Statistica Sinica **23** (2013), 1071-1090
doi:http://dx.doi.org/10.5705/ss.2012.187

MULTIPLE-INFLATION POISSON MODEL WITH L_1 REGULARIZATION

Xiaogang Su[1], Juanjuan Fan[2], Richard A. Levine[2],
Xianming Tan[3], and Arvind Tripathi[1]

[1] *University of Alabama at Birmingham,* [2] *San Diego State University,*
and [3] *McGill University*

Abstract: A multiple-inflation Poisson (MIP) model is put forward for analyzing count data that have multiple inflated values. Analogous to the zero-inflated Poisson model (ZIP; Lambert (1992)), MIP assumes a mixture distribution of Poisson and degenerate distributions, where the probabilities for the inflated values are from a cumulative logit model. We explore the properties of the proposed model, with a detailed treatment given to its maximum likelihood estimation. Moreover, we address variable selection by adopting an L_1 regularization scheme. Both simulation experiments and an analysis of a health care data set are provided to illustrate the multiple-inflation Poisson model.

Key words and phrases: Count data, LASSO, Poisson distribution, variable selection, zero-inflated.

1. Introduction

The zero-inflated Poisson model (ZIP; Lambert (1992)) and its variants have become a popular tool in statistical applications for analyzing count data with a preponderance of zeros. However, we have found ourselves confronted with data displaying multiple inflated values. The illustration in Section 6 considers a healthcare study on the frequency of medical visits which shows an excess of zeros and ones. This data structure may be explained by the fact that a large number of patients may not visit a doctor or a health professional over a given period or require one, with perhaps a single follow-up, visit to diagnose and treat a given ailment. We have confronted similar situations in modeling the number of insurances that are of different types and of different policies and the number of hospitalization days in healthcare applications. Multiple inflated values may also occur when there is a natural "grouping" of the counts. For example, in the National Health and Nutrition Examination Survey (NHANES) data, the number of cigarettes smoked per day, according to self-reporting, is dominated by zeros and twenties, since twenty cigarettes correspond to one pack. Of course these are anecdotal accounts, but nonetheless suggest that an extension of the ZIP model for multiple inflated values is desired.

Our initial exposure to multiple inflated count data came in the analysis of traffic data where, for example, the number of monthly car crashes on high speed roadway segments is mostly zeros, ones, and twos. The transportation literature has expressed serious concern with ZIP models (Lord, Washington, and Ivan (2005, 2007)) stemming primarily from the assumption of a dual-state system of safe and non-safe road zones. One recommended solution is a multiple state crash process; however, as we discuss in more detail after our methodological development, mixture Poisson models are not of practical use in such a setting. Our proposed multiple-inflation Poisson model alleviates this difficulty. A second recommendation is to include a strong set of explanatory variables and we present an L_1 regularization scheme for variable selection to this end.

The paper unfolds as follows. In Section 2 we propose a multiple-inflation Poisson (MIP) model and discuss various issues related to model specifications, mixture model representation, identifiability, and (over and under) dispersion. In Section 3 we address maximum likelihood (ML) estimation under the MIP model and associated computational machinery. In Section 4 we propose the variable selection method for the multiple-inflation Poisson model. In Section 5 we report on simulation experiments designed to evaluate the inferential performance of the multiple-inflation Poisson model relative to competing models in the literature, and the performance of the proposed variable selection routine. In Section 6 an illustration of the proposed model is presented for studying the distribution of patient medical visits with, again, a comparison to competing models in the literature.

2. Multiple-Inflation Poisson (MIP) Model

Consider data that consist of n i.i.d. observations $\{(y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$, where y_i is the count of some event of interest and $\mathbf{X}_i$ is the associated predictor vector. The count response y_i contains a total of M inflated values and, while these inflated values do not have to be consecutive in the model, for notational convenience we denote them as $\{0, 1, \ldots, (M-1)\}$.

2.1. Model specification

The multiple-inflation Poisson model is specified as follows:

$$y_i \sim \begin{cases} m & \text{with probability } p_{im} \text{ for } m=0,\ldots,(M-1), \\ \text{Poisson}(\lambda_i) & \text{with probability } p_{iM}, \end{cases} \tag{2.1}$$

where $\sum_{m=0}^{M} p_{im} = 1$, so that

$$y_i = \begin{cases} m \text{ with probability } p_{im} + p_{iM}\exp(-\lambda_i)\frac{\lambda_i^m}{m!} & \text{for } m=0,\ldots,(M-1) \\ k \text{ with probability } p_{iM}\exp(-\lambda_i)\frac{\lambda_i^k}{k!} & \text{for } k \geq M. \end{cases} \tag{2.2}$$

For regression purposes, we express the mean λ_i of the Poisson model as

$$\log(\lambda_i) = \mathbf{B}_i^T \boldsymbol{\beta} \quad \text{or} \quad \lambda_i = \exp\left(\mathbf{B}_i^T \boldsymbol{\beta}\right) \tag{2.3}$$

and we formulate the p_{im}'s with a cumulative logit or proportional odds model (McCullagh (1980))

$$\text{logit}\left(\Pr\{y_i \leq m\}\right) = \log \frac{\Pr\{y_i \leq m\}}{\Pr\{y_i > m\}} = \mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{m0} \tag{2.4}$$

for $m = 0, 1, \ldots, (M-1)$. Here, both $\mathbf{B}_i$ and $\mathbf{G}_i$ are associated covariate vectors containing selected components from $\mathbf{X}_i$; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_1$ are vectors of regression parameters; the intercepts are separately denoted by γ_{m0}'s. Let $\boldsymbol{\gamma}^T = (\gamma_{00}, \gamma_{10}, \ldots, \gamma_{(M-1)0}, \boldsymbol{\gamma}_1^T)$. The cumulative logit model (2.4) implies that,

$$\begin{cases} p_{i0} = \text{expit}\left(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{00}\right), \\ p_{im} = \text{expit}\left(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{m0}\right) - \text{expit}\left(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{(m-1)0}\right) \text{ for } m=1,\ldots,(M-1), \\ p_{iM} = 1 - \text{expit}\left(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{(M-1)0}\right), \end{cases} \tag{2.5}$$

where $\text{expit}(x) = e^x/(1+e^x)$.

Another way of interpreting the multiple-inflation Poisson model is to view equations in (2.1) as describing a process with $(M+1)$ states; starting with state 0, $y_i = 0$; at state 1, $y_i = 1$, and so on; at state $(M-1)$, $y_i = M-1$; at state M, y_i follows Poisson(λ_i). With a slight abuse of notation, we have used y_i to denote both the random variable and the observed value for the i-th count. Introduce dummy variables z_{im} such that $z_{im} = 1$ if y_i is from the m-th state and 0 otherwise, for $m = 0, 1, \ldots, M$ and $i = 1, \ldots, n$. Thus $\sum_m z_{im} = 1$ and $z_{im}\, z_{im'} = 0$ for any $m \neq m'$. The conditional distribution of $(y_i | z_{i0}, \ldots, z_{iM})$ is thus given by

$$\begin{cases} \Pr(y_i = 0 \,|\, z_{i0} = 1) = 1, \\ \qquad \vdots \\ \Pr(y_i = M-1 \,|\, z_{i(M-1)} = 1) = 1, \\ y_i \,|\, (z_{iM} = 1) \;\sim\; \text{Poisson}(\lambda_i). \end{cases} \tag{2.6}$$

2.2. More on the MIP model specification

The multiple-inflation Poisson model specified in (2.1)−(2.3) is essentially a finite mixture model (FMM) of Poisson and degenerate distributions. A slightly different model formulation can be obtained via a mixture of a discrete distribution over all inflated values $\{0, 1, \ldots, (M-1)\}$ and a Poisson distribution, where the mixture probability is supplied by a Bernoulli model. Specifically,

$$y_i \sim \begin{cases} \text{Discrete}\left\{(0, \ldots, (M-1))\,;\, \left(p'_{i0}, \ldots, p'_{i(M-1)}\right)\right\} & \text{w/ prob. } 1 - p_{iM}, \\ \text{Poisson}(\lambda_i) & \text{w/ prob. } p_{iM}, \end{cases} \tag{2.7}$$

subject to $\sum_{m=0}^{M-1} p'_{im} = 1$. Model (2.7) is equivalent to Model (2.1), since the constraint on p'_{im} can be accounted for by $p'_{im} = p_{im}/(1-p_{iM})$ for $m = 0, \ldots, (M-1)$. To incorporate covariates into (2.7), three regression models are needed:

$$\begin{cases} \log \dfrac{\Pr\{y_i \leq m\}}{\Pr\{M-1 \geq y_i > m\}} = \widetilde{\mathbf{G}}_i^T \widetilde{\boldsymbol{\gamma}}_1 + \widetilde{\gamma}_{m0} \ \text{ for } \ m = 0, 1, \ldots, (M-2), \\ \log(\lambda_i) = \widetilde{\mathbf{B}}_i^T \widetilde{\boldsymbol{\beta}}, \\ \text{logit}(p_{iM}) = \widetilde{\mathbf{H}}_i^T \widetilde{\boldsymbol{\gamma}}_2 + \widetilde{\gamma}_{M0}. \end{cases} \tag{2.8}$$

