Computational Statistics and Data Analysis 99 (2016) 12–24

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Identification of proportionality structure with two-part models using penalization

Kuangnan Fang [a,b], Xiaoyan Wang [c], Ben-Chang Shia [d], Shuangge Ma [a,b,*]

[a] School of Economics, Xiamen University, China
[b] Department of Biostatistics, Yale University, United States
[c] College of Finance and Statistics, Hunan University, China
[d] School of Health Care Administration, Big Data Research Center & School of Management, Taipei Medical University, China

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form 19 October 2015
Accepted 5 January 2016
Available online 13 January 2016

Keywords:
Zero-inflated data
Two-part modeling
Proportionality
Penalization

ABSTRACT

Data with a mixture distribution are commonly encountered. A special example is zero-inflated data, where a proportion of the responses takes zero values, and the rest are continuously distributed. Such data routinely arise in public health, biomedicine, and many other fields. Two-part modeling is a natural choice for zero-inflated data, where the first part of the model describes whether the responses are equal to zero, and the second part describes the continuously distributed responses. With two-part models, an interesting problem is to identify the proportionality structure of covariate effects. Such a structure can lead to more efficient estimates and also provide scientific insights into the underlying data-generating mechanisms. To identify the proportionality structure, we adopt a penalization method. Compared to the alternatives, notable advantages of this method include computational simplicity, solid statistical properties, and others. For inference, we adopt a bootstrap approach. The proposed method shows satisfactory performance in simulation and the analysis of two public health datasets.

1. Introduction

Data with a mixture distribution are commonly encountered. A special example is zero-inflated data. With such data, a proportion of the response values is equal to zero, and the rest of the responses have a continuous distribution. A motivating example is the household medical expenditure dataset analyzed in Section 4.1. This dataset was generated by the China Health and Nutrition Survey (CHNS) study, which was jointly conducted by the Carolina Population Center at the University of North Carolina at Chapel Hill, the National Institute of Nutrition and Food Safety, and the Chinese Center for Disease Control and Prevention. After standard data processing, it is observed that over 70% of the households had no medical expenditure during the study period, and the rest had continuously distributed expenditures (see Fig. 1 in the Appendix). In the literature, there are a large number of examples of zero-inflated data. See for example Cheung (2002), Agarwal et al. (2002), Deb et al. (2006), Bratti and Miranda (2011), Maruotti et al. (2015), and others.

Classic models assume that the zeros and the nonzeros in response come from the same data-generating process and may not be appropriate for data with excessive zeros. Multiple models have been developed to accommodate data withexcessive zeros. Notable examples include the hurdle models (Mullahy, 1986) which have been developed for count data with excessive zeros. Under the hurdle models, the two data-generating processes are not constrained to be the same. The basic idea

[*] Corresponding author at: Department of Biostatistics, Yale University, United States.
E-mail address: shuangge.ma@yale.edu (S. Ma).

http://dx.doi.org/10.1016/j.csda.2016.01.002
0167-9473/© 2016 Elsevier B.V. All rights reserved.

is that a Bernoulli probability governs the binary outcome of whether a count variate has a zero or a positive value. If the value is positive, the hurdle is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model. Another family contains the zero-inflated models (Lambert, 1992), where the response variable is modeled as a mixture of a Bernoulli distribution (a point mass at zero) and a Poisson distribution (or another count distribution with support on non-negative integers). For data analyzed in this study and many others, the positive part has a continuous distribution. For such data, both the hurdle models and zero-inflated models are essentially the two-part models (Han and Kronmal, 2006; Liu et al., 2012). Under the two-part models, the first part describes whether theresponses take zero values (Manning et al., 1987; Olsen and Schafer, 2001). For those responses with nonzero values, the second part of the model describes their distribution. Notable advantages of the two-part models include intuitive interpretations, weak assumptions on the data generating mechanisms, and others. It is noted that in the recent literature, more complicated data structures have been considered. An example is longitudinal data, which have both within-subject correlation and between-subject heterogeneity and need to be accommodated using random effects (Min and Agresti, 2005; Greene, 2009; Alfó and Maruotti, 2010). For the aforementioned models, multiple estimation methods have been developed, including, for example, the quasi-likelihood method (McCulloch and Searle, 2001), penalized quasi-likelihood method (Yau and Lee, 2001), Bayesian method (Ghosh et al., 2006), and others. In most of the existing studies, the focus has been on modeling and estimation, while there is insufficient attention on the structure of covariate effects.

Consider covariate effects under the two-part modeling. The sets of covariates in the two model parts usually have a large overlap and, quite often, are identical. Even though the two parts have different formats, they in fact describe highly related underlying processes: the "growth" of a response from zero to nonzero (in a similar spirit as under the hurdle models), and from a small nonzero value to a large one. It is of interest to examine whether the two covariate effects in the two model parts are partially or completely proportional, that is, the proportionality structure of covariate effects. Research on theproportionality structure can be traced back to Cragg (1971). Lambert (1992) discusses the proportionality constraint for zero-inflated Poisson regression. Han and Kronmal (2006) proposes a hypothesis testing-based approach. Liu et al. (2011) develops a forward stepwise hypothesis testing-based approach and adopts bootstrap to compute the significance level of proportionality. Model selection-based approaches have also been developed. Liu and Chan (2011) develops a model selection criterion based on the marginal likelihood, which is similar to BIC. Liu et al. (2012) adopts a likelihood cross validation-based method.

Studying the proportionality structure is useful in multiple aspects. Statistically, if partial or full proportionality holds, then the model has a smaller number of unknown parameters than the unconstrained model, which can lead to improved efficiency (smaller variation) in estimation. This has been rigorously proved in Han and Kronmal (2006). Other studies, such as Liu et al. (2011), have demonstrated this using extensive numerical studies. Practically, proportionality may provide insights into the underlying data generating mechanisms. Consider for example the scenario with partial proportionality. Then the covariates can be separated into two classes: one has proportional effects in the two model parts and governs the two underlying data generating processes (from zero to nonzero, and from a small nonzero to a large one) in a similar manner, and the other behaves differently under the two processes. This can provide insights into the interconnections among covariates and their associations with the response.

Although the existing methods for determining the proportionality structure have achieved considerable successes, they also have limitations. Specifically, the hypothesis testing-based methods are computationally intensive. In addition, as sequential testing needs to be conducted, the control of type I error is nontrivial. The stepwise methods can be unstable (sensitive to even a small change in data). The BIC-based methods are computationally expensive when there are a moderate to large number of covariates. A common limitation shared by the existing methods is that computationally they do not "scale up" well. That is, their computational cost increases fast as the number of covariates increases. Most of the existing studies have focused on methodological development, and the statistical properties have not been well established.

In this study, we propose adopting penalization to determine the proportionality structure with two-part models. Penalization has been examined in a large number of studies. The goal of this study is not to develop a new penalization technique. Rather it is to apply penalization to a new statistical problem. The adopted method has an intuitive formulation as well as multiple technical advantages. It is computationally simple and can be realized using an effective algorithm. The computational cost increases relatively slowly with the number of covariates. Hence the proposed method can be applicable to data with a large number of covariates. Unlike in many of the existing studies, we take advantage of research on the asymptotics of penalization and rigorously establish the statistical properties, providing a solid ground to the proposed method. In addition, our numerical study suggests superior empirical performance of the penalization method.

2. Identification of proportionality structure using penalization

2.1. Two-part model and proportionality structure

Motivated by the data analyzed in Section 4, we consider the following distribution for the response variable Y

$$f(y) = (1-\phi)\mathbb{1}_{(y=0)} + [\phi \times N(y;\mu,\sigma^2)]\mathbb{1}_{(y>0)}, \quad y \geq 0,\ 0 \leq \phi \leq 1. \tag{1}$$

Under this model, there is a $1-\phi$ point mass at zero. For the positive part, motivated by the histograms in the Appendix, we adopt a normal distribution with mean μ and variance σ^2.

