Hindawi
Journal of Probability and Statistics
Volume 2018, Article ID 2834183, 9 pages
https://doi.org/10.1155/2018/2834183

Research Article

A Note on the Adaptive LASSO for Zero-Inflated Poisson Regression

Prithish Banerjee,[1] Broti Garai,[2] Himel Mallick,[3,4] Shrabanti Chowdhury,[5] and Saptarshi Chatterjee[6]

[1] *JP Morgan Chase & Co., USA*
[2] *NBCUniversal, USA*
[3] *Department of Biostatistics, Harvard T.H. Chan School of Public Health, USA*
[4] *Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, USA*
[5] *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, USA*
[6] *Eli Lilly and Company, USA*

Correspondence should be addressed to Himel Mallick; hmallick@hsph.harvard.edu

Prithish Banerjee, Broti Garai, and Himel Mallick contributed equally to this work.

Received 23 July 2018; Accepted 21 November 2018; Published 30 December 2018

Guest Editor: Ash Abebe

We consider the problem of modelling count data with excess zeros using Zero-Inflated Poisson (ZIP) regression. Recently, various regularization methods have been developed for variable selection in ZIP models. Among these, EM LASSO is a popular method for simultaneous variable selection and parameter estimation. However, EM LASSO suffers from estimation inefficiency and selection inconsistency. To remedy these problems, we propose a set of EM adaptive LASSO methods using a variety of data-adaptive weights. We show theoretically that the new methods are able to identify the true model consistently, and the resulting estimators can be as efficient as oracle. The methods are further evaluated through extensive synthetic experiments and applied to a German health care demand dataset.

1. Introduction

Modern research studies routinely collect information on a broad array of outcomes including count measurements with excess amount of zeros. Modeling such zero-inflated count outcomes is challenging for several reasons. First, traditional count models such as Poisson and Negative Binomial are suboptimal in accounting for excess variability due to zero-inflation [1, 2]. Second, alternative zero-inflated models such as the **Z**ero-**I**nflated **P**oisson (ZIP) [2] and **Z**ero-**I**nflated **N**egative **B**inomial (ZINB) [1] models are computationally prohibitive in the presence of high-dimensional and collinear variables.

Regularization methods have been proposed as a powerful framework to mitigate these problems, which tend to exhibit significant advantages over traditional methods [3, 4]. Essentially all these methods enforce sparsity through a suitable penalty function and identify predictive features by means of a computationally efficient Expectation Maximization (EM) algorithm. Among these, EM LASSO is particularly attractive due to its capability to perform simultaneous model selection and stable effect estimation. However, recent research suggests that EM LASSO may not be fully efficient and its model selection result could be inconsistent [5, 6]. This led to a simple modification of the LASSO penalty, namely, the EM adaptive LASSO (EM AL). EM AL achieves "oracle selection consistency" by allowing different amounts of shrinkage for different regression coefficients.

Previous studies have not, however, investigated the EM AL at sufficient depth to evaluate its properties under diversified and realistic scenarios. It is not yet clear, for example, how reliable the resulting parameter estimates are in the presence of multicollinearity. In particular, the actual variable selection performance of EM AL depends on the proper construction of the data-adaptive weight vector. When the features to be associated possess an inherent collinearity, EM AL is expected to produce suboptimal results, a phenomenon that is especially evident when the sample size is limited [7]. Several remedies have been suggested for linear and generalized linear models (GLMs) such as the standard error-adjusted adaptive LASSO (SEAL) [7, 8]. However, there is a lack of similar published methods for zero-inflated count regression models. In addition, complete software packages of these methods have not been made available to the community.

We address these issues by providing a set of flexible variable selection approaches to efficiently identify correlated features associated with zero-inflated count outcomes in a ZIP regression framework. We have implemented this method as AMAZonn (**A** **M**ulticollinearity-adjusted **A**daptive LASSO for **Z**ero-inflated C**o**u**n**t Regressio**n**). AMAZonn considers two data-adaptive weights: (i) the inverse of the maximum likelihood (ML) estimates (EM AL) and (ii) inverse of the ML estimates divided by their standard errors (EM SEAL). We show theoretically that AMAZonn is able to identify the true model consistently, and the resulting estimator is as efficient as oracle. Numerical studies confirmed our theoretical findings. The rest of the article is organized as follows. The AMAZonn method is proposed in the next section, and its theoretical properties are established in Section 3. Simulation results are reported in Section 4 and one real dataset is analyzed in Section 5. Then, the article concludes with a short discussion in Section 6. All technical details are presented in the Appendix.

2. Methods

2.1. Zero-Inflated Poisson (ZIP) Model. Zero-inflated count models assume that the observations originate either from a "susceptible" population that generates zero and positive counts according to a count distribution or from a "nonsusceptible" population, which produces additional zeros [1, 2]. Thus, while a subject with a positive count is considered to belong to the "susceptible" population, individuals with zero counts may belong to either of the two latent populations. We denote the observed values of the response variable as $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$. Following Lambert [2], a ZIP mixture distribution can be written as

$$P(y_i = k) = \begin{cases} p_i + (1 - p_i) e^{-\lambda_i} & \text{if } k = 0, \\ (1 - p_i) \dfrac{e^{-\lambda_i} \lambda_i^k}{k!} & \text{if } k = 1, 2, \ldots, \end{cases} \quad (1)$$

where p_i is the probability of belonging to the nonsusceptible population and λ_i is the Poisson mean corresponding to the susceptible population for the i^{th} individual $(i = 1, \ldots, n)$. It can be seen from (1) that ZIP reduces to the standard Poisson model when $p_i = 0$. Also, $P(y_i = 0) > e^{-\lambda_i}$, indicating zero-inflation. The probability of belonging to the "nonsusceptible" population, p_i, and the Poisson mean, λ_i, are linked to the explanatory variables through the logit and log links as

$$\operatorname{logit}(p_i) = \mathbf{z}_i' \boldsymbol{\gamma} \text{ and} \quad (2)$$

Table 1: The AMAZonn data-adaptive weights. $\widehat{\beta}_{\text{ML}}$ and $\widehat{\gamma}_{\text{ML}}$ denote the ML estimates based on the unpenalized ZIP model, corresponding to count and zero submodels, respectively. SE denotes the standard errors of the corresponding ML estimates.

