Advances in Applied Mathematics 应用数学进展, 2020, 9(6), 881-891
Published Online June 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.96105

Hans汉斯

Research and Application of Combined Penalty Likelihood Estimation Method Based on Two-Part Model

Xuyu Zhang, Lihua Zhao

School of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
Email: zhagxy1478@163.com, zlh5259@163.com

Received: May 28th, 2020; accepted: Jun. 12th, 2020; published: Jun. 19th, 2020

Abstract

In statistics, the potential model structure and variable selection problems of zero expansion data are often studied by means of zero expansion model. However, in most cases, the non-zero part of the response variable is quantitative data. A simple zero expansion model cannot describe the model structure of such data, and the corresponding parameter estimation method is no longer applicable. In view of this, scholars proposed a two-part model to deal with zero expansion semi-continuous data. In this paper, the combined penalty likelihood estimation method is introduced into the two-part model to study the problem of variable selection. A new penalty likelihood estimation method, NCPM (New Combined Punishment Method), is proposed to deal with high-dimensional statistical analysis problems. The method is applied to Taiyuan precipitation data and its influencing factors are analyzed. The results of simulation and case analysis show that the proposed method is effective and has higher prediction accuracy than the traditional penalty likelihood estimation method.

Keywords

Combined Punishment, Two-Part Model, LLA-CGD (Local Linear Approximation and Coordinate Gradient Descent) Algorithm, Variable Selection, Precipitation

基于两部模型的组合惩罚似然估计方法研究及其应用

张旭宇，赵丽华

太原理工大学数学学院，山西 晋中
Email: zhagxy1478@163.com, zlh5259@163.com

收稿日期：2020年5月28日；录用日期：2020年6月12日；发布日期：2020年6月19日

文章引用：张旭宇，赵丽华．基于两部模型的组合惩罚似然估计方法研究及其应用[J]．应用数学进展，2020，9(6)：881-891. DOI: 10.12677/aam.2020.96105

摘　要

在统计学中，多借助零膨胀模型研究零膨胀数据潜在的模型结构及变量选择问题。然而，在多数情况下，响应变量的非零部分为定量数据，简单的零膨胀模型无法刻画这类数据的模型结构，对应的参数估计方法也不再适用。鉴于此，学者提出处理零膨胀半连续数据的两部模型。本文将组合惩罚似然估计方法引入两部模型，研究其变量选择问题。提出一种新的处理高维统计分析问题的惩罚似然估计方法：NCPM (New Combined Punishment Method)，并将该方法应用于太原市降水量数据，分析其影响因素。模拟及实例分析结果均表明本文的方法行之有效，较传统的惩罚似然估计方法具有更高的预测精度。

关键词

组合惩罚，两部模型，LLA-CGD (Local Linear Approximation and Coordinate Gradient Descent)算法，变量选择，降水量

1. 引言

太原市由于地理环境的影响，形成了北温带大陆性气候，四季分明、冬无严寒、夏无酷暑。但其昼夜温差较大，年际气候变化较大，季风环流交替明显，气象灾害频发。尤其近几年，随着城市化进程的不断加快，公共绿地面积减少，道路等不透水面积不断增大，使得渗透力下降，排水压力加重。2005 年 8 月 16 日，太原市最大降水量达到 28.2 mm/h，导致道路积水严重，交通瘫痪数小时；2016 年 7 月连日暴雨袭击给太原市带来了巨大损失：城市道路淹没，造成交通瘫痪；街边店铺、停泊车辆等不同程度的灌入雨水，给人民带来了不可逆转的财产损失；甚至对市民生命安全产生了威胁。研究降水量的影响因素，不仅可以为农作物的播种、培育提供便利，而且可以给气象部门提前预警，以便人们在暴雨来临之前能够做出有效的预防措施，避免造成人员损伤、财务损失等。

降水量数据是典型的零膨胀半连续数据。这类数据集中“零”值占很大比例，数据中的非零部分服从某一连续分布，所以使用传统的数据模型不能很好地解释这类数据。针对这种特殊的数据类型，专家学者提出一种行之有效的方法——两部模型[1]。两部模型的第一部分用来判断响应变量是否为零，第二部分则用于描述非零响应变量的分布。该模型可以更合理、准确地研究此类特殊数据的内部规律，并在数据预测、检验等方面有着非常重要的作用。针对两部模型，学者们提出多种估计方法，包括：极大似然估计(MLE)、拟似然估计[2] (McCulloch and Searle, 2001)，惩罚似然估计[3] (Yau and Lee, 2001)和贝叶斯估计[4] (Ghosh *et al.*, 2006)等。其中，惩罚似然估计方法使用最为广泛，学者提出多种惩罚似然估计方法，如：Bridge [5] (Frank 和 Friedman, 1993)，MCP [6] (Minimax Concave Penalty)等，已广泛应用到多种领域，并取得良好效果。但当 $p \gg n$ 或者解释变量之间存在较强的相关性时，上述方法的性能会有一定的局限性。为此，Zou [7]等提出了弹性网(Elastic net, Enet)，该方法将 Lasso 和 L_2 混合形成一种新的惩罚，改善了变量间相关性对预测性能造成的影响；Wang [8]等将 SCAD (Smoothly Clipped Absolute Deviation)与 L_2 组合，构成组合惩罚(Combined Penalization, CP)。但针对后来提出的 MCP 函数的组合惩罚问题，目