This reduces to the model in (2.1)−(2.2) if some additional constrains are placed. For example, if $\widetilde{\mathbf{G}}_i = \widetilde{\mathbf{H}}_i = \mathbf{G}_i$, $\widetilde{\mathbf{B}}_i = \mathbf{B}_i$, and $\widetilde{\boldsymbol{\gamma}}_1 = \widetilde{\boldsymbol{\gamma}}_2$, then the parameters in (2.8) have a one-to-one correspondence with those in (2.3)−(2.4), determined by a number of identities as follows:

$$\begin{aligned} &\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}, \quad \frac{\mathbf{G}_i^T \widetilde{\boldsymbol{\gamma}}_1 + \widetilde{\gamma}_{m0}}{\mathbf{G}_i^T \widetilde{\boldsymbol{\gamma}}_1 + \widetilde{\gamma}_{00}} = \frac{\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{m0}}{\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{00}} \quad \text{for } m = 1, \ldots, (M-2), \\ &1 - \text{expit}(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{(M-1)0}) = \text{expit}(\mathbf{G}_i^T \widetilde{\boldsymbol{\gamma}}_1 + \widetilde{\gamma}_{M0}). \end{aligned}$$

We have assumed that the number of inflations, M, is fixed, together with the choices of the inflated values. When $M = 1$, the multiple-inflation Poisson model reduces to the zero-inflated Poisson (ZIP) model (Lambert (1992)). In practice, M can be manifested by inspecting the histogram of the count response or examining the residuals of a loglinear model fit, since the inflated values are typically those that do not fit well. On the other hand, how to determine the inflated values more precisely could be a future research topic that we do not pursue here. It is important to note that the inflated count values do not have to be consecutive. For example, they could be 0's and 10's, instead of 0's and 1's. The multiple-inflation Poisson model specification generalizes well to nonconsecutive scenarios, since the cumulative logit model generally works for ordinal responses following its latent variable justification. In addition, multinomial logistic regression and other models for categorical or ordinal responses can be used instead, but may incur more parameters than the cumulative logit model.

Previous work similar to MIP includes the Poisson mixture model, a general framework where multiple Poisson processes are used to model several states. The multiple-inflation Poisson model can be viewed as a special case of the mixture of Poisson models as discussed in Cameron and Trivedi (1998, Sec. 4.8). In particular, it can be connected to the multinomial-Poisson homogeneous model studied by Baker (1994), Wang et al. (1996), and Lang (2004), where a multinomial distribution models the mixing probabilities for Poisson mixtures. MIP is a simplified version of this multinomial-Poisson model, where all but one of the Poisson distributions are replaced by degenerate distributions. Following the arguments in Lang (2004), both maximum likelihood estimation theory and maximum likelihood fitting techniques can be straightforwardly applied to the multiple-inflation Poisson model. With that being said, the general multinomial-Poisson models are often difficult to fit (Baker (1994)) and their use is not common in applications (Lord, Washington, and Ivan (2005)). Comparatively, the simpler multiple-inflation Poisson model provides a flexible modeling tool ready for practical usage.

2.3. Dispersion

As noted by Greene (1994), the inflated zeros in ZIP models can masquerade as over-dispersion. However, the multiple excess counts in MIP models may induce either over-dispersion or under-dispersion. With the aid of z_{im}'s introduced earlier, it can be verified that

$$\begin{aligned} E(y_i) &= E\{E(y_i|z_{i0}, z_{i1}, \ldots, z_{iM})\} = \sum_{m=0}^{M-1} m \cdot p_{im} + \lambda_i p_{iM}, \qquad (2.9) \\ \mathrm{Var}\,(y_i) &= E\{\mathrm{Var}\,(y_i|z_{i0}, z_{i1}, \ldots, z_{iM})\} + \mathrm{Var}\,\{E(y_i|z_{i0}, z_{i1}, \ldots, z_{iM})\} \\ &= \sum_{m=0}^{M-1} m^2 p_{im}(1-p_{im}) + \lambda_i^2 p_{iM}(1-p_{iM}) + p_{iM}\lambda_i. \end{aligned}$$

Equation (2.9) is useful for prediction purpose. To gain insight, consider the special case $p_{im} \equiv p$ for $m = 0, 1, \ldots, (M-1)$. Here $p < 1/M$ and $p_{iM} = 1 - Mp$, so

$$\mathrm{Var}\,(y_i) = E(y_i) + \left\{ \frac{M\,p\,(M-1)\,(2M-2Mp+p-4)}{6} + \lambda_i Mp(1-Mp) \right\}.$$

A more detailed look indicates that underdispersion occurs, $\mathrm{Var}\,(y_i) < E(y_i)$, when and only when $M = 2$ and $\lambda_i < \sqrt{p/\{2(1-2p)\}}$. With this flexibility, the multiple-inflation Poisson model potentially supplies a competitive method for modeling count data with moderate over- or under-dispersion.

2.4. Identifiability

For finite mixture models, identifiability (see, e.g., Teicher (1961) and Wang et al. (1996)) is an issue that must be addressed before estimation of the involved parameters can be meaningfully discussed.

For the multiple-inflation Poisson model at (2.1)−(2.3), a formal definition is given. Let $\mathbf{G}_i^T$ and $\mathbf{B}_i^T$ denote the i-th row vector in matrix $\mathbf{G}$ and $\mathbf{B}$, respectively, for $i = 1, \ldots, n$. Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T,\ \boldsymbol{\beta}^T)^T$ be the vector containing all parameters.

Definition 1. For given covariate matrices $(\mathbf{B}, \mathbf{G})$, the MIP model is *identifiable* if, for any two sets of parameters $(\boldsymbol{\theta}, \boldsymbol{\theta}^\star)$, y_i has the same distribution for each $i = 1, \ldots, n$ implies that $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$.

Here $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ should be understood up to a permutation in the sense that exchanging the components in $\mathbf{G}_i^T\boldsymbol{\gamma}$ or $\mathbf{B}_i^T\boldsymbol{\beta}$ does not alter the model. The proof of the following proposition is given in the Supplement.

Proposition 1. *If the matrices* $\mathbf{G}$ *and* $\mathbf{B}$ *are both of full column rank, the multiple-inflation Poisson model at* (2.1)−(2.3) *is identifiable.*

3. Maximum Likelihood Estimation

With observed data, the likelihood function of MIP is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \prod_{i=1}^{n} \left\{ \prod_{m=0}^{M-1} \left(p_{im} + p_{iM} \cdot \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!} \right)^{\delta_{im}} \right\} \cdot \left(p_{iM} \cdot \frac{e^{-\lambda_i} \cdot \lambda_i^{y_i}}{y_i!} \right)^{\delta_{iM}}, \quad (3.1)$$

where $\delta_{im} = 1_{\{y_i = m\}}$ for $m = 0, 1, \ldots, (M-1)$, and $\delta_{iM} = 1_{\{y_i \geq M\}}$, the log-likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = \sum_{i=1}^{n} \left\{ \sum_{m=0}^{M-1} \delta_{im} \log \left(p_{im} + p_{iM} \frac{e^{-\lambda_i} \cdot \lambda_i^m}{m!} \right) + \delta_{iM} \log \left(p_{iM} \frac{e^{-\lambda_i} \cdot \lambda_i^{y_i}}{y_i!} \right) \right\} \quad (3.2)$$

where $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ enters the log-likelihood function via (2.3) and (2.5).

In order to obtain the maximum likelihood estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\gamma}}^T, \hat{\boldsymbol{\beta}}^T)$, Newton's method can be applied to maximize L. The gradient of L is provided in Appendix B of the Supplement. Since the Hessian matrix of $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is too complex to be explicitly available, a quasi-Newton method is used. In particular, the BFGS quasi-Newton method would be favored because the low-rank approximation to the Hessian matrix allows its inverse to be conveniently computed. When the number of parameters is very large, the limited-memory variant, L-BFGS (Nocedal (1980), can be used.