Weighting Scheme	Count	Zero
AMAZonn - EM AL	$\dfrac{1}{\lvert\widehat{\beta}_{j\text{ML}}\rvert}$	$\dfrac{1}{\lvert\widehat{\gamma}_{j\text{ML}}\rvert}$
AMAZonn - EM SEAL	$\dfrac{SE(\widehat{\beta}_{j\text{ML}})}{\lvert\widehat{\beta}_{j\text{ML}}\rvert}$	$\dfrac{SE(\widehat{\gamma}_{j\text{ML}})}{\lvert\widehat{\gamma}_{j\text{ML}}\rvert}$

$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta}, \quad (3)$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of covariates for the ith subject $(i = 1, \ldots, n)$ corresponding to the count and zero models, respectively, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_q)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the corresponding regression coefficients including the intercepts.

For n independent observations, the ZIP log-likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = {} & \sum_{y_i=0} \log \left\{ e^{z_i'\boldsymbol{\gamma}} + e^{-e^{x_i'\boldsymbol{\beta}}} \right\} \\ & + \sum_{y_i>0} \left\{ y_i x_i' \boldsymbol{\beta} + e^{-x_i'\boldsymbol{\beta}} \right\} - \sum_{i=1}^{n} \log \left\{ 1 + e^{z_i'\boldsymbol{\gamma}} \right\} \\ & - \sum_{y_i>0} \log(y_i!) . \end{aligned} \quad (4)$$

2.2. The AMAZonn Method. AMAZonn considers two data-adaptive weights in the EM adaptive LASSO framework: (i) the inverse of the maximum likelihood (ML) estimates (EM AL) and (ii) inverse of the ML estimates divided by their standard errors (EM SEAL). As defined by Tang et al. [6], the EM adaptive LASSO formulation for ZIP regression is given by

$$\widehat{\boldsymbol{\theta}}^* = \arg\min \{-L(\boldsymbol{\theta})\} + \nu_1 \sum_{j=1}^{p} w_{1j} \left|\beta_j\right| + \nu_2 \sum_{j=1}^{p} w_{2j} \left|\gamma_j\right|, \quad (5)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}\}$ is the parameter vector of interest with known weights $w_1 = (w_{11}, \ldots, w_{1p})'$ and $w_2 = (w_{21}, \ldots, w_{2p})'$. As noted by Qian and Yang [7], the inverse of the maximum likelihood (ML) estimates as weights may not always be stable, especially when the multicollinearity of the design matrix is a concern. In order to adjust for this instability, AMAZonn additionally considers the inverse of the ML estimates divided by their standard errors as weights. We refer to these two methods as AMAZonn - EM AL and AMAZonn - EM SEAL, respectively (Table 1).

2.3. The EM Algorithm. In order to efficiently estimate the parameters in the above optimization problem (5), we resort to the EM algorithm. To this end, we define a set of latent variables z_i as follows:

$z_i = 1$ if y_i is from the zero state, and

$$z_i = 0 \text{ if } y_i \text{ is from the count state}, \quad i = 1, \ldots, n. \tag{6}$$

We consider the latent variables z_i's as the "missing data" and rewrite the complete-data log-likelihood function in (4) as follows:

$$\begin{aligned} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \big[& z_i X_i \gamma - \log\left(1 + \exp\left(X_i \gamma\right)\right) \\ & + \left(1 - z_i\right)\left\{y_i X_i \beta - \left(y_i + 1\right) \log\left(1 + X_i \beta\right)\right\}\big]. \end{aligned} \tag{7}$$

With the above formulation, the objective function in (5) can be rewritten as

$$Q^*(\boldsymbol{\theta}) = -L(\boldsymbol{\theta}) + \nu_1 \sum_{j=1}^{p} w_{1j} \left|\beta_j\right| + \nu_2 \sum_{j=1}^{p} w_{2j} \left|\gamma_j\right|, \tag{8}$$

which can be iteratively solved as follows:

(1) At iteration t, the **E step** computes the expectation of $Q^*(\boldsymbol{\theta})$ by substituting z_i with its conditional expectation given observed data and current parameter estimates

$$\hat{z}_i^{(t)} = \begin{cases} \left(1 + \left[\dfrac{\exp\left(-X_i \hat{\boldsymbol{\gamma}}^{(t)}\right)}{1 + \exp\left(-X_i \hat{\boldsymbol{\beta}}^{(t)}\right)}\right]\right) & \text{if } y_i = 0, \\ 0 & \text{if } y_i > 0. \end{cases} \tag{9}$$

(2) In the **M step**, the expected penalized complete-data log-likelihood (5) can be minimized the with respect to $\boldsymbol{\theta}$ as

$$\begin{aligned} Q^*\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = & -2E(L\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) + \nu_1 \sum_{j=1}^{p} w_{1j} \left|\beta_j\right| \\ & + \nu_2 \sum_{j=1}^{p} w_{2j} \left|\gamma_j\right|. \end{aligned} \tag{10}$$

(3) Continue this process until convergence, $t = 1, 2, \ldots$.

