第4期 电 子 学 报 Vol.43 No.4
2015年4月 ACTA ELECTRONICA SINICA Apr. 2015

组稀疏模型及其算法综述

刘建伟,崔立鹏,罗雄麟
(中国石油大学(北京)自动化研究所,北京 102249)

摘 要: 稀疏性与组稀疏性在统计学、信号处理和机器学习等领域中具有重要的应用.本文总结和分析了不同组稀疏模型之间的区别与联系,比较了不同组稀疏模型的变量选择能力、变量组选择能力、变量选择一致性和变量组选择一致性,总结了组稀疏模型的各类求解算法并指出了各算法的优点和不足.最后,本文对组稀疏模型未来的研究方向进行了探讨.

关键词: 稀疏性;组稀疏性;变量选择;变量组选择;一致性
中图分类号: TP181 **文献标识码:** **文章编号:** 0372-2112(2015)04-0776-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.04.021

Survey on Group Sparse Models and Algorithms

LIU Jian-wei, CUI Li-peng, LUO Xiong-lin
(*Research Institute of Automation, China University of Petroleum, Beijing 102249, China*)

Abstract: The sparsity and group sparsity have important applications in the statistics, signal processing and machine learning. This paper summarized and analyzed the differences and relations between various group sparse models. In addition, we compared different models' variable selection ability, variable group selection ability, variable selection consistency and variable group selection consistency. We also summarized the algorithms of group sparse models and pointed the advantages and disadvantages of the algorithms. Finally, we point out the future research directions of the group sparse models.

Key words: sparsity; group sparsity; variable selection; variable group selection; consistency

1 引言

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,处理的数据集往往为高维或超高维的,且这些高维数据集通常还具有复杂的结构.虽然数据集的变量空间维数非常大,维数可从几百维到上万维,但只有一小部分变量与要预测的输出变量相关,其余大部分变量为噪声变量.而且相对于变量数来说,样本数很小,直接使用最小二乘等传统的建模工具,得到的是数值计算病态问题,因此变量选择问题显得尤为重要.Tibshirani 提出的套索(least absolute shrinkage and selection operator, Lasso)[1]模型利用 L_1 范数罚在零点处求解次梯度得到稀疏解,使得模型向量的许多分量为零,实现模型稀疏化和变量选择,套索模型为第一种利用正则化方法进行变量选择的模型.然而变量往往具有组结构,例如在基因表达分析中可把属于同一生物学路径的基因看做一个组.套索模型没有将变量的组结构作为先验信息,局限于变量水平上的稀疏性,只有变量选择能力而无变量组选择能力.Yuan 等人利用变量之间存在的组结构作为先验信息,提出了组套索(Group Lasso)[2]模型,该模型利用 $L_{2,1}$ 范数罚作为正则化项,能够实现变量组水平上的稀疏性,具有变量组选择能力.后来具有变量组选择能力的一系列组稀疏模型陆续被提出,其分类如图1所示,其中虚线框内的三种重叠非凸组稀疏模型为尚未被研究的组稀疏模型.

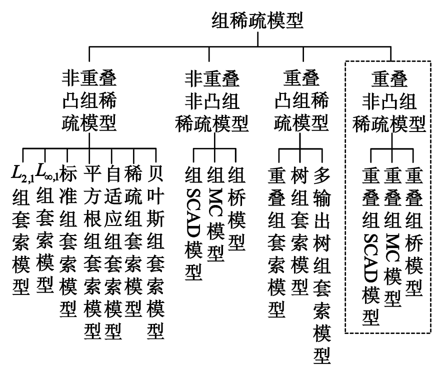

图1 组稀疏模型分类图

2 组稀疏模型

已知线性回归模型为

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

收稿日期:2013-10-23;修回日期:2014-07-03;责任编辑:李勇锋
基金项目:国家自然科学基金(No.21006127)

其中 $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$为设计矩阵,$\boldsymbol{\beta}\in\boldsymbol{R}^{P}$ 为模型向量,$\boldsymbol{y}\in\boldsymbol{R}^{N}$ 为输出向量,$\boldsymbol{\varepsilon}\in\boldsymbol{R}^{N}$ 为噪声向量且 $\varepsilon_n\sim N(0,\sigma^2)$, $n\in\{1,2,\cdots,N\}$, N 为样本数, P 为变量数.组稀疏模型目标函数的一般形式为

$$\arg\min_{\boldsymbol{\beta}\in\boldsymbol{R}^{P}}\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})+\Psi_\lambda(\boldsymbol{\beta})\tag{2}$$