For estimating mixture models, the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin (1977)) is a popular alternative. Despite its slow convergence, the EM algorithm breaks down the estimation problems into familiar subcomponents for simpler statistical programming. The EM procedure can be carried out as follows. With z_{im} at (2.6), suppose that we observe both z_{im} and y_i, referred to as the complete data $(\mathbf{y}, \mathbf{z})$. The corresponding log-likelihood is

$$\begin{aligned} L_c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) &= \sum_{i=1}^{n} \sum_{m=0}^{M} \log f(z_{im}|\boldsymbol{\gamma}) + \sum_{i=1}^{n} \log f(y_i|z_i; \boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \sum_{m=0}^{M} z_{im} \log p_{im} + \sum_{i=1}^{n} z_{iM} \left\{ y_i \mathbf{B}_i^T \boldsymbol{\beta} - \exp(\mathbf{B}_i^T \boldsymbol{\beta}) \right\} \\ &\quad - \sum_{i=1}^{n} z_{iM} \log(y!) \\ &= L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z}) + L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) - \sum_{i=1}^{n} z_{iM} \log(y!). \end{aligned} \tag{3.3}$$

Note that $\boldsymbol{\gamma}$ only shows up in $L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z})$ through the p_{im}'s while $\boldsymbol{\beta}$ is only present in $L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z})$, both objective functions being concave. Thus estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ can be done separately, here alternating between the E step, computing the expected z_{im} given current estimates of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$, and the M step, maximizing $L_c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z})$ and $L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z})$ to update $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ given current estimates of the expected values of z_{im}.

Let $Z_{im} = E(z_{im}|\, y_i, \boldsymbol{\gamma}, \boldsymbol{\beta})$. A detailed description at the k-th iteration is as follows.

◇ *E Step.* Estimate Z_{im} given the current estimates $\hat{\boldsymbol{\gamma}}^{(k-1)}$ and $\hat{\boldsymbol{\beta}}^{(k-1)}$. We have

$$\begin{cases} Z_{im}^{(k)} = \dfrac{\delta_{im}\, p_{im}}{p_{im} + p_{iM}\, e^{-\lambda_i} \lambda_i^m / m!} & \text{for } m = 0, 1, \ldots, (M-1), \\ Z_{iM}^{(k)} = 1 - \sum_{m=0}^{M-1} Z_{im}^{(k)}, \end{cases} \tag{3.4}$$

where p_{im} and λ_i are updated via (2.3) and (2.5).

Replace z_{im} in (3.3) with $Z_{im}^{(k)}$ and let $\mathbf{Z}^{(k)} = (Z_{im}^{(k)})$.

◇ *M Step for $\boldsymbol{\beta}$.* Update $\hat{\boldsymbol{\beta}}^{(k)}$ by maximizing $L_c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{Z}^{(k)})$, equivalent to fitting a weighted Poisson regression model with responses $\mathbf{y}$ and weights $Z_{iM}^{(k)}$.

- ◇ *M Step for $\boldsymbol{\gamma}$.* Update $\hat{\boldsymbol{\gamma}}^{(k)}$ by maximizing $L_c(\boldsymbol{\gamma};\mathbf{y},\mathbf{Z}^{(k)})$. Note that $L_c(\boldsymbol{\gamma};\mathbf{y},\mathbf{Z}^{(k)})$ has a form similar to the log-likelihood obtained from a multinomial logistic model with responses $Z_{im}^{(k)}$'s, except that $Z_{im}^{(k)}$ is not binary. It can be rewritten into a weighted log-likelihood form of a multinomial logistic model. Maximization of $L_c(\boldsymbol{\gamma};\mathbf{y},\mathbf{Z}^{(k)})$ is straightforward by calling a Newton-Raphson routine.

The initial values for $\boldsymbol{\beta}$ can be supplied by the MLE from fitting a truncated Poisson model to data with $y_i \geq M$. Omitting irrelevant terms, the corresponding log-likelihood is

$$L_1(\boldsymbol{\beta};\mathbf{y}) \propto \sum_{i=1}^{n} \delta_{iM}\Big\{y_i \log\lambda_i - \lambda_i - \log\Big(1-\sum_{m=0}^{M-1}\frac{e^{-\lambda_i}\lambda_i^m}{m!}\Big)\Big\}, \qquad (3.5)$$

The score functions are given by $\mathbf{B}^T\mathbf{d} = \mathbf{0}$, where $\mathbf{d}$ is an $n \times 1$ vector with elements

$$d_i = \sum_{i=1}^{n} \delta_{iM}\Big\{(y_i-\lambda_i) - \frac{e^{-\lambda_i}\lambda_i^M/(M-1)!}{1-\sum_{m=0}^{M-1} e^{-\lambda_i}\lambda_i^m/m!}\Big\}.$$

As noted by Lambert (1992), starting values for $\boldsymbol{\gamma}$ are unimportant for ZIP. We suggest using the MLE from fitting a cumulative logit model with ordinal data in which every $y_i \geq M$ is revalued as the M-th ordered category.

The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, $\boldsymbol{\Sigma} = \mathrm{cov}(\hat{\boldsymbol{\theta}})$, can be estimated via the observed Fisher's information matrix $\hat{\boldsymbol{\Sigma}} = \{-\mathbf{H}\}^{-1}$, where $\mathbf{H} = \partial^2 L/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ is the Hessian matrix of the loglikelihood L. Note that the low-rank approximating matrix in BFGS may not converge to the real Hessian matrix when the algorithm stops. R function `optim()` offers a finite difference method for approximating the Hessian.

To sum up, we suggest a three-step procedure to fit the multiple-inflation Poisson (MIP) model: (i) obtain initial estimates of $(\boldsymbol{\beta},\boldsymbol{\gamma})$ by fitting a truncated Poisson model and a cumulative logit model, respectively; (ii) take a few runs ($2 \leq$ nrun ≤ 5) of the EM algorithm to update the estimates; (iii) run the BFGS quasi-Newton to convergence. Unless otherwise modified, the EM algorithm is slow but it helps, with only a few runs, for quickly locating a good neighborhood of $\boldsymbol{\theta}$ in search of a local optimum of L.

With the slightly more general MIP model in (2.7) and (2.8), the likelihood function is

$$l(\widetilde{\boldsymbol{\beta}},\widetilde{\gamma};\mathbf{y}) = \prod_{i=1}^{n}\Big[\prod_{m=0}^{M-1}\Big\{(1-p_{iM})p'_{im}+p_{iM}\frac{e^{-\lambda_i}\cdot\lambda_i^m}{m!}\Big\}^{\delta_{im}}\cdot\Big(\frac{e^{-\lambda_i}\cdot\lambda_i^{y_i}}{y_i!}\Big)^{\delta_{iM}}\Big], \qquad (3.6)$$

where $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\gamma}_{00}, \ldots, \widetilde{\gamma}_{M0}, \widetilde{\boldsymbol{\gamma}}_1, \widetilde{\boldsymbol{\gamma}}_2)^T$ and parameters $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}})$ enter through specifications at (2.8). Optimization of the log-transformed $l(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\gamma}}; \mathbf{y})$ in (3.6) can be done via the BFGS quasi-Newton without added difficulties.

4. Variable Selection via Regularization

Another major obstacle for using ZIP models is the problem of variable selection, since there are two separate model components whose contributions to the count response affect each other. Existing methods such as all-subset selection or stepwise procedures are computationally prohibitive, especially when the number of covariates is large. The same difficulty exists in fitting MIP models. To tackle the problem, Buu et al. (2011) recently introduced the L_1 regularization (Tibshirani (1996)) to ZIP. In this section, we consider a more flexible L_1 regularization method for fitting MIP.

To proceed, assume that each variable X_j has been standardized to have mean 0 and standard deviation 1. Unlike in linear regression, this standardization step does not eliminate the intercepts in MIP model. Here we distinguish intercepts from slopes in $\boldsymbol{\theta}$ by exchanging the positions of its components so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_0^T, \boldsymbol{\theta}_1^T)^T$, where $\boldsymbol{\theta}_0 = (\gamma_{00}, \gamma_{10}, \ldots, \gamma_{(M-1)0}, \beta_0)^T$ contains all intercept terms and $\boldsymbol{\theta}_1 = (\theta_{1j})$ all slopes.

