Network analysis for count data with excess zeros
Hosik Choi, Jungsoo Gim, Sungho Won, You Jin Kim, Sunghoon Kwon and Changyi Park

Abstract
Background: Undirected graphical models or Markov random fields have been a popular class of models for representing conditional dependence relationships between nodes. In particular, Markov networks help us to understand complex interactions between genes in biological processes of a cell. Local Poisson models seem to be promising in modeling positive as well as negative dependencies for count data. Furthermore, when zero counts are more frequent than expected, excess zeros should be considered in the model.

Methods: We present a penalized Poisson graphical model for zero inflated count data and derive an expectation-maximization (EM) algorithm built on coordinate descent. Our method is shown to be effective through simulated and real data analysis.

Results: Results from the simulated data indicate that our method outperforms the local Poisson graphical model in the presence of excess zeros. In an application to a RNA sequencing data, we also investigate the gender effect by comparing the estimated networks according to different genders. Our method may help us in identifying biological pathways linked to sex hormone regulation and thus understanding underlying mechanisms of the gender differences.

Conclusions: We have presented a penalized version of zero inflated spatial Poisson regression and derive an efficient EM algorithm built on coordinate descent. We discuss possible improvements of our method as well as potential research directions associated with our findings from the RNA sequencing data.

Keywords: Count data, EM algorithm, Network, Zero inflation

Background
Graphical models help us to explore relationships between nodes in graphs. Undirected graphical models or Markov random fields have been a popular class of models for representing conditional dependence relationships between nodes. Examples include Gaussian graphical models for continuous data, Ising model for binary data, and multinomial graphical models. These Markov networks help us to understand complex interactions between genes in biological processes of a cell and have been well studied in bioinformatics. Examples of Markov networks in learning the network structure from microarray and next generation sequencing data include [1–4]. For more details on Markov network inference, see those and the references therein.

The main focus of this study is to infer the network structure for a count data. The auto-Poisson model in [5] is a natural extension of univariate Poisson distribution. However it can model only negative dependencies, so that the conditional distributions define a unique joint distribution consistently. Yang et al. [6] propose variants of the auto-Poisson model such as truncated, quadratic, and sub-linear Poisson graphical models (PGM). However none of them provide a satisfactory answer to the question of how to specify a consistent joint graphical model for count data capturing both positive and negative dependencies. Allen and Liu [4] consider a local PGM (LPGM). The LPGM does not have a consistent joint graphical model, but it has the local Markov property and thus the zero coefficient of an edge weight between two nodes implies the conditional independence of the two nodes given the others. Žitnik and Zupan [7] consider a latent factor Poisson model and [8] propose to learn conditional dependence...
structures for binary and Poisson data via marginal loss functions. Also a semiparametric Guassian copula, called the nonparanormal graphical model (NPGM), has been proposed [9].

In practice, zero counts are sometimes more frequent than are expected under a univariate Poisson distribution. In such cases, a zero-inflated Poisson (ZIP) distribution is often adopted. Applications of ZIP models include modeling of defects in quality control [10] and alcoholism and substance abuse in medicine [11]. Extensions of a ZIP model in different frameworks are well-studied in the literature. Dobbie and Welsh [12] extend the two component approach in [13] for serially correlated count data exhibiting extra zeros. Monod [14] develops a zero-inflated spatial Poisson (ZISP) model. Buu et al. [11] study variable selection methods such as LASSO and one-step SCAD for ZIP regression models. For computation, a local linear approximation (LLA) is adopted. The LLA algorithm fails to converge particularly with small sample sizes because it requires fitting penalized ZIP regression models. Wang et al. [15] propose an expectation maximization (EM) algorithm [16] for a penalized ZIP regression model built on coordinate descent algorithms. The EM algorithm seems to have some advantages over the LLA algorithm in numerical convergence and tuning.

