JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie*

Translated By Shuang Xu, henuxs@foxmail.com

*Correspondence:
hastie@stanford.edu
Department of Statistics, Stanford University, 94305 Stanford, USA
Full list of author information is available at the end of the article

Abstract

We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together.The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

我们提出了弹性网, 这是一个新的正则化和变量选择方法. 无论是真实数据还是数值模拟都表明弹性网比 lasso 的效果好, 同时具有类似的稀疏性质. 而且, 弹性网还有群组效应, 即强烈相关的预测变量趋于同时进入或退出模型. 弹性网在预测变量比观测量很大的时候很有用 ($p \gg n$). 相反, lasso 在 $p \gg n$ 的情形中并不十分适用. 本文还提出了一个很效率的算法来计算弹性网的正则化路径, 称为 LARS-EN. 它很像解决 lasso 的 LARS 算法.

Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

1 Introduction and motivation

我们考虑这个回归模型: 给定 p 个预测变量 $\mathbf{x}_1, \cdots, \mathbf{x}_p$, 响应变量 $\mathbf{y}$

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p. \tag{1}$$

模型拟合的过程就是估计系数向量 $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)$. 比如, 普通最小二乘 (OLS) 估计是由最小化残差平方和得到的. 根据不同的背景, 提升模型质量的标准也不一样. 但是都需要注意一下两点:

(a) 对新数据的预测准确性: 我们很难说一个预测力很差的模型是好模型;

(b) 模型的解释力: 学者们偏于选择一个简单的模型, 因为他们更关注于响应变量和协变量之间的关系在预测变量很多时, 选择一个简洁的模型尤其重要.

众所周知 OLS 在以上两个方面的表现都很差. 人们提出了很多罚方法来改善 OLS. 比如, 岭回归 (Hoerl and Kennard, 1988) 的系数在满足其 L_2 范数小于某个数时, 使残差平方和最小. 作为一个连续的收缩方法, 岭回归通过" 偏差-方差权衡" 实现了更好的预测效果. 然而, 岭回归不能产生一个简单的模型, 因为所有考虑的预测变量都被保持在模型中. 相反, Breiman(1996) 提出的子集选择可以产生稀疏模型, 但是由于它是离散的方法, 所以模型的方差极高.

Tibshirani(1996) 提出了一个有效的方法, 称为 lasso. 它对系数的 L_1 范数进行惩罚的 L_1 罚最小二乘. 由于 L_1 惩罚的性质, lasso 可以在实现连续收缩的同时, 自动进行变量选择. Tibshirani (1996) 和 Fu (1998) 对比了 lasso 回归, 岭回归和桥回归 (bridge regression; Frank and Friedman, 1993), 发现三者在不同的情况下各有优劣.

尽管 lasso 在许多情况都表现良好, 但是它有很多限制. 可以考虑下面 3 种情况:

(a) 当 $p > n$ 时, 因为凸优化的性质, lasso 最多选择 n 个变量. 这似乎限制了变量选择方法. 而且, lasso 不是意义明确的, 除非系数的 L_1 范数小于某个值.

(b) 如果有一组变量, 它们两两的相关性很高. 那么 lasso 趋向于选取其中的一个变量, 但是它并不关心选取的是哪一个. 见 Section 2.3.

(c) 对常见的 $n > p$ 的问题. 如果预测变量之间有高度相关性, 经验表明岭回归的预测准确性高于 lasso(Tibshirani, 1996).

情况 (a) 和 (b) 说明 lasso 在某些时候并不十分合适. 我们通过基因微阵列的基因选择问题来论述我们的观点. 一个典型的微阵列数据集有成千上万的预测变量 (基因) 并且样本少于 100 个. 具有相同生物通道 (pathway) 的基因可能相关性很高 (Segal and Conklin, 2003). 我们认为这些基因形成了一个群组. 一个理想的基因选择方法应该具有以下能力: 消除不重要的基因; 一旦群组中的一个基因进入模型, 则整个群组所有基因自动进入 (" 群组选择"). 对 $p \gg n$ 和群组变量的问题, lasso 不是理想的方法, 因为它最多只能选择 p 个变量中的 n 个 (Efron et al., 2004), 并且缺乏揭露群组信息的功能. 至于预测力, 情况 (c) 在回归问题中更是常见. 所以 lasso 方法还有改善的空间.

我们的目标是寻找一个新的方法, 在任何时候都比 lasso 的表现好. 尤其是在上面提到的 3 中情况下, 它既能实现群组选择又具备较高的预测力.

本文提出的新正则化方法称为弹性网 (*elastic net*). 和 lasso 类似, 弹性网在自动选择变量的同时也能实现连续收缩, 并选择相关性的变量群组. 它就像一个可伸缩的渔网, 可以保留所有的大鱼. 数值模拟和真实数据表明弹性网在预测力上优于 lasso.

在 Section 2中我们定义了朴素弹性网 (*naïve elastic net*), 一个使用了弹性网惩罚的罚最小二乘. 我们讨论了群组效应是来自于弹性网惩罚. 在 Section 3中, 研究发现朴素弹性网趋于过度收缩, 接着我们提出了弹性网来修正这个问题. 一个高效的算法 LARS-EN 可以快速计算弹性网的正则化路径, 它的计算量相当于一个 OLS 估计. 在 Section 4给出了前列腺癌的例子, Section 5用数值模拟的例子来比较 lasso 和弹性网. Section 6用弹性网对白血病基因微阵列进行分类和基因选择.

