关于数据归一化和还原：

归一化是将数据缩放到0-1之间，是标准化的一种

数据归一化和数据集的划分问题：

我们训练用的是归一化后的数据，测试数据不做处理，如果直接用归一化的训练集计算得到的模型系数会出问题

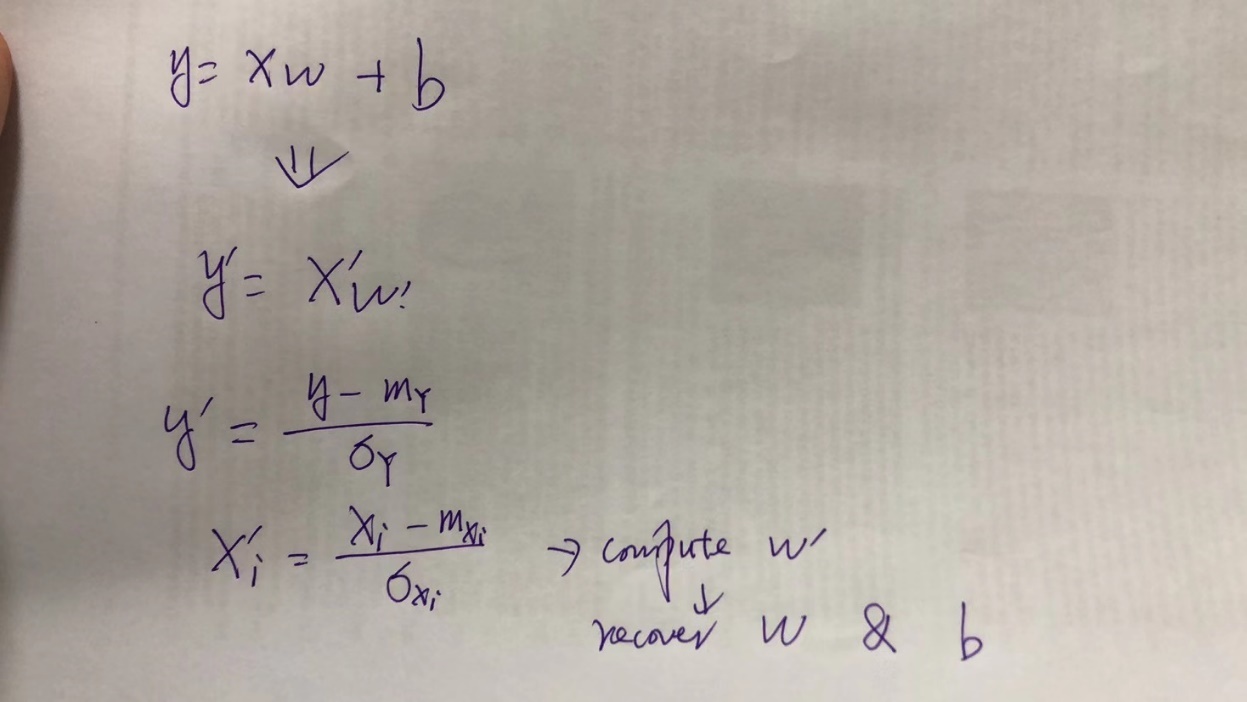
原始数据被划分为训练集和测试集，测试机不作任何处理，训练集先归一化，用归一化之后的训练集拟合模型得到的模型系数只能和归一化的数据对应，不会拟合原始数据，故需要将归一化的模型系数还原到与原始数据对应的系数，还原之后的模型预测测试集？

