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

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 $n \longrightarrow \infty$ 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)

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

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

也可以写成拉格朗日形式为：

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

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**??**来看。

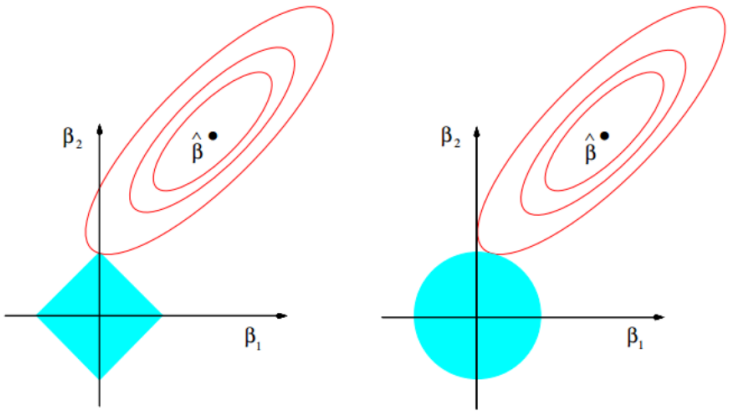

图 1.1: lasso 回归示意图

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:

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

这里 $\lambda >= 0$ 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了神经网络，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除掉一些变量，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:

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

$$p_\lambda\left(|\beta_j|\right) = \lambda^2 - \left(|\beta_j| - \lambda\right)^2 I(|\theta| < \lambda) \tag{1.4}$$

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，$E\hat{\beta} = \beta_{true}$, 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 $P(\mathcal{A}^* \subseteq \mathcal{A}) \to 1$

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[5]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 $X_i (i = 1, 2, ..., n)$, 在 Y 的 X 变量 $X_i (i = 1, 2, ..., n)$ 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：$\overline{Y} = X_k \theta_k$, 其中 $\theta_k = \frac{<X_k, Y>}{\|X_k\|_2}$, 即 $\overline{Y}$ 是 Y 在 X_k 上的投影。因此定义残差：$Y_r es = Y - \overline{Y}$。由于是投影，因此 $Y_r es$ 和 X_k 是正交的。再以 $Y_r es$ 为新的因变量，去掉 X_k 后，剩余的因变量的集合 $X_i, i = 1, 2, 3..., k-1, k+1, ...n$ 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

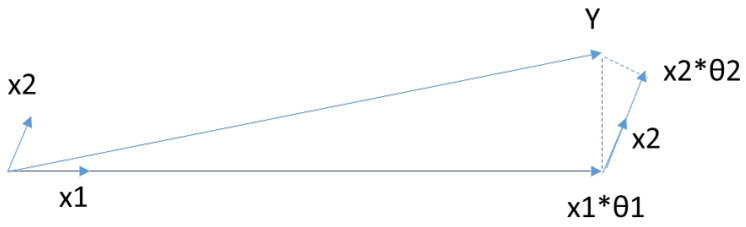

图 1.2: lars 回归示意图

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[6]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)

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

♡

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：

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

在论文中，作者的 Lemma2 揭示了为什么 navie 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。$\hat{\boldsymbol{\beta}}(\text{ENet}) = (1+\lambda_2)\cdot\hat{\boldsymbol{\beta}}(\text{ NaiveENet })$ 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。$\lambda_2 \to \infty$ 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 elasic 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[10]。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 $\eta \in R^d, d \geqslant 1$,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 $K_1, \ldots, K_J$ the Group lasso estimate is defined as the solution to

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

♡

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[7]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[11], 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 $\hat{\boldsymbol{\beta}}(\delta)$ the coefficient estimator produced by a fitting procedure δ Using the language of Fan and Li[3] we call δ an **Oracle** procedure if $\hat{\boldsymbol{\beta}}(\delta)$ (asymptotically) has the following oracle properties:

- Identifies the right subset model, $\left\{j : \hat{\beta}_j \neq 0\right\} = \mathcal{A}$
- Has the optimal estimation rate, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\delta)_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*\right) \rightarrow_d \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$,where $\boldsymbol{\Sigma}^*$ 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)

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

where ω is a known weights vector. 作者在论文中提出的 ω，βOLS 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 $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$

