Advances in Applied Mathematics 应用数学进展, 2021, 10(4), 1243-1254
Published Online April 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.104135

Hans汉斯

零膨胀几何分布的变量选择

文静蕊，赵丽华

太原理工大学数学学院，山西 晋中
Email: 1820804357@qq.com, zlh5259@163.com

收稿日期：2021年3月27日；录用日期：2021年4月15日；发布日期：2021年4月29日

摘　要

在卫生服务和结果研究中，经常遇到计数结果，并且通常零占有有很大比例。零膨胀几何回归模型是分析几何部分过多零的有力工具。在实际建模中，搜集到的变量中可能存在与目标完全无关的变量(冗余变量)或者有些变量已知和目标相关，但实际的影响微乎其微。针对协变量多且具有相关性的问题，本文在似然函数的基础上添加SCAD、MCP和LASSO惩罚，得到基于零膨胀几何回归的惩罚目标函数，然后利用EM算法研究模型的参数估计和变量选择。仿真研究表明：该模型不仅具有准确的参数估计，而且比传统的逐步选择方法更优越。

关键词

零膨胀几何回归，变量选择，LASSO，SCAD，MCP

Variable Selection of Zero-Inflated Geometric Distribution

Jingrui Wen, Lihua Zhao

College of Mathematics, Taiyuan University of Technology, Jinzhong Shanxi
Email: 1820804357@qq.com, zlh5259@163.com

Received: Mar. 27th, 2021; accepted: Apr. 15th, 2021; published: Apr. 29th, 2021

Abstract

In health services and outcome research, count results are often encountered, and there is usually a large proportion of zeros. The zero-inflated geometric regression model is a powerful tool for analyzing excessive zeros in geometrical parts. In actual modeling, there may be variables that are completely unrelated to the target (redundant variables) among the collected variables, or some

文章引用: 文静蕊, 赵丽华. 零膨胀几何分布的变量选择[J]. 应用数学进展, 2021, 10(4): 1243-1254.
DOI: 10.12677/aam.2021.104135

variables are known to be related to the target, but the actual impact is minimal. Aiming at the problem of many covariates and correlations, this paper adds SCAD, MCP and LASSO penalties to the likelihood function to obtain a penalty objective function based on zero-inflated geometric regression, and then uses the EM algorithm to study the parameter estimation and variable selection of the model problem. Simulation research shows that the model not only has accurate parameter estimation, but also is superior to the traditional stepwise selection method.

Keywords

Zero-Inflated Geometric Regression, Variable Selection, LASSO, SCAD, MCP

1. 引言

零膨胀模型广泛存在于各个领域：如社会科学、医学、工业、农业和生态学等。该类数据的特点是观测值中含有过多的零，若用传统的泊松(负二项或二项)回归模型对离散数据建模，拟合的结果就会出现较大偏差。因此，有很多学者开始了零膨胀模型的研究：早在 1960 年，Cohen Jr AC [1]已经注意到零膨胀现象。直到 1986 年，Mullahy [2]对常见的计数数据模型进行替代和测试，提出零膨胀模型。随着 Lambert [3] (1992)的深入研究而广受欢迎。为了研究协变量的影响，他对零部分和非零部分分别建立 logit 连接模型和对数线性模型，提出了具有协变量因子的参数零膨胀泊松回归模型，并将其应用于印制电路板焊接缺陷的实例。随后，一个零膨胀(ZI)回归模型框架被提出并广泛应用于零过多问题。例如：Lee [4] [5] (2012a, b)在流行病学中应用 ZI 模型来研究男性乳头瘤病毒感染；Huang [6]等(2017)提出了一种用于药物安全信号检测的 ZI 模型似然比检验；Liu 和 Powers [7] (2007)提出了一个应用于吸烟行为的 ZI 计数数据的生长曲线模型。而关于零膨胀几何分布的研究寥寥无几；肖翔[8] (2018)开始零膨胀几何分布模型进行参数估计，2019 年，提出零一膨胀几何回归模型[9]，用来拟合数据中过多的零值和一值。因此，本文对零膨胀几何分布的研究很有意义。

变量选择对统计数据分析至关重要，识别重要的预测因子将提高拟合模型的预测性能。为了提高预测精度，选择有意义的变量，统计学家最初提出最优子集法来改进最小二乘估计，当设计矩阵正交时，最佳子集选择在传统的线性模型中的表现与最小二乘法一致。但是这个程序有两个基本限制：一是当预测数很多时，进行子集计算是不可行的；二是子集选择因其固有的离散性而变化极大(Breiman, 1996) [10]。逐步选择经常被用作子集选择的计算替代，然而它也遭受高可变性和经常陷入局部最优解而不是全局最优解的问题。此外，这些选择过程忽略了变量选择阶段的随机误差或不确定性。微小的数据变化就会导致不同的模型(Shen & Ye, 2002) [11]。为了能够自动简捷的选择变量，惩罚函数的思想开始流行，Hoerl 和 Kennard [12] (1970)提出岭回归，Frank 和 Friedman [13] (1993)将最小二乘估计、最优子集和岭回归整合到一个共同的框架中，分析其各自的适用性。Tibshirani [14] (1996)提出最小绝对收缩和选择算子(Lasso)，它是利用 l_1 惩罚通过连续收缩自动选择重要变量，从而保留了最佳子集选择和岭回归的良好特征。在惩罚回归算法中，低于某个阈值的系数被设置为零，否则将向零收缩。Nicolai Meinshausen [15] (2007)表明如果潜在的模型满足某些条件，则用 Lasso 选择变量是一致的。随着对惩罚函数的进一步探索，Fan 和 Li [16] (2001)研究了包括 Lasso 在内的一类惩罚函数，指出 Lasso 可以自动进行变量选择因为 l_1 惩罚在原

