第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

结构稀疏模型及其算法研究进展

刘建伟 崔立鹏 罗雄麟

(中国石油大学(北京)自动化系 北京 102249)

摘 要 结构稀疏模型在统计学、信号处理和机器学习等领域中具有重要的应用。结构稀疏模型主要通过在目标函数中引入会导致组稀疏效果的罚函数来实现特征组结构选择。有趣的是一些组稀疏模型不仅能实现特征组选择,而且同时能够实现组内的特征选择。根据使用的罚函数的类型,结构稀疏模型主要分为组套索模型和非凸罚组稀疏模型两大类。系统地总结了重要的组结构稀疏模型,分析了各种组结构稀疏模型之间的区别与联系,归纳比较了各种组结构稀疏模型的统计特性(例如模型选择一致性、参数估计一致性和 oracle 性质)和组结构稀疏模型的求解算法。当前,结构套索模型主要包括普通组套索模型、$L_{\infty,1}$ 组套索模型、重叠组套索模型、树组套索模型、多输出树组套索模型、混合组套索模型、自适应组套索模型、逻辑斯蒂组套索模型和贝叶斯组套索模型。非凸罚组稀疏模型包括组 SCAD 罚模型、组桥模型和组 MC 罚模型等。求解组稀疏模型的算法有组最小角回归算法、块坐标下降(上升)算法、活动集算法、内点算法、投影梯度算法、谱投影梯度算法、轮换方向乘子算法和块坐标梯度下降算法等,结合组稀疏模型对这些算法进行了详细的分析。在使用上述优化方法前,通常需要对目标函数进行预处理,将不平滑的、非凸的、块坐标不可分离的组稀疏模型的目标函数向平滑、凸、块坐标可分离的方向进行转化,这一步常利用的技巧有变分不等式、Nesterov的平滑近似技巧、局部一阶泰勒展开近似、局部二次近似、对偶范数和对偶函数等。接着给出了最新提出的一些组稀疏模型,如关于广义加模型的组套索模型、复合组桥模型、平方根组套索模型和关于 Tobit 模型的组套索模型等。最后,对组稀疏模型未来的研究方向进行了探讨。

关键词 稀疏,组稀疏,罚函数,组套索,特征组选择,组内特征选择,算法

中图法分类号 TP181 文献标识码 A

Research and Development on Structured Sparse Models and Algorithms

LIU Jian-wei CUI Li-peng LUO Xiong-lin

(Department of Automation, China University of Petroleum, Beijing 102249, China)

Abstract The group sparse model has many important applications in the statistics, signal processing and machine learning. The group sparse model achieves feature group selection through introducing the sparsity-inducing penalty function into the objective function. It's interesting that some group sparse models can achieve feature group selection and feature selection within groups simultaneously. According to the penalty functions, the sparse group models are mainly divided into two categories, i. e., Group Lasso models and the group sparse models with non-convex penalty. This paper systematically summarized important group sparse models and analyzed the differences and relations between various group sparse models. In addition, we summarized and compared the statistical properties (such as model selection consistency, parameter estimation consistency and oracle property) and the solving algorithms for various group sparse models. Roughly speaking, the Group Lasso models include normal Group Lasso model, $L_{\infty,1}$ penalty Group Lasso model, overlapping Group Lasso model, tree guided Group Lasso model, multiple-output tree guided Group Lasso model, mixed Group Lasso model, adaptive Group Lasso model, logistic Group Lasso model and Bayesian Group Lasso model. Algorithms for solving group sparse model are composed of Group LARS, block coordinate descent method (block coordinate ascent method), active set method, interior point method, projected gradient method, spectral projected gradient method, alternating direction method of multipliers and block coordinate gradient descent method. We carried out a detailed analysis of these algorithms for specific group sparse models. Before using the optimization methods above, we must pretreat the objective function, i. e., we must transform the nonsmooth, nonconvex and non-separable penalty function in the objective function of group sparse model into smooth, convex and separable functions. Variational inequalities, Nesterov's smooth approximation techniques, local first-order approximation by Taylor series expansion, local quadratic approximation, the dual norm and dual function are often used in this step. Next, some group sparse mo-

本文受中国石油大学(北京)基础学科研究基金项目(JCXK-2011-07)资助。

刘建伟(1966—),男,博士,副研究员,主要研究方向为智能信息处理、复杂系统分析、预测与控制、算法分析与设计,E-mail:liujw@cup.edu.cn;
崔立鹏(1990—),男,硕士生,主要研究方向为机器学习等;罗雄麟(1963—),男,博士,教授,主要研究方向为智能控制。

dels which are recently proposed were introduced, such as Group Lasso model based on generalized additive model, composite group bridge model, group square-root Lasso model, Group Lasso model based on Tobit model and so on. Finally, we talked the future research directions of the group sparse models.

Keywords Sparsity, Group sparsity, Penalty function, Group lasso, Feature group selection, Feature selection within group, Algorithm

1 引言

随着社交网络、电子商务平台、基因组工程、智慧城市和智能电网等大数据应用的飞速发展，产生了大量几百维到十几万维的高维或超高维数据。这些数据的高维输入变量中只有很少一些变量与输出变量有比较高的相关度，而其余大部分输入变量为冗余或噪声输入变量，直接使用最小二乘等传统的建模工具求解低维空间中的模型是组合优化问题，在所有的输入变量集中选择某些变量子集，得到输出预测准确率最高的输入变量子集。套索模型(Least Absolute Shrinkage and Selection Operator, Lasso)是 Tibshirani 提出的基于正则化方法的能够同时实现稀疏特征变量选择和模型参数估计的方法。线性回归模型 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$ 的套索模型为[1]：

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

其中，$\lambda\geqslant 0$，$\|\cdot\|_2$ 为 L_2 范数，$\|\boldsymbol{\beta}\|_1=\sum_{p=1}^{P}|\boldsymbol{\beta}_p|$ 为模型向量 $\boldsymbol{\beta}\in\mathbf{R}^P$ 的 L_1 范数，$\boldsymbol{X}\in\mathbf{R}^{N\times P}$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_N)^{\mathrm{T}}\in\mathbf{R}^N$，$N$ 为样本个数，P 为输入变量个数。套索模型关于第 p 个输入变量的解为：

$$\hat{\beta}_p=\mathrm{sign}(\hat{\beta}_p^{OLS})(|\hat{\beta}_p^{OLS}|-\lambda)_+$$
$$=\begin{cases}\hat{\beta}_p^{OLS}-\lambda, & \hat{\beta}_p^{OLS}>\lambda\\ 0, & -\lambda\leqslant\hat{\beta}_p^{OLS}\leqslant\lambda\\ \hat{\beta}_p^{OLS}+\lambda, & \hat{\beta}_p^{OLS}<-\lambda\end{cases} \tag{2}$$

其中，$\hat{\beta}_p^{OLS}$ 为最小二乘估计(OLS)解 $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$，$\mathrm{sign}(\cdot)$ 在符号函数，$(|\hat{\beta}_p^{OLS}|-\lambda)_+$ 在 $|\hat{\beta}_p^{OLS}|-\lambda>0$ 时取 $|\hat{\beta}_p^{OLS}|-\lambda$，否则等于 0。显然，由式(2)可以看出套索模型的解具有稀疏性。

套索模型使用 L_1 范数正则化罚实现稀疏化模型，然而有时样本的多个特征变量之间存在组结构，Yuan 等人把输入变量之间的组结构作为先验信息，将输入特征变量分组，使用 L_1 范数罚罚组子向量，提出了组套索模型(Group Lasso)[2]，组套索模型在特征变量组的水平上同时实现了稀疏特征组选择和模型参数估计。为了更好地与其它组套索模型的变种进行区分，本文将 Yuan 等人提出的组套索模型称作普通组套索模型。显然，Tibshirani 提出的套索和 Yuan 等人提出的普通组套索模型之间的区别在于模型选择的效果不同，套索模型导致的模型稀疏性是特征水平上的，未选中的特征变量对应的模型分量 β_p 均为 0，将其叫做特征变量选择；而组套索模型导致的模型稀疏性是特征变量组水平上的，这些未选中的组中的特征变量对应的模型向量 β_j 均为 0，将其叫做特征变量组选择。组套索模型的提出引出了利用结构稀疏化思想进行稀疏特征变量选择的新思路，后来组套索模型的许多变种和基于非凸罚函数的组稀疏模型陆续被提出，将它们统称为组稀疏模型。值得指出的是，有些组稀疏模型能够同时实现特征变量组选择和组内的特征变量选择。组稀疏模型的目标函数的一般形式为：

$$\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\Xi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})=\boldsymbol{\Phi}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})+\lambda\cdot\boldsymbol{\Omega}(\boldsymbol{\beta}) \tag{3}$$

其中，$\boldsymbol{\Phi}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$ 为损失函数，$\boldsymbol{\Omega}(\boldsymbol{\beta})$ 为诱导稀疏的罚函数，$\lambda\geqslant 0$，$\boldsymbol{X}\in\mathbf{R}^{N\times P}$，$\boldsymbol{\beta}=(\beta_1,\cdots,\beta_P)^{\mathrm{T}}\in\mathbf{R}^P$。根据最优化理论，问题(3)等价于以下有约束形式：

$$\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\boldsymbol{\Phi}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$$
$$\boldsymbol{\Omega}(\boldsymbol{\beta})<m^* \tag{4}$$

可以使用交叉校验过程、BIC、AIC 或 C_p 准则选定 m^* 的值。组稀疏模型主要是使用各种损失函数和各种产生组稀疏模型向量的罚函数来构造组稀疏模型。组稀疏模型分为组套索模型和非凸罚组稀疏模型，其中组套索模型中包括普通组套索模型[1-11]、$L_{\infty,1}$ 组套索模型[12-17]、重叠组套索模型[18-21]、树组套索模型[22-25]、多输出树组套索模型[26,27]、混合组套索模型[28-30]、自适应组套索模型[31,32]、逻辑斯蒂组套索模型[33-35]和贝叶斯组套索模型[36,37]等，其中逻辑斯蒂组套索模型是关于逻辑斯蒂回归的组套索模型，其余都是关于线性回归模型的组套索模型。非凸罚组稀疏模型包括组 SCAD 罚模型[38-40,43]、组桥模型[40,41]和组 MC 罚模型[39,40,42,43]，它们的罚函数都是非凸的。组稀疏模型的具体分类如图 1 所示。

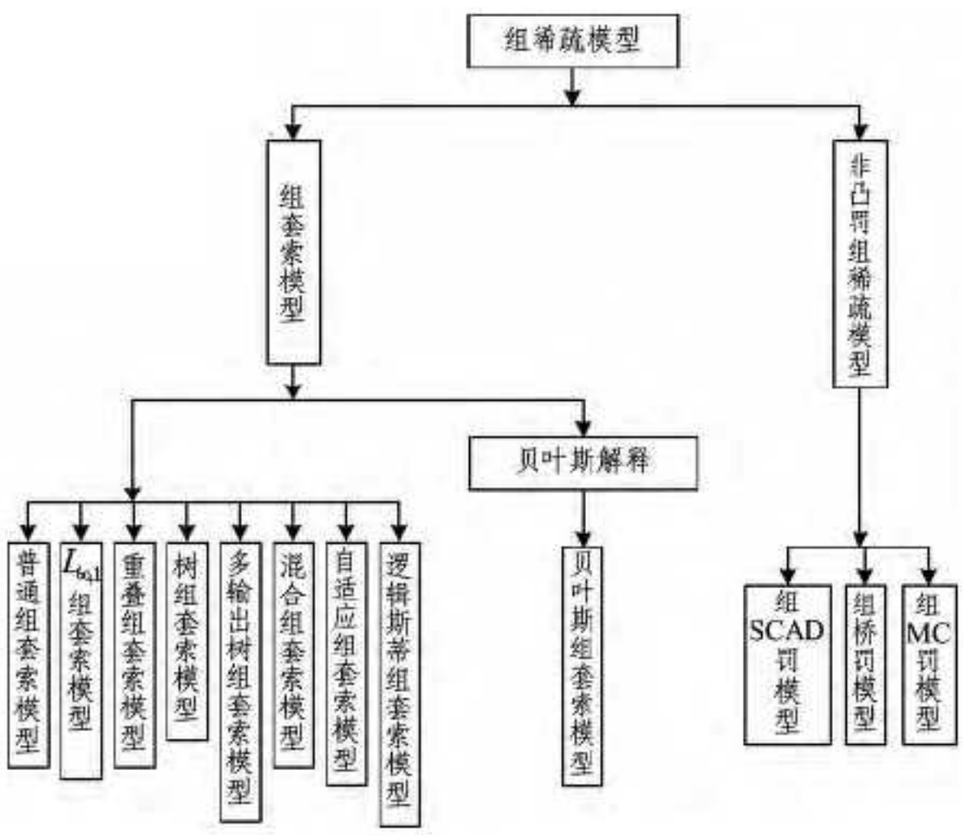

图 1 组稀疏模型

本文第 2 节介绍组稀疏模型，其中第 2.1 节介绍组套索模型及其变种，第 2.2 节介绍非凸罚组稀疏模型；第 3 节介绍组稀疏模型的算法实现，其中第 3.1 节介绍组套索模型及其各种变种形式的算法实现，第 3.2 节介绍非凸罚变量选择模型的算法实现；第 4 节给出了最新提出的一些组稀疏模型；第 5 节指出了组稀疏模型未来的发展方向；最后总结全文。

由于本文涉及的符号较多，为了清晰起见，在此说明本文中主要用到的部分符号所代表的含义。文中加波浪线变量、加星号的变量和不加波浪线且不加星号的变量均为完全不同的变量，例如 L^*、$\tilde{L}$ 和 L 为完全不同的 3 个变量。主要用到的符号的含义如表 1 所列。

表 1　主要符号说明

符号	含义
N	样本数，或对特征的观察次数
$p\in\{1,\cdots,P\}$	对全部特征尚未进行分组时某个特征的索引
$l\in\{1,\cdots,L\}$	对全部特征进行分组后的第 j 个组内的特征的索引
$j\in\{1,\cdots,J\}$	组的索引
$g_j\subseteq\{1,\cdots,P\}$	第 j 个组内特征的索引集
$G=\{g_j \mid j=1,\cdots,J\}$	全部组的索引集的集合
d_j	第 j 个组中特征的个数
P	特征变量的总数
$X_j\in R^{N\times d_j}$	第 j 个组对应的子设计矩阵
$x_p\in R^N$	由设计矩阵 X 中第 p 列元素组成的列向量（设计向量）
$y\in R^N$	输出向量
$\beta_j\in R^{d_j}$	第 j 个组对应的子模型向量
λ	权衡参数
$x^n\in R^N$	对 P 个特征变量的第 n 次观察值（第 n 个样本），即设计矩阵 X 中第 n 行元素组成的行向量

2　组稀疏模型

2.1　组套索模型及其变种

普通组套索模型和 $L_{\infty,1}$ 组套索模型都能实现特征组选择，它们的罚函数同属于 $L_{q,1}$ 范数罚形式，但是它们均不能处理组与组之间的特征存在重叠的情形。Jenatton 等人[18]和 Obozinski 等人[19]针对此问题提出了重叠组套索模型，重叠组套索模型适用于组与组之间包含的特征存在重叠的情形。有时特征可能具有树结构，由 Zhao 等人[22]最先提出的树组套索模型能够处理这种情形。Kim 等人[26]提出的多输出树组套索模型能够处理输出变量之间具有树结构关系的组稀疏模型。由于普通组套索模型只能实现特征组选择而不能实现组内的特征选择，Friedman 等人[28]针对此问题提出了混合组套索模型，混合组套索模型能同时实现特征组选择和组内的特征选择。普通组套索模型一般不具有 oracle 性质，因此 Wang 等人[31]提出了具有 oracle 性质的自适应组套索模型。组套索模型一般使用线性回归模型，Meier 等人[33]对普通组套索模型的损失函数进行改动，提出了逻辑斯蒂组套索模型。Roman[36]从贝叶斯理论的角度考虑组套索模型，指出只要选取适当的先验分布，组套索模型相当于贝叶斯最大后验估计，并提出了贝叶斯组套索模型。