It is to be noted that (10) can be further decomposed as

$$Q^*\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = Q_1^*\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)}\right) + Q_2^*\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(t)}\right), \tag{11}$$

where Q_1^* is the weighted penalized Poisson log-likelihood defined as

$$\begin{aligned} Q_1^*\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}^{(t)}\right) = -2 \Bigg[& \sum_{i=1}^{n} \left(1 - \hat{z}_i^{(t)}\right) \\ & \cdot \left\{y_i X_i \boldsymbol{\beta} - \left(y_i + 1\right) \log\left(1 + X_i \boldsymbol{\beta}\right)\right\} \Bigg] \\ & + \nu_1 \sum_{j=1}^{p} w_{1j} \left|\beta_j\right|, \end{aligned} \tag{12}$$

and Q_2^* is the penalized logistic log-likelihood defined as

$$\begin{aligned} Q_2^*\left(\boldsymbol{\gamma} \mid \boldsymbol{\theta}^{(t)}\right) = -2 & \left[\sum_{i=1}^{n} \hat{z}_i^{(t)} X_i \boldsymbol{\gamma} - \log\left(1 + \exp\left(X_i \boldsymbol{\gamma}\right)\right)\right] \\ & + \nu_2 \sum_{j=1}^{p} w_{2j} \left|\gamma_j\right|, \end{aligned} \tag{13}$$

both of which can be minimized separately using computationally efficient coordinate descent algorithms developed for GLMs [9].

2.4. Selection of Tuning Parameters. We select the tuning parameters based on the minimum BIC [10] criterion, which is known to provide better variable selection performance as compared to other information criteria [11]. This can be effortlessly incorporated in our formulation by using existing implementations for zero-inflated count models [3, 4, 6].

3. Oracle Properties

Recently, Tang et al. [6] showed that the EM adaptive LASSO (i.e., AMAZonn - EM AL) enjoys the so-called oracle properties, i.e., the estimator is able to identify the true model consistently, and the resulting estimator is as efficient as *oracle*. Here we extend these results to the AMAZonn - EM SEAL estimator and show that the AMAZonn - EM SEAL estimator also maintains the same theoretical properties. For the sake of completeness, we provide a combined general proof for both AMAZonn estimators.

Without being too rigorous mathematically, recall that the log-likelihood function for the ZIP regression model is given by

$$\begin{aligned} L\left(\boldsymbol{\theta}; \boldsymbol{v}_i\right) = & \sum_{y_i = 0} \log\left[\psi_i + \left(1 - \psi_i\right) f\left(0; \lambda_i\right)\right] \\ & + \sum_{y_i > 0} \log\left[\left(1 - \psi_i\right) f\left(y_i; \lambda_i\right)\right], \end{aligned} \tag{14}$$

where $\boldsymbol{v}_i$'s are the observed data (i.i.d observations from the ZIP distribution), $f(.; \lambda_i)$ is the probability mass function of Poisson distribution with parameter $\lambda_i = \exp(X_i \boldsymbol{\beta})$ and $\psi_i = \exp(X_i \boldsymbol{\gamma})/(1 + \exp(X_i \boldsymbol{\gamma})), i = 1, \ldots, n$. The corresponding penalized log-likelihood is given by

$$Q(\boldsymbol{\theta}) = -L\left(\boldsymbol{\theta}; \boldsymbol{v}_i\right) + \nu_{1n} \sum_{j=1}^{p} w_{1j} \left|\beta_j\right| + \nu_{2n} \sum_{j=1}^{p} w_{2j} \left|\gamma_j\right|. \tag{15}$$

Let us denote the true coefficient vector as $\boldsymbol{\theta_0} = (\boldsymbol{\beta_0}^T, \boldsymbol{\gamma_0}^T)^T$. Decompose $\boldsymbol{\theta_0} = (\boldsymbol{\theta_{10}}^T, \boldsymbol{\theta_{20}}^T)^T$ and assume that $\boldsymbol{\theta}_{20}^T$ contains all zero coefficients. Let us denote the subset of true nonzero coefficients as $\mathscr{A} = \{j : \theta_{j0} \neq 0\}$ and the subset of selected nonzero coefficients as $\widehat{\mathscr{A}} = \{j : \hat{\theta}_j \neq 0\}$. With this formulation, the Fisher information matrix can be written as

$$I\left(\boldsymbol{\theta_0}\right) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}, \tag{16}$$

where I_{11} is the Fisher information corresponding the true nonzero submodel. The oracle property of AMAZonn may be developed based on certain mild regularity conditions which are as follows:

(**A1**): The Fisher information matrix $I(\boldsymbol{\theta})$ is finite and positive definite for all values of $\boldsymbol{\theta}$.

(**A2**): There exists functions G_{jkl} such that

$$\frac{\partial^3 L(\boldsymbol{\theta}; \boldsymbol{v}_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \leq G_{jkl}(\boldsymbol{v}_i) \quad \forall\boldsymbol{\theta}, \tag{17}$$

where $g_{jkl} = E_{\boldsymbol{\theta}_0}(G_{jkl}(\boldsymbol{v}_i)) < \infty$ for all j, k, l.

Theorem 1. *Under (A1) and (A2), if $\nu_{1n} \longrightarrow \infty$, $\nu_{2n} \longrightarrow \infty$, $\nu_{1n}/\sqrt{n} \longrightarrow 0$, $\nu_{2n}/\sqrt{n} \longrightarrow 0$, then the AMAZonn estimators obey the following oracle properties:*

(1) *consistency in variable selection: $\lim_n P(\widehat{\mathscr{A}} = \mathscr{A}) = 1$, and*

(2) *asymptotic normality of the nonzero coefficients: $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow_d \mathscr{N}(\mathbf{0}, I_{11}^{-1})$.*

4. Simulation Studies

In this section, we conduct simulation studies to evaluate the finite sample performance of AMAZonn. For comparison purposes, the performance of both AMAZonn and EM LASSO is evaluated. For each simulated dataset, the associated tuning parameters are selected by the minimum BIC criterion for all the methods under consideration. All the examples reported in this section are obtained from published papers with slight modifications within the scope of the current study [11, 12].

