



变量选择方法总结

大道至简

作者：冯裕祺

组织：东北大学理学院

时间：Oct 18, 2021

版本：1.0

自定义：信息



秋风萧瑟今又是，换了人间。——毛泽东

特别声明

从研究生入学以来，宽泛地读了很多方向的文章，从机器学习到深度学习，总是觉得自己提不起兴趣。认为这些东西不够**统计**。跟着自己的兴趣，最终选择了变量选择这个自己感兴趣的方向。先将读到的文章做一些总结，来时刻提醒自己。

冯裕祺

Oct 18, 2021

目录

1	变量选择方法回顾	1
1.1	收缩方法	1
1.1.1	lasso 回归	1
1.1.2	岭回归	2
1.1.3	SCAD 方法	2
1.1.4	Oracle 性质	3
1.1.5	最小角回归	3
1.1.6	弹性网模型	4
1.1.7	Group lasso	4
1.1.8	图 LASSO	5
1.1.9	Adaptive lasso	5

第 1 章 变量选择方法回顾

随着科学技术的不断发展，网络通讯速度的飞速提升，各种微型传感器的广泛应用，现在能够获取到的数据的量级和大小是十分恐怖的。简言之，在统计角度来说，我们能够搜集到十分巨大的 X ，但是由于目前计算机算力的限制和算法的局限性，在这种 $n \rightarrow \infty$ 的情况下，许多原有的算法是无法实现的。因此许多统计学家提出了变量选择的方法，简单来说就是从原来的 p 个变量中，通过一些方法选择出 \hat{p} 来代替原来的 p 个变量，从而实现降维的目的，让算法可以顺利地实现。在此，本文提出从统计的角度对相关学者提出的方法进行梳理汇总，来对变量选择这个方向进行整体的归纳。

1.1 收缩方法

1.1.1 lasso 回归

收缩方法 (shrinkage methods) 在回归问题当中，最为常见的是收缩方法。这一类方法通过对回归的目标函数加上正则化项，从而实现了变量选择的目的。其中最先提出的是在 1996 年 Tibshirani^[2]提出的 **lasso** 方法，其表达式具体如下：


定理 1.1 (lasso 方法)

$$\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2$$

subject to $\sum_{j=1}^p |\beta_j| \leq t$

也可以写成拉格朗日形式为：

$$\hat{\beta}^{\text{lasso}} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\} \quad (1.1)$$

 **笔记** 在 lasso 中，其实采用了 l_1 范数的惩罚，可以将 $\lambda \sum_{j=1}^p |\beta_j|$ 视为是惩罚项，正是由于惩罚项的存在，让 lasso 方法可以达到变量选择的目的。

lasso 方法的提出，引起了广泛的引用和探讨，时至今日，lasso 方法仍然有着广泛的应用。lasso 方法为何能够达到变量选择的目的呢，可以通过图 1.1^{lassotu}来看。

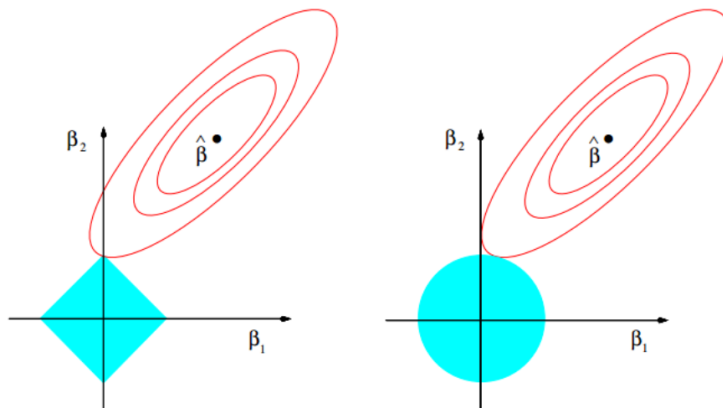


图 1.1: lasso 回归示意图

从上图的左边是 lasso 回归的可行解区域在二元情况下的图示，在 lasso 回归中解的空间是一个矩形，矩形与椭圆线的交点则是 lasso 回归的解。在二元情况下，椭圆线当与 β_2 相交的时候获得一个解，这时 $\beta_1 = 0$ ，从而达到了变量选择的目的。从几何的角度来看，lasso 回归的变量选择的特性相对比较好懂，接下来我们介绍与 lasso 回归类似的 ridge 回归也叫岭回归。