In this paper, we are interested in the construction of graphical models for count data, particularly, with excessive zeros. To this end, we propose a penalized version the ZISP model in [14] called zero inflated local Poisson graphical model (ZILPGM) and derive an EM algorithm built on coordinate descent as in [15]. We show the effectiveness of our method on simulated and real data. In an application to a RNA sequencing data, we investigate the gender effect by comparing the estimated networks according to different genders. It has been well noted that gender is one of the major contributors in the differentiation of gene expression profiles [17, 18] and various sexually dimorphic phenotypes, most of which result from hormonal differences [19]. It was reported that transcriptome study could be predicted to represent a different promising approach for the identification of biological pathways linked to sex hormone regulation and the analysis of associated gene regulatory networks [20]. However, the elucidation of underlying mechanisms of the gender differences is still an area of interest and intense investigation.

The paper is organized as follows. In “Methods” section, we propose a new graph learning method based on ZISP and provide an efficient EM type numerical algorithm. In “Results” section, we compare performances of our method with LPGM on simulated and real data sets. Some discussions and concluding remarks are given in “Conclusions” section.

**Methods**

In this section, we present our graph learning method based on a penalization of the ZISP in [14] and derive an efficient EM algorithm for its computation.

### Zero inflated local Poisson graphical model

Let \( N \) denote the number of observations and \( p \) denote the number of variables or nodes. Denote \( \mathcal{G} = (V, E) \), where \( V = \{1, \ldots, p\} \) is the set of vertices or nodes and \( E \) is the set of edges. We use uppercase letters such as \( X \) and \( Z \) when we refer to random variables. Observations are written in lowercase. For example, \( x_i \) denote \( i \)th observation of \( X \). Vectors and matrices are represented by boldface and blackboard boldface letters, respectively.

Define \( X = (x_{ij})_{N \times p} \), where \( x_{ij} \) is generated from two latent components with zero and Poisson states. Let \( z_{ij} \) be a latent variable such that \( z_{ij} = 1 \) if \( x_{ij} \) is from zero state and \( z_{ij} = 0 \) if \( x_{ij} \) is from Poisson state. \( z_{ij} \) follows a Bernoulli distribution with \( \pi_j \). Let \( I(\cdot) \) denotes an indicator function. Then the ZISP model in [14] is defined by

\[
P(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

where \( \mu_j = \exp(\beta_j + \sum_{k \neq j} \beta_{jk} x_k) \), \( \beta_j \) is an intercept adjusting for \( X_j \), and \( \beta_{jk} \) is the parameter accounting for the conditional relation between \( X_j \) and \( X_k \).

Due to the zero inflation term in the conditional probability, the situation becomes more complicated in our case than in LPGM. Because the important part is the pairwise interaction term in the pairwise-only dependency models, the situation is basically similar. In order to have a valid joint distribution, the coefficient for the interaction term \( \beta_{jk} \) should be non-positive. As in the LPGM, we do not solve the issue of negative parameters in the Poisson graphical model. Note that any existing approaches (e.g. in [6]) do not succeed in giving a satisfactory answer to the consistency issue. Rather, we focus not on the consistency issue but on the practical issue of estimating positive as well as negative dependencies as in LPGM.

In order to learn graph structures, we consider the minimization of the penalized pseudo log-likelihood of (1) in the general weighted LASSO form:

\[
-\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{p} \log \left( \pi_j I(x_{ij} = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!} \right) + \lambda \sum_{j=1}^{p} \sum_{k \neq j} w_{jk} |\beta_{jk}|,
\]

where

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

and

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

where

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

and

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

where

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

and

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

where

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

and

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

where

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]