2.1.1　普通组套索模型

2.1.1.1　普通组套索模型

已知线性回归模型为：

$$y=X\beta+\varepsilon \tag{5}$$

其中

$$X=\begin{pmatrix} X_{11} & \cdots & X_{1P} \\ \vdots & \ddots & \vdots \\ X_{N1} & \cdots & X_{NP} \end{pmatrix}\in R^{N\times P}$$

$$\beta=(\beta_1,\beta_2,\cdots,\beta_P)^{\mathrm{T}}\in R^P$$

$$y=(y_1,y_2,\cdots,y_N)^{\mathrm{T}}\in R^N$$

$\varepsilon\in R^N$ 且 $\varepsilon\sim N(0,\sigma^2 I)$

其中，N 为样本个数，$n\in\{1,2,\cdots,N\}$ 表示样本的下标。P 为特征个数，用 $p\in\{1,\cdots,P\}$ 表示特征的下标。

假定 P 个输入特征变量分为 J 个组子向量 $G=\{g_j \mid j=1,2,\cdots,J\}$，各组子向量中的变量互不相交，正交矩阵 X_j 为第 j 个组子向量对应的子设计矩阵，d_j 表示第 j 个组中特征变量的个数，$\beta\in R^P$ 中的 $\beta_j\in R^{d_j}$ 为第 j 个组 X_j 对应的子模型向量。则普通组套索模型为：

$$\hat{\beta}=\arg\min_{\beta\in R^P}\frac{1}{2}\|y-X\beta\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\beta_j\|_2 \tag{6}$$

其中，$\lambda\geqslant 0$，$\|\beta_j\|_2$ 表示第 j 个组对应的子模型向量 β_j 的 L_2 范数。λ 为权衡参数，用来权衡模型的稀疏性和预测准确性。罚函数项的含义为先对组 j 对应的子模型向量 β_j 进行 L_2 范数运算，再对由子模型向量 β_j（其中 $j\in\{1,\cdots,J\}$）的 L_2 范数值组成的向量求 L_1 范数，上述两个范数的结合 $\sum_{j=1}^{J}\|\beta_j\|_2$ 称作 $L_{2,1}$ 范数，有时简写为 $\|\beta\|_{2,1}$。显然，当每个组只具有一个特征时，普通组套索模型即退化为 Tibshirani 的套索模型。Huang 等人[3]给出了一种叫做强组稀疏性条件（Strong Group Sparsity Condition）的概念，强组稀疏性条件定义为：令 $supp(\beta)$ 表示系数向量 β 的支集，$\tilde{S}\subseteq\{1,2,\cdots,J\}$，$G_{\tilde{S}}=\bigcup_{j\in\tilde{S}}g_j$，若存在一个集合 $\tilde{S}$ 使得 β 满足：

$$supp(\beta)\subseteq G_{\tilde{S}},\ |G_{\tilde{S}}|\leqslant\eta_1,\ |\tilde{S}|\leqslant\eta_2$$

则称模型向量 β 是 (η_2,η_1)-强组稀疏的。Huang 等人还指出当数据满足强组稀疏性条件时普通组套索模型优于 Tibshirani 的套索模型，并指出这种优点体现在普通组套索模型的组结构使得其对噪声干扰具有更强的鲁棒性。

普通组套索模型的特征选择效果如图 2 所示，其中每个椭圆形中的特征对应一个组。图 2 中 14 个特征分为 5 组，每组中的特征都具有某种共性，因此属于各个组的特征应同时选择出来而不是单独选出某个组的某个特征。图 3 中，普通组套索模型选中了两个组 $\{v_3,v_4,v_5,v_6\}$ 和 $\{v_{10},v_{11}\}$，而组 $\{v_1,v_2\}$、$\{v_7,v_8,v_9\}$ 和 $\{v_{12},v_{13},v_{14}\}$ 都被丢弃了，这些被丢弃的组对应的子模型向量 β_j 均为 0。显然，普通组套索模型实现了特征组选择。

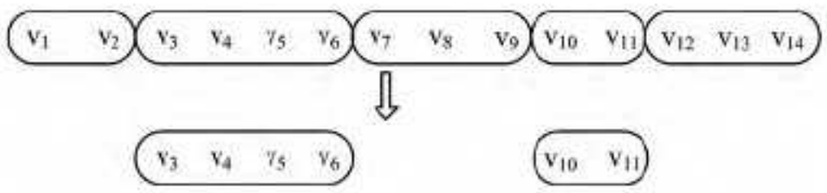

图 2　普通组套索模型的选择效果

2.1.1.2　普通组套索模型的参数估计一致性、模型选择一致性和 oracle 性质

(1)参数估计一致性

对于任意正数 ε^*，有 $\lim_{N\to\infty}P(\|\beta-\hat{\beta}\|_2<\varepsilon^*)=1$ 成立，其中 N 是样本的个数。

(2)模型选择一致性

$$\lim_{N\to\infty}P(\{j:\hat{\beta}_j\neq 0\}=\{j:\beta_j\neq 0\})=1$$

(3)oracle 性质

a)模型选择一致性[44,45]

$$\lim_{N\to\infty}P(\{j:\hat{\beta}_j\neq 0\}=\{j:\beta_j\neq 0\})=1$$

b)参数估计渐近正态性[44,45]

令 $A^*=\{j:\hat{\beta}_j\neq 0\}$，则当 $N\to\infty$ 时，$\sqrt{N}(\hat{\beta}_{A^*}-\beta_{A^*})\to N(0,\Sigma)$。

组套索模型的参数估计一致性、模型选择一致性和 oracle 性质的讨论需要附加假设条件，这些假设条件主要有不可表示条件[46]（irrepresentable condition）、稀疏 Riesz 条件[47]（sparse Riesz condition）和特征值限制条件[48]（restricted eigenvalue condition）等。

2.1.2 $L_{\infty,1}$ 组套索模型

$L_{\infty,1}$ 组套索模型目标函数为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_\infty \tag{7}$$

其中，$\lambda\geqslant 0$，$\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，d_j 表示第 j 个组中特征的个数。$L_{\infty,1}$ 组套索模型可推广为式(8)所示的 $L_{q,1}$ 组套索模型[49-51]：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_q \tag{8}$$

其中，$1\leqslant q\leqslant\infty$，将 $\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_q$ 称作 $L_{q,1}$ 范数。

2.1.3 重叠组套索模型

重叠组套索模型[18]为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2 \tag{9}$$

其中，$\lambda\geqslant 0$，w_{g_j} 表示预先设定好的关于组 g_j 的权，$g_j\subseteq\{1,\cdots,P\}$。$G=\{g_j\mid j=1,\cdots,J\}$ 表示全部组的索引集的集合，且 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$。注意：上述定义中 $g_j\subseteq\{1,\cdots,P\}$ 允许不同组之间的特征出现重叠。

2.1.4 树组套索模型

实际上，许多数据不但具有组结构，而且组之间存在偏序关系，即树结构。当处理这种数据时，需要充分利用树结构作为先验信息。已知共 P 个特征被分为 J 个组 $g_1,\cdots,g_J$，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示某组的索引集，用 $\boldsymbol{\beta}_{g_j}$ 表示组 g_j 对应的子模型向量，$G=\{g_j\mid j=1,\cdots,J\}$ 表示全部组的索引集的集合。G 具有树结构，即对于任意的两个组 $g_j\in G$ 和 $g_j^*\in G$，有 $(g_j\cap g_j^*\neq\varnothing)\Rightarrow(g_j\subseteq g_j^*\cup g_j^*\subseteq g_j)$ 和 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$ 成立。利用上述先验信息得到的树组套索模型[23]为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2 \tag{10}$$

其中，$\lambda\geqslant 0$，w_{g_j} 表示预先设定好的关于组 g_j 的权，G 具有式(10)和式(11)所定义的树结构。

2.1.5 多输出树组套索模型

已知多元线性回归模型为：

$$\begin{pmatrix}Y_{11}&\cdots&Y_{1K}\\ \vdots&\ddots&\vdots\\ Y_{N1}&\cdots&Y_{NK}\end{pmatrix}=\begin{pmatrix}X_{11}&\cdots&X_{1P}\\ \vdots&\ddots&\vdots\\ X_{N1}&\cdots&X_{NP}\end{pmatrix}\begin{pmatrix}\beta_{11}&\cdots&\beta_{1K}\\ \vdots&\ddots&\vdots\\ \beta_{P1}&\cdots&\beta_{PK}\end{pmatrix}+\boldsymbol{W} \tag{11}$$

其中，N 为样本数，P 为特征个数，K 为输出变量个数。$\boldsymbol{X}=\begin{pmatrix}X_{11}&\cdots&X_{1P}\\ \vdots&\ddots&\vdots\\ X_{N1}&\cdots&X_{NP}\end{pmatrix}$ 表示 $N\times P$ 阶的特征矩阵，$\boldsymbol{Y}=\begin{pmatrix}Y_{11}&\cdots&Y_{1K}\\ \vdots&\ddots&\vdots\\ Y_{N1}&\cdots&Y_{NK}\end{pmatrix}$ 表示 $N\times K$ 阶的输出变量矩阵，$\boldsymbol{B}=\begin{pmatrix}\beta_{11}&\cdots&\beta_{1K}\\ \vdots&\ddots&\vdots\\ \beta_{P1}&\cdots&\beta_{PK}\end{pmatrix}$ 为 $P\times K$ 阶的系数矩阵，$\boldsymbol{W}$ 表示噪声矩阵。

多输出树组套索模型假定输出变量之间满足树结构 T 关系。P 个输入特征变量分为 J 个组子特征变量（节点）$g_1,\cdots,g_J$，树的每一个节点表示一个子特征变量组，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示某子特征变量组的下标集合，$\boldsymbol{\beta}_{g_j}$ 表示子特征变量组（节点）g_j 对应的子模型向量，$G=\{g_j\mid j=1,\cdots,J\}$ 表示全部组（节点）的下标集合，且 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$。树的叶子节点对应多输出树组套索模型的单个输出变量，树的内部节点对应一组输出变量，且给每个节点（组）$g_j\in G$ 都赋予一个权 w_{g_j}。对于式(11)中的多元线性回归模型，树结构多输出树组套索模型[26]目标函数为：

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2 \tag{12}$$

其中，$\lambda\geqslant 0$，$\|\cdot\|_F$ 表示 Frobenius 范数，$\lambda\geqslant 0$，w_{g_j} 表示各子特征变量组的权。

式(12)的罚函数 $\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_g^p\|_2$ 包含两部分，内部 $L_{2,1}$ 范数 $\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}^p\|_2$ 和外部 L_1 范数 $\sum_{p\in\{1,\cdots,P\}}$。

2.1.6 混合组套索模型

套索导致的模型稀疏性是特征水平上的稀疏性，能够实现组内的特征选择；普通组套索模型导致的模型稀疏性是组水平上的稀疏性，能够实现组内的特征选择。然而很多时候既希望将重要的特征组选择出来，又希望将特征组中重要的特征选择出来，普通组套索模型显然不能满足这种要求。如果将 L_1 范数罚添加到普通组套索模型的目标函数中，那么既可以实现特征组水平上的模型稀疏性，又可以实现特征水平上的模型稀疏性，即同时实现特征组选择和组内的特征选择，将这种组套索模型叫做混合组套索模型。已知式(5)中的线性回归模型，令 $j\in\{1,\cdots,J\}$ 表示特征组的索引，$p\in\{1,\cdots,P\}$ 表示特征的索引，则混合组套索模型[28]为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2+\lambda_2\|\boldsymbol{\beta}\|_1 \tag{13}$$

其中，$\lambda_1\geqslant 0$，$\lambda_2\geqslant 0$。$\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，$\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$ 的作用为实现特征组选择；$\lambda_2\|\boldsymbol{\beta}\|_1$ 的作用为实现组内的特征选择。混合组套索模型的变量选择效果如图 3 所示，可以看出，其同时实现了特征水平和特征组水平上的稀疏性。实际上混合组套索模型是树组套索模型的一个特例，混合组套索模型相当于 2.1.4 节中树结构为两层时的树组套索模型，$\lambda_1\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$ 对应根节点即一组特征，$\lambda_2\|\boldsymbol{\beta}\|_1$ 对应叶子节点即单个特征。Chatterjee 等人[30]指出混合组套索模型在满足一定条件时具有参数估计一致性，并给出了满足参数估计一致性所需要的假设条件。

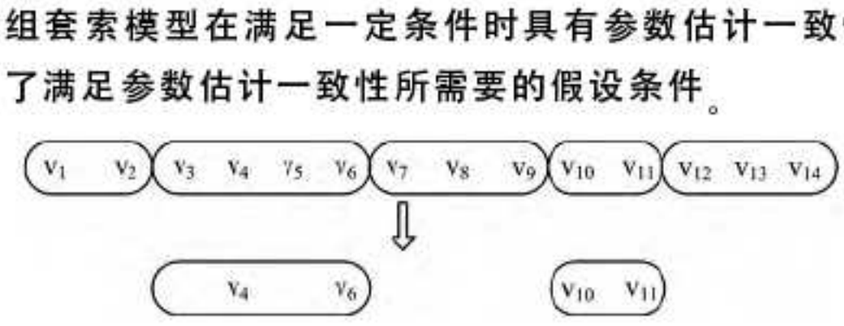

图 3 混合组套索模型的选择效果

2.1.7 自适应组套索模型

由于普通组套索模型对于子模型向量 $\boldsymbol{\beta}_j$ 的惩罚程度不随着子模型向量 $\boldsymbol{\beta}_j$ 的模的大小而自适应地改变，因此会对模较大的子模型向量进行过度的缩小，导致对模较大的子模型

向量的有偏估计。因此，若要令普通组套索模型具有 oracle 性质，一个非常自然的想法就是在罚函数中为不同组对应的子模型向量 $\boldsymbol{\beta}_j$ 合理分配不同的权，令各个子模型向量 $\boldsymbol{\beta}_j$ 所受到的惩罚程度具有“自适应”的性质，对模较大的子模型向量 $\boldsymbol{\beta}_j$ 执行较小程度的惩罚（即在罚函数中给子模型向量 $\boldsymbol{\beta}_j$ 分配较小的权），而对模较小的子模型向量 $\boldsymbol{\beta}_j$ 执行较大程度的惩罚（即在罚函数中给子模型向量 $\boldsymbol{\beta}_j$ 分配较大的权）。自适应组套索就是针对这一问题提出的。已知式(5)中的线性回归模型，自适应组套索模型[31,32]为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}w_j\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \tag{14}$$

其中，$\lambda\geqslant 0$，$\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，$w_j=\dfrac{1}{\|\hat{\boldsymbol{\beta}}_j(GL)\|_2}$ 表示第 j 个组的权，$\hat{\boldsymbol{\beta}}_j(GL)$ 为普通组套索模型