点是奇异的，另一方面 Lasso 收缩程序产生了对大系数的有偏估计，因此就估计风险而言，它可能是次优的。Fan 和 Li [17] (2006)也指出一个好的惩罚函数应该具有 oracle 性(稀疏性、连续性和无偏性)，提出了平滑限幅绝对偏差(SCAD)惩罚。在 2006 年，他们对特征选择进行了全面的综述，并提出了一个统一的惩罚似然框架来处理变量选择问题。Zou [18] (2005)提出了弹性网惩罚，模拟表明当预测数远大于观测数时，该方法比 Lasso 有效。2006 年，提出了用于同时估计和变量选择的自适应 Lasso 惩罚，添加自适应加权 l_1 惩罚满足惩罚函数的神谕性，很好地解决了多重共线性的问题。Zhang [19] (2010)提出的 MCP 惩罚。Buu [20]等(2011)提出一步 SCAD 方法对 ZI 模型进行变量选择；Wang [21]等(2014)对比了 LASSO、SCAD 和 MCP 三种惩罚函数。Chen [22]等(2016)把 SCAD 惩罚应用到重复测量的零膨胀模型中。

2. 惩罚 ZIGe 回归模型

2.1. 惩罚 ZIGe 回归模型的变量选择

对每一个个体 $i=1,2,\cdots,n$，定义其响应变量 Y_i 之间是相互独立的。零膨胀几何分布是退化零分布和几何分布构成的混合分布，那么 Y_i 的概率密度函数为：

$$P\left(Y_i=y_i\right)=\begin{cases}p_i+\left(1-p_i\right)\left(1-\theta_i\right), & y_i=0\\ \left(1-p_i\right)\left(1-\theta_i\right)\theta_i^{y_i}, & y_i=1,2,\cdots\end{cases} \tag{1}$$

其中参数 $p_i\left(0\le p\le 1\right)$ 为容纳过多零的混合概率参数，参数 θ_i 表示几何分布中事件发生的概率。当 $p=0$ 时，ZIGe 分布退化为标准的几何分布；当 $p>0$ 时，表明存在零膨胀。p 越大，说明数据中零的比例越大，非几何分布的数据所占的比例也就越大。在 ZIGe 回归模型中，混合概率 p_i 和事件发生概率 θ_i 都是通过 logistic 回归模型建立连接关系：

$$\begin{cases}logit\left(p_i\right)=W_i^{\mathrm{T}}\gamma\\ logit\left(\theta_i\right)=X_i^{\mathrm{T}}\beta\end{cases} \tag{2}$$

其中 $W_i^{\mathrm{T}}=\left(1,W_{i1},W_{i2},\cdots,W_{iq}\right)$ 和 $X_i^{\mathrm{T}}=\left(1,X_{i1},X_{i2},\cdots,X_{ip}\right)$ 均是协变量随机变量；$\gamma=\left(\gamma_0,\gamma_1,\gamma_2,\cdots,\gamma_q\right)^{\mathrm{T}}$ 和 $\beta=\left(\beta_0,\beta_1,\beta_2,\cdots,\beta_p\right)^{\mathrm{T}}$ 均是可估的未知回归系数，β_0 和 γ_0 是截距项。假设 $\left\{\left(Y_i,X_i,W_i\right),i=1,2,\cdots,n\right\}$ 独立观测的样本量，$\varphi=\left(\beta,\gamma\right)$ 是未知的 $p+q$ 维系数。观测样本的对数似然函数为：

$$\begin{aligned}\ell\left(\varphi;y_i,x_i,w_i\right)&=\sum_{i=1}^{n}\log\left\{I_{(y_i=0)}\left[\frac{p_i}{\left(1-p_i\right)}+\left(1-\theta_i\right)\right]\left(1-p_i\right)+I_{(y_i>0)}\left[\theta_i^{y_i}\left(1-\theta_i\right)\right]\left(1-p_i\right)\right\}\\ &=\sum_{i=1}^{n}\log\left(\exp\left(W_i^{\mathrm{T}}\gamma\right)+\left(1+\exp\left(X_i^{\mathrm{T}}\beta\right)\right)^{-1}\right)-\log\left(1+\exp\left(W_i^{\mathrm{T}}\gamma\right)\right)\\ &+\sum_{i=1}^{n}I_{(y_i>0)}\left[y_iX_i^{\mathrm{T}}\beta-\left(y_i+1\right)\left(1+\exp\left(X_i^{\mathrm{T}}\beta\right)\right)^{-1}\right]\end{aligned} \tag{3}$$

其中 θ 可直接最大化 $\ell\left(\theta\right)$ 或者 EM 算法进行估计。值得注意的是，文章只考虑了 logistic 回归部分和几何分布回归部分具有相同协变量的情况，但是所提出的方法可以很容易地扩展到对于两个回归部分具有不同协变量的情况。

关于变量选择，考虑如下的惩罚零膨胀几何模型：

$$p\ell\left(\varphi;y_i,x_i,w_i\right)=\ell\left(\varphi;y_i,x_i,w_i\right)-p\left(\beta,\gamma\right) \tag{4}$$

其中非负罚函数由下式给出：

$$p(\beta,\gamma)=n\sum_{j=1}^{p_1}p_{a_j}\left(\left|\beta_j\right|\right)-n\sum_{k=1}^{p_2}p_{b_k}\left(\left|\gamma_k\right|\right) \tag{5}$$

其中 $p_{a_j}(\cdot)$ 和 $p_{b_k}(\cdot)$ 是惩罚函数，a_j,b_k 是调节参数。回归系数 β_j,γ_k 可以有不同的惩罚项。Fan [17]所提出的 SCAD 惩罚定义如下：

$$p_\lambda(\xi)=\begin{cases}\lambda|\xi|, & |\xi|\le\lambda\\ \dfrac{\left(a^2-1\right)\lambda^2-\left(|\xi|-a\lambda\right)^2}{2(a-1)}, & \lambda<|\xi|\le a\lambda\\ \dfrac{(a+1)\lambda^2}{2}, & |\xi|>a\lambda\end{cases} \tag{6}$$

即当 $|\theta|$ 较小时，惩罚函数为线性函数；当 $|\theta|$ 较大时，为二次惩罚；而当 $|\theta|$ 很大时，惩罚项为常数。a,λ 均为调节参数，λ 控制所选模型的复杂度，当 $n\to\infty$，λ 趋于 0。通常情况下，设置 $a=3.7$。SCAD 惩罚函数关于 ϕ 的一阶导函数为：

$$p'_\lambda(\xi)=\begin{cases}\lambda, & |\xi|\le\lambda\\ \dfrac{a\lambda-\xi}{a-1}, & \lambda<|\xi|\le a\lambda=\lambda\left\{I(\xi\le\lambda)+\dfrac{(a\lambda-\xi)_+}{(a-1)\lambda}I(\xi>\lambda)\right\}\\ 0, & |\xi|>a\lambda\end{cases} \tag{7}$$