其中 $\Phi(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{y})$为损失函数,$\Psi_\lambda(\boldsymbol{\beta})$为具有稀疏性效果的罚函数,$\lambda\geq0$ 为权衡参数.设变量被分为 J 个组 $G=\{g_j\mid j=1,\cdots,J\}$,$\boldsymbol{X}_j$ 为组 g_j 的子设计矩阵,d_j 为组 g_j 中的变量数,$\boldsymbol{\beta}_j\in\boldsymbol{R}^{d_j}$为组 g_j 的子模型向量,下文中若无声明则沿用上述符号含义.另外,第 2.1 和 2.2 节中组稀疏模型的各个变量组不具有重叠的变量,第 2.3 节中各个变量组可具有重叠的变量.

2.1　非重叠凸组稀疏模型

2.1.1　$L_{\infty,1}$组套索模型和 $L_{2,1}$组套索模型

基于式(1)的 $L_{2,1}$组套索模型和 $L_{\infty,1}$组套索模型的损失函数均为最小二乘损失:

$$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2\tag{3}$$

罚函数为 $\lambda\|\boldsymbol{\beta}\|_{q,1}=\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{\beta}_j\|_q$,当 $q=2$ 时为 $L_{2,1}$范数罚 $\lambda\|\boldsymbol{\beta}\|_{2,1}$,对应 $L_{2,1}$组套索模型[2];当 $q=\infty$时为 $L_{\infty,1}$范数罚 $\lambda\|\boldsymbol{\beta}\|_{\infty,1}$,对应 $L_{\infty,1}$组套索模型[3~8].当每个组只有一个变量时,$L_{2,1}$组套索模型和 $L_{\infty,1}$组套索模型都退化为套索模型.Huang 等人给出了强组稀疏条件(Strong Group Sparsity Condition)[9],并指出当数据集满足该条件时 $L_{2,1}$组套索模型优于套索模型,其优点体现在 $L_{2,1}$组套索模型的组结构使得其对噪声具有更强的鲁棒性.q 除了取 1 和∞外还可取其它值[10~12].

2.1.2　标准组套索模型

基于式(1)标准组套索(Standardized Group Lasso)[13]模型的损失函数为式(3),罚函数为 $\lambda\sum_{j=1}^{J}\sqrt{d_j}\|\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2$. $L_{2,1}$组套索模型要求设计矩阵是正交矩阵,当设计矩阵不满足正交条件时则要进行正交化,但这会使原问题发生改变,而标准组套索模型的优点为其不要求设计矩阵为正交矩阵.

2.1.3　平方根组套索模型

基于式(1)的平方根组套索(Group Square Root Lasso)[14]模型的损失函数为式(3),罚函数为 $L_{2,1}$范数罚 $\lambda\|\boldsymbol{\beta}\|_{2,1}$,平方根组套索模型的优点在于权衡参数 λ 的选择不需知道(1)中噪声变量的标准差 σ.当每个组只含一个变量时,其退化为平方根套索(Square Root Lasso)[15].

2.1.4　自适应组套索模型

基于式(1)的自适应组套索(Adaptive Group Lasso)[16,17]模型的损失函数为式(3),罚函数为 $\lambda\sum_{j=1}^{J}w_j\sqrt{d_j}\|\boldsymbol{\beta}_j\|_2$, $w_j=\|\hat{\boldsymbol{\beta}}_j^{GL}\|_2^{-1}$ 表示第 j 个组的权,$\hat{\boldsymbol{\beta}}_j^{GL}$ 为 $L_{2,1}$组套索模型的解.$L_{2,1}$组套索模型的变量组选择一致性不好,这是因为其过度缩小模较大的子模型向量,导致对模较大的子模型向量的有偏估计.自适应组套索模型在罚函数中为不同子模型向量分配不同的权,对模较大(小)的子模型向量执行较小(大)程度的惩罚,因此具有较好的变量组选择一致性.当每个组只含一个变量时其退化为自适应套索模型[18~20].