♡

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。

1.1.10 dantzig selector

Dantzig selector 由 CANDES 和 TAO[12]在 2007 年提出。其具体形式如下。

定理 1.10 (dantzig selector)

$$\min_{\tilde{\beta}\in\mathbf{R}^p} \|\tilde{\beta}\|_{\ell_1} \quad \text{subject to } \|X^*r\|_{\ell_\infty} \le \left(1+t^{-1}\right)\sqrt{2\log p}\cdot\sigma$$

其中 r 是残差向量 $y-X\tilde{\beta}$,t 是正标量。如果 X 服从 uniform uncertainty principle, 并且如果真实的参数向量 β 是显著稀疏的，那么 ds(dantzig selector) 方法将有大概率服从如下不等式：

$$\|\hat{\beta}-\beta\|_{\ell_2}^2 \le C^2\cdot 2\log p\cdot\left(\sigma^2+\sum_i \min\left(\beta_i^2,\sigma^2\right)\right) \tag{1.8}$$

♡

笔记 ∞ 范数可以视为取最大值的操作。

这里 ds 方法在文中作者提出的优点是其优化问题的解相对好解，有很高的效率。

1.1.11 Irrepresentable Condition

在传统的变量选择命题中会常常提到 Irrepresentable Condition 这个重要条件。由 Zhao 和 Yu[8]在 2006 年提出。这一条件对证明 lasso 方法有十分重要的作用。

1.1.12 MC_+

MC_+ 方法由 Zhang[17]于 2010 年提出。其提出的方法分为两个环节，一个环节是极小极大化凹函数惩罚 (minimax concave penalty，MCP) 和惩罚线性无偏选择 (penalized linear unbiased selection,PLUS)。MCP 方法具体如下：

定理 1.11 (MCP 方法)

考虑惩罚平方回归：

$$L(\mathbf{b};\lambda)\equiv(2n)^{-1}\|\mathbf{y}-\mathbf{Xb}\|^2+\sum_{j=1}^{p}\rho\left(\left|b_j\right|;\lambda\right)$$

MCP 将惩罚项定义为：$\rho(t;\lambda)=\lambda\int_0^t(1-x/(\gamma\lambda))_+dx$，通过正则化参数 $\gamma>0$。该惩罚项将最小化这个极大凹函数:$\kappa(\rho)\equiv\kappa(\rho;\lambda)\equiv\sup_{0<t_1<t_2}\{\dot{\rho}(t_1;\lambda)-\dot{\rho}(t_2;\lambda)\}/(t_2-t_1)$

♡

1.2 最优子集的 splice 算法

最优子集是一个简单并且相对粗暴的变量选取方法，但是如果直接对数据集使用最优子集法进行计算，其计算复杂度将是 $O(2^p)$, 在 2020 年 Zhu[22]对最优子集法提出了一种 splice 算法，将最优子集法的计算复杂度成功的由 N-Phard 问题降为了计算复杂度为多项式的可解问题。

Splice 方法，这里将介绍 Splice 方法。

定理 1.12 (splice 方法)

$$\mathcal{L}_n(\beta)=\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\beta\|_2^2 \tag{1.9}$$

考虑如下优化问题：

$$\min_{\beta}\mathcal{L}_n(\beta),\quad \text{s.t}\ \|\beta\|_0\le s \tag{1.10}$$

给定任意的初始集合 $\mathcal{A}\subset\mathcal{S}=\{1,2,\ldots,p\}$ 并且 $|\mathcal{A}|=s$，记 $\mathcal{I}=\mathcal{A}^c$ 并计算：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}_{\mathcal{I}}=0}\mathcal{L}_n(\boldsymbol{\beta})$$

将 $\mathcal{A}\mathcal{I}$ 分别记为激活集和非激活集。

- 后退损失：对任意的 $j \in \mathcal{A}$, 将 j 去掉的损失为：

$$\xi_j = \mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{\mathcal{A}\backslash\{j\}}\right) - \mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{\mathcal{A}}\right) = \frac{\boldsymbol{X}_j^\top \boldsymbol{X}_j}{2n}\left(\hat{\beta}_j\right)^2$$

- 前进损失：对任意的 $j \in \mathcal{I}$, 将 j 加入的损失为：

$$\zeta_j = \mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{\mathcal{A}}\right) - \mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{\mathcal{A}} + \hat{\boldsymbol{t}}^{\{j\}}\right) = \frac{\boldsymbol{X}_j^\top \boldsymbol{X}_j}{2n}\left(\frac{\hat{d}_j}{\boldsymbol{X}_j^\top \boldsymbol{X}_j/n}\right)^2$$

