kKT Conditions for Zero-Inflated Regression

Zhu Wang
UT Health San Antonio
wangz1@uthscsa.edu

February 19, 2020

This technical report details the calculation of solution path of the **R** package **mpath** on penalized zero-inflated regression, supplementary to the publications Wang et al. (2014, 2015).

1 Zero-inflated Poisson regression

Assume response variable Y has a zero-inflated Poisson distribution, denote d_1+1 and d_2+1-dimensional vectors B_i and G_i. The parameters $\mu = (\mu_1, \mu_2, ...\mu_n)^\intercal$ and π_i are modeled with $\log(\mu_i) = B^\intercal\beta$ and $\text{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = G_i^\intercal\zeta$ for covariate matrix B and G, which can be different. To include an intercept, let $B_{i0} = G_{i0} = 1$. Define $\Phi = (\zeta^\intercal, \beta^\intercal)^\intercal$ with length $d = d_1 + d_2 + 2$. The log-likelihood function $\ell_{ZIP}(\Phi; y)$ is given by

$$\begin{aligned}\ell_{ZIP}(\Phi; y) = &\sum_{y_i=0}\log(\exp(G_i^\intercal\zeta) + \exp(-e^{B_i^\intercal\beta})) + \sum_{y_i>0}(y_iB_i^\intercal\beta - \exp(B_i^\intercal\beta))\\ &- \sum_{i=1}^n\log(1+\exp(G_i^\intercal\zeta)) - \sum_{y_i>0}\log(y_i!).\end{aligned} \tag{1}$$

We minimize a penalized negative log-likelihood function for ZIP model:

$$p\ell_{ZIP}(\Phi; y) = -\ell_{ZIP}(\Phi; y) + p(\zeta, \beta), \tag{2}$$

where

$$p(\zeta, \beta) = n\sum_{j=1}^{d_1}(\alpha_1\lambda_1|\zeta_j| + \frac{\lambda_1(1-\alpha_1)}{2}\zeta_j^2) + n\sum_{k=1}^{d_2}(\alpha_2\lambda_2|\beta_k| + \frac{\lambda_2(1-\alpha_2)}{2}\beta_k^2). \tag{3}$$

In the **R** package **mpath**, α_1, α_2 are labeled as `alpha.zero, alpha.count`, respectively; λ_1, λ_2 are labeled as `lambda.zero, lambda.count`, respectively. A point $\hat\Phi$ is a minimizer of $p\ell_{ZIP}(\Phi; y)$ if and only if $p\ell_{ZIP}(\Phi; y)$ is subdifferentiable at $\hat\Phi$ and $\Phi = 0$ is a subgradient of $p\ell_{ZIP}(\Phi; y)$ at $\hat\Phi$. Take derivatives:

$$
\begin{aligned}
\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \zeta_j} &= \sum_{y_i=0} \frac{\exp(G_i^\intercal \zeta) G_{ij}}{\exp(G_i^\intercal \zeta) + \exp(-\exp(B_i^\intercal \beta))} - \sum_{i=1}^{n} \frac{\exp(G_i^\intercal \zeta) G_{ij}}{1 + \exp(G_i^\intercal \zeta)}, \\
\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \beta_k} &= \sum_{y_i=0} \frac{\exp(-\exp(B_i^\intercal \beta))(-\exp(B_i^\intercal \beta)) B_{ik}}{\exp(G_i^\intercal \zeta) + \exp(-\exp(B_i^\intercal \beta))} \\
&\quad + \sum_{y_i>0} (y_i B_{ik} - \exp(B_i^\intercal \beta) B_{ik}).
\end{aligned} \tag{4}
$$

In this document, we take $j = 1, ..., d_1, k = 1, ..., d_2$ unless otherwise specified. It is clear that $n\alpha_1\lambda_1 \text{sign}(\zeta_j) + \lambda_1(1-\alpha_1)\zeta_j$ is a subgradient of $p(\zeta, \beta)$ at $\hat{\zeta}_j \neq 0$, $n\alpha_2\lambda_2 \text{sign}(\beta_k) + \lambda_2(1-\alpha_2)\beta_k$ is a subgradient of $p(\zeta, \beta)$ at $\hat{\beta}_k \neq 0$, $n\alpha_1\lambda_1 e_1$ for $e_1 \in [-1, 1]$ is a subgradient of $p(\zeta, \beta)$ at $\hat{\zeta}_j = 0$, and $n\alpha_2\lambda_2 e_2$ for $e_2 \in [-1, 1]$ is a subgradient of $p(\zeta, \beta)$ at $\hat{\beta}_k = 0$. If $\hat{\Phi}$ is a minimizer of $p\ell_{ZIP}(\Phi; y)$, then $0 \in \partial p\ell_{ZIP}(\hat{\Phi}; y)$, the subdifferential of $p\ell_{ZIP}(\Phi; y)$ at $\hat{\Phi}$, which leads to the KKT conditions:

$$
\begin{aligned}
&\text{if } \hat{\zeta}_j \neq 0 : -\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \zeta_j} + n\alpha_1\lambda_1 \text{sign}(\zeta_j) + \lambda_1(1-\alpha_1)\zeta_j = 0, \\
&\text{if } \hat{\zeta}_j = 0 : -\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \zeta_j} + n\alpha_1\lambda_1 e_1 = 0, \\
&\text{if } \hat{\beta}_k \neq 0 : -\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \beta_k} + n\alpha_2\lambda_2 \text{sign}(\beta_k) + \lambda_2(1-\alpha_2)\beta_k = 0, \\
&\text{if } \hat{\beta}_k = 0 : -\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \beta_k} + n\alpha_2\lambda_2 e_2 = 0.
\end{aligned} \tag{5}
$$

Therefore, for $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, it must be:

$$
\left| \frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \zeta_j} \right| \leq n\alpha_1\lambda_1, \left| \frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \beta_k} \right| \leq n\alpha_2\lambda_2. \tag{6}
$$

Denote $\lambda_{1,\max}$ and $\lambda_{2,\max}$ the smallest values of λ_1 and λ_2, respectively, such that $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, and $(\lambda_{1,\max}, \lambda_{2,\max})$ can be determined by (6) and the following quantities:

$$
\begin{aligned}
\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \zeta_j} &= \sum_{y_i=0} \frac{\exp(\zeta_0) G_{ij}}{\exp(\zeta_0) + \exp(-\exp(\beta_0))} - \sum_{i=1}^{n} \frac{\exp(\zeta_0) G_{ij}}{1 + \exp(\zeta_0)}, \\
\frac{\partial \ell_{ZIP}(\Phi; y)}{\partial \beta_k} &= \sum_{y_i=0} \frac{\exp(-\exp(\beta_0))(-\exp(\beta_0)) B_{ik}}{\exp(\zeta_0) + \exp(-\exp(\beta_0))} \\
&\quad + \sum_{y_i>0} (y_i B_{ik} - \exp(\beta_0) B_{ik}).
\end{aligned} \tag{7}
$$

There is an alternative approach to construct $(\lambda_{1,\max}, \lambda_{2,\max})$ as in Wang et al. (2014). In mixture models, the EM algorithm is set up by imposing missing data into the problem. Suppose we could observe which zeros came from the zero state and which came from Poisson state; i.e., suppose we knew $z_i = 1$ when y_i is from zero state, and $z_i = 0$ when y_i is from the Poisson state. Denote

$z = (z_1, z_2, ..., z_n)^\intercal$. The complete data (y, z) log-likelihood function can be written as

$$\begin{aligned}\ell^c_{ZIP}(\Phi; y, z) = &\sum_{i=1}^{n} \{z_i G_i^\intercal \zeta - \log(1 + \exp(G_i^\intercal \zeta))\} \\ &+ \sum_{i=1}^{n} (1 - z_i) \{y_i B_i^\intercal \beta - \exp(B_i^\intercal \beta) - \log(y_i!)\}.\end{aligned} \tag{8}$$

The complete data penalized negative log-likelihood function is then given by

$$p\ell^c_{ZIP}(\Phi; y, z) = -\ell^c_{ZIP} + p(\zeta, \beta). \tag{9}$$

Taking derivatives of (8), we obtain

$$\begin{aligned}\frac{\partial \ell^c_{ZIP}(\Phi; y, z)}{\partial \zeta_j} &= \sum_{i=1}^{n} \left\{ z_i G_{ij} - \frac{\exp(G_i^\intercal \zeta) G_{ij}}{1 + \exp(G_i^\intercal \zeta)} \right\}, \\ \frac{\partial \ell^c_{ZIP}(\Phi; y, z)}{\partial \beta_k} &= \sum_{i=1}^{n} (1 - z_i) \{y_i B_{ik} - \exp(B_i^\intercal \beta) B_{ik}\}.\end{aligned} \tag{10}$$