老师说的：

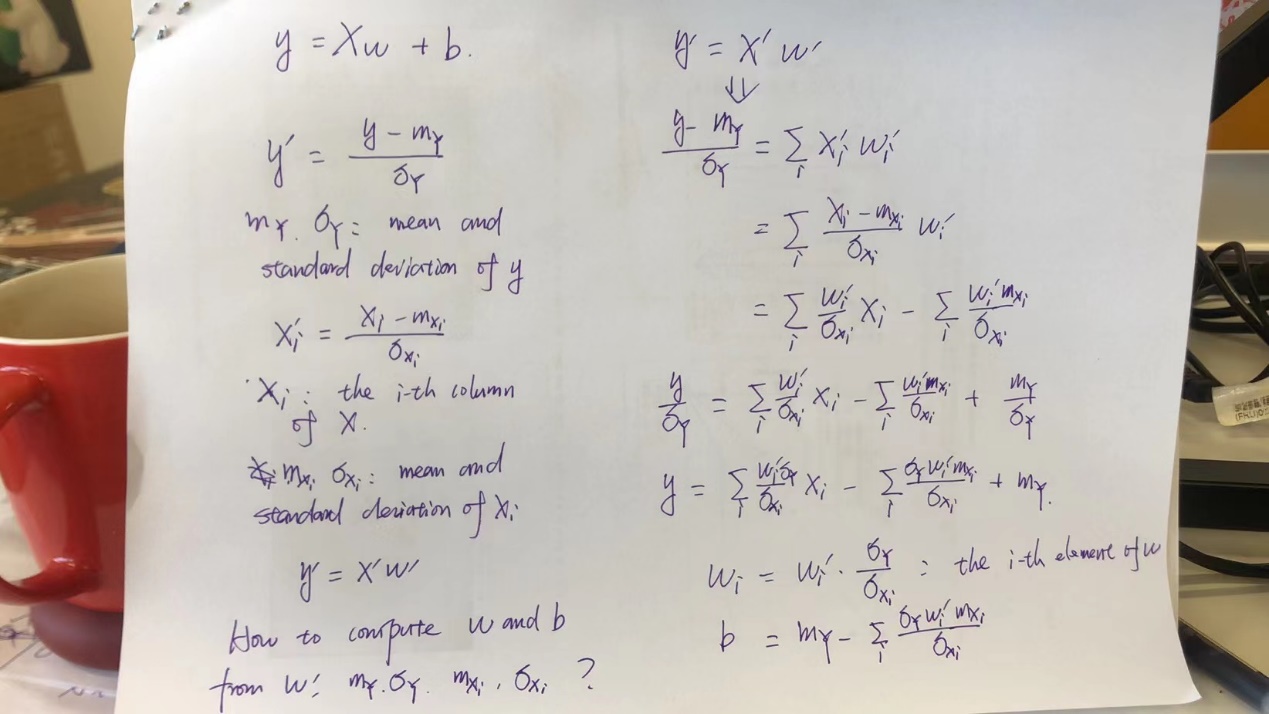
线性回归里面的归一化目的是将线性回归方程的截距消除，y=wx+b变成y’=w’x’；

此时等式中的x’,y’就是归一化的数据，而w’就是用归一化训练集得到的与原始数据不相对应的模型系数（为啥归一化能消除截距b）

与其并行的消除b的简便方法就是将x加一列1，但此时是默认了b和x一样（tls就不能这样干，因为x有误差，但是b并没有误差）



使用归一化消除截距并还原到与原始数据对应的模型系数，这样子的话测试集就不用做任何处理，直接使用还原之后的模型测试



统一归一化：也ok

**MLE推导线性回归：**

最小二乘估计：

但数据都带有一定噪声，假设噪声满足零均值的高斯分布

因此，

则

使用MLE求解W：

首先是最大似然函数：



再求W：



样本权重：三个数据集组成，三批电池，有可能造成估计误差，假设同一批次的电池误差一样，

Tls：以列为单位误差估计

# EM for LS

, where , 样本服从： , LS的目标优化函数为：

为每个样本赋予不同权重（不同样本的重要性不同）：

假设噪声服从零均值的不同方差的高斯分布： , 那么就服从高斯分布 .似然函数为：

最大化似然函数就等于最小化该目标函数 , （此处推理用MLE求解带有高斯噪声的线性回归已证明，另：中也含有变量，直接等于是因为变化大小对结果无太大影响？）对比我们就可以得到：

因为服从的高斯分布标准差不能作为已知条件提前得到，故我们将作为EM算法中的隐变量：

（按照EM算法思想，似然函数是用来更新初始化更新的参数）

1.赋予初始值

2.利用估计隐变量（①基于每个数据样本的残差，将样本使聚类的方法分成几组，每组的标准差一样；②先使用统一的方差，再在迭代中更新每个样本的方差）：

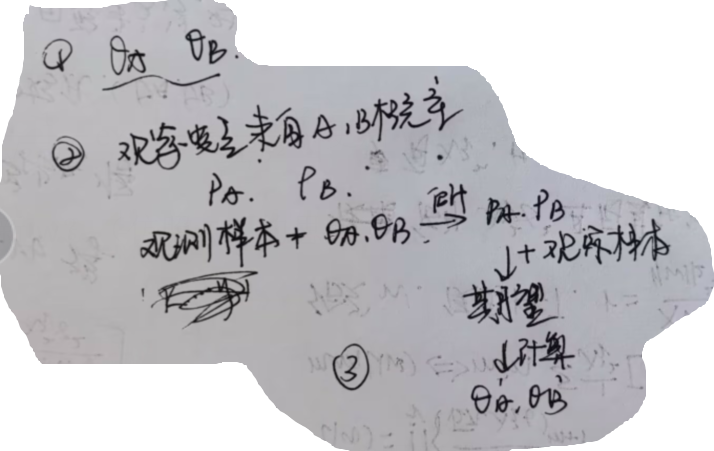
1）聚类：根据判断样本来自聚类之后的哪个样本组别，求出对应的最小？（K-means）

2）最开始使用统一的方差：？？？？更新应该如何更新

3.再利用更新：

利用最大化似然函数更新w？（EM）

利用最小二乘的目标函数更新w？（仿照TLS的思路）



赋予初始值

E-step：求出隐变量，（根据样本值和w初始值判断样本的标准差/方差归属于哪一个聚类组别/使用统一方差的方法时应该如何理解：判断样本的标准差是哪一个具体的值？第二步：根据标准差和样本数据求出期望）