其中,$\hat{\boldsymbol{t}} = \arg\min_{\boldsymbol{t}} \mathcal{L}_n\left(\hat{\boldsymbol{\beta}}^{\mathcal{A}} + \boldsymbol{t}^{\{j\}}\right)$,$\hat{d}_j = \boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$，其中对任意向量 t 和任意集合 $\mathcal{A}$,$t^{\mathcal{A}}$ 意味着第 j 个元素 $\left(\boldsymbol{t}^{\mathcal{A}}\right)_j$ 等于 t_j 如果 $j \in \mathcal{A}$, 否则为 0。

从直觉上，一个大的 $\xi_j or \zeta_j$ 意味着第 j 个变量是潜在重要的。但是不幸的是由于这两种损失的支撑集的大小不一样，无法直接比较。然而，如果我们将 $\mathcal{A}$ 中一些不重要的变量和 $\mathcal{I}$ 中一些重要的变量，进行交换，有可能会得到一个比较好的结果。这个思想驱动了切片方法的提出。

对任意的切片大小 $k \leqslant s$, 定义:

$$\mathcal{A}_k = \left\{ j \in \mathcal{A} : \sum_{i \in \mathcal{A}} 1\left(\xi_j \geq \xi_i\right) \leq k \right\} \tag{1.11}$$

代表在 $\mathcal{A}$ 中 k 个最不相关的变量：

$$\mathcal{I}_k = \left\{ j \in \mathcal{I} : \sum_{i \in \mathcal{I}} \mathrm{I}\left(\zeta_j \leq \zeta_i\right) \leq k \right\} \tag{1.12}$$

代表 k 个最相关的变量在 $\mathcal{I}$。
然后我们将 $\mathcal{A}$ 和 $\mathcal{I}$ 切片，通过交换 $\mathcal{A}_{\|}$ 和 $\mathcal{I}_{\|}$ 从而得到了一个新的激活集：

$$\tilde{\mathcal{A}} = (\mathcal{A}\backslash\mathcal{A}_k) \cup \mathcal{I}_k \tag{1.13}$$

让 $\tilde{\mathcal{I}} = \tilde{\mathcal{A}}^c, \tilde{\beta} = \arg\min_{\beta_{\tilde{\mathcal{I}}=0}} \mathcal{L}_n(\beta)$, 并且 $\tau_s > 0$ 成为阈值。如果 $\tau_s < \mathcal{L}_n(\hat{\boldsymbol{\beta}}) - \mathcal{L}_n(\tilde{\boldsymbol{\beta}})$, 这说明 $\tilde{\mathcal{A}}$ 比 $\mathcal{A}$ 相比起来更好。激活集能通过这种方法反复的更新，直到损失函数不能够通过切片方法提升。这里 τ_s 是一个超参数，τ_s 相对是小的，在这篇文章中 $\tau_s = 0.01s\log(p)\log(\log n)/n$。

接下来的问题是如何确定初始集合 $\mathcal{A}$。在文中，作者提出的方法是选取前 s 个与 y 最相关的变量作为初始激活集合 $\mathcal{A}$。将 k_{max} 成为最大切片大小，$k_{max} \leqslant s$。

这里的方法引起了我的思考，是否可以将选取初始的集合 $\mathcal{A}$ 的方法与之后的提出的 distance correlation 和 ball correlation 相结合。相比较原始的衡量相关性的方式，这样也许会进一步提升该方法的表现。

定义 1.2 (定义初始集合)

在文章中，假设 X 和 y 都经过了标准化处理，在文中应该是使用的 Pearson 相关系数来度量相关程度的。

$$\mathcal{A}^0 = \left\{ j : \sum_{i=1}^p \mid \left(\left| \frac{x_j^\top y}{\sqrt{x_j^\top x_j}} \right| \leq \mid \frac{x_i^\top y}{\sqrt{x_i^\top x_i}} \leq s \right\}, \mathcal{I}^0 = \left(\mathcal{A}^0\right)^c \right.$$

另一个想法是，也许能够有有一种更好的方法来进行切片，原文中作者是定义两种损失，我们能不能有一种更加高效的交换方法，来对 $\mathcal{A}$ 和 $\mathcal{I}$ 来进行比较和交换，从而得到更好的结果。

