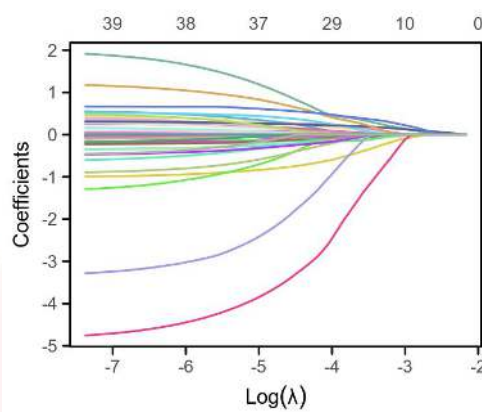


临床意义 - 预后 Lasso 变量轨迹[记录]



网址: <https://www.xiantao.love>



更新时间: 2023.03.07

## 目录

基本概念 .....	3
应用场景 .....	3
分析流程 .....	4
结果解读 .....	5
数据格式 .....	6
参数说明 .....	7
类型 .....	7
线 .....	8
标注信息 .....	9
标题文本 .....	11
风格 .....	12
图片 .....	12
结果说明 .....	13
主要结果 .....	13
方法学 .....	14
如何引用 .....	15
常见问题 .....	16

## 基本概念

- **Lasso 回归**：在线性回归的基础上，通过增加**惩罚项**（ $\lambda \times \text{斜率的绝对值}$ ），减少模型的过拟合，提高模型的泛化能力。另外一种也是通过增加惩罚项来减少模型的过拟合的方法是岭回归，对应的惩罚项是（ $\lambda \times \text{斜率的平方}$ ）。惩罚项在机器学习领域也叫做正则化，其中，Lasso 回归的惩罚项是**L1 正则化**（曼哈顿距离（参数绝对值求和）），而岭回归的惩罚项是**L2 正则化**（欧氏距离（参数平方值求和））
- Lasso 可用于 logistics、Cox 其中，此模块就是 Lasso 在预后中的应用。预后 Lasso 常常出现在构建预后模型或者筛选变量上，最常出现两种图，一种是系数( $\lambda$ )筛选的图，另外一种**是变量轨迹图**。Lasso 的  $\lambda$  筛选一般会采用**交叉验证**的手段进行筛选，常见的会有五折和十折交叉验证。

## 应用场景

将预后 Lasso 系数筛选过程中各个  $\lambda$  值（惩罚项）与各变量的系数值进行可视化，以**构建预后模型或者筛选变量**。当样本较少或者变量较多（少于样本数一半的变量）时，可以用 Lasso 直接构建预后模型或者筛选变量。

## 分析流程

云端数据 → lasso 预后分析 → lasso 变量轨迹可视化

### ➤ 数据格式：

- 云端数据： 这里的数据来自<预后 Lasso 系数筛选 / 预后 Lasso 系数筛选 [云]>保存后的记录，默认选中的是最新保存的记录，保存的记录可以在<历史记录>中找到对应的

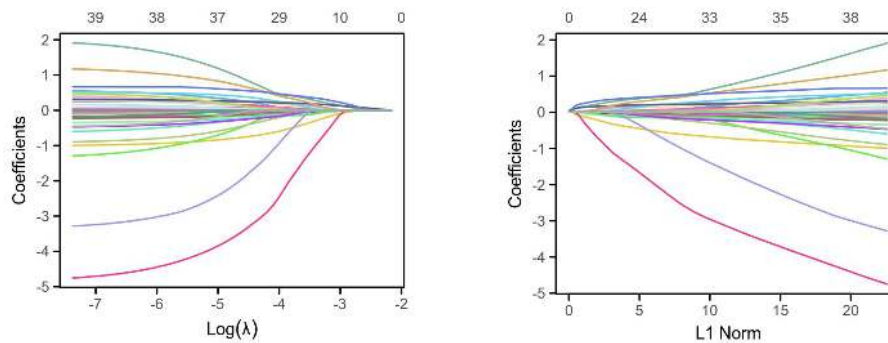
### ➤ Lasso 预后分析：

- 构建 lasso 预后模型
- 计算模型的 lambda 值
- 计算变量的系数值
- 筛选掉 lambda 值对应系数为 0 的变量(系数为 0 表示变量之间不存在相关关系，在预后模型中没有实质上的意义)

### ➤ Lasso 变量轨迹可视化

- 通过不同统计量（lambda 值取对数，或者 L1 Norm (L1 正则化)），分别计算出 x 轴具体的值
- Lasso 预后分析得到的变量系数值作为纵坐标
- 进行可视化