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \tag{15}$$

的解。

显然，自适应组套索模型实际上是两步的普通组套索模型：1)先求解普通组套索模型，得到的解 $\hat{\boldsymbol{\beta}}_j^{GL}$ 作为各组罚函数的权 $w_j=\|\hat{\boldsymbol{\beta}}_j^{GL}\|_2^{-1}$，其中 $j=1,\cdots,J$；2)将权 w_j 代入式(14)中，求解式(14)，这一步就相当于对第一步求得的解进行修正。第一步由普通组套索模型得到各个组对应的子模型向量的初始估计后，把各个组对应的子模型向量的模的倒数作为各个组的权显然是合理的，如此便可实现对模较大的子模型向量执行较小程度的惩罚，而对模较小的子模型向量执行较大程度的惩罚。由于重新对各组的权进行了合理调整，因此自适应组套索模型克服了普通组套索模型一般不具有 oracle 性质的缺点。权 w_j 还可以有其它的分配方法，但对于第一步中的解的要求是其必须具有参数估计一致性，但不要求具有模型选择一致性。显然，普通组套索模型在某些假设条件下（例如 Wei 等人[5]论文中的定理 2.1）的解满足参数估计一致性。若令普通最小二乘估计的解 $\hat{\boldsymbol{\beta}}_j^{OLS}$ 的模的倒数作为权 $w_j=\dfrac{1}{\|\hat{\boldsymbol{\beta}}_j^{OLS}\|_2}$，则其只适用于当 $P<N$（即参数空间维数小于样本空间维数）时的情形，这是因为当 $P>N$（参数空间维数大于样本空间维数）时普通最小二乘估计不具有参数估计一致性。而且，由于普通组套索模型的解是稀疏的，因此第一步采用普通组套索模型的解还可以降低第二步处理时的维数，而采用普通最小二乘估计的解则没有此效果。另外，当每个组只含有一个特征时，自适应组套索模型就退化为自适应套索模型[44]。

2.1.8 逻辑斯蒂组套索模型

逻辑斯蒂回归模型的组套索模型[33]为：

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\Xi(\beta)\\&=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\{-\Phi(\boldsymbol{\beta})+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2\}\end{aligned} \tag{16}$$

其中，$\lambda\geqslant 0$，$\boldsymbol{\beta}_j$ 为第 j 个组的子模型向量，罚函数 $\Phi(\boldsymbol{\beta})$ 为：

$$\Phi(\boldsymbol{\beta})=\sum_{n=1}^{N}\boldsymbol{y}_n\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta}-\log[1+\exp(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta})]$$

逻辑斯蒂组套索模型用于两类和多类分类器设计的特变量组选择。

2.1.9 贝叶斯组套索模型

贝叶斯理论认为，当普通组套索模型的每个组对应的子模型向量都具有式(17)—式(19)中的独立同分布的多维 Laplace 先验分布时，可以把式(6)中的普通组套索模型表示为如下贝叶斯最大后验估计问题：

$$P(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2)=\prod_{n=1}^{N}N(\boldsymbol{y}_n|\boldsymbol{x}^n\boldsymbol{\beta}^{\mathrm{T}},\sigma^2) \tag{17}$$

$$P(\boldsymbol{\beta}_j|\rho)=\mathrm{Laplace}(\boldsymbol{\beta}|0,(\frac{d_j\rho}{\sigma^2})^{-\frac{1}{2}}) \tag{18}$$

$$P(\sigma^2|v_0,s_0^2)=\mathrm{InvGamma}(\sigma^2|v_0,s_0^2) \tag{19}$$

$$P(\rho|r,s)=\mathrm{Gamma}(\rho|r,s) \tag{20}$$

其中，$\boldsymbol{x}^n\in\mathbf{R}^P$ 表示 P 个特征变量在第 $n\in\{1,\cdots,N\}$ 次的观察值即第 $n\in\{1,\cdots,N\}$ 个样本，$\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，d_j 表示第 j 个组中特征的个数。另外，式(17)—式(19)中 β 的多维 Laplace 先验分布可以表示为如下正态分布和 Gamma 混合分布的层次模型：

$$\begin{aligned}P(\boldsymbol{\beta}_j|\rho)&=\mathrm{Laplace}(\boldsymbol{\beta}|0,(\frac{d_j\rho}{\sigma^2})^{-\frac{1}{2}})\\&\propto(\frac{d_j\rho}{\sigma^2})^{-\frac{d_j}{2}}\exp(-(\frac{d_j\rho}{\sigma^2})^{-\frac{1}{2}}\|\boldsymbol{\beta}_j\|_2)\\&\propto\int_0^{\infty}\mathbf{N}(\boldsymbol{\beta}_i|0,\sigma^2\tau_i^2)\mathrm{Gamma}(\tau_i^2|\frac{d_j+1}{2},\frac{2}{d_j\rho})\mathrm{d}\tau_i^2\end{aligned} \tag{21}$$

因此，贝叶斯组套索模型[36]便具有如图 4 所示的层次表示形式。

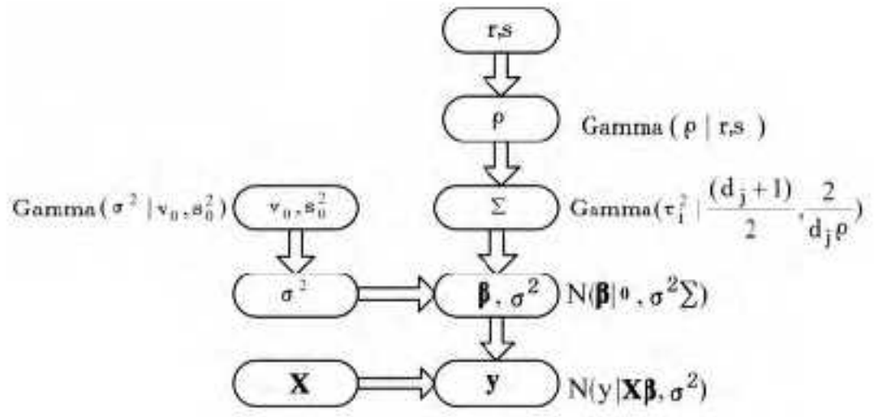

图 4 贝叶斯组套索模型的层次模型表示形式

图 4 中 Σ 为对角矩阵，其对角元素为 τ_i^2。r 和 s 为 Gamma 分布 ρ 的超参数。普通组套索模型不能对回归系数向量的协方差矩阵进行估计，而由式(17)—式(21)可以看出贝叶斯组套索模型可以方便地对其进行估计，这是贝叶斯组套索模型的显著优点。可以使用吉布斯抽样方法得到贝叶斯组套索模型中的参数估计，贝叶斯组套索模型的缺点是对高维数据进行估计时算法复杂性过高。

2.2 非凸罚组稀疏模型

SCAD 罚模型[52]（Smoothly Clipped Absolute Deviation Penalty）和 MC 罚模型[52]（Minimax Concave Penalty）均为非凸罚函数的特征选择模型，它们的提出是为了克服套索模型不具有 oracle 性质的缺点，但只能进行特征选择，而不能进行特征组选择。因此，自然会联想到将 SCAD 罚模型和 MC 罚模型推广为组 SCAD 罚模型[38-40,43]（Group Smoothly Clipped Absolute Deviation Penalty）和组 MC 罚模型[39,40,42,43]（Group Minimax Concave Penalty），以便克服组套索模型不具有 oracle 性质的缺点，同时实现特征组选择和组内的特征选择。另外，桥模型[53]（Bridge Penalty）也是非凸的罚函数，将其推广为组桥模型[40,41]可以同时实现特征组选择和组内的特征选择，并且具有 oracle 性质。

2.2.1 组 SCAD 罚模型

组 SCAD 罚模型分为 L_1 范数组 SCAD 罚模型[39]和 L_2 范数组 SCAD 罚模型[38]。已知式(5)中的线性回归模型，L_1 范数组 SCAD 罚模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\phi_\lambda(\|\boldsymbol{\beta}_j\|_1) \tag{22}$$

L_2 范数组 SCAD 罚模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\phi_\lambda(\|\boldsymbol{\beta}_j\|_2) \tag{23}$$

其中，式(22)和式(23)中 $\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，且

$$\phi_\lambda(\theta)=\begin{cases}\lambda\theta, & 0<\theta\leqslant\lambda\\ -\dfrac{(\theta^2-2\gamma\lambda\theta+\lambda^2)}{2(\gamma-1)}, & \lambda<\theta<\gamma\lambda\\ \dfrac{(\gamma+1)\lambda^2}{2}, & \theta>\gamma\lambda\end{cases} \tag{24}$$

其中，$\gamma>2,\lambda\geqslant 0$。显然，$L_1$ 范数组 SCAD 罚模型和 L_2 范数组 SCAD 罚模型的不同之处在于 L_1 范数组 SCAD 罚模型的罚函数内部为 L_1 范数罚，而 L_2 范数组 SCAD 罚模型的罚函数内部为 L_2 范数罚。L_1 范数组 SCAD 罚模型能同时实现特征组选择和组内的特征选择，而 L_2 范数组 SCAD 罚模型只能实现特征组选择。L_2 范数组 SCAD 罚模型具有 oracle 性质。另外，由式(24)可以看出，当 $\gamma\to\infty$ 时式(24)趋向于 $\phi_\lambda(\theta)=\lambda\theta$，因此式(23)即 L_2 范数组 SCAD 罚模型在 $\gamma\to\infty$ 时趋向于普通组套索模型。

2.2.2 组桥模型

套索只能实现组内的特征选择，普通组套索模型只能实现特征组选择，而组桥模型既能实现组内的特征选择，又能实现特征组选择，而且还具有 oracle 性质。使用式(5)中的线性回归模型，组桥模型[41]为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}c_j\|\boldsymbol{\beta}_j\|_1^\gamma \tag{25}$$

其中，$\lambda\geqslant 0,0<\gamma<1,\boldsymbol{\beta}_j$ 表示第 $j\in\{1,\cdots,J\}$ 个组对应的子模型向量，c_j 的值可以取为 $d_j^{1-\gamma}$。当每个组只含有一个特征时，组桥模型退化为桥模型[53]。

2.2.3 组 MC 罚模型

将 MC 罚模型推广到组变量情形中，将得到复合组 MC 罚模型[42]、L_1 范数组 MC 罚模型[39]和 L_2 范数组 MC 罚模型[40]。使用式(5)中的线性回归模型，复合组 MC 罚模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\varphi_{\lambda,b}\left(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\beta_{jl}|)\right) \tag{26}$$

其中，$\lambda\geqslant 0,\beta_{jl}$ 表示第 j 个组中第 l 个特征的模型分量，$b=\dfrac{d_j a\lambda}{2},a>1,b>1$。

L_1 范数组 MC 罚模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_1) \tag{27}$$

其中，$\lambda\geqslant 0,\gamma>1$。

L_2 范数组 MC 罚模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2) \tag{28}$$

其中，$\lambda\geqslant 0,\gamma>1$。

式(26)—式(28)中

$$\varphi_{\lambda,\gamma}(\theta)=\begin{cases}\lambda\theta-\dfrac{\theta^2}{2\gamma}, & 0<\theta<\gamma\lambda\\ \dfrac{1}{2}\gamma\lambda^2, & \theta>\gamma\lambda\end{cases} \tag{29}$$

显然，三者的不同之处在于：复合组 MC 罚模型的复合罚函数内部函数和外部函数均采用了 MC 罚；而 L_1 范数组 MC 罚模型只在罚函数外部函数使用了 MC 罚，其内部函数为 L_1 范数罚；L_2 范数组 MC 罚模型也只在外部函数使用了 MC 罚，其内部函数为 L_2 范数罚。另外，复合组 MC 罚模型和 L_1 范数组 MC 罚模型能同时实现特征组选择和组内的特征选择，但 L_2 范数组 MC 罚模型只能实现特征组选择。L_2 范数组 MC 罚模型具有 oracle 性质。另外，由式(29)可以看出，当 $\gamma\to\infty$ 时式(29)趋向于 $\varphi_\lambda(\theta)=\lambda\theta$，因此式(28)即 L_2 范数组 MC 罚模型在 $\gamma\to\infty$ 时趋向于普通组套索模型。

3 组稀疏模型的算法实现

3.1 组套索模型的算法实现

组稀疏模型的求解大致分为两个步骤：1)对目标函数进行预处理。将不平滑、非凸、块坐标不可分离的组稀疏模型的目标函数向平滑、凸、块坐标可分离的方向转化，这一步常利用的技巧有变分不等式、Nesterov 提出的平滑近似技巧[54]、局部二次近似(local quadratic approximation)、对偶范数和对偶函数等。2)利用组最小角回归算法[2](Group Least Angle Regression，Group LARS)、块坐标下降(上升)算法、投影梯度算法(Projected Gradient algorithm)、谱投影梯度算法(Spectral Projected Gradient algorithm)、活动集算法(active set algrithm)、轮换方向乘子法[55](Alternating Direction Method of Multipliers，ADMM)或块坐标梯度下降算法[60](block coordinate gradient descent method)等对转换后的目标函数进行求解。

求解组套索模型的各种算法都具有自身的特点和适用条件。组最小角回归算法只适用于解路径为分段线性的组套索模型，这是因为只有当组套索模型的解路径为分段线性时，由组最小角回归算法所求得的解路径才是组套索模型的解路径。块坐标下降(上升)算法只适用于目标函数为块坐标可分离的组套索模型，像重叠组套索和树组套索这样的块坐标不可分离的组套索模型不能利用块坐标下降(上升)算法直接求解。其中函数块坐标可分离指的是函数可以被写作不同坐标块(坐标向量) $\boldsymbol{\beta}_j$ 对应的函数 $f_j(\boldsymbol{\beta}_j)$ 之间的独立相加关系 $f(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_J)=\sum_{j=1}^{J}f_j(\boldsymbol{\beta}_j)$，且不同坐标块(坐标向量) $\boldsymbol{\beta}_j$ 之间不包含共同的自变量。例如式(6)中的普通组套索模型的目标函数就是块坐标可分离的，而式(9)中重叠组套索模型的目标函数中不同组对应的子模型向量 $\boldsymbol{\beta}_{g_j}$ 之间存在共同的自变量，因此是块坐标不可分离的，所以不能利用块坐标下降(上升)算法求解重叠组套索模型。与重叠组套索模型类似，树组套索模型和多输出树组套索模型的目标函数中不同组对应的子模型向量之间也存在共同的自变量，也是块坐标不可分离的，因此也不能利用块坐标下降算法直接对其进行求解。投影梯度算法中的投影算子的计算复杂度很大，故在应用这种算法时往往需要找到高效计算投影算子的方法。另外，投影梯度算法中的梯度下降步骤一般会使得收敛速度较慢，谱投影梯度算法将投影梯度算法的收敛速度进行了改进，加快了收敛速度。活动集算法一般用于大规模复杂问题的求解，一般利用最优性条件如 KKT 条件把大规模复杂问题分解为一系列简单子问题的求解，一般是用时间复杂性增高换取空间复杂性的降低，但也可能在算法迭代过程中因满足最优性条件而提前结束，从而实现同时降低算法的时空复杂性。轮

换方向乘子法为增广拉格朗日方法的推广，它通过引入辅助变量将原问题分解为多个容易求解的子问题。轮换方向乘子法的适用范围很广，它可以用来求解套索模型、普通组套索模型以及重叠组套索模型等。与块坐标下降算法类似，块坐标梯度下降算法也只适用于目标函数块坐标可分离的组套索模型。最后，组稀疏模型权衡参数 λ 的选择可用 Cp 判据、BIC 判据、AIC 判据、GCV 准则和交叉校验等方法确定。