2.1.5　稀疏组套索模型

基于式(1)的稀疏组套索(Sparse Group Lasso)[21~23]模型的损失函数为式(3),罚函数为 $\lambda_1\|\boldsymbol{\beta}\|_{2,1}+\lambda_2\|\boldsymbol{\beta}\|_1$,其中 $L_{2,1}$范数罚的作用为变量组选择,L_1 范数罚的作用为组内的变量选择,因而稀疏组套索模型可同时实现变量选择和变量组选择,克服了套索模型只有变量选择能力和 $L_{2,1}$组套索模型只有变量组选择能力的缺点.

2.1.6　贝叶斯组套索模型

贝叶斯理论认为当每个组的子模型向量都有独立同分布的多维 Laplace 先验分布时,可把 $L_{2,1}$组套索模型表示为贝叶斯最大后验估计[24,25]:

$$P(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta},\sigma^2)=\prod_{n=1}^{N}N(y_n\mid\boldsymbol{x}^n\boldsymbol{\beta}^{\mathrm{T}},\sigma^2)\tag{4}$$

$$P(\boldsymbol{\beta}_j\mid\rho)=\mathrm{Laplace}\left(\boldsymbol{\beta}\mid0,\left(\frac{d_j\rho}{\sigma^2}\right)^{-\frac{1}{2}}\right)\tag{5}$$

$$P(\sigma^2\mid v_0,s_0^2)=\mathrm{InvGamma}(\sigma^2\mid v_0,t_0^2)\tag{6}$$

$$P(\rho\mid r,s)=\mathrm{Gamma}(\rho\mid r,s)\tag{7}$$

其中 $\boldsymbol{x}^n\in\boldsymbol{R}^P$ 表示变量在第 n 次的观察值,$\boldsymbol{y}_n$ 服从正态分布,ρ 服从 Gamma 分布,r 和 s 为 Gamma 分布的超参数,σ^2 为(1)中噪声变量的方差,其服从超参数为 v_0 和 t_0^2 的逆 Gamma 分布.

2.1.7　基于其它回归模型的组稀疏模型

除线性回归模型外,大量研究还将组稀疏模型推广到其他回归模型中,例如逻辑斯蒂回归模型[26~29]、COX 比例风险回归模型[22,30]、Tobit 模型[31,32]和广义加模型[33]等,基于这些回归模型的组稀疏模型呈现出在线性回归模型下所不具有的新特性,例如基于逻辑斯蒂回归模型的组稀疏模型可应用于离散变量建模,适用于分类器设计;基于广义加模型的组稀疏模型具有很好的灵活性,适用于非线性情形下的变量组选择.

2.2　非重叠非凸组稀疏模型

2.2.1　组 SCAD 模型

基于式(1)的组 SCAD (Group Smoothly Clipped Absolute Deviation)模型分为 L_1 组 SCAD 模型[34]和 L_2组 SCAD 模型[35],其损失函数均为式(3),罚函数为

$\sum_{j=1}^{J}\varphi_{\lambda}(\|\boldsymbol{\beta}_j\|_q)$，其中 $\varphi_{\lambda,\gamma}(\cdot)$ 为 SCAD 罚(Smoothly Clipped Absolute Deviation Penalty)[36]. 两者区别在于内部罚函数不同，$q=1$ 时为 L_1 组 SCAD 模型，$q=2$ 时为 L_2 组 SCAD 模型. $\gamma\to\infty$ 时，L_2 组 SCAD 模型变为 $L_{2,1}$ 组套索模型. 在变量(组)选择能力和统计性质方面，L_1 组 SCAD 模型能同时实现变量选择和变量组选择并具有较好的变量组选择一致性，但变量选择一致性较差. L_2 组 SCAD 模型只能实现变量组选择，具有较好的变量组选择一致性.

2.2.2 组 MC 模型

基于式(1)的组 MC(Group Minimax Concave)模型分为三种，其损失函数均为式(3)，当罚函数为 $\sum_{j=1}^{J}\varphi_{\lambda,b}\left(\sum_{l=1}^{L}\varphi_{\lambda,a}(|\beta_{jl}|)\right)$ 时为复合组 MC 模型[37]，当罚函数为 $\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_1)$ 时为 L_1 组 MC 模型[34]，当罚函数为 $\sum_{j=1}^{J}\varphi_{\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2)$ 时为 L_2 组 MC 模型[38]，$\varphi_{\lambda,\gamma}(\cdot)$ 为 MC 罚(Minimax Concave Penalty)[39]. 当 $\gamma\to\infty$ 时 L_2 组 MC 模型变为 $L_{2,1}$ 组套索模型. 复合组 MC 模型和 L_1 组 MC 模型均能同时实现变量组选择和组内变量选择，其中复合组 MC 模型有较好的变量选择一致性和变量组选择一致性，L_1 组 MC 模型只有较好的变量组选择一致性而其变量选择一致性较差. L_2 组 MC 模型只能实现变量组选择且有较好的变量组选择一致性.