Zhang [19]提出的 MCP 惩罚函数定义如下：

$$p_{\lambda,\gamma}(a,c)=\sum_{i=1}^{n}\left(\rho\left(|a_i|;\lambda,\gamma\right)+\rho\left(|c_i|;\lambda,\gamma\right)\right),\ \lambda>0,\gamma>0 \tag{8}$$

$$\rho\left(|\xi|;\lambda,\gamma\right)=\begin{cases}\lambda|\xi|-\dfrac{\xi^2}{2\gamma}, & |\xi|\le\lambda\gamma\\ \dfrac{1}{2}\lambda^2\gamma, & |\xi|>\lambda\gamma\end{cases},\lambda\ge0,\gamma>1,\xi=(a,c) \tag{9}$$

其一阶导数为：

$$\rho\left(|\xi|;\lambda,\gamma\right)=\lambda\left(1-\frac{|\xi|}{\gamma^\lambda}\right)_+\operatorname{sgn}\left(|\xi|\right) \tag{10}$$

Lasso 惩罚定义如下：

$$p\left(\lambda;|\xi|\right)=\lambda|\xi|,\ \lambda\ge0 \tag{11}$$

图 1 展示了 $a=3.7$, $\lambda=1$时的 LASSO 和 SCAD 函数及其导数图。左边是函数 $p_\lambda|\beta|$，右边是函数的一阶导数。非凸 SCAD 惩罚的吸引力在于：当 $|\beta|$ 增加时，它们不会过度惩罚导数为零的较大系数。

Fan 和 Li [16] (2001)中通过相应定理的直接应用，可以推出惩罚零膨胀几何回归模型的 ORACLE 性。因此，如果 $\lambda_{Ge,n}=0$，$\lambda_{BI,n}=0,\sqrt{n}\lambda_{Ge,n}\to\infty,\sqrt{n}\lambda_{BI,n}\to\infty$，则 ZIGe-MCP，ZIGe-SCAD 具有神谕性：在概率趋于 1 时，无效应系数的估计量为 0，有效应系数的估计量具有渐近正态分布，均值为非零系数的真值，方差为非零系数对应的费希尔信息矩阵的子矩阵。

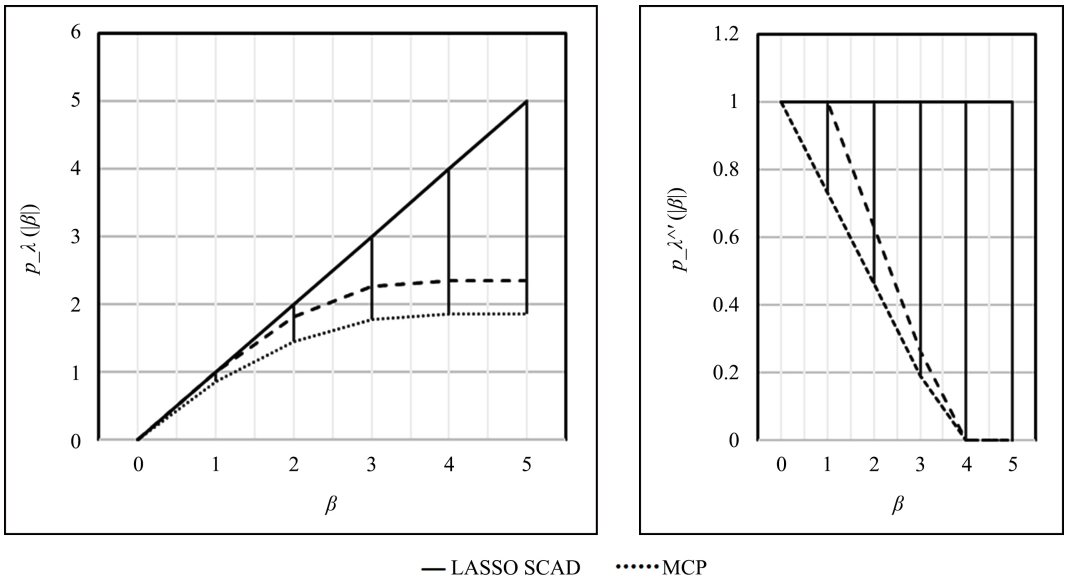

Figure 1. LASSO, MCP and SCAD penalties when $\lambda = 1$ and $a = 3.7$
图 1. $\lambda = 1$, $a = 3.7$ 时的 LASSO、MCP 和 SCAD 惩罚

2.2. EM 算法

根据 Lambert (1992)的文章，零膨胀几何模型可以通过 EM 算法进行最大似然估计拟合。当 Y_i 来自于零部分时，定义未观测到的随机变量 $B_i = 1$。当 Y_i 来自于二项部分时，$B_i = 0$。因为 $B = (B_1, B_2, \cdots, B_n)^{\mathrm{T}}$ 是不可观测的，它通常被视为缺失数据。EM 算法就可以极好的处理缺失数据问题。如果完整数据 (Y_i, B_i) 是可得到的，那么完整函数的对数似然函数为：

$$\ell^c(\varphi; y_i, z_i) = \sum_{i=1}^{n}\left\{B_i W_i^{\mathrm{T}}\gamma - \log\left(1+\mathrm{e}^{W_i^{\mathrm{T}}\gamma}\right) + (1-B_i)\left[y_i X_i^{\mathrm{T}}\beta - (y_i+1)\log\left(1+\mathrm{e}^{X_i^{\mathrm{T}}\beta}\right)\right]\right\} \tag{12}$$

则惩罚完全数据的对数似然函数为：

$$\begin{aligned} p\ell^c(\varphi; y_i, z_i) = &\sum_{i=1}^{n}\left\{B_i W_i^{\mathrm{T}}\gamma - \log\left(1+\exp\left(W_i^{\mathrm{T}}\gamma\right)\right)\right\} - n\sum_{k=1}^{p_1} p_{a_j}\left|\beta_j\right| \\ &+ \sum_{i=1}^{n}(1-B_i)\left[y_i X_i^{\mathrm{T}}\beta - (y_i+1)\log\left(1+\mathrm{e}^{X_i^{\mathrm{T}}\beta}\right)\right] - n\sum_{k=1}^{p_2} p_{b_k}\left|\gamma_k\right| \end{aligned} \tag{13}$$

