正则化稀疏模型

刘建伟[1)] 崔立鹏[1)] 刘泽宇[2)] 罗雄麟[1)]

[1)](中国石油大学(北京)自动化研究所 北京 102249)

[2)](中国科学院软件研究所基础软件国家工程研究中心 北京 100190)

摘 要 正则化稀疏模型在机器学习和图像处理等领域发挥着越来越重要的作用,它具有变量选择功能,可以解决建模中的过拟合等问题. Tibshirani 提出的 Lasso 使得正则化稀疏模型真正开始流行. 文中总结了各种正则化稀疏模型,指出了各个稀疏模型被提出的原因、所具有的优点、适宜解决的问题及其模型的具体形式. 最后,文中还指出了正则化稀疏模型未来的研究方向.

关键词 正则化;稀疏;变量选择;套索;无偏估计;组稀疏;融合套索

中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2015.01307

Survey on the Regularized Sparse Models

LIU Jian-Wei[1)] CUI Li-Peng[1)] LIU Ze-Yu[2)] LUO Xiong-Lin[1)]

[1)](*Research Institute of Automation, China University of Petroleum, Beijing* 102249)

[2)](*National Engineering Research Center for Fundamental Software, Institute of Software, Chinese Academy of Sciences, Beijing* 100190)

Abstract The regularized sparse models are playing a more and more important role in many areas, such as in the machine learning and image processing. The regularized sparse models have the ability of variable selection, so they can solve the over-fitting problem. The Lasso proposed by Tibshirani makes regularized sparse models become popular. This paper summarizes various regularized sparse models and points out the motivation of every regularized sparse model, the advantages of every regularized sparse model, the problems that every regularized sparse model can solve and the form of every regularized sparse model. In the end, we offer the regularized sparse models' research directions in the future.

Keywords regularization; sparse; variable selection; Lasso; unbiased estimation; group sparsity; fused Lasso

1 引 言

机器学习和生物信息学中常常遇到高维小样本数据,高维小样本数据的变量空间维数很高而样本空间维数却很低,因此往往会为建模带来一系列的问题,例如,过少的训练样本会导致过拟合问题,模型的泛化能力很差;训练样本数小于变量数会导致建模过程需要求解病态的欠定方程组,进而导致模型的解不唯一,而且模型的解不连续依赖于定解条件,变量的值发生很小的变化,模型函数的值都会发生很大的变化. 稀疏模型通过降低变量空间维数可有效解决上述问题. 稀疏模型将大量的冗余变量去除,只保留与响应变量最相关的解释变量,简化模型的同时却保留数据集中最重要的信息,有效地解决了高维数据集建模中的诸多问题. 稀疏模型具有更

收稿日期:2014-06-20;最终修改稿收到日期:2014-12-02. 本课题得到国家自然科学基金(21006127)资助. **刘建伟**,男,1966 年生,博士,副研究员,主要研究方向为稀疏学习、智能信息处理、复杂系统的分析、预测与控制、算法分析与设计. E-mail: liujw@cup.edu.cn. **崔立鹏**,男,1990 年生,硕士研究生,主要研究方向为模型稀疏化机器学习、自动控制系统分析与设计、智能信息处理. **刘泽宇**,男,1993 年生,硕士研究生,主要研究方向为机器学习. **罗雄麟**,男,1963 年生,博士,教授,主要研究领域为智能控制、复杂系统分析、预测与控制.

好的解释性,便于数据可视化、减少计算量和传输存储.1996 年 Tibshirani[1]把岭回归估计的 L_2 范数罚正则化项替换为 L_1 范数罚正则化项得到了 Lasso (Least Absolute Shrinkage and Selection Operator, Lasso).L_1 范数罚具有产生稀疏模型的能力,使用 L_1 范数罚作为正则化项的 Lasso 具有变量选择功能和变量空间降维功能.实际上在 Lasso 之前已有能够产生稀疏解的非负绞刑估计(nonnegative garrote estimator)[2]和桥回归(bridge regression)模型[3]被提出,但由于缺少高效的求解算法因而没有引起足够的重视,而直到 Lasso 这种正则化稀疏模型以及可对其有效求解的 LAR 算法(Least Angle Regression, LAR)[4]被提出后,正则化稀疏模型才得到了广泛深入的研究,并在机器学习、数理统计和生物信息学等领域逐渐流行起来.正则化稀疏模型的分类图如图 1 所示.

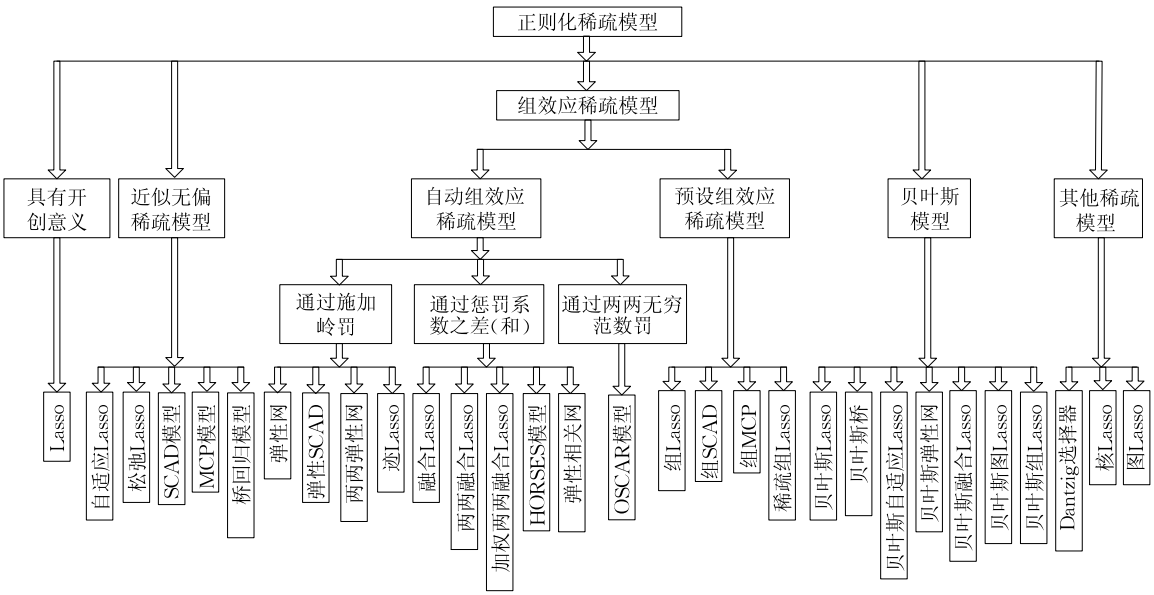

图 1 正则化稀疏模型分类图

本文第 2 节介绍各种正则化稀疏模型,其中首先介绍 Lasso 及其起源,然后介绍近似无偏稀疏模型、自动组效应稀疏模型、预设组效应稀疏模型以及其他一些稀疏模型;第 3 节介绍正则化稀疏模型对应的贝叶斯模型;第 4 节介绍正则化稀疏模型的常见求解算法及其软件包;第 5 节通过一些实验直观地展示几种主要稀疏模型的变量选择效果;第 6 节介绍了稀疏模型在实际中的一些应用;第 7 节指出了正则化稀疏模型未来的发展方向;第 8 节对全文进行总结.

2 正则化稀疏模型

2.1 Lasso 及其起源

已知线性回归模型

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{1}$$

其中 $\boldsymbol{y}\in\boldsymbol{R}^N$ 叫做响应向量,$\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$ 叫做设计矩阵,$\boldsymbol{\beta}\in\boldsymbol{R}^P$ 叫做回归系数向量,$\boldsymbol{\varepsilon}\in\boldsymbol{R}^N$ 叫做误差向量且全部误差变量独立同分布 $\varepsilon_n\sim N(0,\sigma^2)$,$n\in\{1,2,\cdots,N\}$.在 Lasso 被提出之前,最佳子集选择法和逐步回归方法是统计学上传统的变量选择方法,但是这些变量选择方法都是离散的过程,所谓离散的过程指的是初始选择某一个变量或全部变量,然后由某种选择标准依次添加或删除变量的个数,直至选择标准的函数值不再改变为止,最终得到的变量子集参与到模型的拟合过程中.逐步回归法这种离散的变量选择过程抗干扰能力差,数据中微小的扰动都有可能导致完全不同的变量选择结果,并且其变量选择过程与参数估计过程是相互独立的.另外,基于式(1)中线性回归模型的岭回归(ridge regression)是一种典型的正则化模型

$$\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2 \tag{2}$$

岭回归估计对回归系数向量会进行了一定程度的压缩,但并不能将其压缩为零,因此不能产生稀疏解.Breiman 提出的基于式(1)中线性回归模型的非负绞刑估计(non-negative garrote estimator)[2,5-6]为

$$\hat{\boldsymbol{c}}=\arg\min_{\boldsymbol{c}\in\boldsymbol{R}^P}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{B}\boldsymbol{c}\|_2^2+\lambda\|\boldsymbol{c}\|_2^2\right\} \tag{3a}$$

$$\text{s.t.}\quad \forall p,\ c_p\geq 0 \tag{3b}$$

其中 $\boldsymbol{B}=\text{diag}(\hat{\beta}_1^{\text{OLS}},\cdots,\hat{\beta}_P^{\text{OLS}})$,其中 $\hat{\beta}_p^{\text{OLS}}$ 为普通最小

二乘估计的解.非负绞刑估计的解比最佳子集选择法和逐步回归方法稳定且是稀疏的,它的提出比 Lasso 还早.另外,在 Lasso 被提出之前还有一种叫做桥回归(bridge regression)[3,7-9]的稀疏模型被提出:

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\cdot\|\boldsymbol{\beta}\|_1^{\gamma}\right\} \tag{4}$$

其中 $\gamma\in(0,+\infty)$. Lasso 是受到上述子集选择法、岭回归估计、非负绞刑估计和桥回归模型等的启发而被提出的,其通过对回归系数的绝对值之和(即回归系数向量的 L_1 范数)进行惩罚来压缩回归系数的大小,使绝对值较小的回归系数自动被压缩为 0,从而产生稀疏解和实现变量选择,Tibshirani 提出的基于式(1)中线性回归模型的 Lasso 为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\cdot\|\boldsymbol{\beta}\|_1 \tag{5}$$

其中 $\lambda\geqslant 0$ 为可调参数(tuning parameter),$\|\cdot\|_2$ 表示 L_2 范数,$\|\boldsymbol{\beta}\|_1=\sum_{p=1}^{P}|\beta_p|$ 为 L_1 范数罚.与最小二乘法和岭回归不同,Lasso 不具有显式解,其解可通过坐标下降算法(coordinate descent algrithm)[10]进行不断迭代求得,求解 Lasso 的坐标下降算法关于回归系数 β_p 的更新公式为

$$\begin{aligned}\hat{\beta}_p&=\operatorname{sign}(\hat{\beta}_p^{\mathrm{OLS}})(|\hat{\beta}_p^{\mathrm{OLS}}|-\lambda)_+\\&=\begin{cases}\hat{\beta}_p^{\mathrm{OLS}}-\lambda, & \hat{\beta}_p^{\mathrm{OLS}}>\lambda\\ 0, & -\lambda\leqslant\hat{\beta}_p^{\mathrm{OLS}}\leqslant\lambda\\ \hat{\beta}_p^{\mathrm{OLS}}+\lambda, & \hat{\beta}_p^{\mathrm{OLS}}<-\lambda\end{cases}\end{aligned} \tag{6}$$

其中 $\hat{\beta}_p^{\mathrm{OLS}}$ 为普通最小二乘估计(ordinary least square)的解,$\operatorname{sign}(\cdot)$ 为符号函数,$(|\hat{\beta}_p^{\mathrm{OLS}}|-\lambda)_+$ 表示当 $|\hat{\beta}_p^{\mathrm{OLS}}|-\lambda>0$ 时取 $|\hat{\beta}_p^{\mathrm{OLS}}|-\lambda$,否则取 0.坐标下降算法每次只更新一个回归系数,同时固定其他回归系数不变.整个坐标下降算法关于全部回归系数 $\{1,\cdots,P\}$ 循环迭代上述更新过程直到收敛.由式(6)可以看出,落在 $[-\lambda,\lambda]$ 区间内的回归系数向量分量均被置零 0,因此实现了回归系数向量的稀疏化,而回归系数向量的稀疏化使得与回归系数向量分量为零所对应的变量不参与模型的拟合,实现了变量选择效果.实际上,Lasso 是桥回归模型在 $q=1$ 时的一个特例.Lasso 有着诸多优点,其能够同时实现变量选择和参数估计,并且对应的优化问题是凸优化问题.由于 Lasso 这种正则化稀疏模型具有上述诸多优点,所以逐渐流行起来,被广泛应用到机器学习、图像处理、生物信息学和信号处理等多个领域.