前国内并没有具体阐述。本文将该方法引入两部模型对其进行具体说明，并将其应用于太原市降水量数据中，分析其影响因素。

2. 模型及研究方法

2.1. 模型建立

假设 Y 表示事件，根据两部模型的基本思想，Y 的分布可表示为如下形式：

$$f(y)=\pi*\mathbb{I}_{(y=0)}+\left[(1-\pi)*f(y)\right]\mathbb{I}_{(y>0)},\ y\geq 0,0\leq\pi\leq 1 \tag{1}$$

式中 $\mathbb{I}_{(\mathcal{A})}$ 为示性函数，π 为零事件发生的概率，$f(y)$ 为非零事件服从的分布。描述零事件发生可能性的函数包括：*logit*、*probit*、*log-log* 等[9]。本文选用 *logit* 作为模型第一部分的连接函数进行研究，记为：

$$\pi=Pr(y=0\mid X)=\frac{1}{1+\exp\left(\alpha_0+\alpha^{\mathrm{T}}X\right)} \tag{2}$$

其中，$\alpha=\left(\alpha_1,\alpha_2,\cdots,\alpha_p\right)^{\mathrm{T}}$ 表示未知的回归系数，α_0 表示截距项，$X=\left\{x_1,x_2,\cdots,x_p\right\}^{\mathrm{T}}$ 表示 p 维的解释变量。

根据参考文献[10]令 $f(y)$ 为 Gamma 分布，建立对应的概率密度函数为：

$$\begin{gathered}f(y)=\frac{\left(\frac{\mu}{\sigma^2}\right)^{\left(\frac{\mu^2}{\sigma^2}\right)}}{\Gamma\left(\frac{\mu^2}{\sigma^2}\right)}y^{\frac{\mu^2}{\sigma^2}-1}\mathrm{e}^{-\frac{\mu}{\sigma^2}y},\\ \sigma=\exp\left(\beta_0+\beta^{\mathrm{T}}X\right),y>0,\mu>0,\sigma^2>0\end{gathered} \tag{3}$$

其数学期望和方差分别为：μ，σ^2，$\beta=\left(\beta_1,\beta_2,\cdots,\beta_p\right)^{\mathrm{T}}$ 表示未知的回归系数，β_0 表示截距项。

对每个 $i,i=1,2,\cdots,n$ 有：

$$\pi_i=\frac{1}{1+\exp\left(\alpha_0+\alpha^{\mathrm{T}}x_i\right)},\sigma_i=\exp\left(\beta_0+\beta^{\mathrm{T}}x_i\right) \tag{4}$$

可得对应的对数似然函数为：

$$\begin{aligned}L(\theta)=\sum_{i=1}^{n}\Bigg\{&I_{(y_i=0)}\log(\pi_i)+I_{(y_i>0)}\left[\log(1-\pi_i)-\log\left(gamma\left(\frac{\mu^2}{\sigma_i^2}\right)\right)\right.\\&\left.+\frac{\mu^2}{\sigma_i^2}*\log\left(\frac{\mu}{\sigma_i^2}\right)+\left(\frac{\mu^2}{\sigma_i^2}-1\right)*\log(y_i)-y_i*\frac{\mu}{\sigma_i^2}\right]\Bigg\},\theta=(\alpha_0,\alpha,\beta_0,\beta)\end{aligned} \tag{5}$$