Specially, three scenarios are considered: in the data generating models of Simulations 1 and 2, we consider all continuous predictors, whereas in Simulation 3, both continuous and categorical variables are included. For each experimental instance, we randomly partition the data into training and test sets: models are fitted on the training set and prediction errors based on mean absolute scaled error (MASE) are calculated on the held-out samples in the test set. For an exhaustive comparison, we considered three sets of sample sizes $\{n_T, n_P\} = \{200, 200\}, \{500, 500\}$, and $\{1000, 1000\}$, where n_T and n_P represent the size of the training and test data, respectively. The corresponding regression coefficients and intercepts are chosen so that a desired level of sparsity proportion (ϕ) is achieved. In order to remain as model-agnostic as possible, we consider the same set of predictors for both zero and count submodels (i.e., $\mathbf{X} = \mathbf{Z}$). Such models are common in many practical applications where no domain-specific prior information about the zero-inflation mechanism is available. Below we provide the detailed data generation steps for both simulation examples:

Simulation 1.

(1) Generate 40 predictors from the multivariate normal distribution with mean vector $\mathbf{0}$, variance vector $\mathbf{1}$, and variance-covariance matrix V, where the elements of V are $\rho^{|j_1 - j_2|}$ $\forall j_1 \neq j_2 = 1, \ldots, 40$. The values of pairwise correlation ρ varies from 0 (uncorrelated) to 0.4 (moderate collinearity) to 0.8 (high collinearity).

(2) The count and zero regression parameters are chosen as follows:

$$\begin{aligned}
&(\beta_1, \ldots, \beta_8) \\
&\quad = (-1, -0.5, -0.25, -0.1, 0.1, 0.25, 0.5, 0.75)', \\
&(\beta_9, \ldots, \beta_{16}) = (0.2, \ldots, 0.2)', \\
&(\beta_{17}, \ldots, \beta_{40}) = (0, \ldots, 0)', \\
&(\gamma_1, \ldots, \gamma_8) \\
&\quad = (-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4)', \\
&(\gamma_9, \ldots, \gamma_{16}) = (0.2, \ldots, 0.2)', \\
&(\gamma_{17}, \ldots, \gamma_{40}) = (0, \ldots, 0)'.
\end{aligned} \tag{18}$$

(3) The zero-inflated count outcome y is simulated according to (1) with the above parameters and input data.

Simulation 2. It is similar to Simulation 1 except that the count and zero regression parameters are chosen as follows:

$$\begin{aligned}
&(\beta_1, \ldots, \beta_{10}) = (0.05, -0.25, 0.05, 0.25, \\
&\quad - 0.15, 0.15, 0.25, -0.2, 0.25, -0.25)', \\
&(\beta_{11}, \ldots, \beta_{30}) = (-0.2, 0.25, 0.15, \\
&\quad - 0.25, 0.2, 0, \ldots, 0)', \\
&(\beta_{31}, \ldots, \beta_{40}) = (0.27, -0.27, 0.14, 0.2, \\
&\quad - 0.2, 0.2, 0, \ldots, 0)', \\
&(\gamma_1, \ldots, \gamma_{10}) = (-0.5, -0.4, -0.3, -0.2, \\
&\quad - 0.1, 0.1, 0.2, 0.3, 0.4, 0.5)', \\
&(\gamma_{11}, \ldots, \gamma_{30}) = (-0.2, 0.25, 0.15, -0.25, 0.2, 0, \ldots, 0)', \\
&(\gamma_{31}, \ldots, \gamma_{40}) = (0.27, -0.27, -0.14, -0.2, \\
&\quad - 0.2, 0.2, 0, \ldots, 0)'.
\end{aligned} \tag{19}$$

Simulation 3.

(1) First simulate $X_1, \ldots, X_6$ independently from the standard normal distribution. Consider the following as the continuous predictors: $\{X_1\}, \{X_2\}, \{X_3, X_3^2, X_3^3\}, \{X_4\}, \{X_5\}$ and $\{X_6, X_6^2, X_6^3\}$.

(2) Simulate 5 continuous variables from the multivariate normal distribution with mean 0, variance 1, and AR(ρ) correlation structure for varying ρ in $\{0, 0.4,$

Table 2: Results of Simulations 1–3. Average (over 200 replications) of Mean Absolute Scale Errors (MASEs) of AMAZonn and EM LASSO is reported.