2.2　近似无偏稀疏模型

在变量选择中,将与响应变量密切相关的变量叫做目标变量(即期望选择出来的变量),而将其他与响应变量无关的变量叫做噪声变量(即冗余变量).由于 Lasso 对回归系数向量的全部分量都进行相同程度的惩罚,因此除了将噪声变量对应的回归系数压缩为 0 外,还会对目标变量对应的回归系数进行一定程度的压缩,导致了对目标变量回归系数的有偏估计.针对 Lasso 的有偏估计这个缺点,一系列的近似无偏稀疏模型陆续被提出.所谓近似无偏的稀疏模型指的是与 Lasso 相比减弱甚至消除了对于目标变量回归系数压缩的稀疏模型,这些近似无偏的稀疏模型有:自适应 Lasso[11-12]、松弛 Lasso[13]、SCAD 模型[14-15]、MCP 模型[16]和桥回归模型[3,7-10],其中自适应 Lasso 和松弛 Lasso 是凸的,而 SCAD、MCP 和桥回归是非凸的.

(1)自适应 Lasso(adaptive Lasso).自适应 Lasso[11-12]的提出是为了克服 Lasso 估计有偏的缺点.若要令 Lasso 具有估计无偏的特性,一个自然的想法就是令各个变量对应的回归系数所受到的惩罚程度具有自适应的性质,将罚函数中目标变量的回归系数的惩罚权重设定地比较大,而将罚函数中噪声变量的回归系数的惩罚权重设定地比较小,这就是自适应 Lasso 的思想.基于式(1)中线性回归模型的自适应 Lasso 为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{p=1}^{P}\hat{w}_p|\beta_p| \tag{7}$$

其中 $\lambda\geqslant 0$,$\hat{w}_p=1/|\hat{\beta}_p^{\mathrm{OLS}}|^{\gamma}$ 为第 p 个回归系数的权值,$\gamma>0$,$\hat{\beta}_p^{\mathrm{OLS}}$ 为普通最小二乘估计的解.可以看出,自适应 Lasso 稀疏模型构造过程分为两步:① 先做一次普通最小二乘估计,将得到的系数估计值的绝对值的 γ 次方的倒数作为第 p 个变量对应的权值;② 将权值代入式(7)中,求解式(7).自适应组 Lasso 克服了 Lasso 估计有偏的缺点.

(2)松弛 Lasso(relaxed Lasso).另外一种针对 Lasso 有偏估计而被提出的稀疏模型为松弛 Lasso[13],其思想为将变量选择过程和参数估计过程分为两步:第 1 步利用 Lasso 进行变量选择步骤得到变量子集,第 2 步在第 1 步中得到的变量子集的基础上进行参数估计,但第 2 步中的可调参数不像第 1 步中的可调参数那样大,因此减小了对目标变量的回归系数的压缩,但仍有可能将某些冗余变量的回归系数置零.令 M_λ 表示由求解 Lasso 得到的回归系数集:

$$M_\lambda=\{p\mid\hat{\beta}_p\neq 0,p\in\{1,\cdots,P\}\} \tag{8}$$

则基于式(1)中线性回归模型的松弛 Lasso[31]为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|y-X\{\boldsymbol{\beta}\cdot\mathbf{1}_{M_\lambda}\}\|+\phi\lambda\|\boldsymbol{\beta}\|_1 \tag{9}$$

其中 $\lambda\geq 0$，$\phi\in(0,1]$，$p\in\{1,\cdots,P\}$，$\mathbf{1}_{M_\lambda}$ 为关于变量集 $M_\lambda\subseteq\{1,\cdots,P\}$ 的指示函数，即

$$\{\boldsymbol{\beta}\cdot\mathbf{1}_{M_\lambda}\}_p=\begin{cases}0, & p\notin M_\lambda\\ \beta_p, & p\in M_\lambda\end{cases} \tag{10}$$

当 $\phi=1$ 时，松弛 Lasso 就退化为 Lasso. 松弛 Lasso 对回归系数向量进行了两次压缩过程，因此比 Lasso 的解更加稀疏；但松弛 Lasso 与 Lasso 相比减小了对目标变量对应的回归系数的压缩程度，从而提高了 Lasso 的预测准确性.

(3) SCAD(Smoothly Clipped Absolute Deviation penalty). SCAD 模型[14]与也为一种近似无偏稀疏模型，其具有如下的形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{p=1}^{P}\varphi_{\lambda,\gamma}(\beta_p) \tag{11}$$

其中 $p\in\{1,\cdots,P\}$，$\varphi_{\lambda,\gamma}(\cdot)$ 为 SCAD 罚

$$\varphi_{\lambda,\gamma}(\theta)=\begin{cases}\lambda|\theta|, & 0\leq|\theta|\leq\lambda\\ -\dfrac{(|\theta|^2-2\gamma\lambda|\theta|+\lambda^2)}{2(\gamma-1)}, & \lambda<|\theta|<\gamma\lambda\\ \dfrac{(\gamma+1)\lambda^2}{2}, & |\theta|\geq\gamma\lambda\end{cases} \tag{12}$$

其中 $\gamma>2$，$\lambda\geq 0$. SCAD 模型对绝对值小于 λ 的回归系数(即噪声变量的回归系数)的压缩作用与 Lasso 相同，倾向于将这一部分回归系数压缩为零；对于绝对值位于区间$[\lambda,\gamma\lambda]$内的回归系数(即目标变量的回归系数)，随着回归系数绝对值的增大而减小压缩的程度；对于绝对值大于 $\gamma\lambda$ 的回归系数(即目标变量的回归系数)，不再对其进行压缩. 由于减小甚至避免了对目标变量对应的回归系数的压缩，故 SCAD 模型克服了 Lasso 有偏估计的缺点，改善了其参数估计一致性和变量选择一致性. 另外，文献[15]中将损失函数由最小二乘损失函数替换为铰链损失函数，但基于铰链损失函数的 SCAD 与基于最小二乘损失函数的 SCAD 的变量选择效果相同，因为其变量选择效果取决于 SCAD 罚而与损失函数无关.

(4) MCP(Minimax Concave Penalty). MCP 模型[16]也是一种近似无偏稀疏模型，其形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{p=1}^{P}\varphi_{\lambda,\gamma}(\beta_p) \tag{13}$$

其中 $\varphi_{\lambda,\gamma}(\cdot)$ 为 MCP 罚

$$\varphi_{\lambda,\gamma}(\theta)=\begin{cases}\lambda|\theta|-\dfrac{|\theta|^2}{2\gamma}, & |\theta|\leq\gamma\lambda\\ \dfrac{1}{2}\gamma\lambda^2, & |\theta|>\gamma\lambda\end{cases} \tag{14}$$

$\lambda\geq 0$，$\gamma>1$，$I(\cdot)$ 为指示函数. 当 $\gamma\to\infty$ 时，MCP 罚逐渐趋向于 L_1 范数罚，解得的回归系数向量的稀疏性越来越小；当 $\gamma\to 1$ 时，MCP 罚逐渐趋向于 L_0 范数罚，解得的回归系数向量的稀疏性越来越大. MCP 模型对绝对值小于 $\gamma\lambda$ 的回归系数(噪声变量的回归系数)进行压缩，而对绝对值大于 $\gamma\lambda$ 的回归系数(目标变量的回归系数)不再进行压缩. 通过减小对于目标变量回归系数的压缩来实现近似无偏估计.

(5) 桥回归模型. 在第 2.1 节中已提到的桥回归[3,7-9]，实际上是最早被提出的近似无偏稀疏模型，但由于求解算法困难导致一直没有流行起来. 其形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\left\{\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\cdot\|\boldsymbol{\beta}\|_1^\gamma\right\} \tag{15}$$

其中 $\|\boldsymbol{\beta}\|_1^\gamma=\sum_{p=1}^{P}|\beta_p|^\gamma$ 为 L_γ 范数罚，$\gamma\in(0,1]$. 与 SCAD 模型和 MCP 模型类似，桥回归模型也是通过随着回归系数的绝对值的增大而减小压缩的程度来克服 Lasso 有偏估计缺点的.

(6) 小结与分析. 虽然 Lasso 能够得到稀疏回归系数向量，但其有一个显著缺点：Lasso 是有偏估计，只在不可表示条件[17](irrepresentable condition)、稀疏 Riesz 条件[18](sparse Riesz condition)和限制特征值条件[19](restricted eigenvalue condition)等非常强的附加条件下的估计值才是近似无偏的，此时其参数估计一致性和变量选择一致性才存在. 所谓变量选择一致性是指稀疏模型所选出的变量与实际上对响应变量有影响的变量一致，即

$$\lim_{N\to\infty}P(\{p:\hat{\beta}_p\neq 0\}=\{p:\beta_p\neq 0\})=1 \tag{16}$$

Lasso 估计有偏的主要原因在于其不仅对噪声变量对应的回归系数进行压缩，而且对目标变量对应的回归系数也进行压缩，即它对全部变量对应的回归系数向量都进行压缩. 针对 Lasso 估计有偏的问题，提出的正则化稀疏模型大都从调整回归系数向量各分量被惩罚的程度入手，例如自适应 Lasso、松弛 Lasso、SCAD 模型和 MCP 模型，而其中 SCAD 模型和 MCP 模型的罚函数为非凸的 SCAD 罚和 MCP 罚，非凸罚较之凸罚(例如 Lasso 的罚函数就是凸罚)来说在某些方面具有优势，比如利用非凸罚函数得到的正则化稀疏模型往往变量选择一致性不错. 但非凸的罚函数也有弊端，例如非凸性会导致全局最优解不存在，因此在求解优化过程时往往比凸的罚函数难度大.

2.3 自动组效应稀疏模型

Lasso 在进行变量选择时不具有组效应. 所谓

组效应指的是某些变量作为一个整体被同时选中进而参与模型的构造，或同时从模型中移除进而不参与模型的构造，即具有变量组选择的效果. 自动组效应在文献[20]中首先被提出，其含义为某种估计方法令那些彼此之间高度相关的变量的回归系数的绝对值(几乎)相等，从而倾向于将全部高度相关的变量作为一个组同时选中或同时移除. 但自动组效应只能实现对高度相关变量的组选择效果. 自动组效应稀疏模型可被分为三类：一类为通过岭罚实现自动组效应的稀疏模型，包括弹性网[20]、弹性 SCAD[21-22]、两两弹性网[23-24]和迹 Lasso[25]；另一类为通过对回归系数之差施加 L_1 范数罚从而实现自动组效应，包括融合 Lasso[26-31]、两两融合 Lasso[32]、HORSES 模型①、加权两两融合 Lasso[33]和弹性相关网[34]；第三类为利用两两无穷范数罚实现自动组效应，包括 OSCAR 模型[35-36].

2.3.1　通过岭罚实现自动组效应

通过岭罚实现自动组效应的稀疏模型有弹性网、弹性 SCAD、两两弹性网和迹 Lasso.

(1) 弹性网. 弹性网(elastic net)为第一种被提出的具有自动组效应的稀疏模型，其分为朴素弹性网(naive elastic net)和弹性网. 朴素弹性网形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2 \quad (17)$$

其罚函数 $\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2$ 由 L_1 范数罚 $\|\boldsymbol{\beta}\|_1$ 和岭罚(即岭回归的罚函数) $\|\boldsymbol{\beta}\|_2^2$ 组成，被称作弹性网罚函数(elastic net penalty). L_1 范数罚使得朴素弹性网具有稀疏性，岭罚使得弹性网具有自动组效应的特性. 但由于岭罚和 L_1 范数罚都具有压缩回归系数的作用，因此朴素弹性网对回归系数进行了两次压缩，从而导致估计的偏差较大，因此需要对朴素弹性网进行改进. Zou 等人[20]提出的将朴素的弹性网进行一次大小为 $(1+\lambda_2)$ 的比例变换的方法即可解决两次压缩增加的偏差问题，并将比例变换后的朴素弹性网称作弹性网. 令 $\hat{\boldsymbol{\beta}}^{\text{naive}}$ 表示朴素弹性网的解，则弹性网与朴素弹性网之间的关系为 $\hat{\boldsymbol{\beta}}=(1+\lambda_2)\hat{\boldsymbol{\beta}}^{\text{naive}}$.

(2) 弹性 SCAD. 弹性网由于使用了 L_1 范数罚，因此其缺点为有偏估计，Zeng 等人[21]针对此问题提出了弹性 SCAD，其形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{p=1}^{P}\varphi_{\lambda,\gamma}(\beta_p)+\lambda_2\|\boldsymbol{\beta}\|_2^2 \quad (18)$$

其中 $\varphi_{\lambda,\gamma}(\cdot)$ 为 SCAD 罚，$\lambda_1>0$，$\lambda_2>0$，$\gamma>2$. 弹性 SCAD 克服了弹性网有偏估计的缺点，而且由于使用了岭罚，故其具有自动组效应.

