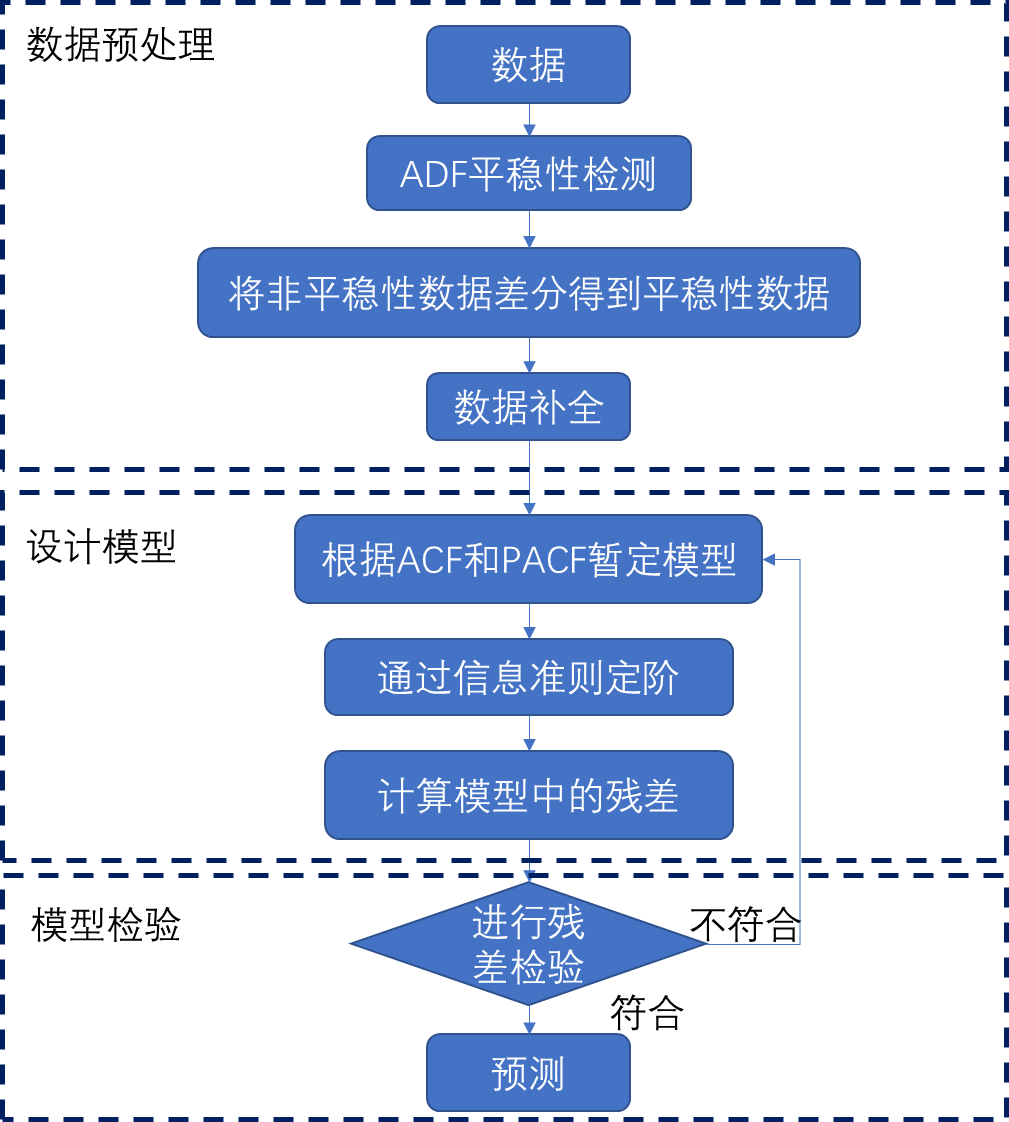
由于波动可能是反向的，因此要考虑相关性时大于零的情况，也需要考虑小于零的情况，因此定义

其中最小值和最大值的指标分别是

令

从定义中可以看出，是一个元组，包含了三种信息，即相关性，波动方向和前后顺序。，越接近于1或-1就表示放大后的波动特征曲线和越相关。正值的表示和的波动方向相同，是正相关，负值的表示和的波动方向相反，是负相关。通过对和即可判断先后顺序。