For subject $i(=1, \ldots, n)$, let $\boldsymbol{x}_i \in R^p$ denote the vector of covariates. Denote $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}^T$ as the design matrix. To incorporate covariate effects, consider the model with

$$\begin{cases} \text{Part 1}: & g(\phi_i) = \alpha_1 + \boldsymbol{x}_i^T \boldsymbol{\beta} \\ \text{Part 2}: & y_i | y_i > 0 = \alpha_2 + \boldsymbol{x}_i^T \boldsymbol{\delta} + \varepsilon_i. \end{cases} \tag{2}$$

In Part 1 of the model, g is the known link function. Choices of the link include logit, probit, log–log, and others. In this study, we choose the commonly adopted logit link where

$$\phi_i = 1/(1 + \exp(-\alpha_1 - \boldsymbol{x}_i^T \boldsymbol{\beta})). \tag{3}$$

α_1 is the intercept, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients. In Part 2 of the model, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T$ is the vector of regression coefficients, α_2 is the intercept, and ε_i is the random error with mean zero and variance σ^2. Here for simplicity of notation, we assume the same set of covariates in both parts. The proposed model and method can be easily modified to accommodate partially matched covariate sets.

Consider a small example with three covariates X_1, X_2, and X_3 and the two covariate effects being $2X_1 + X_2 + X_3$ and $6X_1 + 2X_2 + 3X_3$. Then the second covariate effect can be written as $3 \times (2X_1 + X_2 + X_3) - X_2$ as well as $2 \times (2X_1 + X_2 + X_3) + 2X_1 + X_3$. That is, as pointed out in the literature (Liu et al., 2011), there is an identifiability problem, and an additional constraint is needed. Following the notion of Liu et al. (2011) and references therein, we need to fix an anchor covariate, which has nonzero coefficients in both model parts. Without loss of generality, denote X_1 as the anchor. Published studies have provided discussions on the selection of anchor variable and suggested that it is usually not problematic in data analysis. Practical discussions are also provided in Section 3.1 of this article. Rewrite $\delta_1 = \tau\beta_1$ with $\tau \neq 0$. Under many occasions, sensible results demand $\tau > 0$. Rewrite Part 2 of the model as

$$y_i | y_i > 0 = \alpha_2 + \tau \left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) + \tilde{\boldsymbol{x}}_i^T \boldsymbol{\gamma} + \varepsilon_i, \tag{4}$$

where $\tilde{\boldsymbol{x}}_i = \left(x_{i2}, \ldots, x_{ip}\right)^T$ and $\boldsymbol{\gamma} = \left(\gamma_2, \ldots, \gamma_p\right)^T$. Denote $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \tau, \alpha_1, \alpha_2, \sigma^2, \boldsymbol{\gamma}^T\right)^T$. The log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\theta}) = & -\frac{n_1}{2} \log \sigma^2 - \frac{n_1}{2} \log 2\pi - \sum_i \log\left[1 + \exp\left(\alpha_1 + \boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right] + \sum_{i:y_i>0} \alpha_1 \\ & + \sum_{i:y_i>0} \boldsymbol{x}_i^T \boldsymbol{\beta} - \frac{1}{2\sigma^2} \sum_{i:y_i>0} \left(y_i - \alpha_2 - \tau \boldsymbol{x}_i^T \boldsymbol{\beta} - \tilde{\boldsymbol{x}}_i^T \boldsymbol{\gamma}\right)^2, \end{aligned} \tag{5}$$

where $n_1 = \#\{y_i > 0\}$.

Ignore the intercept terms. Following the definition of Liu et al. (2012) and others, if $\boldsymbol{\gamma} = \mathbf{0}$, then the two covariate effects differ by only a scale constant, that is, they are *fully proportional*. If some components of $\boldsymbol{\gamma}$ are zero, then the two covariate effects are *partially proportional*. Thus determining the proportionality structure amounts to determining which components of $\boldsymbol{\gamma}$ are zero, and it can be formulated as a model (variable) selection problem.

2.2. Penalized estimation

To determine the proportionality structure, we apply penalization. Consider the penalized objective function

$$Q(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - P(\lambda, |\boldsymbol{\gamma}|), \tag{6}$$

where $P(\lambda, |\boldsymbol{\gamma}|) = n\sum_j p\left(\lambda, |\gamma_j|\right)$ is the penalty function, and $\lambda > 0$ is the data-dependent tuning parameter. For penalty, we adopt the MCP (minimax concave penalty) which takes the form

$$p_a(\lambda, v) = \begin{cases} \lambda v - \dfrac{v^2}{2a} & v < a\lambda \\ \dfrac{1}{2} a\lambda^2 & v \geq a\lambda, \end{cases} \tag{7}$$

where $a > 0$ is the regularization parameter. It is noted that other penalties can take the place of MCP. Denote $\hat{\boldsymbol{\theta}}$ as the maximizer of $Q(\boldsymbol{\theta})$. Examining which components of $\hat{\boldsymbol{\gamma}}$ are zero can suggest which covariates have proportional effects.

For computation, consider the following iterative algorithm. (a) Compute the initial value. When the number of covariates is small compared to the sample size, a sensible initial value is the unpenalized estimate. (b) Optimize over $\boldsymbol{\gamma}$. $l(\boldsymbol{\theta})$ is continuously differentiable, and the penalty term is separable. Thus this step of optimization is best realized using the coordinate descent approach. (c) Optimize over other unknown parameters. This can be realized using the Newton–Raphson approach. (d) Repeat (b)–(c) until convergence. Convergence of the coordinate descent in (b) can be established following Tseng (2001). It can be shown that the objective function is non-increasing in (c). Thus the overall convergence can be established. In all of our numerical studies, convergence is achieved in a small number of iterations.

The proposed penalized estimation involves two tuning parameters. Following Breheny and Huang (2011), we set $a = 3$. In numerical study, we have experimented with a few other a values and reached comparable results, as long as a is not too large or too small. For selecting λ, we have numerically experimented with AIC, BIC, and GCV. We observe similar results with AIC and BIC and inferior results with GCV. Thus, we select λ by minimizing

$$\text{BIC}\ (\lambda) = -l\,(\boldsymbol{\theta})\,/n + \log\,(n)\,df\,(\lambda)\,/n, \tag{8}$$

where $df\,(\lambda)$ is the number of nonzero coefficients with a given λ value.

Remarks. The formulation in (6) is appropriate for low-dimensional covariates. It fits the data analyzed in Section 4. Its statistical properties are studied in the next section. Potentially, the proposed method is also applicable to data with a large number of covariates. In the Appendix, we revise (6) to accommodate high-dimensional covariates.

Variance estimation. Although the asymptotic distribution results are derived in the next section, we find that a plug-in variance estimate is difficult. For inference, we adopt the 0.632 bootstrap, which has been used in Huang et al. (2006) and others for inference with penalized estimation. Specifically, (a) B bootstrap samples are generated, each with size $0.632n$, by sampling without replacement from the original data. (b) Penalized estimation is conducted on each bootstrap sample. To reduce computational cost, the λ value selected for the original dataset is applied to all bootstrap samples. (c) The bootstrap estimates are pooled to generate the variance estimate. Although this approach is virtually equivalent to the nonparametric bootstrap, it is computationally more efficient by analyzing a smaller number of subjects.

2.3. Statistical properties

Denote z as the dimensionality of $\boldsymbol{\theta}$ and m as the dimensionality of $\left(\boldsymbol{\beta}^T, \tau, \alpha_1, \alpha_2, \sigma^2\right)^T$. Assume that the true values of $\gamma_2, \ldots, \gamma_{s+1}$ are nonzero, and those of $\gamma_{s+2}, \ldots, \gamma_p$ are zero. Denote $\boldsymbol{\gamma}_1 = (\gamma_2, \ldots, \gamma_{s+1})^T$, $\boldsymbol{\gamma}_2 = (\gamma_{s+2}, \ldots, \gamma_p)^T$, $\boldsymbol{\theta}_1 = \left(\boldsymbol{\beta}^T, \tau, \alpha_1, \alpha_2, \sigma^2, \boldsymbol{\gamma}_1^T\right)^T$, and $\boldsymbol{\theta}_2 = \boldsymbol{\gamma}_2$.