(3) 两两弹性网. 弹性网忽略了不同变量之间相关程度大小各不相同的事实：有些变量之间具有相关性，因此需要施加岭罚；有些变量之间的相关程度较低，故施加岭罚的程度要较小；有些变量之间的相关程度较高，故施加岭罚的程度要较大；甚至有些变量之间并不存在相关性，因此根本不需要施加岭罚，因而弹性网“一刀切”地施加同等程度岭罚的方法是不恰当的. 针对该问题，文献[23-24]提出两两弹性网(pairwise elastic net)，其形式如下：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda|\boldsymbol{\beta}|^{\mathrm{T}}\boldsymbol{P}|\boldsymbol{\beta}| \quad (19)$$

其中 $\lambda\geq0$，$|\boldsymbol{\beta}|$ 表示向量 $\boldsymbol{\beta}$ 的全部元素取绝对之后形成的新的向量，$\boldsymbol{P}$ 为元素非负的对称半正定矩阵

$$\boldsymbol{P}=\boldsymbol{I}+\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}-\boldsymbol{R} \quad (20)$$

两两弹性网中的罚函数 $|\boldsymbol{\beta}|^{\mathrm{T}}\boldsymbol{P}|\boldsymbol{\beta}|$ 可被转化为

$$|\boldsymbol{\beta}|^{\mathrm{T}}\boldsymbol{P}|\boldsymbol{\beta}|=\|\boldsymbol{\beta}\|_2^2+\|\boldsymbol{\beta}\|_1^2-|\boldsymbol{\beta}|^{\mathrm{T}}\boldsymbol{R}|\boldsymbol{\beta}| \quad (21)$$

由式(21)可知关于相关系数矩阵 $\boldsymbol{R}$ 中元素 $\boldsymbol{R}_{ij}$ 的罚函数为

$$\boldsymbol{R}_{ij}(\beta_i^2+\beta_j^2)+(1-\boldsymbol{R}_{ij})(|\beta_i|+|\beta_j|)^2 \quad (22)$$

相关系数 $\boldsymbol{R}_{ij}$ 越小则式(22)就越接近于 $\|\boldsymbol{\beta}\|_1^2$，更倾向于稀疏性；相关系数 $\boldsymbol{R}_{ij}$ 越大则式(22)就越接近于 $\|\boldsymbol{\beta}\|_2^2$，更倾向于组效应. 因而可以看出两两弹性网的自动组效应本质上也是通过岭罚实现的. 而且，两两弹性网的显著特点为利用相关系数矩阵 $\boldsymbol{R}$ 来确定矩阵 $\boldsymbol{P}$，从而使得两两弹性网能够根据变量之间相关系数的大小自适应地决定是否施加岭罚以及施加多大程度的岭罚，从而克服了弹性网死板地施加岭罚的缺点.

(4) 迹 Lasso. 另外一种利用岭罚实现自动组效应的稀疏模型为迹 Lasso(trace Lasso)，其形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\cdot tr(\boldsymbol{X}\,\mathrm{diag}(\boldsymbol{\beta})) \quad (23)$$

其中 $\lambda\geq0$，$\mathrm{diag}(\boldsymbol{\beta})$ 表示由向量 $\boldsymbol{\beta}$ 的全部分量作为对角元素构成的对角阵，$tr(\cdot)$ 表示矩阵迹范数罚(trace norm penalty)，含义为矩阵的全部特征值之和. 迹 Lasso 具有根据变量之间相关系数大小不同而自适应变化的功能. 当全部变量两两都不相关且设计矩阵被单位化时，迹范数退化为 L_1 范数罚：

$$tr(\boldsymbol{X}\,\mathrm{diag}(\boldsymbol{\beta}))=\sum_{i=1}^{P}\|\boldsymbol{X}^{(i)}\|_2|\beta_i|=\|\boldsymbol{\beta}\|_1 \quad (24)$$

① Regression shrinkage and grouping of highly correlated predictors with HORSES. http://arxiv.org/abs/1302.0256, 2013,2,1

此时倾向于实现解的稀疏性.当全部变量两两完全相关且设计矩阵被单位化时,迹范数罚退化为 L_2 范数罚:

$$tr(\boldsymbol{X}\,\mathrm{diag}(\boldsymbol{\beta}))=\|\boldsymbol{x}_1\boldsymbol{\beta}\|_*=\|\boldsymbol{x}_1\|_2\|\boldsymbol{\beta}\|_2 \quad (25)$$

此时倾向于实现自动组效应.显然,迹 Lasso 的自动组效应本质上也是通过岭罚实现的.另外,迹 Lasso 较之于弹性网的优点在于:施加的 L_1 范数罚随变量之间的相关性减小而增大,施加的岭罚随着变量之间的相关性增大而增大,因而其自动组效应性质具有自适应特点.

(5)小结与分析.我们知道,岭回归适合处理变量之间存在高度相关性的数据集,Zou 等人[20]受此启发,将岭回归的罚函数(岭罚)引入到 Lasso 的目标函数中,构造出新的稀疏模型弹性网,使得弹性网兼具 Lasso 的稀疏性和岭回归处理相关性数据的优越性,能够将高度相关的变量作为一个整体同时选中或移除.本节介绍的弹性网、弹性 SCAD、两两弹性网、自适应弹性网和迹 Lasso 这些稀疏模型虽然形式完全不同,但是本质上都是利用岭罚而实现的自动组效应的.另外,弹性网、弹性 SCAD 和自适应弹性网的自动组效应比较死板,对全部变量均施加同等程度的岭罚.而两两弹性网和迹 Lasso 根据变量间相关系数值的变化而自适应地改变所施加的岭罚的大小.

2.3.2 通过惩罚回归系数之差(和)实现自动组效应

通过惩罚回归系数之差实现自动组效应的稀疏模型有融合 Lasso、两两融合 Lasso、HORSES 模型、加权两两融合 Lasso 和弹性相关网.

(1)融合 Lasso.融合 Lasso(fused Lasso)[26-31]假设全部变量是有序的,它不仅对回归系数进行惩罚,还对相邻变量的回归系数之差的绝对值进行惩罚,因此不仅会使解稀疏化,还会使相邻回归系数平坦变化,即得到的解具有分段常数化(piecewise constant solution)的特点.基于式(1)中线性回归模型的融合 Lasso 为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\sum_{p=2}^{P}|\beta_p-\beta_{p-1}| \quad (26)$$

其中 $\lambda_1\geq0,\lambda_2\geq0$,$\sum_{p=2}^{P}|\beta_p-\beta_{p-1}|$ 叫做融合罚(fusion penalty),融合罚的作用即为对相邻变量对应的回归系数之差的绝对值进行惩罚,从而使得融合 Lasso 的解具有分段常数化的特点,而 L_1 范数罚的作用为实现解的稀疏性.显然,由于融合 Lasso 的解具有"分段常数化"性质,它的作用是使相邻回归系数的绝对值几乎相等,所以融合 Lasso 可以被看做具有自动组效应性质.但融合 Lasso 的这种自动组效应性质仅仅局限于前后相邻的变量,为最简单的自动组效应性质.

(2)两两融合 Lasso 与 HORSES 模型.融合 Lasso 的应用具有局限性,其只适用于一维有序的变量,只能在前后相邻变量间实现组效应,两两融合 Lasso(pairwise fused Lasso)[32]将融合 Lasso 推广到变量无序的情形,其思想为对任意两两变量的回归系数之差的绝对值都进行惩罚,因而可以令不相邻变量的回归系数的绝对值趋向于相等,从而将不相邻但具有共性的一些变量成组地选择出来,实现不相邻变量间的自动组效应.两两融合 Lasso 的形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\sum_{j=2}^{P}\sum_{i=1}^{j-1}|\beta_i-\beta_j| \quad (27)$$

其中 $\lambda_1\geq0,\lambda_2\geq0$.罚函数 $\sum_{j=2}^{P}\sum_{i=1}^{j-1}|\beta_i-\beta_j|$ 叫作两两融合罚,两两融合罚的作用为对任意两两变量的回归系数之差的绝对值进行惩罚.另外,HORSES 模型本质上属于两两融合 Lasso,其与两两融合 Lasso 唯一的不同之处在于其要求调节参数 λ_1 大于某个人为选定的正数.另外,两两融合 Lasso 和 HORSES 只实现关于正相关变量的自动组效应.

(3)加权两两融合 Lasso.与两两融合 Lasso 只能实现正相关变量的自动组效应不同,加权两两融合 Lasso[33]将相关系数值引入罚函数中,能够同时实现关于正相关变量和负相关变量的自动组效应,即将正相关变量和负相关变量作为一个整体同时选中或移除.加权两两融合 Lasso 的形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\sum_{j=2}^{P}\sum_{i=1}^{j-1}\frac{1}{1-|\rho_{ij}|}|\beta_i-\mathrm{sign}(\rho_{ij})\beta_j| \quad (28)$$

其中 ρ_{ij} 表示变量 $\boldsymbol{x}_i$ 和变量 $\boldsymbol{x}_j$ 之间的相关系数,且

$$\mathrm{sign}(\rho_{ij})=\begin{cases}1, & \rho_{ij}\geq0\\ -1, & \rho_{ij}<0\end{cases} \quad (29)$$

$\mathrm{sign}(\rho_{ij})$ 有两方面的作用:当变量为正相关时对回归系数之差的绝对值进行惩罚;当变量为负相关时对回归系数之和的绝对值进行惩罚,因而能够实现关于正相关和负相关变量的自动组效应.另外,权重 $1/(1-|\rho_{ij}|)$ 随着相关系数绝对值的增大(减小)而增大(减小),这样可以保证当变量之间的相关性越大(小)时对它们之间回归系数之差的惩罚力度越大

(小)，从而令两个变量对应的回归系数的绝对值越接近(不接近)，即加权两两融合 Lasso 的自动组效应具有随相关系数值自适应变化的特点. 特别地，当 $\rho_{ij}=0$ 时，则弱化为两两融合 Lasso.

(4) 弹性相关网. 另外一种能同时实现正相关和负相关变量自动组效应的稀疏模型为弹性相关网(elastic corr-net)[34]，其形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\phi(\boldsymbol{\beta}) \quad (30)$$

其中 $\lambda_1>0$，$\lambda_2>0$，$\phi(\boldsymbol{\beta})$ 为基于相关系数的罚(correlation based penalty)：

$$\phi(\boldsymbol{\beta})=\sum_{j=2}^{P}\sum_{i=1}^{j-1}\left(\frac{(\beta_i-\beta_j)^2}{1-\rho_{ij}}+\frac{(\beta_i+\beta_j)^2}{1+\rho_{ij}}\right) \quad (31)$$

其实现自动组效应的原理为：当 $\rho_{ij}\to 1$(趋向于完全正相关)时，式(31)中括号里的第一项的权重变得很大，此时主要对回归系数之差进行惩罚；当 $\rho_{ij}\to -1$(趋向于完全负相关)时，式(31)中括号里的第 2 项的权重变得很大，此时主要对回归系数之和进行惩罚. 总之，弹性相关网通过将变量之间的相关系数引入罚函数中，使得到的解中正相关的变量具有几乎相等的回归系数大小，同时负相关的变量具有符号相反而绝对值几乎相等的回归系数，既实现了关于正相关变量的组效应，也实现了关于负相关变量的组效应.

(5) 小结与分析. 本节介绍的稀疏模型本质上都是通过惩罚回归系数之差(和)来实现自动组效应的，其中融合 Lasso 只对有序变量中前后相邻的两个施加融合罚，因此只适用于一维有序变量的情形. 但融合 Lasso 是本小结所介绍的其他全部模型的鼻祖，其余模型的提出均受到其融合罚的启发；两两融合 Lasso 和 HORSES 模型对原始的融合 Lasso 进行了拓展，其对任意两两变量之间均施加融合罚，但只能实现关于正相关变量的自动组效应；加权融合 Lasso 和弹性相关网中均引入了数据集中变量之间的具体相关系数值，因而对变量施加的融合罚随相关系数值不同而自适应变化，且能够实现关于正相关和负相关变量的自动组效应(即将正相关和负相关变量作为一个整体同时选中或移除).

2.3.3　通过两两无穷范数实现自动组效应

Bondell 等人提出的 OSCAR(Octagonal Shrinkage and Clustering Algorithm for Regression)[35-36] 模型也具有自动组效应，其通过无穷范数约束任意两两变量的回归系数的最大值，从而令该两个变量的回归系数趋于相等而实现自动组效应. 基于式(1)中线性回归模型的 OSCAR 模型求解问题为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\left(\|\boldsymbol{\beta}\|_1+\boldsymbol{c}\sum_{j=2}^{P}\sum_{i=1}^{j-1}\max\{|\beta_i|,|\beta_j|\}\right) \quad (32)$$

其中 $\boldsymbol{c}>0$，$\lambda\geq 0$. L_1 范数罚的作用为实现稀疏性，罚函数中的第 2 项叫做两两无穷范数罚(pairwise L_∞ norm penalty)，其作用为约束任意两两变量的回归系数的最大值进而使得 OSCAR 模型具有组效应. OSCAR 模型既能实现关于正相关变量的自动组效应，也能实现关于负相关变量的自动组效应.

2.3.4　自动组效应稀疏模型的总结

岭罚具有组效应性质，弹性网、两两弹性网和基于相关系数的弹性网在本质上都是使用了岭罚才具有组效应性质的. 迹 Lasso 虽然没有直接使用岭罚而是使用的迹范数罚，但当变量间存在的相关性增大时迹范数罚朝着岭罚的方向变化，所以本质上也可将迹 Lasso 归为使用岭罚而具有组效应性质的那一类. 而融合 Lasso、两两融合 Lasso、基于相关系数的两两融合 Lasso 和 HORSES 模型本质上是对回归系数之差或和的绝对值进行惩罚，从而促使回归系数相等，实现组效应. OSCAR 模型通过限制回归系数向量的两两无穷范数罚而实现组效应. 但必须指出的是，自动组效应稀疏模型与下文中的预设组效应稀疏模型不同，自动组效应稀疏模型只是对强相关的变量实现组选择，且其组选择是自动实现的，而预设组效应稀疏模型的组选择特性是通过人为事先定义分组而实现的且可对任意变量组实现组选择.