## 特征提取工具——ARIMA



**第一步：数据处理**

验证数据的平稳性：

方法1直接观察图像，无明显上升或下降趋势

方法2ADF平稳性检测，t统计量的值要小于在99%，95%，90%置信区间下的临界的ADF检验的值，且t统计量对应的概率值尽可能接近0

若不平稳，采用差分法或用平滑法做平稳性处理，但需要注意在处理后可能会出现“NaN”，需要对数据进行补全，由于差分法是为了使数据在0值上下浮动，因此需要进行数据0的补全，但是也可以直接去除序列中的“NaN”。其中平滑法适用于平稳上升的周期性数据。

**第二步：定阶**

ARIMA模型有三个系数。

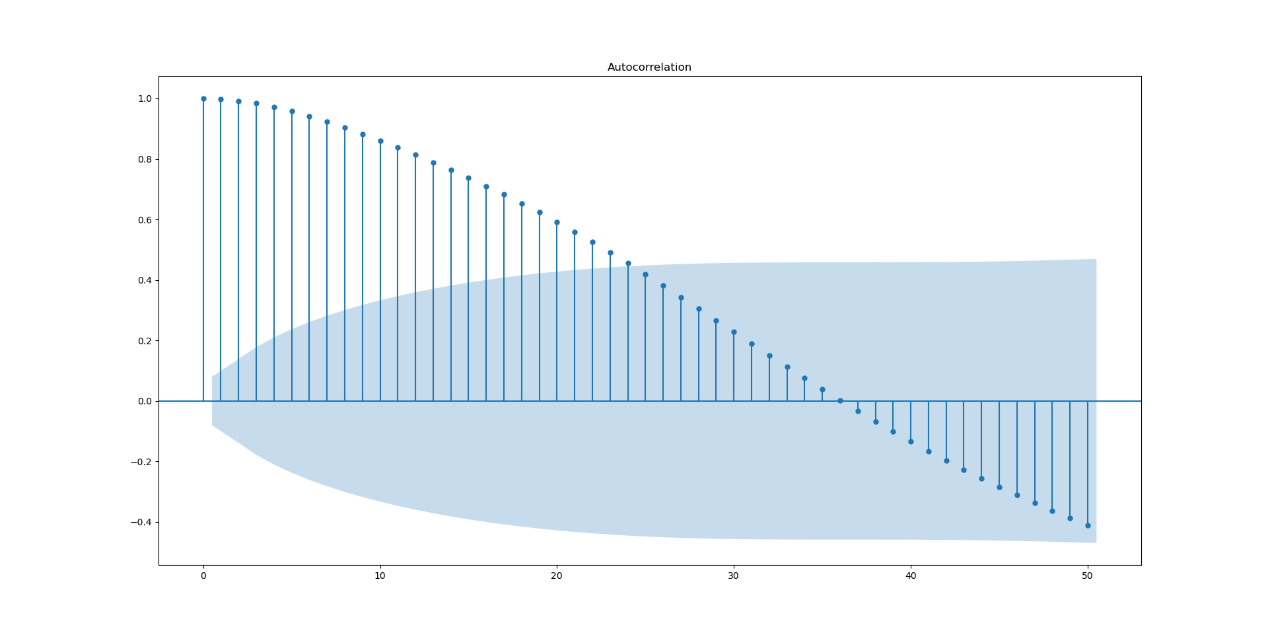
p --代表预测模型中采用的时序数据本身的滞后数(lags) ,也叫做AR/Auto-Regressive项