Use the subscript "0" for the true parameter values. Denote $I(\boldsymbol{\theta}_0)$ as the Fisher information matrix of $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta}_0$ and $I_1(\boldsymbol{\theta}_{10})$ as the Fisher information matrix of $\boldsymbol{\theta}_1$ evaluated at $\boldsymbol{\theta}_{10}$ knowing that $\boldsymbol{\theta}_2 = \mathbf{0}$. We first show that there exists a penalized MLE that converges at the rate of $O_P\left(n^{-1/2} + a_n\right)$, where $a_n = \max_j \left\{p'_\lambda(|\gamma_{j0}|) : \gamma_{j0} \neq 0\right\}$. Here $p'_\lambda\,(v)$ is the derivative of $p_a\,(\lambda, v)$ with respect to v. With the specific form of the penalty, this implies that the penalized estimate is $\sqrt{n}$-consistent if $\lambda \to 0$ as $n \to \infty$. Furthermore, we establish that such a $\sqrt{n}$-consistent estimator satisfies $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$, and $\hat{\boldsymbol{\theta}}_1$ is asymptotically normal with covariance matrix I_1^{-1}, if $n^{1/2}\lambda \to \infty$ as $n \to \infty$. This implies that the penalized MLE performs as well as if $\boldsymbol{\theta}_2 = \mathbf{0}$ were known in advance.

Theorem 1. *Let $\mathbf{V}_i = (\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$. Assume that $\mathbf{V}_i$'s are i.i.d. with density function $f(\mathbf{V}, \boldsymbol{\theta}_0)$. Further assume that conditions* (A)–(C) *listed in the Appendix hold. If $\max_j \left\{p''_\lambda(|\gamma_{j0}|) : \gamma_{j0} \neq 0\right\} \to 0$, where p''_λ is the second order derivative of $p_a\,(\lambda, v)$ with respect to v, then there exists a local maximizer $\hat{\boldsymbol{\theta}}$ of $Q\,(\boldsymbol{\theta})$ such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_P(n^{-1/2} + a_n)$.*

Proof is provided in the Appendix. Theorem 1 establishes that with a properly chosen λ, the penalized estimate is $\sqrt{n}$-consistent. We now establish the sparsity property.

Lemma 1. *Assume that conditions* (A)–(C) *listed in the Appendix are satisfied and that $\liminf_{n\to\infty} \liminf_{v\to 0^+} p'_\lambda(v)/\lambda > 0$. If $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, then with probability tending to* 1, *for any $\boldsymbol{\theta}_1$ that satisfies $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_P(n^{-1/2})$ and any constant C*

$$Q\left(\begin{pmatrix}\boldsymbol{\theta}_1\\ \mathbf{0}\end{pmatrix}\right) = \max_{\|\boldsymbol{\theta}_2\| \le Cn^{-1/2}} Q\left(\begin{pmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{pmatrix}\right).$$

Denote

$$\boldsymbol{\Sigma} = \operatorname{diag}\left\{0, \ldots, 0, p''_\lambda\,(|\gamma_{20}|)\,, \ldots, p''_\lambda\left(|\gamma_{(s+1)0}|\right)\right\},$$

and

$$\boldsymbol{b} = \left(0, \ldots, 0, p'_\lambda\,(|\gamma_{20}|)\operatorname{sgn}(\gamma_{20})\,, \ldots, p'_\lambda\left(|\gamma_{(s+1)0}|\right)\operatorname{sgn}\left(\gamma_{(s+1)0}\right)\right)^T.$$

Here $\operatorname{sgn}(\cdot)$ is the sign function.

Theorem 2. *Assume that conditions* (A)–(C) *listed in the Appendix hold and that $\liminf_{n\to\infty} \liminf_{v\to 0^+} p'_\lambda\,(v)\,/\lambda > 0$. If $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, then with probability tending to* 1, *the $\sqrt{n}$-consistent estimate in Theorem* 1 *satisfies:* (a) *Sparsity: $\hat{\boldsymbol{\theta}}_2 = \mathbf{0}$.* (b) *Asymptotic normality:*

$$\sqrt{n}\,(I_1\,(\boldsymbol{\theta}_{10}) + \boldsymbol{\Sigma})\left\{\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} + (I_1\,(\boldsymbol{\theta}_{10}) + \boldsymbol{\Sigma})^{-1}\,\boldsymbol{b}\right\} \to N\left\{\mathbf{0}, I_1\,(\boldsymbol{\theta}_{10})\right\}$$

in distribution.

3. Simulation

3.1. Selection of the anchor variable

With a practical dataset, the anchor variable needs to be determined. As suggested in Liu et al. (2011), it is possible that multiple covariates can serve as the anchor. Numerically, we have experimented with the following approaches. (a) Fit the unconstrained model, and select a covariate with strong effects in both models. (b) The bias-corrected mean squared error (MSE_c) approach described in Liu et al. (2012). Define $MSE_c = \frac{1}{n_1}\sum_{i=1}^{n}(\hat{\phi}_i\hat{\mu}_i - y_i)^2$, where $\hat{\phi}_i$ and $\hat{\mu}_i$ denote the predicted probability of a nonzero response and its value for subject *i*. In simulation, we generate an independent testing dataset under the same settings as the training dataset and select the covariate that minimizes MSE_c for the testing data as the anchor. In real data analysis, this criterion can be realized using the V-fold cross validation. (c) This approach is similar to (b) except that the criterion used is different. The new criterion is the model error of part 2 defined as $ME_2 = \left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right)^T E\left(\boldsymbol{X}^T\boldsymbol{X}\right)\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0\right)$. Our simulation suggests that approach (c) leads to the best performance (details omitted). In our simulation and data analysis, ME_2 is used as the criterion for choosing the anchor variable.

3.2. Simulation settings

Here we consider examples with a relatively small number of covariates, which matches data analyzed in the next section. In the Appendix, we also present simulation with a large number of covariates. For all examples, we generate covariates from a multivariate normal distribution with marginal mean 0 and variance 1. The correlation coefficient between the *j*th and *k*th covariates is $0.5^{|j-k|}$, and $\sigma = 0.5$. The response values are generated from the model described in Section 2.

Example 1. $p = 8$ and $\tau_0 = 0.2$. Consider a partial proportionality structure, where five covariates have proportional effects. The regression coefficient parameters are $\boldsymbol{\beta}_0 = (-1.5, -1.0, -0.5, 0.5, 1.0, 1.5, 1.7, 1)^T$, $\boldsymbol{\delta}_0 = (-0.3, -0.2, -0.1, -0.1, 0.2, 1.8, 3.34, 2.2)^T$, and $\boldsymbol{\gamma}_0 = (0, 0, 0, 0, 1.5, 3, 2)^T$.