2.2.3 组桥模型

基于式(1)的组桥(Group Bridge)模型分为 L_1 组桥模型[40]、L_2 组桥模型[41,42]与复合组桥模型[43]，其损失函数均为式(3). L_1 组桥模型的罚函数为 $\lambda\sum_{j=1}^{J}d_j^{1-\gamma}\|\boldsymbol{\beta}_j\|_1^{\gamma}$，$L_2$ 组桥模型的罚函数为 $\lambda\sum_{j=1}^{J}d_j^{1-\gamma}\|\boldsymbol{\beta}_j\|_2^{\gamma}$，复合组桥模型的罚函数为 $\lambda\sum_{j=1}^{J}d_j^{1-\gamma_2}\left(\sum_{l=1}^{L}|\beta_{jl}|^{\gamma_1}\right)^{\gamma_2}$，其中 l 为组内单个变量的索引，$\gamma,\gamma_1,\gamma_2\in(0,1]$. 当每个组只含有一个变量时，$L_1$ 组桥模型退化为桥回归(Bridge Regression)模型[44]. 当 $\gamma_1=1$ 时，复合组桥模型退化为 L_1 组桥模型. 当 $\gamma_2=1$ 时，复合组桥退化为桥回归模型. 在变量(组)选择能力和统计性质方面，L_1 组桥模型和复合组桥模型均同时具有变量选择能力和变量组选择能力，但 L_1 组桥模型只具有较好的变量组选择一致性而其变量选择一致性较差，复合组桥模型同时具有较好的变量选择一致性和较好的变量组选择一致性. L_2 组桥模型只具有变量组选择能力且具有较好的变量组选择一致性.

2.2.4 小结与分析

组 SCAD、组 MC 和组桥模型均具有变量组选择能力和较好的变量组选择一致性，但是否具有变量选择能力和较好的变量选择一致性则取决于内部所使用的罚：凡是在罚函数的内部对每个组使用 L_1 范数罚的模型均同时具有变量选择能力和变量组选择能力，并且变量组选择一致性较好，但变量选择一致性较差；凡是在罚函数的内部对每个组使用 L_2 范数罚的模型均只具有变量组选择能力和较好的变量组选择一致性；凡是在罚函数的内部对每个组使用非凸罚的模型均同时具有变量选择能力和变量组选择能力，并且变量选择一致性和变量组选择一致性均较好.

2.3 重叠凸组稀疏模型

2.3.1 重叠组套索模型

非重叠组稀疏模型有局限性，例如在微阵列基因表达数据分析中某基因可同时属于多个组，此时不仅要将组结构作为先验信息，还要把组之间的重叠结构也作为先验信息引入到罚函数中. 重叠组套索模型(Overlap Group Lasso)[45~48]的损失函数为式(3)，罚函数为 $\lambda\sum_{g_j\in G}\|\boldsymbol{\beta}_{g_j}\|_2$，其中不同组 g_j 之间可有重叠的变量，即允许同一个变量属于多个组. 重叠组套索也有统计性质较好的自适应版本：自适应重叠组套索[49].

2.3.2 树组套索模型

有的数据集中各变量组之间为偏序关系，即树组结构. 基于(1)的树组套索模型[50~53]的损失函数为式(3)，罚函数为 $\lambda\sum_{g_j\in G}\|\boldsymbol{\beta}_{g_j}\|_2$，其中各个组 g_j 之间形成树组结构. 实际上树组结构是重叠组结构的一个特例，因而树组套索模型也是重叠组套索模型的特例，前者相当于在后者基础上附加如下三个条件：同一层中各节点不具有重叠的变量；子节点的索引集是父节点索引集的子集；父节点的索引集为其子节点索引集的覆盖集. 树组套索模型的变量组选择效果为：若某组被选中，那么该组的全部父组也被选中；若某组被丢弃，则该组的全部子组也被丢弃.