and

\[
\mathbb{P}(X_j = x_j | X_k = x_k, k \neq j) = \pi_j I(x_j = 0) + (1 - \pi_j) \frac{e^{-\mu_j} \mu_j^{x_{ij}}}{x_{ij}!},
\]
where $\mu_{ij} = \exp(\beta_j + \sum_{k \neq j} \hat{\beta}_{jk} x_{ik})$, $\lambda \geq 0$ is the penalty parameter, and $w_{jk} \geq 0$ is an appropriate weight. As in [4], we can select the tuning parameter using the stability selection criterion in [21]. More specifically, we select the optimal $\lambda$ is selected from 30 equal-spaced grid points in log scale on $[\lambda_{\text{max}}, \lambda_{\text{min}}]$, where $\lambda_{\text{max}} = \max_{j \in [1, \ldots, p]} \max_{k \neq j} \lambda_{\text{max}} \times 10^{-3}$. For each $j$, we fit poisson regression using glmnet for covariates with nonzero coefficients. However, it is difficult to maximize this likelihood directly because the score function of $-\sum_{i \in C_j} \log(\pi_j (1 - \pi_j) + e^{-\mu_{ij}})$ cannot be simplified [14, 22].

Instead of a direct optimization of the likelihood function, we express the likelihood function as a mixture distribution by introducing a latent variable and derive an EM algorithm.

Define $\beta_{-j} = (\beta_0, (\beta_k)_{k \neq j})^T$ and $x_{i, -j} = (1, (x_{ik})_{k \neq j})^T$. The log-likelihood function with respect to complete data can be written as

$$
\ell_j = -\sum_{i=1}^N z_{ij} \log \pi_j - \sum_{i=1}^N (1 - z_{ij}) \pi_j - x_{ij} \pi_j e^{-x_{ij}} - \log \pi_j - \sum_{i=1}^N \log (1 - \pi_j)
$$

for $j = 1, \ldots, p$. However, it is difficult to maximize this likelihood directly because the score function of $-\sum_{i \in C_j} \log(\pi_j (1 - \pi_j) + e^{-\mu_{ij}})$ cannot be simplified [14, 22].

The decomposed likelihood function in the above can be easily maximized via an EM algorithm alternating between the expectation of the complete data likelihood over the latent variable $z_{ij}$ and the maximization of the likelihood given $z_{ij}$.

Define the responsibility of zero state for $j$th variable on $i$th observation at $m$th step as $z_{ij}^{(m)} = \mathbb{E}(z_{ij} | x_{ij}, \beta_{-j}^{(m)})$ and the probability of zero state at $m$th step as

$$
\pi_j^{(m)} = \frac{1}{n} \sum_{i=1}^n \left( I(x_{ij} = 0) - I(x_{ij} = 0) (1 - z_{ij}^{(m)}) \right).
$$

Our EM algorithm alternates the following steps until convergence.

- E-step: Estimate $z_{ij}$ by its conditional mean $z_{ij}^{(m)}$ given data and parameters from the previous step.

- M-step: Estimate $\beta_{-j}^{(m)}$.

Here we set the initial values for our EM iteration as $\pi_j^{(0)} = \text{the number of zeros of } j\text{th variable}/n$ for $j = 1, \ldots, p$ and $\beta_{-j}^{(0)} = \mathbf{0}$.

Now let us discuss the estimation of $\beta_{-j}^{(m)}$ in detail. For each variable, we use the Majorize-Minimization (MM) algorithm in [23], which extends the central idea of EM algorithms to situations not necessarily involving missing data nor even maximum likelihood estimation. A function $g(\theta | \theta_m)$ is said to majorize a function $f(\theta)$ at $\theta_m$ provided that $f(\theta_m) = g(\theta_m | \theta_m)$ and $f(\theta) \leq g(\theta | \theta_m)$ for $\theta \neq \theta_m$.

The key idea is that the surrogate majorizing function $g(\theta | \theta_m)$ is minimized iteratively, instead of the original objective function $f(\theta)$ with the nonquadratic log-likelihood and the nondifferentiable sparsity inducing penalty [23]. The MM algorithm starts from an initial guess, $\theta_0$. Let $\theta_{m+1}$ denote the minimizer of the surrogate $g(\theta | \theta_m)$. Then the following inequalities hold:

$$
f(\theta_{m+1}) \leq g(\theta_{m+1} | \theta_m) \leq g(\theta_m | \theta_m) = f(\theta_m).
$$
The above inequality can easily be shown by definition of $\theta_{m+1}$ and the majorization conditions. The descent property makes the MM algorithm numerically stable [24].