Example 2. $p = 12$ and $\tau_0 = 0.2$. The first three covariates have proportional effects. The regression coefficient parameters are $\boldsymbol{\beta}_0 = (\underbrace{1, \ldots, 1}_{8}, \underbrace{-1, \ldots, -1}_{4})^T$,

$$\boldsymbol{\delta}_0 = (0.2, 0.2, 0.2, -0.8, -0.6, -0.4, -0.2, 0.7, 0.3, 0.5, 0.3, 0.9)^T, \quad \text{and}$$
$$\boldsymbol{\gamma}_0 = (0, 0, -1, -0.8, -0.6, -0.4, 0.5, 0.5, 0.7, 0.9, 1.1)^T.$$

Example 3. Consider a higher dimensionality with $p = 40$. Set the regression coefficient parameters as

$$\boldsymbol{\beta}_0 = \left(\underbrace{0, \ldots, 0}_{10}, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, \underbrace{0, \ldots, 0}_{10}, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\right),$$

and $\boldsymbol{\delta}_0 = \left(\underbrace{0.3, \ldots, 0.3}_{10}, \underbrace{1, \ldots, 1}_{10}, \underbrace{-0.3, \ldots, -0.3}_{10}, \underbrace{-1, \ldots, -1}_{10}\right)^T$. About half of the covariate effects are proportional. With X_1 as anchor, $\tau_0 = 0.3$,

$$\boldsymbol{\gamma}_0 = \left(\underbrace{0, \ldots, 0}_{9}, 0.85, 0.82, 0.79, 0.76, 0.73, 0.7, 0.67, 0.64, 0.61, 0.58, \underbrace{0, \ldots, 0}_{10},\right.$$
$$\left.-1.12, -1.15, -1.18, -1.21, -1.24, -1.27, -1.3, -1.33, -1.36, -1.39\right).$$

For all examples, we set the sample size $n = 200$, 400, and 800. For each simulated dataset (training), a testing dataset is generated independently under the same settings and used to evaluate prediction performance.

Table 1
Simulation results under Examples 1–3: median (sd) for Pro.C and mean (sd) for other measures.

	Method	n	Pro.C	FNR	FDR	ME_2	MSE_c
	True model		4				
Example 1	Unconstrained	200	–	–	–	0.039(0.032)	3.011(0.798)
		400	–	–	–	0.165(0.112)	2.821(0.765)
		800	–	–	–	0.143(0.105)	2.417(0.443)
	Stepwise	200	8(0.000)	0.000(0.000)	0.397(0.110)	0.028(0.025)	3.154(0.892)
		400	6(0.171)	0.400(0.032)	0.500(0.050)	0.136(0.128)	2.720(0.778)
		800	6(0.000)	0.400(0.000)	0.500(0.000)	0.143(0.082)	2.245(0.489)
	Lasso	200	7(0.930)	0.000(0.143)	0.429(0.242)	0.033(0.024)	1.788(0.488)
		400	6(1.799)	0.250(0.236)	0.429(0.278)	0.019(0.011)	1.810(0.362)
		800	4(1.656)	0.250(0.193)	0.000(0.294)	0.008(0.006)	1.667(0.281)
	MCP	200	7(1.427)	0.000(0.557)	0.429(0.314)	0.032(0.028)	1.755(0.536)
		400	4(0.943)	0.000(0.000)	0.000(0.202)	0.017(0.012)	1.654(0.385)
		800	4(0.100)	0.000(0.025)	0.000(0.000)	0.007(0.005)	1.587(0.227)
	True model		2				
Example 2	Unconstrained	200	–	–	–	0.141(0.035)	1.021(0.108)
		400	–	–	–	0.120(0.029)	0.803(0.088)
		800	–	–	–	0.105(0.024)	0.597(0.059)
	Stepwise	200	7(1.887)	1.000(0.363)	0.167(0.062)	0.123(0.039)	0.945(0.312)
		400	7(2.020)	1.000(0.452)	0.200(0.054)	0.107(0.073)	0.772(0.391)
		800	5(0.500)	1.000(0.045)	0.200(0.028)	0.082(0.061)	0.563(0.113)
	Lasso	200	2(3.214)	0.750(0.347)	0.000(0.494)	0.037(0.025)	0.527(0.204)
		400	2(2.030)	0.500(0.298)	0.000(0.359)	0.018(0.010)	0.464(0.100)
		800	2(0.580)	0.500(0.290)	0.000(0.000)	0.008(0.004)	0.416(0.047)
	MCP	200	2(2.141)	0.000(0.151)	0.000(0.310)	0.037(0.025)	0.523(0.134)
		400	2(0.000)	0.000(0.000)	0.000(0.000)	0.015(0.007)	0.419(0.068)
		800	2(0.000)	0.000(0.000)	0.000(0.000)	0.007(0.004)	0.425(0.051)
	True model		19				
Example 3	Unconstrained	200	–	–	–	0.277(0.133)	13.586(3.038)
		400	–	–	–	0.195(0.106)	6.605(1.488)
		800	–	–	–	0.114(0.058)	3.120(0.725)
	Stepwise	200	39(0.234)	0.000(0.000)	0.077(0.004)	0.273(0.109)	13.122(3.118)
		400	39(0.287)	0.000(0.000)	0.077(0.004)	0.168(0.855)	6.108(1.532)
		800	39(0.132)	0.000(0.000)	0.077(0.002)	0.091(0.617)	3.054(0.822)
	Lasso	200	4(5.748)	0.789(0.188)	0.000(0.091)	0.145(0.040)	6.771(2.891)
		400	15(4.905)	0.211(0.258)	0.000(0.000)	0.058(0.015)	4.006(1.440)
		800	17(1.443)	0.105(0.076)	0.000(0.000)	0.029(0.008)	3.135(0.722)
	MCP	200	33(5.184)	0.000(0.101)	0.441(0.117)	0.152(0.058)	8.328(2.380)
		400	19(3.392)	0.000(0.127)	0.000(0.162)	0.055(0.015)	4.011(1.553)
		800	19(0.200)	0.000(0.043)	0.000(0.045)	0.025(0.006)	2.572(0.590)

3.3. Analysis results

We apply the proposed approach to the simulated data. Beyond the MCP penalty, we also apply the Lasso penalty, which is more popular than MCP and can be viewed as an extreme case. For comparison, we consider the unconstrained approach (which does not impose any proportionality structure) and the stepwise approach in Han and Kronmal (2006). To evaluate variable selection/model identification performance, we compute the number of selected proportionality constraints (Pro.C), false negative rate (FNR), and false discovery rate (FDR). To evaluate prediction performance, ME_2 and MSE_c defined above are computed on the independent testing data. Across 200 replicates, we compute median (sd) for Pro.C and mean (sd) for the other measures. The results are shown in Table 1.

We observe that the proposed method satisfactorily identifies the proportionality structure, with a very small number of false findings. Performance improves as sample size increases. For the three simulated examples, performance is satisfactory enough for sample size equal to 400. MCP has more accurate identification results than Lasso. The penalization approach outperforms the stepwise approach with more accurate variable selection. The superior model identification performance of the proposed method also leads to superior prediction performance.

We also evaluate performance of the 0.632 bootstrap. Results for Example 1 are shown in Table 2. For the other two examples, similar results are obtained and omitted here. With 200 replicates and 500 bootstraps per replicate, we see that the bootstrap standard deviation estimates and observed standard deviations match well, which suggests satisfactory performance of the bootstrap.

