Advances in Applied Mathematics 应用数学进展, 2019, 8(8), 1478-1486
Published Online August 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.88173

Hans汉斯

Several Types of Group Variable Selection Methods and Block Coordinate Descent Algorithm

Chunhong Li, Xiaomin Zhong*, Ruixue Zong

School of Mathematics and Information Science, Guangxi University, Nanning Guangxi
Email: 147113569@qq.com, *846345642@qq.com

Received: August 1st, 2019; accepted: August 16th, 2019; published: August 23rd, 2019

Abstract

In complex data, variables often appear in groups. Considering the penalty terms of four different models selected by Lasso, SCAD, Bridge and MCP, their methods in group variables and their block coordinate descent algorithm are studied. The simulation is carried out under the condition of Logistic model. The results show that the Composite MCP penalty method is superior to the other three group punishment methods in predicting ability and variable selection, and is applied to the company advertising data for saling network office software. In the method, the Composite MCP method is the best in the advertisement conversion research. By comparing and selecting the group structure and individual variables that affect the advertisement conversion, it provides a reasonable basis for selecting the delivery strategy.

Keywords

Group Variable Selection, Block Coordinate Descent Algorithm, Logistic Model, Ad Conversion

几类群组变量选择方法及其块坐标下降算法

李春红，钟小敏*，宗瑞雪

广西大学数学与信息科学学院，广西 南宁
Email: 147113569@qq.com, *846345642@qq.com

收稿日期：2019年8月1日；录用日期：2019年8月16日；发布日期：2019年8月23日

摘　要

在复杂数据中变量往往成组出现，考虑了Lasso、SCAD、Bridge及MCP四种不同模型选择的惩罚项，研

*通讯作者。

文章引用：李春红，钟小敏，宗瑞雪．几类群组变量选择方法及其块坐标下降算法[J]．应用数学进展，2019，8(8): 1478-1486. DOI: 10.12677/aam.2019.88173

究了它们在群组变量中的方法及其块坐标下降算法，在Logistic模型的条件下进行模拟，结果表明Composite MCP组惩罚方法在预测能力和变量选择上均优于其他三种群组惩罚方法，并运用到销售网络办公软件公司的广告数据中，结果表示四种方法中Composite MCP方法在广告转化研究中效果是最优的，通过比较，选择出影响广告转化的群组结构及单个变量，为选择投放策略提供合理的依据。

关键词

群组变量选择，块坐标下降算法，Logistic模型，广告转化

1. 引言

科技的进步使得数据收集变得更容易，但也因此导致数据库规模更大、更加复杂。群组变量选择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、信息处理、计算机视觉等领域具有广阔的应用前景。在回归建模中，解释变量大多是存在分组结构的，分组结构可能由于多种原因，并产生完全不同的建模目标。常见的例子包括通过一组指标变量在回归模型中表示多级分类协变量，以及通过一组基函数表示连续变量的效果。利用具有科学意义的先验知识，分组也可以引入模型中。例如，在基因表达分析中，属于相同生物途径的基因可以被认为是一组，在遗传关联研究中，来自同一基因的遗传标记可以被认为是一组。

变量选择思想起源于 Hoed 和 Kenny 提出的岭回归估计，1996 年 Lasso [1] (Least Absolute Shrinkage and Selection Opereator)提出了罚函数，因为缺乏有效的算法所以并没有得到广泛的应用，Efrom [2]等提出的最小角算法(lars)有效解决了这个问题，但是 Lasso 的解不相合，随后 SCAD (Smoothly Clipped Absolute Deviation)和 MCP (Minimax Concave Penalty)惩罚模型的提出有效克服了 Lasso 方法不具有的 Oracle 性质，而这两种方法均是非凸罚函数[3] [4] [5]，同样是非凸罚函数的还有 Bridge 惩罚模型[6]，但是由于 Bridge 是 l_p 罚函数的特点，导致其可操作性差。这几类模型选择方法均只能进行单变量选择而不能进行群组变量选择。通常数据中的变量均具有分组结构，相关的选择问题就演变成选择组而不仅仅是单个变量，在回归模型中，多级分类变量由一组指标变量表示；连续变量可以由多项式的线性组合表示。在这种情况下，变量选择通常对应于整个变量组的选择，而不是单个派生变量。因此学者们将单变量选择的思想扩展到群体选择问题上。2006 年 Yuan 和 Lin 最早提出了能进行组变量选择的 Group Lasso 方法[7]，借鉴 Group Lasso 的思想，将 SCAD、MCP、Bridge 推广到 Group SCAD，Group MCP、Composite MCP 和 Group Bridge 等，以克服 Group Lasso 不具备 Oracle 性质的缺点，并且 Composite MCP 和 Group Bridge 可同时选择重要的群组以及识别组内的重要变量[8] [9] [10]。