2.3.3 多输出树组套索模型

已知多元线性回归模型为

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{W} \tag{8}$$

其中 $\boldsymbol{X}\in\boldsymbol{R}^{N\times P}$，$\boldsymbol{Y}\in\boldsymbol{R}^{N\times K}$，$\boldsymbol{W}\in\boldsymbol{R}^{N\times K}$，$\boldsymbol{B}\in\boldsymbol{R}^{P\times K}$，$N$ 为样本数，P 为自变量数，K 为输出变量数. 假设输出变量具有树组结构，将输出变量分组 $G=\{g_j\mid j=1,\cdots,J\}$，一个组 g_j 对应一个节点，给每个节点赋予一个权 w_{g_j}，$\boldsymbol{\beta}_{g_j}$ 表示组 g_j 的子模型向量，则多输出树组套索模型[54,55]的损失函数为 $\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{XB}\|_F^2$，其中 $\|\cdot\|_F$ 表

示 Frobenius 范数,罚函数为 $\lambda \sum_{p\in\{1,\cdots,P\}} \sum_{g_j\in G} w_{g_j} \| \boldsymbol{\beta}_{g_j}^p \|_2$,其分为内部的 $L_{2,1}$ 范数运算和外部的 L_1 范数运算:内部 $L_{2,1}$ 范数运算表示将 $\boldsymbol{B}$ 第 p 行的模型向量划分为树组结构并对每个组进行 L_2 范数运算,然后在组水平上执行 L_1 范数运算,从而实现行内的组稀疏,即从输出角度而言的稀疏;外部的 L_1 范数运算表示将 $\boldsymbol{B}$ 的一行作为一个组,在该组水平上进行 L_1 范数运算,从而实现从变量角度而言的稀疏.因此,与其它组稀疏模型的不同之处在于多输出树组套索模型能同时实现从变量角度而言的稀疏和输出角度而言的稀疏.

2.3.4 小结与分析

变量往往不只具有简单的非重叠组结构,更常见的是变量组之间交叉重叠甚至为偏序关系的树组结构.稀疏学习从简单的套索模型到实现变量组选择的组套索模型,再到变量组具有重叠结构的重叠组套索模型、变量组为偏序关系的树组套索模型、输出变量具有树组结构的多输出树组套索模型,其正沿着结构稀疏化的方向发展,最近很多研究还将结构稀疏学习应用到概率图模型中,所能揭示的模型结构越来越复杂.

表 1 各种组稀疏模型对比

模型	变量(组)选择能力	变量选择一致性	变量组选择一致性
$L_{2,1}$组套索	变量组	—	差
$L_{\infty,1}$组套索	变量组	—	差
自适应组套索	变量组	—	好
标准组套索	变量组	—	差
平方根组套索	变量组	—	差
稀疏组套索	变量+变量组	差	差
贝叶斯组套索	变量组	—	差
L_1 组 SCAD	变量+变量组	差	好
L_2 组 SCAD	变量组	—	好
L_1 组 MC	变量+变量组	差	好
L_2 组 MC	变量组	—	好
复合组 MC	变量+变量组	好	好
L_1 组桥	变量+变量组	差	好
L_2 组桥	变量组	—	好
复合组桥	变量+变量组	好	好
重叠组套索	变量组	—	差
树组套索	变量组	—	差
多输出树组套索	变量组	—	差

注:表 1 中"—"表示该模型无变量选择能力,故讨论其变量选择一致性无意义.

各模型的对比如表 1 所示.除自适应组套索模型和非凸组稀疏模型外,其它组稀疏模型的变量(组)选择一致性较差,只在附加条件下变量(组)选择一致性才较好,这些条件有不可表示条件(Irrepresentable Condition)[56,57]、稀疏 Riesz 条件(Sparse Riesz Condition)[58,59]和限制特征值条件(Restricted Eigenvalue Condition)[60,61]等,其中变量选择一致性的定义为

$$\lim_{N\to\infty} P(\{p:\hat{\beta}_p\neq 0\}=\{p:\beta_p\neq 0\})=1 \tag{9}$$

变量组选择一致性的定义为

$$\lim_{N\to\infty} P(\{j:\hat{\boldsymbol{\beta}}_j\neq 0\}=\{j:\boldsymbol{\beta}_j\neq 0\})=1 \tag{10}$$

3 组稀疏模型的求解

组稀疏模型的求解常分为两步:对目标函数进行预处理,将不平滑、非凸、变量块不可分离的目标函数向平滑、凸、变量块可分离的方向转化,然后对转换后的目标函数进行求解.另外,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.