1.1.2 岭回归

岭回归 (ridge regression) 于 1970 年由 Hoerl 和 Kennard^[1]提出，其本质思想还是对回归加上正则化项，不同于 lasso 回归，岭回归的正则化项是 l_2 范数，因此岭回归也有了不同于 lasso 回归的一些性质。^[1]

定理 1.2 (岭回归)

岭回归 (Ridge regression) 根据回归系数的大小加上惩罚因子对它们进行收缩。岭回归的系数使得带惩罚的残差平方和最小：

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (1.2)$$

这里 $\lambda \geq 0$ 是控制收缩程度的参数： λ 越大，收缩的程度越大，每个系数都向 0 收缩。通过参数的平方和来惩罚的想法也用在神经网络，也被称作权重衰减。

需要注意的是，岭回归由于其采用了 l_2 范数的惩罚项，从图 1.1 右图可以看出，岭回归的解的形状是圆形，当椭圆线与其相交时候， $\beta_1 \beta_2$ 都会有值，因此岭回归并不会完全的去掉一些变量，而是通过对变量的重新组合在达到其降维的目的。

1.1.3 SCAD 方法

Fan 和 Li^[3]于 2001 年提出了 SCAD 方法 (Smoothly Clipped Absolute Deviation Penalty)，其思想是将惩罚项拓展为如下形式：

定理 1.3 (SCAD)

SCAD 的目标函数如下：

$$\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n \sum_{j=1}^d p_{\lambda} (|\beta_j|) \quad (1.3)$$

$$p_{\lambda} (|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda) \quad (1.4)$$

Fan 在这篇十分重要的论文中提出了变量选择的三个性质，他认为只有具有这三个优良性质的变量选择方法才是好的方法。

- 无偏性
- 稀疏性
- 有门限

接下来我们将具体说明为什么有这三个性质的变量选择方法才是好的方法。

定理 1.4 (无偏性)

将使用变量选择方法筛选出的系数集合记为 $\hat{\beta}$ ，将真实的系数集合记为 β_{true} ，无偏性是指， $E\hat{\beta} = \beta_{\text{true}}$ ，简单来说就是使用变量筛选方法筛选出的变量是真实变量集合的无偏估计量。

定理 1.5 (稀疏性)

使用变量选择方法筛选出的系数是有门限的, 应该自动的将小的估计系数设置为 0, 从而可以减少模型的复杂性。

定理 1.6 (连续性)

估计出的估计量在样本应该是连续的, 为了避免在预测过程中的不固定性。这里的个人的翻译感觉不是很好, 原文如下: The resulting estimator is continuous in data z to avoid instability in model prediction.

Fan 和 Li 在论文中提出的这三条性质基本奠定了今后变量选择方法的评判指标, 如果一个变量选择方法能够满足这三个性质则说明其是一个好的变量选择方法。这三条性质的提出被大家广泛接受, 并到现在也是十分重要的。

1.1.4 Oracle 性质

在 Fan 和 Li 的经典提出 SCAD 方法的论文中, 还提到了一个十分重要的性质, 就是 **Oracle** 性质, 其具体内容如下。

定义 1.1 (Oracle 性质)

将使用变量筛选方法得到的系数集合记为 \mathcal{A}^* , 将真实的变量的集合记为 \mathcal{A} , **Oracle** 性质则为 $P(\mathcal{A}^* \subseteq \mathcal{A}) \rightarrow 1$