该方法有开源的 python 包 abess，但是其底层是通过 c 来构建的，如何改进 c 也是一个问题，也许通过改进一下处理步骤也可以得到比较好的效果。

1.3 Regularization Paths for Generalized Linear Models via Coordinate Descent

在 2010 年，由 Friedman，Hastie 和 Tibshirani[16]提出了一种针对带有凸惩罚函数 (convex penalties) 的广义线性模型的快速算法。并且在 R 中通过 **glmnet** 包进行实现，从而拓展了 LASSO,Elastic net 和 Ridge regression 等的应用范围，其具体思想在这里省略，如果有兴趣可以具体查阅原文。

1.4 重要不等式

在使用传统方法进行变量选择时，会使用到几个常用并且十分重要的概率不等式。其形式如下所示：

定理 1.13 (重要不等式)

$Y_1, \ldots Y_n$ 是相互独立的随机变量且均值为 0, 记 $S_n = \sum_{i=1}^n Y_i$。

1.**Hoeffding inequality**: 如果 $Y_i \in [a_i, b_i]$, 则有

$$P\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

2.**Berstein' s inequality**, 如果 $E\left|Y_i\right|^m \leq m! M^{m-2} v_i/2$, 则有

$$P\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(v_1 + \cdots + v_n + Mt\right)}\right).$$

3.**Sub-Gaussian** 如果 $E\exp\left(aY_i\right) \leq \exp\left(v_i a^2/2\right)$ 则有

$$P\left(|S_n| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(v_1 + \cdots + v_n\right)}\right)$$

♡

这些不等式的具体推导及应用可以详看 Peter Buhlmann 和 Sara van de Geer[18]的书中第十四章，其中有详细的证明推导过程。这里比较详细的证明过程在维基百科中有，证明连接

1.5 总结

惩罚方法也称作收缩方法是一类经典的变量选择方法，其从提出到后续不断地完善经历了一个完整的过程，并且逐渐成为变量选择的主流思想。随着机器学习和深度学习的不断发展，其思想也在不断地扩散，影响着其他的领域。在机器学习和深度学习都有很深的应用，在这些领域当中，惩罚的思想更多的被称作**正则化**。对于本章提出的这些方法，我们可以从各个角度来进行分析。其中比较详细的可以参考 Hasite,Tibshirani 和 Firedman 的《The Elements Of Statistical Learning》一书[4]。该书中对比较高阶的统计方法都有介绍，并且可以有很好的启发。这里推荐中文版的该书，中文版连接。该作者在对书中的各种定理给出了详细的证明，适合进一步阅读提升。

第 2 章 Screening 方法

随着相关学者对收缩方法的不断研究，Fan 和 Lv[14]在 2008 年提出了 Screening 方法，该方法从另一个角度对变量选择这个命题进行了阐释。其提出的 Screening 方法，在最初是一个看起来十分简单的想法。其具体内容将在稍后给出，并且提出了 **Screening 性质**。这个性质在我看来有点类似于之前提出的 **Oracle 性质**[3]。

2.1 Sure independence screening

该方法的主要目的是将 p(例如 $exp\{O(n^{\epsilon})\}$) 维从一个大规模的维度将低到相对大的维度 d(e.g$o(n)$) 通过一种快速且有效的方法。其主要思想是从相关性来进行的，将那些与因变量弱相关的变量筛选出去。通过这种筛选方法作者提出了一种性质，所有的重要的变量在筛选后存在的概率趋近于 1。

定理 2.1 (screening property)

记 $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_j \neq 0\}$ 是真实的稀疏集合，并且其稀疏程度为 $s = |\mathcal{M}_*|$。其余的 $p - s$ 个变量也能与通过与其他变量的相关从而与因变量产生相关。记 $\omega = (\omega_1, ..., \omega_p)^T$ 是一个 p 维向量由 componentwise regression 得到

$$\omega = X^T y$$

, 其中 X 是 $n \times p$ 列标准话的矩阵，因此 ω 是变量与因变量的相关系数。

对任意给定的 $\gamma \in (0, 1)$, 将 p 个以降序排序从而定义出子模型为：

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all } \} \tag{2.1}$$

其中 $[\gamma n]$ 是 γn 的整数部分。这是一个前向的方式来将全部模型 $\{1, ..., p\}$ 压缩到 $\mathcal{M}_\gamma, d = [\gamma n] < n$