最小角算法的提出使得 Lasso 为人所知，随后 Group Lasso 提出的相应的最小角算法也推广成为组最小角算法，但是这两种算法只是用于解路径分段线性的正则模型。若目标函数中变量块之间是独立的且不包含相同的自变量，则组最小角算法就不适应了。因而对坐标下降算法的改进推广称为块坐标下降算法，有效解决了这个问题[7]。此方法同样适用于 Group SCAD，Group MCP、但是由于 SCAD、MCP 罚函数是非凸的，需相应地转化成凸的模型求解。Group Bridge 和 Composite MCP 需要对群组和组内变量进行选择，

所以在算法上需要分成内部循环和外部循环进行，并且需要适当做一些调整，例如一阶泰勒展开等。

群组变量选择方法在实际中应用广泛，学者在这方面的研究也层出不穷。在销售网络办公软件公司的广告数据中，研究影响广告转化的因素涉及的变量存在分组结构。本文研究几种群组变量选择方法并将其应用到 Logistic 模型中，研究该问题的影响因素、选择显著的广告转化群组结构和单个变量，并比较几种群组变量选择方法的优良性。

2. 群组变量选择方法

2.1. Logistic 回归模型

假设 (x_i, y_i) 为独立同分布的观测值，其中 $i=1,2,\cdots,n$，$x=(x_1,x_2,\cdots,x_i)^{\mathrm{T}}$ 为解释变量，令 $y=(y_1,y_2,\cdots,y_n)$，$y_i \in \{0,1\}$。Logistic 回归模型为：

$$\pi(x)=p(y=1|x)=\frac{\mathrm{e}^{\eta}}{1+\mathrm{e}^{\eta}}=\frac{\mathrm{e}^{x^{\mathrm{T}}\beta}}{1+\mathrm{e}^{x^{\mathrm{T}}\beta}} \tag{1}$$

其中 $\beta=(\beta_0,\beta_1,\cdots,\beta_p)^{\mathrm{T}}$，$\beta_0$ 为常数，$\beta_1,\cdots,\beta_p$ 表示回归系数。在 Logistic 回归分析中通常是通过最大似然法实现参数估计，Logistic 回归模型的对数似然函数为：

$$l(\beta)=\sum_{i=1}^{n}\left(y_i x_i^{\mathrm{T}}\beta-\log\left(1+\exp\left(x_i^{\mathrm{T}}\beta\right)\right)\right) \tag{2}$$

在对数似然函数中加入不同的群组变量选择的惩罚项便可得到不同的变量选择方法。

2.2. 群组变量选择方法

1) Group Lasso 方法

Group Lasso 方法对 Lasso 的惩罚项做了修改，是 Lasso 方法的推广，在高维数据中，能够选择自变量分成 m 个不同组的整组变量，如果每个组中只包含一个变量，则目标函数(3)简化为通常的 Lasso 解决，所以该方法应用也更加广泛。即使模型复杂度随着样本量的增加而增加 Group Lasso 估计也渐近一致[11]。但 Group Lasso 与 Lasso 具有相同的缺点，即变量选择不一致并且倾向于过度收缩大系数。出现这些缺点是因为 Group Lasso 的惩罚率不随组系数的大小而变化，这导致大系数的偏差估计，为了补偿过度收缩，组套索倾向于将虚假系数包括在模型中。自适应组套索通过应用不同的调整参数并因此对分组系数应用不同的收缩量来弥补这些缺点，就像自适应套索对个体协变量一样。但是自适应组套索仍然没有完成双层选择[12] [13]。Group Lasso 在 Logistic 模型下其估计为：

$$\hat{\beta}^{\mathrm{GLasso}}=\arg\min_{\beta}\left\{-l(\beta)+\lambda\sum_{j=1}^{J}\sqrt{p_j}\left\|\beta_j\right\|_2\right\} \tag{3}$$