d --代表时序数据需要进行几阶差分化，才是稳定的，也叫Integrated项。

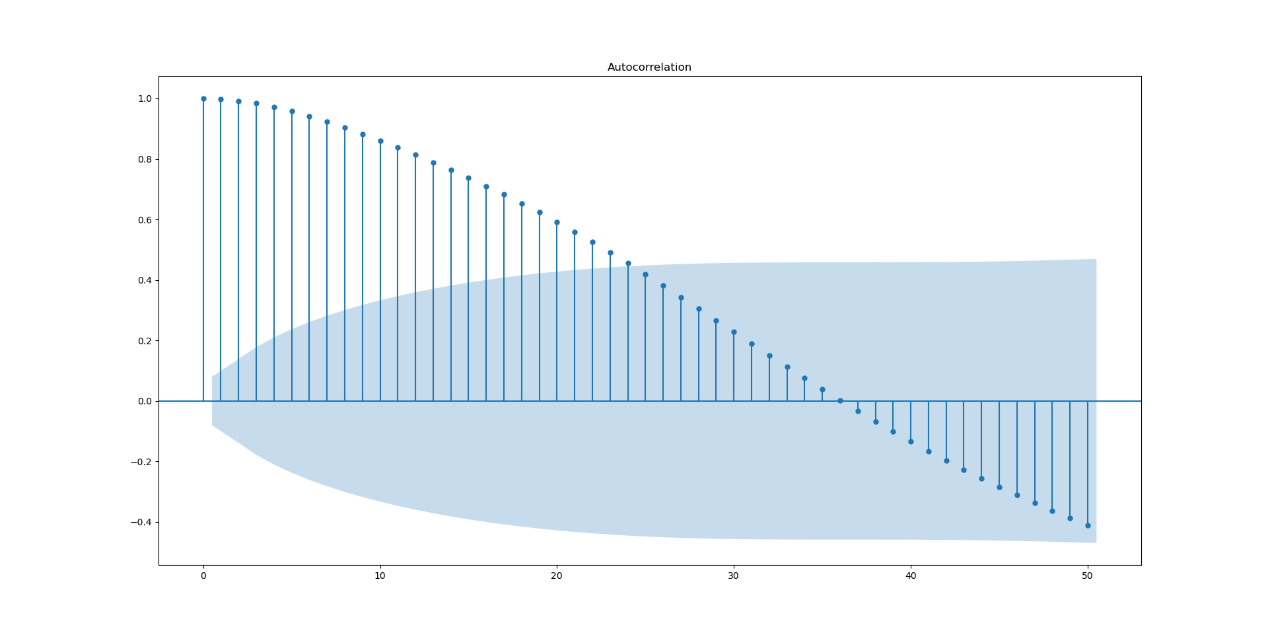
q --代表预测模型中采用的预测误差的滞后数(lags)，也叫做MA/Moving Average项

需要判断p、q的选取

方法1：通过观察自相关图（有序的随机变量序列与其自身相比较，反应同一序列在不同时序的取值之间的相关性）和偏相关图（计算两个变量之间的相关性，是剔除了中间变量的干扰之后，得到两个变量之间的相关程度）确认。对于ARIMA模型，当自相关q或偏相关p阶后衰减趋于零即为所取值。