The KKT conditions of a minimizer $\hat{\Phi}$ of $p\ell^c_{ZIP}(\Phi; y, z)$ are given by:

$$\begin{aligned}&\text{if } \hat{\zeta}_j \neq 0 : -\frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \zeta_j} + n\alpha_1 \lambda_1 \text{sign}(\zeta_j) + \lambda_1 (1 - \alpha_1) \zeta_j = 0, \\ &\text{if } \hat{\zeta}_j = 0 : -\frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \zeta_j} + n\alpha_1 \lambda_1 e_1 = 0, \\ &\text{if } \hat{\beta}_k \neq 0 : -\frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \beta_k} + n\alpha_2 \lambda_2 \text{sign}(\beta_k) + \lambda_2 (1 - \alpha_2) \beta_k = 0, \\ &\text{if } \hat{\beta}_k = 0 : -\frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \beta_k} + n\alpha_2 \lambda_2 e_2 = 0.\end{aligned} \tag{11}$$

Therefore, for $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, it must be:

$$\left| \frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \zeta_j} \right| \leq n\alpha_1 \lambda_1, \left| \frac{\partial \ell^c_{ZIP}(\Phi; y)}{\partial \beta_k} \right| \leq n\alpha_2 \lambda_2. \tag{12}$$

The EM algorithm estimates z_i at iteration m by its conditional mean $z_i^{(m)}$ given below:

$$z_i^{(m)} = \begin{cases} \left[1 + \exp(-G_i^\intercal \zeta^{(m)} - \exp(B_i^\intercal \beta^{(m)})\right]^{-1}, & \text{if } y_i = 0, \\ 0, & \text{if } y_i = 1, 2, ... \end{cases} \tag{13}$$

Let $\zeta^{(m)} = \zeta, \beta^{(m)} = \beta$, then (13) becomes

$$z_i = \begin{cases} \left[1 + \exp(-G_i^\intercal \zeta - \exp(B_i^\intercal \beta)\right]^{-1}, & \text{if } y_i = 0, \\ 0, & \text{if } y_i = 1, 2, ... \end{cases} \tag{14}$$

It is a simple exercise to show that the right hand side of (10) is the same as that of (4) once (14) is plugged into (10). Hence, the KKT conditions (11) are the same as (5) once (14) is plugged into (10). These connections offer a different method to derive $(\hat{\lambda}_{1,\max}, \hat{\lambda}_{2,\max})$ such that $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$ hold (Wang et al., 2014). We first estimate ζ_0, β_0 for an intercept-only ZIP model, then (14) reduces to

$$z_i = \begin{cases} [1 + \exp(-\zeta_0 - \exp(\beta_0)]^{-1}, & \text{if } y_i = 0, \\ 0, & \text{if } y_i = 1, 2, ... \end{cases} \tag{15}$$

Plugging in (15), $(\hat{\lambda}_{1,\max}, \hat{\lambda}_{2,\max})$ are computed based on (12). Furthermore, we have shown that $(\lambda_{1,\max}, \lambda_{2,\max}) = (\hat{\lambda}_{1,\max}, \hat{\lambda}_{2,\max})$ holds, a special case of a more general result (**?**).

2 Zero-inflated negative binomial regression

Assume response variable Y has a zero-inflated negative binomial distribution, denote d_1+1 and d_2+1-dimensional vectors B_i and G_i, respectively. As before, the first entry of these vectors is 1. In ZINB regression, assume $\log(\mu_i) = B_i^{\intercal}\beta$ and $\log(\frac{p_i}{1-p_i}) = G_i^{\intercal}\zeta$ where $\zeta = (\zeta_0, \zeta_1, ..., \zeta_{d_1})$ and $\beta = (\beta_0, \beta_1, ..., \beta_{d_2})$ are unknown parameters. Here ζ_0 and β_0 are intercepts. For n independent random samples, denote $\Phi = (\zeta^{\intercal}, \beta^{\intercal}, \theta)^{\intercal}$, the log-likelihood function is then given by