这样的关联学习方法将变量的重要性通过其与因变量的边际相关进行排列，并且过滤出那些与因变量弱相关的变量。从本质上来说，SIS 方法更像是一种两阶段方法，首先使用 SIS 方法进行一个初步的选择，将筛选后的变量再使用别的方法进行一步筛选，如下图所示，改图来自 Fan 和 Lv[14]的原文，从一个直观的角度想我们阐述了 SIS 方法的工作原理。

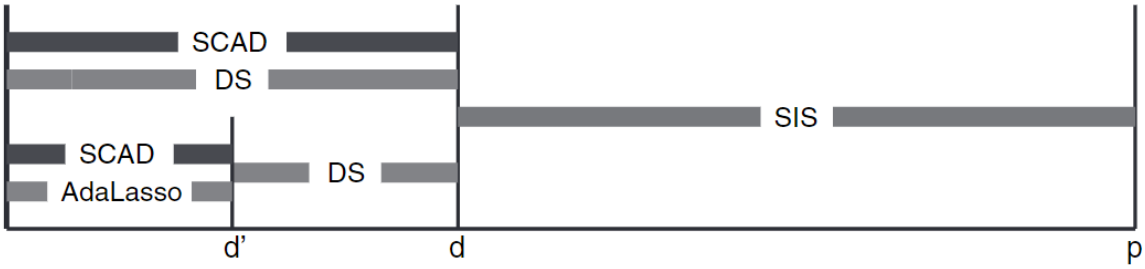

图 2.1: SIS 示意图

假设数据的维度是 p, 样本个数是 n, 并且 $p >> n$，首先使用 SIS 方法将数据的维度从 p 降到 d, 然后在 d 上使用如 SCAD,dantzig selector(DS)，adaptive lasso(adaLasso)，也可以进一步对 d 维变量使用 lasso，dantzig selector 这种硬门限方法来进一步降低维度到 d'。

这样的思想可以让 SIS 方法灵活的应用到超高维变量选择和模型选择当中，能够明显的加快其速度。在原文的第五章中[14]，作者在一些条件下证明了 SIS 方法的一些优良性质。如下：

命题 2.1

在条件 1-4 下, 如果 $2\kappa+\tau<1$, 因此 $\theta<1-2\kappa-\tau$, 当 $\gamma\ cn^{-\theta}, c>0$, 对 $C>0$,

$$P\left(\mathcal{M}_{*}\subset\mathcal{M}_{\gamma}\right)=1-O\left[\exp\left\{-Cn^{1-2\kappa}/\log(n)\right\}\right] \tag{2.2}$$

Screening 方法问世之后，引起了广泛的讨论，并且许多学者在此基础上对其进一步进行改进。

2.2 DC-SIS

在 Fan 等人提出 SIS 方法之后，Li,ZHONG 和 ZHU[19]提出了一种对 SIS 方法改进的其称为 Feature Screening via Distance Correlation Learning 文中简称为 distance correlation sure independence screening(DC-SIS)。其主要思想受到 Szekely，Rizzo 和 Bakirov[13], 还有 Székely 和 Rizzo[15]的启发，使用了一种全新的度量两个随机向量相关性的方法。其优良性是他们提出的距离相关是当且仅当两个随机向量是独立的时候其值为 0。

接下来我们介绍距离相关 (DISTANCE CORRELATION)。

定理 2.2 (DC)

记 $\phi_{\mathbf{u}}(\mathbf{t})$ 和 $\phi_{\mathbf{v}}(\mathbf{s})$ 分别是随机向量 $\mathbf{u}$ 和 $\mathbf{v}$ 的特征函数。记 $\phi_{u,v}(t,s)$ 为 u 和 v 的联合特征函数。他们定义 u 和 v 的距离协方差定义为：

$$\mathrm{d\,cov}^{2}(\mathbf{u},\mathbf{v})=\int_{R^{d_u+d_v}}\left\|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t},\mathbf{s})-\phi_{\mathbf{u}}(\mathbf{t})\phi_{\mathbf{v}}(\mathbf{s})\right\|^{2}w(\mathbf{t},\mathbf{s})d\mathbf{t}d\mathbf{s} \tag{2.3}$$