因此，惩罚对数似然函数可看作是关于 B_i 的线性函数，$p\ell^c(\gamma; y_i, z_i)$ 和 $p\ell^c(\beta; y_i, z_i)$ 分别用 logistic 回归和加权几何回归求得最大值。利用 EM 算法：E 步通过给定观测数据和先前参数估计 $\varphi^{(m)}$ 来计算条件期望 δ 的条件期望；M 步求使 $\ell^c(\varphi; y_i, z_i)$ 极大化的 φ，确定第 $i+1$ 次迭代的参数估计值 $\varphi^{(i+1)} = (\beta^{i+1}, \gamma^{i+1})$。当 φ 收敛时，迭代停止。从初始值 $\varphi^{(0)} = (\beta^{(0)}, \gamma^{(0)})$ 开始，EM 算法在以下三个步骤中迭代进行，直到收敛：

1) E 步：给定观测数据并假设当前的估计 $\varphi^{(m)}$ 提供模型的真实参数，迭代 m 次时 δ 的条件期望由下式给出：

$$\hat{B}_i^{(m)} = \begin{cases} \left(\left(1+\exp\left(W_i^{\mathrm{T}}\hat{\gamma}^{(m)}\right)^{-1}\right)\left(1+\exp\left(X_i^{\mathrm{T}}\hat{\beta}^{(m)}\right)\right)^{-1}\right)^{-1}, & y_i = 0 \\ 0, & y_i > 0 \end{cases} \tag{14}$$

因此，可以计算完整数据对数似然的期望值：

$$\mathcal{Q}\left(\varphi|\hat{\varphi}^{(m)}\right)=E\left(p\ell^{c}\left(\varphi;y_i,z_i\right)\middle|y_i,\hat{\varphi}^{(m)}\right)=\mathcal{Q}_1\left(\gamma|\hat{\varphi}^{(m)}\right)+\mathcal{Q}_2\left(\beta|\hat{\varphi}^{(m)}\right) \tag{15}$$

其中：

$$\mathcal{Q}_1\left(\gamma|\hat{\varphi}^{(m)}\right)=\sum_{i=1}^{n}\hat{B}_i^{(m)}W_i^{\mathrm{T}}\gamma-\log\left(1+\exp\left(W_i^{\mathrm{T}}\gamma\right)\right)-n\sum_{k=1}^{p_1}p_{a_j}\left|\beta_j\right| \tag{16}$$

$$\mathcal{Q}_2\left(\beta|\hat{\varphi}^{(m)}\right)=\sum_{i=1}^{n}\left(1-\hat{B}_i^{(m)}\right)\left[y_iX_i^{\mathrm{T}}\beta-\left(y_i+1\right)\log\left(1+e^{X_i^{\mathrm{T}}\beta}\right)\right]-n\sum_{k=1}^{p_2}p_{b_k}\left|\gamma_k\right| \tag{17}$$

2) M 步(关于 γ)：

$$\hat{\gamma}^{(m+1)}=\arg\underbrace{\max}_{\gamma}\mathcal{Q}\left(\gamma|\hat{\gamma}^{(m)}\right) \tag{18}$$

3) M 步(关于 β)

$$\hat{\beta}^{(m+1)}=\arg\underbrace{\max}_{\beta}\mathcal{Q}\left(\beta|\hat{\beta}^{(m)}\right) \tag{19}$$

$\mathcal{Q}_1\left(\gamma|\theta^{(k)}\right)$是一个加权惩罚逻辑对数似然函数，$\mathcal{Q}_2\left(\beta|\theta^{(k)}\right)$是一个惩罚逻辑回归对数似然函数，重复 E 步和 M 步直到收敛。

2.3. 调整参数的选择

惩罚参数$\left(a_j,b_k\right)$由 BIC 准则选取：

$$BIC=-2\ell\left(\hat{\varphi};a_j,b_k\right)+\log\left(n\right)df \tag{20}$$

其中$\hat{\varphi}$是估计参数；$\left(a_j,b_k\right)$是调整参数；$\ell\left(\cdot\right)$是对数似然函数；$df=\sum_{j=0}^{q1}I\left(\beta_j\neq0\right)+\sum_{k=0}^{q2}I\left(\gamma_k\neq0\right)$是自由度。首先基于成对的收缩参数构造一个解路径，算法生成两个递减序列：$a_j^{(1)}>a_j^{(2)}>\cdots>a_j^{(M)}$和$b_k^{(1)}>b_k^{(2)}>\cdots>b_k^{(M)}$。然后对序列进行配对：$\left(a_j^{(1)},b_k^{(1)}\right),\cdots,\left(a_j^{(M)},b_k^{(M)}\right)$。原则上选择大的值$\left(a_j^{(1)},b_k^{(1)}\right)$使得除了截距项外的所有系数都趋于零；而一种改进的方法表明最小值$\left(a_j^{(1)},b_k^{(1)}\right)$可从 LASSO 惩罚的数据中计算得到。关于 SCAD 惩罚和 MCP 惩罚，本文也遵循这种策略。

3. 仿真模拟

3.1. 模型和数据生成

$$\begin{cases}logit\left(\theta_i\right)=\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_{20}x_{i20}\\logit\left(p_i\right)=\gamma_1w_{i1}+\gamma_2w_{i2}+\cdots+\gamma_{20}w_{i20}\end{cases}$$

本文采用了 Anne Buu 提出的数据生成方案，在仿真研究中，协变量取自多元正态分布$N_{20}\left(0,\Sigma\right)$，协方差阵Σ包含元素$\rho^{|i-j|}\left(i,j=1,\cdots,20\right)$。协变量之间的相关性$\rho$分别设置为0.4和0.8；样本量$n$ = 100, 300。计数部分和退化零部分的协变量相同，选择这些参数是为了获得不同数量的协变量和不同比例的零膨胀率。

