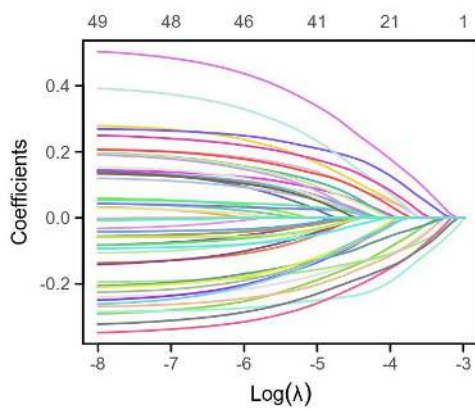


临床意义 — 诊断 Lasso 变量轨迹



网址: <https://www.xiantao.love>



更新时间: 2023.05.29

目录

基本概念	3
应用场景	3
分析流程	4
结果解读	6
数据格式	7
参数说明	8
类型	8
线	9
标注信息	10
标题文本	12
风格	13
图片	13
结果说明	14
主要结果	14
方法学	15
如何引用	16
常见问题	17

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 在诊断中的应用。诊断 Lasso 常常出现在构建诊断模型或者筛选变量上，最常出现两种图，一种是系数(λ)筛选的图，另外一种为变量轨迹图。Lasso 的 λ 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

将诊断 Lasso 系数筛选过程中各个 λ 值（惩罚项）与各变量的系数值进行可视化，以**构建诊断模型或者筛选变量**。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 直接构建诊断模型或者筛选变量。

分析流程

上传数据 → 数据处理(清洗) → lasso 诊断分析 → lasso 变量轨迹可视化

➤ 数据格式: xlsx / csv / txt 文件格式:

- 第 1 列数据作为结局变量(事件发生情况), 可以是数值类型也可以是分类型数据, 需要是二分类类型, 用 (0 和 1, 0 表示未发生事件, 1 表示发生了事件), 默认会把先出现的组作为参考组。注: 第 1 列不能都是删失

	A	B	C	D	E	F	G	H	I
1	event	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
2	1	-0.022280617	-2.770435561	-0.467054253	0.659710026	-0.515117091	0.026623033	0.964868218	-1.064910322
3	1	-1.183217086	-0.316118003	0.60603766	-0.303598493	0.148727871	-1.21013808	1.213376601	-0.195115978
4	0	-0.622974392	1.845886201	1.517668885	-0.839217309	-0.273699539	-1.914794488	0.986931949	-0.185920375
5	1	-0.961432206	-0.136129034	0.707027039	-2.240777738	-0.115896498	-1.678393183	0.581343256	1.248839562
6	1	-2.009057914	0.754472479	-1.360111214	0.743456609	1.242012981	0.373015672	0.65579689	-0.654581201
7	0	0.79356585	-0.236620914	-0.501233774	0.938055954	-1.219659013	-1.625081979	0.328081467	1.04612915
8	1	-0.291946728	-0.194660575	0.208889948	-0.744460884	-1.593752647	-0.118096617	-1.29451323	1.312739415
9	1	0.709801941	-0.255890999	1.345437412	-1.040223718	-0.040457512	0.702152223	-0.592764399	2.245095696
10	1	0.257086806	0.372722112	-0.013511554	-1.046213702	0.86385945	0.767475738	0.847607122	-1.561368255
11	1	2.504925108	1.371715847	-0.114156296	-1.722567123	-0.052316113	-1.274151487	1.361514751	-0.678908796
12	1	0.44265243	-0.456238875	-0.015031461	1.584885619	0.05882662	-1.299684282	-1.501592696	-1.009258186

- 第 2 列开始直至后面每一列都代表一个样本/变量/分子, 必须是数值类型数据

➤ 数据处理: 分别对第 1 列 (事件)、第 2 列开始后的所有变量进行清洗 (去掉数据中的非数值或者不符合条件的数据)

