Journal of Applied Statistics, 2014
Vol. 41, No. 4, 879–894, http://dx.doi.org/10.1080/02664763.2013.858672

Taylor & Francis
Taylor & Francis Group

Variable selection approach for zero-inflated count data via adaptive lasso

Ping Zeng[a,b], Yongyue Wei[a], Yang Zhao[a], Jin Liu[a], Liya Liu[a], Ruyang Zhang[a], Jianwei Gou[a], Shuiping Huang[b] and Feng Chen[a]*

[a]*Department of Epidemiology and Biostatistics, Nanjing Medical University, Nanjing, People's Republic of China;* [b]*Department of Epidemiology and Biostatistics, Xuzhou Medical College, Xuzhou, People's Republic of China*

(*Received 1 February 2013; accepted 21 October 2013*)

This article proposes a variable selection approach for zero-inflated count data analysis based on the adaptive lasso technique. Two models including the zero-inflated Poisson and the zero-inflated negative binomial are investigated. An efficient algorithm is used to minimize the penalized log-likelihood function in an approximate manner. Both the generalized cross-validation and Bayesian information criterion procedures are employed to determine the optimal tuning parameter, and a consistent sandwich formula of standard errors for nonzero estimates is given based on local quadratic approximation. We evaluate the performance of the proposed adaptive lasso approach through extensive simulation studies, and apply it to analyze real-life data about doctor visits.

Keywords: adaptive lasso; variable selection; zero-inflated Poisson; zero-inflated negative binomial; coordinate descent algorithm; Taylor approximation

AMS Subject Classifications: Primary: 62J07; Secondary: 62P10

1. Introduction

Building models for count data is common in medical, social, and economic sciences [3,23,44]. In practice, count data with excess zeros are often observed, i.e. the data display much more zeros than are consistent with either the Poisson or negative binomial model. This phenomenon is referred to as zero-inflation [22]. The zero-inflated Poisson (ZIP) model with a fixed mixture proportion was first proposed by Mullahy [27]. Lambert [22] made a generalization by specifying a logistic model for the mixture proportion. Greene [17] further extended Lambert's ZIP model to the zero-inflated negative binomial (ZINB) model. Since then the zero-inflated count models have been intensively studied [21,39,45].

Usually, an important task of the zero-inflated count models is to select variables with significant impacts on the number of occurrences of an event. The traditional variable selection methods

*Corresponding author. Email: fengchen@njmu.edu.cn

(e.g. best subset searching and stepwise procedures) can be employed but may suffer from some shortcomings [1,12]. For example, these methods are unstable due to the separation of selection and estimation [1]. Recently, a penalization-based technique, known as the least absolute shrinkage and selection operator (lasso) [37], was developed for simultaneous shrinkage estimation and variable selection. The lasso and its variants, such as elastic net [49], adaptive lasso [48], and relaxed lasso [26], have been widely applied in general linear models [8,37], survival models [24,35,38,46], and generalized linear models [15,31]. However, unlike the models mentioned above, the zero-inflated count models are finite mixture of regression (FMR) models [25] with two components, each of them may have different explanatory variables, which makes the variable selection much more complex. Applying lasso-type variable selection method to the zero-inflated count models is attractive, but to the best of our knowledge, few papers have been published previously on this topic. This motivates us to develop an adaptive lasso approach for the variable selection of the ZIP and ZINB models.

The rest of the article is organized as follows. In Section 2 we give a brief introduction to the ZIP and ZINB models. In Section 3 we investigate the adaptive lasso approach for the two models. We perform extensive simulations to evaluate the performance of the proposed approach in Section 4. Section 5 is an application on real data about the number of doctor visits. In Section 6 we give discussion and emphasize the areas for further investigation.

2. ZIP and ZINB

Let y be the response variable taking only on non-negative integers, then the ZIP model [22] is

$$f_{\mathrm{ZIP}}(y) = \varphi I(y = 0) + (1 - \varphi) f_{\mathrm{Pois}}(y), \tag{1}$$

where φ is the mixture proportion and $I(\cdot)$ denotes an indicator equaling 1 when $y = 0$ and 0 otherwise, and $f_{\mathrm{Pois}}(y)$ denotes the Poisson density

$$f_{\mathrm{Pois}}(y) = \frac{e^{-\mu}\mu^y}{y!}, \ \log(\mu) = \sum_{j=0}^{p} x_j \beta_j = \mathbf{x}'\boldsymbol{\beta}, \tag{2}$$

with mean $E(y) = \mu$ and variance $\mathrm{Var}(y) = \mu$. Here $\mathbf{x} = [x_0, x_1, \ldots, x_p]$ denotes a set of explanatory variables and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]$ is a $p + 1$ dimensional vector of the unknown regression coefficients. We assume that $x_0 = 1$ represents the intercept term β_0. The ZIP model has mean $E(y) = (1 - \varphi)\mu$ and variance $\mathrm{Var}(y) = [(1 - \varphi)\mu](1 + \varphi\mu)$.

The ZINB model is

$$f_{\mathrm{ZINB}}(y) = \varphi I(y = 0) + (1 - \varphi) f_{\mathrm{NB}}(y), \tag{3}$$

where $f_{\mathrm{NB}}(y)$ denotes the negative binomial density [3]

$$f_{\mathrm{NB}}(y) = \frac{\Gamma(y + \alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu}\right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu}\right)^{y}, \tag{4}$$

with mean $E(y) = \mu$ and variance $\mathrm{Var}(y) = \mu + a\mu^2$. Here $a \geq 0$ is called the dispersion parameter, and μ is related to $\mathbf{x}$ by a log-link function as the same way as in the Poisson. The ZINB model has mean $E(y) = (1 - \varphi)\mu$ and variance $\mathrm{Var}(y) = [(1 - \varphi)\mu][1 + \mu(\varphi + \alpha)]$.

The mixture proportion φ is usually parameterized via a logit link function [22]

$$\text{logit}(\varphi) = \sum_{g=0}^{q} z_g \gamma_g = \mathbf{z}'\boldsymbol{\gamma}, \tag{5}$$

in which $\mathbf{z} = [z_0, z_1, \ldots, z_q]$ denotes the zero-inflated explanatory variables, and $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \ldots, \gamma_q]$ is a $q+1$ dimensional vector of the corresponding coefficients. We assume that $z_0 = 1$ represents the intercept term γ_0.