2 Naïve elastic net

2.1 Definition

假设数据集有 n 条观测和 p 个预测变量. 令 $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1 | \cdots | \mathbf{x}_p)$, 其中列向量 $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})^T, j = 1, \cdots, p$ 表示第 j 个变量的 n 条观测. 经过中心化和标准化处理, 我们假设响应变量有零均值, 预测变量是标准的:

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} x_{ij}^2 = 1 \ (j = 1, 2, \cdots, p). \tag{2}$$

对任意非负的 λ_1 和 λ_2, 我们定义朴素弹性网准则 (罚函数)

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1, \tag{3}$$

其中 $|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2, |\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$. 回归系数 $\hat{\boldsymbol{\beta}}$ 的朴素弹性网估计对应 eq. (3) 的最小值

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\}. \tag{4}$$

这个过程可以看为一个罚最小二乘方法. 令 $\alpha = \lambda_2(\lambda_1 + \lambda_2)$; 然后 $\hat{\boldsymbol{\beta}}$ 在 eq. (3) 中解等价于优化问题

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta}|^2, \quad \text{subject to } (1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \le t \text{ for some } t. \tag{5}$$

我们称 $(1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2$ 为弹性网惩罚, 是 lasso 惩罚和岭惩罚的凸组合. 当 $\alpha = 1$ 时, 朴素弹性网退化为岭回归, 本文仅考虑 $\alpha < 1$ 的情况. 对所有的 $\alpha \in [0, 1)$, 弹性网惩罚函数的奇点是 0(没有一阶导数), 并且 $\alpha > 0$ 时是严格凸的. 所以弹性网继承了 lasso 和岭回归的特征. 注意, lasso 惩罚 (即 $\alpha = 0$ 的弹性网惩罚) 是凸的, 但不是严格凸的. 在 Fig. 1中可以清晰的看出来.

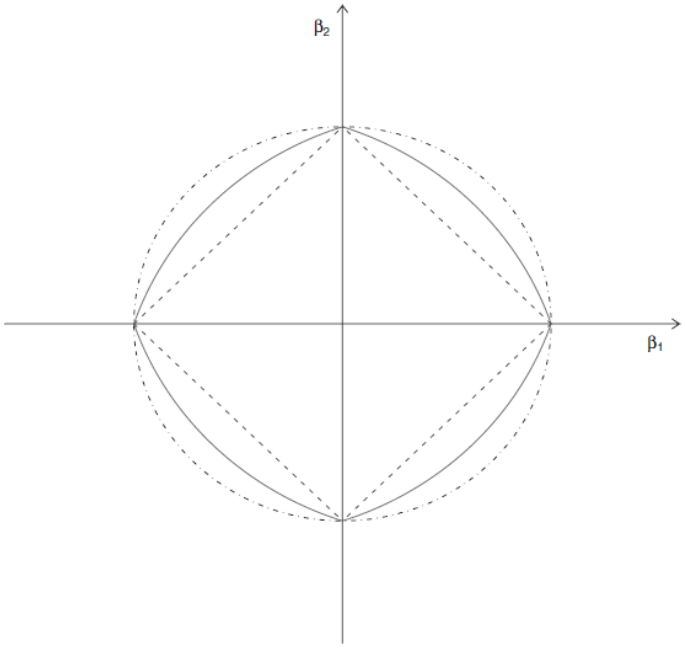

Figure 1 三个惩罚函数的等高线图. · — · — ·— 岭惩罚, - - - - - -lasso 惩罚; ——弹性网惩罚 ($\alpha = 0.5$). 我们可以在节点处看出 lasso 惩罚的一阶不可导性质, 等高线的边可以看出弹性网和岭回归的严格凸性; 凸性的强弱与 α 有关.

2.2 Solution

本节提出解决朴素弹性网的高效算法. 可以证明最小化 eq.(3) 等价于 lasso 型的优化问题. 所以朴素弹性网也具有 lasso 的计算便利性.

Lemma 2.1 给定数据集 $(\mathbf{y},\mathbf{X})$ 和 (λ_1,λ_2), 定义 $(\mathbf{y}^*,\mathbf{X}^*)$ 为

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}, \mathbf{y}^*_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}.$$

令 $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ 和 $\boldsymbol{\beta}^* = \sqrt{(1+\lambda_2)}\boldsymbol{\beta}$. 则朴素弹性网惩罚函数可以写成

$$L(\gamma,\boldsymbol{\beta}) = L(\gamma,\boldsymbol{\beta}^*) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1.$$

令 $\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} L\{(\gamma,\boldsymbol{\beta}^*)\}$; 则 $\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\boldsymbol{\beta}}^*$

这个的证明仅仅是简单的代数方法, 这里不再赘述. Lemma 2.1表明通过把原始数据扩展为增广数据, 朴素弹性网问题可以转变为 lasso 问题. 注意增广数据有 $n+p$ 条观测, 矩阵 $\mathbf{X}^*$ 的秩为 p, 所以朴素弹性网的所有解会潜在地选择所有 p 个预测变量. 这个性质克服了我们在情况 (a) 中所描述的 lasso 的缺点. Lemma 2.1也说明朴素弹性网可以像 lasso 一样自动实现变量选择. 下节中我们论述朴素弹性网”群组选择”的性质 (lasso 并不具备这个性质).

在正交设计中, 我们可以给出朴素弹性网的解

$$\hat{\beta}_i(\text{naïve elastic net}) = \frac{(|\hat{\beta}_i(\text{OLS})| - \lambda_1/2)_+}{1+\lambda_2}\text{sgn}\{\hat{\beta}_i(\text{OLS})\}, \tag{6}$$

其中 $\hat{\boldsymbol{\beta}}(\text{OLS}) = \mathbf{X}^T\mathbf{y}$; z_+ 表示其正数部分, 即如果 $z>0, z_+ = z$, 否则 $z_+ = 0$. 岭回归的解为 $\hat{\boldsymbol{\beta}}(\text{ridge}) = \hat{\boldsymbol{\beta}}(\text{OLS})/(1+\lambda_2)$, lasso 的解为

$$\hat{\beta}_i(\text{lasso}) = (|\hat{\beta}_i(\text{OLS})| - \lambda_1/2)_+\text{sgn}\{\hat{\beta}_i(\text{OLS})\}.$$

Fig 2展示了三种罚方法在正交设计中的特点, 其中朴素弹性网可以视为两阶段收缩: 在 lasso 的临界点之内, 该系数收缩为零; 否则, 使用岭收缩.

2.3 The grouping effect

在”大 p 小 n”的问题中 (West et al., 2001), 变量的群组性质尤其需要考虑. 比如, 主成分分析 (PCA) 已经用于寻找相关性高的基因 (Hastie et al., 2000; Díaz-Uriarte, 2003). 还有树的方法通过监督式学习选择出由层次聚类发现的基因 (Hastie et al., 2003). Segal and Conklin (2003) 使用正则化过程选择出了成组的基因. 我们考虑这类罚方法

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta}|^2 + \lambda J(\boldsymbol{\beta}), \tag{7}$$

其中 $J(\cdot)$ 关于 $\boldsymbol{\beta}\neq\mathbf{0}$ 为非负函数.