2.2. 研究方法

对 2.1 节所述模型中的未知参数采用基于组合惩罚函数的极大似然方法进行估计，即对目标函数：

$$Q(\theta)=-L(\theta)+J_{\lambda,v}(\theta) \tag{6}$$

$$J_{\lambda,v}(\theta)=p_{\lambda,\gamma}(\theta)+\frac{v}{2}\theta^2 \tag{7}$$

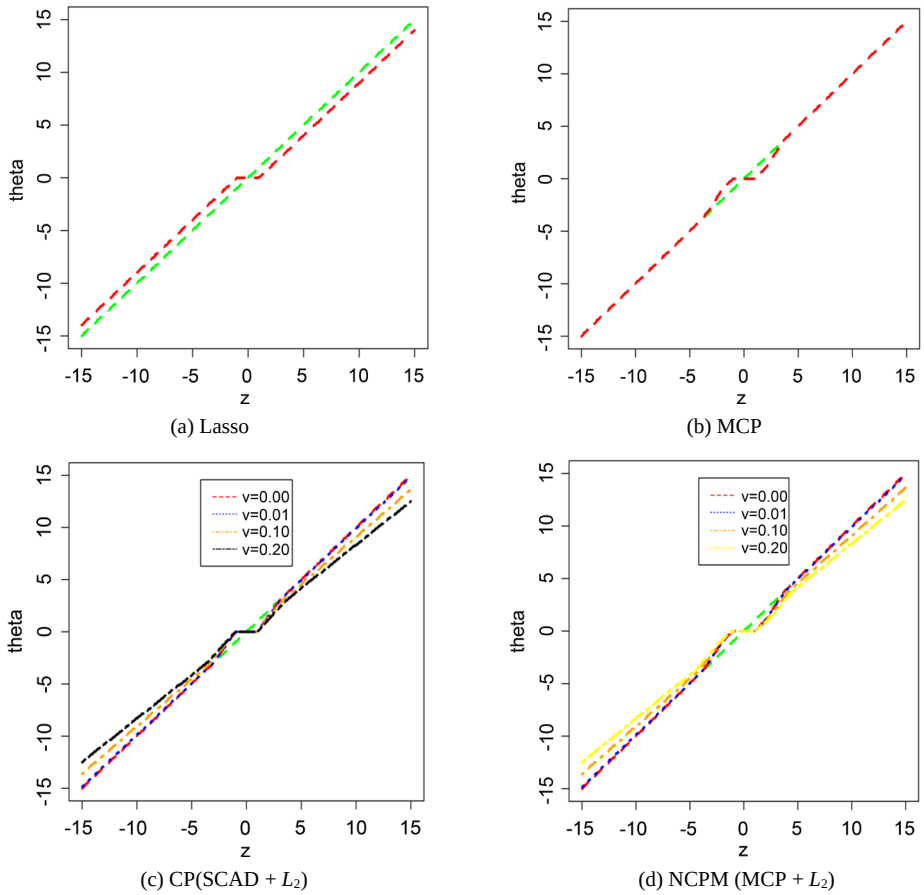
求最小值。(6，7)式中的 λ, v 称为调整参数，γ 为正则化参数。惩罚函数中只考虑系数 α 和 β。$p_{\lambda,\gamma}(\theta)$ 选取 MCP 函数，其形式如(8)。对正则化参数 γ，Breheny and Huang (2011)的模拟中建议 $\gamma=3$，所以本文取 $\gamma=3$，并且尝试令 γ 取了几个不同的值，得到的结果基本一致[11]。一般来说，调整参数的选择方法有很多，包括：AIC，BIC，GCV 和 CV 等。本文利用 10 次 5 折交叉验证法确定调整参数 λ 的值。另外，还可以选用其他惩罚函数代替 MCP。例如 Hard、Lasso (Least Absolute Shrinkage and Selection Operator)、Ridge 和 SCAD 等。

MCP 函数形式如下：

$$p_{\lambda,\gamma}(\alpha,\beta)=\sum_{j=1}^{p}\left(\rho\left(|\alpha_j|;\lambda,\gamma\right)+\rho\left(|\beta_j|;\lambda,\gamma\right)\right),\ \lambda>0,\gamma>0$$

$$\rho\left(|\theta|;\lambda,\gamma\right)=\begin{cases}\lambda|\theta|-\dfrac{\theta^2}{2\gamma}, & |\theta|\le\lambda\gamma\\ \dfrac{1}{2}\lambda^2\gamma, & |\theta|>\lambda\gamma\end{cases},\ \lambda\ge 0,\gamma>1,\theta=(\alpha,\beta) \tag{8}$$

图 1 展示了几种惩罚函数的阈值函数图。结果表明，MCP、CP 和提出的 NCPM 均具有稀疏性和连续型。组合惩罚函数对于 $\forall v>0$ 都不会产生近似无偏估计量，而当 $v\to 0$ 时会产生渐近无偏估计量。



(a) Lasso　(b) MCP

(c) CP(SCAD + L_2)　(d) NCPM (MCP + L_2)