3. Adaptive lasso

3.1 *Adaptive lasso for ZIP and ZINB*

Let $L(\boldsymbol{\theta})$ be the log-likelihood function of the ZIP or ZINB model, where $\boldsymbol{\theta}$ are the parameters of interest. For convenience, we suppose that the last two elements of $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$ are β_0 and γ_0 for the ZIP model, or that the last three elements of $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}]$ are β_0, γ_0, and α for the ZINB model, and that the first $p+q$ elements of $\boldsymbol{\theta}$ are β_js $(j = 1, 2, \ldots, p)$ and γ_gs $(g = 1, 2, \ldots, q)$. The lasso estimators [37] are given by

$$\hat{\boldsymbol{\theta}}_{\text{lasso}}(\lambda) = \arg\min \left\{ -L(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p+q} |\theta_d| \right\}. \tag{6}$$

Note that we do not penalize β_0 and γ_0 as well as α [37]. The second term in Equation (6) is called L_1 penalty, due to which the lasso can shrink small coefficients to be exactly zeros. Here $\lambda \geq 0$ is a tuning parameter controlling the amount of shrinkage. However, the lasso imposes the same penalty on all the regression coefficients, which over-penalizes the important ones and accordingly results in biased estimators [12,48]. The adaptive lasso [48] offers an effective way to fix this bias. It has been shown that the adaptive lasso enjoys the oracle property [48], i.e. the adaptive lasso is consistent in variable selection, and its estimators are asymptotically normal and unbiased. More explicitly, it works as well as knowing the true model in advance [10,48].

The adaptive lasso estimators [48] are defined as

$$\hat{\boldsymbol{\theta}}_{\text{alasso}}(\lambda) = \arg\min \left\{ -L(\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p+q} \tau_d |\theta_d| \right\}, \tag{7}$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{p+q})$ are adaptive weights, which are usually set to $1/|\hat{\boldsymbol{\theta}}|$, here $\hat{\boldsymbol{\theta}}$ are any consistent estimators to $\boldsymbol{\theta}$, e.g. the maximum likelihood estimators (MLEs) $\hat{\boldsymbol{\theta}}_{\text{ml}}$ [48].

3.2 *Taylor approximation algorithm*

Theoretically the adaptive lasso estimators in Equation (7) can be obtained by minimizing the penalized log-likelihood function. However, it will be highly difficult because of the mixture form. Here we utilize an efficient approach based on a second-order Taylor series expansion of $L(\boldsymbol{\theta})$ with regard to $\hat{\boldsymbol{\theta}}_{\text{ml}}$

$$L(\boldsymbol{\theta}) \approx L(\hat{\boldsymbol{\theta}}_{\text{ml}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})' L'(\hat{\boldsymbol{\theta}}_{\text{ml}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})' L''(\hat{\boldsymbol{\theta}}_{\text{ml}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}}), \tag{8}$$

where L' and L'' are the first and second derivatives, respectively. In the right side of Equation (8), $L(\hat{\boldsymbol{\theta}}_{\text{ml}})$ is a constant and $L'(\hat{\boldsymbol{\theta}}_{\text{ml}}) = 0$. So the penalized log-likelihood function is equivalent to

$$\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}})'[-L''(\hat{\boldsymbol{\theta}}_{\text{ml}})](\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\text{ml}}) + \lambda \sum_{d=1}^{p+q} \tau_d \, |\theta_d|. \tag{9}$$

Equation (9) was called least-squares approximation (LSA) in [41]. Let $\hat{\Sigma}$ be the estimated variance–covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{ml}}$, which can be available from many statistical packages and provides a way to estimate the Hessian matrix L'', that is, $L'' = -\hat{\Sigma}^{-1}$. In our article, we use the R package pscl [18,32] to obtain $\hat{\Sigma}$. Then a pseudo-data can be constructed as

$$\mathbf{X}^* = \hat{\boldsymbol{\Sigma}}^{-1/2}, \mathbf{Y}^* = \hat{\boldsymbol{\Sigma}}^{-1/2}\hat{\boldsymbol{\theta}}_{\text{ml}}. \tag{10}$$

Here $\mathbf{X}^*$ is a square matrix of order $p + q + 2$ for the ZIP model or $p + q + 3$ for the ZINB model, and $\boldsymbol{Y}^*$ is a vector corresponding to $\mathbf{X}^*$. Accordingly, Equation (7) can be re-expressed as

$$\hat{\boldsymbol{\theta}}_{\text{alasso}}(\lambda) \approx \arg\min \left\{ \frac{1}{2}(\boldsymbol{Y}^* - \mathbf{X}^*\boldsymbol{\theta})'(\boldsymbol{Y}^* - \mathbf{X}^*\boldsymbol{\theta}) + \lambda \sum_{d=1}^{p+q} \boldsymbol{\tau}_{\text{d}} \, |\theta_d| \right\} \tag{11}$$

This is the familiar least squares with adaptive Lasso penalization. Various efficient algorithms can be used to conduct the minimization of Equation (11), such as the homotopy algorithm [29,30], the least angle regression (Lars) algorithm [8], the predictor–corrector algorithm [31], and the coordinate descent algorithm [14,15,35].

We adopt the coordinate descent algorithm here to carry out the adaptive lasso estimation via the soft threshold rule [7,15]

$$S\{x, \lambda\} = \begin{cases} x - \lambda, & \text{if} \quad x > 0 \text{ and } \quad \lambda < |x|, \\ x + \lambda, & \text{if} \quad x < 0 \text{ and } \quad \lambda < |x|, \\ 0 & \text{if} \quad \lambda \geq |x|. \end{cases} \tag{12}$$

Then the coordinate descent algorithm [14,15] for Equation (11) is

$$\hat{\theta}_{\text{alasso}\,d}(\lambda) = \frac{S\left\{\sum_{i=1}^{D} x_{id}^* \left[y_i^* - \sum_{l \neq d}^{D} x_{il}^* \hat{\theta}_{\text{alasso}\,l}(\lambda)\right], \lambda\tau_d\right\}}{\sum_{i=1}^{D} x_{id}^{*2}}, \quad d = 1, 2, \ldots, D, \tag{13}$$

where D is the total number of the parameters. The following steps give a simple way of performing the Taylor approximation algorithm.

(I) Obtain $\hat{\Sigma}$, $\hat{\Sigma}^{-1}$, and set $\tau = 1/\hat{\boldsymbol{\theta}}_{\text{ml}}$.
(II) Construct the pseudo-data $\mathbf{X}^*$ and $\boldsymbol{Y}^*$.
(III) Minimize Equation (11) via the coordinate descent algorithm described in Equations (12) and (13), which yields $\hat{\boldsymbol{\theta}}_{\text{alasso}}(\lambda)$.

The initial values can be set to $\hat{\boldsymbol{\theta}}_{\text{ml}}$. The adaptive weights for β_0, γ_0, and α are set to 1, and λ is set to 0 due to not being unpenalized. Clearly the matrix $\hat{\Sigma}$, or equivalently the negative Hessian matrix $-L''$, plays a crucial role in the Taylor approximation algorithm. When the different components of the ZIP or ZINB model are poorly separated, $\hat{\Sigma}$ may not be obtained accurately [33]. Thus this algorithm can encounter severe numerical problem in practice. We further discuss this issue through simulation in Section 4.