2.4　预设组效应稀疏模型

具有代表性的预设组效应稀疏模型有组 Lasso(Group Lasso)、组 SCAD(Group SCAD)模型、组 MCP(Group MCP)模型和稀疏组 Lasso(Sparse Group Lasso，SGL)等，三者均只具有组稀疏性，而稀疏组 Lasso 既具有变量稀疏性又具有变量组稀疏性. 预设组效应稀疏模型需要预先设定分组，分组情况完全由人为决定，被设定为同一个组的变量的回归系数将同时为零或非零，即同时被预设组效应稀疏模型选中或移除.

(1) 组 Lasso. 实践中常常遇到变量之间具有组结构的情形，例如在多因子方差分析中因子对应的多个哑变量可被视为一个变量组，针对该问题组 Lasso[37] 被提出. 假设存在 P 个变量，预先人为地将它们划分成 J 个变量组，则组 Lasso 为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_{2,1} \quad (33)$$

其中 $\|\boldsymbol{\beta}\|_{2,1}=\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$，$\boldsymbol{\beta}_j$ 为第 j 个变量组对应的回归系数向量. 组结构还有一些特殊情形，例如不同变量组所包含的变量出现重复的情形[38-40]和变量组之间的关系为树结构[41-43]的情形，组 Lasso 正在从简单的组结构向着复杂的重复组(overlap group)结构和树组(tree-guided group)结构等方向发展.

(2)组 SCAD 与组 MCP. 与 Lasso 类似，组 Lasso 的估计也是有偏的，针对该问题，文献[44]利用 SCAD 罚构造出组 SCAD 模型，文献[45]利用 MCP 罚构造出组 MCP 模型. 组 SCAD 与组 MCP 具有如下统一的形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2) \tag{34}$$

$\varphi_{\lambda,\gamma}(\cdot)$ 为式(12)中的 SCAD 罚或式(14)中的 MCP 罚. 因为利用了非凸的 SCAD 罚和 MCP 罚，所以组 SCAD 与组 MCP 克服了组 Lasso 估计有偏的缺点.

(3)稀疏组 Lasso. Simon 等人[46]将 Lasso 与组 Lasso 的罚函数结合到一起提出了稀疏组 Lasso. 稀疏组 Lasso 的形式为

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_{2,1}+\lambda_2\|\boldsymbol{\beta}\|_1 \tag{35}$$

其中 $\lambda_1\|\boldsymbol{\beta}\|_{2,1}$ 使得稀疏组 Lasso 具有组稀疏性，而 $\lambda_2\|\boldsymbol{\beta}\|_1$ 使得稀疏组 Lasso 的某些分组内的变量也具有稀疏性.

(4)小结与分析. 预设组效应的实现必须预先将变量进行分组，并且分组情况完全由实验者自主决定，因而其与自动组效应的最大区别在于可实现关于任意分组的组选择效果，而自动组效应稀疏模型只能实现关于高度相关变量的组选择. 值得指出的是，同时结合了 Lasso 罚函数和组 Lasso 罚函数的稀疏组 Lasso 既能实现变量选择也能实现变量组选择. 另外，组 SCAD 和组 MCP 克服了组 Lasso 的有偏估计缺点.

2.5 其他稀疏模型

(1) Dantzig 选择器(Dantzig Selector). 基于式(1)中线性回归模型的 Dantzig 选择器[47-53]为如下的凸规划问题：

$$\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\|\boldsymbol{\beta}\|_1 \tag{36a}$$

$$\text{s.t.}\quad\|\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\|_\infty\leq\lambda_D \tag{36b}$$

其中 $\lambda_D>0$，且

$$\|\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\|_\infty=\sup_{1\leq i\leq P}|(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}))_i| \tag{37}$$

Dantzig 选择器也可以被写作如下无约束的形式：

$$\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\|\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\|_\infty+\tilde{\lambda}_D\|\boldsymbol{\beta}\|_1 \tag{38}$$

其中 $\tilde{\lambda}_D>0$. 显然，Dantzig 选择器与 Lasso 的不同之处在于 Lasso 的目标函数是最小化输出误差的 L_2 范数的平方，而 Dantzig 选择器的目标函数是最小化输出误差向量与设计矩阵乘积的 L_∞ 范数，两者共同之处在于均使用了 L_1 范数罚 $\|\boldsymbol{\beta}\|_1$，因此 Dantzig 选择器也能实现稀疏解，具有类似于 Lasso 的变量选择功能. 文献[49]指出当 $P\leq N$(参数空间维数小于样本空间维数)时，Dantzig 选择器和 Lasso 的解路径相同的充分条件为矩阵 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ 满足对角占优条件：

$$\boldsymbol{M}_{ii}>\sum_{i\neq j}|\boldsymbol{M}_{ij}| \tag{39}$$

其中 $\boldsymbol{M}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$，$i,j\in\{1,\cdots,P\}$. 特别地，在二维的参数空间中(即 $P=2$ 时)对角占优条件总是成立的，因此 Dantzig 选择器与 Lasso 得到的解在二维的参数空间中总是等价的. 但是文献[49]中给出的 Lasso 与 Dantzig 选择器两者产生的解等价的条件只局限于 $P\leq N$ 的情形，为此文献[53]给出了在任意情形下(无论是 $P\leq N$ 还是 $P>N$) Lasso 与 Dantzig 选择器等价的条件：对于 Lasso

$$\min_{\boldsymbol{\beta}\in\boldsymbol{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_L\|\boldsymbol{\beta}\|_1 \tag{40}$$

令 I_L 表示由 Lasso 得到的解 $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ 中的非零元素的索引集，令 $\tilde{\boldsymbol{X}}_L\in\boldsymbol{R}^{N\times|I_L|}$ 表示由索引集 I_L 对应的变量组成的矩阵，令 $\boldsymbol{X}_L\in\boldsymbol{R}^{N\times|I_L|}$ 表示将 $\tilde{\boldsymbol{X}}_L$ 中每个变量乘以 $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ 的分量 β_i 的符号后形成的新的矩阵，假设设计矩阵 $\boldsymbol{X}$ 是满秩矩阵且秩为 $|I_L|$，在式(36)中的可调参数 λ_D 与式(40)中 Lasso 的可调参数 λ_L 满足 $\lambda_D=\lambda_L$ 时，若

$$\boldsymbol{u}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{1}\geq\boldsymbol{0} \tag{41a}$$

$$\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}\|_\infty\leq 1 \tag{41b}$$

成立，则 $\hat{\boldsymbol{\beta}}_{\mathrm{Dantzig}}=\hat{\boldsymbol{\beta}}_{\mathrm{LASSO}}$，其中 $\boldsymbol{1}$ 为 $|I_L|$ 维的全部元素都为 1 的向量，$\boldsymbol{u}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{1}\geq\boldsymbol{0}$ 表示向量 $\boldsymbol{u}$ 中的元素全部都大于等于 0. 此外，文献[53]还指出，当令 $\lambda_D=\lambda_L$ 时，式(40)中的 Lasso 的解总是式(36)中的 Dantzig 选择器的一个可行解(虽然不一定是最优解)，因此当 Lasso 和 Dantzig 选择器的解不同时，Dantzig 选择器的解要比 Lasso 的解更具稀疏性. 文献[49]指出 Dantzig 选择器解路径的波动性比较严重，Lasso 的解路径比 Dantzig 选择器的解路径要平滑地多，并且 Lasso 关于 $\boldsymbol{\beta}$ 的均方误差 $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2$ 和 $\boldsymbol{X}\boldsymbol{\beta}$ 的均方误差 $\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2$ 与 Dantzig 选择器相比差不多，有时候还明显小于 Dantzig 选择器的均方误差，尤其当信噪比较高或变量间的相关

性较高时 Dantzig 选择器的均方误差要比 Lasso 的均方误差大很多. 但 Dantzig 选择器与 Lasso 相比最大的优点在于其要求解的优化问题是一个线性规划问题，求解非常方便.

(2) 核 Lasso. 将式(1)中的线性模型推广为如下的基于核函数的非线性模型：

$$\boldsymbol{y}=\sum_{n=1}^{N}\beta_n k(\boldsymbol{x},\boldsymbol{x}_n)+\boldsymbol{\varepsilon} \tag{42}$$

其中 $\boldsymbol{\varepsilon}$ 为噪声向量，$k(\boldsymbol{x},\boldsymbol{x}_n)$ 为核函数. 已知设计矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\boldsymbol{R}^{N\times P}$，其中 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\in\boldsymbol{R}^P$，$N$ 为样本数，P 为变量数，假设 $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I})$，则基于式(42)中非线性模型的核 Lasso(kernel Lasso)[54-56]为

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^N}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{K}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\\&=\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^N}\frac{1}{2}\sum_{n=1}^{N}\Big(y_n-\sum_{m=1}^{N}\beta_m k(\boldsymbol{x}_n,\boldsymbol{x}_m)\Big)^2+\lambda\sum_{m=1}^{N}|\beta_m|\end{aligned} \tag{43}$$

其中 $\boldsymbol{K}$ 为由核函数 $k(\boldsymbol{x}_n,\boldsymbol{x}_m)$ 作为元素组成的核矩阵 $\boldsymbol{K}=[k(\boldsymbol{x}_n,\boldsymbol{x}_m)]_{n,m=1}^{N}$，$n\in\{1,\cdots,N\}$，$m\in\{1,\cdots,N\}$. 核 Lasso 的提出是因为 Lasso 只能处理线性回归模型下的变量选择问题，存在一定的局限性，在处理非线性问题时线性 Lasso 的预测效果较差. 而引入核函数是处理非线性问题的典型方法，对于线性不可分的情形，核函数利用非线性变换将数据从低维空间映射到高维空间，在高维空间中实现分类. 核函数的引入使得核 Lasso 具有处理非线性情形下变量选择问题的能力，克服了 Lasso 不适用于处理非线性情形下变量选择问题的缺点.

(3) 图 Lasso. 已知

$$\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(N)}\sim N_P(\boldsymbol{0},\boldsymbol{\Sigma}) \tag{44}$$

则基于式(44)的图 Lasso(Graphical Lasso)[57-60]为

$$\hat{\boldsymbol{\Sigma}}^{-1}=\arg\max_{\boldsymbol{\Sigma}^{-1}>0}\log|\boldsymbol{\Sigma}^{-1}|-tr(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})-\lambda\|\boldsymbol{\Sigma}^{-1}\|_1 \tag{45}$$

其中

$$\boldsymbol{S}=\sum_{n=1}^{N}\boldsymbol{X}^{(n)}\boldsymbol{X}^{(n)\mathrm{T}}/N \tag{46}$$

$\|\boldsymbol{\Sigma}^{-1}\|_1=\sum_{j=1}^{P}\sum_{i=1}^{P}|\boldsymbol{\Sigma}_{ij}^{-1}|$ 使得图 Lasso 具有稀疏解. 近年来将稀疏方法应用于概率图模型的结构和参数学习中是研究的热点，很多学者还将图 Lasso 推广到有向无环图[61-62]、Ising 模型[63-64]、半参数图模型[65-67]、Poisson 图模型[68-69]、节点不能观图模型[70-71]和时变图模型[72-74]等其他概率图模型中，图 Lasso 的应用领域正在快速扩展.

2.6　各稀疏模型的对比

各正则化稀疏模型的特点如表 1 所示，其中“稀疏性”指变量水平上的稀疏性，“预设组稀疏性”和“自动组稀疏性”指变量组水平上的组稀疏性.