Table 2
Simulation results of the 0.632 bootstrap under Example 1.

	n		δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8
Lasso	200	SD est	0.077	0.078	0.076	0.079	0.087	0.086	0.087	0.075
		Mean SD	0.080	0.076	0.070	0.071	0.081	0.087	0.088	0.078
	400	SD est	0.041	0.041	0.040	0.036	0.041	0.043	0.044	0.031
		Mean SD	0.038	0.040	0.037	0.038	0.039	0.044	0.043	0.034
	800	SD est	0.029	0.026	0.027	0.029	0.024	0.031	0.029	0.030
		Mean SD	0.028	0.027	0.025	0.028	0.025	0.030	0.029	0.031
MCP	200	SD est	0.053	0.058	0.047	0.042	0.060	0.072	0.063	0.062
		Mean SD	0.051	0.059	0.049	0.044	0.055	0.076	0.061	0.059
	400	SD est	0.035	0.037	0.036	0.037	0.035	0.041	0.042	0.040
		Mean SD	0.037	0.038	0.037	0.034	0.033	0.041	0.039	0.037
	800	SD est	0.026	0.026	0.025	0.026	0.026	0.032	0.033	0.029
		Mean SD	0.025	0.028	0.027	0.025	0.028	0.032	0.032	0.028

4. Data analysis

4.1. The CHNS medical cost data

It has been observed that, in multiple Asian countries, high medical cost has been an important contributing factor for poverty, and medical cost has been rising significantly in China (Fang et al., 2012). It is of interest to analyze the associated factors of medical cost in China. Here the data are extracted from the CHNS database (http://www.cpc.unc.edu/projects/china), which has been analyzed in multiple published studies. In this analysis, medical cost includes that of self-care, hospitalization, and extras beyond those two. Whether a subject has nonzero medical cost depends on whether he/she has been sick or injured during the four-week period prior to data collection. Potentially relevant covariates include monthly income (which includes wage or salary, subsidies and bounds of both primary and secondary occupations), years of education (from 0 to 26), age (from 18 to 76), square of age, Hukou (0: urban, 1: rural), gender (male is used as the reference), current health status (excellent, good, fair, and poor), marital status (never married, married, and divorced), and medical insurance (0: no, 1: yes). The first four covariates are continuous, current health status is ordinal, and the rest are nominal. The ordinal and nominal covariates are transformed into dummy variables. After removing those with invalid and missing measurements, we obtain 396 observations. Among them, 280 (70.7%) have zero medical cost. Descriptive statistics are provided in Table 6 (Appendix). The histogram in Fig. 1 clearly shows the zero-inflated nature of data. It also suggests a logarithm transformation of the response variable.

When applying the proposed approach, as simulation suggests almost dominatingly better performance of MCP over Lasso, we apply MCP only. With the approach described in Section 3.1, square of age is selected as the anchor variable. Seven other covariates are identified as having proportional effects, namely income, year of education, age, Hukou, gender, marital status (never), and medical insurance. With the proportionality structure, these covariates can be potentially viewed as forming a latent variable (or a composite index), affecting medical cost. The other covariates also affect cost, however, through a different channel. Estimation using the proposed method suggests that a subject with higher income is more likely to have nonzero medical cost and also have a higher value of medical cost. Education is negatively associated with the probability of nonzero cost and also the value of nonzero cost. Both age and square of age have positive effects. Similar observations have been made in the literature (Quan and Ai, 2008; Fang et al., 2012). Subjects living in the rural areas are less likely to have nonzero medical cost or have higher cost. This observation is consistent with the literature (Quan and Ai, 2008). Females are more likely to have nonzero and higher values of medical cost. Health status is negatively associated with the probability and value of medical cost. Those who are never married have a higher probability of nonzero cost. Those who have insurance have a high probability of nonzero medical cost and also tend to spend more. The "directions" of the estimates have intuitive interpretations.

Beyond the proposed approach, we also apply the stepwise approach and the unconstrained model. With the stepwise approach, all covariates are concluded as having proportional effects. For all three methods, the estimates and standard errors are provided in Table 3. Different methods lead to different constraint structures and comparable but different estimates. For example, for the variable health status (good) and the logit part of the model, the three estimates are −2.144, −2.014, and −2.126, respectively. The proposed method leads to the smallest estimated standard errors, which suggests more efficient estimation and fits the pattern observed in the published studies. For example, for the aforementioned estimates, the standard errors are 0.369, 0.602, and 0.621, respectively. The improved efficiency comes from a smaller number of parameters to be estimated as well as the regularization nature of estimation. To further compare the three methods, we adopt a cross validation-based approach and evaluate prediction performance. Beyond ME_2 and MSE_c described above, we also compute CE_1, the prediction error for the logit part of the model. Table 3 suggests that the proposed method has the best prediction performance, although when factoring in variance, the improvement in prediction is not significant.

Table 3
Analysis of the CHNS dataset using different methods. In each cell, estimate (sd).

Covariate	Proposed		Stepwise		Unconstrained	
	Logit	Linear	Logit	Linear	Logit	Linear
Intercept	0.925(0.453)	3.233(0.290)	0.645(0.485)	5.015(0.861)	0.981(0.772)	2.905(0.873)
Income	0.016(0.252)	0.006(0.089)[a]	0.010(0.256)	0.003(0.087)[a]	0.005(0.288)	0.066(0.143)
Education	−0.098(0.104)	−0.037(0.111)[a]	−0.092(0.110)	−0.024(0.123)[a]	−0.091(0.122)	−0.081(0.205)
Age	0.060(0.118)	0.023(0.049)[a]	0.070(0.119)	0.018(0.056)[a]	0.066(0.128)	0.011(0.197)
Age^2	0.268(0.127)	0.102(0.101)[a]	0.281(0.132)	0.074(0.094)[a]	0.278(0.198)	0.065(0.134)
Hukou (Rural)	−0.605(0.232)	−0.229(0.217)[a]	−0.610(0.239)	−0.160(0.226)[a]	−0.606(0.242)	−0.234(0.351)
Gender (Female)	0.007(0.198)	0.002(0.051)[a]	0.012(0.207)	0.003(0.084)[a]	−0.017(0.216)	0.169(0.325)
Health status (Excellent)	−2.652(0.362)	−0.053(0.347)	−2.518(0.618)	−0.661(0.806)[a]	−2.636(0.633)	0.067(0.945)
Health status (Good)	−2.144(0.369)	−0.050(0.375)	−2.014(0.602)	−0.529(0.628)[a]	−2.126(0.621)	−0.163(0.743)
Health status (Fair)	−1.541(0.357)	−0.134(0.269)	−1.403(0.498)	−0.369(0.483)[a]	−1.522(0.641)	0.072(0.645)
Marital status (Never)	0.629(0.197)	0.238(0.216)[a]	0.622(0.323)	0.163(0.263)[a]	0.576(0.546)	0.578(0.304)
Marital status (Married)	−0.023(0.459)	1.467(0.188)	0.150(0.476)	0.039(0.175)[a]	−0.059(0.517)	1.672(0.235)
Medical insurance (Yes)	0.459(0.223)	0.174(0.203)[a]	0.450(0.221)	0.118(0.312)[a]	0.419(0.325)	0.365(0.319)
σ	1.697(0.081)		1.753(0.081)		1.690(0.086)	
τ	0.379(0.377)		0.263(0.506)			
CE_1	0.309(0.031)		0.310(0.029)		0.312(0.031)	
ME_2	3.811(1.289)		3.854(1.544)		4.077(0.838)	
MSE_c	5.247(0.583)		5.341(0.502)		5.966(1.771)	

[a] Proportional effects.

Table 4
Analysis of the RCHS dataset using different methods. In each cell, estimate (sd).

Covariate	Proposed		Stepwise		Unconstrained	
	Logit	Linear	Logit	Linear	Logit	Linear
Intercept	−0.085(0.149)	5.031(0.204)	−0.131(0.126)	5.026(0.221)	−0.472(0.201)	5.112(0.179)
Household size	0.006(0.010)	0.027(0.041)[a]	0.005(0.013)	0.017(0.044)[a]	0.124(0.040)	−0.040(0.042)
Number of people aged 65+	−0.029(0.009)	−0.128(0.038)[a]	−0.037(0.012)	−0.12(0.035)[a]	0.079(0.042)	−0.149(0.033)
Education of household head	0.391(0.033)	0.111(0.025)	0.059(0.021)	0.182(0.032)[a]	0.420(0.264)	0.115(0.025)
Marital status (Never)	0.101(0.063)	0.434(0.250)[a]	0.143(0.082)	0.471(0.262)[a]	−0.007(0.179)	0.442(0.209)
Marital status (Married)	0.074(0.051)	0.316(0.187)[a]	0.096(0.061)	0.312(0.193)[a]	−0.075(0.351)	0.299(0.162)
Marital status (Divorced)	−0.059(0.056)	−0.258(0.251)[a]	−0.073(0.081)	−0.22(0.24)[a]	−0.121(0.029)	−0.263(0.207)
Household income	0.132(0.029)	0.576(0.028)	0.178(0.035)	0.567(0.031)[a]	0.101(0.029)	0.574(0.030)
Number of insured people	0.040(0.013)	0.173(0.039)[a]	0.051(0.014)	0.164(0.041)[a]	−0.010(0.039)	0.174(0.044)
Health status	−0.244(0.039)	0.128(0.036)	−0.019(0.008)	0.064(0.032)	−0.237(0.040)	0.123(0.030)
σ	1.327(0.024)		1.328(0.016)		1.33(0.014)	
τ	4.528(1.055)		3.332(0.795)			
CE_1	0.399(0.058)		0.398(0.064)		0.405(0.060)	
ME_2	1.854(0.374)		1.977(0.361)		1.858(0.359)	
MSE_c	11.362(0.996)		11.696(1.117)		13.052(1.646)	

[a] Proportional effects.