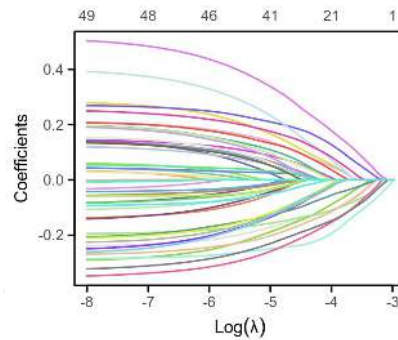
➤ Lasso 诊断分析:

- 构建 lasso 诊断模型
- 计算模型的 lambda 值
- 通过 lambda 值计算变量的系数值

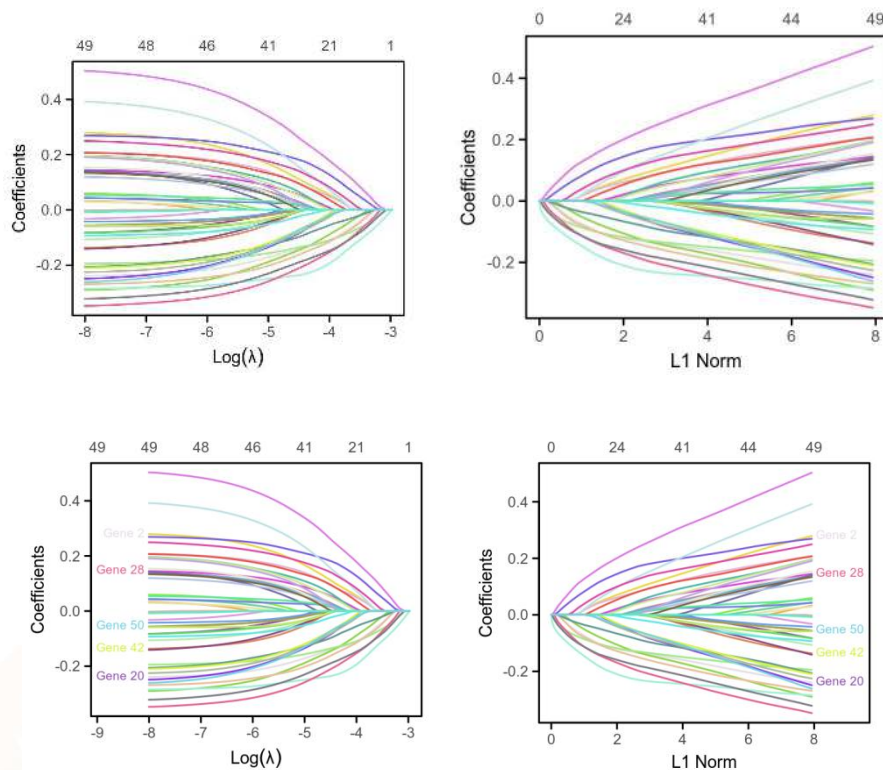
- 筛选掉 λ 值对应系数为 0 的变量(系数为 0 表示变量之间不存在相关关系，在诊断模型中没有实质上的意义)

➤ Lasso 变量轨迹可视化

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



结果解读



- 左图横坐标表示 λ 对数值 ($\log(\lambda)$) 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 λ 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 λ 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 λ 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 2”) 系数最先发生改变, 随着 λ 的减小, 非 0 变量的数目逐渐增多