定性地讲, 如果一组高度相关的变量的绝对值趋于相等, 则这个回归方法会展现出群组效应. 特别在极端情况下, 一些变量就是完全相同的, 则回归方法应该赋予它们相等的系数.

Lemma 2.2 假设 $\mathbf{x}_i = \mathbf{x}_j,\ i,j\in\{1,\cdots,p\}$.

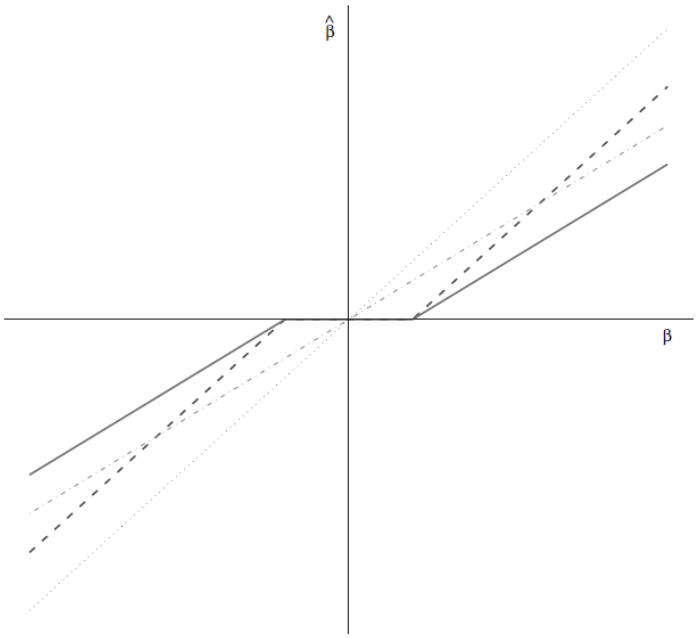

Figure 2 正交设计中 4 种估计的精确解. 岭回归 (· – · – ·–), lasso(- - - - -), 弹性网 (——), OLS(· · · · · ·). 收缩系数 $\lambda_1 = 2; \lambda_2 = 1$

(a) 如果 $J(\cdot)$ 是严格凸的, 则 $\hat{\beta}_i = \hat{\beta}_j, \forall \lambda > 0$.
(b) 如果 $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$, 则 $\hat{\beta}_i \hat{\beta}_j \geq 0$, 且 *eq.(7)* 可以写成

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j, \end{cases}$$

其中 $s \in [0, 1]$.

Lemma 2.2展示了严格凸惩罚函数和 lasso 惩罚函数之间的区别. 严格凸性保证了群组效应. 相反, lasso 甚至不能有唯一解. $\lambda_2 > 0$ 的弹性网惩罚是严格凸的, 所以具有群组选择的性质.

Theorem 2.3 给定数据 $(\mathbf{y}, \mathbf{X})$ 和参数 (λ_1, λ_2), 响应变量 $\mathbf{y}$ 是中心化的, 预测变量 $\mathbf{X}$ 是标准化的. 令 $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ 为朴素弹性网估计. 假设 $\hat{\boldsymbol{\beta}}_i(\lambda_1, \lambda_2)\hat{\boldsymbol{\beta}}_j(\lambda_1, \lambda_2) > 0$. 定义

$$D_{\lambda_1, \lambda_2}(i, j) = \frac{1}{|\mathbf{y}|_1} |\hat{\boldsymbol{\beta}}_i(\lambda_1, \lambda_2) - \hat{\boldsymbol{\beta}}_j(\lambda_1, \lambda_2) > 0|;$$

则

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - p)},$$

其中 $\rho = \mathbf{x}_i^T \mathbf{x}_j$ 表示样本相关系数.

$D_{\lambda_1,\lambda_2}(i,j)$ 是个无单位的量, 描述了变量 i 和 j 的系数路径的差别. 如果 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 是高度相关的, 例如 $\rho \approx 1$(如果 $\rho \approx -1$, 则考虑 $-\mathbf{x}_j$), Theorem 2.3说 i,j 之间的系数路径差别几乎为 0. 上面不等式的上限提供了朴素弹性网群组效应的一个定量描述.

Lasso 没有群组效应. Section 1中的情况 (b) 在实际中经常发生. 一个理论的解释在 Efron et al. (2004) 中给出. 简单地讲, 我们考虑一个 $p = 2$ 的线性模型. Tibshirani (1996) 给出了 $(\hat{\beta}_1, \hat{\beta}_2)$ 的精确表达式, 我们易得 $|\hat{\beta}_1 - \hat{\beta}_2| = |cos(\theta)|$, 其中 θ 是 $\mathbf{y}$ 和 $\mathbf{x}_1 - \mathbf{x}_2$. 我们很容易构建一组数据, 满足 $\rho = \text{corr}(\mathbf{x}_1, \mathbf{x}_2) \to 1$ 但 $cos(\theta)$ 的值不等于 0.

2.4 Bayesian connections and the L_q-penalty

桥回归 (bridge regression;Frank and Friedman, 1993 & Fu, 1998) 把 eq.(7) 中的惩罚函数变为 $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_p^p = \sum_{j=1}^p |\beta_j|^q$ 即可. 这是岭回归 ($q = 2$) 和 lasso($q = 1$) 的广义形式. 桥估计可以看为一种贝叶斯模式, 即我们已经知道了系数服从的分布. 其先验概率函数为

$$p_{\lambda,q}(\boldsymbol{\beta}) = C(\lambda, q) \exp(-\lambda |\boldsymbol{\beta}|_q^q). \tag{8}$$

岭回归 ($q = 2$) 对应了高斯先验概率, lasso($q = 1$) 对应了拉普拉斯 (或称双指数) 先验概率. 弹性网惩罚对了一种新的先验概率

$$p_{\lambda,q}(\boldsymbol{\beta}) = C(\lambda, q) \exp\left\{-\lambda \left[\alpha |\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1\right]\right\}, \tag{9}$$

是高斯概率和拉普拉斯概率的折中. 尽管当 $1 < q < 2$ 的桥估计与弹性网有很多相似之处, 但是它们之间有本质的区别. 弹性网产生稀疏解, 然而桥估计并不能. Fan and Li (2001) 证明了在 $L_q(q \geq 1)$ 罚函数族中只有等于 1(lasso) 能够产生稀疏解. 桥估计 ($1 < q \leq 2$) 估计会保留所有的变量在模型中. 因为本文最初目的是通过罚方法自动实现变量选择, 所以虽然 $L_q(1 < q < 2)$ 图像与弹性网很像, 但我们并不对其研究.