表 1　各正则化稀疏模型的特点归纳

模型	稀疏性	预设组稀疏性	自动组稀疏性	无偏性	说明
非负绞刑估计	√	×	×	×	提出比 Lasso 早，具有稀疏性
Lasso	√	×	×	×	具有开创性的意义，使得正则化稀疏模型真正开始广泛流行
自适应 Lasso	√	×	×	√	近似无偏估计，克服了 Lasso 有偏估计的缺点
松弛 Lasso	√	×	×	√	近似无偏估计，克服了 Lasso 有偏估计的缺点
SCAD 模型	√	×	×	√	近似无偏估计，克服了 Lasso 有偏估计的缺点
MCP 模型	√	×	×	√	近似无偏估计，克服了 Lasso 有偏估计的缺点
桥回归	√	×	×	√	近似无偏估计，克服了 Lasso 有偏估计的缺点
弹性网	√	×	√	×	克服 Lasso 不适合处理共线性数据集的缺点
弹性 SCAD	√	×	√	√	克服 Lasso 不适合处理共线性数据集的缺点且具有估计无偏性
两两弹性网	√	×	√	×	引入相关性信息，克服 Lasso 不适合处理共线性数据集的缺点
迹 Lasso	√	×	√	×	罚函数中引入设计矩阵，能根据相关性信息自适应地施加岭罚
融合 Lasso	√	×	√	×	对相邻变量的系数进行融合，其解具有分段常数化的特点
两两融合 Lasso	√	×	√	×	与融合 Lasso 区别在于其对全部变量两两之间的系数都进行融合
HORSES	√	×	√	×	对全部变量两两之间的系数都进行融合
加权两两融合 Lasso	√	×	√	×	在两两融合 Lasso 的基础上引入了数据集中的相关性信息
OSCAR	√	×	√	×	通过两两无穷范数罚同时实现对正负相关变量的组效应
弹性相关网	√	×	√	×	引入相关性信息，克服 Lasso 不适合处理共线性数据集的缺点
组 Lasso	×	√	×	×	克服了 Lasso 不能进行变量组选择的缺点
组 SCAD	×	√	×	√	克服组 Lasso 估计有偏的缺点
组 MCP	×	√	×	√	克服组 Lasso 估计有偏的缺点
稀疏组 Lasso	√	√	×	×	同时具有稀疏性和组稀疏性
核 Lasso	√	×	×	×	克服 Lasso 不适合处理非线性情形的缺点
图 Lasso	√	×	×	×	简化概率图模型的结构
Dantzig 选择器	√	×	×	×	优化问题为线性规划问题，求解方便

3 贝叶斯模型

早在 1996 年 Tibshirani[1] 就指出基于式(1)中线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 的 Lasso 可以被表示为一个最大后验估计,其中 Lasso 的正则化项对应最大后验估计中的先验分布,Lasso 的损失函数项对应最大后验估计中的似然函数,文献[75-76]根据这一思想进一步具体地提出了贝叶斯 Lasso(Bayesian Lasso). 贝叶斯 Lasso 实质上是关于参数向量 $\boldsymbol{\beta}$ 的最大后验估计,其似然函数和先验分布分别如式(47)和(48)所示:

$$P(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta})=\prod_{n=1}^{N}N(y_n\mid\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\beta},\sigma^2) \tag{47}$$

$$P(\boldsymbol{\beta}\mid\sigma^2)=\prod_{p=1}^{P}\frac{\lambda}{2\sigma}\mathrm{e}^{-\lambda|\beta_p|/\sigma} \tag{48}$$

其中 $\boldsymbol{x}_n$ 和 y_n 表示第 n 个样本,且各样本间是相互独立的,y_n 服从均值为 $\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{\beta}$ 和方差为 σ^2 的正态分布,方差 σ^2 即为式(1)中误差变量 $\varepsilon_n\sim N(0,\sigma^2)$ 的方差. 需注意的是,式(48)中参数向量 $\boldsymbol{\beta}$ 的先验分布是条件 Laplace 分布,而且取 σ^2 所服从的分布为均匀分布,如下文中式(51e)所示,文献[46]指出这样做能够保证后验分布的单峰性,若后验分布不具有单峰性会导致吉布斯抽样算法收敛太慢. 由式(47)和(48)便可以求得参数向量 $\boldsymbol{\beta}$ 的后验概率分布

$$P(\boldsymbol{\beta}\mid\boldsymbol{y})=P(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta})P(\boldsymbol{\beta}\mid\sigma^2) \tag{49}$$

根据贝叶斯理论,在贝叶斯估计中有了参数向量 $\boldsymbol{\beta}$ 的后验概率分布后,若给定某可信水平 $1-\alpha$,则计算回归系数 β_p 的贝叶斯可信区间(Bayesian credible interval)是非常直接的,因此贝叶斯 Lasso 能够提供全部回归系数 $\beta_1,\cdots,\beta_P$ 的贝叶斯可信区间,这是贝叶斯 Lasso 的显著优点,因为 Lasso 这种非贝叶斯形式的稀疏模型只能给出回归系数的点估计值而不能给出其区间估计. 另外,贝叶斯 Lasso 通过吉布斯抽样等算法可以求得式(1)中误差变量 ε_n 的方差 σ^2 的估计值,这也是 Lasso 所不能做到的. 文献[77]中进一步指出,将式(48)中的 Laplace 分布表示为如下 Gaussian scale mixture 分布的形式会有效降低计算的复杂度:

$$P(\boldsymbol{\beta}\mid\sigma^2)=\prod_{p=1}^{P}\int_0^{\infty}\frac{\lambda}{\sqrt{2\pi\sigma^2\tau_p^2}}\mathrm{e}^{-\lambda|\beta_p|^2/2\sigma^2\tau_p^2}\frac{\lambda^2}{2}\mathrm{e}^{-\frac{\lambda^2}{2}\tau_{p2}^2}\mathrm{d}\tau_p^2$$

$$=\prod_{p=1}^{P}\int_0^{\infty}\boldsymbol{N}(\beta_p\mid 0,\sigma^2\tau_p^2)\mathrm{Gamma}\left(\tau_p^2\mid 1,\frac{\lambda^2}{2}\right)\mathrm{d}\tau_p^2 \tag{50}$$

其中 $\boldsymbol{N}(\beta_p\mid 0,\sigma^2\tau_p^2)$ 表示 β_p 服从均值为零且方差为 $\sigma^2\tau_p^2$ 的正态分布,$\mathrm{Gamma}\left(\tau_p^2\mid 1,\frac{\lambda^2}{2}\right)$ 表示 τ_p^2 服从参数为 1 和 $\frac{\lambda^2}{2}$ 的伽马分布,显然式(48)将 $\boldsymbol{\beta}$ 的先验分布进一步转化成了两层结构:顶层分布为正态分布,底层分布为伽马分布. 因此整个贝叶斯 Lasso 的最大后验估计形式可以被表示为如下的层次贝叶斯模型(Hierarchical Bayesian Model):

$$\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta},\sigma^2\sim\boldsymbol{N}_N(\boldsymbol{X\beta},\sigma^2\boldsymbol{I}_N) \tag{51a}$$

$$\boldsymbol{\beta}\mid\sigma^2,\tau_1^2,\cdots,\tau_P^2\sim\boldsymbol{N}_P(\boldsymbol{0}_P,\sigma^2\boldsymbol{D}_\tau) \tag{51b}$$

$$\boldsymbol{D}_\tau=\mathrm{diag}(\tau_1^2,\cdots,\tau_p^2) \tag{51c}$$

$$\sigma^2,\tau_1^2,\cdots,\tau_P^2\sim\pi(\sigma^2)\mathrm{d}\sigma^2\prod_{p=1}^{P}\frac{\lambda^2}{2}\mathrm{e}^{-\lambda^2\tau_p^2/2}\mathrm{d}\tau_p^2 \tag{51d}$$

$$\pi(\sigma^2)=1/\sigma^2 \tag{51e}$$

$$\sigma^2,\tau_1^2,\cdots,\tau_P^2>0 \tag{51f}$$

其中式(51e)表示 σ^2 服从均匀分布,$\boldsymbol{I}_N$ 为 $N\times N$ 阶的单位矩阵. 可以看出,关于 $\boldsymbol{\beta}$ 的整个最大后验估计形式被表示成了一个具有 3 层分布结构的层次模型:第 1 层分布为响应向量所服从的均值为 $\boldsymbol{X\beta}$ 且协方差矩阵为 $\sigma^2\boldsymbol{I}_n$ 的正态分布,第 2 层分布为回归系数向量所服从的均值为零向量且协方差矩阵为 $\sigma^2\boldsymbol{D}_\tau$ 的正态分布,第 3 层分布为 τ_p^2 所服从的参数为 1 和 $\frac{\lambda^2}{2}$ 伽马分布和 σ^2 所服从的均匀分布 $\pi(\sigma^2)=1/\sigma^2$. 文献[75]中利用吉布斯抽样方法求解贝叶斯 Lasso 中的参数,但吉布斯抽样方法对于高维数据集来说计算复杂度很高,期望最大化算法(Expectation Maximation,EM)的计算复杂度往往比吉布斯抽样方法低,因此未来探索如何利用 EM 算法求解贝叶斯 Lasso 是一个值得研究的方向.

实际上,很多正则化稀疏模型均有其对应的贝叶斯模型,均可以被表示为一个最大后验估计形式,例如桥回归模型的最大后验估计——贝叶斯桥(Bayesian Bridge)[75]、自适应 Lasso 的最大后验估计形式——贝叶斯自适应 Lasso(Bayesian Adaptive Lasso)[78]、弹性网的最大后验估计形式——贝叶斯弹性网(Bayesian Elastic Net)[79]、组 Lasso 的最大后验估计形式——贝叶斯组 Lasso(Bayesian Group Lasso)[77,80] 以及图 Lasso 的最大后验估计形式——贝叶斯图 Lasso(Bayesian Graphical Lasso)[81] 等,虽然各个正则化稀疏模型的最大后验估计形式不同(主要是先验分布不同),但其原理是相同的,即正则化稀疏模型的损失函数项对应最大后验估计形式中的似然函数. 正则化稀疏模型的正则化项对应最大后验估计形式中的先验分布,各正则化稀疏模型的最大后验估计形式均可以用类似于式(51a)~(51f)的层次贝叶斯模型并利用吉布斯抽样算法求解,具

体形式请参考相应的文献，在此不再赘述.

另外，贝叶斯模型推理中的先验信息多种多样，远不止目前已经构造的稀疏模型中的正则化项的形式. 从贝叶斯角度设计新的先验信息从而得到新的正则化稀疏模型是一个重要的研究方向.

4　正则化稀疏模型的求解算法

正则化稀疏模型本质上为一个最优化问题. 在 Lasso 被提出的前几年，由于缺少对其高效求解的算法，所以一直没有广泛流行. 直到 LAR 算法(Least Angle Regression，LAR)的提出，使得 Lasso 的求解方便而快捷，Lasso 等一系列正则化稀疏模型开始被广泛地研究. 文献[4]中指出，在一定条件下 LAR 算法的解路径与 Lasso 的解路径一致，因而可以通过 LAR 算法来求解 Lasso 的解. 另外，LAR 算法的变体组 LAR 算法(Group Least Angle Regression)[37]可用来求解组 Lasso.

坐标下降[10,82](coordinate descent)及其变体组坐标下降[37](Block coordinate descent)可用来求解 Lasso、自适应 Lasso、非负绞刑估计、弹性网、组 Lasso 和稀疏组 Lasso 等问题，但不同的稀疏模型的坐标下降算法的具体形式不同，例如 Lasso 和组 Lasso 等稀疏模型直接利用坐标下降算法或组坐标下降算法求解即可，而求解稀疏组 Lasso 需要利用两层迭代的结构才行，其中外层迭代为针对 $L_{2,1}$ 范数罚的组坐标下降算法，内层迭代为针对 L_1 范数罚的坐标下降算法.

对于 SCAD 模型与 MCP 模型这类非凸的稀疏模型来说，无法直接应用凸优化方法求解，Fan 等人提出的 LQA(Local Quadratic Approximation)[14]方法可有效解决优化问题目标函数中的非凸非光滑难题，但 LQA 方法往往较为耗时.

ADMM 方法(Alternating Direction Method of Multipliers)[83]可以被用来求解 Lasso 和组 Lasso 等诸多稀疏模型问题. 已知优化问题：

$$\min_{\boldsymbol{x}\in\boldsymbol{R}^P} g(\boldsymbol{x})+h(\boldsymbol{z}) \tag{52a}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{x}=\boldsymbol{c} \tag{52b}$$

其中 $g(\cdot)$ 和 $h(\cdot)$ 均为凸的，$\boldsymbol{x}\in\boldsymbol{R}^P$，$\boldsymbol{z}\in\boldsymbol{R}^Q$，$\boldsymbol{A}\in\boldsymbol{R}^{N\times P}$，$\boldsymbol{B}\in\boldsymbol{R}^{N\times Q}$，$\boldsymbol{c}\in\boldsymbol{R}^N$. 对应的增广 Lagrange 函数为

$$L_\rho(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y})=f(\boldsymbol{x})+g(\boldsymbol{z})+\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c})+\frac{\rho}{2}\|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{z}-\boldsymbol{c}\|_2^2 \tag{53}$$

其中 $\rho>0$，$\boldsymbol{y}\in\boldsymbol{R}^N$，ADMM 算法由如下的迭代组成：

$$\boldsymbol{x}^{(k+1)}=\arg\min_{\boldsymbol{x}\in\boldsymbol{R}^P} L_\rho(\boldsymbol{x},\boldsymbol{z}^{(k)},\boldsymbol{y}^{(k)}) \tag{54}$$

$$\boldsymbol{z}^{(k+1)}=\arg\min_{\boldsymbol{z}\in\boldsymbol{R}^P} L_\rho(\boldsymbol{x}^{(k+1)},\boldsymbol{z},\boldsymbol{y}^{(k)}) \tag{55}$$

$$\boldsymbol{y}^{(k+1)}=\boldsymbol{y}^{(k)}+\rho(\boldsymbol{A}\boldsymbol{x}^{(k+1)}+\boldsymbol{B}\boldsymbol{z}^{(k+1)}-\boldsymbol{c}) \tag{56}$$