M-step:更新，（根据E-step中求得的期望更新w）

（K-means和GMM的隐变量都是来自不同类别的样本中心点/参数不同的高斯模型概率，但此处样本的标准差就是隐变量，且将其融入了求解w的公式中）

EM公式：

E-step：（此处公式中的对i求和，是对于的i求和，使用聚类即为类别总数，使用统一方差即以单个样本为单位，为样本总数）



·聚类：仿照k-means方法先对样本点进行聚类？初始的标准差该如何定义？(based on the residual of each data sample)；K-means更新隐变量

，此处更新隐变量是min？？？

·统一方差：we can start from a unified variance and the variance of each data sample is adapted in each iteration，仍然是更新隐变量函数如何定义？min？

M-step：（EM算法最大化似然函数更新参数）

# EM for TLS

, 样本是带有噪声的，需要矫正，使 ,前者是不带有噪声的Xi.假设 and , follow the Gaussian distribution ,

the probability of data sample conditioned on the model is , where and .

Thus, the likelihood is :

Similarly, we can assume that and follow . Thus, the likelihood is updated to

Maximize the log-likelihood is equivalent to minimize

Given a , use KKT to compute , . We can thus estimate for each data sample and use it to optimize in next iteration. Given , how to compute ? Just scale each data sample by multiplying both and to

初始化：赋予初始值，

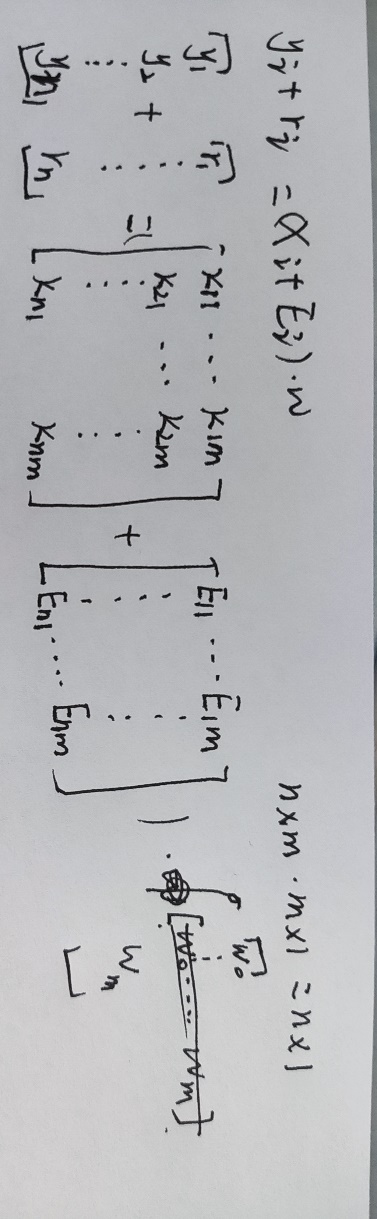
E-step：利用估计隐变量（按照EM公式求期望，该如何类比EM样本来自哪个类别）

M-step:再利用更新：

1. 使用什么函数更新隐变量
2. 什么函数更新w

(multiplying both and to ：min, , .)

# Different Noise Level on Each Column（列）



We assume , follow the Gaussian distribution , and follows the Gaussian distribution . the probability of data sample conditioned on the model is , where and .

Thus, the likelihood is :

Maximize this likelihood is equivalent to minimize

The corresponding optimization problem is

, s.t.

We consider two scenarios:

## When the variances of all columns in are identical such that

Let , , s.t.

,

KKT Condition: (1), (2), (3)

From (1) and (2): , , (4)

From (3) and (4): , , (5)

From (4) and (5): (6)

Put (5) and (6) into the objective function:

(5)

From (4) and (5): (6)

-> -> （同时乘以因为大于零min目标不变）

Compute by solving the standard TLS problem, and then compute

Initialize and

E: Estimate and , , , compute （隐变量是and ，此处根据设定，求出隐变量就可以得到的值）

M: Compute based on given （根据求是利用tls求解的结果，说明了，M-step就是最小化函数？？？？？？？）

## When the variances of all columns in are not assumed to be identical

, s.t.

, s.t.

, s.t. ,

,

KKT:

(1)

(2)

(3)

From (1) and (2): , , (4)

From (3) and (4): , , , (5)

From (4) and (5): (6)

Put (5) and (6) into the objective function:

（是对角矩阵，就相当于乘以一个单位矩阵）

Let , the problem is equivalent to

Initialize and

E: Estimate variances , , update

M: Compute based on current

（隐变量从最初的和变成了求两者之间的关系,求解过程只需要知道良知的关系，对原始值无需定义，M-step求解可以直接用，但在求解过程中是最大化似然函数？？？？？）