3 Elastic net

3.1 Deficiency of the naïve elastic net

作为一种自动变量选择方法, 朴素弹性网克服了 lasso 在情况 (a) 和 (b) 中的限制. 然而, 经验表明朴素弹性网的表现经常不如人意, 除非它很接近岭或 lasso(见 Section 4& 5). 这就是我们称它”朴素”的原因.

考虑到模型的预测力, 一个罚方法的参数一般需要通过偏差-方法权衡来获取最佳值. 朴素弹性网估计是一个两阶段过程: 先固定 λ_2 寻找岭回归系数, 然后我们沿着 lasso 系数的路径进行 lasso 型收缩. 这经常会导致收缩量太多. 对比单纯的 lasso 或岭, 这两次收缩不会帮助我们减少很多模型方差, 反而会增加不必要的模型偏差. 下节我们对这两次收缩进行修正来提升朴素弹性网的预测力.

3.2 The elastic net estimate

我们继续 Section 2.2的记号. 给定数据 (**y**,**X**), 罚参数 (λ_1, λ_2) 和增广数据 $(\mathbf{y}^*, \mathbf{X}^*)$, 则朴素弹性网可以转变为 lasso 型问题

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} |\mathbf{y}^* - \mathbf{X}^* \boldsymbol{\beta}^*|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}} |\boldsymbol{\beta}^*|_1. \tag{10}$$

修正的弹性网估计定义如下

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = \sqrt{(1+\lambda_2)}\hat{\boldsymbol{\beta}}^*. \tag{11}$$

而 $\hat{\boldsymbol{\beta}}(\text{naïve elastic net}) = \{1/\sqrt{1+\lambda_2}\}\,\hat{\boldsymbol{\beta}}^*$; 因此

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = (1+\lambda_2)\hat{\boldsymbol{\beta}}(\text{naïve elastic net}). \tag{12}$$

所以弹性网系数是重新调节的朴素弹性网系数.

这个变换保持了朴素弹性网的变量选择性质, 而且是减少收缩的最简单的方式. 所以弹性网继承了 Section 2中论述的朴素弹性网的所有优良性质. 我们还发现与 lasso 和岭相比, 弹性网往往表现更好.

上面我们使用 $1+\lambda_2$ 作为调节因子还有一个原因. 正交设计中, 预测变量都是正交的, 此时 lasso 估计是最优解 (Donoho et al., 1995), 这意味着朴素弹性网不是最优的. 经过因子 $1+\lambda_2$ 的调节, 弹性网自动成为最优解了.

实际上, 选取 $1+\lambda_2$ 作为调节因子的动机来自于对岭回归的作用矩阵的分解. 我们知道岭回归的系数估计为 $\hat{\boldsymbol{\beta}}(\text{ridge}) = (\mathbf{X}^T\mathbf{X}+\lambda_2\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. 记作用在响应变量的矩阵为 $\mathbf{R} = (\mathbf{X}^T\mathbf{X}+\lambda_2\mathbf{I})^{-1}\mathbf{X}^T$. 其中由于预测变量是标准化的, 所以 $\mathbf{X}^T\mathbf{X}$ 就是预测变量的协方差矩阵, 即 $\mathbf{X}^T\mathbf{X}_{ij} = \rho_{ij}$, 对角线都为 1.

$$\mathbf{R} = \frac{1}{1+\lambda_2}\mathbf{R}^* = \frac{1}{1+\lambda_2}\begin{pmatrix} 1 & \frac{\rho_{12}}{1+\lambda_2} & \cdots & \cdots & \frac{\rho_{1p}}{1+\lambda_2} \\ & 1 & \cdots & \cdots & \frac{\rho_{2p}}{1+\lambda_2} \\ & & \ddots & & \vdots \\ & & & 1 & \frac{\rho_{p-1,p}}{1+\lambda_2} \\ & & & & 1 \end{pmatrix}^{-1} \mathbf{X}^T \tag{13}$$

$\mathbf{R}^*$ 与 OLS 的作用矩阵类似, 但是矩阵被因子 $1/1+\lambda_2$ 收缩, 我们称之为去相关 (即降低相关性). 所以从 eq.(13) 中我们可以知道岭回归能够解决多重共线性的原因: 岭回归使用 $1/(1+\lambda_2)$ 对变量进行收缩, 才能实现去相关.

当我们糅合岭回归和 lasso 时, 岭的去相关步骤是多余的. 因为尽管岭收缩 $1/(1+\lambda_2)$ 可以有效控制模型方差, 但是 lasso 收缩既可以控制方差又能得到稀疏解. 为了避免双重收缩, 我们对朴素弹性网乘上因子 $(1+\lambda_2)$ 以消除岭收缩的不良影响.

从此开始我们令 $\hat{\boldsymbol{\beta}}$ 代表 $\hat{\boldsymbol{\beta}}$(elastic net). 下面的定理给出了朴素弹性网的另外一种表述, 这里去相关的步骤更明显.

Theorem 3.1 给定数据 $(\boldsymbol{y},\boldsymbol{X})$ 和参数 (λ_1,λ_2), 则弹性网估计如下

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\left(\frac{\mathbf{X}^T\mathbf{X}+\lambda_2\mathbf{I}}{1+\lambda_2}\right)\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda_1|\boldsymbol{\beta}|_1. \tag{14}$$

易见 *lasso* 估计为

$$\hat{\boldsymbol{\beta}}(\text{lasso}) = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda_1|\boldsymbol{\beta}|_1. \tag{15}$$

Theorem 3.1说明了弹性网是稳定版本的 lasso. 注意相关性矩阵 Σ 的样本相关矩阵为 $\hat{\Sigma} = \mathbf{X}^T\mathbf{X}$, 则

$$\frac{\mathbf{X}^T\mathbf{X}+\lambda_2\mathbf{I}}{1+lambda_2} = (1-\gamma)\hat{Sigma} + \gamma\mathbf{I}.$$

这里 $\gamma = \lambda_2/(1+\lambda_2)$ 对 $\hat{\Sigma}$ 朝着单位阵的方向进行收缩. 对朴素弹性网重新调节等价于用收缩的样本相关矩阵代替样本相关矩阵. 在线性判别分析中, 这个做法可以提升预测准确性 (Friedman, 1989; Hastie et al., 2001).