3.1 预处理方法

常见预处理方法有 Nesterov 平滑近似[62]、局部二次近似(Local Quadratic Approximation,LQA)[36]、局部线性近似(Local Linear Approximation,LLA)[63]、对偶范数和对偶函数[50,55]等.Nesterov 平滑近似的优点为将不平滑的优化问题转化为平滑的优化问题.对偶范数和对偶函数方法可解决变量块不可分离问题.LQA 和 LLA 用于组 SCAD 模型等非凸模型的目标函数的预处理,其中 LQA 方法将原目标函数用一个凸二次函数近似表示,但由于涉及海森矩阵的重复求逆问题因而计算复杂度大,其还需要设定一个对算法的收敛性影响很大的初始解,其另一缺点是一旦某变量被剔除,该变量将不再出现在最后的模型中.LLA 将原目标函数近似为一次函数,其也需要设定一个初始解,而且初始解的设定对算法性能影响很大.

3.2 求解算法

常见的求解算法有组最小角回归(Group Least Angle Regression,GLAR)[2]、块坐标下降(Block Coordinate Descent, BCD)[2,64,65]、块坐标梯度下降(Block Coordinate Gradient Descent, BCGD)[26,66]、谱投影梯度法(Spectral Projected Gradient, SPG)[6]、活动集方法(Active Set)[67]和轮换方向乘子法(Alternating Direction Method of Multipliers, ADMM)[68].GLAR 的前身是最小角回归(Least Angle Regression,LAR)[69],其能得到权衡参数在整个取值范围内变化时的整个解路径,这是其相对于其它算法的突出优点,但其只适用于解路径为分段线性的模型.BCD 在求解优化问题时每次只涉及单个坐标块,固定其余全部坐标块,因此大大简化了优化问题,但其只适用于目标函数为变量块可分离的模型.BCGD 首先将优化问题的目标函数利用一个严格凸二次函数近似,然后对该近似函数执行块坐标下降算法求解梯度方向,再利用非精确的线搜索结合上一步求解的梯度方向执行梯度下降步骤,其具有高度并行化的特点,适用于求解大规模问题.活动集方法一般用于大规模复杂问题的求解,利用最优性条件把大规模复杂问题分解为一系列

简单子问题的求解.ADMM 为增广拉格朗日方法的推广,该方法通过引入辅助变量将难求解的原问题分解为若干便于求解的子问题,并且每个子问题的解都是显式解,其适用范围非常广,可用来求解套索模型、$L_{2,1}$组套索模型、重叠组套索模型和树组套索模型等.SPG 在迭代过程中采用非单调线搜索技术,不要求每次迭代后目标函数值都下降,只要求在规定的最近某些次迭代时目标函数的值下降即可.

4 结论和展望

本文总结了各种组稀疏模型,还对各种组稀疏模型求解前的预处理方法和优化求解算法进行了归纳总结,并给出了组稀疏模型未来的研究方向.组稀疏模型是当前高维数据建模的重要研究方向,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位置.但该领域还存在一些有待于研究的问题:

问题 1:将组稀疏模型扩展到 Probit 回归模型、索引模型(Index Model)、部分线性模型(Partially Linear Models)、变系数模型(Varying Coefficient Models)、加速失效时间模型(Accelerated Failure Time Model)等情形,极大地丰富组稀疏模型.

问题 2:使用 AIC 等判据选择模型的权衡参数时,要求估计噪声变量的协方差矩阵和自由度,虽然交叉校验方法不需要估计噪声变量的协方差矩阵和自由度,但是交叉校验方法会导致很大的计算复杂度.Meinshausen 等人[70]基于重抽样(resample)方法提出选择模型的权衡参数的稳定选择(Stability Selection)方法,该方法不要求对噪声变量的协方差矩阵和自由度进行估计,未来需要做的工作是将该方法应用到组稀疏模型的权衡参数选择中,并与 Cp 判据、BIC 判据、AIC 判据、GCV 准则和交叉校验等方法进行对比分析.

问题 3:目前尚无学者将 MCBP 罚(Minimax Concave Bridge Penalty)[71]、对数罚[72]、稀疏桥罚(Sparse Bridge Penalty)[73]和反正切罚[74]等非凸罚用于构造重叠组结构的组稀疏模型,我们猜想利用其构造出的重叠结构组稀疏模型应该同时具有变量选择和变量组选择能力,但该问题有待于进一步研究.