## 结果解读



- 左图横坐标表示  $\lambda$  对数值 ( $\log(\lambda)$ )，纵坐标表示变量的系数值
- 右图横坐标表示向量中各非零元素的绝对值之和 (L1 正则化)，纵坐标表示变量的系数值
- 上方的横坐标的数字代表每个  $\lambda$  下对应的系数非 0 的变量个数
  - 这些数字对应的值是说：不同  $\lambda$  值计算得到模型中所有变量系数不为 0 的变量的个数，而不是所有的变量(要是数值与变量个数对应不上，则是因为缺少的那些变量间不存在相关关系(系数为 0)被筛选掉了)
  - 由于可视化结果是 ggplot2 格式，故不能展示全部的数值
- 图中每条线对应一个变量随 Lasso 惩罚项的  $\lambda$  系数 ( $\log$  后) 的系数变化情况。如左图，可以看到最下边线条对应的变量 (“Gene 28”) 系数最先发生改变，随着  $\lambda$  的减小，非 0 变量的数目逐渐增多
- 图中特定的线还可以标注具体的变量名，这个可以在参数部分的选项卡中输入相关变量名进行可视化

## 数据格式

这里的数据来自<预后 Lasso 系数筛选 / 预后 Lasso 系数筛选[云]>保存后的记录, 默认选中的是最新保存的记录, 保存的记录可以在<历史记录>中找到对应的

数据参数

预后 Lasso

预后lasso系数筛选 / 预后lasso系数筛选-云 @1.0 / 2023-03-07 08:48:39

根据需要可视化的项目选择好对应的云端数据记录

(此样本数据使用的是<预后 Lasso 系数筛选[云]>中的食管癌样本数据, 如下: )

数据参数

云端数据

食管癌 / TCGA / TCGA-ESCA / RNAseq / STAR / TPM @过滤:去除正常+去除无临床信息 @处理:log2(v...

## 参数说明

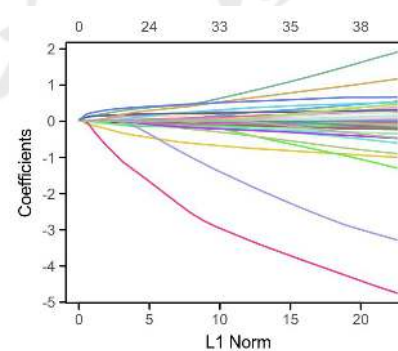
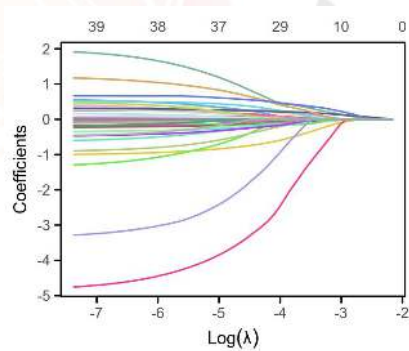
(说明：标注了颜色的为常用参数。)

## 类型

类型

x轴统计量 log(lambda)

- x 轴统计量：可以选择 lasso 变量轨迹的方法：可选择  $\log(\lambda)$  或者 L1 Norm (L1 正则化)，如下：左侧为  $\log(\lambda)$ ，右侧为 L1 Norm



## 线

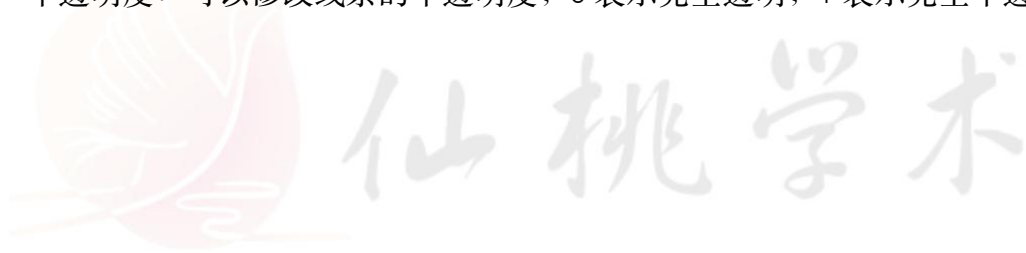
线 

线条类型 实线 

线条粗细 0.75pt 

不透明度 1

- 线条类型：可选择变量轨迹对应线条的类型，可以是实线（默认）也可以是虚线
- 线条粗细：对应图中各个变量系数轨迹的线条的粗细，默认为 0.75
- 不透明度：可以修改线条的不透明度，0 表示完全透明，1 表示完全不透明