4.2. The RCHS health insurance expenditure data

China has one of the largest health insurance systems in the world. In rural China, commercial health insurance is very much undeveloped. The dominating form of basic health insurance is the new rural cooperative medical system (NCMS). The Rural China Health Survey (RCHS) study was conducted by the Data Mining Research Center at the Xiamen University of China in 2012. In this study, it is of interest to identify the factors associated with health insurance expenditure.

The survey was conducted in the rural areas of five cities in the Fujian Province, including Fuzhou, Quanzhou, Zhangzhou, Nanping, and Sanming. As household remained as the basic unit for health insurance and health expenditure in rural China, data were collected and analyzed at the household level. The survey has a satisfactory response rate of 76%. After removing invalid observations and those with missing measurements, we obtain 561 valid samples. The response variable is the household health insurance expenditure during a period of twelve months prior to survey. There are a total of 228 (40.6%) zeros. The histogram in Fig. 2 (Appendix) also suggests the zero-inflated nature of data and a logarithm transformation. The following covariates are included in analysis: household size, number of household members aged 65+, education of household head, marital status of household head, household income in twelve months, number of people insured, and health status of household head. Descriptive statistics are shown in Table 7 (Appendix). This dataset has also been analyzed in the literature (Yi et al., 2016).

With the approach described in Section 3.1, household income is chosen as the anchor variable. Six covariates are identified as having proportional effects, including household size, number of household members aged 65+, marital status (never, married, divorced), and number of people insured. The detailed estimation and inference results are shown in Table 4. It is observed that the levels of household head education, household income, and number of people insured are positively associated with the probability of nonzero health insurance expenditure as well as its level if nonzero. Education enhances the risk awareness of a household head. Higher income improves the purchasing power of a household. Expenditure increases as the number of people insured increases. Households with heads never married have a higher probability of health insurance spending as well as higher values of expenditure. This observation is consistent with that in the previous section. Households with divorced heads are the least likely to spend on health insurance, and the amount is low. Health status is negatively associated with the probability of spending on health insurance. People with worse health status are more likely to purchase insurance. The number of household members aged 65+ years is negatively associated with the probability and the value of health insurance expenditure. Under the current NCMS regulations, the elderly only need to pay a small amount to obtain basic health insurance, and the government pays the rest of the premium. Overall, the findings are consistent with the literature and intuitive.

In Table 4, we also present the analysis results using the two alternative methods. The patterns are similar to those for the previous dataset. The standard errors under the proposed method are in general smaller, suggesting more efficient estimation. In the evaluation of prediction performance, the proposed method has CE_1 similar to that of the stepwise approach and smaller than that of the unconstrained model. It has the smallest ME_2 and MSE_c.

4.3. Remarks

The two datasets have been analyzed in published studies, which contain more detailed information on the study design and data collection. The analysis in this study takes a different angle, focuses on the proportionality structure of the covariate effects, and may provide additional insights beyond the existing studies. In both datasets, the response variable is expenditure. The histograms in the Appendix suggest that the logarithm transformation is reasonable. For simplicity of interpretation, we take this transformation and then adopt a normal model for the positive response values. We note that such a transformation and model may not provide the best fit of data. As a limitation of this study, we are not able to find a simple and interpretable alternative model. We note that the logarithm transformation and normal model have also been adopted in published studies.

5. Discussion

Two-part modeling provides an effective tool for zero-inflated data and other data types with mixture distributions. For such data under many scenarios, it is of interest to examine the proportionality structure of the covariate effects. A properly identified proportionality structure can lead to more efficient estimation and may also have important scientific implications. In this study, we have applied penalization for identifying the proportionality structure. The proposed approach has an intuitive formulation and computational advantages. As shown in the Appendix, it can be easily applied under high-dimensional settings, which cannot be achieved using the alternatives. We have rigorously established the consistency properties. The alternative studies have mostly focused on methodological development. With a solid theoretical basis, the proposed method can be preferred over the alternatives. Simulation shows the superior performance of the proposed method. In data analysis, the proposed method generates interpretable and more efficient estimation. It also has satisfactory prediction performance.

To match the datasets analyzed, we adopt the logistic and linear models. These two models can be replaced with other sensible models. Specifically, there are a large number of generalized linear models for modeling binary responses. When the nonzero responses are counts (as has been encountered in a large number of published studies), the second part of the model can be replaced by a Poisson regression and others. Once the likelihood function is properly constructed, the proposed approach will be applicable. The adopted MCP can also be replaced by other penalty functions. Our theoretical investigation establishes the asymptotic estimation and proportionality structure identification consistency. As the analyzed datasets have a relatively small number of covariates, we have focused on the "classic" asymptotics. Theoretical investigation with high-dimensional data will be pursued in future studies.

Acknowledgments

We thank the associate editor and three reviewers for their careful review and insightful comments, which have led to a significant improvement of the manuscript. This study was supported by the National Natural Science Foundation of China (71471152, 71201139, and 71303200) and National Social Science Foundation of China (13&ZD148, and 13CTJ001).

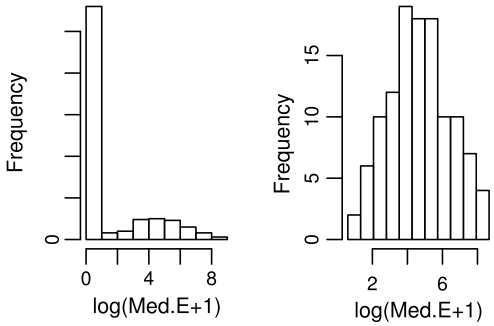

Fig. 1. Analysis of the CHNS data: histograms for all subjects (left panel) and for those with nonzero medical cost (right panel). Med.E: the response variable.

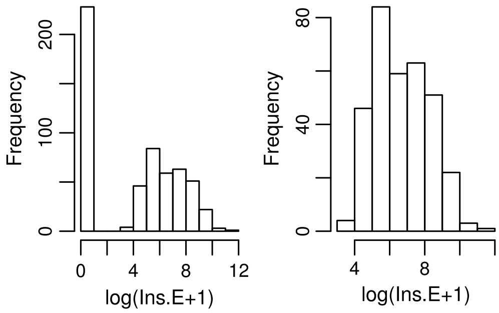

Fig. 2. Analysis of the RCHS data: histograms for all subjects (left panel) and for those with nonzero health insurance expenditure (right panel). Ins.E: the response variable.

Appendix

A.1. Accommodating high-dimensional covariates

In formulation (6), regularization (penalization) is imposed on $\boldsymbol{\gamma}$ for identifying the proportionality structure. When the number of covariates is moderate to large, regularized estimation is in general needed. In addition, not all covariates may be relevant, and variable selection is needed. Motivated by such concerns, we consider the following objective function

$$Q(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - P_1(\lambda_1, |\boldsymbol{\gamma}|) - P_2(\lambda_2, |\boldsymbol{\beta}|), \tag{9}$$

where P_1 and P_2 are two penalty functions. For computational simplicity, we take both of them to be MCP.