另外一类求解稀疏模型的算法为近似梯度方法(proximal gradient method)及其变种，近似梯度方法又叫做广义梯度方法(generalized gradient method)或近似算子方法(proximal operator method). 近似梯度方法一般要求解的优化问题的形式为

$$\min_{\boldsymbol{x}\in\boldsymbol{R}^P} g(\boldsymbol{x})+h(\boldsymbol{x}) \tag{57}$$

其中 $g(\boldsymbol{x})$ 为可微的凸函数，$h(\boldsymbol{x})$ 为任意的凸函数，不要求 $h(\boldsymbol{x})$ 为可微的. 求解上式中优化问题的近似梯度算法的迭代公式为

$$\boldsymbol{x}^{(k)}=prox(\boldsymbol{x}^{(k-1)}-t_k\cdot\nabla g(\boldsymbol{x}^{k-1})),$$

其中 t_k 为步长，$prox(\cdot)$ 为近似算子：

$$prox(\boldsymbol{u})=\arg\min_{\boldsymbol{z}\in\boldsymbol{R}^P}\frac{1}{2t}\|\boldsymbol{u}-\boldsymbol{z}\|^2+h(\boldsymbol{z}) \tag{58}$$

当 $h(\boldsymbol{x})=0$ 时，近似梯度算法就是经典的梯度方法；当 $h(\boldsymbol{x})$ 为指示函数 $h(\boldsymbol{x})=I_C(\boldsymbol{x})$ 时，近似梯度算法退化为投影梯度法；当 $h(\boldsymbol{x})$ 为 L_1 范数罚 $h(\boldsymbol{x})=\|\boldsymbol{x}\|_1$ 时，近似梯度算法特化为所谓的 ISTA 算法(Iterative Soft-thresholding Algorithm，ISTA)[84]. Nesterov[85-87]对近似梯度方法的收敛速度进行了改善. 近似梯度算法及其变种形式广泛应用于稀疏优化问题的求解中，例如文献[88]利用近似梯度方法的变种求解以融合罚为罚函数，文献[89]利用近似梯度方法的变种求解以迹范数罚为罚函数的稀疏模型，文献[90]利用近似梯度方法的变种求解重叠组 Lasso. 另外，著名的稀疏优化问题求解软件包 SLEP(Sparse Learning with Efficient Projections)中的算法基本上都采用了近似梯度方法及其变种形式来求解稀疏模型，更详细的说明请参考该软件包提供的说明书①. 另外，更多关于稀疏模型求解方法的研究综述，读者可参考文献[91].

5　实　验

本节通过高维小样本不相关数据集实验和高维小样本相关数据集实验来展示 Lasso、SCAD、组 Lasso、稀疏组 Lasso 和弹性 SCAD 这 5 种具有代表性的正则化稀疏模型的变量选择效果进行对比，其中 Lasso 和 SCAD 代表了只在变量水平上实现稀疏性的稀疏模型，其他的稀疏模型例如自适应 Lasso、

① http://www.public.asu.edu/~jye02/Software/SLEP/

松弛 Lasso 和 MCP 模型等的变量选择效果与 Lasso 和 SCAD 的变量选择效果类似;组 Lasso 代表了在组水平上实现稀疏性的预设组效应稀疏模型,其他的稀疏模型像组 SCAD 模型和组 MCP 模型的变量选择效果与组 Lasso 类似;稀疏组 Lasso 代表了同时实现变量水平上稀疏性和组变量水平上稀疏性的稀疏模型;弹性 SCAD 代表了具有关于高度相关变量的组选择能力的自动组效应稀疏模型;其他的稀疏模型像弹性网、OSCAR 模型、HORSES 模型、两两弹性网和两两融合 Lasso 等与弹性 SCAD 的变量选择效果一致.

(1)高维小样本不相关数据集实验. 首先通过不相关数据集实验来对比基于最小二乘损失函数的 3 种稀疏模型:Lasso(使用了 L_1 范数罚)、Group Lasso(使用了 $L_{2,1}$ 范数罚)与 Sparse Group Lasso(同时使用了 L_1 范数罚和 $L_{2,1}$ 范数罚)的变量选择效果. 当今稀疏模型能够实现的稀疏效果有 3 种:普通的稀疏性、组稀疏性和双稀疏性,其中普通的稀疏性是指稀疏效果只在变量水平上实现,具有代表性的稀疏模型为 Lasso,组稀疏性是指在变量组的水平上实现稀疏效果,具有代表性的稀疏模型为 Group Lasso;双稀疏性是指既在变量组水平上实现变量组选择,又在组内的变量中实现变量选择,即既有稀疏性又具有组稀疏性,具有代表性的稀疏模型为 Sparse Group Lasso. 因此,本实验中我们选择 Lasso、Group Lasso 和 Sparse Group Lasso 这 3 种具有代表性的模型来说明稀疏模型的变量选择效果. 首先生成一模拟数据集,该数据集包含 20 个样本和 100 个变量:变量 1,变量 2,…,变量 10,…,变量 100,且各个变量之间不具有相关性,其真实的系数向量为

$$(-2,-1,0,1,0,0,1,1,0,0,0,\cdots,0),$$

即变量 1、变量 2、变量 4、变量 7 和变量 8 的回归系数非零,为想要选择出来的重要变量,而其余变量均为不期望选择出来的噪声变量. 对于 Group Lasso 和 Sparse Group Lasso 来说,其分组情况为

分组 1=(变量 1,变量 2,变量 3,变量 4),

分组 2=(变量 5,变量 6),

分组 3=(变量 7,变量 8),

分组 4=(变量 9,变量 10),

分组 5=(变量 11,变量 12,…,变量 100).

对于 Lasso 来说不进行任何分组. 3 种稀疏模型所拟合出的回归系数向量如表 2 和表 3 所示,但由于空间有限,我们只列出前 10 个变量变量 1,变量 2,…,变量 10 的回归系数,从前 10 个变量的回归系数就足以展示 3 种稀疏模型的变量选择功能了. 从表 2 和表 3 的对比可以看出,Group Lasso 具有组稀疏性,因为其各个分组中的回归系数同时为零或同时非零,尤其从分组 1 中的回归系数可以看出虽然变量 3 为不期望选择出来的噪声变量,但由于分组时将变量 3(噪声变量)与期望选择出来的变量 1、变量 2 和变量 4 分在了同一组,所以其也不得不被选择出来,所以这 4 个变量的回归系数同时不为零. 再观察表 3,其中的 Sparse Group Lasso,Sparse Group Lasso 与 Group Lasso 的分组情况完全一致,但得到的稀疏性效果却不同:Sparse Group Lasso 得到的回归系数向量中变量 3(噪声变量)的系数为零,而 Group Lasso 却不能将分组 1 中变量 3 的回归系数置零,这说明 Sparse Group Lasso 的优点为具有组内变量水平上的稀疏性. 本实验中利用了求解 Lasso 的软件包 glmnet①;求解 Group Lasso 利用了软件包 gglasso②;求解 Sparse Group Lasso 利用了软件包 SGL③. 总之,glmnet、gglasso 和 SGL 软件包均来自 CRAN 网站(The Comprehensive R Archive Network,CRAN).

表 2 Lasso 拟合出的回归系数向量

模型	截距	变量 1	变量 2	变量 3	变量 4	变量 5	变量 6	变量 7	变量 8	变量 9	变量 10	…
Lasso	0.803	−0.724	−0.136	0	0.254	−0.016	0	0	0.721	0	0	…

表 3 Group Lasso 与 SGL 拟合出的回归系数向量

模型	截距	分组 1				分组 2		分组 3		分组 4		…
		变量 1	变量 2	变量 3	变量 4	变量 5	变量 6	变量 7	变量 8	变量 9	变量 10	
GL	0.145	−1.789	−0.822	0.049	0.933	0	0	0.742	0.981	0	0	…
SGL	0.916	−7.062	−2.498	0	3.714	0	0	1.793	4.315	0	0	…

(2)高维小样本相关数据集实验. 稀疏模型是否具有自动组效应由罚函数部分决定,而与损失函数无关,下面通过高度相关数据集实验来对比基于

① http://cran.r-project.org/web/packages/glmnet/index.html
② http://cran.r-project.org/web/packages/gglasso/index.html
③ http://cran.r-project.org/web/packages/SGL/index.html

铰链损失函数的Lasso、SCAD和弹性SCAD的变量选择效果，三者分别使用了 L_1 范数罚、SCAD罚和弹性SCAD罚，而损失函数均为铰损失函数. 首先生成一个含有若干高度相关变量的模拟数据集，该模拟数据集包含 $n=50$ 个样本和 $p=300$ 个变量，其中前5个变量两两之间具有高度的相关性且相关系数 $\rho=0.9$. 进行60次实验，将60次实验结论的平均值列入表4中，其中 *tol*、*sig* 和 *unsig* 分别表示基于铰链损失函数的Lasso、SCAD、弹性SCAD选中的变量数、选中的重要变量数、选中的噪声变量数. 通过表4中的实验结果可以看出Lasso与SCAD具有明显的变量选择能力，但面对5个高度相关的变量却只能选择出其中的一小部分；弹性SCAD也具有明显的变量选择能力，而且几乎将高度相关的重要变量都选出来了. 因此可以得出结论：弹性SCAD具有关于高度相关变量的变量组选择能力，而Lasso和SCAD具有变量选择能力但却不具有变量组选择能力. 与上一节中的实验一样，本实验中求解各稀疏模型的软件包来自网站CRAN，求解SCAD和弹性网的软件包在CRAN网站[①]上有许多.

表4　Lasso、SCAD和弹性SCAD的变量选择效果对比

模型	样本数 n	变量数 p	选出变量数 *tol*	选出的重要变量数 *sig*	选出的噪声变量数 *noi*
Lasso	50	300	17.48	2.16	15.32
SCAD			3.36	1.53	1.83
弹性SCAD			16.82	4.97	11.85

6　正则化稀疏模型的应用

总体来说，Group Lasso适用于具有组稀疏性的数据集的变量选择，例如在基因微阵列分析中属于同一个生物学路径的基因可被归类为一个基因组，在基因关联研究中某基因的全部基因标记可被视为一个基因标记组；弹性网、两两弹性网、迹Lasso、OSCAR模型和HORSES模型等适用于具有高度相关变量的数据集的变量选择；融合Lasso适用于处理变量的回归系数具有光滑性的数据集，例如在微阵列比较基因组杂交(Array-based Comparative genomic hybridization，arrayCGH)分析中，相邻基因往往被认为具有共性，往往将相邻基因的回归系数近似相等作为先验信息；核Lasso适合处理非线性情形下的变量选择问题；图Lasso适合处理具有网络结构的数据集. 下面介绍一些稀疏模型的代表性的应用情况.

6.1　在生物信息学和医药学中的应用

稀疏模型在生物信息学中有大量的应用. 随着科学技术的发展，生物医学领域中数据的规模、多样性以及复杂性快速增长，形成了海量数据的状况. 然而，虽然生物医药数据是海量的，但往往只有一小部分是有意义的数据，即生物医药数据可以被稀疏化. 在微阵列比较基因组杂交(Array-based Comparative genomic hybridization，arrayCGH)中，相邻基因往往被认为具有共性，因此将相邻基因的回归系数近似相等作为先验信息是合理的. Liu等人[88]将融合罚应用于arrayCGH中对膀胱癌的级别分类问题(tumor grade classification)，实验证明利用融合罚的分类准确性比利用 L_1 范数罚的分类准确性要高出6个百分点. 科学家通常通过脑部图像来诊断阿尔茨海默(Alzheimer's Disease)疾病，一种合理的假设是阿尔茨海默疾病的早期损害集中于脑部的某个区域，然而传统方法在诊断阿尔茨海默病时往往忽略了立体像素之间的关联性，针对上述问题，Xin等人[92]利用基于两两融合罚的逻辑斯蒂回归方法对阿尔茨海默疾病的诊断并且通过实验证明收到了良好的效果，诊断的准确性要比 L_1 逻辑斯蒂回归、支持向量机以及文献[93]和文献[94]中的方法都高. Allen等人[95]将图Lasso推广到随机变量服从泊松分布的情形，并且将其应用到乳腺癌微核苷酸(microRNA)的网络结构的学习中；Zhong等人[96]将Lasso、OSCAR模型、融合Lasso和弹性网应用于乳腺癌数据的基因选择中. 实验证明OSCAR模型的测试准确性最高. 另外，还有很多文献将稀疏模型应用于生物信息学与医药学中[97-98].

6.2　在信号去噪中的应用

文献[10]中提出了FLSA(fused lasso signal approximator)方法，该方法利用了融合Lasso的罚函数，其与经典的全变差(total variation)去噪方法的区别在于多了一项 L_1 范数罚，FLSA方法可对一维的信号进行去噪. 文献[10]中还提出了二维的FLSA方法，该方法假设对象为一个网格，对网格结构的横向和纵向均施加融合Lasso罚，因而能够对二维的图像进行去噪，文献[99]中应用二维的融合Lasso进行图像去噪获得了良好的效果. Wang等人[100]将Lasso和自适应Lasso应用到图像去噪领域：传统去噪方法考虑的都是去除高斯噪声的影响，而Wang等人考虑了一种脉冲噪声(impulse

① http://cran.r-project.org/

noise),脉冲噪声是数字系统中常见的一种噪声,常常由相机中传感器发生故障、模数转换错误以及硬件中存储位置错误等造成. Wang 等人通过实验对最小二乘法、Lasso 和自适应 Lasso 的去噪效果进行了对比,实验结果表明自适应 Lasso 的去噪效果最好;Selesnick 等人[101]在经典的全变差去噪基础上进一步假设相邻变量的回归系数之差具有重叠的组结构,将该重叠组结构作为先验信息引入去噪方法中,实验结果表明得到的去噪信号比全变差方法更加平滑.