3.1.1 普通组套索模型的算法实现

普通组套索模型的算法实现主要有组最小角回归算法、块坐标下降算法、活动集算法和轮换方向乘子法等。

Efron 等人[56]提出最小角回归算法来进行变量选择。已知线性回归模型为：

$$y=\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_P x_P+\tilde{r}$$

模型向量为 $\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_P)^{\mathrm{T}}$，$\tilde{r}$ 表示残差，其中共 P 个特征 $x_1,x_2,\cdots,x_P$，不妨假设最后选出的特征为 3 个，下面阐述最小角回归算法的原理：开始时判断全部特征 $x_1,x_2,\cdots,x_P$ 中谁与输出 y 之间的夹角最小，不妨假设判断结果为 $x_{\tilde{a}}$ 与输出 y 之间的夹角最小，然后就在 $x_{\tilde{a}}$ 的方向上前进（即增大 $\hat{u}_1=\hat{\gamma}_{\tilde{a}}x_{\tilde{a}}$ 中 $x_{\tilde{a}}$ 对应的系数 $\hat{\gamma}_{\tilde{a}}$），直到另一个特征（不妨假设这个特征为 $x_{\tilde{b}}$）与当前残差向量 $\tilde{r}_1=\bar{y}_2-\hat{u}_1$ 的夹角等于 $x_{\tilde{a}}$ 与当前残差向量 $\tilde{r}_1$ 的夹角为止，那么此时残差向量 $\tilde{r}_1$ 的方向即为 $x_{\tilde{a}}$ 与 $x_{\tilde{b}}$ 的角平分线 $x_{\tilde{a}}+x_{\tilde{b}}$ 的方向，$\bar{y}_2$ 表示输出 y 到由 $x_{\tilde{a}}$ 与 $x_{\tilde{b}}$ 所张成的空间的投影。然后在残差向量 $\tilde{r}_1$ 的方向上前进（即增大 $\hat{u}_2=\hat{u}_1+\hat{\gamma}_2 i_2$ 中的 $\hat{\gamma}_2$，其中 i_2 为单位向量 $\frac{x_{\tilde{a}}+x_{\tilde{b}}}{\|x_{\tilde{a}}+x_{\tilde{b}}\|_2}$），直到另一个特征（不妨假设这个特征为 $x_{\tilde{c}}$）与当前残差向量 $\tilde{r}_2=\bar{y}_3-\hat{u}_2$ 的夹角等于 $x_{\tilde{a}}$ 与 $x_{\tilde{b}}$ 各自与当前残差向量 $\tilde{r}_2$ 的夹角为止，此时的残差向量 $\tilde{r}_2$ 的方向即为 $x_{\tilde{a}}$、$x_{\tilde{b}}$ 和 $x_{\tilde{c}}$ 的空间角平分线 $x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}$ 的方向，其中 $\bar{y}_3$ 表示输出 y 到由 $x_{\tilde{a}}$、$x_{\tilde{b}}$ 和 $x_{\tilde{c}}$ 所张成的空间的投影。然后在残差向量 $\tilde{r}_2$ 的方向上前进（即增大 $\hat{u}_2+\hat{\gamma}_3 i_3$ 中的 $\hat{\gamma}_3$，其中 i_3 为单位向量 $\frac{x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}}{\|x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}\|_2}$）直到 $\hat{u}_2+\hat{\gamma}_3 i_3=\bar{y}_3$ 成立为止。由上面过程中每一步得到的 $\hat{\gamma}_1$、$\hat{\gamma}_2$ 和 $\hat{\gamma}_3$ 即可得到模型向量 $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3)^{\mathrm{T}}$：

$$\hat{\beta}_1=\hat{\gamma}_1+\frac{\hat{\gamma}_2}{\|x_{\tilde{a}}+x_{\tilde{b}}\|_2}+\frac{\hat{\gamma}_3}{\|x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}\|_2}$$

$$\hat{\beta}_2=\frac{\hat{\gamma}_2}{\|x_{\tilde{a}}+x_{\tilde{b}}\|_2}+\frac{\hat{\gamma}_3}{\|x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}\|_2}$$

$$\hat{\beta}_3=\frac{\hat{\gamma}_3}{\|x_{\tilde{a}}+x_{\tilde{b}}+x_{\tilde{c}}\|_2}$$

以被选出的特征数为 3 为例的最小角回归算法的空间几何解释如图 5 所示。当被选出的特征多于 3 个时，算法的原理与上述被选出的特征数为 3 时的原理相同，第 4 个特征后的算法过程类似于上述前 3 个特征的算法过程，向后依次类推即可。Yuan 等人[2]将上述最小角回归算法（LARS）推广为组最小角回归算法（Group LARS）来求解普通组套索模型。组最小角回归算法与最小角回归算法的原理类似，不同的是需要将上述算法中的“特征 x_p”替换为“特征组 X_j”以及将“残差向量 $\tilde{r}$ 与特征 x_p 的夹角”变为“残差向量 $\tilde{r}$ 与特征值 X_j 的夹角”等，并且此时残差向量 $\tilde{r}$ 与 X_j 之间夹角 α^* 的大小用 $\cos^2\alpha^*=\frac{\|X_j^{\mathrm{T}}\tilde{r}\|^2}{d_j\|\tilde{r}\|^2}$ 来衡量。当设计矩阵 X 为正交矩阵时普通组套索模型的解路径是分段线性的，因此可以利用上述组最小角回归算法求解普通组套索模型。

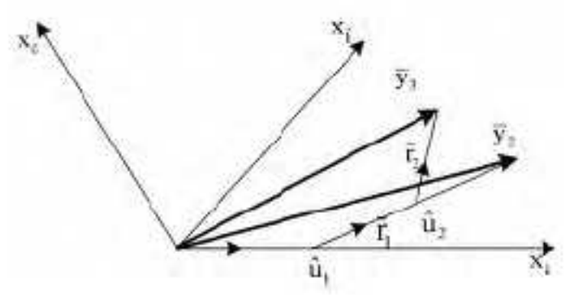

图 5 最小角回归算法（LARS）的空间几何解释（以被选出的特征数为 3 为例）

由于普通组套索模型的目标函数是块坐标可分离的，因此可以利用块坐标下降算法对其求解。Yuan 等人[2]沿用文献[57]中 shooting 算法的思想给出了一种块坐标下降算法来求解普通组套索模型，这种算法也要求设计矩阵 X 为正交矩阵。Yuan 等人指出，对于普通组套索模型：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|y-X\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \tag{30}$$

来说，利用 KKT 条件可以得到其最优解 $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_J^{\mathrm{T}})^{\mathrm{T}}$ 的取值情况：

1) $\|-X_j^{\mathrm{T}}(y-X\boldsymbol{\beta})\|_2\leqslant\lambda\sqrt{d_j}$ 时，$\boldsymbol{\beta}_j=0$；

2) $-X_j^{\mathrm{T}}(y-X\boldsymbol{\beta})+\frac{\lambda\sqrt{d_j}\boldsymbol{\beta}_j}{\|\boldsymbol{\beta}_j\|_2}=0$，$\boldsymbol{\beta}_j\neq0$；

综合这两种情况，又由于 $X_j^{\mathrm{T}}X_j=I$，则普通组套索模型解的表达式为：

$$\boldsymbol{\beta}_j=\left(1-\frac{\lambda\sqrt{d_j}}{\|X_j^{\mathrm{T}}(y-X\boldsymbol{\beta}_{(-j)})\|_2}\right)+X_j^{\mathrm{T}}(y-X\boldsymbol{\beta}_{(-j)}) \tag{31}$$

其中，$\boldsymbol{\beta}_{(-j)}=(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_{j-1}^{\mathrm{T}},0,\boldsymbol{\beta}_{j+1}^{\mathrm{T}},\cdots,\boldsymbol{\beta}_J^{\mathrm{T}})$。对于每个组 $j=1,\cdots,J$，依次选择不同组索引 $j=1,\cdots,J$ 迭代求解上式，直到满足终止条件即可得到 $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_J^{\mathrm{T}})^{\mathrm{T}}$ 的最优解。显然，这是一种块坐标下降算法，其中一个组对应一个“坐标块”。块坐标下降算法在 $j=1,\cdots,J$ 的执行过程中每次都固定除第 j 个组之外的所有其它组对应的子模型向量的值，只把第 j 个组子模型向量作为未知变量，进行迭代求解，如算法 1 所示。

算法 1 普通组套索模型的块坐标下降算法

1. 初始化模型参数向量 $\boldsymbol{\beta}$

2. for j=1,…,J

若 $\|-\mathbf{X}_j^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\|\leqslant\lambda\sqrt{d_j}$，则 $\boldsymbol{\beta}_j\leftarrow0$；

否则，令

$$\boldsymbol{\beta}_j\leftarrow\left(1-\frac{\lambda\sqrt{d_j}}{\|\mathbf{X}_j^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{(-j)})\|_2}\right)+\mathbf{X}_j^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_{(-j)})$$

其中

$$\boldsymbol{\beta}_{(-j)}=(\boldsymbol{\beta}_1^{\mathrm{T}},\cdots,\boldsymbol{\beta}_{j-1}^{\mathrm{T}},0,\boldsymbol{\beta}_{j+1}^{\mathrm{T}},\cdots,\boldsymbol{\beta}_J^{\mathrm{T}})$$

end

3. 重复第 2 步直到满足预先设定的收敛条件

可以利用活动集算法[8]求解普通组套索模型。活动集算法从只含一个组的活动集开始，不断根据违反最优性条件最大原则依次增加新的组进入活动集，直到所有的组均满足最优性条件为止。求解普通组套索的活动集算法如算法 2 所示。

算法 2　普通组套索模型的活动集算法

1. 初始化活动集 ACT

2. 在当前活动集上求解普通组套索模型，即求解以下问题：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$$

$$\text{s.t. } \boldsymbol{\beta}_j=0, j\in \mathrm{ACT}^{\mathrm{C}}$$

其中 $\mathrm{ACT}^{\mathrm{C}}$ 表示活动集 ACT 的补集。得到在当前活动集上的解 $\boldsymbol{\beta}^*$，并更新活动集 $\mathrm{ACT}=\{j\in\mathrm{ACT}:\boldsymbol{\beta}_j^*\neq 0\}$。

3. 判断在当前活动集上的解 $\boldsymbol{\beta}^*$ 是否满足条件

$$\Lambda^*(\nabla\Phi(\boldsymbol{\beta}^*))\leqslant\lambda$$

其中，$\Phi(\boldsymbol{\beta}^*)=\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2$，$\Lambda^*$ 表示对偶范数。若满足，则表明 $\boldsymbol{\beta}^*$ 为普通组套索模型：$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$ 的最优解；否则，将 $\mathrm{ACT}^{\mathrm{C}}$ 中违反条件 $\Lambda^*(\nabla f(\boldsymbol{\beta}^*))\leqslant\lambda$ 程度最大的组添加到活动集 ACT 中。转到第 2 步。

轮换方向乘子法也可用于求解普通组套索模型。对于普通组套索模型：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$$

引入辅助向量 z 得普通组套索模型有约束的形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{z}_j\|_2$$

$$\text{s.t. } \boldsymbol{z}_j=\boldsymbol{\beta}_j, j\in\{1,\cdots,J\}$$

上述约束形式的增广拉格朗日形式为：

$$L_\rho(\boldsymbol{\beta},z,\xi)=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{z}_j\|_2+\sum_{j=1}^{J}\xi_j(\boldsymbol{z}_j-\boldsymbol{\beta}_j)+\frac{\rho}{2}\sum_{j=1}^{J}\|\boldsymbol{z}_j-\boldsymbol{\beta}_j\|^2$$

然后利用轮换方向法：

$$\boldsymbol{\beta}^{\tilde{k}+\tilde{1}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}L_\rho(\boldsymbol{\beta},z,\xi^{\tilde{k}})$$

$$z^{\tilde{k}+1}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}L_\rho(\beta^{\tilde{k}+1},z,\xi^{\tilde{k}})$$

$$\xi_j^{\tilde{k}+1}=\xi_j^{\tilde{k}}+\rho(z_j^{\tilde{k}+1}-\beta^{\tilde{k}+1})$$

进行求解。

另外，对于损失函数任意且不要求设计矩阵为正交矩阵的普通组套索模型来说，在利用块坐标下降算法对其求解的难点在于其第 j 个组的优化问题 $\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$ 的解没有显式形式。Yang 等人[9]针对此问题给出了一种叫做 QM 条件的假设，他们指出当不要求设计矩阵为正交矩阵的普通组套索模型满足此假设时其解具有显式形式，并给出了此时解的显式形式，然后利用块坐标下降算法求解。他们还给出了满足 QM 条件的一些损失函数，例如平方损失函数、Huberized 铰链损失函数、平方铰链损失函数和逻辑斯蒂损失函数等。Qin 等人[10]针对参数空间维数较小的情况给出了一种求解普通组套索模型的新的块坐标下降算法，他们将每个“块”对应的优化问题转化为求解一个信赖域子问题，并且利用牛顿方法求解此信任域子问题。但是当参数空间维数较大时牛顿方法的计算复杂度较大，因此 Qin 等人将块坐标下降算法中独立求解各特征组对应的优化问题的思想应用到迭代收缩阈值算法中，令每个特征组对应的梯度下降步骤具有自己独立的步长，独立地求解每个组对应的梯度下降问题。Rakotomamonjy[11]利用 Landweber 迭代算法[58]求解普通组套索模型，并在计算复杂度方面比较了求解普通组套索模型的块坐标下降法和 Landweber 迭代算法，指出这两种算法的计算复杂度虽然大致相等，但块坐标算法比 Landweber 迭代算法更快。

3.1.2　$L_{\infty,1}$ 组套索模型的算法实现

$L_{\infty,1}$ 组套索模型的求解算法有内点算法、投影梯度算法和谱投影梯度算法。

Turlach 等人[12]将 $L_{\infty,1}$ 组套索模型的目标函数转化为一个凸二次规划问题，并提出用内点算法对其求解，但是这种方法要求计算目标函数的 Hessian 矩阵，因此计算复杂度和执行所需的存储空间往往很大。Quattoni 等人[14]指出当损失函数采用铰链损失函数的特殊情况时，可以将原 $L_{\infty,1}$ 范数的正则化问题表示为一个线性规划形式，但这只在特征数目较小的情形下可行。Vogt 等人[17]利用活动集算法求解 $L_{\infty,1}$ 组套索模型，其原理与 3.1.1 节中所述的活动集算法相同，不再赘述。

另外，可以利用投影梯度算法（projected gradient method）求解 $L_{\infty,1}$ 组套索模型。已知凸函数 $f(v)$，C 为闭凸集，令 Π_C 表示 v 到 C 上的投影即投影算子。当为欧氏投影时，投影算子为：

$$\Pi_C(v)=\arg\min_{u\in C}\|u-v\|_2^2$$

则投影梯度算法迭代公式为：

$$v_{\tilde{k}+1}=\Pi_C(v_{\tilde{k}}-\tilde{t}_{\tilde{k}}\nabla f(v))$$

其中，$\tilde{t}_{\tilde{k}}$ 为步长。但是投影梯度算法存在两方面的问题：1）每次选择最快下降方向会导致收敛速度变慢；2）投影步骤的计算复杂度过高。Schmidt 等人[15]和 Quattoni 等人[16]分别针对上述两个问题对投影梯度算法进行了改进。针对上述第一个难题，Schmidt 等人利用如下的谱投影梯度算法来加速算法的收敛。谱投影梯度法在投影梯度法的基础上进行了两项改变：1）梯度迭代过程 $v_{\tilde{k}+1}=v_{\tilde{k}}-\tilde{\lambda}_{\tilde{k}}\tilde{d}_{\tilde{k}}$ 的步长 $\tilde{\lambda}_{\tilde{k}}$ 选择过程采用了非单调线搜索技术，即每次迭代后目标函数不一定要减小，只要求在最近规定的某些次迭代目标函数减小即可；2）利用谱梯度法确定投影步骤的步长 $\tilde{t}_{\tilde{k}}$。Quattoni 等人则针对上述第二个难题给出了计算到 $L_{\infty,1}$ 范数球投影的高效方法。关于式(11)中多元线性回归模型的多输出 $L_{\infty,1}$ 组套索模型为：