其中 d_u 和 d_v 是 $\mathbf{u},\mathbf{v}$ 的维度。并且：

$$w(\mathbf{t},\mathbf{s})=\left\{c_{d_u}c_{d_v}\|\mathbf{t}\|_{d_u}^{1+d_u}\|\mathbf{s}\|_{d_v}^{1+d_v}\right\}^{-1} \tag{2.4}$$

其中 $c_d=\pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$。在文章中 $\|\mathrm{a}\|_d$ 表示欧几里得范数对 $\mathbf{a}\in\mathbb{R}^d$ 并且 $\|\phi\|^2=\phi\bar{\phi}$。

$\mathbf{u}$ 和 $\mathbf{v}$ 的距离相关如下：

$$\mathrm{d\,corr}(\mathbf{u},\mathbf{v})=\frac{\mathrm{d\,cov}(\mathbf{u},\mathbf{v})}{\sqrt{\mathrm{d\,cov}(\mathbf{u},\mathbf{u})\mathrm{d\,cov}(\mathbf{v},\mathbf{v})}} \tag{2.5}$$

并且其在论文中提出了皮尔逊相关系数与 DC 的关系如下：

$$\mathrm{dcorr}(U,V)=\left\{\frac{\rho\arcsin(\rho)+\sqrt{1-\rho^2}-\rho\arcsin(\rho/2)-\sqrt{4-\rho^2}+1}{1+\pi/3-\sqrt{3}}\right\}^{1/2} \tag{2.6}$$

其值是对 $|\rho|$ 严格增的。

这里发现一个很好的博客记录了各种方法博客地址。DC-SIS 方法的 R 包是 VariableScreening

2.3 Ball distance

在有关距离的方法提出之后 Pan 等人提出了一种 ball distance[20], 通过提出的这种方法来进行 Screening。其 R 包为 Ball。目前的想法是，是否能够将 splice 的思想与这几种方法相结合，从而提升变量选择的精确度。目前的想法有，splice + distan correlation , splice + ball distance 等，先进行程序模拟试试。

这里作者提出是先使用 Screening 方法进行初步的筛选之后然后使用传统的变量筛选方法进行进一步选择。Pan 提出的 Ball distance 相比 DC-SIS 和传统的 SIS 方法其优点是可以识别出复杂的关系在更少的假定条件之下。其工作主要基于 2018 年提出的 Ball correlation, 其相比传统的 Pearson 相关系数的的优点是：该相关系数在 0 1 之间，**当且仅当两个随机向量是无关的时候其值为 0**。该性质使我们能够利用 BC(Ball correlation) 计算出自变量和因变量的 BC 值，对其排序，从而实现一个 Screening 的过程。

作者进一步提出了 BC-SIS 方法的几个优点

- 有强大的 strong screening consistency property

- 具有鲁棒性
- 可以对复杂的因变量和自变量进行筛选

接下来给出两个随机变量 X 和 Y 之间的 Ball 协方差的定义

定理 2.3 (ball cov)

Ball covariance:

$$\mathrm{BCov}^2(X,Y) = \iint_{U\times V} [\theta - \mu\otimes\nu]^2 \left(\bar{B}_{\zeta_X}(x_1,x_2)\times\bar{B}_{\zeta_Y}(y_1,y_2)\right)\theta(dx_1,dy_1)\,\theta(dx_2,dy_2) \tag{2.7}$$

Ball correlation:

$$\mathrm{BCor}^2(X,Y) = \mathrm{BCov}^2(X,Y)/\sqrt{\mathrm{BCov}^2(X,X)\times\mathrm{BCov}^2(Y,Y)} \tag{2.8}$$

♡

接下来的步骤与传统的 SIS 方法基本一致。

第 3 章 变量选择应用

之前介绍了常见的变量选择方法，其思想不仅在传统的统计建模中获得了广泛使用，并且在新兴的领域如机器学习，深度学习都有很好的拓展和提升。接下来将举例一些将传统变量选择方法与机器学习方法结合的例子。

3.1 A new adaptive weighted imbalanced data classifier via improved support vector machines with high-dimension nature

Qi[21]等人在 2019 年提出一种改进的支持向量机，其主要思想是将 Zou[6]等人提出的弹性网模型的惩罚项加入到支持向量机的目标函数之中。其与 Wang[9]等人提出的 DrSVM 的区别最主要在于，Wang 等人按照弹性网的思想将超平面系数中加入 l_1 和 l_2 惩罚项，而 Qi 等人是按照这一思想，将支持向量机的松弛变量 ξ 也加入 l_1 和 l_2 范数的惩罚，但是从文章中没有看出其理论性质的具体证明，但是效果还是不错。其具体形式如下：