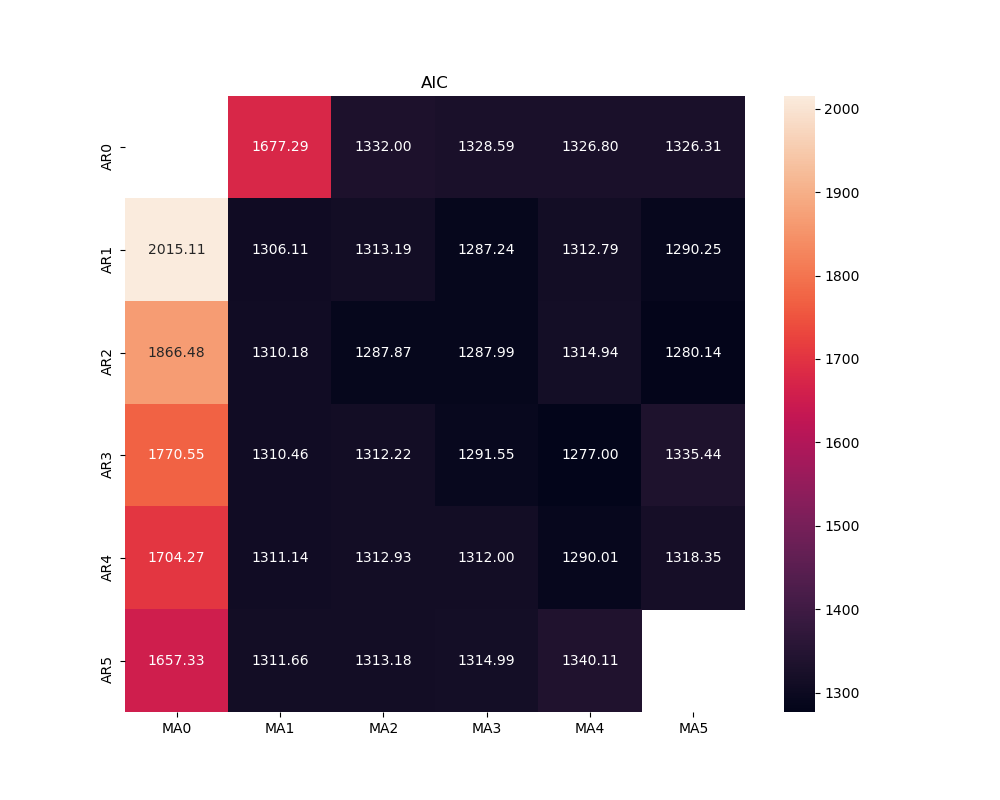


Gz自相关图



Gz偏相关图

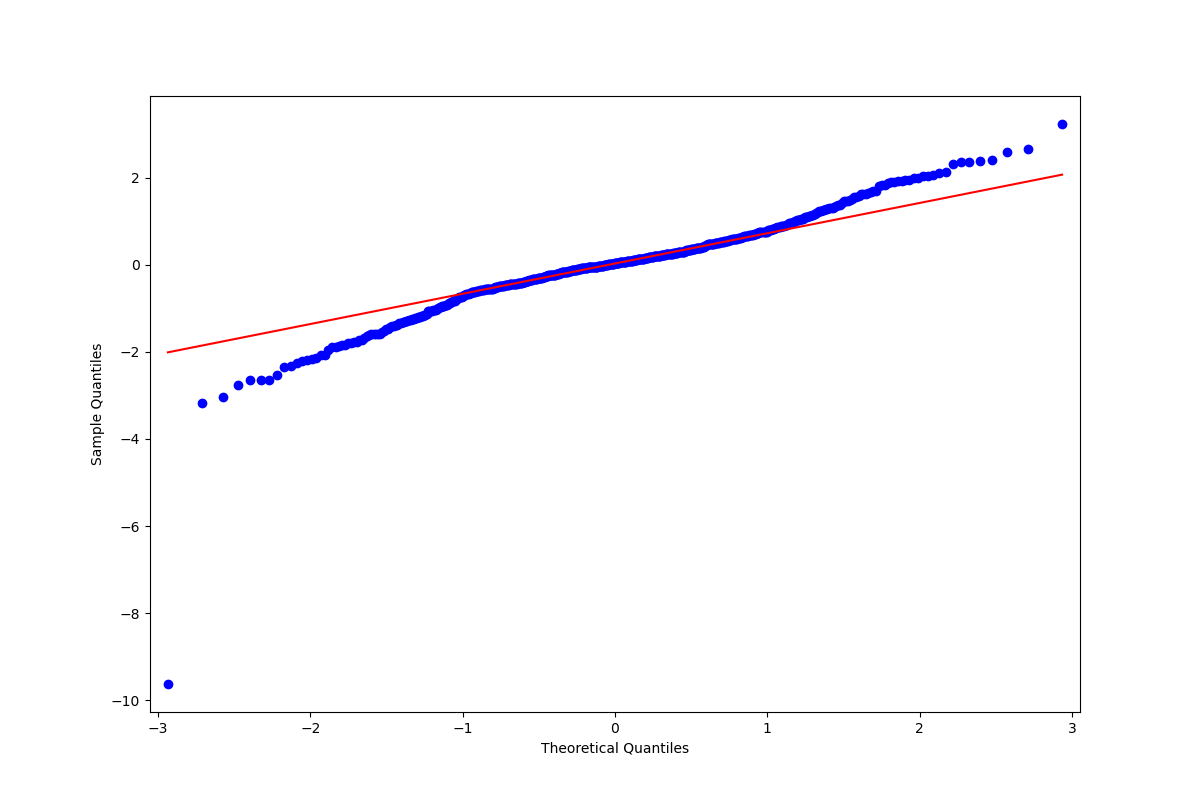
方法2：通过信息准则AIC，BIC和HQIC进行定阶，取值小的作为模型阶数



Pm根据AIC和BIC定阶热力图

**第三步：构建模型并评估**

评估有三种：一评估残差是否满足正态分布（QQ图）二评估残差的自相关性（D-W检验）、三是残差的白噪声检验（Ljungbox检验）



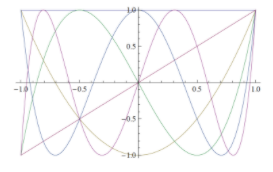
Gz预测结果QQ图，大部分位于直线，满足正态分布

## 切比雪夫多项式

使用正交多项式做最佳平方逼近，但使用一般的广义多项式做基底，求解最佳平方逼近多项式，当较大时，系数矩阵是“高度病态的”，求方程组的解，舍入误差很大，但对函数族，使得任意一对和两两带权正交，则系数矩阵中可化成对角阵，不用解线性方程组。

