**问题一**

首先对给定三个人员的数据进行预处理，处理数据的异常值和缺失值，题目中设备采样率为 100Hz，说明数据为连续型，可以用插值方法处理缺失值，如lagrange插值法；数据异常值可以用格拉布斯准则判断，具体方法：

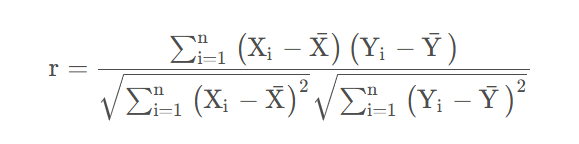
1. 计算数据平均值x\_和标准差s，计算每个数据偏离值xi，然后计算Gi值Gi=(xi-x\_)/s；
2. 自定义检出水平α（0<α<1），置信概率P=1-α。
3. 查拉布斯表获得临界值。
4. 比较临界值G计算值Gi，大于临界值的Gi为异常数据，将其剔除。

然后对各个数据特征进行提取过程中，发现每列数据具有周期性，在Matlab中做出每列数据的统计频率直方图，如Person1\_SY1\_acc\_x(g)有：



用该频率直方图算出每组数据每列的平均值（即提取特征数据）可以将一组数据转化为一个1✖6的数据表格，即可实现数据的降维。

下一步，使用皮尔逊相关性系数来计算数据之间的相关性：



得到一个相关性系数矩阵C（60✖60）。要将一个人的60个数据合理分为12组，可以将每组的5个数据之间的相关性系数相加，最后得到12的相关性系数总和最大即为最优，可以使用优化规划模型来解决：

**1. 定义变量**

设X为一个 60 x 12 的二进制矩阵，其中X[i][j]表示第i个数据是否被分到第j组（1表示分到该组，0表示不分到）。

**2. 目标函数**

构建一个目标函数，表示12组每组相关性系数的总和。

**3. 约束条件**

（1）每组必须包含5个数据：

（2）不能重复选择数据：

**4. 求解**

线性规划求解。

将三个人的数据分别分为12组后，求他们的各自的12组数据的均值，分别得到3组每组12个一维数据。然后，取Person1和Person2数据求其皮尔逊相关性系数，得到一个相关性系数矩阵E（12✖12），再利用上面的优化规划模型将24个数据分为12组，填入表中，即完成了问题一中两个实验人员的分组；最后，取Person3数据和Person2数据重复上述操作，将所得Person3分组填入表中，即得问题一的解。

**问题二**

要建立人员活动状态的判别模型，首先对该10名实验人员的数据进行问题一中的数据清洗和降维处理，提取特征数据，即问题一中的降维处理，然后将提取的特征划分为训练集和测试级，训练一个随机森林模型。

1. 用问题一的分类模型对10名实验人员的数据分类，分析分类准确度。
2. 先对该实验人员的数据进行清洗和降维处理，然后运用随机森林模型判别其活动状态。

**问题三**

要研究活动状态数据与实验人员的年龄、身高、体重的关系，通过数据进行人员画像，先分别提取未知实验人员数据特征（用实验一中降维方法），再求10位实验人员的12类活动数据的均值，然后分别计算各实验人员与未知实验人员的皮尔逊相关性系数总和，认为总和最高时，未知实验人员与该实验人员相匹配，即为判别结果。