其中 β_j 表示第 l 组的回归系数，p_j 为 β_j 的长度，λ 为调节参数。$\|\cdot\|_2$ 是 2-范数。

2) Group SCAD 方法

2007 年 Group SCAD 是由 wang [9]等提出，并证明了其低维 Oracle 性质，Group SCAD 惩罚方法加载 Logistic 模型下其估计为：

$$\hat{\beta}^{\mathrm{GLasso}}=\arg\min_{\beta}\left\{-l(\beta)+\lambda\sum_{j=1}^{J}\sqrt{p_j}\left\|\beta_j\right\|_2\right\} \tag{4}$$

其中 $P_{\lambda,a}(\beta)$ 是 SCAD 的惩罚函数形式如下：

$$p_{\lambda,a}(\beta)=\begin{cases}\lambda|\beta| & 0\le|\beta|\le\lambda\\ -\dfrac{\beta^2-2a\gamma|\beta|+\gamma^2}{2(a+1)} & \lambda\le|\beta|a\gamma\\ \dfrac{(a+1)\lambda^2}{2} & \text{其他}\end{cases} \tag{5}$$

这里 $a\ge 2,\lambda\ge 0$，当 $\lambda\to\infty$ 时 Group SCAD 惩罚模型就变成了 Group Lasso 形式，惩罚函数(5)是一种非凸的惩罚函数，该函数在 $a\lambda$ 出有两个节点，其中 a 是另一个调整参数，Fan 表明使用 $a=3.7$ 时效果最好。由于在组水平上使用了非凸的 SCAD 惩罚函数，因此 Group SCAD 惩罚模型具有群组变量选择一致性。

3) Group Bridge 方法

双层变量选择的有趣之处是不仅能筛选出重要的群组结构还能筛选出重要的单个变量，Group Bridge 方法是最早实现双层变量选择方法，Group Bridge 结合了两种惩罚，即群组结构选择是 Bridge 惩罚和群组内单变量选择是 Lasso 惩罚。Group Bridge 在某些正则条件下具有群组 Oracle 性质，但是组内不具备相合性。Group Bridge 允许存在单个大预测器以持续降低其组中其他变量的进入阈值。此属性可防止 Group Bridge 在选择单个变量时实现一致性。假设已知有 J 组分组，$M_1M_2\cdots M_J$，每组变量数有 j 个，令 $\beta_{M_j}=(\beta_j)_{j\in M_j}$ 是 β 对应变量的子向量，所以在 Logistic 模型下的惩罚估计如下：

$$\hat{\beta}^{\text{Gbridge}}=\arg\min_{\beta}\left\{-l(\beta)+\lambda\sum_{j=1}^{J}\tau_j\left\|\beta_{M_j}\right\|_1^{\gamma}\right\} \tag{6}$$

其中 λ 是罚参数，τ_j 是 β_{M_j} 的调整参数，当 $0<\gamma<1$ 即可实现双重变量选择。

4) Composite MCP 方法

Composite MCP 的一个有趣的特殊情况是使用 MCP 作为外部和内部惩罚，(这种惩罚在 Huang 和 *Breheny* 中被称为 Group MCP)，而用 Composite MCP 不仅能更好地反映框架而且只能进行组间选择的 Group MCP 区分开来。Composite MCP 的两个特点为通过允许协变量增大来避免过度收缩，以及允许组内部保持稀疏。并且具有组间向量选择和组内分量选择一致性。Composite MCP 在 Logistic 模型下的估计为：

$$\hat{\beta}^{\text{CMCP}}=\arg\min_{\beta}\left\{-l(\beta)+\sum_{l=1}^{m}f_{\lambda,b}\left(\sum_{l=1}^{m}f_{\lambda,a}(|\beta_{lk}|)\right)\right\} \tag{7}$$

其中

$$f_{\lambda,\omega}(\beta)=\begin{cases}\lambda\beta-\dfrac{\beta^2}{2\omega},\beta\le\omega\lambda\\ \dfrac{1}{2}\omega\lambda^2,\beta\le\omega\lambda\end{cases} \tag{8}$$

式子(8)是 MCP 的惩罚形式，a,b 分别是组外和组内的调整参数，$\lambda\ge 0$，β_{lk} 表示的是第 l 组的第 k 个变量的回归系数。对于逻辑回归损失函数的 Composite MCP 惩罚，使用值 $a=30$ 时响应变量总是在相同的范围内；因此无论是在模拟研究中还是实例分析中，$a=30$ 似乎是我们遇到的所有逻辑回归问题的适当值。基于上述四种群组变量选择方法都可以用算法实现下面给出对应算法研究。