模拟 1：设置参数使得响应变量 Y 的零膨胀率为 29%，$\rho=0.4$，样本量 n 分别为 100 和 300，重复 100 次实验；模型中的真实参数向量为：

$$\beta=\left(1.5,0,0,0,-0.22,0,0,0,-0.25,0,0.20,0,0.30,-0.32,0,0,0,0.20,0,0,0\right)$$

$$\gamma=(-1.05,0,0.45,0,-0.3,0,0,0,0,0,-0.33,0,-0.39,0,0,0.3,0,0,0,0.36,0)$$

模拟 2：设置参数使得响应变量 Y 的零膨胀率为 29%，$\rho=0.8$，样本量 n 分别为 100 和 300，重复 100 次实验；模型中的真实参数向量为：

$$\beta=(1.5,0,0,0,-0.22,0,0,0,-0.25,0,0.20,0,0.30,-0.32,0,0,0,0.20,0,0,0)$$

$$\gamma=(-1.05,0,0.45,0,-0.3,0,0,0,0,0,-0.33,0,-0.39,0,0,0.3,0,0,0,0.36,0)$$

模拟 3：设置参数使得响应变量 Y 的零膨胀率为 55%，$\rho=0.4$，样本量 n 分别为 100 和 300，重复 100 次实验；模型中的真实参数向量为：

$$\beta=(1.10,0,0,0,-0.36,0,0,0,0,0,0,0,-0.32,0,0,0,0,0,0,0)$$

$$\gamma=(0.3,-0.48,0,0,0,0.4,0,0,0,0,0.44,0,0.44,0,0,0,0,0,0,0,0)$$

模拟 4：设置参数使得响应变量 Y 的零膨胀率为 55%，$\rho=0.8$，样本量 n 分别为 100 和 300，重复 100 次实验；模型中的真实参数向量为：

$$\beta=(1.10,0,0,0,-0.36,0,0,0,0,0,0,0,-0.32,0,0,0,0,0,0,0)$$

$$\gamma=(0.3,-0.48,0,0,0,0.4,0,0,0,0,0.44,0,0.44,0,0,0,0,0,0,0,0)$$

3.2. 仿真结果

所选择的统计评估方法有惩罚 ZIGe 回归，显著水平分别为 0.01，0.05 和 0.1573 的逐步后退法以及假设非零系数已知的惩罚 ZIGe 回归模型。$BE(\alpha)$ 是从全 ZIGe 模型中去除不重要的变量(包括 Ge 和零分量中的所有预测变量)，然后用剩余的预测因子重新调整模型，重复该过程，直到所有变量都基于 α 水平显著。估计的精度度通过所选择的模型和全 ZIGe 模型之间的参数均方误差比来衡量。为了评估变量选择的性能，我们计算了灵敏度(正确识别的非零系数数量的比例)和特异性(正确识别的零系数数量的比例)。为了比较预测精度，我们从每个模型中生成了与训练数据相同数量的测试观测值，并使用从训练数据中估计的参数来计算对数似然值并用箱线图来表示。参数估计和变量选择的结果总结在表 1 和表 2 以及附录(附表 1 和附表 2)里面。

通过表 1、表 2、附表 1、附表 2 的模拟结果得到：

1) 在零膨胀几何模型中，无论是零部分还是几何部分，LASSO 在均方误差和特异性上都比 SCAD 和 MCP 表现更好，在敏感性上相反，ZIGe-MCP 和 ZIGe-SCAD 方法之间没有明显的差别。BE 方法的结果因显著性水平而异：在显著性水平较小时($\alpha=0.01$)，用较小的均方误差产生了更精确的估计，但可能选择较少的变量，导致不太精确的灵敏度；在显著性水平较大时($\alpha=0.1573$)，BE 方法保留了更多的预测变量并且提高了灵敏度。就均方误差而言，BE (0.01)与 ZIGe-MCP 和 ZIGe-SCAD 相比具有竞争力。然而在所有的模拟情况中，BE (0.01)在零部分的敏感度最小，在小样本情况下尤其明显。

2) 随着样本量的增加，所有方法都有较好的性能(均方误差值降低)，零分量的灵敏度随之增加。表中的均方误差被定义为一个比率，而不是绝对值。当相关性增加时，我们在表 1 和附表 1 或者表 2 和附表 2 的结果中并未表明敏感性有明显变化。

3) 图 2 和图 3 展示了预测的对数似然值。惩罚的零膨胀几何模型比 BE 方法的预测效果要稍好点。在模拟 1 中，样本量为 100 时，LASSO 惩罚的表现力相对来说最好的，SCAD 没有优势，BE 方法随着显著性水平的增加，预测效果逐渐变差；而当样本量增加为 300 时，几种惩罚方法的效果都有所提高。在模拟 4 中，ρ 值变为 0.8，零膨胀率由原来的 29%变为 55%，模型具有较好的预测能力。

Table 1. Results of simulation 1 (ρ = 0.4)
表 1. 模拟 1 (ρ = 0.4)的结果

Method	n = 100			n = 300		
	MSE	Sensitivity	Specificity	MSE	Sensitivity	Specificity
			Geometric component			
BE (0.01)	0.666 (0.497)	0.369 (0.303)	0.951 (0.074)	0.538 (0.411)	0.841 (0.172)	0.98 (0.043)
BE (0.05)	0.704 (0.408)	0.558 (0.3)	0.812 (0.14)	0.592 (0.336)	0.916 (0.125)	0.923 (0.092)
BE (0.1573)	0.824 (0.391)	0.707 (0.243)	0.648 (0.174)	0.719 (0.29)	0.956 (0.099)	0.809 (0.121)
ZIGe-LASSO	0.606 (0.277)	0.246 (0.316)	0.952 (0.095)	0.519 (0.405)	0.454 (0.252)	0.948 (0.077)
ZIGe-MCP	0.824 (0.391)	0.707 (0.243)	0.648 (0.174)	0.548 (0.372)	0.859 (0.185)	0.965 (0.06)
ZIGe-SCAD	0.683 (0.405)	0.463 (0.285)	0.886 (0.103)	0.584 (0.37)	0.889 (0.165)	0.935 (0.085)
ZIGe-ORACLE	0.161 (0.111)	1 (0)	1 (0)	0.232 (0.147)	1 (0)	1 (0)
			Zero component			
BE (0.01)	0.041 (0.059)	0.069 (0.101)	0.992 (0.022)	0.655 (0.434)	0.292 (0.192)	0.984 (0.033)
BE (0.05)	0.096 (0.137)	0.274 (0.224)	0.903 (0.107)	0.602 (0.299)	0.533 (0.19)	0.927 (0.08)
BE (0.1573)	0.337 (0.316)	0.485 (0.224)	0.653 (0.182)	0.676 (0.252)	0.719 (0.179)	0.797 (0.123)
ZIGe-LASSO	0.036 (0.051)	0.168 (0.137)	0.962 (0.081)	0.519 (0.405)	0.454 (0.252)	0.948 (0.077)
ZIGe-MCP	0.205 (0.252)	0.506 (0.238)	0.691 (0.238)	0.635 (0.276)	0.733 (0.184)	0.777 (0.176)
ZIGe-SCAD	0.095 (0.124)	0.257 (0.173)	0.92 (0.096)	0.619 (0.349)	0.429 (0.215)	0.962 (0.053)
ZIGe-ORACLE	0.041 (0.055)	1 (0)	1 (0)	0.177 (0.125)	1 (0)	1 (0)