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda\|\boldsymbol{B}\|_{L_1/L_\infty}$$

$$=\arg\min_{\boldsymbol{B}}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda\sum_{p=1}^{P}\|\boldsymbol{\beta}_p\|_\infty$$

不妨假设某矩阵 $\boldsymbol{B}$ 的元素均为正，令 Θ 表示 $L_{\infty,1}$ 范数球，则求解系数矩阵 $\boldsymbol{B}$ 到 $L_{\infty,1}$ 范数球 Θ 的投影 $\boldsymbol{B}'$

$$\boldsymbol{B}'=\Pi_\Theta(\boldsymbol{B})$$

$$=\arg\min_{\boldsymbol{B}'\in\Theta}\|\boldsymbol{B}'-\boldsymbol{B}\|_2^2$$

$$=\min_{\boldsymbol{B}'\in\Theta}\sum_{\substack{p\in\{1,\cdots,P\},\\k\in\{1,\cdots,K\}}}(\boldsymbol{B}'_{p,k}-\boldsymbol{B}_{p,k})^2$$

等价于求解如下的二次规划问题所得到的矩阵 $\boldsymbol{B}'$：

$$\arg\min_{\mu,\boldsymbol{B}'}\frac{1}{2}\sum_{\substack{p\in\{1,\cdots,P\},\\k\in\{1,\cdots,K\}}}(\boldsymbol{B}'_{p,k}-\boldsymbol{B}_{p,k})^2 \tag{32}$$

$$\text{s.t. } \forall p,k,\boldsymbol{B}'_{p,k}\leqslant\mu_p\sum_{p\in\{1,\cdots,P\}}\mu_p=H^*,\forall p,k$$

$$\boldsymbol{B}'_{p,k}\geqslant 0,\forall p$$

$$\mu_p\geqslant 0$$

式(32)的拉格朗日形式为：

$$L(\boldsymbol{B}',\mu,\zeta,\lambda,\hbar,\chi)=$$

$$\frac{1}{2}\sum_{\substack{p\in\{1,\cdots,P\},\\k\in\{1,\cdots,K\}}}(\boldsymbol{B}'_{p,k}-\boldsymbol{B}_{p,k})^2+\sum_{\substack{p\in\{1,\cdots,P\},\\k\in\{1,\cdots,K\}}}\zeta_{p,k}(\boldsymbol{B}'_{p,k}-\mu_p)+$$

$$\lambda\Big(\sum_{p\in\{1,\cdots,P\}}\mu_p-H^*\Big)-\sum_{\substack{p\in\{1,\cdots,P\},\\k\in\{1,\cdots,K\}}}\hbar_{p,k}\boldsymbol{B}'_{p,k}-\sum_{p\in\{1,\cdots,P\}}\chi_p\mu_p \tag{33}$$

再由互补松弛条件，式(33)的解等价于求解满足如下问题的 μ 和 λ：

$$\sum_i\mu_i=H^*\sum_{p,\boldsymbol{B}_{p,k}\geqslant\mu_p}(\boldsymbol{B}_{p,k}-\mu_p)=\lambda,\ \forall\ p \tag{34}$$

$$\text{s.t. } \mu_p>0$$

$$\sum_{k\in\{1,\cdots,K\}}\boldsymbol{B}_{p,k}\leqslant\lambda,\ \forall\ p$$

$$\text{s.t. } \mu_p=0$$

$$\forall\ p\mu_p\geqslant 0;\lambda\geqslant 0$$

此时，原来的二次规划问题已转换为线性规划问题，而且解路径是分段线性的，因此很容易求解。Quattoni 等人给出的上述算法是关于多输出 $L_{\infty,1}$ 组套索模型的，由于单输出为多输出的特例，因此其也适用于单输出的情况。

3.1.3 重叠的组套索模型的算法实现

块坐标下降算法不适用于重叠组套索模型的求解，这是因为重叠组套索模型的目标函数中不同组之间具有共同的自变量，因此是块坐标不可分离的。而且，重叠组套索模型的目标函数还是不平滑的。对重叠组套索模型的求解有轮换方向乘子算法和 Nesterov 提出的平滑近似技巧。

Boyd 等人[56]和 Yuan 等人[21]均指出引入辅助向量后可以直接利用轮换方向乘子法直接求解重叠组套索模型。对于重叠组套索模型：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2$$

引入辅助向量 z 得如下重叠组套索模型的有约束形式：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{z}_{g_j}\|_2$$

$$\text{s.t. } \boldsymbol{z}_{g_j}=\boldsymbol{\beta}_{g_j},\ g_j\in G$$

上述约束形式的增广拉格朗日形式为：

$$L_\omega(\boldsymbol{\beta},\boldsymbol{z},\xi)=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{z}_{g_j}\|_2+\sum_{g_j\in G}\xi_{g_j}^{\mathrm{T}}(\boldsymbol{z}-\boldsymbol{\beta}_{g_j})+\frac{\tilde{\omega}}{2}\sum_{g_j\in G}\|\boldsymbol{z}_{g_j}-\boldsymbol{\beta}_{g_j}\|^2$$

再利用轮换方向法分别求解以下最优化问题：

$$\boldsymbol{\beta}^{\tilde{k}+1}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}L_\rho(\boldsymbol{\beta},\boldsymbol{z}^{\tilde{k}},\xi^{\tilde{k}})$$

$$\boldsymbol{z}^{\tilde{k}+1}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}L_\rho(\boldsymbol{\beta}^{\tilde{k}+1},\boldsymbol{z},\xi^{\tilde{k}})$$

$$\xi_{g_j}^{\tilde{k}+1}=\xi_{g_j}^{\tilde{k}}+\tilde{\omega}(\boldsymbol{z}_{g_j}^{\tilde{k}+1}-\boldsymbol{\beta}_{g_j}^{\tilde{k}+1})$$

即得到重叠组套索模型的解。

Chen 等人[59]提出利用 Nesterov 提出的平滑近似技巧来对重叠组套索模型的目标函数进行预处理，即利用 Nesterov 提出的平滑近似技巧将不平滑的问题转化为平滑的问题，从而得到不平滑罚函数的平滑近似表达式。这个过程中由于引入了对偶范数，因此同时将原来块坐标不可分离的问题转化成了块坐标可分离的问题。经过上述预处理后，再利用梯度法求解经过上述变换后的平滑问题。已知重叠组套索模型的罚函数为：

$$\Omega(\boldsymbol{\beta})=\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2 \tag{35}$$

引入对偶范数

$$\|\boldsymbol{\beta}_{g_j}\|_2=\max_{\|\alpha_{g_j}\|_2\leqslant 1}\boldsymbol{\alpha}_{g_j}^{\mathrm{T}}\boldsymbol{\beta}_{g_j} \tag{36}$$

代入式(35)可得：

$$\begin{aligned}\Omega(\boldsymbol{\beta})&=\lambda\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}_{g_j}\|_2\\&=\lambda\sum_{g_j\in G}w_{g_j}\max_{\|\alpha_{g_j}\|_2\leqslant 1}\boldsymbol{\alpha}_{g_j}^{\mathrm{T}}\boldsymbol{\beta}_{g_j}\\&=\max_{\boldsymbol{\beta}\in Q}\sum_{g_j\in G}\lambda w_{g_j}\boldsymbol{\alpha}_{g_j}^{\mathrm{T}}\boldsymbol{\beta}_{g_j}\\&=\max_{\boldsymbol{\alpha}\in Q}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\beta}\end{aligned} \tag{37}$$

其中

$\boldsymbol{\alpha}=[\boldsymbol{\alpha}_{g_1}^{\mathrm{T}},\cdots,\boldsymbol{\alpha}_{g_j}^{\mathrm{T}}]^{\mathrm{T}}$ 且 $Q\equiv\{\boldsymbol{\alpha}\mid\|\boldsymbol{\alpha}_{g_j}\|_2\leqslant 1,\forall\ g_j\in G\}$。矩阵 $\boldsymbol{M}\in\boldsymbol{R}^{\Sigma_{g_j\in G}|g_j|\times P}$ 的元素为：

$$\boldsymbol{M}_{(l,g_j),p}=\begin{cases}\lambda w_{g_j}, & \text{当 } l=p \text{ 时}\\ 0, & \text{其他情况}\end{cases} \tag{38}$$

其中，l 表示组内特征的索引。此时 $\max_{\boldsymbol{\alpha}\in Q}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\beta}$ 已变为关于 $\boldsymbol{\alpha}$ 的块坐标可分离的，但仍不平滑。因此再引入 $\max_{\boldsymbol{\alpha}\in Q}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\beta}$ 的近似平滑函数：

$$f_\mu(\boldsymbol{\beta})=\max_{\boldsymbol{\alpha}\in Q}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\beta}-\aleph\cdot\frac{1}{2}\|\boldsymbol{\alpha}\|_2^2) \tag{39}$$

将其作为 $\max_{\boldsymbol{\alpha}\in Q}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{\beta}$ 的近似表达式，解决了不平滑的问题。故重叠组套索模型的目标函数此时变为：

$$\begin{aligned}\min_\beta\tilde{f}(\beta)&=\frac{1}{2}\|y-X\beta\|_2^2+f_\aleph(\beta)\\&=\frac{1}{2}\|y-X\beta\|_2^2+\max_{\alpha\in Q}(\alpha^{\mathrm{T}}M\beta-\aleph\cdot\frac{1}{2}\|\alpha\|_2^2)\end{aligned} \tag{40}$$

此时的目标函数为平滑且块坐标可分离的，可以利用梯度法求解。另外，由于树组套索模型是重叠组套索模型的一个特例，因此这种方法当然也可以用于求解 3.1.4 节中的树组套索模型。

3.1.4 树组套索模型的算法实现

由于树组套索模型是重叠组套索模型的特例，因此求解重叠组套索模型的算法也可以用来求解树组套索模型。例如，利用 Nesterov 的平滑近似技巧将树组套索模型的目标函数转换为平滑的，然后再利用梯度算法求解。树组套索模型的目标函数是块坐标不可分离的，不能利用块坐标下降算法对其直接进行求解，因此 Jenatton 等人[23]提出通过对偶变换得到块坐标可分离的对偶问题，再利用块坐标上升算法求解。已知树组套索模型为：

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^P}f(\boldsymbol{\beta})+\lambda\Omega(\boldsymbol{\beta}) \tag{41}$$

其中，$f(\boldsymbol{\beta})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2$，$\Omega(\boldsymbol{\beta})=\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\alpha}_{g_j}\|_2$。经过 $\hat{\boldsymbol{\beta}}$ 处的二阶泰勒展开，上述问题等价于以下问题：

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^P}f(\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathrm{T}}\nabla f(\hat{\boldsymbol{\beta}})+\lambda\Omega(\boldsymbol{\beta})+\frac{L^*}{2}\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_2^2 \tag{42}$$

其中，$L^*>0$ 为 Lipschitz 常数。式(42)又可以写作如下形式：

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{\beta}-(\hat{\boldsymbol{\beta}}-\frac{1}{L^*}\nabla f(\hat{\boldsymbol{\beta}}))\|_2^2+\frac{\lambda}{L^*}\Omega(\boldsymbol{\beta}) \tag{43}$$

式(43)等价于：

$$prox_\lambda(\boldsymbol{v})=\arg\min_{\boldsymbol{v}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{u}-\boldsymbol{v}\|_2^2+\lambda^*\Omega(\boldsymbol{v}) \tag{44}$$

其中，$\boldsymbol{u}=\hat{\boldsymbol{\beta}}-\frac{1}{L^*}\nabla f(\hat{\boldsymbol{\beta}})$，$\boldsymbol{v}=\boldsymbol{\beta}$，$\boldsymbol{\Omega}(\boldsymbol{v})=\sum_{g_j\in G}w_{g_j}\|\boldsymbol{v}_{g_j}\|_2$。

Jenatton 等人指出式(44)的对偶问题为：

$$\max_{\tilde{\xi}}-\frac{1}{2}(\|\boldsymbol{u}-\sum_{g_j\in G}\tilde{\boldsymbol{\xi}}_{g_j}\|_2^2-\|\boldsymbol{u}\|_2^2)$$
$$\text{s.t. }\forall g_j\in G,\|\tilde{\boldsymbol{\xi}}_{g_j}\|_*\leqslant\lambda^* w_{g_j} \tag{45}$$

若 $l\notin g_j$，则 $\tilde{\xi}^l_{g_j}=0$，其中，$\|\cdot\|_*$ 表示对偶范数，$\tilde{\xi}^l_{g_j}$ 表示向量 $\tilde{\boldsymbol{\xi}}_{g_j}$ 的第 l 个分量。经过上述对偶变换后，将原来的无约束最小化问题即式(44)转化为了有约束最大化问题即式(45)。无约束形式的原问题即式(44)中的罚函数关于原变量 $\boldsymbol{v}$ 是块坐标不可分离的，而有约束形式的对偶问题即式(45)中的约束项关于对偶变量 $\tilde{\xi}$ 是块坐标可分离的，因此经过对偶变换后符合块坐标上升算法的适用条件，可以结合块坐标上升算法与梯度法对对偶问题即式(45)进行求解，其中块坐标上升算法为外层循环，梯度法为内层循环。

3.1.5 多输出树组套索模型的算法实现

多输出树组套索模型的罚函数具有不平滑性和块坐标不可分离性，不能利用块坐标下降算法直接对多输出树组套索模型求解。对于多输出树组套索模型的求解往往需要先对其目标函数进行预处理后再进行求解：一种方法为将其罚函数经变分替代后转化为平滑问题，再利用轮转变量寻优方法求解；或将利用 Nesterov 的近似平滑技巧将不平滑的问题转化为平滑的问题再求解。

Kim 等人[26]首先将多输出树组套索模型目标函数中的罚项取平方得：

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda(\sum_p\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2)^2 \tag{46}$$

Kim 等人指出此操作并不改变问题的解路径。但是式(46)中的罚函数仍然是不平滑的函数，因此再根据如下的变分不等式：

$$(\sum_p\sum_{g_j\in G}w_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2)^2\leqslant\sum_p\sum_{g_j\in G}\frac{w^2_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2^2}{d_{p,g_j}} \tag{47}$$

得到罚项的替代函数

$$\sum_p\sum_{g_j\in G}\frac{w^2_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2^2}{d_{p,g_j}} \tag{48}$$

即将原来不平滑的目标函数转化为平滑的目标函数。由于上述过程引入了一个新的变量 d_{p,g_j}，因此此时为两个自变量，再利用轮转变量寻优方法进行求解即可。

Kim 等人[27]利用 Nesterov 的近似平滑技巧求解多输出树组套索模型，方法类似于 3.1.3 节中求解重叠组套索模型的方法。他们首先把多输出树组套索模型的罚函数按照叶子节点和内部节点分解为两部分：

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G_{int}}w_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2+$$
$$\lambda\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G_{leaf}}w_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2 \tag{49}$$

其中，G_{int} 表示内部节点的集合，G_{leaf} 表示叶子节点的结合。求解的难点在于内部节点对应的那一项

$$\lambda\sum_{p\in\{1,\cdots,P\}}\sum_{g_j\in G_{int}}w_{g_j}\|\boldsymbol{\beta}^p_{g_j}\|_2 \tag{50}$$

是块坐标不可分离且不平滑的，与 3.1.3 节中所采用的目标函数预处理技巧相同，这里也要利用对偶范数和 Nesterov 的近似技巧将式(50)转化为块坐标可分离且平滑的函数。经过上述对目标函数的预处理后，Kim 等人指出此时可以利用快速迭代收缩阈值算法[61](Fast Iterative Shrinkage-Thresholding Algorithm)对预处理后的目标函数进行求解。