For computation, consider the following iterative algorithm: (a) Compute the initial value. A simple choice is the ridge estimate where squared penalties are imposed on both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. (b) Optimize over $\boldsymbol{\gamma}$ using the coordinate descent approach. (c) Optimize over $\boldsymbol{\beta}$ using the coordinate descent approach. (d) Optimize over the other unknown parameters using the Newton–Raphson approach. (e) Repeat steps (b)–(d) until convergence. The convergence properties can be established accordingly. Tuning parameters are selected using BIC in a similar manner as in Section 2.2.

Simulation (Example 4). Consider simulation with high-dimensional covariates. Building on Example 1, we add 360 zero elements to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$, so that there are a total of 368 covariates. Covariate values are generated in a similar manner as under Example 1. Set $n = 200$.

For the two penalty functions, we apply (a) Lasso + Lasso and (b) MCP + MCP. With high-dimensional data, the unconstrained model is not feasible. In addition, the stepwise variable selection has been suggested as unreliable in the literature. Thus they are not applied. Simulation suggests that the penalization method is computationally feasible: the analysis of one replicate takes about 200 s on a regular laptop. The variable selection results for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are reported in Table 5. It is observed that both penalty combinations can correctly identify the majority of zero values. The FNRs and FDRs are low. Overall the MCP + MCP method has better variable selection performance. Prediction performance of the two penalty combinations is comparable.

A.2. Proof

Before presenting the proofs, we first state the sufficient regularity conditions. Denote by Ω the parameter space for $\boldsymbol{\theta}$.

Table 5
Simulation results under Example 4: mean(sd) over 200 replicates.

Method	Parameter	# of zeros	FNR	FDR	ME_2	MSE_c
True model	$\boldsymbol{\gamma}$	364				
	$\boldsymbol{\beta}$	360				
Lasso + Lasso	$\boldsymbol{\gamma}$	353.0(10.9)	0.03(0.03)	0(0)	0.255(0.185)	6.884(9.238)
	$\boldsymbol{\beta}$	299.4(42.5)	0.169(0.118)	0.001(0.001)		
MCP + MCP	$\boldsymbol{\gamma}$	362.9(0.8)	0.003(0.002)	0(0)	0.315(0.283)	5.655(2.456)
	$\boldsymbol{\beta}$	325.8(74.2)	0.106(0.201)	0.011(0.008)		

Table 6
Analysis of the CHNS dataset: descriptive statistics.

Variable	Mean	Minimum	Maximum
Medical cost	122.654	0	3 500
Income	1313	110	18 000
Education	15.030	0	26
Age	40.580	18	76
Hukou(Rural)	0.318	0	1
Gender (Female)	0.374	0	1
Health status (Excellent)	0.189	0	1
Health status (Good)	0.523	0	1
Health status (Fair)	0.255	0	1
Marital status (Never)	0.116	0	1
Marital status (Married)	0.854	0	1
Medical insurance (Yes)	0.674	0	1

Table 7
Analysis of the RCHS dataset: descriptive statistics.

Variable	Mean	Minimum	Maximum
Health insurance expenditure	1526.922	0	70 000
Household size	4.333	1	11
Number of people aged 65+	0.504	0	4
Education of household head	2.961	1	6
Marital status (Never)	0.023	0	1
Marital status (Married)	0.920	0	1
Marital status (Divorced)	0.014	0	1
Household income	55 220	1000	300 000
Number of insured people	4.164	0	11
Health status	1	3.57	5

Regularity conditions

(A) The observations $\boldsymbol{V}_i$'s are independently and identically distributed with probability density $f(\boldsymbol{V}, \boldsymbol{\theta}_0)$ with respect to some measure μ. For $\boldsymbol{\theta} \in \Omega, f(\boldsymbol{V}, \boldsymbol{\theta})$ has a common support, and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f satisfy the equations

$$E_{\boldsymbol{\theta}}\left(\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_j}\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0,$$

and

$$I_{tj}(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}}\left(\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_t}\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_j}\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -E\left(\frac{\partial^2 \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \theta_t \partial \theta_j}\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

(B) The Fisher information matrix

$$I(\boldsymbol{\theta}_0) = E\left\{\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \log f(\boldsymbol{V}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T\right\}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

is finite and positive definite.

(C) There exists an open subset ω of Ω that contains the true parameter value $\boldsymbol{\theta}_0$, such that for almost all $\boldsymbol{V}$, the density $f(\boldsymbol{V}, \boldsymbol{\theta})$ admits all third-order derivatives $\frac{\partial^3 \log f(\boldsymbol{V},\boldsymbol{\theta})}{\partial\theta_t\partial\theta_j\partial\theta_k}$ for all $\boldsymbol{\theta} \in \omega$. Furthermore, there exist functions M_{tjk} such that

$$\left| \frac{\partial^3 \log f\,(\boldsymbol{V}, \boldsymbol{\theta})}{\partial\theta_t\partial\theta_j\partial\theta_k} \right| \le M_{tjk}(\boldsymbol{V})$$

for all $\boldsymbol{\theta} \in \omega$.

Proof of Theorem 1. Let $\alpha_n = n^{-1/2} + a_n$. We need to show that for any given $\eta > 0$, there exists a constant C such that

$$P\left\{ \sup_{\|\boldsymbol{u}\|=C} Q(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) < Q(\boldsymbol{\theta}_0) \right\} \ge 1 - \eta.$$

This implies that, with probability at least $1-\eta$, there exists a local maximum in the ball $\{\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u} : \|\boldsymbol{u}\| \le C\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\alpha_n)$.

Using $p_a(\lambda, 0) = 0$, we have

$$\begin{aligned} D_n(\boldsymbol{u}) &\equiv Q(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) - Q(\boldsymbol{\theta}_0) \\ &\le l(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) - l(\boldsymbol{\theta}_0) - n \sum_{j=m+1}^{m+s} \left[p_a(\lambda, |\theta_{j0} + \alpha_n u_j|) - p_a(\lambda, |\theta_{j0}|) \right]. \end{aligned}$$

Let $l'(\boldsymbol{\theta}_0)$ be the gradient vector of l. Applying Taylor expansion to $l(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u})$ at $\boldsymbol{\theta}_0$, we have

$$D_n(\boldsymbol{u}) \le \alpha_n l'(\boldsymbol{\theta}_0)^T\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T I(\boldsymbol{\theta}_0)\boldsymbol{u} n\alpha_n^2 \{1 + o_P(1)\} - \sum_{j=m+1}^{m+s} \left[n\alpha_n p'_\lambda(|\theta_{j0}|)\mathrm{sgn}(\theta_{j0})u_j + n\alpha_n^2 p''_\lambda(|\theta_{j0}|)u_j^2 \{1 + o_P(1)\} \right].$$

Note that $n^{-1/2}l'(\boldsymbol{\theta}_0) = O_P(1)$. Thus, the first term on the right-hand side of the inequality is of the order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly on $\|\boldsymbol{u}\| = C$. Note that the third term in the inequality is bounded by

$$\sqrt{s} n\alpha_n a_n \|\boldsymbol{u}\| + n\alpha_n^2 \max_j \left\{ |p''_\lambda(|\theta_{j0}|)| : \theta_{j0} \neq 0, j = m+1, \ldots, m+s, \right\} \|\boldsymbol{u}\|^2.$$

This is also dominated by the second term of the inequality. Hence, by choosing a sufficiently large C, we have

$$P\left\{ \sup_{\|\boldsymbol{u}\|=C} Q(\boldsymbol{\theta}_0 + \alpha_n\boldsymbol{u}) < Q(\boldsymbol{\theta}_0) \right\} \ge 1 - \eta.$$