3.3 Connections with univariate soft thresholding

Lasso 是弹性网的一个特例 ($\lambda_2 = 0$). 另外一个有趣的特例是 $\lambda_2 \to \infty$. 由 Theorem 3.1, 当 $\lambda_2 \to \infty$ 时, $\hat{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}}(\infty)$, 这里

$$\hat{\boldsymbol{\beta}}(\infty) = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^T\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda_1|\boldsymbol{\beta}|_1.$$

$\hat{\boldsymbol{\beta}}(\infty)$ 有一个简单的近似形式

$$\hat{\beta}(\infty)_i = \left(|\mathbf{y}^T\mathbf{x}_i| - \frac{\lambda_1}{2}\right)_+ \operatorname{sgn}(\mathbf{y}^T\mathbf{x}_i), \quad i = 1, 2, \cdots, p. \tag{16}$$

注意到 $\mathbf{y}^T\mathbf{x}_i$ 是第 i 个变量的一元回归系数, $\hat{\boldsymbol{\beta}}(\infty)$ 就是把软门限应用到一元回归系数. Eq.(16) 称为一元软门限 (univariate soft thresholding, UST).

UST 忽略了预测变量之间的相关性, 在处理时当做独立变量. 尽管这种处理不太合适, 但是 UST 应用到了许多方法之中, 比如基因微阵列的显著性分析 (Tusher et al., 2001), 最近邻收缩分类器 (Tibshirani et al., 2002). 弹性网是联系 lasso 和 UST 的桥梁.

3.4 Computation: the algorithm LARS-EN

我们提出计算弹性网的高效算法 LARS-EN, 这个算法基于 LARS(Efron et al. 2004). 他们证明了: 从零开始, lasso 的系数路径以可以预测的分段线性方式变化. 他们提出的 LARS 算法与 OLS 的计算复杂度是等阶的. 通过 Lemma 2.1, 对于一个固定的 λ_2, 弹性网就等价于增广数据的 lasso 问题. 所以 LARS 可以直接用于计算弹性网的解路径, 且计算量与 OLS 相当. 但是对于 $p \gg n$ 问题, 增广数据集有 $p+n$ 条观测和 p 个变量, 这可能影响计算效率.

我们利用 $\mathbf{X}^*$ 的结构稀疏性质来简化计算量, 这对 $p \gg n$ 问题很重要. 具体而言, 如 Efron et al. (2004) 所讲, 在第 k 步, 我们要对矩阵 $\mathbf{G}_{A_k} = \mathbf{X}_{A_k}^{*T}\mathbf{X}_{A_k}^*$, 其中 A_k 是进入方程的变量集合. 这个计算通过对 $\mathbf{G}_{A_{k-1}}$ 的 Cholesky 分解可以获得. 注意, 对任意的集合 A

$$\mathbf{G}_A = \frac{1}{1+\lambda_2}(\mathbf{X}_A^T\mathbf{X}_A + \lambda_2\mathbf{I})$$

所以这相当于 $\mathbf{X}_{A_{k-1}}^T\mathbf{X}_{A_{k-1}} + \lambda_2\mathbf{I}$ 的 Cholesky 分解的迭代过程. 可以证明我们可以直接使用 $\mathbf{X}_{A_{k-1}}^T\mathbf{X}_{A_{k-1}}$ 进行迭代 (Golub and Van Loan, 1983). 所以我们没有必要精确地使用 $\mathbf{X}^*$ 计算 LARS 算法. 简化的方式可以有效得到系数的路径.

算法 LARS-EN 逐步更新弹性网拟合. 在 $p \gg n$ 问题中, 比如基因微阵列, 我们没必要把这个算法进行到底. 真实数据和数值模拟表明在 LARS-EN 的早期阶段就能得到最优解. 如果我们在第 m 步停止算法, 则只需要 $O(m^3 + pm^2)$ 次操作.

3.5 Choice of tuning parameters

本节讨论如何选取弹性网中的调节参数 (tuning parameter). 尽管我们能够得到系数路径, 但是我们并不知道选取哪一对 (λ_1, λ_2) 是最好的. 在 lasso 中通常使用系数的 L_1 范数 t, 或 lasso 与 OLS 系数的范数之比 s 作为调节参数. 相对应地, 我们可以使用 (λ_2, s) 或 (λ_2, t). 前者的优点在于 $s \in [0, 1]$. LARS 是一个向前逐步可加拟合过程, 与 $\varepsilon - L_2$ *boosting* 类似 (Efron et al., 2004). 这个角度下, 算法的步数 k 也可以看成一种调节参数. 类似地, 固定 λ_2 时 LARS-EN 进行的步数 k 也可以作为 λ_2 后的第二参数. 上面三种调节参数对应了三种对 Fig. 3的解释角度.