Figure 1. Threshold Function Diagram of Lasso, MCP, CP and NCPM
图 1. Lasso、MCP、CP 和 NCPM 的阈值函数图

2.3. 求解算法

Algorithm1 非凸优化问题的 LLA-CGD 算法

输入：目标函数 $Q(\theta)$，初始值点 $\theta^{[0]}\in\mathbb{R}^n$，$k_{\max}$：最大迭代次数，正则化参数：$\lambda,v,\gamma$，步长 $\epsilon,(\epsilon^0=1)$，$\delta=1\mathrm{e}-8$，$M(\theta)=-L(\theta)+\frac{v}{2}\sum_{j=1}^{p}(\alpha_j^2+\beta_j^2)$

输出：$Q(\theta)$ 的极小值点 $\hat{\theta}$

1:for $k=0,1,\cdots$ do
2: while $k<k_{\max}$ do
3: for $j=1,2,\cdots,2p+2$ do
4: $\min\left\{Q(\theta_j)=M(\theta_j)+\nabla\rho\left(\left|\theta_j^{[k]}\right|\right)|\theta_j|\right\}$，其中：$\nabla\rho\left(\left|\theta_j^{[k]}\right|\right)$ 表示 $\rho(|\theta|;\lambda,\gamma)$ 在 $\theta=\theta_j^{[k]}$ 处的一梯度，$\theta_j^{[k]}$ 表示 $\theta_j, j=1,2,\cdots,2p+2$ 的第 k 个估计。
5: 计算 $h_j^{[k]}=\min\left(\max\left(\nabla^2M\left(\theta_j^{[k]}\right),c_{\min}\right),c_{\max}\right)$，$0<c_{\min}<c_{\max}<\infty$。取 $c_{\min}=0.5$，$c_{\max}=10^8$
6: 计算 $d_j^{[k]}=\arg\min_d\left\{\nabla M\left(\theta_j^{[k]}\right)d+\frac{1}{2}h_j^{[k]}d^2+\nabla\rho\left(\left|\theta_j^{[k]}\right|\right)\left|\theta_j^{[k]}+d\right|\right\}, d=\theta_j-\theta_j^{[k]}$
7: if $\left|\nabla M\left(\theta_j^{[k]}\right)-h_j^{[k]}\theta_j^{[k]}\right|\le\nabla\rho$
8: $d_j^{[k]}=-\theta_j^{[k]}$
9: else
10: $d_j^{[k]}=-\frac{1}{h_j^{[k]}}\left[\nabla M\left(\theta_j^{[k]}\right)+\nabla\rho\left(\left|\theta_j^{[k]}\right|\right)\mathrm{sgn}\left(-\nabla M\left(\theta_j^{[k]}\right)+h_j^{[k]}\theta_j^{[k]}\right)\right]$
11: end if
12: 利用 Armijo 准则计算 $\epsilon_j^{[k]}$，$\epsilon_j^{[k]}=\epsilon^0 0.5^{k+1}$
13: if $Q\left(\theta_j^{[k]}+\epsilon_j^{[k]}d_j^{[k]}\right)-Q\left(\theta_j^{[k]}\right)<0.1\epsilon_j^{[k]}\Delta_j^{[k]}$，其中：$\Delta_j^{[k]}=\nabla M\left(\theta_j^{[k]}\right)d_j^{[k]}+\rho\left(\left|\theta_j^{[k]}\right|\right)\left(\left|\theta_j^{[k]}+d_j^{[k]}\right|-\left|\theta_j^{[k]}\right|\right)$
14: $\theta_j^{[k+1]}=\theta_j^{[k]}+\epsilon_j^{[k]}d_j^{[k]}$
15: end if
16: end for
17: if $\left\|\theta^{[k+1]}-\theta^{[k]}\right\|_\infty<\delta$
18: break, $\hat{\theta}=\theta^{[k+1]}$
19: else
20: $k=k+1$
21: end if
22: end while
23:end for

3. 数值模拟及结果分析

3.1. 评价标准

本文选取一些常用的评价指标来衡量模型的泛化能力，具体指标如下：

1) 预测均方误差：$\mathrm{PMSE}=\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-y_i\right)^2$ (9)

2) 标准误差：$\mathrm{SE}=\sqrt{\frac{\sum_{i=1}^{m}\left(f(x_i)-y_i\right)^2}{m}}$ (10)

3) Freq：重要变量被优先选择的平均次数；

4) N_{nz}：最终模型中非零系数个数；

5) 平均误差：$\mathrm{ME}=\left\|\theta-\hat{\theta}\right\|_2^2$ (11)

6) AUC：指 ROC 曲线下的面积

7) 错误发现率：$\mathrm{FDR}=\dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}$ (12)