3. 块坐标下降算法

块坐标下降算法是一种用于拟合具有分组惩罚模型的有效方法，这个算法基于坐标下降算法提出的，目的是计算 Group Lasso 的解，块坐标下降算法是一个优化单个组变量的目标函数并循环通过这些组直到收敛，这种算法适用于 Group Lasso，Group SCAD 模型，因为这两种模型都具有单组模型的简单闭合表达式。根据线性模型下的 Group Lasso 算法的闭合表达式为：$s(z,\lambda)=s(\|z\|,\lambda)\frac{z}{\|z\|}$，该表达式用于计算 Group Lasso 的解，它是软阈值算子的多变量版本，软阈值应用于向量的长度，同时保持其方向不变[8]。将线性模型推广到 Logistic 模型中 Group Lasso 算法的闭合表达式为：

$$s(z_j,\lambda_j)=\frac{1}{v}s(v\|z_j\|,\lambda_j)\frac{z_j}{\|z_j\|} \tag{9}$$

以下呈现的是 Group Lasso 在 Logistic 模型中的算法过程：

步骤一：初始化 $\beta=\beta^0$

步骤二：更新 $\eta=X^{\mathrm{T}}\beta$ 和 $\pi=\sum_{i=1}^{n}\frac{e^{\eta_i}}{1+e^{\eta_i}}$

步骤三：计算残差向量 $r=(y-\pi)/v$

步骤四：迭代直至收敛：$j=1,2,\cdots,J$

a) 更新 $z_j=x_j^{\mathrm{T}}r+\beta_j$

b) 更新 $\tilde{\beta}^{(j)}\leftarrow s(uz_j,\lambda_j)/v$

c) 更新 $r'=r-x_j^{\mathrm{T}}\left(\tilde{\beta}^{(j)}-\beta^{(j)}\right)$

重复步骤三和步骤四直到收敛。

群组变量选择是基于不同的惩罚项提出的，其计算过程主体相似，可以通过置换步骤四中的更新步骤，得到不同算法。相应的 Group SCAD 更新表达式为：

$$\tilde{\beta}^{(j)}=\frac{1}{v}F_s(uz_j,\lambda_j,\gamma)=\frac{1}{v}F_s(u\|z_j\|,\lambda_j,\gamma) \tag{10}$$

当 $0<\gamma<1$ 时 Group Bridge 模型是非凸的需要将转化成凸的形式求解，即将原模型的问题求解转化成下式：

$$\arg\min\left\{-l(\beta)+\sum_{j=1}^{J}\theta_j^{1-1/\gamma}c_j^{1/\gamma}\|\beta_j\|_1+\sum_{j=1}^{J}\tau\theta_j\right\} \tag{11}$$

该式的解 $\hat{\beta}$ 就是 Group Bridge 模型的解。求解过程可分为两层循环，外层循环即利用 β_j 和 θ_j 循环变量寻求最优方法；内部循环采用块坐标下降算法，从 $j=1$ 迭代到 $j=J$，下列为内部循环 θ_j 的更新公式[14]：

$$\begin{aligned}&\theta_j=\arg\min-l(\beta)+\theta_j^{1-1/\gamma}c_j^{1/\gamma}\|\beta_j\|_1+\tau\theta_j\\&\Rightarrow\theta_j=c_j\left(\frac{1-\gamma}{\tau\gamma}\right)^{\gamma}\|\beta_j\|_1^{\gamma}\end{aligned} \tag{12}$$

β_j 的更新公式：

$$\theta_j=\arg\min_{\beta_j\in R^{d_j}}-l(\beta)+\theta_j^{1-1/\gamma}c_j^{1/\gamma}\|\beta_j\|_1+\tau\theta_j \tag{13}$$