Table 2. Results of simulation 4 (ρ = 0.8)
表 2. 模拟 4 (ρ = 0.8)的结果

Method	n = 100			n = 300		
	MSE	Sensitivity	Specificity	MSE	Sensitivity	Specificity
			Geometric component			
BE (0.01)	0.059 (0.053)	0.295 (0.282)	0.903 (0.127)	0.177 (0.222)	0.611 (0.359)	0.949 (0.059)
BE (0.05)	0.089 (0.107)	0.467 (0.322)	0.731 (0.174)	0.353 (0.297)	0.668 (0.339)	0.868 (0.101)
BE (0.1573)	0.205 (0.279)	0.638 (0.298)	0.565 (0.178)	0.64 (0.289)	0.742 (0.317)	0.722 (0.137)
ZIGe-LASSO	0.033 (0.027)	0.295 (0.334)	0.932 (0.076)	0.118 (0.095	0.837 (0.285)	0.887 (0.076)
ZIGe-MCP	0.047 (0.042)	0.707(0.243)	0.648(0.174)	0.158 (0.203)	0.566 (0.377)	0.957 (0.045)
ZIGe-SCAD	0.06 (0.059)	0.34 (0.347)	0.909 (0.077)	0.191 (0.218)	0.648 (0.368)	0.93 (0.069)
ZIGe-ORACLE	0.007 (0.007)	1 (0)	1 (0)	0.018 (0.017)	1 (0)	1 (0)
			Zero component			
BE (0.01)	8e−05 (0.00012)	0.084 (0.124)	0.98 (0.037)	0.287 (0.202)	0.376 (0.233)	0.962 (0.043)
BE (0.05)	0.00035 (0.0005)	0.247 (0.206)	0.87 (0.145)	0.355 (0.244)	0.526 (0.251)	0.896 (0.086)
BE (0.1573)	0.00479 (0.00708)	0.482 (0.246)	0.609 (0.188)	0.53 (0.284)	0.634 (0.241)	0.756 (0.132)
ZIGe-LASSO	6e−05 (8e−05)	0.14 (0.192)	0.969 (0.04)	0.191 (0.134)	0.365 (0.239)	0.952 (0.056)
ZIGe-MCP	5e−04 (0.00073)	0.358 (0.242)	0.877 (0.11)	0.234 (0.154)	0.487 (0.25)	0.95 (0.044)
ZIGe-SCAD	3.3e−04 (4.8e−04)	0.185 (0.183)	0.957 (0.055)	0.252 (0.188)	0.349 (0.223)	0.973 (0.039)
ZIGe-ORACLE	7e−05 (1.1e−04)	1 (0)	1 (0)	0.046 (0.037)	1 (0)	1 (0)