3.3 *Tuning parameter selection*

In the literature the optimal adaptive lasso estimators are generally determined by cross-validation (CV), generalized cross-validation (GCV), or using the idea of Stein's unbiased risk estimate [37]. The CV procedure requires splitting the data by random [9], thus the selected tuning parameter is unstable. Here, following the approach of Fan and Li [10], we use the GCV procedure [16]

$$\mathrm{GCV}(\lambda) = -\frac{L[\hat{\boldsymbol{\theta}}_{\mathrm{alasso}}(\lambda)]}{n\,[1 - e(\lambda)/n]^2}, \tag{14}$$

where $e(\lambda)$ is the effective number of parameters. For simplicity, we estimate $e(\lambda)$ with the number of nonzero coefficients (except intercepts and α) in the ZIP and ZINB models as done in [36].

The Bayesian information criterion (BIC) [34] procedure is also considered

$$\mathrm{BIC}(\lambda) = -2L[\hat{\boldsymbol{\theta}}_{\mathrm{alasso}}(\lambda)] + e(\lambda)\log(n). \tag{15}$$

It has been shown that the BIC is consistent in variable selection [28,42,43]. We calculate GCV (or BIC) over a grid of candidate values for λ, and select the optimal tuning parameter $\lambda_1 = \arg\min \mathrm{GCV}(\lambda)$(or $\lambda_1 = \arg\min \mathrm{BIC}(\lambda)$).

3.4 *Standard error formula*

At present, little literature is available on the standard errors for the adaptive lasso estimates. The bootstrap method [9] used by Tibshirani [37] is computationally intensive. Alternatively, one can first select variables by using the adaptive lasso and then obtains coefficients and their standard errors with maximum likelihood method as suggested in [8]. However, doing so will lose the continuous nature of variable selection and parameter estimation shared by adaptive lasso. A sandwich formula of standard errors was provided in [10–12] based on local quadratic approximation (LQA). This formula has been proven to be consistent [12]. Suppose that $\hat{\boldsymbol{\theta}}_1$ (with T elements) are the nonzero values of adaptive lasso estimates based on the optimal tuning parameter λ_1. We use the following LQA sandwich formula:

$$\hat{V}(\hat{\boldsymbol{\theta}}_1) = [L''(\hat{\boldsymbol{\theta}}_1) - \lambda_1 \Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}_1)]^{-1} \hat{V}[L'(\hat{\boldsymbol{\theta}}_1)][L''(\hat{\boldsymbol{\theta}}_1) - \lambda_1 \Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}_1)]^{-1}, \tag{16}$$

where $\Sigma_{\lambda_1}(\hat{\boldsymbol{\theta}}_1) = \mathrm{diag}[\tau_{1t}/|\hat{\theta}_{1t}|]\,(t = 1, 2, \ldots, T)$ with the notation diag denoting diagonal matrix, and $\boldsymbol{\tau}_1$ are the corresponding adaptive weights of $\hat{\boldsymbol{\theta}}_1$. The standard errors for the zero estimates are generally set to zero [10,37,48].

4. Simulations

To evaluate the performance of the adaptive lasso, extensive simulation studies are performed. Both the mean-squared error (MSE)

$$\mathrm{MSE} = \sum\nolimits_d (\hat{\theta}_d - \theta_d)^2, \tag{17}$$

and the true (false) positive are used as measurements of the performance. The true positive indicates the number of nonzero coefficients correctly estimated to be nonzero, and the false positive indicates the number of zero coefficients incorrectly estimated to be nonzero. In the article, TP1 and TP0 represent true positive for the count and zero components, respectively. FP1 and FP0 represent false positive for the count and zero components, respectively. The number of runs is 100.

4.1 *Simulation 1*

The first simulation is designed to assess that on what degree the proposed approach is influenced by the divergence between different mixture components. Following Redner and Walker [33], if the difference of population means of the two components is small, then it is called poor separation, otherwise well separation. The explanatory variables $\mathbf{x} = \mathbf{z}$ are generated from the standard normal distribution with pairwise correlation between x_j and x_l to be $0.5^{|j-l|}$, but for the ZIP model their absolute values (i.e. $|\mathbf{x}|$ and $|\mathbf{z}|$) are used in this simulation. We set $\boldsymbol{\gamma} = [0, 0, 0, 0, 1.5, -1.0, -0.5]$, which leads to a population mean of about 0.497 for the zero component. The values of $\boldsymbol{\beta}$ are specified in order to simulate ZIP datasets that are poor-separated or well-separated (Table 1). The sample size n is 500. The results of Simulation 1 are listed in Table 2.

From Table 2, it is observed that the false positives significantly increase as the situation changes from well-separation to poor-separation although the true positives almost remain the same. For example, when the situation is well-separated (Case a), few zero responses are generated from the Poisson component because a Poisson distribution with large mean has very small probability of presence of zeros. Under this situation, we can identify both the nonzero and zero coefficients with rather low errors. However, when the situation is poor-separation (Case d) where the Poisson component has a small mean, a large number of zero responses are generated from the Poisson component. Under this situation, it is difficult to distinguish which components a zero response belongs to. As a result the false positives for both the zero and Poisson components are high.

Checking whether the smallest eigenvalue of $-L''$s far away from zero [33] provides a way to investigate the behavior of the Taylor approximation algorithm. Figure 1 shows that the size of the smallest eigenvalue varies with the extent of the separated situation. If the smallest eigenvalue is much bigger than zero (Figure 1(a)), the inverse of $-L''$ can be calculated stably and accurately, then the algorithm works well. On the other hand, if the smallest eigenvalue is very close to zero (Figure 1(b)), $-L''$ may be nearly singular, hence the algorithm behaves poorly. Clearly the characteristic of $-L''$ determines the applicability of the proposed method.

Table 1. Designs of poor and well separations for Simulation 1.

Situations	Values of $\boldsymbol{\beta}$	μ_0	$\{y = 0\}^{\#}$	μ_1
Case a	[0, 1.5, 1.0, 0.5, 0, 0, 0]	0.497	0.508	41.762
Case b	[0, 1.5, 1.0, −0.5, 0, 0, 0]	0.497	0.538	14.021
Case c	[0, 1.5, −1.0, −0.5, 0, 0, 0]	0.497	0.689	1.816
Case d	[0, -1.5, −1.0, −0.5, 0, 0, 0]	0.497	0.928	0.162

Note: μ_0 is the population mean of the zero component, μ_1 is the population mean of the Poisson component, # indicates the proportion that y is equal to zero. These values are obtained by Monte Carlo simulation. Here Case a is the most well-separated, Case d is the most poor-separated, and Cases b and c are in between the two extremes.

Table 2. Average numbers of the true positive and false positive for Simulation 1.

