LASSO

中国人民大学《复杂数据统计方法-基于R的应用》人大经济论坛

岭回归的问题：

岭参数计算方法太多，差异太大

根据岭迹图进行变量筛选，随意性太大

岭回归返回的模型（如果没有经过变量筛选）包含所有的变量

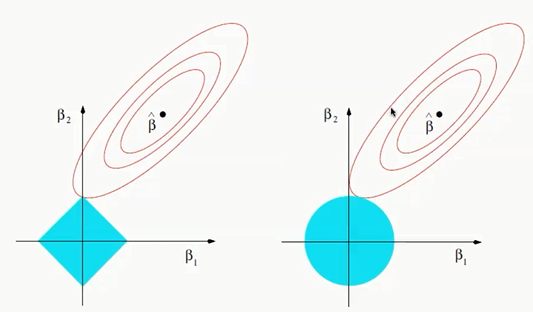
Lasso 1996由Tibshirani提出（The Least Absolute Shrinkage and Selectionator operator）算法

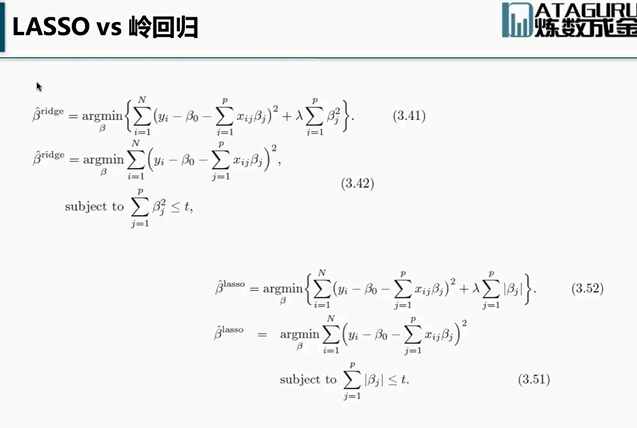
通过钩子哦一个一阶惩罚函数获得一个精炼的模型；通过最终确定一些指标（变量）的系数为零（岭回归估计系数等于0的机会微乎其微，造成筛选变量困难），解释力很强。