$$\begin{aligned} \ell_{ZINB}(\Phi; y) = &\sum_{y_i=0} \log \left[p_i + (1-p_i)(\frac{\theta}{\mu_i + \theta})^{\theta} \right] \\ &+ \sum_{y_i>0} \log \left[(1-p_i) \frac{\Gamma(\theta + y_i)}{\Gamma(y_i+1)\Gamma(\theta)} (\frac{\mu_i}{\mu_i+\theta})^{y_i} (\frac{\theta}{\mu_i+\theta})^{\theta} \right], \end{aligned}$$

where $\mu_i = \exp(B_i^{\intercal}\beta)$ and $p_i = \frac{\exp(G_i^{\intercal}\zeta)}{1+\exp(G_i^{\intercal}\zeta)}$. The derivatives are given by:

$$\begin{aligned} \frac{\partial \ell_{ZINB}(\Phi; y)}{\partial \zeta_j} &= \sum_{y_i=0} \frac{\frac{\partial p_i}{\partial \zeta_j} - \frac{\partial p_i}{\partial \zeta_j}(\frac{\theta}{\mu_i+\theta})^{\theta}}{p_i + (1-p_i)(\frac{\theta}{\mu_i+\theta})^{\theta}} - \sum_{y_i>0} \frac{\partial p_i}{\partial \zeta_j} \frac{1}{1-p_i}, \\ \frac{\partial \ell_{ZINB}(\Phi; y)}{\partial \beta_k} &= \sum_{y_i=0} \frac{-(1-p_i)\frac{\partial u_i}{\partial \beta_k} \dfrac{\theta \left(\frac{\theta}{\mu_i+\theta}\right)^{\theta}}{\mu_i+\theta}}{p_i + (1-p_i)(\frac{\theta}{\mu_i+\theta})^{\theta}} + \sum_{y_i>0} \frac{\partial u_i}{\partial \beta_k} (\frac{y_i}{\mu_i} - \frac{y_i+\theta}{\mu_i+\theta}), \end{aligned} \tag{16}$$

where

$$\begin{aligned} \frac{\partial p_i}{\partial \zeta_j} &= \frac{G_{ij}\exp(G_i^{\intercal}\zeta)}{(1+\exp(G_i^{\intercal}\zeta))^2}, \\ \frac{\partial u_i}{\partial \beta_k} &= B_{ik}\exp(B_i^{\intercal}\beta). \end{aligned}$$

For variable selection, consider minimizing a penalized negative loss function:

$$p\ell_{ZINB}(\Phi; y) = -\ell(\Phi) + p(\zeta, \beta), \tag{17}$$

where $p(\zeta, \beta)$ is given by (3). The KKT conditions for a minimizer $\hat{\Phi}$ of $p\ell_{ZINB}(\Phi)$ can be derived. Therefore, for $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, it must be:

$$\left| \frac{\partial \ell_{ZINB}(\Phi; y)}{\partial \zeta_j} \right| \le n\alpha_1\lambda_1, \left| \frac{\partial \ell_{ZINB}(\Phi; y)}{\partial \beta_k} \right| \le n\alpha_2\lambda_2. \tag{18}$$

Denote $\lambda_{1,\max}$ and $\lambda_{2,\max}$ the smallest values of λ_1 and λ_2, respectively, such that $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, and $(\lambda_{1,\max}, \lambda_{2,\max})$ can be determined by (16), (18) and the following quantities:

$$p_i = \frac{\exp(\zeta_0)}{1+\exp(\zeta_0)}, \frac{\partial p_i}{\partial \zeta_j} = \frac{G_{ij}\exp(\zeta_0)}{(1+\exp(\zeta_0))^2}, \mu_i = \exp(\beta_0), \frac{\partial u_i}{\partial \beta_k} = B_{ik}\exp(\beta_0). \tag{19}$$