The objective to maximize is $l_j^2 = -\sum_{i=1}^{N} (1 - z_{ij}) (x_{ij} \mu_{ij} - \exp (x_{ij}^T \beta_{\cdot j} - \log x_{ij}^T))$ whose first and second derivatives with respect to $\beta_{\cdot j}$ are

$$\frac{\partial l_j}{\partial \beta_{\cdot j}} = -\sum_{i=1}^{N} (1 - z_{ij}) (x_{ij} - \mu_{ij}) x_{i\cdot j},$$

$$\frac{\partial^2 l_j}{\partial \beta_{\cdot j}^2} = \sum_{i=1}^{N} (1 - z_{ij}) \mu_{ij} x_{i\cdot j} x_{i\cdot j}^T.$$ 

Let $X_{\cdot j} = \begin{pmatrix} (x_{1\cdot j}, \ldots, x_{N\cdot j})^T \end{pmatrix}$. Define $B^{(m)} = \begin{pmatrix} (1 - z_{ij}) (x_{ij} - \mu_{ij})^T, \ldots, (1 - z_{nj}) (x_{nj} - \mu_{nj}) \end{pmatrix}$ and $E^{(m)} = X_{\cdot j}^T \text{diag}(1 - z_{ij}) \mu_{ij}, \ldots, (1 - z_{nj}) \mu_{nj} X_{\cdot j}$. If we ignore additive constants, the quadratic approximation of the objective function at $\hat{\beta}^{(m)}_{\cdot j}$ yields

$$l_j^2 \approx \frac{1}{2} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})^T E^{(m)} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})$$

$$- (B^{(m)})^T X_{\cdot j} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})$$

$$\leq \frac{\sigma^{(m)}}{2} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})^T X_{\cdot j} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})$$

$$- (B^{(m)})^T X_{\cdot j} (\beta_{\cdot j} - \hat{\beta}^{(m)}_{\cdot j})$$

for an appropriate $\sigma^{(m)}$. To find an appropriate upper bound, we may set $\sigma^{(m)}$ as the maximum of $(1 - z_{ij}^{(m)}) \mu_{ij}^{(m)}$ for $i = 1, \ldots, N$. We can easily show that

$$\sigma^{(m)} X_{\cdot j} X_{\cdot j} - E^{(m)}$$

is a positive definite matrix. The upper bound can be expressed as

$$l_j^2 \leq \frac{\sigma^{(m)}}{2} \|w_{\cdot j}^{(m)} - X_{\cdot j} \beta_{\cdot j}\|_2^2,$$ 

where $w_{\cdot j}^{(m)} = X_{\cdot j} \hat{\beta}_{\cdot j}^{(m)} + \sigma^{(m)-1} b^{(m)}$. The majorized problem is written as

$$\min_{\beta_{\cdot j} \in \mathbb{R}^p} \left( \frac{1}{2} \|w_{\cdot j}^{(m)} - X_{\cdot j} \beta_{\cdot j}\|_2^2 + \frac{\lambda}{\sigma^{(m)}} \sum_{k 
eq j} w_{jk} |\beta_k| \right). \quad (4)$$

Up to a constant depending not on $\beta_{\cdot j}$ but on $\hat{\beta}_{\cdot j}^{(m)}$, the function in the minimization problem (4) majorizes $l_j^2$. Hence we achieve the property, guaranteeing the convergence of the algorithm for $\beta_{\cdot j}^{(m)}$ in M-step.

**Results**

In this section, we illustrate that our method is effective through a simulation study by comparing the performances of our method, LPGM, and NPGM on simulated data. Then we apply our method to a RNA sequencing data. Also we investigate the gender effect by comparing the estimated networks according to different genders.