With MIP model either at (2.2)–(2.3) or at (2.8), L_1 regularization solves

$$\min_{\boldsymbol{\theta}} \quad -L(\boldsymbol{\theta}) + \sum_j \lambda_j \left|\theta_{1j}\right|, \tag{4.1}$$

where $L(\boldsymbol{\theta})$ is the log-likelihood function and $\lambda_j \geq 0$ are the tuning parameters. Note that the penalty is only applied to the slopes in $\boldsymbol{\theta}_1$. The solution to (4.1) can be conveniently obtained via a local quadratic approximation to $L(\boldsymbol{\theta})$. Given some initial value $\widetilde{\boldsymbol{\theta}}$ that is close to the minimizer of (4.1),

$$L(\boldsymbol{\theta}) \approx Q(\boldsymbol{\theta}) \;=\; L(\widetilde{\boldsymbol{\theta}}) + \widetilde{\mathbf{g}}^T(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) + \frac{1}{2} \cdot (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \widetilde{\mathbf{H}} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}), \tag{4.2}$$

where

$$\widetilde{\mathbf{g}} = \dot{L}(\widetilde{\boldsymbol{\theta}}) = \left.\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \quad \text{and} \quad \widetilde{\mathbf{H}} = \ddot{L}(\widetilde{\boldsymbol{\theta}}) = \left.\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}}$$

are the gradient and the Hessian matrix of $L(\boldsymbol{\theta})$ evaluated at $\widetilde{\boldsymbol{\theta}}$. Replacing $L(\boldsymbol{\theta})$ in (4.1) by the RHS of (4.2) yields a quadratic programming problem

$$\min_{\boldsymbol{\theta}} \quad -\widetilde{\mathbf{g}}^T(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) - \frac{1}{2} \cdot (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})^T \widetilde{\mathbf{H}} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) + \sum_j \lambda_j \left|\theta_{1j}\right|, \tag{4.3}$$

where the irrelevant term $L(\widetilde{\boldsymbol{\theta}})$ has been omitted. Optimization in (4.3) can be done conveniently using the efficient LARS algorithm (Efron et al. (2004)). Since only the slopes in $\boldsymbol{\theta}_1$ are penalized, some technical details for handling this issue are given in Appendix C of the Supplement.

For the choices of λ_j, it is convenient to write $\lambda_j = \lambda\, a_j$ for $\lambda \geq 0$ and $a_j > 0$. Zou (2006)'s adaptive lasso sets $a_j = 1/\left|\widetilde{\theta}_{1j}\right|$; we use this in our simulations and data analyses. Another choice is to use $a_j = \dot{\rho}_j(|\widetilde{\theta}_{1j}|)$, where $\rho_j(\cdot)$ is the smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001). This amounts to the one-step sparse estimator of Zou and Li (2008) where a local linear approximation is applied to approximate the SCAD penalty. Yet another possibly is to adopt Khalili and Chen (2007), who studied the L_1 regulation for finite mixture models (FMM) with constant mixing probabilities. With this approach, an additional proportion factor, that corresponds to the mixing probability, can be assigned to the penalties of β's since they are present only in the Poisson model component. The penalty function has the form $\sum_j \lambda_j|\gamma_{1j}| + \sum_j p_{\cdot M}\, \lambda_j\, |\beta_{1j}|$, where $p_{\cdot M} = \sum_{i=1}^{n} p_{iM}$ has to be estimated by plugging in an initial $\widetilde{\boldsymbol{\theta}}$. Selection of λ can be tuned via some model selection criterion such as AIC (Akaike (1974)), BIC (Schwarz (1978)), or generalized cross-validation (GCV).

As for the initial estimator $\widetilde{\boldsymbol{\theta}}$, one common choice is the MLE from the full MIP model where all predictors are included in both the cumulative logit model and the Poisson model. In this case, the gradient vector $\hat{\mathbf{g}} = \mathbf{0}$. Thus (4.3) can be further simplified as

$$\min_{\boldsymbol{\theta}} \quad -\frac{1}{2} \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\mathbf{H}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \sum_j \lambda_j\, |\theta_{1j}|\,, \tag{4.4}$$

With this approach, the so-called 'oracle' property, consisting of both consistency in variable selection and consistent estimation of non-zero coefficients, can be established, following arguments in either Wang and Leng (2007) or Zou and Li (2008). On the other hand, this approach requires that MLE $\hat{\boldsymbol{\theta}}$ be available for the full model. This however is not the case when severe multi-collinearity exists or the MIP dimension $(2p + M + 1)$ is higher than n. A reasonable solution is use an L_2-regularized (ridge) estimator for $\widetilde{\boldsymbol{\theta}}$ instead. In this case, the first-order term in (4.3) remains. L_2 regularization is easier to solve, often with closed-form solutions available.

5. Simulation

This section reports on simulated experiments designed to evaluate MIP model estimation, compare MIP with other count models, and assess the performance of L_1 regularization in selecting variables. The data were generated from the following MIP model with $M = 2$:

$$\text{Model A:} \begin{cases} \text{Cumulative Logit: } \operatorname{logit}\{\Pr(Y \le 0)\} = \gamma_{00} + \gamma_1 X_1 + \gamma_2 X_3, \\ \qquad\qquad\qquad\quad \operatorname{logit}\{\Pr(Y \le 1)\} = \gamma_{01} + \gamma_1 X_1 + \gamma_2 X_3, \\ \text{Loglinear:} \qquad\quad \log(\lambda) \qquad\qquad\quad = \beta_0 + \beta_1 X_2 + \beta_2 X_3, \end{cases} \tag{5.1}$$

where $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{01}, \gamma_1, \gamma_2)^T = (-3, -1.5, 3, 2)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T = (-2, 3, 2)^T$. Each data set includes five covariates $(X_1, X_2, X_3, X_4, X_5)$ that were independently generated as uniform $(0, 1)$. However, only X_1 and X_3 are effective in the cumulative logit model component, while X_2 and X_3 are effective in the loglinear model component. Two sample sizes $n = 100$ and $n = 300$ were considered, and, for each model configuration, a total of $1,000$ runs were made.

To evaluate the ML estimation, we fit the true MIP model in (5.1) to each simulated data set and obtained the estimated parameters and their standard errors. Table 1(a) reports the sample mean and sample standard deviation (SD) of the parameter estimates as well as the sample mean of the standard error estimates over 1,000 simulation runs. Here, the finite difference approximation implemented in the R (R Development Core Team (2012)) function `optim()` was used to compute the Hessian matrix and hence the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. Note that the standard error is essentially an asymptotic approximation to the standard deviation of the sampling distribution of the MLE. Thus a comparison of the SD column with the standard error column in Table 1(a) allows us to evaluate the performance of the asymptotic results. In addition, we computed the coverage rate of the 95% Wald confidence interval (CI) that is presented in the last column.

From the results, the maximum likelihood procedure seems to work very well in estimating the true parameter and providing confidence interval coverage within the nominal levels. The average estimates are reasonably close to their true values. The average standard errors are also close to the sample SDs obtained from the simulation runs. It is interesting to see that the standard errors for $\boldsymbol{\beta}$ (in the loglinear model component) are generally smaller than those of $\boldsymbol{\gamma}$ in the cumulative logit model component. This is because relatively more observations in each category are needed in order to achieve high accuracy and high precision in estimating the cumulative logit model. It is not surprising that the performance improves with the increased sample size.

Next, we compared MIP with four other choices of count models: loglinear, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) regression models. To do so, we generated an independent test sample beforehand, which consisted of $n_1 = 500$ observations. With each

Table 1. Fitting MIP Based on 1,000 Runs: (a) Maximum likelihood estimation; (b) Comparison with other models. In each simulation run, data were generated according to (5.1). The Average and SD columns give the sample mean and standard deviation of estimates over 1,000 runs for each parameter. The Averaged SE column gives the standard errors averaged over simulations runs. The 95% CI coverage rate is supplied in the last column.