- 图中特定的线还标注了变量名，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

	A	B	C	D	E	F	G	H	I
1	event	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
2	1	-0.022280617	-2.770435561	-0.467054253	0.659710026	-0.515117091	0.026623033	0.964868218	-1.064910322
3	1	-1.183217086	-0.316118003	0.60603766	-0.303598493	0.148727871	-1.21013808	1.213376601	-0.195115978
4	0	-0.622974392	1.845886201	1.517668885	-0.839217309	-0.273699539	-1.914794488	0.986931949	-0.185920375
5	1	-0.961432206	-0.136129034	0.707027039	-2.240777738	-0.115896498	-1.678393183	0.581343256	1.248839562
6	1	-2.009057914	0.754472479	-1.360111214	0.743456609	1.242012981	0.373015672	0.65579689	-0.654581201
7	0	0.79356585	-0.236620914	-0.501233774	0.938055954	-1.219659013	-1.625081979	0.328081467	1.04612915
8	1	-0.291946728	-0.194660575	0.208889948	-0.744460884	-1.593752647	-0.118096617	-1.29451323	1.312739415
9	1	0.709801941	-0.255890999	1.345437412	-1.040223718	-0.040457512	0.702152223	-0.592764399	2.245095696
10	1	0.257086806	0.372722112	-0.013511554	-1.046213702	0.86385945	0.767475738	0.847607122	-1.561368255
11	1	2.504925108	1.371715847	-0.114156296	-1.722567123	-0.052316113	-1.274151487	1.361514751	-0.678908796
12	1	0.44265243	-0.456238875	-0.015031461	1.584885619	0.05882662	-1.299684282	-1.501592696	-1.009258186

数据要求：

- 列数：至少需要 3 列以上的数据，最多 300 列（299 个变量）的数据，行数：至少需要 20 个以上的样本(20 行)，暂时支持最多 3000 个以上的样本
 - 第 1 列表示事件发生的情况（或结局），二分类类型
 - ◆ 可以是数值类型也可以是分类类型数据
 - ◆ 不能含有无法识别的特殊字符或者非法字符
 - 第 2 列及以后每一列数据都需要是数值类型
 - ◆ 不能含有非数值类型数据，或者混合数值与非数值类型数据
- 列名(样本名)不能重复

参数说明

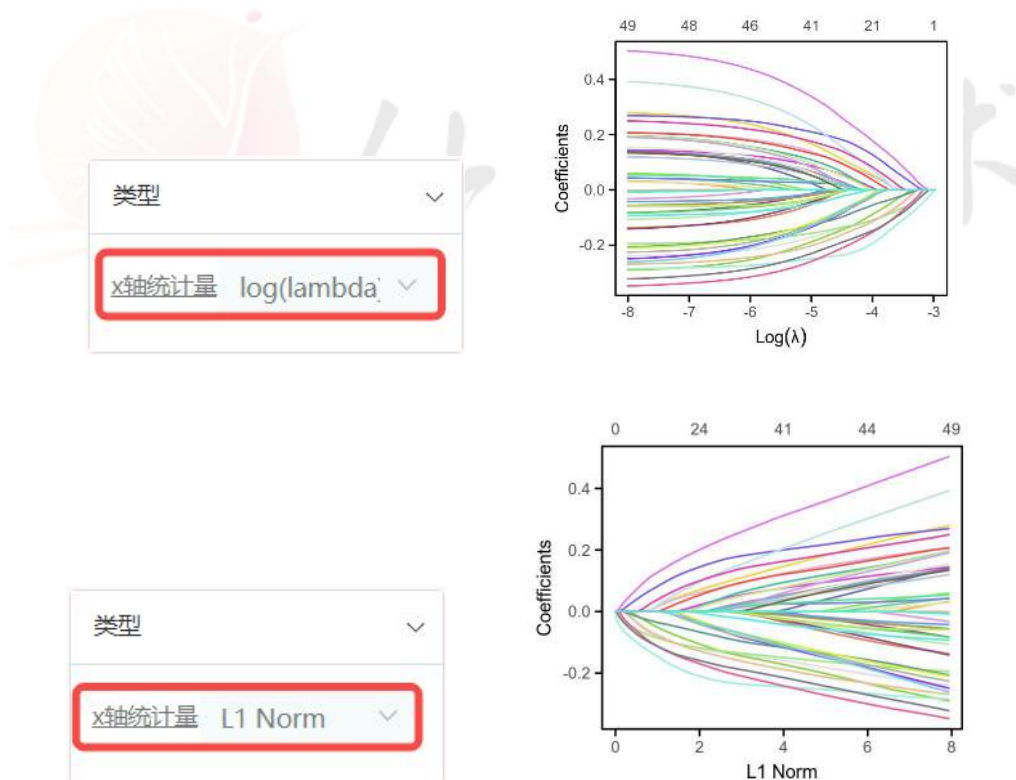
(说明：标注了颜色的为常用参数。)