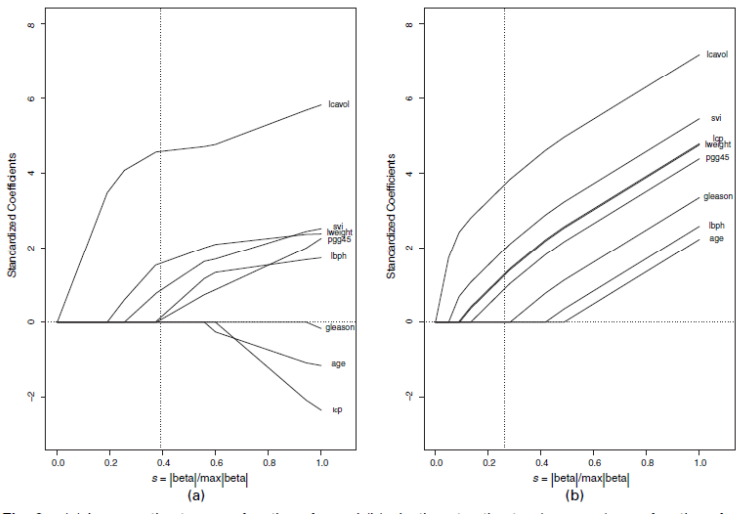

Figure 3 (a)Lasso 的系数路径 (b) 弹性网的系数路径: 两者都是分段线性的, 这是 LARS 族算法的特征; 解路径也表明了弹性网相等于一元软门限 (: 最终的模型)

人们已经对调节参数的最佳选择建立了一套比较完善的方法 (Hastie et al. (2001), chapter 7). 如果我们只有训练集可用, 则 10-折交叉验证是一个比较受欢迎的方法. 注意我们有两个参数, 所以我们是对一个二维曲面进行交叉验证. 在这个平面上, 我们对 λ_2 取相对较小的一组网格点. 对每一个 λ_2, 可以用 10-折 CV 选择出最优的 λ_2(或 s, k). 最后再选择最小 CV 误差对应的 λ_2.

对每一个 λ_2, 10 折-CV 的计算量相当于 10 次 OLS 估计. 所以 2 维 CV 在一般的 $n > p$ 问题中比较节省计算量. 在 $p \gg n$ 问题中, 计算量随着 p 以线性增长, 仍然容易控制, 因为在实际中, 在早期算法就可以停止了. 比如, $n = 30,\ p = 5000$, 我们只希望不超过 200 个变量进入模型, 我们可以在 500 步后停止算法. 在 500 步之内选择最优的步数.

下文中, λ_2 不再使用下标.

4 Prostate cancer example

数据来自前列腺癌的研究 (Stamey et al., 1989). 预测变量有 8 个指标: log(cancer volume)(lcavol), log(prostate weight)(lweight), age, log(benign prostatic hyperplasia amount)(lbph), seminal vesicle invasion(svi), log(capsular penetration)(lcp), Gleason score(gleason), percentage Gleason scores 4 or 5(pgg45). 响应变量是 logarithm of prostate-specific antigen (lpsa).

使用 OLS/ridge/lasso/naïve elastic net/elastic net 进行回归. 数据分为两部分: 67 条观测的训练集和 30 条观测的测试集. 我们使用测试集的预测均方误差作为评价标准.

Table 1 说明弹性网无论在稀疏性还是在预测力上都是最优的. 朴素弹性网和岭在本例中相似, 没有实现变量选择. 弹性网比 lasso 的预测误差低了 24%. 另外这里的弹性网相当于 UST. 因为 λ 非常大 (1000). Fig. 3是它们的系数路径.

它们预测变量的相关系数矩阵表明有相当的相关性, 最高是 0.76. 换言之, lasso 失败的原因在于变量的相关性. 我们猜测只要岭回归能改善 OLS, 弹性网就能改善 lasso. 下节我们会使用数值模拟证明该论点.

Table 1. Prostate cancer data: comparing different methods

Method	*Parameter(s)*	*Test mean-squared error*	*Variables selected*
OLS		0.586 (0.184)	All
Ridge regression	$\lambda=1$	0.566 (0.188)	All
Lasso	$s=0.39$	0.499 (0.161)	(1,2,4,5,8)
Naïve elastic net	$\lambda=1, s=1$	0.566 (0.188)	All
Elastic net	$\lambda=1000, s=0.26$	0.381 (0.105)	(1,2,5,6,8)

5 A simulation study

本节论证弹性网无论在预测力还是在变量选择上都优于 lasso. 模拟的数据来源于模型

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\varepsilon, \quad \varepsilon \sim N(0,1).$$

下面有 4 个例子. 前 3 个例子与论文 Tibshirani(1996) 一致. 第四个例子的数据出现了群组变量.

在每一个例子中, 我们的数据由训练集和与其独立的验证集、测试集. 我们对测试集计算测试误差 (均方误差). 并使用 $\cdot/\cdot/\cdot$ 来描述训练集/验证集/测试集的样本数量. 下面是 4 个例子的具体信息.

(a) 例子 1, 有 50 组数据, 每组数据有 20/20/200 的观测量. 令 $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $\sigma = 3$. 变量之间两两相关系数为 $\text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$.

(b) 例子 2, 与上相同, 但 $\beta_j = 0.85, \forall j$.

(c) 例子 3, 有 50 组数据, 每组数据有 100/100/400 的观测量. 令

$$\boldsymbol{\beta} = (\underbrace{0,\cdots,0}_{10}, \underbrace{2,\cdots,2}_{10}, \underbrace{0,\cdots,0}_{10}, \underbrace{2,\cdots,2}_{10})$$

$\sigma = 15, \text{corr}(\mathbf{x}_i, \mathbf{x}_j) = 0.5$

(d) 例子 4, 有 50 组数据, 每组数据有 50/50/400 的观测量. 令

$$\boldsymbol{\beta} = (\underbrace{3,\cdots,3}_{15}, \underbrace{0,\cdots,0}_{25})$$

$\sigma = 15$. 预测变量由如下模型生成

$$\begin{aligned}
\mathbf{x}_i &= Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1,\ldots,5,\\
\mathbf{x}_i &= Z_2 + \varepsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6,\ldots,10,\\
\mathbf{x}_i &= Z_3 + \varepsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11,\ldots,15,\\
\mathbf{x}_i &\sim N(0,1), \quad \mathbf{x}_i \text{ independent identically distributed}, \quad i = 16,\ldots,40,
\end{aligned}$$

ε_i^x $i.i.d \sim N(0, 0.01), i = 1, \cdots, 15$. 在这些模型中, 我们有 3 组同等重要的变量, 每个组内有 5 个. 后 25 个变量为纯噪声. 一个理想的模型应该仅选择前 15 个, 噪声的系数为零.

Table2 和 Fig. 4为结果. 第一, 我们发现朴素弹性网要么表现很差 (1), 要么和岭回归表现相同 (2/3), 或 lasso(4). 在所有的例子中, 弹性网显著地比 lasso 准确, 甚至当 lasso 表现比岭回归好的时候. 预测准确性分别高了 18%、18%、13%、27%. 这说明共线性问题中弹性网由于 lasso.