			Simulation 1			Simulation 2			Simulation 3		
ρ	ϕ	n	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO	AMAZonn - EM SEAL	AMAZonn - EM AL	EM LASSO
0.0	0.3	200	0.91	0.92	0.91	0.60	0.61	0.62	0.97	1.03	1.00
		500	0.90	0.90	0.91	0.60	0.60	0.61	0.97	0.99	1.00
		1000	0.91	0.91	0.92	0.58	0.58	0.60	0.97	0.98	0.98
	0.4	200	1.12	1.13	1.12	0.75	0.75	0.76	1.18	1.23	1.23
		500	1.05	1.05	1.06	0.73	0.73	0.74	1.11	1.17	1.20
		1000	1.03	1.03	1.04	0.71	0.71	0.72	1.11	1.16	1.16
	0.5	200	1.28	1.28	1.27	0.87	0.87	0.87	1.40	1.46	1.46
		500	1.16	1.16	1.17	0.84	0.84	0.85	1.28	1.33	1.36
		1000	1.15	1.15	1.19	0.80	0.80	0.82	1.23	1.30	1.31
0.4	0.3	200	1.05	1.06	1.09	0.63	0.63	0.63	0.96	1.01	0.99
		500	1.04	1.04	1.05	0.61	0.61	0.62	0.95	0.97	0.99
		1000	0.96	0.96	0.98	0.58	0.58	0.59	0.97	0.98	0.98
	0.4	200	1.21	1.22	1.22	0.75	0.75	0.76	1.19	1.22	1.23
		500	1.18	1.18	1.21	0.71	0.71	0.72	1.14	1.19	1.22
		1000	1.13	1.14	1.18	0.68	0.68	0.70	1.13	1.18	1.17
	0.5	200	1.42	1.43	1.42	0.83	0.84	0.83	1.34	1.40	1.43
		500	1.26	1.26	1.32	0.80	0.81	0.82	1.27	1.32	1.35
		1000	1.23	1.23	1.30	0.75	0.75	0.77	1.27	1.34	1.33
0.8	0.3	200	1.32	1.31	1.36	0.62	0.63	0.63	0.96	1.00	1.01
		500	1.13	1.13	1.23	0.59	0.59	0.61	0.97	0.99	1.01
		1000	1.13	1.13	1.21	0.56	0.56	0.58	0.95	0.96	0.96
	0.4	200	1.52	1.52	1.58	0.71	0.72	0.72	1.18	1.21	1.23
		500	1.31	1.32	1.45	0.68	0.68	0.69	1.12	1.19	1.20
		1000	1.24	1.24	1.37	0.64	0.64	0.64	1.12	1.17	1.16
	0.5	200	1.56	1.58	1.61	0.78	0.78	0.78	1.37	1.42	1.44
		500	1.44	1.45	1.65	0.73	0.73	0.76	1.29	1.34	1.39
		1000	1.33	1.36	1.52	0.69	0.70	0.69	1.26	1.33	1.34

0.8} as before, and quantile-discretize each of them into 5 new variables based on their quantiles: $(-\infty, \Phi^{-1}(1/5)]$, $(\Phi^{-1}(1/5), \Phi^{-1}(2/5)]$, $(\Phi^{-1}(2/5), \Phi^{-1}(3/5)]$, $(\Phi^{-1}(3/5), \Phi^{-1}(4/5)]$, and $(\Phi^{-1}(4/5), \infty)$, leading to a total of 20 categorical variables.

(3) With the above input data and parameters, the zero-inflated count outcome y is simulated according to (1), where the two sets of regression parameters are chosen as follows:

$$\begin{aligned}(\beta_1, \ldots, \beta_{10}) &= \left(0, 0, 0.1, 0.2, 0.1, 0, 0, \frac{2}{3}, -1, \frac{1}{3}\right),\\ (\beta_{11}, \ldots, \beta_{30}) &= (-2, -1, 1, 2, 0, \ldots, 0),\\ (\gamma_1, \ldots, \gamma_{10}) &= \left(0, 0, 0.1, 0.2, 0.1, 0, 0, \frac{2}{3}, -1, \frac{1}{3}\right),\\ (\gamma_{11}, \ldots, \gamma_{30}) &= (-2, -1, 1, 2, 0, \ldots, 0).\end{aligned} \tag{20}$$

The resulting performance measures iterated over 200 replications (Table 2) reveal that AMAZonn performs as well as or better than EM LASSO in most of the simulation scenarios. For highly collinear designs, AMAZonn - EM SEAL stands out to be the best estimator for almost every sample size and zero-inflation proportion, highlighting the benefit of incorporating data-adaptive weights based on both ML estimates and their standard errors. This phenomenon is also apparent in the analysis of German health care data in Section 5, where the parameter estimates from the AMAZonn - EM SEAL method appear to be more parsimonious than those from other methods.

5. Application to German Health Care Demand Data

Next, we apply our method to the German health care demand data [3], a subset of the German Socioeconomic Panel (GSOEP) dataset [13], which has also been used for

Table 3: Summary of predictors in German health care demand data.

Variables	Mean (sd) or Frequency	Description
health	6.84 (2.19)	health satisfaction: 0 (low) - 10 (high)
handicap	216 / 1596	1 : handicap, 0 : otherwise
hdegree	6.16 (18.49)	degree of handicap in percentage points
married	1257 / 555	1 : married, 0 : otherwise
schooling	11.83 (2.49)	years of schooling
hhincome	4.52 (2.13)	household income per month in German marks/1000
children	703 / 1109	1 : children under 16 in household, 0 : otherwise
self	153 / 1659	1 : self-employed, 0 : otherwise
civil	198 / 1614	1 : civil servant, 0 : otherwise
bluec	566 / 1246	1 : blue collar employee, 0 : otherwise
employed	1506 / 306	1 : employed, 0 : otherwise
public	1535 / 277	1 : public health insurance, 0 : otherwise
addon	33 / 1779	1 : addon insurance, 0 : otherwise
age30	1480 / 332	1 if age $\geq$ 30
age35	1176 / 636	1 if age $\geq$ 35
age40	919 / 893	1 if age $\geq$ 40
age45	716 / 1096	1 if age $\geq$ 45
age50	535 / 1227	1 if age $\geq$ 50
age55	351 / 1461	1 if age $\geq$ 55
age60	147 / 1665	1 if age $\geq$ 60

Table 4: Model selection performance of EM LASSO and AMAZonn on German health care data.