问题 4:树组套索、多输出树组套索、L_1 组 SCAD 模型和 L_1 组 MC 模型的变量选择一致性有待于研究,尤其在变量数远远大于样本数时的一致性有待于研究.另外,最近文献[75~77]中给出了比原限制特征值条件更强的条件,组稀疏模型在这些更强的条件下的变量选择一致性和变量组一致性如何?

参考文献

[1] Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society: Series B, 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, 2006, 68(1): 49 – 67.

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

[4] Tropp J A. Algorithms for simultaneous sparse approximation [J]. Signal Processing, 2006, 86(3): 589 – 602.

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

[6] Schmidt M W, Murphy K P, Fung G, Rosales R. Structure learning in random fields for heart motion abnormality detection [A]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition[C]. Alaska, USA: IEEE, 2008. 1 – 8.

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

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

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

[10] Sra S. Fast projections onto $\ell_{1,q}$-norm balls for grouped feature selection[J]. Lecture Notes in Computer Science, 2011: 305 – 317.

[11] Kowalski M. Sparse regression using mixed norms[J]. Applied and Computational Harmonic Analysis, 2009, 27(3): 303 – 324.

[12] Rakotomamonjy A, et al. Lp-Lq penalty for sparse linear and sparse multiple kernel multi-task learning[J]. IEEE Transactions on Neural Networks, 2011, 22(8): 1307 – 1320.

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

[14] 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.

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

[16] Wang H, Leng C. A note on adaptive group lasso[J]. Computational Statistics and Data Analysis, 2008, 52(12): 5277 –

5286.

[17] Wei F, Huang J. Consistent group selection in high-dimensional linear regression[J]. Bernoulli, 2010, 16(4): 1369 – 1384.

[18] Zou H. The adaptive lasso and its oracle properties[J]. Journal of the American statistical association, 2006, 101(476): 1418 – 1429.

[19] Zhang H H, Lu W. Adaptive lasso for Cox's proportional hazards model[J]. Biometrika, 2007, 94(3): 691 – 703.

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

[21] Simon N, et al. A sparse-group lasso[J]. Journal of Computational and Graphical Statistics, 2013, 22(2): 231 – 245.

[22] Chatterjee S, Steinhaeuser K, Banerjee A, et al. Sparse group lasso: consistency and climate applications[A]. Proceedings of the 12th SIAM International Conference on Data Mining[C]. California, USA: Omnipress, 2012.47 – 58.

[23] Zhu X, Huang Z, et al. Video-to-shot tag allocation by weighted sparse group lasso[A]. Proceedings of the 19th ACM International Conference on Multimedia[C]. Scottsdale: ACM, 2011.1501 – 1504.

[24] Raman S, Fuchs T J, Wild P J, Dahl E, Roth V. The Bayesian group-Lasso for analyzing contingency tables[A]. Proceedings of the 26th Annual International Conference on Machine Learning[C]. Montreal, Canada: ACM, 2009.881 – 888.

[25] Chandran M. Analysis of Bayesian group-Lasso in regression models[D]. Florida, USA: University of Florida, 2011.

[26] Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression[J]. Journal of the Royal Statistical Society: Series B, 2008, 70(1): 53 – 71.

[27] Kim Y, Kim J, et al. Blockwise sparse regression[J]. Statistica Sinica, 2006, 16(2): 375 – 390.

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

[29] Wu F, Yuan Y, Zhuang Y. Heterogeneous feature selection by group lasso with logistic regression[A]. Proceedings of the International Conference on Multimedia[C]. Firenze, Italy: ACM, 2010.983 – 986.

[30] Wang S, et al. Hierarchically penalized Cox regression with grouped variables[J]. Biometrika, 2009, 96(2): 307 – 322.

[31] 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: 80 – 89.

[32] Ji Y, Lin N, Zhang B. Model selection in binary and tobit quantile regression using the Gibbs sampler[J]. Computational Statistics & Data Analysis, 2012, 56(4): 827 – 839.

[33] Yin J, Chen X, Xing E P. Group sparse additive models[A]. Proceedings of the 29th International Conference on Machine Learning[C]. Scotland, UK: Omnipress, 2012.871 – 878.