8) 假反例率：$\mathrm{FNR}=\dfrac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}}$ (13)

上式 m 表示样本量；y_i，$f\left(x_i\right)$ 分别表示第 i 个样本的预测值和真实值；θ 表示参数真实值，$\hat{\theta}$ 表示参数的估计值；TP 为真正例；FP 为假正例；FN 为假反例；TP 为真正例。PMSE、SE、ME 均是衡量样本真实值与预测值之间偏差的综合指标，FDR、FNR 是性能度量指标，这些指标的值越小，表明模型描述数据的精确度越高。ROC 曲线是以假正例率为横坐标，真正例率为纵坐标作图得到。若一种方法对应的 ROC 曲线被另一条曲线完全“包住”，则后者的性能优于前者[12]。

3.2. 模拟及结果

根据前述模型随机生成模拟数据集。下面给出的四个不同场景，仅通过改变解释因子的维数以及设计矩阵的相关结构来实现，每个场景重复模拟 100 ($N=100$)次。在场景 1 和 2 中，分别生成训练集和测试集，且每生成一个训练集，在相同设置下独立生成相应的测试集，用于评估所得模型的预测性能。每个训练集由 100 个独立的观察值组成，而每个测试有 500 个独立的样本；在场景 3 和 4 中，整体生成模拟数据，不区分训练集和测试集。

模拟 1 在该例中，令 $p=50$，$\alpha_j=-0.5,-1,3,1,0.2$，$j=1,11,31,41,50$。$\alpha_j=0$，$j\neq1,11,31,41,50$；$\beta_j=1.49,0.4,-0.01,0.39,-0.06$，$j=1,\cdots,5$。$\beta_j=0$，$j\neq1,\cdots,5$。另外，假设：$X=\left(X_1,X_2,\cdots,X_p\right)\sim N_p\left(0,\Sigma\right)$，$\Sigma_{i,j}=\rho^{|i-j|}$，$\rho$ 分别取 0.1、0.4、0.7。

模拟 1 旨在解释因子相关程度不同的情况下，将各惩罚方法的预测精度和变量选择性能进行比较。作为参考，还考虑了 Oracle 估计结果，即所有重要变量都事先已知的岭回归。

Table 1. Simulation results of example 1
表 1. 例 1 模拟结果

ρ	Method	PMSE(SE)	Freq.(%)	N_{nz}	FDR	FNR
$\rho=0.1$	Oracle	2.821(3.118)	100.000	10.000	0.100	0.530
	Ridge	3.235(5.184)	84.833	50.000	0.077	0.732
	NCPM	2.873(2.380)	88.667	9.731	0.080	0.490
	MCP	2.656(3.038)	87.000	9.435	0.180	0.483
	Enet	3.011(4.090)	87.000	29.890	0.190	0.610
$\rho=0.4$	Oracle	2.172(1.532)	100.000	10.000	0.040	0.500
	Ridge	2.413(1.825)	81.000	50.000	0.273	0.600
	NCPM	2.141(1.689)	86.500	9.950	0.030	0.480
	MCP	2.245(1.680)	84.320	9.950	0.020	0.520
	Enet	2.391(1.799)	83.647	34.310	0.160	0.598

Continued

	Oracle	1.587(0.571)	100.000	10.000	0.060	0.560
	Ridge	1.810(0.204)	74.167	50.000	0.172	0.632
$\rho=0.7$	NCPM	0.527(0.358)	92.483	10.180	0.030	0.390
	MCP	1.553(0.553)	89.076	8.990	0.060	0.530
	Enet	1.742(0.943)	81.833	26.740	0.140	0.590

从表 1 中报告的预测精度结果可以看出，在 ρ 的取值不同的情况下，Ridge 是四种惩罚方法中效率最低的。显然，当真实模型中有多余变量时，简约的回归模型比非简约的回归模型具有更强的预测能力。此外，Enet 的预测精度略低于 MCP 和 NCPM，这可能是 Lasso 惩罚的偏差性引起的。可以看出，这些方法中 MCP 和 NCPM 是最优的。至于变量选择的性能：岭回归使用所有变量预测，Enet 倾向于包含太多冗余变量，而 MCP 和 NCPM 通过对未知参数的值进行压缩，实现变量选择。因此，MCP 和 NCPM 是这四种方法中最好的变量选择方法。