Methods	BIC	Time (in seconds)
EM LASSO	9062.744	50.252
AMAZonn - EM AL	9002.487	26.215
AMAZonn - EM SEAL	**8982.924**	26.528

illustration purposes in previous studies [3, 14]. The original data contains number of doctor office visits for 1, 812 West German men aged 25 to 65 years in the last three months of 1994 (response variable of interest), which is supplemented with complementary information on twelve annual waves from 1984 to 1995 including health care utilization, current employment status, and insurance arrangements under which subjects are protected [3]. The goal of the original study was to investigate how the employment characteristics of the German nationals are related to their health care demand. The distribution of the dependent variable (Figure 1) reveals that many doctor visits are zeros (41.2%), confirming that classical methods such as Poisson regression are inappropriate for modeling this outcome.

In the model fitting process, along with the original variables, the interactions between age groups and health condition are also considered, resulting in 28 candidate predictors (Table 3). The fitting results from the full models indicate that both EM adaptive LASSO methods provide competitive model selection performance (Table 4), often leading to sparser model selection than EM LASSO (Table 5). In addition, the AMAZonn - EM SEAL method appears to choose even fewer numbers of variables. Such feature of AMAZonn - EM SEAL can be appealing in many practical situations, where data collinearity between variables is a concern and a more aggressive feature selection is desired. While the computational overheads of both EM adaptive LASSO methods are similar, they are an order of magnitude faster than EM LASSO (Table 4), further confirming that AMAZonn offers a viable alternative to existing methods.

6. Discussion

In recent years, there has been a huge influx of zero-inflated count measurements spanning several disciplines including biology, public health, and medicine. This has motivated the widespread use of zero-inflated count models in many practical applications such as metagenomics, single-cell RNA sequencing, and health care research. In this article, we propose the AMAZonn method for adaptive variable selection in ZIP regression models. Both our simulation and real data experience suggest that AMAZonn can outperform EM LASSO under a variety of regression settings while maintaining the desired theoretical properties and computational convenience. Our preliminary results are rather encouraging, and for practical purposes, we provide a publicly available R package implementing this method: https://github.com/himelmallick/AMAZonn.

We envision a number of improvements that may further refine AMAZonn's performance. While AMAZonn relies on ML estimates to construct the weight vector, these estimates may not be available in ultrahigh dimensions [7]. Alternative initialization schemes could further improve on this such as the ridge estimates [15]. Extension to other zero-inflated models such as marginalized zero-inflated count regression [16, 17], two-part and hurdle models [18], and multiple-inflation models [19] can form a useful basis for further

Table 5: Estimated coefficients from the best-fitting ZIP models in German health care demand data analysis.

(a)

Methods								Count Coefficients						
	(Intercept)	hlth	handicap	ddegree	married	schooling	hhincome	children	self	civil	bluec	employed	public	addon
EM LASSO	2.322	-0.14	0.207	-0.002	-0.97	0.0	0.0	0.078	-0.178	-0.166	0.038	-0.106	0.089	0.205
AMAZonn - EM AL	2.305	-0.135	0.111	0.0	-0.947	0.0	0.0	0.079	-0.234	-0.245	0.0	-0.059	0.043	0.205
AMAZonn - EM SEAL	2.378	-0.142	0.098	0.0	-0.066	0.0	0.0	0.046	-0.189	-0.222	0.0	-0.055	0.0	0.14

Methods								Count Coefficients						
	ag30	ag35	ag40	ag45	ag50	ag55	ag60	ag30:hlth	ag35:hlth	ag40:hlth	ag45:hlth	ag50:hlth	ag55:hlth	ag60:hlth
EM LASSO	0.0	0.0	0.0	0.586	0.0	-0.27	0.081	0.0	0.0	-0.006	-0.076	0.006	0.082	-0.034
AMAZonn - EM AL	0.0	0.0	-0.047	0.769	0.0	-0.402	0.099	0.0	0.0	0.0	-0.101	0.0	0.106	-0.034
AMAZonn - EM SEAL	0.0	0.0	0.0	0.586	0.0	-0.25	0.0	0.0	0.0	0.0	-0.081	0.0	0.081	-0.017

(b)

Methods								Zero Coefficients						
	(Intercept)	hlth	handicap	ddegree	married	schooling	hhincome	children	self	civil	bluec	employed	public	addon
EM LASSO	-2.193	-0.262	-0.098	-0.003	-0.121	0.0	-0.012	0.253	0.112	0.134	0.0	0.0	-0.012	0.0
AMAZonn - EM AL	-2.226	-0.261	-0.162	0.0	0.0	0.0	0.0	0.163	0.0	0.0	0.0	0.0	0.0	0.0
AMAZonn - EM SEAL	-2.403	-0.283	0.0	0.0	-0.053	0.0	0.0	0.238	0.0	0.0	0.0	0.0	0.0	0.0

Methods								Zero Coefficients						
	ag30	ag35	ag40	ag45	ag50	ag55	ag60	ag30:hlth	ag35:hlth	ag40:hlth	ag45:hlth	ag50:hlth	ag55:hlth	ag60:hlth
EM LASSO	0.0	0.0	0.0	0.0	-0.459	0.0	-0.217	0.013	0.0	0.005	0.0	0.0	0.023	0.0
AMAZonn - EM AL	0.047	0.0	0.065	0.009	-0.527	0.0	-0.198	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AMAZonn - EM SEAL	0.0	0.0	0.0	0.0	-0.443	0.0	0.0	0.009	0.0	0.0	0.0	0.0	0.0	0.0