6.3 在信号重建中的应用

压缩感知无疑是稀疏模型应用的前沿阵地,例如利用了 L_1 范数罚的基追踪方法为压缩感知中图像重建的重要方法. 然而,很多情况下信号的稀疏结构是组稀疏的,因此文献[102]将 L_1 范数罚替换为组 Lasso 的罚函数、稀疏组 Lasso 的罚函数和重叠组 Lasso 的罚函数,将组稀疏结构、双水平稀疏结构(组稀疏而且组内元素也稀疏)和重叠组稀疏结构引入到图像重建中. 实验结果证明基于双水平稀疏结构的信号重建方法和基于重叠组稀疏结构的信号重建方法均优于基于单纯组稀疏结构的信号重建方法. Rao 等人①指出当信号的具有重叠组稀疏结构时,将该重叠组稀疏结构作为先验信息进行信号重建所需的测量值比不利用重叠稀疏结构作为先验的标准压缩感知方法更少,并且给出了所需测量值的边界. 以往的压缩感知方法均假设信号是时不变的,而文献[103]将组 Lasso 的罚函数和融合 Lasso 的罚函数应用到时变信号的重建当中.

6.4 在人脸与语音识别中的应用

文献[104]中提出基于稀疏表示的分类算法(Sparse Representation-based Classification, SRC)应用于人脸识别中,该方法利用了 L_1 范数罚;而文献[105]则将组稀疏结构引入人脸识别中;进一步地,文献[106]将更复杂的数组稀疏结构引入 SRC 中实现人脸识别,并且通过实验证明该方法取得了比 SRC 更佳的识别效果;文献[107]将迹范数罚(trace norm penalty)应用于人脸识别中,提出了一种叫做监督迹 Lasso(Supervised Trace Lasso, SSL)的方法,但值得指出的是该方法中迹范数罚的构造与上文中迹 Lasso 中的迹范数罚的构造方式不完全相同. Tan 等人[108]将组 Lasso 的组稀疏思想应用于语音识别领域,将组稀疏思想与文献[109]中的稀疏贝叶斯学习(Sparse Bayesian Learning)方法结合,提出了一种新的组稀疏贝叶斯学习(Group Sparse Bayesian Learning)方法,并通过实验证明组稀疏结构的语音识别方法(组稀疏贝叶斯学习方法)准确率明显高于不具有组稀结构的语音识别方法(稀疏贝叶斯学习方法);另外,他们还提出了一种叫做组弹性网(Group Elastic Net)的新稀疏模型,该模型是将弹性网中的 L_1 范数罚替换为组 Lasso 的 $L_{2,1}$ 范数罚而得到的,他们的实验还表明组弹性网模型的识别准确率高于弹性网的识别准确率,而且弹性网和 Lasso 的识别准确率都比稀疏贝叶斯方法高.

7 未来研究方向

7.1 拓展正则化稀疏模型

尽管已经提出了很多正则化稀疏模型,但它们仍然存在各自的缺点,如何克服这些正则化稀疏模型的缺点? 一种方法为改进已有的罚函数,例如 Zou 等人将 L_1 范数罚和 L_2 范数罚结合到一起提出的弹性网兼具 L_1 范数罚的稀疏性和 L_2 范数罚的组效应性质,类似的,未来通过改进已有的正则化稀疏模型也许可以得到大大优于已有正则化稀疏模型的新的正则化稀疏模型. 例如,将非凸的 capped-L_1 罚函数[110-111]、对数罚[112-114](log penalty)、对数和罚函数(Log-Sum Penalty, LSP)[115]、对数指数和罚[116](Log-Exp-Sum penalty)、Geman 罚函数[117](Geman Penalty, GP)等非凸罚函数推广到变量组选择情形下从而得到相应的组稀疏模型的问题值得研究. 当然,另一种方法为针对已有正则化稀疏模型存在的缺点直接设计新的罚函数并将其应用到正则化稀疏模型中.

7.2 其他回归模型上的推广

正则化稀疏模型在线性回归模型中已经得到了极其广泛的应用,并且其中一些在广义线性回归模型中得到了一定程度的应用,但还有很多正则化稀疏模型在其他某些回归模型中的应用有待于更深入更广泛的研究,未来我们可以将已有的正则化稀疏模型推广到其他回归模型中,以便处理在相应回归模型下的变量选择问题. 例如,将融合 Lasso 推广到 COX 比例风险回归模型的情形和将 MCP 模型推广到负二项回归模型的情形等问题有待于研究. 研究损失函数的改变,对稀疏模型统计特性和变量选择特性的影响也是值得研究的问题.

① Tight measurement bounds for exact recovery of structured sparse signals. http://arxiv.org/abs/1106.4355, 2011, 10, 18

7.3 正则化稀疏模型的一致性

很多正则化稀疏模型的一致性尚未被研究，例如，基于相关系数的弹性网和两两弹性网的变量选择一致性和参数估计一致性等的研究仍为空白. 另外，正则化稀疏模型的一致性大都在不可表示条件和限制特征值条件等假设条件下进行探讨，然而这些假设条件在实际应用中指导性不大，而且目前主要使用人工产生的数据集对变量选择一致性进行验证，对变量选择一致性进行验证的基准实际数据集仍没有建立，这些方面还需要继续深入地研究.

7.4 贝叶斯观点重新审视正则化稀疏模型

根据贝叶斯的观点，用损失函数＋正则化罚项构造模型即相当于用似然函数＋正则化罚项构造模型，而正则化罚函数本质上对应于贝叶斯模型推理中的先验信息. 贝叶斯模型推理中的先验信息多种多样，远不止目前已经构造的稀疏模型中的罚函数形式，而且针对不同的问题应使用不同的先验信息，到底使用哪种先验信息，怎样判别先验信息形式的好坏，都是值得探讨的问题.

7.5 稀疏模型问题的再思考

最原始的稀疏模型或变量选择的问题本质上应该是组合优化问题，即给定 P 维变量的采样样例，如何从 P 维变量中选择 k 维变量子集参与到回归系数向量的求解当中去，使得这 k 维变量子集是真正与输出最相关的变量，且在求得的回归系数向量尽可能低的维数情况下，对输出的预测尽可能的准确. 这里牵涉到到底选择多大的变量子集，选择的变量子集是否是真正产生输出变量的原因变量，选择的变量是否能够产生足够准确的回归系数向量的问题，这些问题的权衡有时候可能是矛盾的，有的时候变量选择正确并不一定能够使得回归系数向量对输出的预测更准确. 真的是在损失函数或似然函数上增加一个罚项就能实现上述所有目标了吗？如何在目标函数中引入更加精细化的约束项，体现上述所有要求，是值得认真思考的问题. 而且损失函数或似然函数与稀疏罚项之间存在一个比例分配系数(即可调参数 λ)，使得整个问题的求解需要遍历整个比例分配系数取值范围的解路径，才能确定这个比例分配系数适当的值，除了某些特殊情况外，这使得参数选择算法复杂性太高，在实际应用中不太可行. 如何解决这些问题仍需要研究.

7.6 稀疏方法在支持向量机中的应用

稀疏方法已被应用到支持向量机(support vector machine)领域用来同时进行分类和变量选择，例如使用了 L_1 范数罚的 L_1 SVM[118] 和使用了 SCAD 罚的 SCAD SVM[15] 不仅仅为分类器，其还具有稀疏性，具有变量选择功能；而使用了“L_1 范数罚＋L_2 范数罚的”DrSVM[119-120] (doubly regularized support vector machine)同时具有 3 种能力：分类、变量选择和组效应. 未来可以探讨融合 Lasso 的融合罚、两两融合罚、迹范数罚和 $L_{2,1}$ 范数罚等其他罚函数应用于支持向量机领域中得到的分类器具有何种特性.

7.7 模型稀疏化的质疑

模型稀疏化有许多优点，但是到底是否应该稀疏化，稀疏化到什么程度才合适？最近有学者[121]对稀疏模型提出质疑，他们认为稀疏性必然带来模型不稳定，模型不稳定使得留一误差估计值不准确，最终使得模型的泛化误差不好，因此模型的稀疏化和稳定性是一对不可调和的矛盾，在使用时必须在两者之间做出恰当的权衡. 另外，我们不禁要问，除了牺牲算法的稳定性以外，得到模型稀疏化的同时是否还会付出更多其他方面的代价？

8 结论与展望

本文对各种正则化稀疏模型进行了综述，指出了各个模型提出的原因、所具有的优缺点和应用中应该注意的问题. 纵观正则化稀疏模型的发展历程，其大都是根据人们期望的新特性，合理地在已有模型的目标函数基础上进行改动或直接设计新的目标函数，大致思想为或将先验信息加入到罚函数中，或将不同的罚函数进行整合，或将其推广到其他的回归模型情形下等. 正则化稀疏模型为进行变量选择的有效方法，可以解决建模过程中由于高维数据集造成的过拟合问题和数值计算病态等问题，其在机器学习和图像处理等领域势必将发挥越来越重要的作用.

参 考 文 献

[1] Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288

[2] Breiman L. Better subset regression using the nonnegative garrote. Technometrics, 1995, 37(4): 373-384

[3] Frank L L E, Friedman J H. A statistical view of some chemometrics regression tools. Technometrics, 1993, 35(2): 109-135

[4] Efron B, Hastie T, Johnstone I, et al. Least angle regression. The Annals of Statistics, 2004, 32(2): 407-499

[5] Yuan M, Lin Y. On the non-negative garrotte estimator. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007, 69(2): 143-161

[6] Xiong S. Some notes on the nonnegative garrote. Technometrics, 2010, 52(3): 349-361

[7] Fu W J. Penalized regressions: The bridge versus the Lasso. Journal of Computational and Graphical Statistics, 1998, 7(3): 397-416

[8] Knight K, Fu W. Asymptotics for Lasso-type estimators. Annals of Statistics, 2000, 28(5): 1356-1378

[9] Huang J, Horowitz J L, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. The Annals of Statistics, 2008, 36(2): 587-613

[10] Friedman J, Hastie T, Höfling H, et al. Pathwise coordinate optimization. The Annals of Applied Statistics, 2007, 1(2): 302-332

[11] Zou H. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association, 2006, 101(476): 1418-1429

[12] Huang J, Ma S, Zhang C H. Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica, 2008, 18(4): 1603-1618

[13] Meinshausen N. Relaxed Lasso. Computational Statistics & Data Analysis, 2007, 52(1): 374-393

[14] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348-1360

[15] Zhang H H, Ahn J, Lin X, et al. Gene selection using support vector machines with non-convex penalty. Bioinformatics, 2006, 22(1): 88-95

[16] Zhang C H. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 2010, 38(2): 894-942

[17] Zhao P, Yu B. On model selection consistency of Lasso. The Journal of Machine Learning Research, 2006, 7(11): 2541-2563

[18] Zhang C H, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear regression. The Annals of Statistics, 2008, 36(4): 1567-1594

[19] Bickel P J, Ritov Y, Tsybakov A B. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics, 2009, 37(4): 1705-1732

[20] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320

[21] Zeng L, Xie J. Group variable selection via SCAD-L_2. Statistics, 2014, 48(1): 49-66

[22] Becker N, Toedt G, Lichter P, et al. Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. BMC Bioinformatics, 2011, 12(1): 1-13

[23] Lorbert A, Ramadge P J. The pairwise elastic net support vector machine for automatic fMRI feature selection// Proceedings of the IEEE International Conference on Speech and Signal Processing. Vancouver, Canada, 2013: 1036-1040

[24] Lorbert A, Eis D, Kostina V, et al. Exploiting covariate similarity in sparse regression via the pairwise elastic net// Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Italy, 2010: 477-484

[25] Grave E, Obozinski G R, Bach F R. Trace Lasso: A trace norm regularization for correlated designs//Proceedings of the Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems. Granada, Spain, 2011: 2187-2195

[26] Tibshirani R, Saunders M, Rosset S, et al. Sparsity and smoothness via the fused Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(1): 91-108

[27] Rinaldo A. Properties and refinements of the fused Lasso. The Annals of Statistics, 2009, 37(5B): 2922-2952

[28] Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused Lasso. Biostatistics, 2008, 9(1): 18-29

[29] Rapaport F, Barillot E, Vert J P. Classification of arrayCGH data using fused SVM. Bioinformatics, 2008, 24(13): 375-382

[30] Zhou J, Liu J, Narayan V A, et al. Modeling disease progression via fused sparse group Lasso//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 1095-1103

[31] Ye G B, Xie X. Split Bregman method for large scale fused Lasso. Computational Statistics & Data Analysis, 2011, 55(4): 1552-1569

[32] Hoefling H. A path algorithm for the fused Lasso signal approximator. Journal of Computational and Graphical Statistics, 2010, 19(4): 984-1006