(a) ML Estimation

Sample		True	Estimates		Averaged	95% CI
Size	Parameter	Value	Average	SD	SE	Coverage
$n = 100$	γ_{00}	−3.0	−3.335	1.131	1.078	0.938
	γ_{01}	−1.5	−1.743	1.073	1.029	0.950
	γ_1	3.0	3.249	1.137	1.083	0.950
	γ_2	2.0	2.224	1.221	1.153	0.940
	β_0	−2.0	−2.100	0.606	0.611	0.969
	β_1	3.0	3.060	0.680	0.663	0.956
	β_2	2.0	2.042	0.606	0.576	0.940
$n = 300$	γ_{00}	−3.0	−3.115	0.609	0.582	0.948
	γ_{01}	−1.5	−1.587	0.638	0.559	0.946
	γ_1	3.0	3.114	0.628	0.588	0.936
	γ_2	2.0	2.064	0.644	0.621	0.945
	β_0	−2.0	−2.031	0.327	0.326	0.948
	β_1	3.0	3.022	0.354	0.354	0.948
	β_2	2.0	2.012	0.310	0.300	0.942

(b) Averaged Squared Loss (ASL)

Sample	Models				
Size (n)	MIP	Loglinear	NB	ZIP	ZINB
100	0.142	0.204	0.179	0.230	0.235
300	0.053	0.092	0.091	0.109	0.129

simulation data set, we fit all four count models and then applied the fitted model to predict the counts in the test sample. The average squared loss (ASL)

$$\mathrm{ASL} = \sum_{i=1}^{n_1} \{\mathrm{E}(y_i) - \hat{y}_i\}^2 \qquad (5.2)$$

was recorded. Equation (2.9) is used for prediction with an MIP model; thus ASL provides a direct assessment. To make it fair, $\{X_1, X_2, X_3\}$ were used in fitting the three comparison model. Table 1(b) presents the averaged ASL over 1,000 runs, together with the associated standard deviation. As expected, MIP has the smallest ASL.

Finally, we investigated the L_1 regularization in variable selection with MIP. In this simulation, the MLE from fitting the full MIP model was used in the

Table 2. Variable Selection via L_1 Regularization Based on 1,000 Runs. Five covariates $\{X_1, X_2, X_3, X_4, X_5\}$ were independently uniform (0,1). Model A corresponds to the confounded case where X_1 and X_3 are involved in the cumulative logit model, while X_2 and X_3 are present in the log-linear model. Model B corresponds to the unconfounded case where X_1 and X_3 are involved in the cumulative logit model, while X_2 and X_4 are present in the log-linear model.

	Sample	Selection	Correct Selections %			mean ASL	
Model	Size (n)	Criterion	γ's	β's	both	selected	true
A	100	AIC	0.379	0.485	0.201	0.260	0.157
		BIC	0.340	0.698	0.247	0.401	
	200	AIC	0.607	0.571	0.371	0.134	0.146
		BIC	0.577	0.868	0.488	0.194	
	300	AIC	0.569	0.593	0.387	0.108	0.055
		BIC	0.668	0.854	0.593	0.150	
B	100	AIC	0.424	0.552	0.266	0.429	0.203
		BIC	0.423	0.800	0.342	0.565	
	200	AIC	0.587	0.605	0.397	0.145	0.100
		BIC	0.694	0.890	0.613	0.149	
	300	AIC	0.635	0.657	0.477	0.088	0.064
		BIC	0.848	0.906	0.784	0.150	

quadratic approximation and the adaptive lasso penalty was used. We took sample sizes $n = 100$, 200, and 300, and two model selection criteria, AIC and BIC. In addition to Model A, we considered model B which has the same cumulative logit model as Model A in (5.1), but X_3 in the loglinear model is replaced with X_4. Model B is a scenario in which the covariates in the two model components are not confounded. Table 2 reports the proportions of correct selections in the cumulative logit model component, the loglinear model components, and both. This helps with the consistency assessment in model selection. At the same time, we recorded the average squared loss (ASL) given in (5.2), commonly used for assessing model selection criterions in terms of efficiency (see, e.g., Shao (1997)). For comparison, we also computed the ASL from the true model.

Several observations are in order. First of all, variable selection is more difficult with MIP, compared to e.g., simulation results reported on linear regression in the literature (McQuarrie and Tsai (1998)). However, the results are generally consistent with the L_1 regularized ZIP as recently reported by Buu et al. (2011). Secondly, variable selection is more difficult for the cumulative logit model component than for the loglinear model component. The selection performance improves when the covariates in the two model components are not confounded. Thirdly, the performance improves with increased sample sizes as expected. In general, both AIC and BIC work better with the unconfounded

Table 3. Variable Description for Health Visit Data.

Var	Name	Description
1	`SEX`	1 if female and 0 if male
2	`AGE`	Age in years divided by 100
3	`INCOME`	Annual income in Australian dollars divided by 1,000
4	`HSCORE`	General health questionnaire score using Goldberg's method, with high score indicating bad health.
5	`CHCOND1`	1 if chronic condition(s) but not limited in activity, 0 otherwise
6	`CHCOND2`	1 if chronic condition(s) and limited in activity, 0 otherwise
7	`PREVIATE.INS`	1 if covered by private insurance company, 0 covered by government
8	`HVISITS`	Number of visits to doctors or health professionals in the past two weeks.

Model B in terms of variable selection, but show mixed results with the average squared loss. Fourthly, BIC considerably outperforms AIC in terms of correct variable selection. This can be explained by the moderately large sample sizes and the relatively strong signals in the models we considered and the consistent selection criterion BIC. On the other hand, AIC compares favorably to BIC in terms of average squared loss, since AIC is an efficient model selection criterion.

6. Data Analysis Example

To illustrate, we compiled a data set from `Racd3.asc` of Cameron and Trivedi (1998), available at `http://cameron.econ.ucdavis.edu/racd/racddata.html`. The data set contains 5,190 observations and eight variables extracted from the ABS (Australian Bureau of Statistics) 1977-78 Health survey and restricted to single people over 18 years of age. Table 3 provides a brief description of these variables. More details about the data set can be found in Cameron and Trivedi (1998) and references listed there. The objective is to establish a predictive model for `HVISITS`, the frequency of visits to doctors and/or health professionals in the past two weeks.

Figure 1 plots the histogram of `HVISITS`, which shows excess zeros and ones. Also superimposed on the histogram are frequency distributions from a fitted Poisson, zero-inflated Poisson, and Poisson model with two inflations at 0 and 1. These fitted distributions are obtained in the spirit of goodness-of-fit involved in, e.g., normality tests, without taking covariates into account. Specifically, the Poisson distribution component is obtained by setting its mean as the sample average of counts after excluding the pre-specified inflated values. Then we simulate counts, of the same number as the sample size, from each fitted model

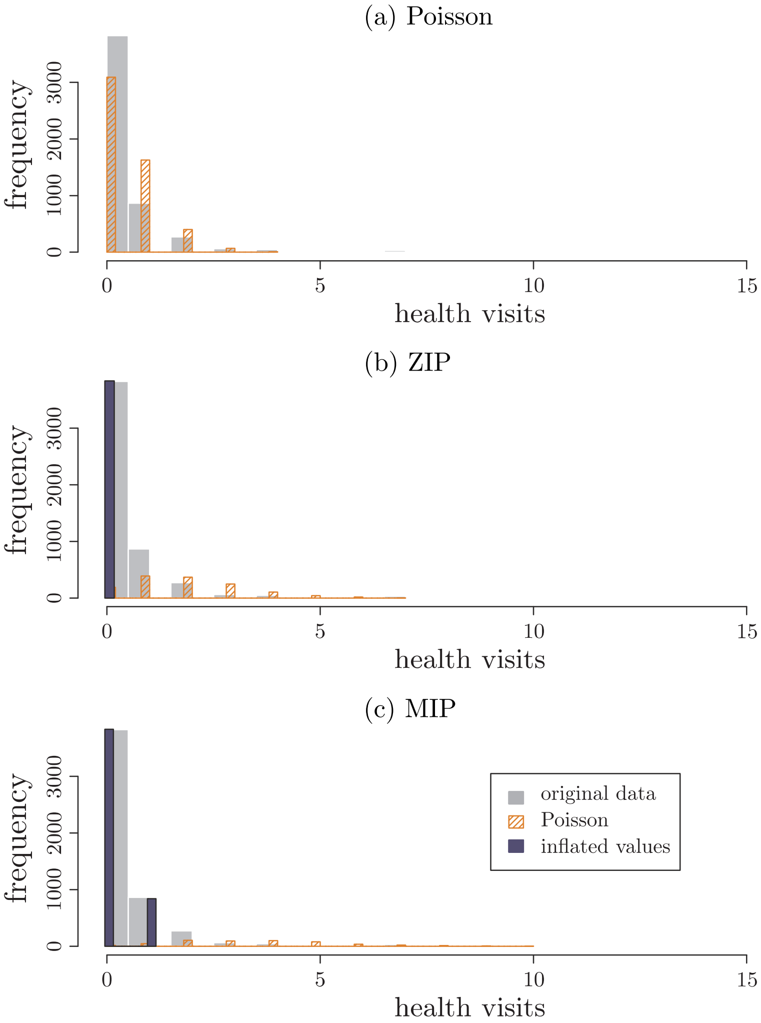

Figure 1. Frequency Distribution of Health Visits to Medical Doctors and Health Professionals, Superimposed with Frequency Distributions from Different Good-of-Fit Models: (a) Poisson Model; (b) Zero-Inflated Poisson (ZIP); (c) Multiple-Inflation Poisson (MIP) with $M = 2$.

distribution. It can be seen that the Poisson model does not fit well to data. ZIP would miss out considerably on the prediction of ones. Comparatively, the multiple-inflation Poisson model with $M = 2$ is a better choice.