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，如下：左侧为 log(lambda)，右侧为 L1 Norm



线

线

线条类型

实线

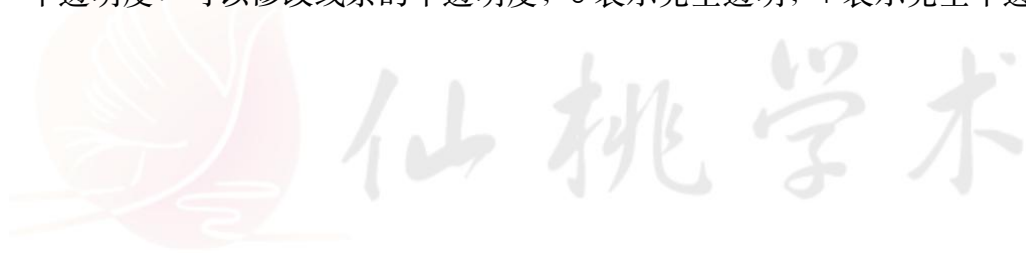
线条粗细

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

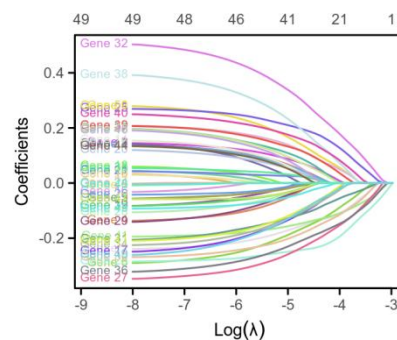
类型选择

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

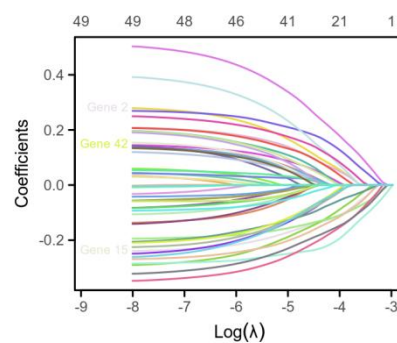
标注下面特定

特定变量

Gene 2
Gene 15
XXXX
Gene 42
Gene 20

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 λ 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

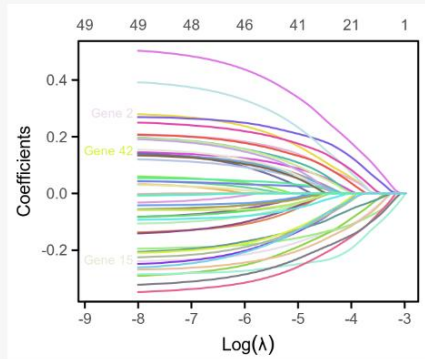
结果说明

主要结果

诊断lasso变量轨迹

诊断Lasso变量轨迹: 主要用来展示变量系数动态筛选过程

· 模型对应二分类结局: 0 vs. 1 (其中参考组: 0)[影响lasso回归非零系数的正负和分析预测值]



诊断Lasso变量轨迹.pdf

诊断Lasso变量轨迹.tiff

· 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：glmnet（用于分析及可视化）

处理过程：

- (1) 使用 glmnet 包对清洗过后的数据进行分析得到变量系数值、lambda 对数值、L1 正则化值等
- (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)。

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. 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 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

2. 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：由于图的文字是不会被压缩的，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。

3. 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 λ 的系数自始至终都是 0, 则不会在图中出现。

4. 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。

5. 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 λ 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