	GCV				BIC			
Situations	TP1	TP0	FP1	FP0	TP1	TP0	FP1	FP0
Case a	3.00	2.70	0.03	0.08	3.00	2.75	0.06	0.12
Case b	3.00	2.72	0.15	0.14	3.00	2.72	0.15	0.12
Case c	3.00	2.63	0.43	0.37	3.00	2.60	0.25	0.24
Case d	2.45	2.64	2.26	2.41	1.85	0.96	1.22	0.95

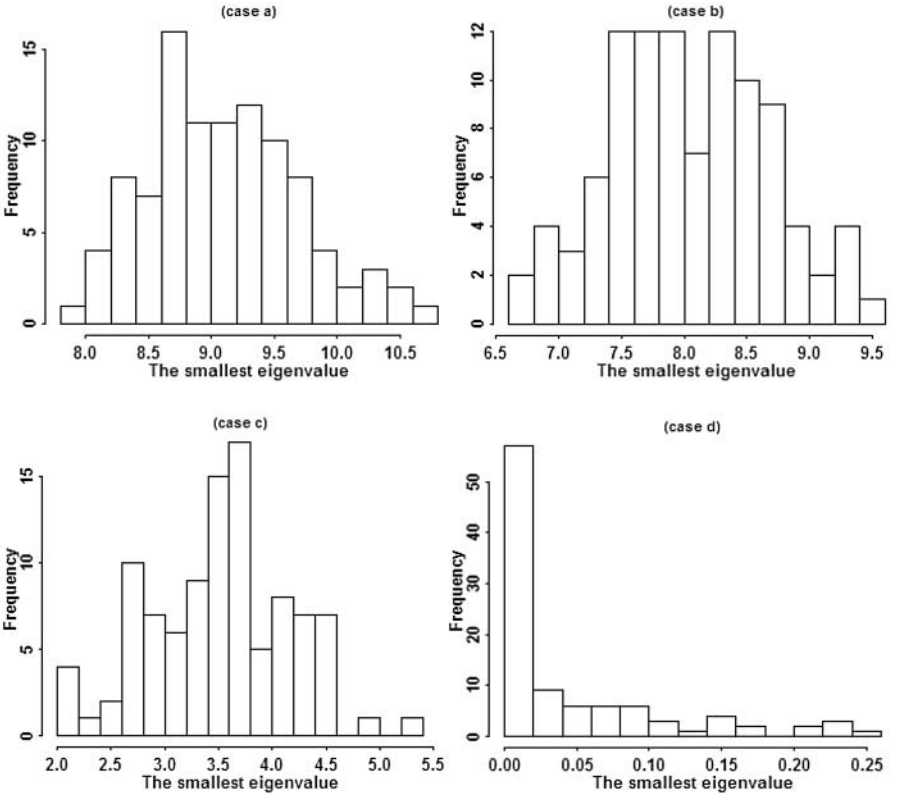

Figure 1. The smallest eigenvalues of $-L''$.

Examining the standard errors offers an empirical guideline for practical users. That is, if some standard errors are abnormally large, then it is a signal that some eigenvalues of $-L''$ are very small. Note that the variance–covariance matrix of the estimates is the inverse matrix of $-L''$. Our experience from simulation shows that the zero component in the ZIP usually displays abnormal standard errors as well as estimates when the dataset is poor-separated.

4.2 *Simulation 2*

Simulation 2 is to evaluate the performance of the adaptive lasso ZIP model with various mixture proportions φ. The explanatory variables $\mathbf{x} = \mathbf{z}$ are generated from the standard normal distribution with pairwise correlation between x_j and x_l to be $0.5^{|j-l|}$. We consider two situations,

$$\begin{cases} \boldsymbol{\beta} = [0, 2.0, -1.5, -1.0, -0.5, \underbrace{0, \ldots, 0}_{6}] \\ \boldsymbol{\gamma} = [\gamma_0, \underbrace{0, \ldots, 0}_{6}, 2.0, -1.5, -1.0, -0.5] \end{cases} \quad \text{and} \quad \begin{cases} \boldsymbol{\beta} = [0, 2.0, -1.5, -1.0, -0.5, \underbrace{0, \ldots, 0}_{16}] \\ \boldsymbol{\gamma} = [\gamma_0, \underbrace{0, \ldots, 0}_{16}, 2.0, -1.5, -1.0, -0.5] \end{cases}$$

The intercept γ_0 is chosen so that φ is approximately equal to 27%, 50%, and 73%, respectively, which are obtained by Monte Carlo simulation. The sample size n is 500 and 1000. The results are displayed in Tables 3–5. Tables 3–5, Oracle indicates the oracle model including only variables with nonzero coefficients, Full indicates the full model including all the variables, both are estimated via the maximum likelihood method. GCV indicates the adaptive lasso based

Table 3. Medians of MSE for situation 1 in Simulation 2.

	$n = 500$			$n = 1,000$		
Methods	MSET	MSE1	MSE0	MSET	MSE1	MSE0
$\varphi = 27\%$						
Oracle	0.331	0.003	0.326	0.198	0.001	0.196
Full	1.148	0.006	1.142	0.495	0.003	0.493
GCV	0.854	0.003	0.852	0.339	0.001	0.338
BIC	0.690	0.004	0.683	0.265	0.002	0.263
$\varphi = 50\%$						
Oracle	0.257	0.004	0.244	0.110	0.002	0.109
Full	0.839	0.012	0.820	0.331	0.005	0.322
GCV	0.447	0.005	0.445	0.160	0.002	0.159
BIC	0.419	0.006	0.409	0.164	0.003	0.160
$\varphi = 73\%$						
Oracle	0.316	0.010	0.287	0.150	0.004	0.147
Full	0.985	0.040	0.945	0.474	0.011	0.464
GCV	0.588	0.017	0.555	0.249	0.005	0.239
BIC	0.562	0.017	0.528	0.257	0.005	0.240

Table 4. Medians of MSE for situation 2 in Simulation 2.

	$N = 500$			$n = 1,000$		
Methods	MSET	MSE1	MSE0	MSET	MSE1	MSE0
$\varphi = 27\%$						
Oracle	0.402	0.003	0.398	0.164	0.001	0.162
Full	2.991	0.017	2.972	0.924	0.006	0.916
GCV	1.704	0.004	1.702	0.637	0.001	0.635
BIC	1.949	0.004	1.943	0.702	0.001	0.700
$\varphi = 50\%$						
Oracle	0.268	0.005	0.251	0.138	0.002	0.136
Full	1.810	0.033	1.734	0.652	0.010	0.638
GCV	0.724	0.007	0.713	0.262	0.002	0.260
BIC	0.934	0.007	0.925	0.345	0.002	0.343
$\varphi = 73\%$						
Oracle	0.343	0.014	0.312	0.176	0.001	0.174
Full	2.728	0.131	2.554	0.981	0.007	0.975
GCV	0.796	0.033	0.742	0.588	0.001	0.586
BIC	1.287	0.021	1.258	0.748	0.001	0.746

onGCV, and BIC indicates the adaptive lasso based on BIC. MSET represents the total MSE, and MSE1 and MSE0 represent the MSE for the Poisson (or negative binomial) and zero components, respectively.

Some observations from Tables 3 and 4 are listed as follows. (1) The adaptive lasso models exhibit much better than the full models. As the sample size increases, the adaptive lasso estimates converge to the oracle estimates, especially in the Poisson component. (2) MSE1 increases as φ increases, while MSE0 is the smallest when $\varphi = 50\%$, as a result the total performances (i.e. MSET) of all the methods are the best when $\varphi = 50\%$. The similar phenomenon was also observed in [20]. Städler *et al.* [36] argued that the balanced cases (i.e. φ not too large and not too small, for example, $\varphi = 50\%$) often worked better than very unbalanced cases. (3) The performance in the Poisson component is typically better than in the zero component, suggesting that it is more difficult to estimate the zero component in this simulation. (4) As the noise (i.e. the proportion of zero coefficients) increases, the performance gets worse. (5) The results of GCV and BIC are comparable with respect to MSE.