Table 4(a) provides the fitting results of the full MIP model. Next, this full model was refit with standardized covariates. The estimated parameters, together with the estimated variance-covariance matrix, were then used for quadratic approximation in L_1 regularization. The resultant regularization path is plotted

Table 4. Fitting the MIP Model ($M = 2$) with the Health Visit Data: (a) The fitted full MIP model and the best model selected via L_1 regularization; (b) Mean Squared Errors (MSE) computed via 3-fold cross validation (CV) for competitive models.

(a) The Full and best MIP Model

			Full Model			L_1-Regularized Model		
	Variable		Estimate	S.E.	P-Value	Estimate	S.E.	P-Value
Cumulative	`INTERCEPT`	γ_{00}	2.393	0.146	0.0000	2.323	0.088	0.0000
Logit		γ_{01}	3.452	0.155	0.0000	3.382	0.103	0.0000
	`SEX`	γ_1	−0.338	0.073	0.0000	−0.341	0.070	0.0000
	`AGE`	γ_2	−1.431	0.198	0.0000	−1.361	0.176	0.0000
	`INCOME`	γ_3	0.011	0.109	0.9203	—	—	—
	`HSCORE`	γ_4	−0.158	0.014	0.0000	−0.157	0.014	0.0000
	`CHCOND1`	γ_5	−0.471	0.079	0.0000	−0.477	0.078	0.0000
	`CHCOND2`	γ_6	−1.006	0.105	0.0000	−1.002	0.104	0.0000
	`PRIVATE.INS`	γ_7	−0.066	0.090	0.4664	—	—	—
Log-Linear	`INTERCEPT`	β_0	0.936	0.129	0.0000	0.884	0.104	0.0000
	`SEX`	β_1	0.024	0.062	0.7005	—	—	—
	`AGE`	β_2	0.280	0.165	0.0902	0.388	0.145	0.0074
	`INCOME`	β_3	−0.215	0.098	0.0287	−0.256	0.090	0.0044
	`HSCORE`	β_4	0.031	0.009	0.0010	0.032	0.009	0.0008
	`CHCOND1`	β_5	0.033	0.079	0.6784	—	—	—
	`CHCOND2`	β_6	0.291	0.085	0.0007	0.282	0.062	0.0000
	`PRIVATE.INS`	β_7	−0.086	0.070	0.2221	—	—	—

(b) MSE via 3-fold CV

MIP	Loglinear	NB	ZIP	ZINB
1.638	1.648	1.674	1.646	1.694

in Figure 2. The best MIP model, selected by minimum BIC, is also presented in the right panel of Table 4 (a). `SEX`, `AGE`, `HSCORE`, `CHCOND1`, and `CHCOND2` shows up significantly in the cumulative logit model while `AGE`, `INCOME`, `HSCORE`, and `CHCOND2` are selected in the loglinear Poisson model component.

The final model selected via L_1 regularization seems very interpretable. In the cumulative logit model, the regression coefficient corresponds to the amount of change in the ordered logit scale of the dependent variable level, $\log\{P(Y \leq m)/P(Y > m)\}$, with a one-unit increase in the predictor while holding constant the other variables in the model. This explains why the signs of $\hat{\gamma}_4$ and $\hat{\beta}_4$ for `HSCORE` are opposite in the two model components. It is worth noting that the two binary variables `CHCOND1` and `CHCOND2` correspond to prevalence and severity of chronic conditions. It is interesting to see that both variables show up in the cumulative logit model while only `CHCOND2` shows up in the log-linear model. This might be interpreted as the fact that both prevalence and severity

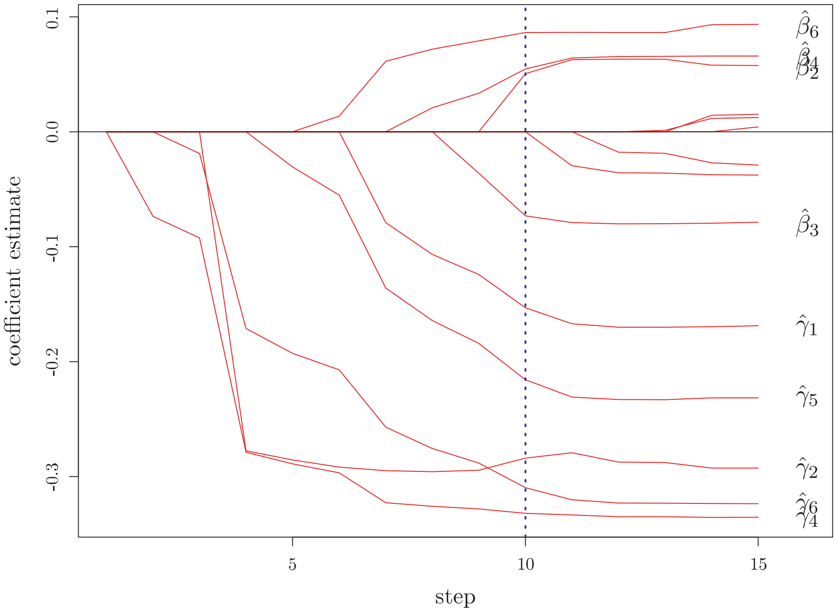

Figure 2. Regularization Path of Adaptive LASSO with Quadratic Approximation: Variable Selection for Fitting MIP with the Health Care Data. The dotted line highlights the best model choice via minimum BIC.

of chronic conditions are important predictors of whether or not a person would have a health visit in the past two weeks and severer chronic conditions are more likely to induce multiple health visits.

For comparison purposes, we also fit the loglinear, NB, ZIP, and ZINB models. The stepwise procedure was used for selecting variables. For brevity, the fitted model are omitted from the presentation. Among these four model choices, Vuong's test (1989), performed in a pairwise manner, strongly indicates that ZIP is the best. In order to make comparison with the multiple-inflation Poisson model, a $V-$fold cross-validation method was employed. In this approach, the sample was randomly divided into $V = 3$ folds. Observations in each fold were predicted via the model fitted with the remaining data. Figure 3 plots the density curve of the fitted values from each best-fit model, as opposed to the histogram of the original data. The overall model performance is summarized by a mean squared error (MSE) quantity, $\text{MSE} = \sum_{i=1}^{n}\{y_i - \hat{y}_i^{(\text{cv})}\}^2$, where $\hat{y}_i^{(\text{cv})}$ denotes the cross-validated prediction for y_i. The results, as given in Table 4(b), clearly suggest preference for the MIP model in this predictive task.

7. Discussion

The L_1 regularized MIP model offers enhanced flexibility and feasibility for

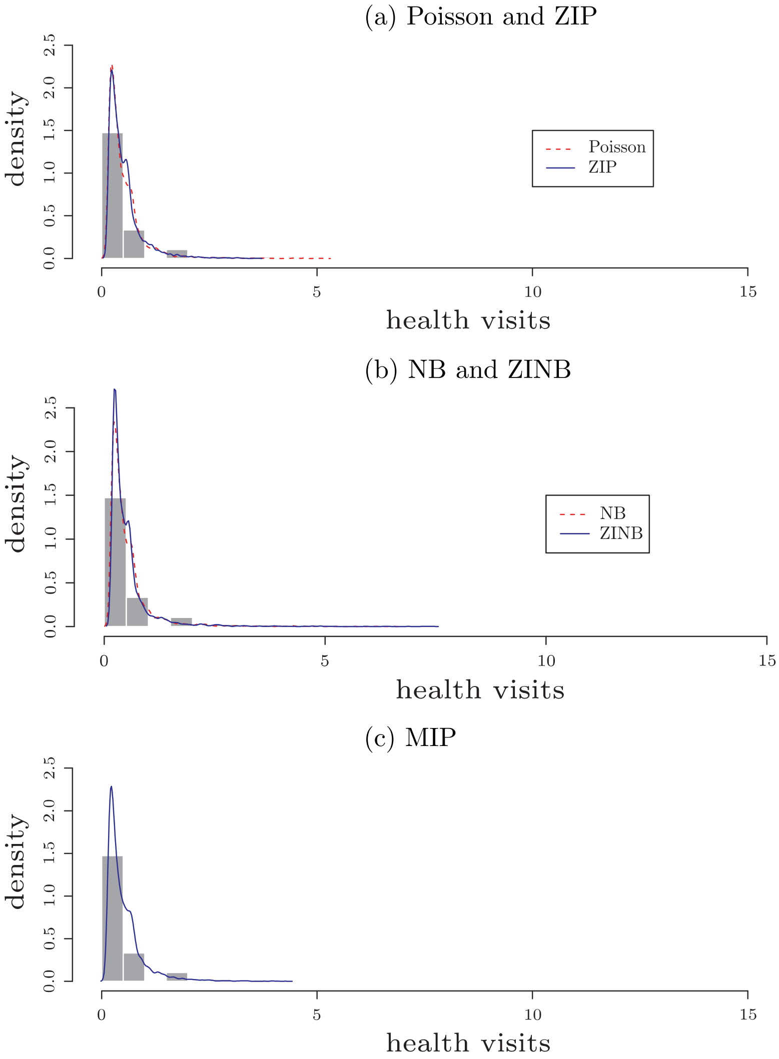

Figure 3. Density distribution of Health Visits to Medical Doctors and Health Professionals, Superimposed with the smoothed densities of fitted values obtained from the best-fit Poisson, ZIP, NB, ZINB, and MIP models.