笔记 Oracle 性质十分重要, 基本在之后的变量选择相关论文中, 都需要证明其提出的方法是符合 **Oracle** 性质的。

Oracle 性质向我们展示了变量选择方法的一种神奇的特性, 也正是因为这个神奇的特性, 能够让我们对**变量选择**这个方向有了更深的理论指导。从数理科学的角度来说明我们筛选出来的变量是真实可靠的。

1.1.5 最小角回归

最小角回归 (least angle regression) 由 Efron 等人^[4]提出。类似于向前逐步回归 (Forward Stepwise) 的形式。从解的过程上来看它是 lasso regression 的一种高效解法。可以将 LASSO 回归认为是最小角回归的一个变种。

首先来看前向选择算法 (Forward Selection) 算法。前向选择算法的原理是一种典型的贪心算法, 要解决的问题对 $Y = X\theta$ 这样的线性关系, 如何求解系数向量 θ 的问题, 其中 Y 为 $m \times 1$ 的向量, X 为 $m \times n$ 的矩阵, θ 为 $n \times 1$ 的向量, m 为样本数量, n 为维度特征。

把矩阵 X 看作 n 个 $m \times 1$ 的向量 $X_i (i = 1, 2, \dots, n)$, 在 Y 的 X 变量 $X_i (i = 1, 2, \dots, n)$ 中, 选择和目标 Y 最为接近 (余弦距离最大) 的一个变量 X_k , 用 X_k 来逼近 Y 得到下式: $\bar{Y} = X_k \theta_k$, 其中 $\theta_k = \frac{\langle X_k, Y \rangle}{\|X_k\|_2^2}$, 即 \bar{Y} 是 Y 在 X_k 上的投影。因此定义残差: $Y_{res} = Y - \bar{Y}$ 。由于是投影, 因此 Y_{res} 和 X_k 是正交的。再以 Y_{res} 为新的因变量, 去掉 X_k 后, 剩余的因变量的集合 $X_i, i = 1, 2, 3, \dots, k-1, k+1, \dots, n$ 为新的自变量的集合, 重复投影与残差的操作, 直到残差为 0, 或者所有的自变量都选择完毕, 停止算法。当 X 只有二维时, 如上图所示, 和 Y 最接近的是 X_1 , 首先

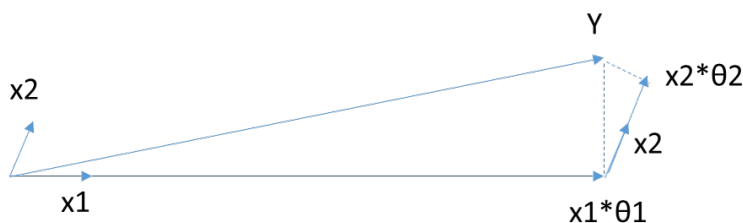


图 1.2: lars 回归示意图

在 X_1 上面投影，残差入上图长虚线。此时 $X_1\theta_1$ 模拟了 Y ， θ_1 模拟了 θ (仅仅模拟了一个维度)。接着发现最接近的是 X_2 ，此时用残差接着在 X_2 投影，残差如图中短虚线，由于没有其他自变量了，此时 $X_1\theta_1 + X_2\theta_2$ 模拟了 Y 对应的模拟了两个维度 θ 即为最终结果。此算法对每个变量只需要执行一次操作，效率高，速度快。但也容易看出，当自变量不是正交的时候，由于每次都是在做投影，所有算法只能给出一个局部近似解。因此，这个简单的算法太粗糙，还不能直接用于我们的 Lasso 回归。

最小角回归在网上有许多教程，这里有比较好的[最小角回归说明](#)。Lasso 回归是在 ridge 回归的基础上发展起来的，如果模型的特征非常多，需要压缩，那么 Lasso 回归是很好的选择。一般的情况下，普通的线性回归模型就够了。