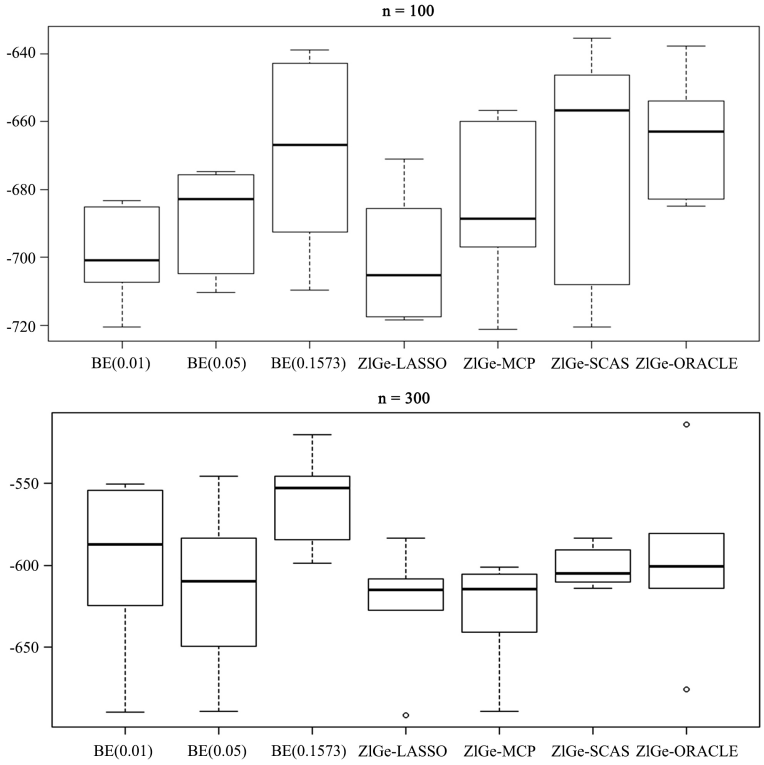

Figure 2. The log-likelihood value of simulation 1
图 2. 模拟 1 的对数似然值

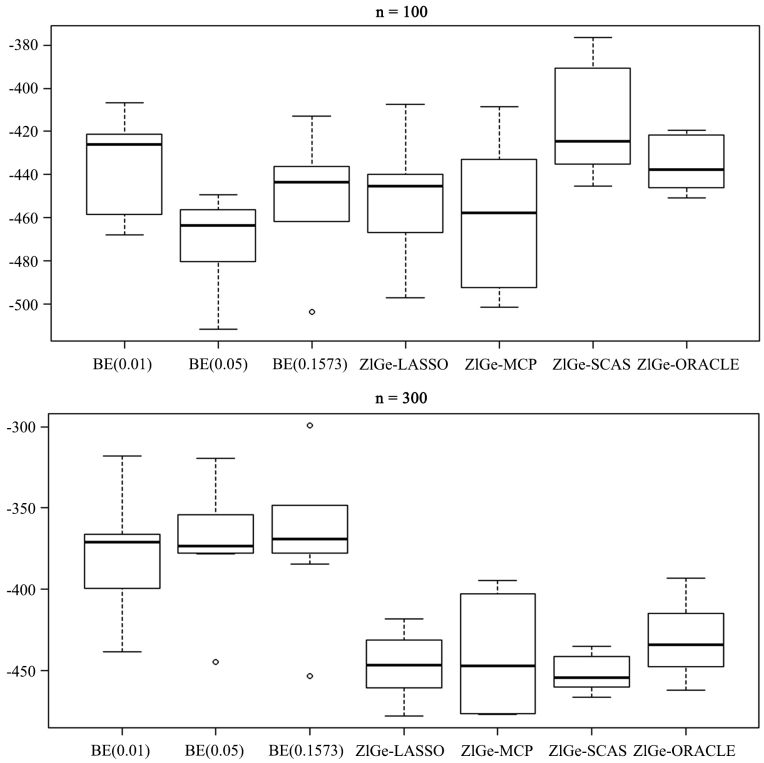

Figure 3. The log-likelihood value of simulation 4
图 3. 模拟 4 的对数似然值

4. 讨论

预测因子的数量很大，而样本量却是小样本或者中等样本时，混合模型的变量选择是一个具有挑战性的问题，本文研究了 ZIGe 模型中变量选择的 EM 的性能。在实际建模中发现，数据中的协变量多且具有相关性，有的协变量对模型建立的效果影响微乎其微，有的协变量之间存在多重共线性。针对此问题，在最大似然函数的基础上添加 SCAD、MCP 和 LASSO 惩罚，得到基于零膨胀几何回归的惩罚目标函数，然后利用 EM 算法研究模型的参数估计和变量选择问题。仿真结果表明：该模型不仅具有准确的参数估计，而且比传统的逐步选择方法更优越。

在今后的研究中，可考虑如下四个方向的发展：

一是零膨胀几何模型推广。本文只考虑了零膨胀模型，若数据相关，可以推广到随机效应零膨胀几何回归模型或者半参数零膨胀模型中；若数据中零值和一值都很多，可推广到零一膨胀几何回归模型中。

二是惩罚函数的选择，本文考虑了常见的三种惩罚函数，后续可以尝试弹性网惩罚，自适应弹性网惩罚、贝叶斯惩罚等技术。

三是参数估计的求解方法，本文选择 EM 算法，另一种算法是采用对数似然函数的泰勒近似。但是，这种方法需要反演 Hessian 矩阵，相对比较复杂，但值得尝试。

四是未找到合适的数据进行实例分析，希望在下一步的研究中可以找到适合的数据，将理论运用到实际中，证实理论的正确性。

致　谢

衷心地感谢我的研究生导师赵丽华老师，在整个学习及论文写作过程中，认真的指导；在生活中，经常告诫我不要有太大压力，做好自己，让我在漫漫学涯中感受到了温暖和肯定。其次，感谢给予转载和引用权的文献及研究思想和设想的所有者们，正是借助你们的肩膀，我才能够更好地完成论文的撰写。

参考文献

[1] Cohen Jr., A.C. (1960) Estimating the Parameters of a Modified Poisson Distribution. *Journal of the American Statistical Association*, **55**, 139-143. https://doi.org/10.1080/01621459.1960.10482054

[2] Mullahy, J. (1986) Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, **33**, 341-365. https://doi.org/10.1016/0304-4076(86)90002-3

[3] Lambert, D. (1992) Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14. https://doi.org/10.2307/1269547

[4] Lee, J.H., Han, G., Fulp, W.J., *et al*. (2012) Analysis of Overdispersed Count Data: Application to the Human Papillomavirus Infection in Men (HIM) Study. *Epidemiology & Infection*, **140**, 1087-1094. https://doi.org/10.1017/S095026881100166X

[5] Lee, S.M., Li, C.S., Hsieh, S.H., *et al*. (2012) Semiparametric Estimation of Logistic Regression Model with Missing Covariates and Outcome. *Metrika*, **75**, 621-653. https://doi.org/10.1007/s00184-011-0345-9

[6] Huang, L., Zheng, D., Zalkikar, J., *et al*. (2017) Zero-Inflated Poisson Model Based Likelihood Ratio Test for Drug Safety Signal Detection. *Statistical Methods in Medical Research*, **26**, 471-488. https://doi.org/10.1177/0962280214549590

[7] Liu, H. (2007) Growth Curve Models for Zero-Inflated Count Data: An Application to Smoking Behavior. *Structural Equation Modeling*: *A Multidisciplinary Journal*, **14**, 247-279. https://doi.org/10.1080/10705510709336746

[8] 肖翔, 刘福窑. 零膨胀几何分布的参数估计[J]. 上海工程技术大学学报, 2018, 32(3): 267-271+277.

[9] 肖翔. 0-1 膨胀几何分布回归模型及其应用[J]. 系统科学与数学, 2019, 39(9): 1486-1499.

[10] Breiman, L. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, **37**, 373-384. https://doi.org/10.1080/00401706.1995.10484371

[11] Shen, X. and Ye, J. (2002) Adaptive Model Selection. *Journal of the American Statistical Association*, **97**, 210-221. https://doi.org/10.1198/016214502753479356

[12] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[13] Frank, L.L.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[14] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[15] Meinshausen, N. (2007) A Note on the Lasso for Gaussian Graphical Model Selection. *Statistics and Probability Letters*, **78**, 880-884. https://doi.org/10.1016/j.spl.2007.09.014

[16] 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

[17] Fan, J. and Li, R. (2006) Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the International Congress of Mathematicians*, Madrid, 22-30 August 2006, 595-622.