Table 5. Average numbers of the true positive and false positive for Simulation 2 with $n = 500$ (1,000).

	Situation 1		Situation 2	
	GCV	BIC	GCV	BIC
$\varphi = 27\%$				
TP1	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)
TP0	3.62 (3.85)	3.82 (3.96)	3.56 (3.77)	3.49 (3.70)
FP1	0.55 (0.33)	1.80 (1.30)	1.53 (0.73)	1.29 (0.44)
FP0	0.54 (0.31)	1.63 (1.14)	0.86 (0.41)	0.61 (0.25)
$\varphi = 50\%$				
TP1	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)
TP0	3.80 (3.96)	3.90 (3.99)	3.68 (3.93)	3.58 (3.86)
FP1	0.69 (0.61)	1.51 (1.13)	1.86 (1.07)	1.02 (0.45)
FP0	0.62 (0.46)	1.46 (1.09)	1.18 (0.73)	0.55 (0.25)
$\varphi = 73\%$				
TP1	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)
TP0	3.86 (3.95)	3.88 (3.97)	3.83 (3.82)	3.55 (3.72)
FP1	1.54 (1.09)	1.62 (1.21)	4.50 (1.15)	1.34 (0.61)
FP0	1.56 (1.17)	1.67 (1.34)	3.45 (0.82)	0.36 (0.44)

The observations from Table 5 are presented as follows. (1) The average numbers of the true positive are rather high in all the situations, but the true positives of the zero component are slightly lower than those of the Poisson component. (2) The average numbers of the false positive for GCV increase as φ increases, and also rise as the noise increases, but the performances of BIC are the best when φ is balanced, similar to the results in [20]. The average numbers of the false positive for BIC decrease as the noise increases. (3) It is interesting that the false positives for the Poisson and zero components are roughly equal in all the cases, the false positives of GCV are smaller than those of BIC when the noise is weaker (e.g. $p = q = 10$), the false positives of the Poisson component are greater than those of the zero component, and the false positives of GCV are greater than those of BIC when the noise is relatively stronger (e.g. $p = q = 20$). (4) The tuning parameters of BIC are always greater than those of GCV (data not shown), suggesting that BIC imposes heavier penalties on the coefficients than GCV, which in turn leads to more parsimonious models and also gives an explanation for why the false positives of BIC are smaller than those of GCV when the noise is strong. (5) As expected, when the sample size becomes larger, the false positive reduces.

4.3 *Simulation 3*

To further evaluate the performance of the adaptive lasso ZIP model, we simulate datasets with more practical correlation structure from the doctor visits data [19]. The explanatory variables $\mathbf{x} = \mathbf{z}$ are age, health, handicap, hdegree, married, schooling, hhincome, children, self, civil, bluec, employed, public, and addon, respectively, see Table 10 for further information. The coefficients are $\boldsymbol{\beta} = [2.235, 0.006, -0.167, 0.118, 0, 0, 0, 0, 0, 0, 0, -0.044, 0.057, 0]$ and $\boldsymbol{\gamma} = [-2.440, 0, 0.281, 0, 0, 0, 0, 0, 0.142, 0, 0, 0, 0, 0, 0]$. These are in fact the coefficients of the adaptive lasso ZIP model displayed in Table 12. The sample size is 1812. The medians of MSE are displayed in Table 6. TP1, TP0, FP1, and FP0 for the GCV (BIC) procedure are 2.89 (2.89), 1.08 (1.08), 0.25 (0.25), and 0.11 (0.10), respectively. The similar conclusions to those of Simulation 2 can be achieved in Simulation 3, except that the true positives for the Poisson component are much lower than those in Simulation 2 although the sample size is much larger. This may be

due to the more complicated correlation structure in the real data and to the smaller true nonzero coefficients.

4.4 *Simulation 4*

To evaluate the performance of the adaptive lasso ZINB model, we also simulate datasets based on the doctor visits data. The coefficients are $\boldsymbol{\beta} = [2.217, 0.006, -0.188, 0.194, 0, 0, 0, 0, 0, -0.104, 0, 0, 0, 0, 0]$ and $\boldsymbol{\gamma} = [-3.126, 0, 0.296, 0, 0, 0, 0, 0, 0, 0.368, 0, 0, 0, 0, 0, 0]$. These are in fact the coefficients of the adaptive lasso ZINB model displayed in Table 13. We set the dispersion parameter $\alpha = 0.10$ and 1.00, respectively. The simulation results are displayed in Tables 7–9.

Table 6. Medians of MSE for ZIP model based on the real doctor visits data.

Methods	MSET	MSE1	MSE0
Oracle	0.039	0.008	0.028
Full	0.715	0.041	0.675
GCV	0.083	0.020	0.056
BIC	0.083	0.020	0.056

Table 7. Medians of MSE for ZINB model based on the real doctor visits data.

	$\alpha = 0.10$			$\alpha = 1.00$		
Methods	MSET	MSE1	MSE0	MSET	MSE1	MSE0
Oracle	0.070	0.012	0.052	0.153	0.039	0.097
Full	1.079	0.084	0.952	2.685	0.374	2.149
GCV	0.190	0.031	0.144	0.605	0.100	0.499
BIC	0.190	0.031	0.144	0.595	0.103	0.480

Table 8. Average numbers of the true positive and false positive for ZINB model based on the real doctor visits data.

	$\alpha = 0.10$		$\alpha = 1.00$	
	GCV	BIC	GCV	BIC
TP1	2.80	2.80	2.08	2.09
TP0	1.58	1.59	1.46	1.49
FP1	0.22	0.24	0.61	0.70
FP0	0.15	0.15	0.54	0.66

Table 9. Estimates of the dispersion parameter in ZINB model.

	$\alpha = 0.10$				$\alpha = 1.00$			
	Oracle	Full	GCV	BIC	Oracle	Full	GCV	BIC
Min	0.065	0.068	0.067	0.066	0.645	0.664	0.684	0.666
P2.5	0.066	0.070	0.070	0.067	0.785	0.808	0.857	0.824
Median	0.094	0.099	0.099	0.097	0.965	0.968	1.067	1.012
Mean	0.094	0.098	0.098	0.096	0.964	0.986	1.082	1.013
P97.5	0.122	0.126	0.127	0.124	1.227	1.223	1.407	1.305
Max	0.135	0.141	0.139	0.138	1.296	1.284	1.465	1.270

Simulation 4 shows the similar observations to those of Simulation 3. As the dispersion parameter α increases, the performances become worse, i.e. the MSE increases, the number of the true positive reduces and the number of the false positive increases. A possible reason is that a larger α tends to impose greater variation to the data, which makes the variable selection more difficult. The estimate of α in the adaptive lasso model is very close to those in the oracle and full models. However, when $\alpha = 1.00$, the confidence interval estimated via the adaptive lasso is slightly longer than those via the oracle and full models, this may reflect the uncertainty in selection of the tuning parameter λ.