而上式相当于求解 Lasso 惩罚，可以直接利用最小角算法求解。

Composite MCP 的求解方法也是采用块坐标下降法，但是由于惩罚函数的特殊性首先需要对原惩罚项进行特殊预处理，在块坐标下降算法执行的 $i+1$ 轮迭代中，将 Composite MCP 惩罚函数 $f_{\lambda,b}\left(\sum_{l=1}^{m} f_{\lambda,a}\left(\left|\beta_{lk}\right|\right)\right)$ 在 $\left|\beta_{lk}^{(\tilde{t})}\right|$ 处进行一阶泰勒展开[10]：

$$f_{\lambda,b}\left(\sum_{l=1}^{m} f_{\lambda,a}\left(\left|\beta_{lk}^{(\tilde{t})}\right|\right)\right)+f'_{\lambda,b}\left(\sum_{l=1}^{m} f_{\lambda,a}\left(\left|\beta_{lk}^{(\tilde{t})}\right|\right)\right)\cdot f'_{\lambda,a}\left(\left|\beta_{lk}^{(\tilde{t})}\right|\right)\left(\beta_{lk}^{(\tilde{t}+1)}-\beta_{lk}^{(\tilde{t})}\right) \tag{14}$$

其中 $\left|\beta_{lk}^{(\tilde{t})}\right|$ 和 $\left|\beta_{lk}^{(\tilde{t}+1)}\right|$ 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 $\left|\beta_{lk}^{(\tilde{t})}\right|$ 可以被看作常数，常常可以忽略不计。从而得到惩罚函数的近似表达式：

$$f'_{\lambda,b}\left(\sum_{k^*=1}^{K} f_{\lambda,a}\left(\left|\beta_{lk}^{*(\tilde{t})}\right|\right)\right)f'_{\lambda,a}\left(\left|\beta_{lk}^{(\tilde{t})}\right|\right)\left|\beta_{lk}^{(\tilde{t}+1)}\right| \tag{15}$$

(15)式是关于 $\left|\beta_{lk}^{(\tilde{t})}\right|$ 的线性函数，此时求解最优化问题相当于求解一个 Group Lasso。

4. 模拟研究

考虑本文研究的 Logistic 模型，其条件概率可以表示为：

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+X\beta+\varepsilon$$

其中 $X_i\sim N(0,\Sigma)$，$\operatorname{cov}\left(X_i,X_j\right)=0.1^{|i-j|}$，误差项 $\varepsilon\sim N(0,1)$。

考虑组结构的不同，分别对以下两种情况的组结构进行模拟研究：

结构一：考虑解释变量之间存在组结构、组内没有零系数的情况，$J=60,P=300$，考虑变量组数为 60 组，每组 5 个变量，回归参数为 $\beta_1=(0.5,1,1.5,2,2.5)^{\mathrm{T}}$，$\beta_2=(2,2,2,2,2)^{\mathrm{T}}$，$\beta_3=(-3,-1,2,1,3)^{\mathrm{T}}$，$\beta_4=\cdots=\beta_{60}=(0,0,0,0,0)^{\mathrm{T}}$。

结构二：考虑解释变量之间存在组结构但是具体分组不同，组内存在零系数的情况，$J=74,P=300$，考虑变量组数为 60 组，前 4 组每组 5 个变量，后 70 组大小为 4 回归参数为：$\beta_1=(-3,-2,-1,1,2)^{\mathrm{T}}$，$\beta_2=(-3,-2,2,1,0)^{\mathrm{T}}$，$\beta_3=(1,2,0,0,0)^{\mathrm{T}}$，$\beta_4=(0,0,0,0,0)^{\mathrm{T}}$，$\beta_5=(0.5,1,1.5,2)^{\mathrm{T}}$，$\beta_6=(2,2,0,0)^{\mathrm{T}}$，$\beta_7=(-3,0,0,0)^{\mathrm{T}}$，$\beta_8=\cdots=\beta_{74}=(0,0,0,0)^{\mathrm{T}}$。

通过计算机模拟上述两种数据类型，样本容量取 $n=200$ 和 $n=500$，重复 1000 次实验，该模拟实验借助 *R* 语言中的 grperg 数据包实现，取相应的平均值作为参考。选取模型误差，正确选择 0 (真实为 0，估计也为 0 的变量个数)，错误选择 0 (真实为 0，估计不为 0 的变量个数)，选取组数的四项平均值作为测度来比较文中四种群组变量选择方法。结果如表 1 所示，结果表明 Group Bridge 和 Composite MCP 在模型误差、正确选择 0，错误选择 0、选取组数四个方面平均值上表现明显比 Group Lasso 和 Group SCAD 效果要好，Group Lasso 方法会选择更多的变量，把更多不重要变量选入模型，模型可解释性不高，Group SCAD 相对好一些而对比两种双层变量选择方法，Composite MCP 方法表现更好，并且随着样本容量增大预测越准确。