3.1.6 混合组套索模型的算法实现

Friedman 等人[28]提出用两层坐标下降算法求解混合组套索模型，外层迭代为组水平对应的块坐标下降算法，内层迭代为变量水平对应的坐标下降算法。假定 $\boldsymbol{X}_j$ 表示第 j 个组，x_l 表示组 $\boldsymbol{X}_j$ 内的第 l 个特征，即

$$\boldsymbol{X}_j=(x_1,\cdots,x_l,\cdots,x_L)$$

且 $\boldsymbol{\beta}_j=\boldsymbol{\theta}^*=(\theta_1^*,\cdots,\theta_l^*,\cdots,\theta_L^*)$。

令 $\boldsymbol{\beta}_{(-j)}=(\beta_1,\cdots,\beta_{j-1},0,\beta_{j+1},\cdots,\beta_J)$ 表示去掉第 j 个组后的模型向量，令 $\boldsymbol{r}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_{(-j)}$，则式(13)目标函数式的子梯度方程为：

$$-\boldsymbol{x}_l^T(\boldsymbol{r}-\sum_{l\in\{1,\cdots,L\}}x_l\theta_l^*)+\lambda_1 e_l+\lambda_2 t_l=0 \tag{51}$$

其中，若 $\theta_j^*\neq0$，则 $e_l=\frac{\theta_l^*}{\|\boldsymbol{\theta}^*\|}$；若 $\theta_j^*=0$，则 e 为满足 $\|e\|_2\leqslant1$ 的向量。若 $\theta_l^*\neq0$，则 $t_l\in\text{sign}(\theta_l^*)$；$\theta_l^*=0$，则 $t_l\in[-1,1]$。若令 $\tilde{a}=\boldsymbol{X}_j\boldsymbol{r}$，则 θ_l^* 为零的充分必要条件为方程 $\tilde{a}_l=\lambda_1 e_l+\lambda_2 t_l$ 的解满足 $\|e\|_2\leqslant1$ 且 $t_l\in[-1,1]$，这等价于在 $t_l\in[-1,1]$ 区间内求解以下最小化问题：

$$\Im(t)=(1/\lambda_1^2)\sum_{l=1}^{k}(\tilde{a}_l-\lambda_2 t_l)^2=\sum_{l=1}^{L}e_l^2$$

得到 $\Im(t)$ 的最优解后，然后检查是否满足 $\Im(\hat{t})\leqslant1$。若 $\Im(\hat{t})\leqslant1$，则

$$\hat{\boldsymbol{\beta}}_j=0 \tag{52}$$

否则求解下面的最小化问题

$$\min\frac{1}{2}\sum_{n=1}^{N}(r_n-\sum_{l=1}^{L}x_{nl}\theta_l^*)^2+\lambda_1\|\boldsymbol{\theta}^2\|_2+\lambda_2\sum_{l=1}^{L}|\theta_l^*| \tag{53}$$

令 $\boldsymbol{r}_l=\boldsymbol{r}-\sum_{l^*\neq l}\boldsymbol{x}_{l^*}\theta_{l^*}$，由式(53)的子梯度方程可得，若 $|\boldsymbol{r}_l^T\boldsymbol{r}_l|<\lambda_2$，则

$$\hat{\theta}_l^*=0 \tag{54}$$

否则，求解式(53)的最优化问题得到 $\hat{\theta}_l^*$ 的值：

$$\hat{\theta}_l^*\leftarrow\min\frac{1}{2}\sum_{n=1}^{N}(r_n-\sum_{l=1}^{L}x_{nl}\theta_l^*)^2+\lambda_1\|\boldsymbol{\theta}^*\|_2+\lambda_2\sum_{l=1}^{L}|\theta_l^*| \tag{55}$$

显然，可以对上述过程套用利用块坐标下降算法求解，即对于 $j=1,\cdots,J$ 和 $l=1,\cdots,L$，迭代式(52)到式(55)直到满足预先设定的收敛条件。

另外，Chatterjee 等人[30]指出混合组套索是树组套索的特例，其可以看作是两层的树结构：其中树的根节点对应特征组，叶子节点对应特征。所以他们指出可以利用求解树组套索的算法来求解混合组套索模型。

3.1.7 自适应组套索模型的算法实现

自适应组套索模型的算法实现分为两步，首先求解普通组套索模型得到其模型参数的初始估计 $\hat{\boldsymbol{\beta}}^{GL}$：

$$\hat{\boldsymbol{\beta}}^{GL}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \tag{56}$$

然后将初始估计 $\hat{\boldsymbol{\beta}}^{GL}$ 作为权，即令 $w_j=\|\hat{\boldsymbol{\beta}}_j^{GL}\|_2^{-1}$，$j=1,\cdots,J$，再继续求解如下的普通组套索模型：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}w_j\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2 \tag{57}$$

显然,上面两步均相当于求解普通组套索模型,可以直接利用 3.1.1 节中所述的块坐标下降算法或组最小角回归算法等方法求解,因此求解自适应组套索模型的计算本质上是求解两次普通组套索模型。

3.1.8 逻辑斯蒂组套索模型的算法实现

因为逻辑斯蒂组套索模型的解路径不是分段线性的,所以不能利用组最小角回归算法对其进行求解。Kim 等人[34]提出利用投影梯度算法求解逻辑斯蒂组套索模型,但是他们的算法只在步长很小时才收敛。由于很小的步长又会导致收敛速度慢,因此需要在收敛性和收敛速度之间进行取舍。Meier 等人[33]针对此问题提出利用块坐标下降算法和块坐标梯度下降算法求解逻辑斯蒂组套索模型,并指出,当模型参数空间维数不太大时可用类似于 3.1.1 节中求解普通组套索模型时的块坐标下降算法求解,但是在关于第 j 个组的优化问题中逻辑斯蒂组套索模型的解 $\boldsymbol{\beta}_j$ 不具有式(31)那样的显式更新公式,在算法中逻辑斯蒂组套索模型的解的更新公式只能表示为 $\boldsymbol{\beta}_j\leftarrow\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\Xi(\boldsymbol{\beta})$,因此这一步需要数值迭代计算方法,只适用于中小规模的变量选择问题。针对上述缺点,Meier 等人针对模型参数空间维数很大时的情况提出利用 Tseng 等人[60]的块坐标梯度下降算法(Block Coordinate Gradient Descent Method)求解逻辑斯蒂组套索模型,其基本思想为:对于逻辑斯蒂组套索模型的目标函数

$$\min_{\boldsymbol{\beta}\in\mathbf{R}^P}-\Phi(\boldsymbol{\beta})+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2$$

来说,块坐标梯度下降算法中关于第 j 个组的梯度 $\{\tilde{\boldsymbol{d}}|\tilde{\boldsymbol{d}}_{j^*}=0,j^*\neq j\}$ 由如下的最优化问题求得:

$$\begin{aligned}&\arg\min_{\{\tilde{\boldsymbol{d}}|\tilde{\boldsymbol{d}}_{j^*}=0,j^*\neq j\}}\Xi(\beta+\tilde{d})\\&=\arg\min_{\{\tilde{\boldsymbol{d}}|\tilde{\boldsymbol{d}}_{j^*}=0,j^*\neq j\}}-\{\Phi(\beta)+\tilde{d}^{\mathrm{T}}\nabla\Phi(\beta)+\frac{1}{2}\tilde{d}^{\mathrm{T}}H\tilde{d}\}+\\&\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j+\tilde{d}\|_2\end{aligned}\tag{58}$$

其中,H 为 Hessian 矩阵。得到关于第 j 个组的梯度 $\{\tilde{\boldsymbol{d}}|\tilde{\boldsymbol{d}}_{j^*}=0,j^*\neq j\}$ 后,然后执行关于第 j 个组的梯度下降更新公式:

$$\boldsymbol{\beta}\leftarrow\boldsymbol{\beta}+\tilde{t}\tilde{\boldsymbol{d}}\tag{59}$$

此即梯度下降步骤,其中步长 $\tilde{t}$ 可以通过 Armijo 线搜索得到。当 $j=1,\cdots,J$ 时不断迭代式(58)和式(59),直到满足预先设定的收敛条件即可。

上述块坐标下降算法和块坐标梯度下降算法不仅适用于逻辑斯蒂组套索模型,还适用于其他广义线性模型的组套索模型,但均要求设计矩阵 $\boldsymbol{X}$ 为正交矩阵。另外,Roth 等人[8]提出利用 3.1.1 节中所述的活动集算法求解关于广义线性模型的组套索模型,其原理与 3.1.1 节中求解普通组套索模型的原理相同,不再赘述。

3.2 非凸罚组稀疏模型的算法实现

与组套索模型不同,非凸罚组稀疏模型的罚函数是非凸的,因此非凸性使得非凸罚组稀疏模型的求解更加困难,往往需要先利用局部近似方法对其目标函数进行预处理,再结合块坐标下降算法对其进行求解。由于非凸罚组稀疏模型目标函数中的罚函数是非凸的,因此一般来说会得到局部的最优解或近似最优解。

3.2.1 组 SCAD 罚模型的算法实现

Wang 等人[38]和 Huang 等人[40]给出了 L_2 范数组 SCAD 罚模型的求解算法。Jiang 等人[39]给出了 L_1 范数组 SCAD 罚模型的求解算法。

Wang 等人提出先对 L_2 范数组 SCAD 罚模型的罚函数进行预处理,然后结合块坐标下降算法求解。他们利用文献[45]中处理 SCAD 罚模型的局部二次近似思想,指出 L_2 范数组 SCAD 罚模型的非凸罚函数 $\phi_\lambda(\|\boldsymbol{\beta}_j\|_2)$ 在某一满足 $\|\boldsymbol{\beta}_j^0\|_2>0$ 的已知点 $\boldsymbol{\beta}_j^0$ 的邻域内的近似表达式为:

$$\begin{aligned}\phi_\lambda(\|\boldsymbol{\beta}_j\|_2)\approx\phi_\lambda(\|\boldsymbol{\beta}_j^0\|_2)+\frac{1}{2}\frac{\phi_\lambda'(\|\boldsymbol{\beta}_j^0\|_2)}{\|\boldsymbol{\beta}_j^0\|_2}[\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\beta}_j-\\\boldsymbol{\beta}^{0}{}_j^{\mathrm{T}}\boldsymbol{\beta}_j^0]\end{aligned}$$

其中,$\|\boldsymbol{\beta}_j^0\|_2>0$,然后用牛顿法求解。注意,$\boldsymbol{\beta}_j$ 的初始估计应由岭回归估计得到,即将原目标函数中的罚函数 $\phi_\lambda(\|\boldsymbol{\beta}_j\|_2)$ 替换为 $\|\boldsymbol{\beta}_j\|_2^2$ 时相应的优化问题的解。但是,利用局部二次近似方法和牛顿方法的计算复杂度很高,因此 Huang 等人给出了 L_2 范数组 SCAD 罚模型的显式解。由文献[52]可得当设计矩阵 $\boldsymbol{X}$ 为正交矩阵时,SCAD 罚模型关于第 p 个特征的解具有如下的显式形式:

$$\begin{aligned}&\hat{\beta}_p^{SCAD}(z^*;\lambda,\gamma)\\&=\begin{cases}S(z^*;\lambda), & \text{当}\ \|z^*\|_2\leqslant 2\lambda\ \text{时}\\\dfrac{\gamma-1}{\gamma-2}S(z^*;\dfrac{\gamma\lambda}{\gamma-1}), & \text{当}\ 2\lambda<\|z^*\|_2\leqslant\gamma\lambda\ \text{时}\\z^*, & \text{当}\ \|z^*\|_2>\gamma\lambda\ \text{时}\end{cases}\end{aligned}$$

其中,$S(\cdot)$ 为软阈值算子。

$$S(z^*;\tilde{\sigma})=\begin{cases}z^*-\tilde{\sigma}, & z^*>\tilde{\sigma}\\0, & |z^*|\leqslant\tilde{\sigma}\\z^*+\tilde{\sigma}, & z^*<-\tilde{\sigma}\end{cases}$$

Huang 等人指出,当设计矩阵 $\boldsymbol{X}$ 为正交矩阵时,将上述 SCAD 罚模型关于第 p 个特征的显式形式的解推广到多维情形后即为 L_2 范数组 SCAD 罚模型关于第 j 个组的解:

$$\begin{aligned}&\hat{\boldsymbol{\beta}}_j^{GSCAD}(z^*;\lambda,\gamma)\\&=\begin{cases}\boldsymbol{S}(z^*;\lambda), & \text{当}\ \|z^*\|_2\leqslant 2\lambda\ \text{时}\\\dfrac{\gamma-1}{\gamma-2}\boldsymbol{S}(z^*;\dfrac{\gamma\lambda}{\gamma-1}), & \text{当}\ 2\lambda<\|z^*\|_2\leqslant\gamma\lambda\ \text{时}\\z^*, & \text{当}\ \|z^*\|_2>\gamma\lambda\ \text{时}\end{cases}\end{aligned}\tag{60}$$

其中,z^* 为线性回归问题 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 的普通最小二乘估计,$\boldsymbol{S}(\cdot)$ 为多维软阈值算子。

$$\begin{aligned}\boldsymbol{S}(z^*;\tilde{\sigma})&=S(\|z^*\|;\tilde{\sigma})\frac{z^*}{\|z^*\|}\\&=\begin{cases}(\|z^*\|-\tilde{\sigma})\dfrac{z^*}{\|z^*\|}, & z^*>\tilde{\sigma}\\0, & |z^*|\leqslant\tilde{\sigma}\\(\|z^*\|+\tilde{\sigma})\dfrac{z^*}{\|z^*\|}, & z^*<-\tilde{\sigma}\end{cases}\end{aligned}$$

由式(2)得到了式(60)中 L_2 范数组 SCAD 罚模型关于第 j 个组的显式形式的解,然后再结合块坐标下降算法对全部的组 $j=1,\cdots,J$ 不断迭代直到收敛即可。

Jiang[39]给出了 L_1 范数组 SCAD 罚模型的求解算法,其原理与上述 Huang 等人[40]求解 L_2 范数组 SCAD 罚模型的算法相同,也是先推导出 L_1 范数组 SCAD 罚模型的显式解,再结合块坐标下降算法求解问题的最优解,此处不再赘述。

3.2.2 组桥模型的算法实现

当 $0<\gamma<1$ 时组桥模型的罚函数是非凸的，Huang 等人[41]等价地将其转化成凸的形式进行求解。他们指出当 $\hat{\boldsymbol{\beta}}$ 和 $\hat{\boldsymbol{\theta}}$ 为最优化问题

$$\min S(\boldsymbol{\beta})=\min \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2+\sum_{j=1}^{J}\theta_j^{1-1/\gamma}c_j^{1/\gamma} \| \boldsymbol{\beta}_j \|_1+\tau\sum_{j=1}^{J}\theta_j \tag{61}$$

的解时，相应地 $\hat{\boldsymbol{\beta}}$ 就为组桥模型的解。于是就将求解原组桥模型的最优解问题转化为求解式(61)二元凸优化问题，求解的方法为两层循环：外部循环利用关于两个自变量 $\boldsymbol{\beta}_j$ 和 θ_j 的轮转变量寻优方法，在轮转变量寻优方法的内部利用块坐标下降算法从 $j=1$ 一直迭代到 $j=J$。其中在内部循环块坐标下降算法中参数 θ_j（其中 $j=1,\cdots,J$）的更新公式为：

$$\theta_j=\arg\min_{\theta_j}\frac{1}{2} \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2+\theta_j^{1-1/\gamma}c_j^{1/\gamma} \| \boldsymbol{\beta}_j \|_1+\tau\theta_j$$

$$\Rightarrow\theta_j=c_j\left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma} \| \boldsymbol{\beta}_j \|_1^{\gamma}$$