定义多项式序列，，称为切比雪夫多项式。

该多项式满足如下递推关系



其中在区间上带权函数正交，即

## 希尔伯特—黄

**经验模态分解（EMD分解）**

1. 找出信号中所有局部最大值并用三次样条函数连接成上包络，同理利用三次样条插值函数连接所有局部极小值构成的下包络；
2. 求出上下包络的平均值记为，并求原始信号与包络均值之间的差值
3. 如果满足IMF固有模态函数的两个条件
4. 的平均值趋近于0
5. 原始信号的极值点的个数与原始信号同的交点个数值差小于等于1

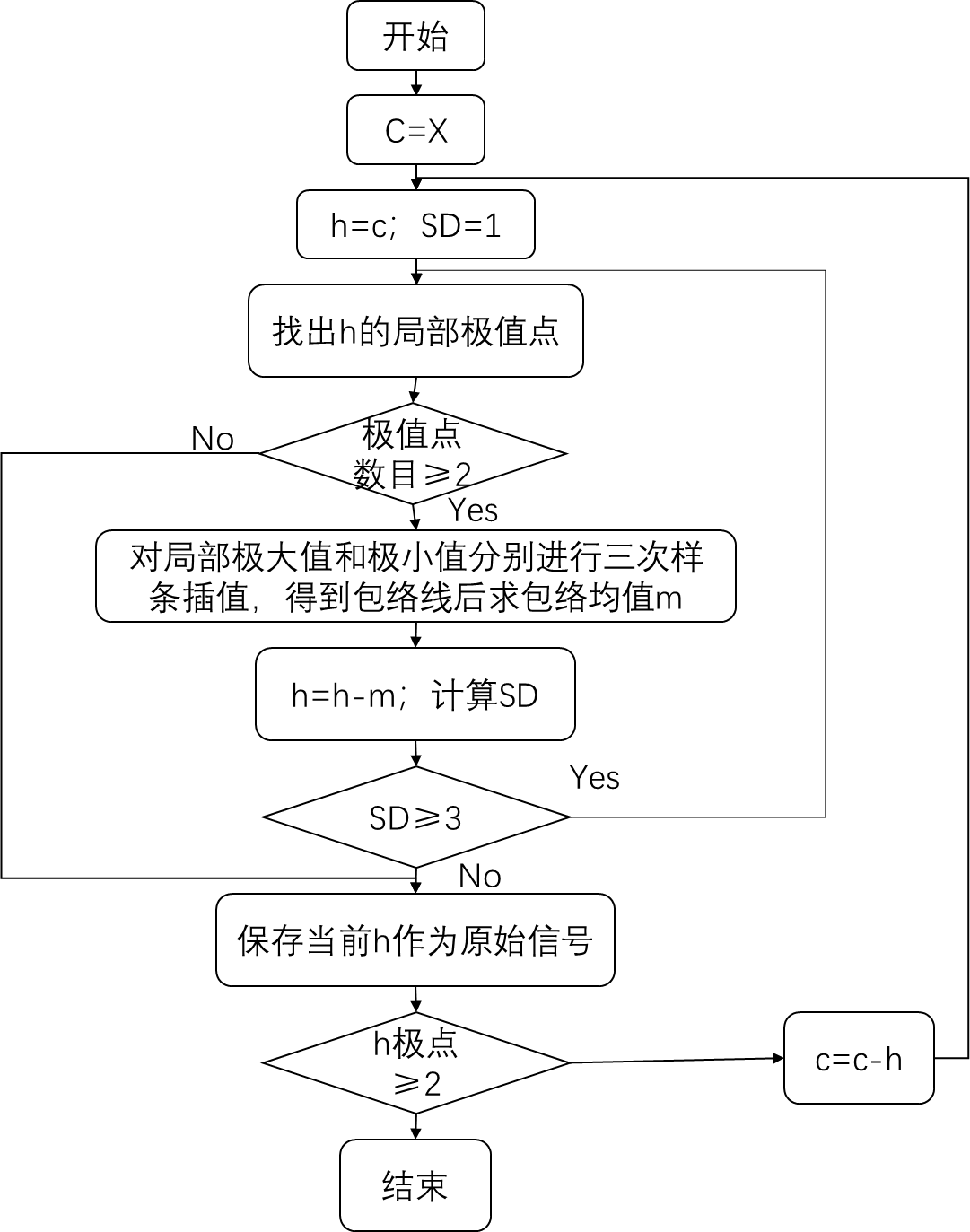
那么就是求得的第一个IMF分量，否则将作为原始信号重复1-2的步骤，直到第k次迭代后差值成为了一个IMF分量，记为