[34] Jiang D F. Concave selection in Generalized linear models [D]. Iowa, USA: University of Iowa, 2012.

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

[36] 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.

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

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

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

[40] Huang J, Ma S, Xie H, Zhang C H. A group bridge approach for variable selection[J]. Biometrika, 2009, 96(2): 339 – 355.

[41] Ma S, Huang J, Song X. Integrative analysis and variable selection with multiple high-dimensional data sets[J]. Biostatistics, 2011, 12(4): 763 – 775.

[42] Ma S, Dai Y, Huang J, et al. Identification of breast cancer prognosis markers via integrative analysis[J]. Computational Statistics & Data Analysis, 2012, 56(9): 2718 – 2728.

[43] Seetharaman I. Consistent bi-level variable selection via composite group bridge penalized regression[D]. Kansas, USA: Kansas State Univesity, 2013.

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

[45] 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.

[46] Mosci S, et al. A primal-dual algorithm for group sparse regularization with overlapping groups[A]. Proceedings of Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems [C]. Canada: Curran Associates, 2010.2604 – 2612.

[47] Percival D. Theoretical properties of the overlapping groups Lasso[J]. Electronic Journal of Statistics, 2012, 6: 269 – 288.

[48] 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.

[49] Percival D. Theoretical properties of the overlapping groups lasso[J]. Electronic Journal of Statistics, 2012, 6: 269 – 288.

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

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

[52] Martins A F T, Smith N A, et al. Structured sparsity in structured prediction[A]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing[C]. Massachusetts, America: ACL, 2011. 1500 – 1511.

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

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

[55] 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.

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

[57] Bach F R. Consistency of the group lasso and multiple kernel learning[J]. The Journal of Machine Learning Research, 2008, 9: 1179 – 1225.

[58] 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.

[59] Wei F, Huang J. Consistent group selection in high-dimensional linear regression[J]. Bernoulli, 2010, 16(4): 1369 – 1384.

[60] 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.

[61] Lounici K, Pontil M, Van D G, Tsybakov A B. Oracle inequalities and optimal inference under group sparsity[J]. The Annals of Statistics, 2011, 39(4), 2164 – 2204.

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

[63] Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models[J]. Annals of statistics, 2008, 36(4): 1509 – 1533.

[64] Yang Y, Zou H. A fast unified algorithm for solving group-lasso penalized learning problems[J]. Journal of Computational and Graphical Statistics, 2012: 328 – 361.

[65] Qin Z, Scheinberg K, Goldfarb D. Efficient block-coordinate descent algorithms for the group lasso[J]. Mathematical Programming Computation, 2013: 143 – 169.

[66] Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization[J]. Mathematical Programming, 2009, 117: 387 – 423.

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

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

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

[70] Meinshausen N, et al. Stability selection[J]. Journal of the Royal Statistical Society: Series B, 2010, 72(4): 417 – 473.

[71] Choon C L. Minimax concave bridge penalty function for variable selection[D]. Singapore: National University of Singapore, 2012.

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

[73] Kwon S, Kim Y, Choi H. Sparse bridge estimation with a diverging number of parameters[J]. Statistics and Its Interface, 2012, 6: 231 – 242.

[74] Candes E J, Wakin M B, Boyd S P. Enhancing sparsity by reweighted L_1 minimization[J]. Journal of Fourier Analysis and Applications, 2008, 14(5 – 6): 877 – 905.

[75] Van D G, Bühlmann P. On the conditions used to prove oracle results for the lasso[J]. Electronic Journal of Statistics, 2009, 3: 1360 – 1392.

[76] Zhang T. Some sharp performance bounds for least squares regression with L_1 regularization[J]. The Annals of Statistics, 2009, 37(5A): 2109 – 2144.

[77] Ye F, Zhang C H. Rate minimaxity of the lasso and Dantzig selector for the Lq loss in Lr balls[J]. The Journal of Machine Learning Research, 2010, 11: 3519 – 3540.

作者简介

刘建伟 男,1966 年出生.博士,中国石油大学(北京)副研究员,主要研究方向包括智能信息处理、机器学习、算法分析与设计等.

E-mail: liujw@cup.edu.cn

崔立鹏 男,1990 年出生.2012 年毕业于河北大学自动化专业,现为中国石油大学(北京)地球物理与信息工程学院硕士研究生,研究方向为机器学习.