模拟 2 在该例中，令 $p=50$，$\alpha_j=-0.5,-1,3,1,0.2,-0.14$，$j=1,\cdots,6$。$\alpha_j=0$，$j\neq 1,\cdots,6$；$\beta_j=1.49,0.5,0.4,-0.01,0.39,-0.06$，$j=1,\cdots,6$。$\beta_j=0$，$j\neq 1,\cdots,6$。令 $X_j, j=1,\cdots,50$ 为一组$i.i.d$的随机变量，且 $X_j\sim N(0,1)$。本例设置两种情况：

1) 当 $j=7,\cdots,12$ 时，令 $X_j=X_{j-6}+\eta_j$，其中，$\eta_j(i.i.d)$ 且 $\eta_j\sim N(0,0.01)$；

2) 当 $j=7,\cdots,28$ 时，令 $X_j=X_{j+22}+\eta_j$，其中，$\eta_j(i.i.d)$ 且 $\eta_j\sim N(0,0.01)$；

(1)与(2)的区别在于：情况(1)中，重要变量间存在相关关系，情况(2)中，非重要变量间具有相关关系。

Table 2. Simulation results of example 2
表 2. 例 2 模拟结果

场景	方法	PMSE	(SE)	N_{nz}	Freq.(%)
	Oracle	2.095	1.554	12.000	100.000
	Ridge	2.927	1.635	50.000	78.583
1	NCPM	2.155	1.581	11.864	87.083
	MCP	2.154	1.579	11.927	88.500
	Enet	2.351	1.653	18.635	81.677
	Oracle	0.724	0.976	6.000	100.000
	Ridge	1.237	1.371	50.000	56.833
2	NCPM	0.841	0.513	6.039	91.459
	MCP	1.652	1.008	44.830	0.000
	Enet	1.037	1.023	25.379	89.176

模拟 2 旨在比较当解释变量间存在强相关性时，各惩罚方法性能的优越性。表 2 展示了实例 2 的模拟结果。MCP 在场景 1 的变量选择中是最好的，但是在场景 2 中，比其他变量选择要差得多，得到的结果完全是误导性的。在场景 2 中，使用组合惩罚(Enet 和 NCPM 两种方法)的性能比使用单一惩罚要好得多；NCPM 对这两种情况都是最好的。从这个例子中，我们可以看出，在 MCP 中增加 L_2 惩罚可以显著降低解释因子之间的高共线性带来的风险。

接下来的两个例子旨在比较两种组合惩罚方法(CP 和 NCPM)在 ρ 较大情况下的效果。

模拟 3 在该例中，令 $p=200$，剩余条件与模拟 1 相同。

模拟 4 在该例中，令 $X_j(i.i.d)$，$j=1,\cdots,p$ 且 $X_j \sim N(0,1)$。分别，令 $p=100,300,500$。$\alpha_j=-0.5,-1,3,1,0.2$，$j=1,11,31,41,50$。$\alpha_j=0$，$j\neq 1,11,31,41,50$；$\beta_j=1.49,0.4,-0.01,0.39,-0.06$，$j=1,\cdots,5$。$\beta_j=0$，$j\neq 1,\cdots,5$

表 3 总结了组合惩罚方法的模拟结果。在表格中，我们分别列出了系数 α 和 β 在不同衡量指标下的具体情况。模拟结果表明：两种方法的变量选择均有很好的效果；增加解释变量间的相关性对两种方法的变量选择性能影响不大，这表明它们均有很好的处理解释变量间共线性的能力；NCPM 比 CP 给出了更精确的变量选择结果。例如：例 3 $\rho=0.4$ 的情境下(表 3 中第 3、4 行)，系数 α 在使用 NCPM 方法时，得到的 FDR、FNR、ACU 的值分别为：0.03、0.39、0.93，使用 CP 方法得到的值分别为 0.05、0.43、0.92。并且发现，NCPM 方法在 ME、PMSE 上均有所改善。另外，对系数 β 也有类似的发现。