Proof of Lemma 1. It is sufficient to show that as $n \to \infty$, with probability tending to 1, for any $\boldsymbol{\theta}_1$ satisfying $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_{10}\| = O_P(n^{-1/2})$ and for some small $\xi_n = Cn^{-1/2}$ and $j = m+s+1, \ldots, z$,

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial\theta_j} \begin{cases} <0, & \text{for } 0 < \theta_j < \xi_n \\ >0, & \text{for } -\xi_n < \theta_j < 0. \end{cases}$$

To show the first inequality, by Taylor's expansion, we have

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta})}{\partial\theta_j} &= \frac{\partial l(\boldsymbol{\theta})}{\partial\theta_j} - np'_\lambda(|\theta_j|)\mathrm{sgn}(\theta_j) \\ &= \left.\frac{\partial l(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\sum_{t=1}^{z} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_t}(\theta_t - \theta_{t0})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \left.\sum_{t=1}^{z}\sum_{k=1}^{z} \frac{\partial^3 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_t\partial\theta_k}(\theta_t - \theta_{t0})(\theta_k - \theta_{k0})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} - np'_\lambda(|\theta_j|)\mathrm{sgn}(\theta_j), \end{aligned}$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Note that by standard arguments,

$$\left.\frac{1}{n}\frac{\partial l\,(\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = O_P\left(n^{-1/2}\right),$$

and

$$\left.\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_t}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = E\left.\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_t\partial\theta_j}\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_P(1).$$

By the assumption that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O_P(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\theta})}{\partial\theta_j} = n\lambda \left\{ -\lambda^{-1}p'_\lambda(|\theta_j|)\mathrm{sgn}(\theta_j) + O_P(n^{-1/2}/\lambda) \right\}.$$

When $\liminf_{n\to\infty} \liminf_{v\to 0^+} p'_\lambda(v)/\lambda > 0$, the sign of the derivative is determined by that of θ_j. The proof is completed.

Proof of Theorem 2. It can be easily shown that there exists a $\hat{\boldsymbol{\theta}}_1$ in Theorem 1 that is a $\sqrt{n}$-consistent optimizer of $Q\left(\left(\boldsymbol{\theta} = \binom{\boldsymbol{\theta}_1}{\mathbf{0}}\right)\right)$ and satisfies the likelihood equations

$$\left.\frac{\partial Q(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\binom{\hat{\boldsymbol{\theta}}_1}{\mathbf{0}}} = 0, \quad \text{for } j = 1, \ldots, m+s.$$

With the consistency of $\hat{\boldsymbol{\theta}}_1$, for $j = m+1, \ldots, m+s$,

$$\begin{aligned}\left.\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\binom{\hat{\boldsymbol{\theta}}_1}{\mathbf{0}}} - np'_\lambda(|\hat{\theta}_j|)\mathrm{sgn}(\hat{\theta}_j) &= \left.\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \sum_{t=1}^{m+s}\left\{\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_t}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_P(1)\right\}(\hat{\theta}_t - \theta_{t0}) \\ &\quad - n\left(p'_\lambda\left(|\theta_{j0}|\right)\mathrm{sgn}\left(\theta_{j0}\right) + \left\{p''_\lambda\left(|\theta_{j0}|\right) + o_P\left(1\right)\right\}\left(\hat{\theta}_j - \theta_{j0}\right)\right) \\ &= 0,\end{aligned}$$

and for $j = 1, \ldots, m$,

$$\left.\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\binom{\hat{\boldsymbol{\theta}}}{\mathbf{0}}} = \left.\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \sum_{t=1}^{m+s}\left\{\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_t}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + o_P(1)\right\}(\hat{\theta}_t - \theta_{t0}) = 0.$$

It follows from Slutsky's theorem and the central limit theorem that

$$\sqrt{n}\left(I_1\left(\boldsymbol{\theta}_{10}\right) + \boldsymbol{\Sigma}\right)\left\{\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} + \left(I_1\left(\boldsymbol{\theta}_{10}\right) + \boldsymbol{\Sigma}\right)^{-1}\boldsymbol{b}\right\} \to N\left\{\mathbf{0}, I_1\left(\boldsymbol{\theta}_{10}\right)\right\}$$

in distribution.

References

Agarwal, D.K., Gelfand, A.E., Citron-Pousty, S., 2002. Zero-inflated models with application to spatial count data. Environ. Ecol. Stat. 9, 341–355.

Alfó, M., Maruotti, A., 2010. Two-part regression models for longitudinal zero-inflated count data. Canad. J. Statist. 38 (2), 197–216.

Bratti, M., Miranda, A., 2011. Endogenous treatment effects for count data models with endogenous participation or sample selection. Health Econ. 20 (9), 1090–1109.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann. Appl. Stat. 5 (1), 232–253.

Cheung, Y.B., 2002. Zero-inflated models for regression analysis of count data: a study of growth and development. Stat. Med. 21 (10), 1461–1469.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39, 829–844.

Deb, P., Munkin, M.K., Trivedi, P.K., 2006. Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. J. Appl. Econometrics 21 (7), 1081–1099.

Fang, K., Shia, B., Ma, S., 2012. Health Insureance coverage and impact: a survey in three cities in China. PLoS One 7 (6), e39157.

Ghosh, S.K., Mukhopadhyay, P., Lu, J.C., 2006. Bayesian analysis of zero-inflated regression models. J. Statist. Plann. Inference 136, 1360–1375.

Greene, W., 2009. Models for count data with endogenous participation. Empir. Econom. 36 (1), 133–173.

Han, C., Kronmal, R., 2006. Two-part models for analysis of Agatston scores with possible proportionality constraints. Comm. Statist. Theory Methods 35, 99–111.

Huang, J., Ma, S., Xie, H., 2006. Regularized estimation in the accelerated failure time model with high dimensional covariates. Biometrics 62, 813–820.

Lambert, D., 1992. Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34, 1–14.

Liu, H., Chan, K., 2011. Generalized additive models for zero-inflated data with partial constraints. Scand. J. Statist. 38, 650–665.

Liu, A., Kronmal, R., Zhou, X., Ma, S., 2011. Determination of proportionality in two-part models and analysis of Multi-Ethnic Study of Atherosclerosis (MESA). Stat. Interface 4, 475–487.

Liu, H., Ma, S., Kronmal, R., Chan, K., 2012. Semiparametric zero-inflated modeling in multi-ethnic study of atherosclerosis (MESA). Ann. Appl. Stat. 6, 1236–1255.

Manning, W.G., Duan, N., Rogers, W.H., 1987. Monte Carlo evidence on the choice between sample selection and two-part models. J. Econometrics 35 (1), 59–82.

Maruotti, A., Raponi, V., Lagona, F., 2015. Handling endogeneity and nonnegativity in correlated random effects models: Evidence from ambulatory expenditure. Biom. J. http://dx.doi.org/10.1002/bimj.201400121.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, New York, Chichester.

Min, Y., Agresti, A., 2005. Random effect models for repeated measures of zero-inflated count data. Stat. Model.: Int. J. 5 (1), 1–19.

Mullahy, J., 1986. Specification and testing of some modified count data models. J. Econometrics 3, 341–365.

Olsen, M.K., Schafer, J.L., 2001. A two-part random-effects model for semicontinuous longitudinal data. J. Amer. Statist. Assoc. 96 (454), 730–745.

Quan, X., Ai, C., 2008. Determinants of Chinese Residents' demand for medical care—an application of semiparametric estimation of ordered probit model. Stat. Res. 25, 40–45.

Tseng, P., 2001. Convergence of a Block Coordinate Descent method for nondifferentiable minimization. J. Optim. Theory Appl. 109, 475–494.

Yau, K.K., Lee, A.H., 2001. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Stat. Med. 20, 2907–2920.

Yi, H., Zhang, J., Ma, C., Ma, S., 2016. Utilization of the NCMS and its association with expenditures: observations from rural Fujian, China. Public Health 130, 84–86.