1.1.6 弹性网模型

在 LASSO 和 Ridge 回归提出之后，Zou 等人^[5]于 2005 年提出了弹性网模型 (elastic net)。该方法本质上还是收缩惩罚的思想，其思想是将 l_1 范数和 l_2 范数相结合。

定理 1.7 (弹性网模型)

$$\hat{\beta}(\text{Naive ENet}) = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \quad (1.5)$$

弹性网模型发明的动机：1. 模型的预测准确率和模型的可解释性是回归模型的两个重要的部分。2. LASSO 方法提升了最小二乘估计和岭回归估计，并且他的收敛性和变量选择的性质同时提高了模型的预测能力和模型的可解释性。3. LASSO 方法不能够解决群组效应。如果有一组变量，他们两两之间相关系数很大，LASSO 方法倾向于只从其中选择一个变量。4. 解决群组效应是十分重要的，例如在基因选择问题中。5. 弹性网能够解决群组效应问题。弹性网像一个有弹性的渔网，抓住所有的大鱼。

考虑如下的惩罚回归模型：

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|^2 + \lambda J(\beta)$$

在论文中，作者的 Lemma2 揭示了为什么 naive elastic 方法能够处理群组效应。1. 严格的凸函数保证了能够从群组效应中选择出变量。2. naive elastic 惩罚是严格凸函数，因为有二次项部分存在，LASSO 是凸函数，但因为没有二次项所以不是严格凸函数。3. LASSO 方法没有唯一解，因此不能很好地解决群组效应。

弹性网模型的缺点：1. 通过实证表明，naive elastic 的表现并不是完全令人满意的。因为在其中有两个收敛的过程 (LASSO 和 ridge)，双重的收敛导致了不必要的偏置 (bias)。变形后的 Naive elastic net 有更好的表现。 $\hat{\beta}(\text{ENet}) = (1 + \lambda_2) \cdot \hat{\beta}(\text{NaiveENet})$ 2. 这样做能够提升表现的原因是：1. 撤销了收敛 2. 对正交设计矩阵，LASSO 方法的解是极小极大优化问题，为了让 elastic net 达到同样的极小极大优化，我们需要加上系数。3. $\lambda_2 = 2$, elastic net 方法就是 LASSO。 $\lambda_2 \rightarrow \infty$ 此时等同于单变量软门限 (soft thresholding)。

总结：

1. Elastic net 方法提供了一个模型系数的稀疏解，并且能够解决群组效应。
2. Elastic net 方法的系数计算方法是基于 LARS 方法的。
3. 数据结果和模拟计算证实了 elastic net 方法是优于 LASSO 方法的。
4. 对 Elastic net 方法而言，需要通过训练集和交叉验证的方法来选取两个参数。
5. Elastic net 方法是为回归模型提出的，但是也可以拓展到分类问题。

1.1.7 Group lasso

Group lasso 由 Yuan 和 Lin 在 2006 年提出^[7]。作者在文章中认为 lasso 和 LARS 方法具有很好的性质但是他们是被设计用来选择独立的变量。其定义如下。

定理 1.8 (Group lasso)

For a vector $\eta \in R^d$, $d \geq 1$, and a symmetric $d \times d$ positive definite matrix K , denote:

$$\|\eta\|_K = (\eta' K \eta)^{1/2}$$

We write $\|\eta\| = \|\eta\|_{I_d}$ for brevity. Given positive definite matrices K_1, \dots, K_J the Group lasso estimate is defined as the solution to

$$[\text{Grouplasso}] \frac{1}{2} \left\| Y - \sum_{j=1}^J X_j \beta_j \right\|^2 + \lambda \sum_{j=1}^J \|\beta_j\|_{K_j} \quad (1.6)$$

容易看出, group lasso 是对 lasso 的一种推广, 即将特征分组后的 lasso。显然, 如果每个组的特征个数都是 1, 则 group lasso 就回归到原始的 lasso。该段摘自 csdn 博客, [博客地址](#)。