Table 3. Simulation 3 and Simulation 4 Results
表 3. 模拟 3、模拟 4 结果

例子	场景	估计方法	α					β				
			FDR	FNR	AUC	ME	PMSE	FDR	FNR	AUC	ME	PMSE
例 3 (p)	0.1	NCPM	0.090	0.439	0.890	0.182	0.385	0.300	0.171	0.900	0.016	0.321
		CP	0.092	0.524	0.700	0.285	1.285	0.326	0.282	0.870	0.169	1.746
	0.4	NCPM	0.030	0.390	0.930	0.141	0.457	0.280	0.142	0.920	0.016	0.208
		CP	0.050	0.430	0.920	0.165	1.741	0.289	0.236	0.900	0.125	1.711
	0.7	NCPM	0.029	0.536	0.918	0.132	0.872	0.230	0.274	0.960	0.029	0.151
		CP	0.011	0.374	0.890	0.152	2.239	0.247	0.196	0.890	0.120	1.501
例 4 (p)	100	NCPM	0.050	0.020	0.960	0.016	0.341	0.330	0.250	0.930	0.052	0.088
		CP	0.060	0.433	0.910	0.022	1.046	0.300	0.324	0.870	0.089	0.114
	300	NCPM	0.020	0.060	0.980	0.017	0.351	0.260	0.020	0.970	0.031	0.062
		CP	0.090	0.371	0.950	0.059	0.832	0.285	0.143	1.000	0.054	0.105
	500	NCPM	0.018	0.134	0.990	0.003	0.556	0.180	0.000	0.990	0.024	0.041
		CP	0.023	0.520	0.970	0.041	0.610	0.244	0.025	1.000	0.049	0.067

4. 实例分析

4.1. 降水量定义

降水量是指从天空落到地面的液态或固态(经融化后)水，未经流失，在水平面上积聚的深度(https://baike.baidu.com/item)。1 mm 的降水量是指：在 666.7 m^2 上，降水总深度达到 1 mm。

4.2. 数据来源及预处理

降水量相关数据源于气象数据库 https://www.aqistudy.cn/historydata/index.php，https://www.wunderground.com/history。本文主要收集了山西省太原市 2017 年 1 月~2018 年 12 月期间风速、能见度、气压、日照时长、相对湿度、AQI、PM2.5 以及 PM10 的日平均值。具体说明见表 4。

Table 4. Description of relevant variables
表 4. 相关变量说明

变量	符号	单位	说明	类型	均值	Min	Max
Y	降水量	mm	连续型变量	数值型	0.821	0.00	27.94
x_1	T_{max}	℃	日最高温度	数值型	18.231	−7.22	37.78
x_2	T_{min}	℃	日最低温度	数值型	5.239	−18.33	23.44
x_3	T_{Dr}	℃	日期温差	数值型	−4.786	−16.11	7.22
x_4	T_{mean}	℃	日平均气温	数值型	11.747	−11.11	30.00
x_5	T_{Dp}	℃	露点温度	数值型	0.747	−30.00	22.22
x_6	WS	km/h	风速	数值型	21.009	6.44	49.89
x_7	VS	km	能见度	数值型	24.679	3.1	30.96
x_8	P	hPa	气压	数值型	1021.987	1002.03	1049.78
x_9	SD	h	日照时长	数值型	12.186	9.53	14.77
x_{10}	RH	%	相对湿度	数值型	52.038	14	93
x_{11}	AQI	无	空气质量指数	数值型	109.426	0	468
x_{12}	PM2.5	μg/m^3	细颗粒物	数值型	57.370	0	377
x_{13}	PM10	μg/m^3	可吸入颗粒物	数值型	121.869	0	473

其中，2018 年中有 65 天降水量大于 0；2017 年中有 58 天降水量大于 0。

对数据收集整理后，共得到 730 组观测值。其中 607 (约占 83.2%)组内降水量的取值为 0。另外，为了消除量纲的影响，需要对变量进行标准化处理。本文使用 z-score 标准化法对原始数据预处理，且经过 z-score 标准化处理的数据服从 $N(0,1)$。

z-score 标准化公式为：

$$\tilde{x}_i = \frac{x_i - \overline{x}}{\sigma}, \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{14}$$

4.3. 模型建立

根据参考文献[13]得到：Gamma 分布对降水量非零部分的数据拟合效果最佳。因此，本节使用前述的 Logit-Gamma 模型进行研究。

4.4. 结果分析及模型比较

运用 NCPM 方法对实例数据估计，并与 Enet 方法对比，结果见表 5。

Table 5. Example analysis results (Estimated value (Standard error))
表 5. 实例分析结果(估计值(标准误差))

解释变量		NCPM		Enet	
		Logistic (α)	Gamma (β)	Logistic (α)	Gamma (β)
T_{max}	x_1				
T_{min}	x_2			0.274	1.816

Continued

T_{Dr}	x_3			0.017	0.739
T_{mean}	x_4		3.522	0.019	0.001
T_{Dp}	x_5	4.533		0.548	3.144
WS	x_6	1.732	0.422	0.113	1.774
VS	x_7			−0.033	0.159
P	x_8			−4.667	−16.322
SD	x_9	−8.493	−1.663	−1.120	−8.449
RH	x_{10}	12.958	3.606	1.326	11.225
AQI	x_{11}	−1.266	−0.505	0.006	1.236
PM2.5	x_{12}	−2.848	−0.461	−0.144	−2.695
PM10	x_{13}	−2.089	−0.539	0.077	1.943