## 标注信息

标注

类型选择 不标注

特定变量

标注大小 5pt

- 类型选择：可以选择是否在图中进行变量标注，默认为不进行标注，还可以选择全部标注，或者标注特定变量，如下：

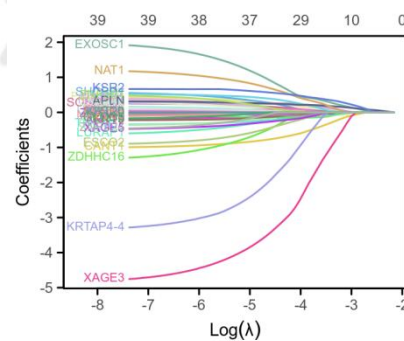
### ■ 标注全部变量

标注

类型选择 标注全部变量

特定变量

标注大小 5pt



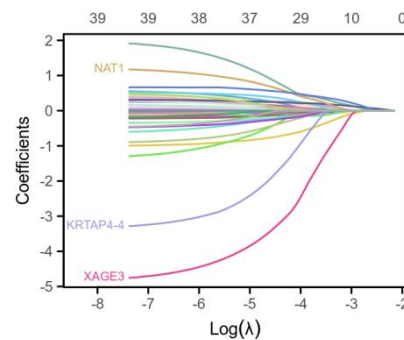
### ■ 标注特定变量

标注

类型选择 标注下面特定

特定变量
NAT1
KRTAP4-4
XAGE3

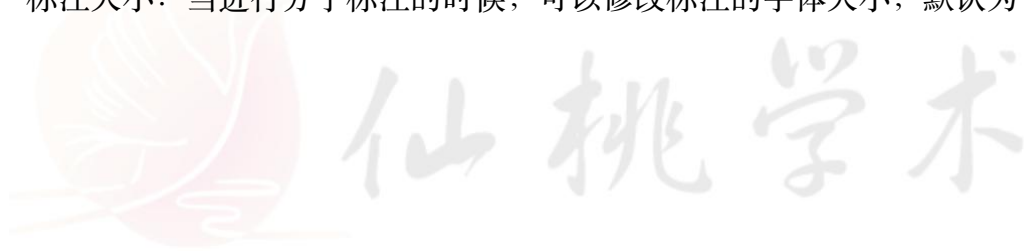
标注大小 5pt



- 特定变量：变量名（分子），可以输入想要标注的变量名才会进行标注，一行为一个变量，用回车键换行。需要和所选择的云端记录对应 Lasso 系数筛选中上传的数据的变量要一致。如果某个变量在 Lasso 模型内不管 lambda 如何改变，始终系数都是 0，则无法在图中进行标注。结果如上：



- 标注大小：当进行分子标注的时候，可以修改标注的字体大小，默认为 5pt



## 标题文本

标题	
大标题	大标题内容
x轴标题	x轴标题内容
y轴标题	y轴标题内容

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如  $\{2\}$ ；如果需要下标，可以用两个英文输入法下的中括号括住，比如  $[2]$

## 风格



风格

边框 ☒

网格 ☐

文字大小 7pt

- 外框：是否添加外框，默认添加
- 网格：是否添加网格
- 文字大小：控制整体文字大小，默认为 7pt

## 图片



图片

宽度 (cm) 6

高度 (cm) 5

字体 Arial

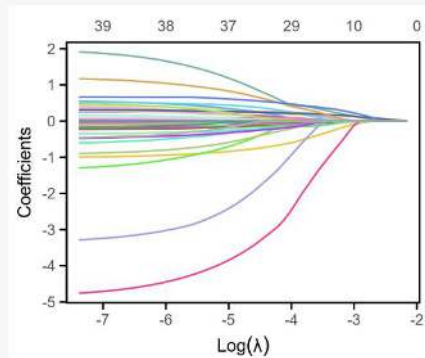
- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

## 结果说明

## 主要结果

### 预后Lasso变量轨迹-记录

预后lasso变量轨迹: 对变量系数的变化情况进行可视化



[预后lasso变量轨迹.pdf](#)

[预后lasso变量轨迹.tiff](#)

[预后lasso变量轨迹.pptx](#)

- 横坐标表示lambda对数值 ( $\log(\lambda)$ ), 纵坐标表示变量的系数值
- 上横坐标表示此模型中非0系数的变量个数, 每一条曲线表示每一个变量系数的变化轨迹
- 随着参数的不断增大, 变量的系数最终被压缩为0, 说明比较重要

## 方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包：survival 包（用于构建模型）、glmnet（用于分析）

处理过程：

- (1) 使用 survival 包对清洗过后的数据进行模型构建，使用 glmnet 包进行分析得到 lambda 值、各变量的系数值等
- (2) 对数据进行可视化