擅长处理具有多重共线性的数据，与岭回归一样是有偏估计。

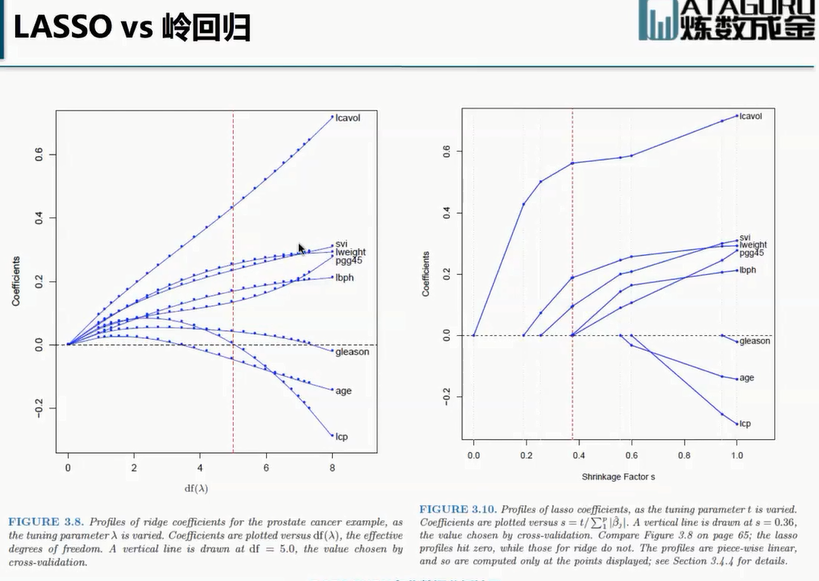
为什么Lasso能直接筛选变量？

右边的图形就是岭回归当中的。左边的图形是lasso的图形表示（菱形）、





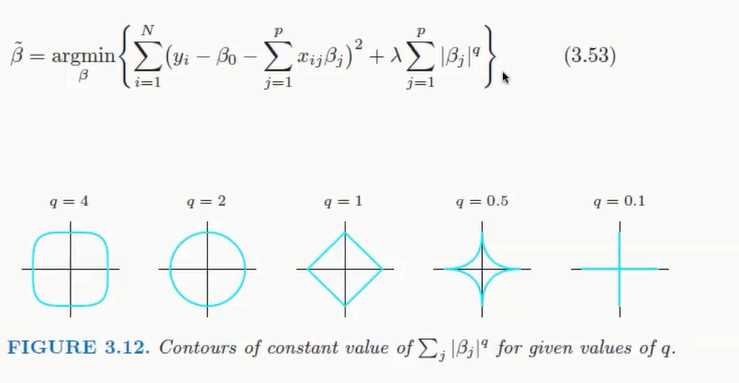
上图左上角岭回归，右下角lasso。



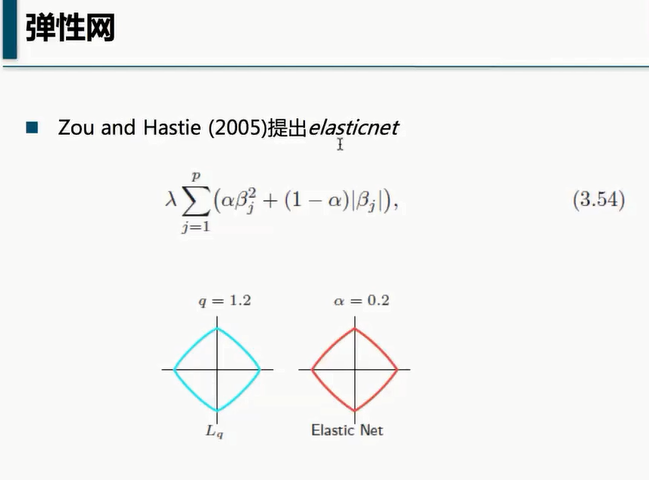
上图左边图形是岭回归的，横轴是自由度，和岭系数k有关。

右边图形描述lasso，从右到左是一个收缩的过程，可以想象成菱形再收缩。横轴是

========更一般化的模型==============



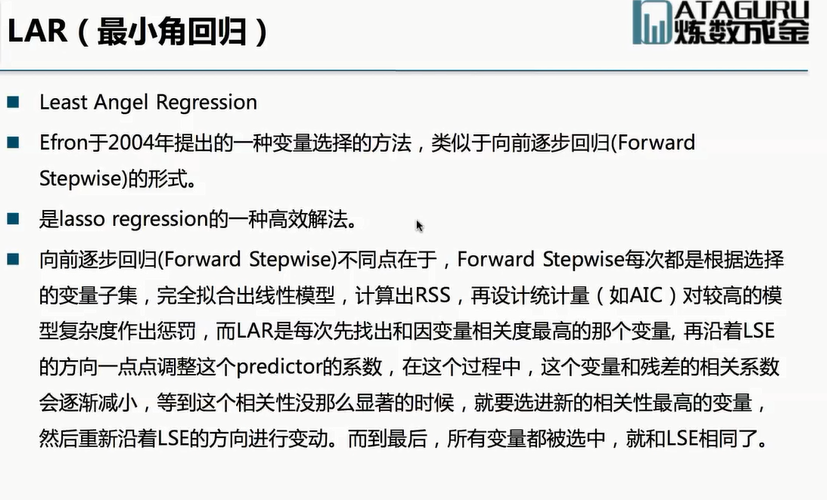
===========弹性网===========