modeling count data with multiple inflations and selecting variables efficiently. The multiple-inflation Poisson model is essentially a finite mixture model of multinomial and Poisson distributions. The likelihood function associated with finite mixture models is typically not log-concave, which implies that the fitting algorithm may be trapped into some local maximum. Thus trying out multiple starting points or resorting to global optimization techniques could be helpful. The R codes for fitting L_1 regularized MIP models are provided as part of the online supplemental material, available in *Statistica Sinica* website.

The multiple-inflation Poisson model supplies several immediate interesting avenues for future research. First, determination of M and the inflated values may be more strictly calibrated. Vuong's test (1989) can be extended to compare multiple-inflation Poisson models with other model choices. Also, the Poisson model component in MIP can be replaced with a negative binomial model for data with over-dispersion. Extensions of MIP to dependent data can be studied as well.

Acknowledgement

We are grateful to the Editor, an associate editor, and an anonymous referee for their helpful and constructive comments that have greatly improved an earlier draft.

References

Akaike, H. (1974). A new look at model identification. *IEEE Trans. Automat. Control* **19**, 716-723.

Baker, S. G. (1994). The multinomial-Poisson transformation. *The Statistician* **43**, 495-504.

Buu, A., Johnson, N. J., Li, R., and Tan, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statist. medicine* **30**, 2326-2340.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-489.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Department of Economics, Stern School of Business, New York University.

Khalili, A. and Chen, J. H. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102**, 1025-1038.

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

Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *Ann. Statist.* **32**, 340-383.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* **37**, 35-46.

Lord, D., Washington, S. P., and Ivan, J. N. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis and Prevention* **39**, 53-57.

McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42**, 109-42.

McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific, Singapore.

Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Math. Computation* **35**, 773-782.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221-264.

Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32**, 244-248.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, H. and Leng, C. (2007). Unified LASSO estimation by least squares approximation. *J. Amer. Statist. Assoc.* **102**, 1039-1048.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics* **52**, 381-400.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307-333.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1566.

School of Nursing, University of Alabama, Birmingham, AL 35294, U.S.A.

E-mail: xgsu@uab.edu

Department of Mathematics and Statistics, San Diego State University, CA 92182, U.S.A.

E-mail: jjfan@sciences.sdsu.edu

Department of Mathematics and Statistics, San Diego State University, CA 92182, U.S.A.

E-mail: ralevine@sciences.sdsu.edu

McGill University Health Center, Montreal, Quebec, Canada H3H 2R9.

E-mail: Xianming.Tan@clinepi.mcgill.ca

Department of Biostatistics, University of Alabama, Birmingham, AL 35294, U.S.A.

E-mail: arvind.biostat@gmail.com

(Received June 2012; accepted August 2012)

Statistica Sinica: Supplement

MULTIPLE-INFLATION POISSON MODEL WITH L_1 REGULARIZATION

Xiaogang Su[1], Juanjuan Fan[2], Richard A. Levine[2],Xianming Tan[3], and Arvind Tripathi[1]

University of Alabama at Birmingham[1]*, San Diego State University*[2]*, and McGill University*[3]

Supplementary Material

Several technical details are included in Sections this supplementary material. Besides, the R codes written for fitting the proposed MIP model with L_1 regularization are also supplied.

S1 Proof of Proposition 1

Write the distribution of y_i in a mixture model form $\sum_{m=0}^{M} p_m f_m(y_i)$, where $f_m(y_i)$ is the density function of either Poisson(λ_i) for $m = M$ or the degenerate distribution otherwise. Teicher (1961) establishes the identifiability for arbitrary (and hence finite) mixtures of Poisson and degenerate distributions without covariates. Suppose that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\star$ give the same density, that is, $\sum_{m=0}^{M} p_m f_m(y) = \sum_{m=0}^{M^\star} p_m^\star f_m^\star(y)$ for any y. Teicher's (1961) results imply that $M = M^\star$, $p_m = p_m^\star$, and $\lambda = \lambda^\star$. It follows that $\mathbf{G}_i^T \boldsymbol{\gamma} = \mathbf{G}_i^T \boldsymbol{\gamma}^\star$ and $\mathbf{B}_i^T \boldsymbol{\beta} = \mathbf{B}_i^T \boldsymbol{\beta}^\star$ for $i = 1, \ldots, n$. Equivalently, $\mathbf{G}\,(\boldsymbol{\gamma} - \boldsymbol{\gamma}^\star) = \mathbf{B}\,(\boldsymbol{\beta} - \boldsymbol{\beta}^\star) = \mathbf{0}$. However, both $\mathbf{G}$ and $\mathbf{B}$ are of full column rank. We must have $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$. ▮

S2 The Gradient $\mathbf{g}$ and $\mathbf{g}_i$

Denote $\pi_{im} = \text{expit}(\mathbf{G}_i^T \boldsymbol{\gamma}_1 + \gamma_{m0})$, $\pi'_{im} = e^{-\lambda_i} \lambda_i^m / m!$, and $A_{im} = p_{im} + p_{iM}\, \pi'_{im}$ for $m = 0, 1, \ldots, (M-1)$. Let $\mathbf{d}_{1i}$ be an M-dimensional vector with elements

$$\begin{aligned}
d_{1im} &= \pi_{im}(1-\pi_{im}) \cdot \left\{ \frac{\delta_{im}}{A_{im}} - \frac{\delta_{i(m+1)}}{A_{i(m+1)}} \right\} \text{ for } m = 0, 1, \ldots, (M-2), \\
d_{1i(M-1)} &= \pi_{i(M-1)}(1-\pi_{i(M-1)}) \cdot \left\{ -\sum_{m=0}^{M-2} \delta_{im} \frac{\pi'_{im}}{A_{im}} + \delta_{i(M-1)} \frac{(1-\pi'_{i(M-1)})}{A_{i(M-1)}} - \frac{\delta_{iM}}{p_{iM}} \right\}.
\end{aligned}$$

Let $\mathbf{d}_2$ be an n-dimensional vector with elements

$$\begin{aligned} d_{2i} \;&=\; \delta_{i0} \cdot \frac{\pi_{i0}(1-\pi_{i0}) - \pi_{i(M-1)}(1-\pi_{i(M-1)})\,\pi'_{i0}}{A_{i0}} \;-\; \delta_{iM} \cdot \pi_{i(M-1)} \\ &\quad+\; \sum_{m=1}^{M-1} \delta_{im} \cdot \frac{\pi_{im}(1-\pi_{im}) - \pi_{i(m-1)}(1-\pi_{i(m-1)}) - \pi_{i(M-1)}(1-\pi_{i(M-1)})\pi'_{im}}{A_{im}}, \end{aligned}$$

for $i = 1, \ldots, n$. Let $\mathbf{d}_3$ be an n-dimensional vector with elements

$$d_{3i} \;=\; \sum_{m=0}^{M-1} \delta_{im} \cdot \frac{\pi'_{im}\,(m-\lambda_i)}{A_{im}} \;+\; \delta_{iM} \cdot (y_i - \lambda_i).$$