## 如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分 (Acknowledge) 致谢仙桃学术 ([www.xiantao love](http://www.xiantao love))。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。



## 常见问题

### 1. 为什么进入到模块没有数据进行分析?

答：此模块是记录类的模块，数据是自动上传的（[这里的数据来自<预后 Lasso 系数筛选 / 预后 Lasso 系数筛选\[云\]>保存后的记录](#)，默认选中的是最新保存的记录，[保存的记录可以在<历史记录>中找到对应的](#)），所以进行该模块分析之前需要渠道预后 lasso 系数筛选模块保存结果到历史记录，然后在进入到此模块就会有相应的数据了，如下步骤：

#### ① 进行预后 Lasso 系数筛选 / 预后 Lasso 系数筛选[云] 分析，保存结果

**主要参数** 保存参数 重置参数

方法 > 点 > 误差线 > 标题 > 风格 >

图片 >

确认

**保存结果** **下载整份报告**

**主要结果** **补充结果** **方法学**

**预后lasso系数筛选**

预后lasso系数筛选: 可视化lasso系数筛选过程中各个lambda值 (惩罚值) 对应的统计量(似然偏差值或C指数)

作用: 预后lasso常用于构建预后模型或者筛选变量

交叉验证: 十折交叉验证 || 种子号: 2022

---

**主要参数** 保存参数 重置参数

预后参数 > ID列表 > 方法 > 点 > 误差线 >

标题 > 风格 > 图片 >

确认

**保存结果** **下载整份报告**

**主要结果** **补充结果** **方法学**

**预后lasso系数筛选-云**

预后lasso系数筛选: 可视化lasso系数筛选过程中各个lambda值 (惩罚值) 对应的统计量(似然偏差值或C指数)

作用: 预后lasso常用于构建预后模型或者筛选变量

交叉验证: 十折交叉验证 || 种子号: 2022

预后类型: OS[Overall Survival]

#### ② 进入此模块，选择数据



## 数据参数

预后  
Lasso

①

预后lasso系数筛选 / 预后lasso系数筛选-云 @1.0 / 2023-03-07 08:48:39

点击选择数据

## 云端数据

	记录名称	来源模块	时间	补充说明
<input checked="" type="checkbox"/>		预后lasso系数筛选 @1.0	2023-03-07 09:15:24	数据记录可以在历史记录中找到
<input type="checkbox"/>	预后lasso系数筛选	预后lasso系数筛选-云 @1.0	2023-03-07 08:48:39	数据记录可以在历史记录中找到
<input type="checkbox"/>		预后lasso系数筛选-云 @1.0	2023-03-06 18:05:40	数据记录可以在历史记录中找到

2. 右侧的参数中输入了变量名，但是没有在图中进行标注？

答：变量名必须与上传数据中的变量名（除了第 1、2 列）一致，并且输入变量名的时候应该一个变量名一行，然后换行输入下一个变量名

3. 图中标注的部分超过了外框？图中标注的内容有重叠，如何解决？

答：由于图的文字是会被压缩的，所以只能通过增加图片的宽度或者高度来解决，或者减少需要标注的分子数量或缩短标注分子的名字。

#### 4. 为什么上传的数据的变量数目和图中对应的最大的变量数不一致?

答：图中最上方的横坐标对应的最多变量的个数对应的是非 0 系数的变量个数。  
如果某些变量在 Lasso 的不同 lambda 的系数自始至终都是 0, 则不会在图中出现。

#### 5. 如何修改某条线/某个变量对应的颜色?

答：当数据记录在「Lasso 变量轨迹」模块被保存时，也会一同保存一份随机生成的颜色。这个颜色跟对应的数据记录是绑定的，也就是一份数据记录对应一份颜色，无法进行修改。如果想要更换某些变量对应的颜色，可以在「Lasso 变量轨迹」模块中重新保存一份数据，对应 Lasso 变量轨迹的整个颜色都会改变。

#### 6. 为什么图上方非 0 系数变量的个数与数据中的变量个数对应不上? 为什么看不到所有的数字，只是一小部分?

答：

①图上方的这些数字对应的值是说：不同 lambda 值计算得到模型中所有变量系数不为 0 的变量的个数，而不是所有的变量(要是数值与变量个数对应不上，则是因为缺少的那些变量间不存在相关关系(系数为 0)被筛选掉了，或者变量在数据处理过程中就被筛选掉了)

②由于可视化结果是 ggplot2 格式，故不能展示全部的数值