其迭代终止准则是使得

1. 从原始信号中减去得到第一阶剩余信号
2. 将剩余信号作为原始信号进行1-4过程，即

终止准则是当第阶剩余信号足够，不能够在提取IMF，综上，原始信号的分解为

EMD的分解流程图



希尔伯特Hilbert算法

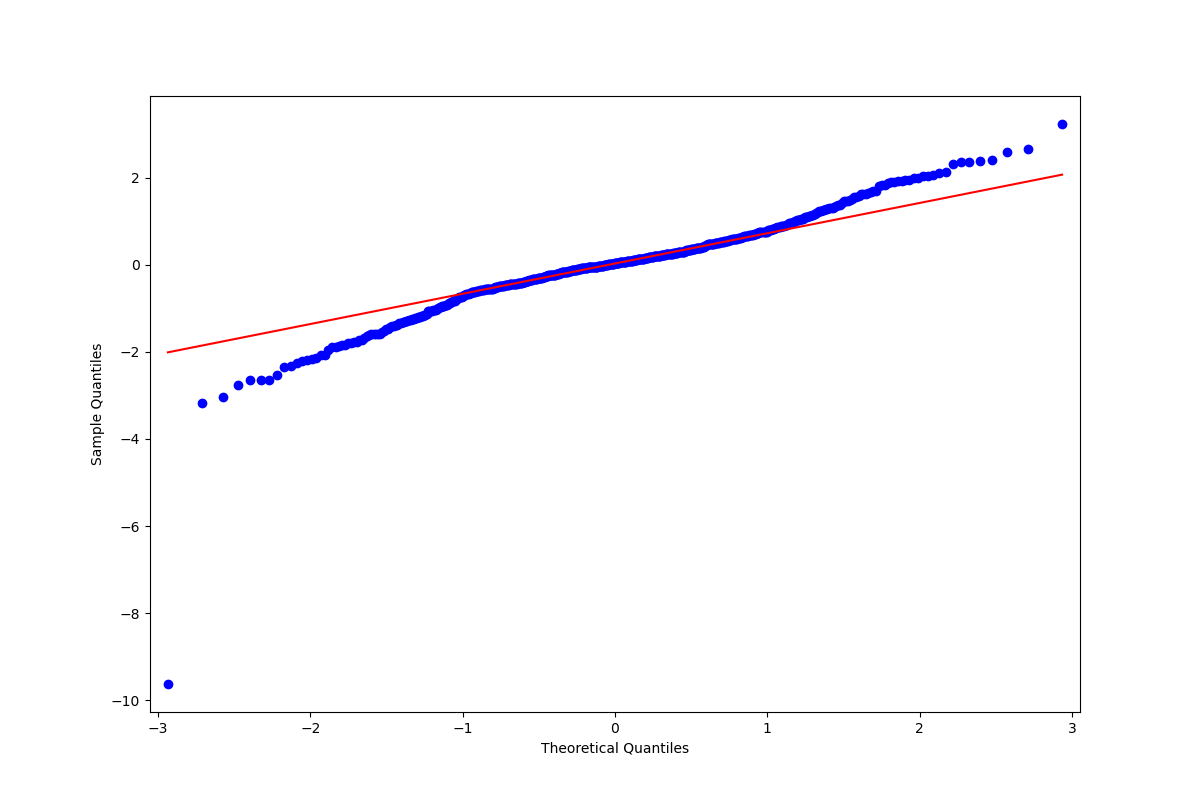
现有信号，定义信号的希尔伯特变换为

故是和的卷积。求解。

引入欧拉公式

由欧拉公式可知，当，即，即，整理得

可知，当频率大于0时，相位向左移90度；反之，向右移动90度。

从下图可以看出，Gz信号是由频率25.43hz、1.42hz、2.31hz等几个信号组合而成的