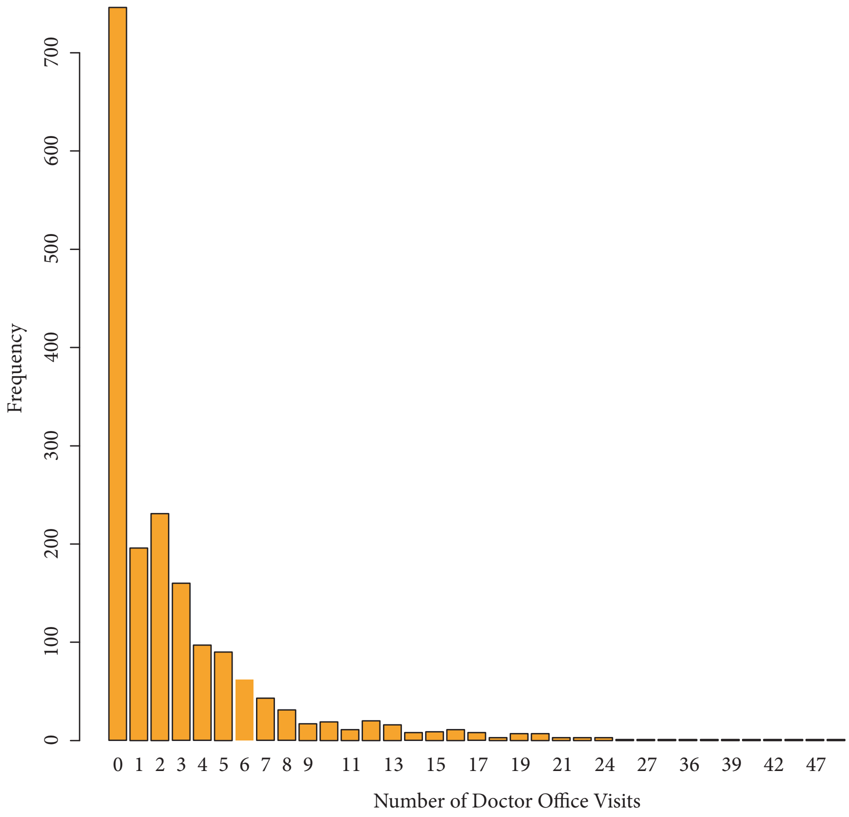


FIGURE 1: Number of doctor office visits in the German health care data.

investigations. Although we only focused on variable selection for fixed effects models, future work could include an extension to other regularization problems such as grouped variable selection [12, 20] as well as sparse mixed effects models [21].

Appendix

Proof. It is to be noted that both logistic and Poisson distributions belong to the exponential family. Since the objective function in (10) can be decomposed into weighted logistic and Poisson log-likelihoods (each belonging to the GLM family without the penalties), Theorem 1 is the direct application of Theorem 4 in Zou [22]. Therefore, if $\nu_{1n} \longrightarrow \infty$, $\nu_{2n} \longrightarrow \infty$, $\nu_{1n}/\sqrt{n} \longrightarrow 0$, and $\nu_{2n}/\sqrt{n} \longrightarrow 0$, then both the AMAZonn - EM AL and AMAZonn - EM SEAL estimators hold the oracle properties: with probability tending to 1, the estimate of zero coefficients is 0, and the estimate for nonzero coefficients has an asymptotic normal distribution with mean being the true value and variance which approximately equals the submatrix of the Fisher information matrix containing nonzero coefficients. Hence the proof is complete. □

Data Availability

The German Healthcare dataset used in the paper is publicly available from others (https://cran.r-project.org/web/packages/HDtweedie/index.html) and the software is publicly available at https://github.com/himelmallick/AMAZonn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Prithish Banerjee, Broti Garai, and Himel Mallick contributed equally to this work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript. This work was supported in part by the research computing resources acquired and managed by University of Alabama at Birmingham IT Research Computing. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Alabama at Birmingham.

References

[1] W. H. Greene, *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*, New York University, New York, NY, 1994.

[2] D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[3] Z. Wang, S. Ma, and C.-Y. Wang, “Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany,” *Biometrical Journal*, vol. 57, no. 5, pp. 867–884, 2015.

[4] Z. Wang, S. Ma, C.-Y. Wang, M. Zappitelli, P. Devarajan, and C. Parikh, “EM for regularized zero-inflated regression models with applications to postoperative morbidity after cardiac surgery in children,” *Statistics in Medicine*, vol. 33, no. 29, pp. 5192–5208, 2014.

[5] H. Mallick and H. K. Tiwari, “EM adaptive LASSO-a multilocus modeling strategy for detecting SNPs associated with zero-inflated count phenotypes,” *Frontiers in Genetics*, vol. 7, 2016.

[6] Y. Tang, L. Xiang, and Z. Zhu, “Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models,” *Risk Analysis*, vol. 34, no. 6, pp. 1112–1127, 2014.

[7] W. Qian and Y. Yang, “Model selection via standard error adjusted adaptive lasso,” *Annals of the Institute of Statistical Mathematics*, vol. 65, no. 2, pp. 295–318, 2013.

[8] Z. Y. Algamal and M. H. Lee, “Adjusted Adaptive LASSO in High-dimensional Poisson Regression Model,” *Modern Applied Science (MAS)*, vol. 9, no. 4, 2014.

[9] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software* , vol. 33, no. 1, pp. 1–22, 2010.

[10] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, “A group bridge approach for variable selection,” *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

[12] S. Chatterjee, S. Chowdhury, H. Mallick, P. Banerjee, and B. Garai, “Group regularization for zero-inflated negative binomial regression models with an application to health care demand in Germany,” *Statistics in Medicine*, vol. 37, no. 20, pp. 3012–3026, 2018.

[13] R. T. Riphahn, A. Wambach, and A. Million, “Incentive effects in the demand for health care: A bivariate panel count data estimation,” *Journal of Applied Econometrics*, vol. 18, no. 4, pp. 387–405, 2003.

[14] M. Jochmann, “What belongs where? Variable selection for zero-inflated count models with an application to the demand for health care,” *Computational Statistics*, vol. 28, no. 5, pp. 1947–1964, 2013.