Table3 是弹性网的稀疏解. 与 lasso 相比, 弹性网趋于选择更多的变量. 这是群组效应的结果. 特别在例子 4 中, 弹性网的表现简直超神.

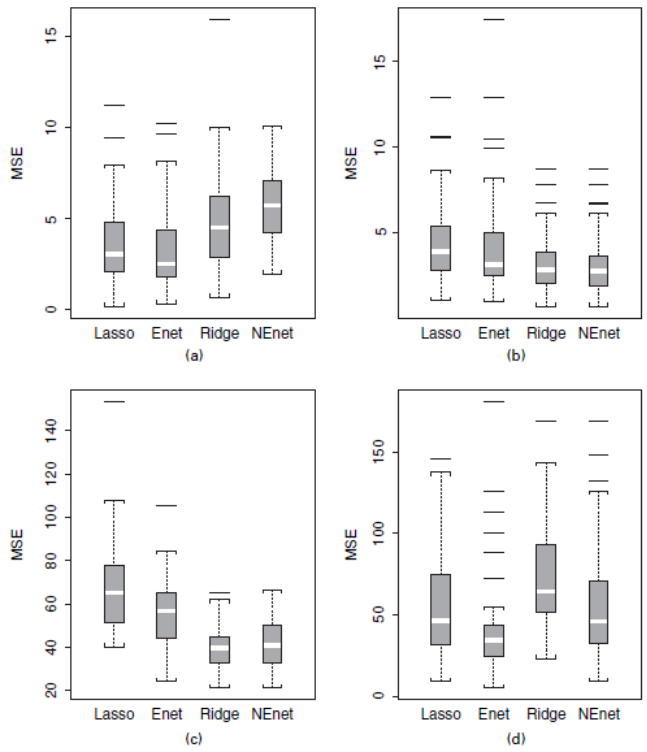

Figure 4 对比四种方法.(a-d) 依次为例子 1-4.

Table 2. Median mean-squared errors for the simulated examples and four methods based on 50 replications†

Method	*Results for the following examples:*			
	Example 1	*Example 2*	*Example 3*	*Example 4*
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
Ridge regression	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Naïve elastic net	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

†The numbers in parentheses are the corresponding standard errors (of the medians) estimated by using the bootstrap with $B = 500$ resamplings on the 50 mean-squared errors.

Table 3. Median number of non-zero coefficients

Method	*Results for the following examples:*			
	Example 1	*Example 2*	*Example 3*	*Example 4*
Lasso	5	6	24	11
Elastic net	6	7	27	16

下面是一个理想的例子, 表现了弹性网和 lasso 的重要差别. 令 $Z_i(i = 1, 2)$ $i.i.d$ $U(0, 20)$. 响应变量 $\mathbf{y}$ 产生于 $N(Z_1 + 0.1Z_2, 1)$. 假设预测变量

$$\mathbf{x}_1 = Z_1 + \varepsilon_1, \quad \mathbf{x}_2 = -Z_1 + \varepsilon_2, \quad \mathbf{x}_3 = Z_1 + \varepsilon_3,$$
$$\mathbf{x}_4 = Z_2 + \varepsilon_4, \quad \mathbf{x}_5 = -Z_2 + \varepsilon_5, \quad \mathbf{x}_6 = Z_2 + \varepsilon_6,$$

其中 ε_i $i.i.d$ $N(0, 1/16)$. 产生 100 条观测. 变量 $\mathbf{x}_i(i = 1, 2, 3)$ 为一个组, 其余为一个组. 组内的相关系数几乎为 1, 组间相关系数几乎为 0. Fig. 5为这个例子的 lasso 和弹性网系数路径.

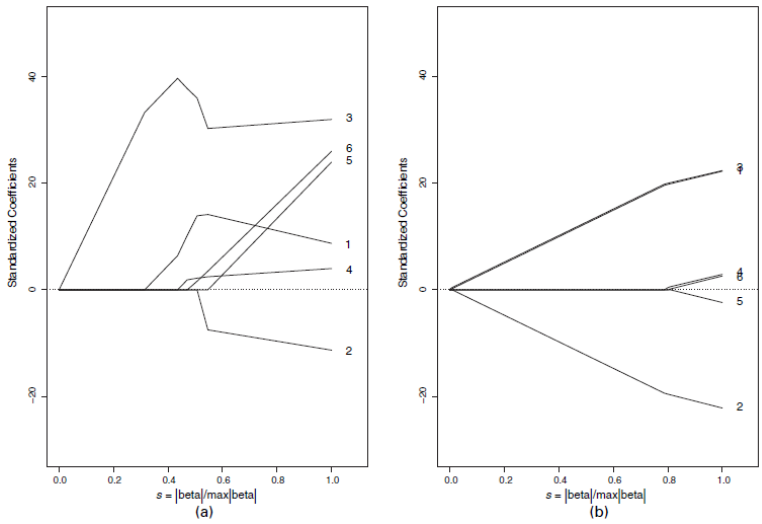

Figure 5 (a)Lasso 和 (b) 弹性网 ($\lambda_2 = 0.5$) 的解路径: lasso 的解路径不稳定, 且路径并没有揭露变量相关性的信息; 相反, 弹性网更加平滑, 并且明显展示出了群组效应. 去相关产生了群组效应和稳定的解路径.

6 Microarray classification and gene selection

典型的基因微阵列具有成千上万的基因并少于 100 条观测. 因为这种数据的特殊性, 我们认为一种好的分类方法应该具有如下性质:

(a) 这个过程本身就能进行基因选择.

(b) 不被 $p \gg n$ 限制.

(c) 具有相同生物通道的基因应该能被自动选择出来.

虽然很多分类方法具有很低的错判率, 但是可能不满足上面的性质. Lasso 擅长 (a) 但不满足 (b,c). 支持向量机 (SVM; Guyon et al., 2002) 和罚逻辑回归 (Zhu and Hastie, 2004) 虽然是很有效地分类器, 但在自动选择基因方面有所欠缺.