Consider an EM algorithm to optimize (17). Let $z_i = 1$ if y_i is from the zero state and $z_i = 0$ if y_i is from the NB state. Since $z = (z_1, ..., z_n)^T$ is not observable, it is often treated as missing data. The EM algorithm is particularly attractive to missing data problems. If complete data (y, z) are available, the complete data log-likelihood function is given by

$$\ell^c_{ZINB}(\Phi; y) = \sum_{i=1}^{n} \left\{ (z_i G_i^\intercal \zeta - \log(1+\exp(G_i^\intercal \zeta)) + (1-z_i)\log(f(y_i;\beta,\theta)) \right\}, \tag{20}$$

and the complete data penalized negative log-likelihood function is given by

$$p\ell^c_{ZINB}(\Phi; y, z) = -\ell^c_{ZINB}(\Phi; y, z) + p(\zeta, \beta),$$

where $f(y_i;\beta,\theta) = \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)}(\frac{\mu_i}{\mu_i+\theta})^{y_i}(\frac{\theta}{\mu_i+\theta})^{\theta}$ and $\mu_i = \exp(B_i^\intercal\beta)$. Taking derivatives of (20), we obtain

$$\begin{aligned} \frac{\partial \ell^c_{ZINB}(\Phi; y, z)}{\partial \zeta_j} &= \sum_{i=1}^{n} \left\{ z_i G_{ij} - \frac{\exp(G_i^\intercal \zeta) G_{ij}}{1+\exp(G_i^\intercal \zeta)} \right\}, \\ \frac{\partial \ell^c_{ZINB}(\Phi; y, z)}{\partial \beta_k} &= \sum_{i=1}^{n} \left\{ (1-z_i)\frac{\partial u_i}{\partial \beta_k}(\frac{y_i}{\mu_i} - \frac{y_i+\theta}{\mu_i+\theta}) \right\}. \end{aligned} \tag{21}$$

The KKT conditions of a minimizer $\hat{\Phi}$ of $\ell^c_{ZINB}(\Phi; y, z)$ can be derived. Therefore, for $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$, it must be:

$$\left| \frac{\partial \ell^c_{ZINB}(\Phi; y)}{\partial \zeta_j} \right| \le n\alpha_1\lambda_1, \left| \frac{\partial \ell^c_{ZINB}(\Phi; y)}{\partial \beta_k} \right| \le n\alpha_2\lambda_2. \tag{22}$$

The conditional expectation of z_i at iteration m is provided by

$$z_i^{(m)} = \begin{cases} \left(1+\exp(-G_i^\intercal \zeta^{(m)}) \left[\frac{\theta}{\exp(B_i^\intercal \beta^{(m)})+\theta}\right]^{\theta}\right)^{-1}, & \text{if } y_i = 0 \\ 0, & \text{if } y_i > 0. \end{cases} \tag{23}$$

Let $\zeta^{(m)} = \zeta, \beta^{(m)} = \beta$, then (23) becomes

$$z_i = \begin{cases} \left(1+\exp(-G_i^\intercal \zeta) \left[\frac{\theta}{\exp(B_i^\intercal \beta)+\theta}\right]^{\theta}\right)^{-1}, & \text{if } y_i = 0 \\ 0, & \text{if } y_i > 0. \end{cases} \tag{24}$$

It is simple to show that the right hand side of (21) is the same as that of (16) once (24) is plugged into (21). Hence, the KKT conditions (22) are the same as (18) once (24) is plugged into (21). These connections offer a different method to derive $(\hat{\lambda}_{1,\max}, \hat{\lambda}_{2,\max})$ such that $\hat{\zeta}_j = 0, \hat{\beta}_k = 0$ hold (Wang et al., 2015). We first estimate ζ_0, β_0 for an intercept-only ZINB model, then (24) becomes

$$z_i = \begin{cases} \left(1 + \exp(-\zeta_0)\left[\frac{\theta}{\exp(\beta_0)+\theta}\right]^{\theta}\right)^{-1}, & \text{if } y_i = 0 \\ 0, & \text{if } y_i > 0. \end{cases} \tag{25}$$

Plugging in (25), $(\lambda_{1,\max}, \lambda_{2,\max})$ are computed based on (22). Furthermore, we have shown that $(\lambda_{1,\max}, \lambda_{2,\max}) = (\hat{\lambda}_{1,\max}, \hat{\lambda}_{2,\max})$ holds.

References

Zhu Wang, Shuangge Ma, Ching-Yun Wang, Michael Zappitelli, Prasad Devarajan, and Chirag Parikh. EM for regularized zero inflated regression models with applications to postoperative morbidity after cardiac surgery in children. *Statistics in Medicine*, 33(29):5192–5208, 2014. URL http://dx.doi.org/10.1002/sim.6314.

Zhu Wang, Shuangge Ma, and Ching-Yun Wang. Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. *Biometrical Journal*, 33(29):5192–208, 2015. URL http://dx.doi.org/10.1002/bimj.201400143.