1.1.8 图 LASSO

图 LASSO 方法于 2006 年由 MEINSHAUSEN 和 BUHLMANN^[6]提出。对这个方法不是很了解, 尤其是图这一块, 具体详见[博客](#)。

1.1.9 Adaptive lasso

Adaptive lasso 于 2006 年由 Zou 提出^[8], 其本质是对 lasso 方法的一种改进提升。在这篇论文当中作者对 Oracle 性质有了更明确的定义。

命题 1.1 (Oracle 性质)

Let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and further assume that $|\mathcal{A}| = p_0 < p$ Thus the true model depends only on a subset of the predictors. Denote by $\hat{\beta}(\delta)$ the coefficient estimator produced by a fitting procedure δ Using the language of Fan and Li^[3] we call δ an **Oracle** procedure if $\hat{\beta}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\{j : \hat{\beta}_j \neq 0\} = \mathcal{A}$
- Has the optimal estimation rate, $\sqrt{n}(\hat{\beta}(\delta)_{\mathcal{A}} - \beta_{\mathcal{A}}^*) \rightarrow_d N(0, \Sigma^*)$, where Σ^* is the covariance matrix knowing the true subset model.

Adaptive lasso 其对传统 lasso 提升最大的是具有了无偏性。传统的 lasso^[2], 虽然能够得到稀疏的解, 但是得到的解并不具有无偏性, 而无偏性是十分重要的性质。具体的实现方法如下。

定理 1.9 (Adaptive lasso)

$$\arg \min_{\beta} \left\| \mathbf{y} - \sum_{j=1}^p \mathbf{x}_j \beta_j \right\|^2 + \lambda \sum_{j=1}^p w_j |\beta_j| \quad (1.7)$$

where ω is a known weights vector. 作者在论文中提出的 ω , β_{OLS} 是对输入的变量进行最小二乘拟合得到的系数, 选择一个合适的正整数 $\gamma > 0$, 定义 $\hat{\omega} = 1/|\hat{\beta}|^\gamma$

从上可以看出 adaptive lasso 是对 lasso 的惩罚项添加了系数 ω , 通过添加系数的方法, Zou 证明了其具有 Oracle 性质。

参考文献

- [1] HOERL A E, KENNARD R W. Ridge Regression: Applications to Nonorthogonal Problems[J]. *Technometrics*, 1970, 12(1): 69-82.
- [2] TIBSHIRANI R. Regression Shrinkage and Selection Via the Lasso[J]. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58(1): 267-288. DOI: [10.1111/j.2517-6161.1996.tb02080.x](#).
- [3] FAN J, LI R. Variable selection via nonconcave penalized likelihood and its oracle properties[J]. *Journal of the American Statistical Association*, 2001, 96(456): 1348-1360. DOI: [10.1198/016214501753382273](#).
- [4] EFRON B, HASTIE T, TIBSHIRANI J R. Least Angle Regression[J]. *Annals of Statistics*, 2004, 32(2): 407-451.
- [5] ZOU H, HASTIE T. Erratum: Regularization and variable selection via the elastic net (*Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2005) 67 (301-320))[J]. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2005, 67(5): 768. DOI: [10.1111/j.1467-9868.2005.00527.x](#).
- [6] MEINSHAUSEN N, BÜHLMANN P. High-dimensional graphs and variable selection with the Lasso[J]. *Annals of Statistics*, 2006, 34(3): 1436-1462. arXiv: [0608017 \[math\]](#). DOI: [10.1214/009053606000000281](#).
- [7] YUAN M, LIN Y. Model selection and estimation in regression with grouped variables[J]. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2006, 68(1): 49-67. DOI: [10.1111/j.1467-9868.2005.00532.x](#).
- [8] ZOU H. The adaptive lasso and its oracle properties[J]. *Journal of the American Statistical Association*, 2006, 101(476): 1418-1429. DOI: [10.1198/016214506000000735](#).