定义 3.1 (elastic net support vector machine)

训练集为 (Y, X) 并有 N 个样本，特征数量为 p,C_1 和 C_2 为调和参数，也是超参数，即可定义为：

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left(\|\omega\|_2^2 + b^2\right) + \frac{C_1}{2}\xi^{\mathrm{T}}\xi + C_2\xi \\ &\text{s.t.} \quad \begin{cases} D(X\omega + eb) - e + \xi \geq \mathbf{0} \\ \xi \geq \mathbf{0} \end{cases} \end{aligned} \tag{3.1}$$

其中：$D = \text{diag}\,(y_1, \ldots, y_N)$，$\xi = (\xi_1, \ldots, \xi_N)^{\mathrm{T}}$ 为松弛变量。e 和 0 为单位向量和 0 向量。

其思想其实和我们在第 1.6 节提出的弹性网模型1.7相同，分别加入 L_1 和 L_2 范数并加入调和系数 C_1 和 C_2，这里的主要区别是将惩罚项的想法加入到了惩罚系数 ξ 中。

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] HASTIE T J, TIBSHIRANI R J, FRIEDMAN J H. The Elements Of Statistical Learning[J]. Elements, 2001, 1.

[5] EFRON B, HASTIE T, TIBSHIRANI J R. Least Angle Regression[J]. Annals of Statistics, 2004, 32(2): 407-451.

[6] 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.

[7] 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.

[8] PENG Z, YU B. On Model Selection Consistency of Lasso[J]. Journal of Machine Learning Research, 2006, 7(12): 2541-2563.

[9] WANG L, ZHU J, ZOU H. The doubly regularized support vector machine[J]. Statistica Sinica, 2006: 589-615.

[10] 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.

[11] 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.

[12] CANDES E, TAO T. The Dantzig selector: Statistical estimation when p is much larger than n[J]. Annals of Statistics, 2007, 35(6): 2313-2351. arXiv: 0506081 [math]. DOI: 10.1214/009053606000001523.

[13] SZÉKELY G, RIZZO M L, BAKIROV N K. Measuring and testing dependence by correlation of distances[J]. Annals of Statistics, 2007, 35(6): 2769-2794.

[14] FAN J, LV J. Sure independence screening for ultrahigh dimensional feature space[J]. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 2008, 70(5): 849-911. arXiv: 0612857 [math]. DOI: 10.1111/j.1467-9868.2008.00674.x.

[15] SZÉKELY G, RIZZO M L. Brownian distance covariance[J]. Annals of Applied Stats, 2009, 3(4): 1236-1265.

[16] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1-22. DOI: 10.18637/jss.v033.i01.

[17] ZHANG C H. Nearly unbiased variable selection under minimax concave penalty[M]. [S.l. : s.n.], 2010: 894-942. DOI: 10.1214/09-AOS729.

[18] BÜHLMANN P, GEER S. Statistics for High-Dimensional Data: Methods, Theory and Applications[M]. [S.l.]: Statistics for High-Dimensional Data: Method, Theory, 2011.

[19] LI R, ZHONG W, ZHU L. Feature screening via distance correlation learning[J]. Journal of the American Statistical Association, 2012, 107(499): 1129-1139. arXiv: 1205.4701. DOI: 10.1080/01621459.2012.695654.

[20] Pan, Wenliang, Tian, et al. BALL DIVERGENCE: NONPARAMETRIC TWO SAMPLE TEST[J]. The Annals of Statistics: An Official Journal of the Institute of Mathematical Statistics, 2018.

[21] QI K, YANG H, HU Q, et al. A new adaptive weighted imbalanced data classifier via improved support vector machines with high-dimension nature[J/OL]. Knowledge-Based Systems, 2019, 185: 104933. https://doi.org/10.1016/j.knosys.2019.104933. DOI: 10.1016/j.knosys.2019.104933.

[22] ZHU J, WEN C, ZHU J, et al. A polynomial algorithm for best-subset selection problem[J]. Proceedings of the National Academy of Sciences of the United States of America, 2021, 117(52): 33117-33123. DOI: 10.1073/PNAS.2014241117.