[15] A. E. Hoerl and R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[16] D. L. Long, J. S. Preisser, A. H. Herring, and C. E. Golin, “A marginalized zero-inflated Poisson regression model with overall exposure effects,” *Statistics in Medicine*, vol. 33, no. 29, pp. 5151–5165, 2014.

[17] V. A. Smith and J. S. Preisser, “Direct and flexible marginal inference for semicontinuous data,” *Statistical Methods in Medical Research*, vol. 26, no. 6, pp. 2962–2965, 2016.

[18] V. A. Smith, B. Neelon, J. S. Preisser, and M. L. Maciejewski, “A marginalized two-part model for longitudinal semicontinuous data,” *Statistical Methods in Medical Research*, vol. 26, no. 4, pp. 1949–1968, 2017.

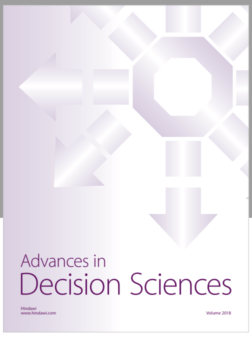
[19] X. Su, J. Fan, R. A. Levine, X. Tan, and A. Tripathi, “Multiple-inflation Poisson model with L_1 regularization,” *Statistica Sinica*, vol. 23, no. 3, pp. 1071–1090, 2013.

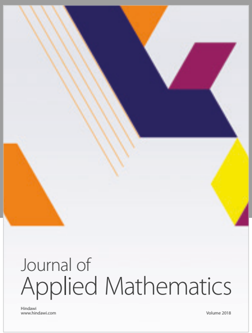
[20] S. Chowdhury, S. Chatterjee, H. Mallick, H. Banerjee, and B. Garai, “Group regularization for zero-inflated poisson regression models with an application to insurance ratemaking,” *Journal of Applied Statistics*, 2018, In Press.

[21] A. Groll and G. Tutz, “Variable selection for generalized linear mixed models by L_1-penalized estimation,” *Statistics and Computing*, vol. 24, no. 2, pp. 137–154, 2014.

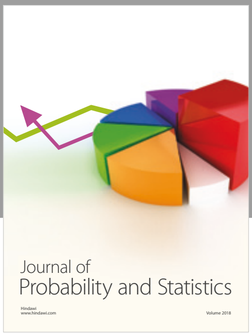
[22] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

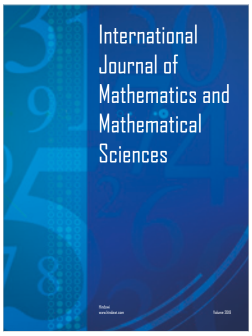
Advances in
Operations Research


Advances in
Decision Sciences

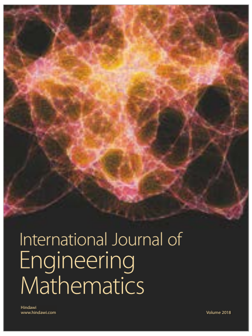

Journal of
Applied Mathematics

The Scientific
World Journal

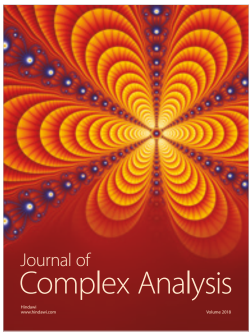

Journal of
Probability and Statistics

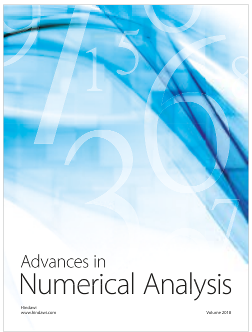

International
Journal of
Mathematics and
Mathematical
Sciences

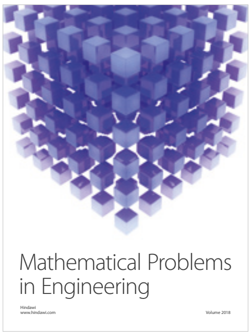
Journal of
Optimization

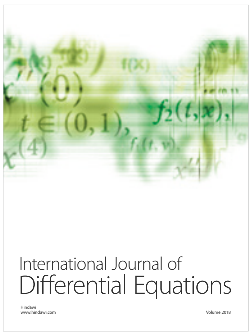

International Journal of
Engineering
Mathematics

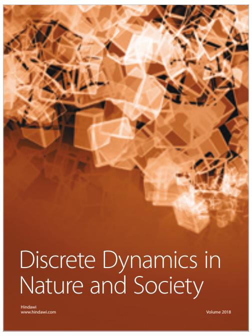
International Journal of
Analysis

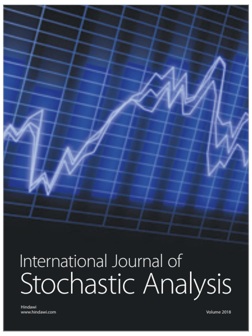

Journal of
Complex Analysis


Advances in
Numerical Analysis


Mathematical Problems
in Engineering

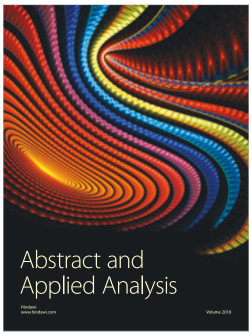

International Journal of
Differential Equations

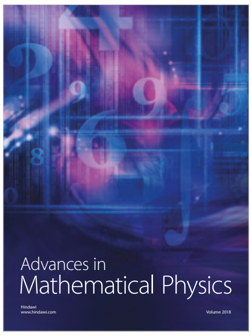

Discrete Dynamics in
Nature and Society


International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces


Abstract and
Applied Analysis


Advances in
Mathematical Physics