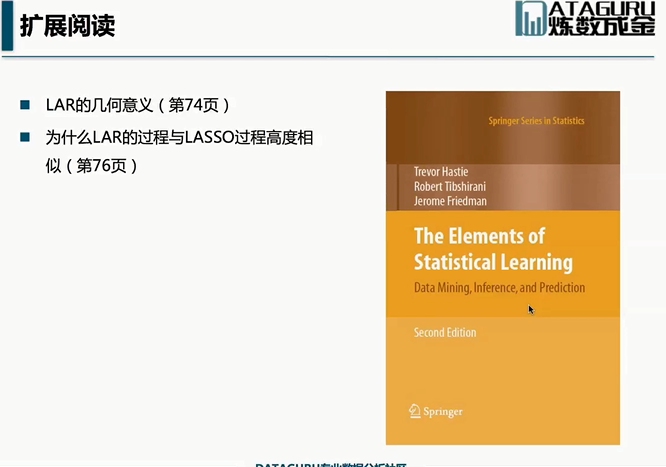


既有岭回归的部分又有lasso的部分，画出来的形状是上图右侧。

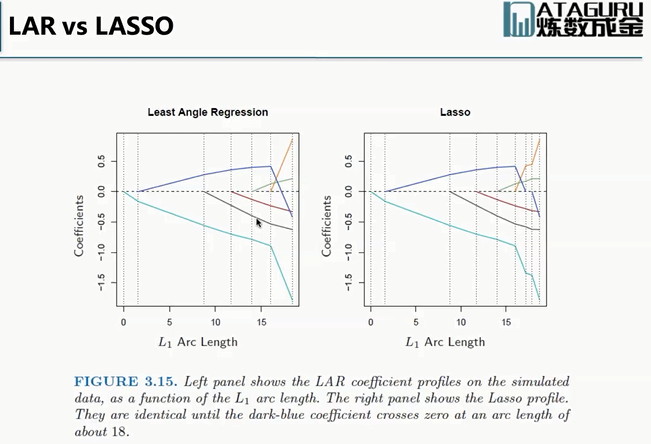
筛选变量非常有效，而且损失的精度不会太多。

==========最小角回归====================

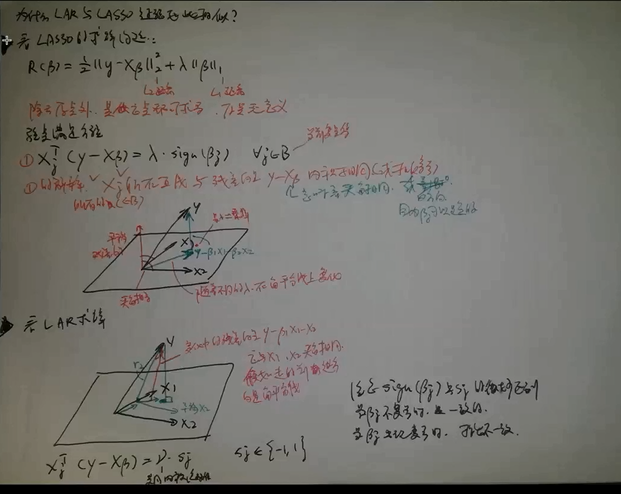


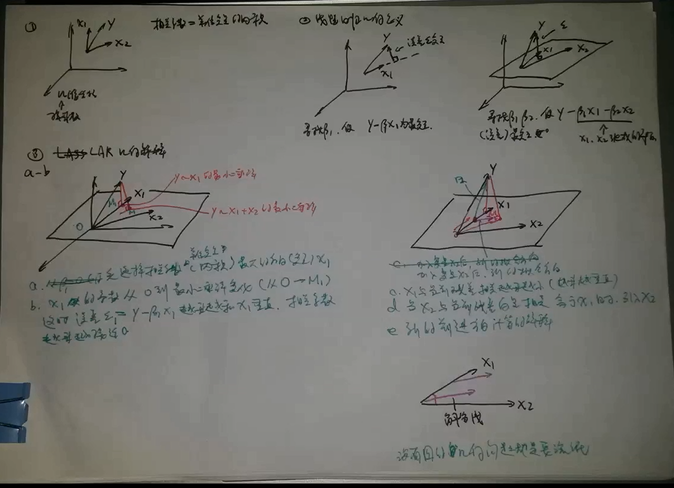


最小角回归最早是为了解决线性回归问题，它给出一种在几何上有非常清晰的几何意义，很有解释能力的一种方法，描述再几何上是怎样逼近最小二乘法解，是面向传统的回归问题。类似于线性回归当中向前逐步回归。LAR描画出来的解的过程和Lasso是完全一样的，如下图：