由表 5 新提出方法得到的结果可知:气温与降水量成正相关，这一现象主要受全球变暖的影响。气候变暖，气温升高，水循环加快，大气中的水蒸气增多，降水量也随之增大。风速与降水量成正比。这是因为风速越大，单位时间内进入空气中的水分子越多，蒸发量就越大，导致降水量越大。日照时长与降水量成负相关；空气相对湿度与降水量成正相关，这些结论与实际相符。PM2.5、PM10 的浓度以及 AQI 均与降水量呈负相关。这表明：降水能有效去除大气中的颗粒物，降低空气中 PM2.5、PM10 的浓度，从而起到净化空气的作用，导致 AQI 的值降低。

另外发现，新提出的方法得到的模型更简约，模型可解释性更高。如：表 5 中 PM10 的 Logit 部分，用两种方法获得的估计值分别：−2.089、0.077；PM10 的 Gamma 部分，用两种方法得到的估计值分别为：−0.539、1.943，显然利用 Enet 方法得到的估计与实际意义不符。

5. 结束语

本文以变量选择切入，从理论和数值模拟两方面系统地研究组合惩罚函数的极大似然估计方法在 Logit-Gamma 两部模型中的表现。具体总结如下：

利用 L_2 在高度相关解释变量间的良好表现，将基于 SCAD + L_2 惩罚函数的极大似然估计方法扩展到 MCP 函数，提出一种新的处理高维统计分析问题的惩罚似然估计方法，NCPM 极大似然估计方法。该方法改善了变量间相关性对模型稳定性、精确度的影响。模拟研究表明，当 $p \gg n$ 或解释变量间的相关性较强时，该方法高效便捷且易于实现。

关于组合惩罚似然估计两部模型的变量选择，本文采用 LLA-CGD 算法。该算法解决了目标函数非线性问题，同时实现计算可行性。数值模拟结果显示，该算法选择效果良好，为两部模型的变量选择提供了新思路。

将提出的 NCPM 方法应用于 Logit-Gamma 两部模型，分析太原市降水量的影响因素。结果显示，是否降水主要受露点温度、风速、日照时长、空气相对湿度、PM2.5 及 PM10 浓度等的影响；当降水产生时，降水量多少更易受日平均气温、风速、日照时长、空气相对湿度、PM2.5 及 PM10 浓度、AQI 等的影响。最后与 Enet 方法对比，进一步证实了提出方法具有估计的稳定性、模型可解释性等优势。

参考文献

[1] Manning, W.G., Duan, N. and Rogers, W.H. (1987) Monte-Carlo Evidence on the Choice between Sample Selection and 2-Part Models. *Journal of Econometrics*, **35**, 59-82. https://doi.org/10.1016/0304-4076(87)90081-9

[2] McCulloch, C.E. and Searle, S.R. (2001) Generalized, Linear, and Mixed Models. A Wiley-Interscience Publication John Wiley & Sons INC, New York, 23-24. https://doi.org/10.1002/0471722073

[3] Yan, K.K.W. and Lee, A.H. (2001) Zero-Inflated Poisson Regression with Random Effects to Evaluate an Occupational Injury Prevention Programme. *Statistics in Medicine*, **20**, 2907-2920. https://doi.org/10.1002/sim.860

[4] Xu, X. and Ghosh, M. (2015) Bayesian Variable Selection and Estimation for Group Lasso. *Bayesian Analysis*, **10**, 1727-1734. https://doi.org/10.1214/14-BA929

[5] Frank, I. and Friedman, I. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-148. https://doi.org/10.1080/00401706.1993.10485033

[6] 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

[7] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] Wang, X.M., Park, T. and Carriere, K.C. (2010) Variable Selection via Combined Penalization for High-Dimensional Data Analysis. *Computational Statistics and Data Analysis*, **54**, 2230-2243. https://doi.org/10.1016/j.csda.2010.03.026

[9] Duan, N. and Morris, C.N. (1983) A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business& Economic Statistics*, **1**, 115-126. https://doi.org/10.2307/1391852

[10] Wang, X.M., Park, T. and Carriere, K.C. (2010) Variable Selection via Combined Penalization for High-Dimensional Data Analysis. *Computational Statistics and Data Analysis*, **54**, 2230-2243. https://doi.org/10.1016/j.csda.2010.03.026

[11] Breheny, P. and Huang, J. (2010) Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *Annals of Applied Statistics*, **5**, 232-253. https://doi.org/10.1214/10-AOAS388

[12] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 33-35.

[13] 丁裕国. 降水量r分布模式的普适性研究[J]. 1994, 18(5): 552-560.