5. 实例分析

来源于某销售网络办公软件广告数据，随机选取了 2018 年 7 月至 8 月的 15000 条数据。利用组变量

选择方法加载到 Logistic 回归模型中对广告成本因素、广告性质、广告访问数据、广告历史数据、广告浏览数据 、广告点击数据 6 组共 21 个影响广告转化的因素进行数据拟合估计。

由于部分变量是数值数据，它们单位不一致，并且取值有较大差异，因此我们首先对数值数据进行标准化，然后通过 R 软件得到各个组变量选择方法对系数压缩后的结果见表 2。并且分别计算了不同方法下的均方根误差(RMSE)和 AUC 值结果见表 3。

Table 1. Simulation results
表 1. 模拟结果

方法	$n = 200$				$n = 500$			
	模型误差	正确选择 0	错误选择 0	选取组数	模型误差	正确选择 0	错误选择 0	选取组数
结构一：$J = 60, P = 300$								
GLasso	1.213	243.28	0.00	5.72	0.726	248	0.00	7.25
GSCAD	0.725	253.32	0.00	5.56	0.614	256.40	0.00	5.32
GBridge	0.604	280.54	0.04	3.09	0.356	281.06	0.03	3.03
CMCP	0.585	278.40	0.06	3.13	0.273	283.22	0.04	3.09
结构二：$J = 74, P = 300$								
GLasso	2.751	75.68	0.00	7.48	1.102	98.22	0.00	7.25
GSCAD	2.323	101.44	0.24	7.13	0.925	108.56	0.00	7.12
GBridge	2.019	256.32	0.63	6.43	0.617	292.18	0.21	6.38
CMCP	1.838	272.68	0.78	6.25	0.541	286.32	0.34	6.19

Table 2. Estimates by method factor
表 2. 各方法系数估计值

		Group Lasso	Group SCAD	Group Bridge	Composite MCP
	截距	−0.5512	−0.2625	−0.5174	−0.1281
*x*1 成本因素	*x*11 曝光数	0.4012	0.3258	0.2982	0.3215
	*x*12 平均点击价格	1.1531	1.0246	1.1428	1.1433
	*x*13 平均转化成本	0.0327	0.0436	0	0
	*x*14 平均位置排名	2.4502	2.3125	2.2985	3.0136
*x*2 广告性质	*x*21 关键词个数	−0.0325	−0.1001	−0.0958	−0.1032
	*x*22 是否含商标	0.2018	0.2249	0.2014	0.2257
	*x*23 是否含动词	−0.0362	−0.0237	0	0
	*x*24 是否包含价格	0.1028	0.1671	0.2749	0.3518
	*x*25 是否涉及地域信息	−0.1335	−0.1264	−0.2071	−0.2403
	*x*26 是否为特定产品	0.1243	0.1127	0.2180	−0.2349
*x*3 访问数据	*x*31 平均访问页数	1.0028	1.1025	1.1079	1.4513
	*x*32 访问次数	1.2407	0.9678	1.1009	1.0005
	*x*33 平均访问时长	−0.8721	−0.9124	−0.7149	−0.7706
	*x*34 跳出率	−0.1035	−0.1257	−0.1795	−0.1628

Continued

	*x*41 历史点击率	0	0	0	0
*x*4 历史数据	*x*42 历史转化率	0	0	0.0641	0.0776
	*x*43 转化次数	0	0	0.0103	0
*x*5 浏览数据	*x*51 展现量	0	0	0	0
	*x*52 浏览量	0	0	0	0
*x*6 点击数据	*x*61 点击量	0.0125	0	0	0
	*x*62 点击率	0.0317	0	0	0.0742

Table 3. Model comparison
表 3. 模型比较

	AUC	RMSE
Group Lasso	0.803	0.032
Group SCAD	0.809	0.032
Group Bridge	0.826	0.030
Composite MCP	0.858	0.029

从表 1 可得：Group Lasso 方法选择了 4 组 17 个变量，表示转化价格、广告性质、访问数据、点击数据是影响广告转化的组变量，而 Group SCAD 选择方法把点击数据这一组变量系数都压缩为 0，该方法认为这一组变量对广告转化影响不大。Group Bridge 和 Composite MCP 都保留了 15 个变量，但是 Composite MCP 保留了更多的组内信息。

从表 2 可得：Group Lasso 方法选择了 4 组 17 个变量，表示转化价格、广告性质、访问数据、点击数据是影响广告转化的组变量。而 Group SCAD 选择方法把点击数据这一组变量系数都压缩为 0 ，该方法认为这一组变量对广告转化影响不大。Group Bridge 和 Composite MCP 都保留了 15 个变量，但是 Composite MCP 保留了更多的组内信息。