Denoting $L = \sum_{i=1}^{n} L_i$, let $\mathbf{g} = \dot{L}(\boldsymbol{\theta}) = \partial L/\partial \boldsymbol{\theta}$ and $\mathbf{g}_i = \dot{L}_i(\boldsymbol{\theta}) = \partial L_i/\partial \boldsymbol{\theta}$ for $i = 1, \ldots, n$. Then $\mathbf{g}$ and $\mathbf{g}_i$ are given by

$$\mathbf{g_i} = \begin{pmatrix} \mathbf{d}_{1i} \\ d_{2i}\,\mathbf{G}_i \\ d_{3i}\,\mathbf{B}_i \end{pmatrix} \qquad \text{and} \qquad \mathbf{g} = \sum_{i=1}^{n} \mathbf{g}_i = \begin{pmatrix} \sum_{i=1}^{n} \mathbf{d}_1 \\ \mathbf{G}^T\mathbf{d}_2 \\ \mathbf{B}^T\mathbf{d}_3 \end{pmatrix}.$$

Despite the tedious expressions, involvement of the indicator functions δ_{im}'s greatly simplifies the computation of $\mathbf{g}$ and $\mathbf{g}_i$. Note that the reason we write out $\mathbf{g}_i$ separately is because they supply alternative ways of computing the variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$, albeit not pursued in our implementation. Specifically, one way is via the inverse of the outer product of gradients (OPG), $\hat{\boldsymbol{\Sigma}} = \left(\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i^T\right)^{-1}$, where $\hat{\mathbf{g}}_i$ denotes the $\mathbf{g}_i$ value when evaluated at MLE $\hat{\boldsymbol{\theta}}$. This result can be combined with the Hessian for the sandwich-typed estimator of $\boldsymbol{\Sigma}$. ▮

S3 Details for Solving the L_1 Regularization in Section 4

After position exchange in the components of $\boldsymbol{\theta}$, partition $\widetilde{\mathbf{g}}$ and $\widetilde{\mathbf{H}}$ as

$$\widetilde{\mathbf{g}} = \begin{bmatrix} \widetilde{\mathbf{g}}_0 \\ \widetilde{\mathbf{g}}_1 \end{bmatrix} \;\text{ and }\; \widetilde{\mathbf{H}} = \begin{bmatrix} \widetilde{\mathbf{H}}_{00} & \widetilde{\mathbf{H}}_{01} \\ \widetilde{\mathbf{H}}_{10} & \widetilde{\mathbf{H}}_{11} \end{bmatrix}.$$

Equation (18) in Section 4 becomes

$$\begin{aligned} -\,\widetilde{\mathbf{g}}_0^T(\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}_0) - \widetilde{\mathbf{g}}_1^T(\boldsymbol{\theta}_1 - \widetilde{\boldsymbol{\theta}}_1) - \frac{1}{2}\,(\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}_0)^T\widetilde{\mathbf{H}}_{00}(\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}_0) \;-\; (\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}_0)^T\widetilde{\mathbf{H}}_{01}(\boldsymbol{\theta}_1 - \widetilde{\boldsymbol{\theta}}_1) \\ -\,\frac{1}{2}\,(\boldsymbol{\theta}_1 - \widetilde{\boldsymbol{\theta}}_1)^T\widetilde{\mathbf{H}}_{11}(\boldsymbol{\theta}_1 - \widetilde{\boldsymbol{\theta}}_1) + \sum_j \lambda_j|\theta_{1j}|. \qquad \text{(S3.1)} \end{aligned}$$

Given $\boldsymbol{\theta}_1$, solving (S3.1) for $\boldsymbol{\theta}_0$ yields

$$\boldsymbol{\theta}_0^{\star} = \widetilde{\boldsymbol{\theta}}_0 + \widetilde{\mathbf{H}}_{00}^{-1}\left\{\widetilde{\mathbf{g}}_0 + \widetilde{\mathbf{H}}_{01}(\boldsymbol{\theta}_1 - \widetilde{\boldsymbol{\theta}}_1)\right\}. \qquad \text{(S3.2)}$$

Bringing $\boldsymbol{\theta}_0^\star$ into (S3.1) leads to a profile objective function for $\boldsymbol{\theta}_1$

$$-\frac{1}{2}\,(\boldsymbol{\theta}_1-\widetilde{\boldsymbol{\theta}}_1)^T\left(\widetilde{\mathbf{H}}_{11}-\widetilde{\mathbf{H}}_{10}\widetilde{\mathbf{H}}_{00}^{-1}\widetilde{\mathbf{H}}_{01}\right)(\boldsymbol{\theta}_1-\widetilde{\boldsymbol{\theta}}_1)-\left(\widetilde{\mathbf{g}}_1-\widetilde{\mathbf{H}}_{10}\widetilde{\mathbf{H}}_{00}^{-1}\widetilde{\mathbf{g}}_0\right)^T(\boldsymbol{\theta}_1-\widetilde{\boldsymbol{\theta}}_1)+\sum_j\lambda_j|\theta_{1j}|. \tag{S3.3}$$

Recall $\lambda_j=\lambda\,c_j$. Let $\mathbf{C}=\text{diag}(c_j)$. Denote $\vartheta_{1j}=\theta_{1j}\,c_j$ or, in matrix form, $\boldsymbol{\theta}_1=\mathbf{C}^{-1}\boldsymbol{\vartheta}_1$. Let

$$\mathbf{A}=-\mathbf{C}^{-1}\left(\widetilde{\mathbf{H}}_{11}-\widetilde{\mathbf{H}}_{10}\widetilde{\mathbf{H}}_{00}^{-1}\widetilde{\mathbf{H}}_{01}\right)\mathbf{C}^{-1}\quad\text{and}\quad\mathbf{b}=\mathbf{A}\mathbf{C}\widetilde{\boldsymbol{\theta}}_1+\mathbf{C}^{-1}\left(\widetilde{\mathbf{g}}_1-\widetilde{\mathbf{H}}_{10}\widetilde{\mathbf{H}}_{00}^{-1}\widetilde{\mathbf{g}}_0\right). \tag{S3.4}$$

Assuming $\widetilde{\mathbf{H}}$ is negative definite, $\mathbf{A}$ must be positive definite. Let $\mathbf{A}^{1/2}$ denote its square root or the Cholesky factor. Minimizing (S3.3) is equivalent to

$$\min_{\boldsymbol{\vartheta}_1}\quad\frac{1}{2}\parallel\mathbf{A}^{1/2}\boldsymbol{\vartheta}_1-\mathbf{A}^{-1/2}\mathbf{b}\parallel^2+\lambda\sum_j|\vartheta_{1j}|, \tag{S3.5}$$

whose solution $\boldsymbol{\vartheta}_1^\star$ can be solved directly via LARS. The solution to (S3.3) is then given by $\boldsymbol{\theta}_1^\star=\mathbf{C}^{-1}\boldsymbol{\vartheta}_1^\star$. Bringing $\boldsymbol{\theta}_1^\star$ into (S3.2) gives $\boldsymbol{\theta}_0^\star$.

If the MLE $\hat{\boldsymbol{\theta}}$ is used for the initial $\widetilde{\boldsymbol{\theta}}$, then $\widetilde{\mathbf{g}}=\mathbf{0}$ and further simplification can be made to $\mathbf{b}=\mathbf{A}\mathbf{C}\hat{\boldsymbol{\theta}}_1$ in (S3.4) and $\boldsymbol{\theta}_0^\star=\hat{\boldsymbol{\theta}}_0+\hat{\mathbf{H}}_{00}^{-1}\hat{\mathbf{H}}_{01}(\boldsymbol{\theta}_1^\star-\hat{\boldsymbol{\theta}}_1)$ in (S3.2). The Cholesky decomposition of $\hat{\mathbf{H}}=\mathbf{R}^T\mathbf{R}$ can be conveniently used to aid in the computation. Write the Cholesky factor as $\mathbf{R}=[\mathbf{R}_0\ \mathbf{R}_1]$ so that $\mathbf{R}_0$ denotes the first $(M+1)$ columns of $\mathbf{R}$ and $\mathbf{R}_1$ for the rest. It can be easily seen that $\hat{\mathbf{H}}_{11}-\hat{\mathbf{H}}_{10}\hat{\mathbf{H}}_{00}^{-1}\hat{\mathbf{H}}_{01}=\mathbf{R}_1^T\mathbf{R}_1-\mathbf{R}_1^T\mathbf{P}_0\mathbf{R}_1$, where $\mathbf{P}_0=\mathbf{R}_0^T(\mathbf{R}_0^T\mathbf{R}_0)^{-1}\mathbf{R}_0$ is the project matrix onto the column space of $\mathbf{R}_0$. This leads to an algorithm similar to that given in Appendix A of Buu et al. (2011). ∎

References

Buu, A., Johnson, N. J., Li, R., and Tan, X. (2011). New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in medicine* **30**, 2326–2340.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics* **32**, 244–248.