弹性网可以克服这些困难, 我们用白血病数据证明. 这个数据集有 7129 个基因和 72 条观测. 训练集有 38 个样本, 来自两种类型, 其中 27 个来自于 acute lymphoblastic

leukaemia, 另外 11 个来自于 acute myeloid leukaemia. 我们的目标是基于这 7219 个基因的表达水平预测病人属于哪个类型. 剩余的 34 个为测试集, 用来测试预测准确性. 响应变量用 0-1 编码代表两种类型. 分类函数为示性函数 $I(\text{fitted value} > 0.5)$. 使用 10-折 CV 选择调节参数.

在 200 步之后, 我们停止算法. 如 Fig. 6所示, 算法的步数为调节参数, 最终的弹性网分类器 ($\lambda = 0.01, k = 82$) 的 10-折 CV 误差为 3/38, 测试误差为 0/34. Fig. 7为系数路径. Table 4 给出了弹性网/Golub/SVM/罚逻辑回归/最近邻收缩中心的比较. 弹性网最优.

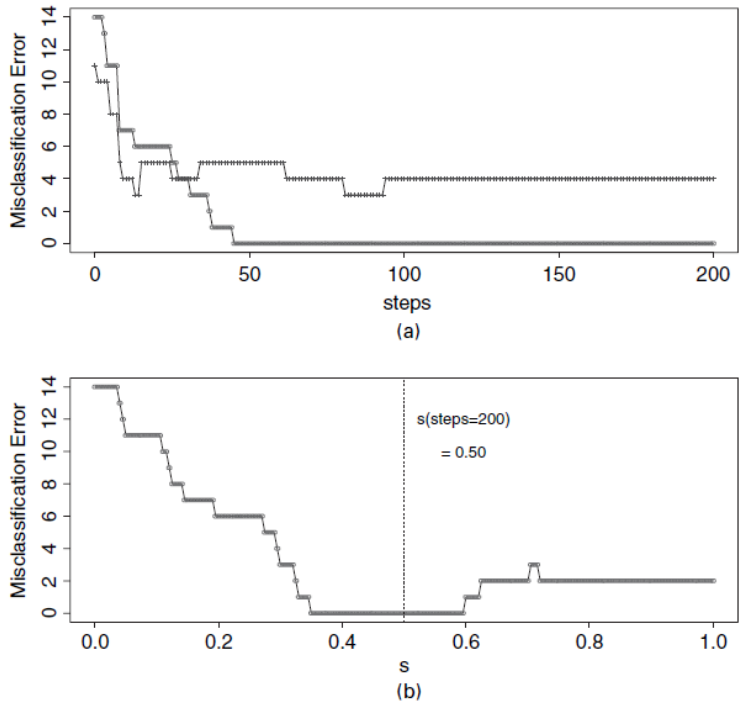

Figure 6 白血病分类的例子 ($\lambda = 0.01$): (a) 在 200 步时停止, (b) 弹性网的完整路径

Table 4. Summary of the leukaemia classification results

Method	*Tenfold CV error*	*Test error*	*Number of genes*
Golub	3/38	4/34	50
Support vector maching–recursive feature elimination	2/38	1/34	31
Penalized logistic regression–recursive feature elimination	2/38	1/34	26
Nearest shrunken centroids	2/38	2/34	21
Elastic net	3/38	0/34	45

7 Discussion

本文提出的弹性网是一种新的收缩和变量选择方法. 弹性网会产生一个预测力强的稀疏模型, 并具有群组效应. 真实数据和数值模拟证明了弹性网表现良好, 优于 lasso. 尤其在一个 2 分类问题中, 弹性网表现出了很低的错判率和自动选择基因.

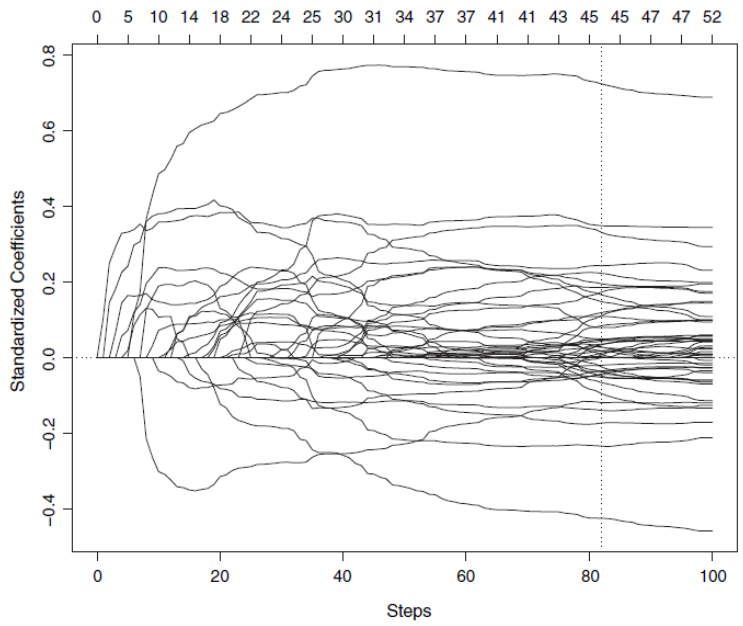

Figure 7 白血病数据: 坐标系上方的数字代表对应模型包含的变量数, 最优模型在第 82 步给出, 有 45 个变量进入; 注意训练集有 38 条观测, 所以 lasso 最多选择 38 个基因; 相反, 弹性网选了多于 38 个基因, 不受样本量的限制; $\lambda = 0.01$ 由 10-折 CV 决定; 如果 λ 变大, 则群组效应变大.

尽管我们的方法提出动机来自回归问题, 但是选取合适的损失函数, 弹性网惩罚可以用于分类问题. 比如, Fig. 6和 boosting 很类似: 测试误差一直下降, 然后有一部分是平缓的, 最后稍微上升 (Hastie et al., 2001). 这不是巧合. 实际上弹性网和极大边际解释 (maximum margin explanation; Rosset et al., 2004) 有密切的关系. 在接下来的论文中, 我们会论述弹性网惩罚在分类中的应用.

可以把弹性网视为 lasso 的扩展, 是一个很有价值的工具. 最近 lasso 被用于解释 boosting: 虽然没有明确地使用 lasso 罚函数但是 boosting 在高维数据的表现与其相似. 我们的结果为 lasso 提供了另外一种视角, 和一种改善它的方式.