5. Illustrate example

5.1 *Study description*

We use the doctor visits data, a subsample of the German Socioeconomic Panel which is available from the R package zic [19,32], to illustrate the proposed method. There are 14 candidate variables (Table 10), leading to $2^{24} (\approx 1.68 \times 10^7)$ submodels. The response variable is the number of doctor visits observed among a total of 1812 German males in 1994, including 746 (41.2%) zeros, and Figure 2 is the histogram. In this application the adaptive Lasso results of GCV and BIC are almost identical, so to save space, in the following only the results of GCV are presented.

5.2 *Results*

Figure 2 shows the histogram of the number of doctor visits with the predicted probabilities obtained by the adaptive lasso estimates. The predicted probability [23] is computed as $\Sigma P(Y = y)/n$. However, none of these models provides a perfect matching to the observed values. The predicted probability of the negative binomial is close to the observed value, but a high prediction occurs for $y = 1$. The ZINB model over-estimates the probability of zeros. The Poisson and ZIP models perform the worst.

Tables 11–13 present the estimated coefficients of these models, to save space, only the nonzero values are displayed. The Pearson's χ^2 statistic [6] is calculated as $\Sigma[y - \hat{E}(y)]^2/\hat{V}\text{ar}(y)$, χ^2 follows a chi-squared distribution with the degrees of freedom equal to the number of observations minus the effective number of parameters, provided that the means are not too small [6]. Hence when the value of χ^2 is approximately close to the degrees of freedom, then it is an implication

Table 10. Descriptions for the doctor visits data.

Variables	Mean ± sd (or frequency)	Descriptions
Docvisits	2.96 ± 5.22	Number of doctor visits in last 3 months
Age	41.65 ± 11.58	Years
Health	6.84 ± 2.19	Health satisfaction, 0 (low) - 10 (high)
Handicap	216 vs. 1596	1 if handicapped, 0 otherwise
hdegree	6.16 ± 18.49	Degree of handicap in percentage points
married	1257 vs. 555	1 if married, 0 otherwise
schooling	11.83 ± 2.49	Years of schooling
hhincome	4.52 ± 2.13	Household monthly net income (German marks/1000)
children	703 vs. 1109	1 if children under 16 in the household, 0 otherwise
self	153 vs. 1659	1 if self-employed, 0 otherwise
civil	198 vs. 1614	1 if civil servant, 0 otherwise
bluec	566 vs. 1246	1 if blue collar employee, 0 otherwise
employed	1506 vs. 306	1 if employed, 0 otherwise
public	1535 vs. 277	1 if public health insurance, 0 otherwise
addon	33 vs. 1779	1 if add-on insurance, 0 otherwise

Table 11. Estimated coefficients for the doctor visits data using the Poisson and negative binomial models.

	Poisson		Negative binomial	
	Adaptive lasso	MLE	Adaptive lasso	MLE
Intercept	2.155(0.055)	2.013 (0.068)	2.702(0.100)	2.807(0.105)
Age	0.010(0.001)	0.012 (0.001)	0	0
health	−0.250	−0.248	−0.271	−0.284
Handicap	0.178(0.024)	0.227 (10)	0	0
α	–	–	1.495(0.031)	1.587(0.081)
λ_1	26.970	–	16.061	–
χ^2	9575.824	9927.306	2079.006	1940.061
Log-likelihood	−5438.658	−5434.1539	−3720.099	−3718.840

of adequacy for the chosen model. Whereas due to the large amount of zeros in the data, χ^2 may not follow the chi-squared distribution. Thus here χ^2 is only used as a reference measurement of model assessment. Compared to their full models (data not shown), the adaptive lasso models select much less significant variables.

In Table 11, the Poisson model selects three variables (i.e. age, health, and handicap), the negative binomial model only selects one variable (i.e. health). However, the overall prediction of the negative binomial is better than that of the Poisson (Figure 2). The selected variables of the ZIP are age, health, handicap, employed, public in the Poisson component, and health, children in the zero component (Table 12). The selected variables of the ZINB are consistent with those of the ZIP except the variable employed not included in the negative binomial component of the ZINB (Table 13). The results presented in columns of MLE in Tables 11–13 are the maximum likelihood estimates obtaining by fitting on the selected variables. The differences are small. The standard errors are presented in parentheses. For the adaptive lasso, the standard errors are calculated via Equation (16). For the MLE, the standard errors are calculated via maximum likelihood.

To compare the four adaptive lasso count models, we perform the Vuong [40] test because these models are non-nested in the sense of containing different variables and thus having different parameter spaces, see further details of Vuong test in [23,40]. The Vuong tests have values of 10.19 ($p < .001$) for the negative binomial vs. the Poisson, 4.81 ($p < .001$) for the ZINB vs. the negative binomial, 14.13 ($p < .001$) for the ZIP vs. the Poisson, and 5.94 ($p < .001$) for the ZINB vs. the ZIP, respectively. The asymptotic distribution of Vuong test statistic follows the

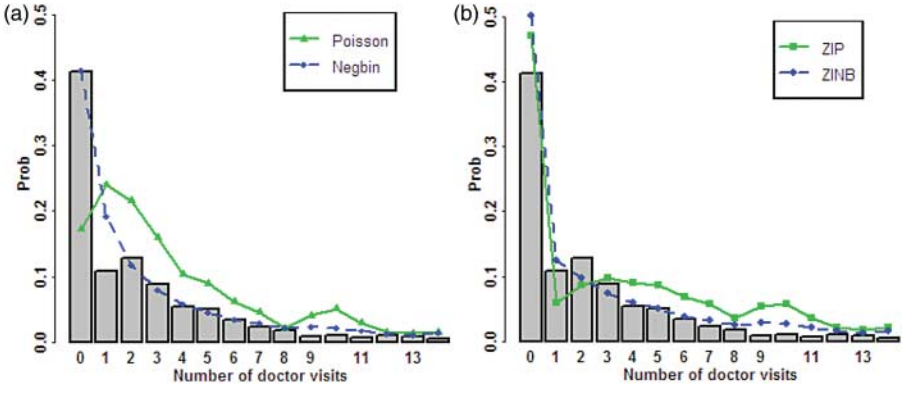

Figure 2. The histogram of the number of doctor visits with the predicted probabilities obtained by the adaptive lasso estimates. The left panel (a) is for the Poisson and negative binomial models, and the right panel (b) is for the ZIP and ZINB models.

Table 12. Estimated coefficients for the doctor visits data using the ZIP model.