参数 $\boldsymbol{\beta}_j$（其中 $j=1,\cdots,J$）的更新公式为：

$$\boldsymbol{\beta}_j=\arg\min_{\boldsymbol{\beta}_j\in\mathbf{R}^{d_j}}\frac{1}{2} \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2+\theta_j^{1-1/\gamma}c_j^{1/\gamma} \| \boldsymbol{\beta}_j \|_1+\tau\theta_j \tag{62}$$

式(62)相当于求解套索问题，可以利用最小角回归算法求解。

3.2.3 组 MC 罚模型的算法实现

Breheny 等人[42]给出了复合组 MC 罚模型的求解算法，他们利用块坐标下降算法求解复合组 MC 罚模型，但是需要先对其目标函数的罚函数进行预处理。在块坐标下降算法执行过程的第 $k+1$ 轮迭代中（索引 j 从 1 一直取到 J 的整个过程被称作一轮迭代，即块坐标下降算法遍历全部的组 $1,2,\cdots,J$ 一次），他们将复合组 MC 罚模型目标函数中的罚函数 $\varphi_{\lambda,b}\left(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\beta_{jl}|)\right)$ 在上一轮的值 $|\beta_{jl}^{(\tilde{k})}|$ 处进行一阶泰勒展开：

$$\varphi_{\lambda,b}\left(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\beta_{jl}^{(\tilde{k})}|)\right)+\varphi'_{\lambda,b}\left(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\beta_{jl}^{(\tilde{k})}|)\right)\varphi'_{\lambda,a}(|\beta_{jl}^{(\tilde{k})}|)(|\beta_{jl}^{(\tilde{k}+1)}|-|\beta_{jl}^{(\tilde{k})}|)$$

其中 $|\beta_{jl}^{(\tilde{k})}|$ 和 $|\beta_{jl}^{(\tilde{k}+1)}|$ 分别对应块坐标下降算法中上一轮的迭代和这一轮的迭代，对于这一轮迭代来说上一轮的值 $|\beta_{jl}^{(\tilde{k})}|$ 可以被看作常数，又由于忽略常数项不改变最优解，因此可以将常数 $|\beta_{jl}^{(\tilde{k})}|$ 对应的项忽略，从而得到罚函数的近似表达式：

$$\varphi'_{\lambda,b}\left(\sum_{l^*=1}^{L}\varphi_{\lambda,a}(|\beta_{jl^*}^{(\tilde{k})}|)\right)\varphi'_{\lambda,a}(|\beta_{jl}^{(\tilde{k})}|)|\beta_{jl}^{(\tilde{k}+1)}| \tag{63}$$

式(63)是关于 $|\beta_{jl}^{(\tilde{k}+1)}|$ 的线性函数，此时求解最优化问题相当于求解一个套索(Lasso)问题即引言中的式(1)，可直接由式(2)得到式(63)的解。

Huang 等人[40]给出了 L_2 范数组 MC 罚模型的显式解。由文献[53]可得当设计矩阵 $\boldsymbol{X}$ 为正交矩阵时 MC 罚模型关于第 p 个特征的解具有如下的显式形式：

$$\hat{\beta}_p^{MCP}(z^*;\lambda,\gamma)=\begin{cases}\dfrac{\gamma}{\gamma-1}S(z^*;\lambda), & \text{当 } \| z^* \|_2\leqslant\gamma\lambda \text{ 时}\\ z^*, & \text{当 } \| z^* \|_2>\gamma\lambda \text{ 时}\end{cases}$$

其中，$S(\cdot)$ 为软阈值算子。Huang 等人指出，当设计矩阵 $\boldsymbol{X}$ 为正交矩阵时，将上述 MC 罚模型关于第 p 个特征的显式形式的解推广到多维情形后即为 L_2 范数组 MC 罚模型关于第 j 个组的解：

$$\hat{\beta}_j^{GMCP}(z^*;\lambda,\gamma)=\begin{cases}\dfrac{\gamma}{\gamma-1}\boldsymbol{S}(z^*;\lambda), & \text{当 } \| z^* \|_2\leqslant\gamma\lambda \text{ 时}\\ z^*, & \text{当 } \| z^* \|_2>\gamma\lambda \text{ 时}\end{cases}$$

其中，z^* 为线性回归问题 $\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 的普通最小二乘估计，$\boldsymbol{S}(\cdot)$ 为多维软阈值算子。有了上述关于第 j 个组的解的显式形式，然后再结合块坐标下降算法对全部的组 $j=1,\cdots,J$ 不断迭代直到满足预先设定的收敛条件为止。

Jiang[39]给出了 L_1 范数组 MC 罚模型的求解算法，其原理与上述 Huang 等人给出的 L_2 范数组 MC 罚模型的求解算法类似，也是先推导出 L_1 范数组 MC 罚模型的显式解，然后再结合坐标下降算法求解，不再赘述。

4 最新提出的一些组稀疏模型

4.1 标准组套索模型

普通组套索模型要求设计矩阵是正交矩阵，当不满足上述条件时，需要将设计矩阵变换为正交矩阵，但是这会使原问题发生改变。另外，之前的很多论文对设计矩阵是否为正交矩阵的问题含糊不清，在参数空间维数不大于样本空间维数的条件下，Simon 等人[62]针对上述问题提出了标准组套索模型(standardized Group Lasso)：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2} \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2+\lambda\sum_{j=1}^{J}\sqrt{d_j} \| \boldsymbol{X}_j\boldsymbol{\beta}_j \|_2 \tag{64}$$

上述标准组套索模型对设计矩阵是否为正交矩阵无要求，不必将设计矩阵变换为正交矩阵。他们还将 Yuan 等人提出的普通组套索模型命名为非标准组套索模型(nonstandardized Group Lasso)，同时指出了标准组套索模型的解与非标准组套索模型的解之间的关系：标准组套索模型的求解等价于先将非标准组套索模型的设计矩阵变换为正交矩阵，再求解非标准组套索模型，最后将得到的解变换回原始空间。

4.2 自适应的重叠组套索模型

Percival 等人[63]给出了自适应的重叠组套索模型。已知共 P 个特征被分为 J 个组 $g_1,\cdots,g_J$，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示组的索引集，用 $\boldsymbol{\beta}_{g_j}$ 表示组 g_j 对应的子模型向量，$G=\{g_j \mid j=1,\cdots,J\}$ 表示全部组的索引集的集合，且 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$。已知式(5)中的线性回归模型，自适应重叠组套索模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2} \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2+\lambda_{g_j}\sum_{g_j\in G} \| \boldsymbol{\beta}_{g_j} \| \tag{65}$$

其中，$\lambda_{g_j}=\dfrac{1}{\| \boldsymbol{\beta}_{g_j}^{OLS} \|^{\gamma}}$，$\boldsymbol{\beta}^{OLS}=(\boldsymbol{X}^{\mathrm{T}}X)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ 为最小二乘估计的解，$\gamma>0$，$g_j\subseteq\{1,\cdots,P\}$。与自适应组套索模型的原理类似，自适应的重叠组套索模型实质上是求解两次重叠组套索模型。

4.3 平方根组套索模型

Bunea 等人[64]在平方根套索模型[65](square-root Lasso)的基础上提出了平方根组套索模型(group square-root Lasso)：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{\sqrt{N}} \| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2+\frac{\lambda}{N}\sum_{j=1}^{J}\sqrt{d_j} \| \boldsymbol{\beta}_j \|_2 \tag{66}$$

其中，N 为样本数。平方根组套索模型将普通组套索模型的损失函数由 $\| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2^2$ 替换为其平方根 $\| \boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta} \|_2$，但是并

未丧失普通组套索模型所具有的性质。Bunea 等人指出平方根组套索模型不仅在估计的准确性、预测的准确性和子集恢复的准确性方面与普通组套索模型一致，而且其优点在于使得权衡参数 λ 的选择独立于噪声水平，这种优点尤其在 $P>N$（参数空间维数大于样本空间维数）的情况下体现地更加显著，因为此时对于噪声水平的估计很困难。

4.4 多输出融合混合组套索模型

Zhou 等人[66]将 L_1 范数、$L_{2,1}$ 范数和融合套索[67]（fused Lasso）罚函数组合到一起，提出了多输出融合混合组套索模型。已知多输出回归模型式(11)，其对应的多输出融合混合组套索模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}\|_F^2+\lambda_1\|\boldsymbol{B}\|_1+\lambda_2\|\boldsymbol{R}\boldsymbol{B}^{\mathrm{T}}\|_1+\lambda_3\|\boldsymbol{B}\|_{2,1} \tag{67}$$

其中，$\boldsymbol{R}$ 为 $(K-1)\times K$ 阶的稀疏矩阵，$\boldsymbol{R}_{i,i}=1$ 且 $\boldsymbol{R}_{i,i+1}=-1$，其余元素均为 0。$\|\boldsymbol{B}\|_1$ 为 L_1 范数罚，作用为实现变量水平上的稀疏性，即特征选择；$\|\boldsymbol{B}\|_{2,1}$ 为 $L_{2,1}$ 范数罚，作用为实现组水平上的稀疏性，即特征组选择；$\|\boldsymbol{R}\boldsymbol{B}^{\mathrm{T}}\|_1$ 为融合套索的罚函数，作用为使模型向量中索引值相邻的解取类似矩形波形状的值，即索引值相邻的解的值波动不太大。

4.5 复合组桥模型

Seetharaman[68]提出了复合组桥模型：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}c_j\Big(\sum_{l=1}^{L}\|\boldsymbol{\beta}_l\|_{\gamma_1}\Big)^{\gamma_2} \tag{68}$$

其中，$c_j=|A_j|^{1-\gamma_2}$，$\gamma_1\in(0,1]$，$\gamma_2\in(0,1]$。显然，当 $\gamma_1=1$ 时，其退化为组桥模型；当 $\gamma_2=1$ 时，其退化为桥模型。与组桥模型一样，复合组桥模型也能同时实现特征选择和特征组选择。

4.6 关于逻辑斯蒂回归模型的混合组套索模型

Simon 等人[29]将混合组套索模型推广到逻辑斯蒂回归模型中，给出了关于逻辑斯蒂回归模型的混合组套索模型。假设 $\boldsymbol{y}\in\mathbf{R}^N$，$\boldsymbol{X}\in\mathbf{R}^{N\times P}$，则关于逻辑斯蒂回归模型的混合组套索模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{N}\Big(\sum_{n=1}^{N}\log(1+\exp(\boldsymbol{X}_n^{\mathrm{T}}\beta)+y_n\boldsymbol{X}_n^{\mathrm{T}}\beta)\Big)+\lambda_1\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2+\lambda_2\|\boldsymbol{\beta}\|_1 \tag{69}$$

4.7 关于广义逻辑斯蒂回归模型的组套索模型

Kwemou 等人[69]考虑了广义的逻辑斯蒂回归模型：

$$P(y_n=1)=\frac{\exp(f(\boldsymbol{x}_n))}{1+\exp(f(\boldsymbol{x}_n))} \tag{70}$$

其中，$\boldsymbol{x}_{1n},\cdots,\boldsymbol{x}_{Pn}$ 为 P 个特征，输出为 $y_n\in\{0,1\}$，$n=1,\cdots,N$ 为样本个数，$f(\cdot)$ 为未知的函数。假定 P 个特征被分为 J 个组 $g_1,\cdots,g_J$，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示某组的索引集，$\bigcap_{g_j\in G}g_j=\phi$。令 p 代表划分组之前的特征的索引，令 l 表示划分组之后的特征的索引。他们提出通过如下一系列已知函数 $\{\phi_1,\cdots,\phi_P\}$ 的线性组合来逼近未知函数 $f(\cdot)$：

$$f_\beta(\cdot)=\sum_{p=1}^{P}\boldsymbol{\beta}_p\phi_p(\cdot)=\sum_{j=1}^{J}\sum_{l\in g_j}\boldsymbol{\beta}_l\phi_l(\cdot) \tag{71}$$

使得待估量由 $f(\cdot)$ 转变为 $\boldsymbol{\beta}$，并给出了关于广义的逻辑斯蒂回归模型的组套索模型：

$$\hat{f}_\beta=\arg\min_{f_\beta\in\Gamma}\frac{1}{N}\sum_{n=1}^{N}\log(1+\exp(f(x_n)))-y_nf(x_n)+\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\beta_j\|_2 \tag{72}$$

其中，$\Gamma\subseteq\{f_\beta(\cdot)=\sum_{j=1}^{J}\sum_{l\in g_j}\boldsymbol{\beta}_l\phi_l(\cdot)\}$，$g_j\subseteq\{1,\cdots,P\}$。

4.8 关于 COX 比例风险回归模型的混合组套索模型

Simon 等人[29]还将混合组套索模型推广到 COX 比例风险回归模型中。假设有 N 个观测，$n\in\{1,\cdots,N\}$，δ_n 为第 n 次观测时的删失指示变量，对于非删失变量来说 $\delta_n=1$，否则 $\delta_n=0$。Simon 等人给出的关于 COX 比例风险回归模型的混合组套索模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{N}\log\Big(\sum_{n\in D}\sum_{j\in R_n}\exp(\boldsymbol{X}_j^{\mathrm{T}}\beta)-\boldsymbol{X}_j^{\mathrm{T}}\beta\Big)+\lambda_1\sum_{j=1}^{J}\sqrt{d_j}\|\beta_j\|_2+\lambda_2\|\beta\|_1 \tag{73}$$

其中，$D=\{n|\delta_n=1\}$，$R_n=\{j|y_j\geqslant y_n\}$。

4.9 关于 Tobit 模型的组套索模型

Liu 等人[70]将组套索模型应用到 Tobit 模型中，已知 Tobit 模型：

$$y_n^+=(\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\beta}+e_n)^+,\ n=1,\cdots,N \tag{74}$$

其中，$y_n^+=\max(y_n,0)$。他们定义关于 Tobit 模型的组套索模型为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{N}\sum_{n=1}^{N}|y_n^+-\sum_{j=1}^{J}\boldsymbol{X}_{nj}^{\mathrm{T}}\boldsymbol{\beta}_j|+\lambda\sum_{j=1}^{J}\|\boldsymbol{\beta}_j\|_2 \tag{75}$$

他们通过仿真和实际数据分析得出结论：在识别有效的变量组方面，Tobit 模型组套索模型优于 Wang 等人[71]提出的关于 Tobit 模型的套索模型。

4.10 关于广义加模型的组套索模型

Yin 等人[72]提出了关于广义加模型的组套索模型。已知广义加模型为：

$$y=\sum_{p=1}^{P}f_p(x_p)+\epsilon$$

其中，$f_1,\cdots,f_P$ 为平滑函数，$x_1,\cdots,x_P$ 为自变量，y 为响应变量。将 $x_1,\cdots,x_P$ 划分为 J 个组，组级的索引为 $j\in\{1,\cdots,J\}$，第 j 个组内的变量的索引为 $l\in\{1,\cdots,L\}$。Yin 等人提出的关于广义加模型的组套索模型为：

$$\hat{f}=\arg\min_f\frac{1}{2}E[(y-\sum_{p=1}^{P}f_p(x_p))^2]+\lambda\sum_{j=1}^{J}\sqrt{\sum_{l=1}^{L}E[f_l^2(x_l)]} \tag{76}$$