[18] 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

[19] 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

[20] Buu, A., Johnson, N.J., Li, R., *et al*. (2011) New Variable Selection Methods for Zero-Inflated Count Data with Applications to the Substance Abuse Field. *Statistics in Medicine*, **30**, 2326-2340. https://doi.org/10.1002/sim.4268

[21] Wang, Z., Ma, S., Wang, C.Y., *et al*. (2014) EM for Regularized Zero-Inflated Regression Models with Applications to Postoperative Morbidity after Cardiac Surgery in Children. *Statistics in Medicine*, **33**, 5192-5208. https://doi.org/10.1002/sim.6314

[22] Chen, T., Wu, P., Tang, W., *et al*. (2016) Variable Selection for Distribution-Free Models for Longitudinal Zero-Inflated Count Responses. *Statistics in Medicine*, **35**, 2770-2785. https://doi.org/10.1002/sim.6892

附录

Table A1. Results of simulation 2 (ρ = 0.8)
附表 1. 模拟 2 (ρ = 0.8)的结果

Method	n = 100			n = 300		
	MSE	Sensitivity	Specificity	MSE	Sensitivity	Specificity
Geometric component						
BE (0.01)	0.297 (0.21)	0.183 (0.215)	0.916 (0.105)	0.719 (0.368)	0.467 (0.225)	0.937 (0.063)
BE (0.05)	0.498 (0.311)	0.354 (0.271)	0.775 (0.137)	0.62 (0.334)	0.607 (0.246)	0.88 (0.107)
BE (0.1573)	0.758 (0.277)	0.537 (0.233)	0.595 (0.138)	0.74 (0.263)	0.726 (0.225)	0.757 (0.133)
ZIGe-LASSO	0.214 (0.09)	0.102 (0.16)	0.95 (0.075)	0.737 (0.338)	0.307 (0.298)	0.917 (0.087)
ZIGe-MCP	0.245 (0.145)	0.12 (0.14)	0.929 (0.071)	0.691 (0.252)	0.398 (0.268)	0.939 (0.057)
ZIGe-SCAD	0.308 (0.178)	0.181 (0.17)	0.902 (0.095)	0.654 (0.276)	0.488 (0.284)	0.917 (0.071)
ZIGe-ORACLE	0.086 (0.071)	1 (0)	1 (0)	0.111 (0.063)	1 (0)	1 (0)
Zero component						
BE (0.01)	0.031 (0.039)	0.06 (0.103)	0.965 (0.053)	0.287 (0.202)	0.376 (0.233)	0.962 (0.043)
BE (0.05)	0.106 (0.144)	0.238 (0.159)	0.852 (0.122)	0.355 (0.244)	0.526 (0.251)	0.896 (0.086)
BE (0.1573)	0.354 (0.36)	0.44 (0.215)	0.653 (0.151)	0.53 (0.284)	0.634 (0.241)	0.756 (0.132)
ZIGe-LASSO	0.017 (0.023)	0.062 (0.106)	0.969 (0.051)	0.207 (0.145)	0.176 (0.181)	0.951 (0.061)
ZIGe-MCP	0.14 (0.167)	0.366 (0.214)	0.707 (0.176)	0.352 (0.256)	0.445 (0.208)	0.764 (0.148)
ZIGe-SCAD	0.043 (0.049)	0.134 (0.142)	0.926 (0.071)	0.313 (0.169)	0.252 (0.167)	0.918 (0.069)
ZIGe-ORACLE	0.018 (0.026)	1 (0)	1 (0)	0.074 (0.059)	1 (0)	1 (0)

Table A2. Results of simulation 3 (ρ = 0.4)
附表 2. 模拟 3 (ρ = 0.4)的结果

Method	n = 100			n = 300		
	MSE	Sensitivity	Specificity	MSE	Sensitivity	Specificity
Geometric component						
BE (0.01)	0.185 (0.156)	0.406 (0.369)	0.903 (0.127)	0.177 (0.235)	0.824 (0.273)	0.968 (0.053)
BE (0.05)	0.563 (0.283)	0.562 (0.33)	0.672 (0.201)	0.404 (0.353)	0.918 (0.201)	0.879 (0.113)
BE (0.1573)	0.784 (0.201)	0.735 (0.282)	0.428 (0.163)	0.614 (0.279)	0.956 (0.161)	0.748 (0.129)
ZIGe-LASSO	0.063 (0.051)	0.337 (0.405)	0.953 (0.082)	0.243 (0.202)	0.862 (0.278)	0.924 (0.087)
ZIGe-MCP	0.111 (0.089)	0.418 (0.364)	0.933 (0.072)	0.1 (0.133)	0.846 (0.243)	0.974 (0.048)
ZIGe-SCAD	0.151 (0.15)	0.541 (0.364)	0.884 (0.099)	0.167 (0.214)	0.856 (0.239)	0.958 (0.056)
ZIGe-ORACLE	0.023 (0.025)	1 (0)	1 (0)	0.042 (0.042)	1 (0)	1 (0)
Zero component						
BE (0.01)	1.97e−02 (3e−02)	0.078 (0.118)	0.984 (0.035)	0.403 (0.308)	0.58 (0.241)	0.987 (0.029)
BE (0.05)	3e−02 (0.05347)	0.344 (0.268)	0.879 (0.12)	0.406 (0.319)	0.731 (0.212)	0.922 (0.085)
BE (0.1573)	3.6e−02 (0.484)	0.551 (0.228)	0.618 (0.193)	0.607 (0.24)	0.83 (0.174)	0.788 (0.134)
ZIGe-LASSO	8e−05 (1e−04)	0.148 (0.179)	0.985 (0.036)	0.381 (0.343)	0.614 (0.306)	0.973 (0.043)
ZIGe-MCP	2.5e−04 (4e−04)	0.508 (0.245)	0.779 (0.2)	0.327 (0.282)	0.745 (0.201)	0.938 (0.09)
ZIGe-SCAD	1.7e−04 (2e−04)	0.316 (0.22)	0.947 (0.079)	0.382 (0.316)	0.59 (0.234)	0.981 (0.036)
ZIGe-ORACLE	6e−05 (8e−05)	1 (0)	1 (0)	0.119 (0.091)	1 (0)	1 (0)