介绍LAR之前呢，首先解释下相关系数的几何意义：



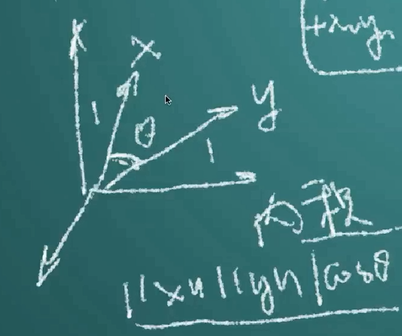


相关系数：假如有两个变量x，y；协方差(y,x)除以（根号下方差x\*根号下方差y）

Cov(y,x)/(根号下x的方差\*根号下y的方差)

Cov(y,x)=x1y1+x2y2+…xnyn (前提是xy都已经做了中心化和标准化)

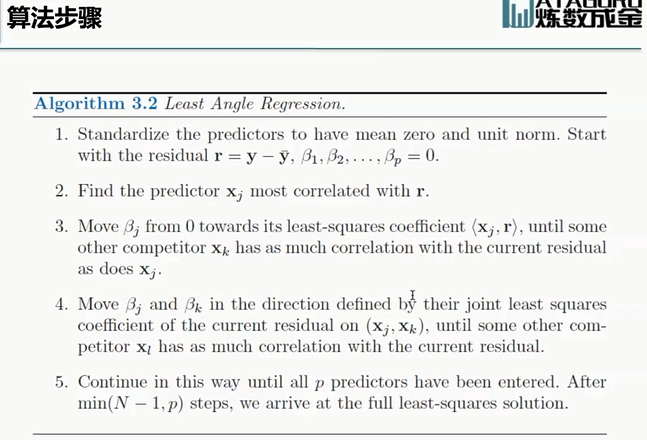
此为两个向量的内积，两个向量的模相乘再乘以costheta,如下图：

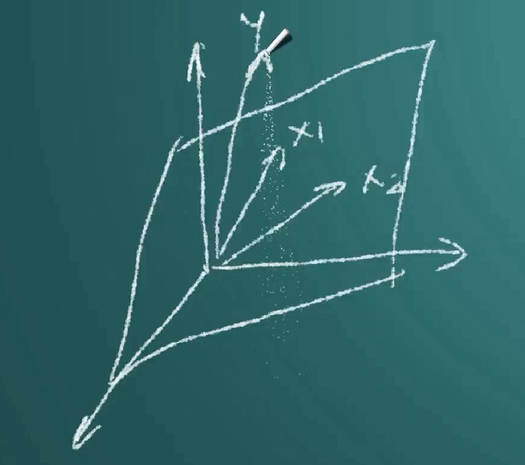


**相关系数的几何意义**：再样本作了中心化标准化之后呢，他们之间的相关系数就等于他们之间夹角的余弦。当两个向量快重合时，theta就是0，算出来的相关系数就是1.

他们呈现高度相关，如果他们相关程度很低，则两个向量越接近于垂直，相关程度越低。当夹角是180度时，相关程度依然很高，但是确是负相关，x越大，y越小。

最小角算法：





最小二乘解就是从y的顶点到平面x1x2的距离。

最小角回归是从一堆变量当中选择一个变量，它和残差变量的相关系数是最大的。

一开始残差向量是y，从x1、x2……当中选择一个变量，它和y的相关系数是最大