其中，$E[\cdot]$ 为期望算子，$j\in\{1,\cdots,J\}$ 表示第 j 个组的索引，$l\in\{1,\cdots,L\}$ 表示组中第 l 个函数的索引。若每个平滑函数 f_j 都能用自变量的线性组合形式表示，那么上述关于加模型的组套索模型即退化为普通组套索模型形式。另外，若每个组中都只包含一个函数，其就退化为加模型套索[73]。由于广义加模型的组套索模型为非参数形式，因此它有很好的灵活性，且同时选择和拟合一簇平滑函数，结合了广义加模型和普通组套索模型两者的优点，适用于非线性变量选择的情形。

5 未来的研究方向

5.1 存在组重叠的非凸组 SCAD 罚模型、组桥模型和组 MC 罚模型

罚函数为非凸的组 SCAD 罚模型、组桥模型和组 MC 罚模型在组之间变量存在重叠的情况时的模型和求解算法有待研究。假定 P 个特征被分为 J 个组 $g_1,\cdots,g_J$，其中 $g_j\subseteq\{1,\cdots,P\}$ 表示某组的索引集，用 β_{g_j} 表示组 g_j 对应的子模型向量，$G=\{g_j|j=1,\cdots,J\}$ 表示全部组的索引集的集合，且 $\bigcup_{g_j\in G}g_j=\{1,\cdots,P\}$。重叠 L_1 范数组 SCAD 罚模型求解的最

优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}\phi_\lambda(\|\boldsymbol{\beta}_{g_j}\|_1) \tag{77}$$

重叠 L_2 范数组 SCAD 罚模型求解的最优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}\phi_\lambda(\|\boldsymbol{\beta}_{g_j}\|_2) \tag{78}$$

其中，$\lambda\geqslant 0$，$\varphi_\lambda(|\theta|)$ 如式(24)所示。

重叠组桥模型求解的最优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}c_{g_j}\|\boldsymbol{\beta}_{g_j}\|_1^\gamma \tag{79}$$

其中，$\lambda\geqslant 0$。

重叠复合组 MC 罚模型求解的最优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}\varphi_{\lambda,b}\Big(\sum_{l=1}^{L}\varphi_{\lambda,a}(\|\boldsymbol{\beta}_{g_j l}\|)\Big) \tag{80}$$

重叠 L_1 范数组 MC 罚模型求解的最优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_{g_j}\|_1) \tag{81}$$

重叠 L_2 范数组 MC 罚模型求解的最优化问题为：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{g_j\in G}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_{g_j}\|_2) \tag{82}$$

其中，$\varphi_{\lambda,\gamma}(\theta)$ 如式(29)所示。上述各问题有待研究。

5.2 组稀疏罚函数的叠加与多重嵌套

将不同的罚函数组合到一起形成新的组稀疏模型有待研究，有如下两种组合方法。

5.2.1 叠加形式

比如将 SCAD 罚和 $L_{2,1}$ 范数罚相加到一起：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{j=1}^{J}(\|\boldsymbol{\beta}_j\|_2)+\lambda_2\sum_{p=1}^{P}\phi_\lambda(|\boldsymbol{\beta}_p|)$$

其中，$\phi_\lambda(\theta)$ 如式(24)所示。

5.2.2 嵌套形式

比如将 SCAD 罚函数和桥罚函数复合嵌套到一起，外部用 SCAD 罚，内部用 MC 罚：

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbf{R}^P}\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}\phi_{\lambda,b}\Big(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\boldsymbol{\beta}_{jl}|)\Big)$$

其中 $\varphi_{\lambda,\gamma}(\theta)$ 如式(29)所示。

5.3 非线性贝叶斯组套索

Raman 等人提出的贝叶斯组套索只涉及线性模型，未来可以把核函数引入到贝叶斯组套索中，建立非线性贝叶斯组套索模型。另外，Raman 等人利用吉布斯抽样方法对贝叶斯组套索中的参数进行估计，但是吉布斯抽样方法的计算量很大，EM 算法对贝叶斯组套索进行求解会大大加快计算速度，但 EM 算法可能陷入局部最优解，如何构造具有全局近似解的 EM 算法如随机近似 EM 算法值得研究。

5.4 树组套索、多输出树组套索、L_1 范数组 SCAD 和 L_1 范数组 MC 罚的模型选择一致性问题

研究不可表示条件[46]、稀疏 Riesz 条件[47]和特征值限制条件[48]树组套索、多输出树组套索、L_1 范数组 SCAD 罚模型和 L_1 范数组 MC 罚模型选择一致性。

结束语　本文对组稀疏模型中的普通组套索模型、$L_{\infty,1}$ 组套索模型、重叠组套索模型、树组套索模型、多输出树组套索模型、混合组套索模型、自适应组套索模型、逻辑斯蒂组套索模型和贝叶斯组套索模型，以及组 SCAD 罚模型、组桥模型和组 MC 罚模型等非凸罚组稀疏模型进行了详细的讨论。值得指出的是，其中混合组套索模型、L_1 范数组 SCAD 罚模型、组桥模型、复合组 MC 罚模型和 L_1 范数组 MC 罚模型能够同时实现特征组选择和组内特征选择，而其它的组稀疏模型只能实现特征组选择。我们还对各种组稀疏模型的求解算法，如最小角回归算法、块坐标下降(上升)算法、活动集算法、内点算法、投影梯度算法、谱投影梯度算法、轮换方向乘子算法和块坐标梯度下降算法等算法结合组稀疏模型进行了详细的分析，对各种罚函数近似方法如变分替代、对偶问题变换、罚函数对偶范数变换、一阶泰勒展开近似、Nesterov 近似技巧和局部二次近似等结合组稀疏模型进行了阐述，对最新提出的一些组稀疏模型进行了阐述，并给出了组稀疏模型未来的研究方向。

组稀疏模型是最简单的结构稀疏化模型，是构造其它复杂结构稀疏化模型的基础。组稀疏模型及其算法是当前高维数据建模的主要研究方向，在数理统计、模式识别、机器学习、信号处理、计算机视觉和生物信息学等领域具有广阔的应用前景，势必在以后的高维数据建模方法中占有重要的位置。

参 考 文 献

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

[2] 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

[3] Huang J, Zhang, T. The benefit of group sparsity[J]. The Annals of Statistics, 2010, 38(4): 1978-2004

[4] Bach F R. Consistency of the Group Lasso and multiple kernel learning[J]. The Journal of Machine Learning Research, 2008, 9: 1179-1225

[5] Wei F, Huang J. Consistent group selection in high-dimensional linear regression[J]. Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability, 2010, 16(4): 1369-1384

[6] Lounici K, Pontil M, Van De Geer S, et al. Oracle inequalities and optimal inference under group sparsity[J]. The Annals of Statistics, 2011, 39(4): 2164-2204

[7] Liu H, Zhang J. Estimation consistency of the Group Lasso and its applications[C]// Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Florida, USA: the MIT Press, 2009: 376-383

[8] Roth V, Fischer B. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms[C]// Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008: 848-855

[9] Yang Y, Zou H. A Fast Unified Algorithm for Solving Group-Lasso Penalized Learning Problems[J]. Journal of Computational and Graphical Statistics, 2015, 25(6): 1129-1141

[10] Qin Z, Scheinberg K, Goldfarb D. Efficient block-coordinate descent algorithms for the Group Lasso[J]. Mathematical Programming Computation, 2013, 5(2): 143-169

[11] Rakotomamonjy A. Surveying and comparing simultaneous sparse approximation (or group-Lasso) algorithms[J]. Signal

processing,2011,91(7):1505-1526

[12] Turlach B A, Venables W N, Wright S J. imultaneous variable selection[J]. Technometrics,2005,47(3):349-363

[13] Tropp J A. Algorithms for simultaneous sparse approximation Part II:Convex relaxation[J]. Signal Processing,2006,86(3):589-602

[14] Quattoni A, Collins M, Darrell T. Transfer learning for image classification with sparse prototype representations[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA:IEEE,2008. 1-8

[15] Schmidt M W, Murphy K P, Fung G, et al. Structure learning in random fields for heart motion abnormality detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Alaska, USA:IEEE,2008:1-8

[16] Quattoni A, Carreras X, Collins M, et al. An efficient projection for $L_{1,\infty}$ regularization[C]// Proceedings of the 26th Annual International Conference on Machine Learning. Quebec, Canada: ACM,2009:857-864

[17] Vogt J E, Roth V. The group-Lasso: $l_{1,\infty}$ regularization versus $l_{1,2}$ regularization[C]// Proceedings of the 32nd DAGM conference on Pattern recognition. Darmstadt, Germany:Springer-Verlag,2010:252-261

[18] Jenatton R, Audibert J Y, Bach F. Structured variable selection with sparsity-inducing norms[J]. The Journal of Machine Learning Research,2011,12:2777-2824

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

[20] Percival D. Theoretical properties of the overlapping groups Lasso[J]. Electronic Journal of Statistics,2012,6(1):269-288

[21] Yuan L, Liu J, Ye J. Efficient Methods for Overlapping Group Lasso[J]. IEEE transactions on pattern analysis and machine intelligence,2013,35(9):2104-2116

[22] Zhao P, Rocha G, Yu B. Grouped and hierarchical model selection through composite absolute penalties[J]. The Annals of Statistics,2009,37(6A):3468-3497

[23] Jenatton R, Mairal J, Obozinski G, et al. Proximal methods for hierarchical sparse coding[J]. Journal of Machine Learning Research,2011,12(2):2297-2334

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

[25] Jawanpuria P, Nath J S, Ramakrishnan G. Generalized hierarchical kernel learning[J]. The Journal of Machine Learning Research,2015,16(1):617-652

[26] Kim S, Xing E P. Tree-guided Group Lasso for multi-task regression with structured sparsity[C]// Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress,2010:543-550

[27] Kim S, Xing E P. Tree-guided Group Lasso for multi-response regression with structured sparsity, with an application to eQTL mapping[J]. The Annals of Applied Statistics,2012,6(3):1095-1117

[28] Prechmann P, Bronstein A M, Sapiro G. Learning efficient sparse and low rank models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(9):1821-1833

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

[30] Chatterjee S, Steinhaeuser K, Banerjee A, et al. Sparse Group Lasso: Consistency and Climate Applications[C]// Proceedings of the 12th SIAM International Conference on Data Mining. California, USA:Omnipress,2012:47-58

[31] Wang H, Leng C. A note on adaptive Group Lasso[J]. Computational Statistics and Data Analysis,2008,52(12):5277-5286

[32] Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high-dimensional additive models[J]. Journal of the American Statistical Association,2014,109(507):1270-1284

[33] Meier L, Van De Geer S, Bühlmann P. The Group Lasso for logistic regression[J]. Journal of the Royal Statistical Society:Series B (Statistical Methodology),2008,70(1):53-71

[34] Kim S, Xing E P. Tree-guided Group Lasso for multi—task regression with structured sparsity[C]// Proceedings of the 27th International Conference on Machine Learning. Haifa, Israel: Omnipress,2010:543-550

[35] Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies[J]. Bioinformatics,2012,28(10):1368-1375

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

[37] Chandran M. Analysis of Bayesian Group-Lasso in Regression Models[D]. Florida: University of Florida,2011

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

[39] Jiang D. Concave selection in generalized linear models[D]. Iowa: University of Lowa, USA,2012

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

[41] Geng Z, Wang S, Yu M. Group variable selection via convex log-exp-sum penalty with application to a breast cancer survivor study[J]. Biometrics,2015,71(1):53-62

[42] Breheny P, Huang J. Penalized methods for bi-level variable selection[J]. Statistics and its Interface,2009,2(3):369-380

[43] Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors[J]. Statistics and Computing,2015,25(2):173-187

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

[45] 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

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

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

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

[49] Parikh N, Boyd S P. Proximal Algorithms[J]. Foundations and Trends in optimization, 2014, 1(3): 127-239

[50] Zhou Z, Zhang Q, So A M. $\ell_{1,p}$-Norm Regularization: Error Bounds and Convergence Rate Analysis of First-Order Methods[C]// Proceedings of the 32nd International Conference on Machine Learning (ICML-15). Lille, France: MIT Press, 2015: 1501-1510

[51] Rakotomamonjy A, Flamary R, Gasso G, et al. L_p-L_q penalty for sparse linear and sparse multiple kernel multi-task learning[J]. IEEE Transaction on Neural Networks, 2011, 22(8): 1307-1320

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

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

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

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

[56] Efron B, Hastie T, Johnstone I. Least angle regression[J]. The Annals of statistics, 2004, 32(2): 407-499

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

[58] Fornasier M, Rauhut H. Recovery algorithms for vector-valued data with joint sparsity constraints[J]. SIAM Journal on Numerical Analysis, 2008, 46(2): 577-613

[59] Chen X, Lin Q, Kim S. Smoothing proximal gradient method for general structured sparse regression[J]. The Annals of Applied Statistics, 2012, 6(2): 719-752

[60] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization[J]. Computational Optimization and Applications, 2010, 47(2): 179-206

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

[62] Simon N, Tibshirani R. Standardization and the group Lasso penalty[J]. Statistica Sinica, 2012, 22(3): 983-1001

[63] Percival D. Theoretical properties of the overlapping groups Lasso[J]. Electronic Journal of Statistics, 2012, 6(5): 269-288

[64] Bunea F, Lederer J, She Y. The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms[J]. IEEE Transactions on Information Theory, 2014, 60(2): 1313-1325

[65] Belloni A, Chernozhukov V, Wang L. Square-root Lasso: pivotal recovery of sparse signals via conic programming[J]. Biometrika, 2011, 98(4): 791-806

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

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

[68] Seetharaman I. Consisitent bi-level variable selection via composite group bridge regression[D]. Kansas State: Kansas State University, 2013

[69] Nardi Y, Rinaldo A. On the asymptotic properties of the group lasso estimator for linear models[J]. Electronic Journal of Statistics, 2008, 2(1): 605-633

[70] Liu X, Wang Z, Wu Y. Group variable selection and estimation in the tobit censored response model[J]. Computational Statistics and Data Analysis, 2013, 60(2): 80-89

[71] W Zhan-feng, W Yao-hua, Z Lin-cheng. A LASSO-Type Approach to Variable Selection and Estimation for Censored Regression Model[J]. Chinese Journal of Applied Probability and Statistics, 2010, 26(1): 66-80

[72] Yin J, Chen X, Xing E P. Group Sparse Additive Models[C]// Proceedings of the 29th International Conference on Machine Learning. Omnipress. Scotland, UK: 2012: 871-878

[73] Ravikumar P, Lafferty J, Liu H. Sparse additive models [J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2009, 71(5): 1009-1030

[74] Meinshausen N, Bühlmann P. Stability selection[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010, 72(4): 417-473

[75] Borgi M A, Labate D, El Arbi M. Sparse multi-stage regularized feature learning for robust face recognition[J]. Expert Systems with Applications, 2015, 42(1): 269-279

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

[77] Asif M S, Romberg J. Fast and accurate algorithms for re-weighted-norm minimization[J]. IEEE Transactions on Signal Processing, 2013, 61(23): 5905-5916

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

[79] Zhao Y B, Li D. Reweighted ℓ_1-Minimization for Sparse Solutions to Underdetermined Linear Systems[J]. SIAM Journal on Optimization, 2012, 22(3): 1065-1088

[80] Simon N, Friedman J, Hastie T. A sparse-group lasso[J]. Journal of Computational and Graphical Statistics, 2013, 22(2): 231-245

[81] Bühlmann P. Statistical significance in high-dimensional linear models[J]. Bernoulli, 2013, 19(4): 1212-1242

[82] Zhang C H, Zhang S S. Confidence intervals for low dimensional parameters in high dimensional linear models[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2014, 76(1): 217-242

[83] Ye F, Zhang C H. Rate Minimaxity of the Lasso and Dantzig Selector for the L_q Loss in L_r Balls[J]. The Journal of Machine Learning Research, 2010, 11: 3519-3540