表 3 可得：几种方法的均方误差都比较小，表明几种方法在该问题上都表现良好，而 Composite MCP 的均方误差相对最小，并且 Composite MCP 的 AUC 值为 0.858，相对其他方法数值偏大。因此可认为在几种组变量选择方法中，Composite MCP 方法表现最佳。

投放广告者目的就是想让广告转化成为公司的盈利，不转化的广告是没有意义的。在 Composite MCP 加载到 Logistic 回归模型中，该方法所选取的变量以及系数估计结果显示，影响 *x*1 成本因素的变量有三个，把不显著的平均转化成本压缩为 0，*x*14 平均位置排名的系数为 3.0136，表明位置排名对广告转化的影响很大，越靠前越容易转化。据调查不管是广告或者搜索引擎，浏览者往往更倾向与点击排名靠前的广告或者推送。而 *x*12 平均点击价格和 *x*11 曝光数对广告转化都有积极的影响。

广告的关键词是吸引顾客的关键，研究结果表明关键词中包含商标以及价格会促进广告的转化，关键词个数、设计地域信息、特定产品和是否包含动词对广告转化是有消极的作用。关键词个数越多则包含的信息更加丰富，但是重点不突出反而会影响广告的转化。涉及地域信息会影响广告的转化，热门地区的广告显然更加容易转化。平均访问页数越多、访问次数频繁是对广告转化是有积极影响的。平均访问时长和跳出率对广告转化是有消极的影响，同时在另外两个组变量中选择了历史转化率和点击率，即表明这两个变量对广告转化有一定的影响。

6. 结语

文章研究了四种群组变量选择方法及块坐标下降算法，模拟验证了四种方法的模型选择优良性，结果表明了无论是在模型误差、正确选择 0 个数还是组数的选取上，Composite MCP 惩罚方法均比其他方法表现更好，因此不管在预测能力还是在变量选择上，Composite MCP 都是表现最优，并将几种方法运用到广告数据的实例中，在广告转化的研究中有实际的意义。

基金项目

国家自然科学基金资助项目(No.71462002)。

参考文献

[1] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*, **5**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[2] Efron, B., Haistie, T., Johnstone, I., *et al*. (2004) Least Angle Regression. *The Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[3] Zhang, C.H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[4] Zhang, C.H. (2007) Penalized Linear Unbiased Selection. Rutgers University, Department of Statistics and Biostatistics Technical Report, Newark.

[5] Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348-1360. https://doi.org/10.1198/016214501753382273

[6] Fu, W.J. (1998) Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416. https://doi.org/10.1080/10618600.1998.10474784

[7] Yuan, M. and Lin, Y. (2006) Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society*: *Series B*, **68**, 49-67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

[8] Huang, J., Breheny, P. and Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, **27**, 481-499. https://doi.org/10.1214/12-STS392

[9] Huang, J., Ma, S., Xie, H., *et al*. (2009) A Group Bridge Approach for Variable Selection. *Biometrika*, **96**, 339-355. https://doi.org/10.1093/biomet/asp020

[10] Breheny, P. and Huang, J. (2009) Penalized Methods for Bi-Level Variable Selection. *Statistics and Its Interlace*, **2**, 369-380. https://doi.org/10.4310/SII.2009.v2.n3.a10

[11] Nardi, Y. and Rinaldo, A. (2008) On the Asymptotic Properties of the Group Lasso Estimator for Linear Models. *Electronic Journal of Statistics*, **2**, 605-633. https://doi.org/10.1214/08-EJS200

[12] Wang, H. and Leng, C. (2008) A Note on Adaptive Group LASSO. *Computational Statistics & Data Analysis*, **52**, 5277-5286. https://doi.org/10.1016/j.csda.2008.05.006

[13] Zhang, C.-H. and Huang, J. (2008) The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression. *Annals of Statistics*, **36**, 1567-1594. https://doi.org/10.1214/07-AOS520

[14] Geng, Z., Wang, S. and Yu, M. (2015) Group Variable Selection via Convex Log-Exp-Sum Penalty with Application to a Breast Cancer Survivor Study. *Biometrics*, **71**, 53-62. https://doi.org/10.1111/biom.12230

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2324-7991，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：aam@hanspub.org