[33] Daye Z J, Jeng X J. Shrinkage and model selection with correlated variables via weighted fusion. Computational Statistics & Data Analysis, 2009, 53(4): 1284-1298

[34] El Anbari M, Mkhadri A. Penalized regression combining the L_1 norm and a correlation based penalty. Sankhya B: The Indian Journal of Statistics, 2014, 76 (1): 82-102

[35] Bondell H D, Reich B J. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. Biometrics, 2008, 64(1): 115-123

[36] Zeng X R, Figueiredo M A. Solving OSCAR regularization problems by fast approximate proximal splitting algorithms. Digital Signal Processing, 2014, 31(1): 124-135

[37] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49-67

[38] Mosci S, Villa S, Verri A, et al. A primal-dual algorithm for group sparse regularization with overlapping groups// Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2010: 2604-2612

[39] Jacob L, Obozinski G, Vert J P. Group Lasso with overlap and graph Lasso//Proceedings of the 26th Annual International Conference on Machine Learning. Quebec, Canada, 2009: 433-440

[40] Bach F, Jenatton R, Mairal J, et al. Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning, 2012, 4(1): 1-106

[41] Jenatton R, Mairal J, Obozinski G, Bach F. Proximal methods for hierarchical sparse coding. Journal of Machine Learning Research, 2011, 12(7): 2297-2334

[42] Liu J, Ye J. Moreau-Yosida regularization for grouped tree structure learning//Proceedings of the Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Vancouver, Canada, 2010: 1459-1467

[43] Huang J, Zhang T, Metaxas D. Learning with structured sparsity. The Journal of Machine Learning Research, 2011, 12(9): 3371-3412

[44] Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics, 2007, 23(12): 1486-1494

[45] Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Statistical Science, 2012, 27(4): 481-499

[46] Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-Group Lasso. Journal of Computational and Graphical Statistics, 2013, 22(2): 231-245

[47] Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 2007, 35(6): 2365-2369

[48] Antoniadis A, Fryzlewicz P, Letué F. The Dantzig selector in Cox's proportional hazards model. Scandinavian Journal of Statistics, 2010, 37(4): 531-552

[49] Meinshausen N, Rocha G, Yu B. Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig. The Annals of Statistics, 2007, 35(6): 2373-2384

[50] Romberg J K. The Dantzig selector and generalized thresholding //Proceedings of the 42nd Annual Conference on Information Sciences and Systems. Princeton, USA, 2008: 22-25

[51] Asif M S, Romberg J. On the Lasso and Dantzig selector equivalence//Proceedings of the 45th Annual Conference on Information Sciences and Systems. Baltimore, USA, 2010: 1-6

[52] Koltchinskii V. The Dantzig selector and sparsity oracle inequalities. Bernoulli, 2009, 15(3): 799-828

[53] James G M, Radchenko P, Lv J. DASSO: Connections between the Dantzig selector and Lasso. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2009, 71(1): 127-142

[54] Wang G, Yeung D Y, Lochovsky F H. The kernel path in kernelized LASSO//Proceedings of the 11th International Conference on Artificial Intelligence and Statistics. San Juan, Puerto Rico, 2007: 580-587

[55] Roth V. The generalized Lasso. IEEE Transactions on Neural Networks, 2004, 15(1): 16-28

[56] Gao J, Kwan P W, Shi D. Sparse kernel learning with LASSO and Bayesian inference algorithm. Neural Networks, 2010, 23(2): 257-264

[57] Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94(1): 19-35

[58] Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. The Journal of Machine Learning Research, 2008, 9(3): 485-516

[59] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical Lasso. Biostatistics, 2008, 9(3): 432-441

[60] Dahl J, Vandenberghe L, Roychowdhury V. Covariance selection for nonchordal graphs via chordal embedding. Optimization Methods & Software, 2008, 23(4): 501-520

[61] Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika, 2010, 97(3): 519-538

[62] Fu F, Zhou Q. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. Journal of the American Statistical Association, 2013, 108(501): 288-300

[63] Ravikumar P, Wainwright M J, Lafferty J D. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. The Annals of Statistics, 2010, 38(3): 1287-1319

[64] Wainwright M J, Ravikumar P, Lafferty J D. High-dimensional graphical model selection using L_1-regularized logistic regression//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2006: 1465-1472

[65] Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. The Journal of Machine Learning Research, 2009, 10(10): 2295-2328

[66] Liu H, Han F, Yuan M, et al. High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics, 2012, 40(4): 2293-2326

[67] Xue L, Zou H. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. The Annals of Statistics, 2012, 40(5): 2541-2571

[68] Yang E, Allen G, Liu Z, et al. Graphical models via generalized linear models//Proceedings of the 25th Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1358-1366

[69] Yang E, Ravikumar P, Allen G I, et al. On Poisson graphical models//Proceedings of the 26th Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2013: 1718-1726

[70] Chandrasekaran V, Parrilo P A, Willsky A S. Latent variable graphical model selection via convex optimization. The Annals of Statistics, 2012, 40(4): 1935-1967

[71] Ma S, Xue L, Zou H. Alternating direction methods for latent variable Gaussian graphical model selection. Neural Computation, 2013, 25(8): 2172-2198

[72] Kolar M, Song L, Ahmed A, et al. Estimating time-varying networks. The Annals of Applied Statistics, 2010, 4(1): 94-123

[73] Zhou S, Lafferty J, Wasserman L. Time varying undirected graphs. Machine Learning, 2010, 80(2-3): 295-319

[74] Kolar M, Xing E P. On time varying undirected graphs// Proceedings of the International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, USA, 2011: 407-415

[75] Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association, 2008, 103(482): 681-686

[76] Hans C. Bayesian Lasso regression. Biometrika, 2009, 96(4): 835-845

[77] Chandran M. Analysis of Bayesian Group-Lasso in Regression Models [Ph. D. dissertation]. University of Florida, Florida, USA, 2011

[78] Leng C, Tran M N, Nott D. Bayesian adaptive Lasso. Annals of the Institute of Statistical Mathematics, 2014, 66(2): 221-244

[79] Li Q, Lin N. The Bayesian elastic net. Bayesian Analysis, 2010, 5(1): 151-170

[80] Raman S, Fuchs T J, Wild P J, et al. The Bayesian group-Lasso for analyzing contingency tables//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 881-888

[81] Wang H. Bayesian graphical Lasso models and efficient posterior computation. Bayesian Analysis, 2012, 7(4): 867-886

[82] Wu T T, Lange K. Coordinate descent algorithms for Lasso penalized regression. The Annals of Applied Statistics, 2008, 2(1): 224-244

[83] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 2011, 3(1): 1-122

[84] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202

[85] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983, 27(2): 372-376

[86] Nesterov Y. Smooth minimization of non-smooth functions. Mathematical Programming, 2005, 103(1): 127-152

[87] Nesterov Y. Gradient methods for minimizing composite objective function. Mathematical Programming, 2013, 140(1): 125-161

[88] Liu J, Yuan L, Ye J. An efficient algorithm for a class of fused Lasso problems//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 323-332

[89] Ji S, Ye J. An accelerated gradient method for trace norm minimization//Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada, 2009: 457-464

[90] Yuan L, Liu J, Ye J. Efficient methods for overlapping group Lasso//Proceedings of the Advances in Neural Information Processing Systems. Granada, Spain, 2011: 352-360

[91] Tao Qing, Gao Qian-Kun, Jiang Ji-Yuan, Chu De-Jun. Survey of solving the optimization problems for sparse learning. Journal of Software, 2013, 24(11): 2498-2507(in Chinese) (陶卿, 高乾坤, 姜纪远, 储德军. 稀疏学习优化问题的求解综述. 软件学报, 2013, 24(11): 2498-2507)

[92] Xin B, Kawahara Y, Wang Y, et al. Efficient generalized fused Lasso with its application to the diagnosis of Alzheimer's disease//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2014: 2163-2169

[93] Bo C, Zhang D Q, Shen D G. Domain transfer learning for MCI conversion prediction. Medical Image Computing and Computer-Assisted Intervention, 2012, 15(1): 82-90

[94] Chu C, Hsu A L, Chou K H, et al. Does feature selection improve classification accuracy impact of sample size and feature selection on classification using anatomical magnetic resonance images. Neuroimage, 2012, 60(1): 59-70

[95] Allen G I, Liu Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data// Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. Philadelphia, USA, 2012: 1-6

[96] Zhong L W, Kwok J T. Efficient sparse modeling with automatic feature grouping. IEEE Transactions on Neural Networks and Learning Systems, 2012, 23(9): 1436-1447

[97] Ma Shuangge, Song Xiao, Huang Jian. Supervised group Lasso with applications to microarray data analysis. BMC Bioinformatics, 2007, 8(60): 1-17

[98] Wang D, Eskridge K M, Crossa J. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. Journal of Agricultural, Biological, and Environmental Statistics, 2011, 16(2): 170-184

[99] Tibshirani R, Taylor J. The solution path of the generalized Lasso. Annals of Statistics, 2010, 39(3): 1335-1371

[100] Wang L, Zhu J. Image denoising via solution paths. Annals of Operations Research, 2010, 174(1): 3-17

[101] Selesnick I W, Chen P Y. Total variation denoising with overlapping group sparsity//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 5696-5700

[102] Gishkori S, Leus G. Compressed sensing for block-sparse smooth signals//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, 2014: 4166-4170

[103] Angelosante D, Giannakis G B, Grossi E. Compressed sensing of time-varying signals//Proceedings of the 16th International Conference on Digital Signal Processing. Santorini, Greece, 2009: 1-8

[104] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227

[105] Fusco G, Zini L, Noceti N, et al. Structured multi-class feature selection for effective face recognition//Proceedings of the 17th International Conference on Image Analysis and Processing, Naples, Italy, 2013: 410-419

[106] Jia K, Chan T H, Ma Y. Robust and practical face recognition via structured sparsity//Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 2012: 331-344

[107] Lai J, Jiang X. Supervised trace Lasso for robust face recognition//Proceedings of the IEEE International Conference on Multimedia and Expo. Chengdu, China, 2014: 1-6

[108] Tan Q F, Narayanan S S. Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(4): 1337-1346

[109] Tipping M E. Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research, 2001, 1(6): 211-244

[110] Zhang T. Analysis of multi-stage convex relaxation for sparse regularization. The Journal of Machine Learning Research, 2010, 11(3): 1081-1107

[111] Zhang T. Multi-stage convex relaxation for feature selection. Bernoulli, 2013, 19(5B): 2277-2293

[112] Mazumder R, Friedman J H, Hastie T. SparseNet: Coordinate descent with nonconvex penalties. Journal of the American Statistical Association, 2011, 106(495): 1125-1138

[113] Gao C, Wang N, Yu Q, et al. A feasible nonconvex relaxation approach to feature selection//Proceedings of the 25th AAAI Conference on Artificial Intelligence. San Francisco, USA, 2011: 1-6

[114] Zhang Z, Tu B. Nonconvex penalization using Laplace exponents and concave conjugates//Proceedings of the Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 611-619

[115] Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted L_1 minimization. Journal of Fourier Analysis and Applications, 2008, 14(5): 877-905

[116] Geng Z, Wang S, Yu M, et al. Group variable selection via convex Log-Exp-Sum penalty with application to a breast cancer survivor study. University of Wisconsin-Madison, Madison Wisconsin, America: Technical Report No. 1175R, 2013

[117] Geman D, Yang C. Nonlinear image recovery with half-quadratic regularization. IEEE Transactions on Image Processing, 1995, 4(7): 932-946

[118] Zhu J, Rosset S, Hastie T, et al. 1-norm support vector machines//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2003, 15: 49-56

[119] Wang L, Zhu J, Zou H. The doubly regularized support vector machine. Statistica Sinica, 2006, 16(2): 589-615

[120] Wang L, Zhu J, Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics, 2008, 24(3): 412-419

[121] Xu H, Caramanis C, Mannor S. Sparse algorithms are not stable: A no-free-lunch theorem. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(1): 187-193

LIU Jian-Wei, born in 1966, Ph.D., associate professor. His main research interests include intelligent information processing, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing.

CUI Li-Peng, born in 1990, M. S. candidate. His main research interests include sparsity model of machine learning, analysis and design of automatic control system, intelligent information processing.

LIU Ze-Yu, born in 1993, M. S. candidate. His main research interest is machine learning.

LUO Xiong-Lin, born in 1963, Ph. D., professor. His main research interests include intelligent control, and analysis, prediction, controlling of complicated nonlinear system.

Background

The problem of sparse learning is becoming more and more popular in bioinformatics, machine learning, signal processing and artificial intelligence. The regularized sparse models produce sparse solutions automatically and thus the models are simplified to a certain degree. Therefore, they can solve the over-fitting problem in machine learning effectively. In addition, the regularized sparse models have the ability of variable selection and thus play an important role in bioinformatics. In a word, the problem of sparse learning is a hot and important topic in many areas.

This paper gives a systematical survey of the regularized sparse models. We point the motivations and characteristics of the different regularized sparse models. In addition, we summarize and compare all the regularized sparse models in tabular form.

This work is supported by the National Natural Science Foundation of China (21006127).