	Adaptive lasso		MLE	
	Poisson	Zero	Poisson	Zero
Intercept	2.235(0.060)	−2.440	2.045(0.093)	−2.787
Age	0.006(0.001)	0	0.007(0.001)	0
Health	−0.167	0.281(0.003)	−0.165	0.311(0.028)
Handicap	0.118(0.027)	0	0.106(0.039)	0
Children	0	0.142(0.004)	0	0.413(0.107)
Employed	−0.044	0	−0.096	0
Public	0.057(9)	0	0.248(0.049)	0

Note: For the adaptive lasso, $\lambda_1 = 10.606$, $\chi^2 = 4083.218$, and the log-likelihood value is -4499.285. For the MLE, $\chi^2 = 4192.805$ and the log-likelihood value is -4496.992. The standard errors are in parentheses.

Table 13. Estimated coefficients for the doctor visits data using the ZINB model.

	Adaptive lasso		MLE	
	negbin	Zero	negbin	Zero
intercept	2.217(0.125)	−3.126	2.091 (0.158)	−3.931
Age	0.006(0.002)	0	0.009 (0.003)	0
Health	−0.188	0.296 (0.003)	−0.190	0.355 (0.050)
Handicap	0.194(0.056)	0	0.265 (0.089)	0
Children	0	0.368 (0.025)	0	0.648 (0.174)
Self	−0.104	0	−0.349	0
α	0.865(0.061)		0.815 (0.076)	

Note: Here negbin indicates the negative binomial component of ZINB model. For the adaptive lasso, $\lambda_1 = 4.636$, $\chi^2 = 1905.565$, and the log-likelihood value is −3670.249. For the MLE, $\chi^2 = 1987.937$ and the log-likelihood value is −3659.411. The standard errors are in parentheses.

standard normal under some regularity conditions. The first model is favored if the Vuong statistic is large enough, say 2; the second model is favored if the Vuong statistic is small enough, say -2; and otherwise neither model is preferred. Consequently, the ZINB model is favored even when multiple comparisons are taken into account. The Pearson's χ^2 statistic for the ZINB model is 1905.565, much less than those of other models and approximately equal to the degrees of the freedom 1805.

To gain further insight into the fittings of these models, we perform analyses of k-fold CV [9]. We compute the mean log-likelihood value for each model on the testing data. This value was called the predictive log-likelihood by Khalili and Chen [20]. A large predictive log-likelihood value indicates a better fitting. For stability, we repeat the k-fold CV analyses 10 times and average the results. The predictive log-likelihood values are given in Table 14. It is clear that the adaptive lasso ZINB outperforms the others.

The ZINB model behaves relatively well according to the results of the Vuong test, the χ^2 statistic, the log-likelihood value, and the predictive log-likelihood value, although none of the four models offers an adequate fit to the doctor visits data in terms of the prediction probability. So we give the explanations based on the results of the ZINB model. Four variables (i.e. age, health, handicap, and self) are included in the negative binomial component of the ZINB. The variables age and handicap have positive coefficients, indicating that the older or the handicapped individuals visit a doctor more frequently. This is reasonable because these people have much more healthy requirements. While the variables health and self have negative coefficients, indicating

Table 14. Predictive log-likelihood values.

	GCV				BIC			
k	Poisson	negbin	ZIP	ZINB	Poisson	negbin	ZIP	ZINB
10	−551.1	−371.2	−458.2	−368.1	−551.1	−372.4	−457.3	−369.0
20	−275.6	−185.4	−229.2	−184.2	−275.4	−186.2	−228.7	−184.4

Note: Here negbin indicates the negative binomial model.

that, on average, individuals who have a higher health satisfaction or are self-employed would have few number of doctor visits. It is natural that a subject with satisfied health does not see a doctor as frequently as the one with less satisfied health, and that the subject who is self-employed does not often go to hospital partly due to being busy. Besides health, the variable children with a positive coefficient is also included in the zero component, showing that the individual would tend not to go to hospital if he has children under 16, perhaps that he needs to look after his children and consequently has less spare time is one of the main causes. Here, the excess zeros in the doctor visits data may come from the healthy individuals, or from the sick ones but not visiting a doctor.

To verify the adequacy of these models, we perform residual analyses. The standardized residual is calculated as $[y - \hat{E}(y)]/\sqrt{\hat{V}\text{ar}(y)}$. For all the four models, the plots of residual against the number of doctor visits shows that the residuals have obvious trends along the responses (Figures not shown), suggesting possible missing of some important explanatory variables. For instance, the distance from home to hospital and available health resources can have important influences on visiting a doctor or not. Further work to discover the reasons of deficiency will be warranted.

6. Discussion

This article has discussed the adaptive lasso approach for the zero-inflated count data. The Taylor approximation algorithm is straightforward to deal with the complex penalized log-likelihood function and also computationally efficient as long as the Hessian matrix can be well estimated. The simulation results have shown that the adaptive lasso works well to identify the important variables.

More recently, another popular penalization-based method, the smoothly clipped absolute deviation (SCAD) designed by Fan and Li [10], has been employed in the ZIP model [2]. Both the SCAD and the adaptive lasso enjoy the oracle property [10,48], but compared to the adaptive lasso, the SCAD is a bit computationally difficult because of its non-convex penalty function. To address this difficulty, the one-step sparse estimation [50] based on LSA is used in [2], which leads to an adaptive lasso estimator with the adaptive weights defined by the first derivative of the SCAD penalized function [50]. Compared to the work in [2], we give more extensive simulations, specifically about the applicability of the Taylor approximation algorithm; we also discuss the ZINB model. But the substantial comparison of the two methods in the zero-inflated count data will be an interesting problem.

Both the ZIP and ZINB models belong to the family of FMR models [25] for which the variable selection has drawn intensive research interest recently [20,28,36]. Instead of the Taylor approximation algorithm adopted here, the well-known expectation–maximization (EM) algorithm [5] is used to perform the penalized estimation in those studies. Although the Taylor approximation algorithm has the advantage of computational simplify and fast convergence, as shown in Simulation 1 it is numerically instable if the Hessian matrix cannot be obtained accurately [41]. The EM algorithm may provide a more natural and appropriate choice especially when the count and zero

components are poorly separated. Developing EM algorithm for the adaptive lasso zero-inflation count models is our ongoing work.

The proposed approach relies on the choice of tuning parameter [13,28,42,43,47]. Our simulation has shown that the GCV and BIC procedures perform well. For FMR models the variable selection is generally more challenging because the selections of mixture components and variables need to be conducted simultaneously [20,28,36]. The performance of variable selection for FMR models is dependent on several factors, among which the accurate estimation of the number of mixture components is crucially important [4,20,28]. In the ZIP and ZINB models, we assumed only two components existing in the data a priori. This assumption, however, may be not true in real-life data. There can be more than two components, or the mixture may have forms of Poisson–Poisson or Poisson–negative binomial. This can be one of the reasons that lead to the failure of the models employed in Section 5. Naik *et al.* [28] recently proposed a new efficient mixture regression criterion extended from AIC based on a clustering penalty function and showed that it significantly outperformed both the AIC and BIC. Exploring this new criterion for the zero-inflated count models is our future topic.

Acknowledgements

The authors thank two reviewers for their helpful comments and suggestions which substantially improve the manuscript and thanks also go to the editor for his support and encouragement. This research was supported in part by National Natural Science Foundation of China (No. 30571619, 81373102), College Philosophy and Social Science Foundation from Education Department of Jiangsu Province of China (No. 2013SJB790059, 2013SJD790032), Research Foundation from Xuzhou Medical College (No. 2012KJ02), and Research and Innovation Project for College Graduates of Jiangsu province of China (No. CXLX13_574).

References

[1] L. Breiman, *Heuristics of instability and stabilization in model selection*, Ann. Stat. 24 (1996), pp. 2350–2383.
[2] A. Buu, N. J. Johnson, R. Li, and X. Tan, *New variable selection methods for zero-inflated count data with applications to the substance abuse field*, Stat. Med. 30 (2011), pp. 2326–2340.
[3] C.A. Cameron and P.K. Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, New York, 1998.
[4] J. Chen and A. Khalili, *Order selection in finite mixture models with a nonsmooth penalty*, J. Am. Stat. Assoc. 103 (2008), pp. 1674–1683.
[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B. 39 (1977), pp. 1–38.
[6] A.J. Dobson and A.G. Barnett, *An Introduction to Generalized Linear Models*, 3rd ed., Chapman and Hall, New York, 2008.
[7] D.L. Donoho and J.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika, 81 (1994), pp. 425–455.
[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Stat. 32 (2004), pp. 407–499.
[9] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
[10] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), pp. 1348–1360.
[11] J. Fan and R. Li, *Variable selection for Cox's proportional hazards model and Frailty model*, Ann. Stat. 30 (2002), pp. 74–99.
[12] J. Fan and H. Peng, *Nonconcave penalized likelihood with diverging number of parameters*, Ann. Stat. 32 (2004), pp. 928–961.
[13] Y. Fan and C.Y. Tang, *Tuning parameter selection in high dimensional penalized likelihood*, J. R. Stat. Soc. Ser. B. 75 (2013), pp. 531–552.
[14] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, *Pathwise coordinate optimization*, Ann. Appl. Stat. 1 (2007), pp. 302–332.
[15] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, J. Stat. Softw. 33 (2010), pp. 1–22.
[16] G.H. Golub, M. Heath, and G. Wahba, *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–223.

[17] W.H. Greene, *Accounting for Excess of Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*, New York University, 1994.
[18] S. Jackman, *pscl: Political Science Computational Laboratory*, 2012. Available at http://cran.r-project.org/web/packages/pscl/index.html
[19] M. Jochmann, *zic: Bayesian Inference for Zero-Inflated Count Models*, 2012. Available at http://cran.r-project.org/web/packages/zic/index.html
[20] A. Khalili and J. Chen, *Variable selection in finite mixture of regression models*, J. Am. Stat. Assoc. 102 (2007), pp. 1025–1038.
[21] B.M.G. Kibria, K. Månsson, and G. Shukur, *Some ridge regression estimators for the zero-inflated Poisson model*, J. Appl. Stat. 40 (2012), pp. 721–735.
[22] D. Lambert, *Zero-inflated poisson regression with an application to defects in manufacturing*, Technometrics, 34 (1992), pp. 1–14.
[23] S.J. Long, *Regression Models for Categorical and Limited Dependent Variables*, Sage Publications, Thousand Oaks, CA, 1997.
[24] W. Lu and H.H. Zhang, *Variable selection for proportional odds model*, Stat. Med. 26 (2007), pp. 3771–3781.
[25] G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, 2000.
[26] N. Meinshausen, *Relaxed lasso*, Comput. Stat. Data Anal. 52 (2007), pp. 374–393.
[27] J. Mullahy, *Specification and testing of some modified count data models*, J. Econometrics, 33 (1986), pp. 341–365.
[28] P.A. Naik, P. Shi, and C.L. Tsai, *Extending the akaike information criterion to mixture regression models*, J. Am. Stat. Assoc. 102 (2007), pp. 244–254.
[29] M.R. Osborne, B. Presnell, and B.A. Turlach, *A new approach to variable selection in least squares problems*, IMA J. Numer. Anal. 20 (2000), pp. 389–403.
[30] M.R. Osborne, B. Presnell, and B.A. Turlach, *On the LASSO and its dual*, J. Comput. Graph. Stat. 9 (2000), pp. 319–337.
[31] M.Y. Park and T. Hastie, *L1 regularization path algorithm for generalized linear models*, J. R. Stat. Soc. Ser. B. 69 (2007), pp. 659–677.
[32] R Core Team, *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*, 2013; softwareavailable at http://www.R-project.org/
[33] R. Redner and H. Walker, *Mixture densities, maximum likelihood and the EM algorithm*, SIAM Rev. 26 (1984), pp. 195–239.
[34] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978), pp. 461–464.
[35] N. Simon, J.H. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for Cox's proportional hazards model via coordinate descent*, J. Stat. Softw. 39 (2011), pp. 1–13.
[36] N. Städler, P. Bühlmann, and S. van de Geer, *$\ell 1$-penalization for mixture regression models*, TEST, 19 (2010), pp. 209–256.
[37] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. Roy. Stat. Soc. Ser. B. 58 (1996), pp. 267–288.
[38] R. Tibshirani, *The lasso method for variable selection in the Cox model*, Stat. Med. 16 (1997), pp. 385–395.
[39] D. Todem, Y. Zhang, A. Ismail, and W. Sohn, *Random effects regression models for count data with excess zeros in caries research*, J. Appl. Stat. 37 (2010), pp. 1661–1679.
[40] Q.H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, Econometrica, 57 (1989), pp. 307–344.
[41] H. Wang and C. Leng, *Unified LASSO estimation by least squares approximation*, J. Am. Stat. Assoc. 102 (2007), pp. 1039–1048.
[42] H. Wang, B. Li, and C. Leng, *Shrinkage tuning parameter selection with a diverging number of parameters*, J. R. Stat. Soc. Ser. B. 71 (2009), pp. 671–683.
[43] H. Wang, R. Li, and C. L. Tsai, *Tuning parameter selectors for the smoothly clipped absolute deviation method*, Biometrika, 94 (2007), pp. 553–568.
[44] R. Winkelmann, *Econometric Analysis of Count Data*, Springer-Verlag, Berlin, 2008.
[45] K.Y. Wong and K.F. Lam, *Modeling zero-inflated count data using a covariate-dependent random effect model*, Stat. Med. 32 (2013), pp. 1283–1293.
[46] H.H. Zhang and W. Lu, *Adaptive Lasso for Cox's proportional hazards model*, Biometrika, 94 (2007), pp. 691–703.
[47] Y. Zhang, R. Li, and C.L. Tsai, *Regularization parameter selections via generalized information criterion*, J. Am. Stat. Assoc. 105 (2010), pp. 312–323.
[48] H. Zou, *The adaptive lasso and its oracle properties*, J. Am. Stat. Assoc. 101 (2006), pp. 1418–1429.
[49] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B. 67 (2005), pp. 301–320.
[50] H. Zou and R. Li, *One-step sparse estimates in nonconcave penalized likelihood models*, Ann. Stat. 36 (2008), pp. 1509–1566.