**Simulation**

To simulate data from a Poisson network with excess zeros, we modify the data generation scheme in [4] slightly. Let $X \in \{0, 1, \ldots, \infty\}^{m \times p}$ denote $n$ independent observations from a Poisson network with $p$ nodes. The data generation model is given as

$$X = YB + E,$$

where $Y$ is a $n \times (p_{C2} + pC_2)$ matrix with $Y_{ij} \sim \text{Poisson}(\lambda_{\text{true}})$ and $E$ is a $n \times p$ matrix with $E_{ij} \sim \text{Poisson}(\lambda_{\text{noise}})$. The coefficient matrix $B$ encoding the true underlying graph structure denoted by the adjacency matrix $A \in \{0, 1\}^{p \times p}$ is defined as

$$B = \left[ I_{p}; P \odot (1_p \text{tri}(A)^T) \right]^T,$$

where $P$ is the $p \times pC_2$ pairwise permutation matrix, $\odot$ denotes the element-wise product, and $\text{tri}(A)$ is the $pC_2 \times 1$ vectorized upper triangular portion of the adjacency matrix. Each of off-diagonal elements in $A$ is randomly generated from Bernoulli($\rho$), where $\rho$ is the sparsity parameter for the network defined as the number of active edges in $A$ divided by the number of all possible edges between the nodes. In order to make $X_{ij}$'s to have excess zeros, we multiply each of $X_{ij}$ by a random variate from Bernoulli($\pi$) for $i = 1, \ldots, N$ and $j = 1, \ldots, p$. As an abuse of notation, we denote the final matrix containing zero inflated Poisson counts as $X$.

The Poisson rates were set as $\lambda_{\text{true}} = 1.5$ and $\lambda_{\text{noise}} = 0.5$. And we have experimented at different levels of $N(= 50, 100, 150)$, $p(= 10, 20, 30)$, $\pi(= 0\%, 10\%, 20\%)$, and $\rho(= .2, .3, .4)$. At each experimental condition, we generated data according to the above scheme and compared the areas under the curve (AUC) from ZILPGM, LPGM, and NPGM. AUC can be obtained in this way. If we regard active and in-active edges in $A$ as positive and negative examples in a binary classification, then we can compute true positive rate (TPR) as the fraction of edges found by a method that are in the true underlying network structure $A$. False positive rate (FPR) can obtained analogously. Receiver operating characteristic (ROC) curve and AUC can be obtained from TPR and FPR. To assess the variabilities, we replicated the process of generating data and computing AUC’s 100 times. In Tables 1 and 2, average AUC’s of ZILPGM and LPGM with their standard
Table 1 Average AUCs for ZILPGM, LPGM, and NPGM on simulated data with their standard errors in parentheses

| $\rho$ | $N$ | $\rho$ | ZILPGM | LPGM | NPGM | ZILPGM | LPGM | NPGM | ZILPGM | LPGM | NPGM |
|-------|-----|-------|--------|------|------|--------|------|------|--------|------|------|
|       |     |       |        |      |      |        |      |      |        |      |      |
| 0%    | 50  | 3     | .9945  | .944  | .926 | .9826  | .865 | .845 | .804   | .785 | .748 |
|       |     |       | (.0009)| (.0009)| (.0026)| (.0079)| (.0090)| (.0095)| (.0092) | (.0100)| (.0104)|
| 10%   | 4   | 3     | .8972  | .8774 | .888 | .761 | .724 | .689 | .682   | .637 | .586 |
|       |     |       | (.0053)| (.0053)| (.0061)| (.0073)| (.0087)| (.0092)| (.0085)| (.0090)| (.0094)|
| 20%   | 2   | 3     | .7748  | .7749 | .753 | .694 | .652 | .591 | .642   | .610 | .534 |
|       |     |       | (.0075)| (.0076)| (.0083)| (.0089)| (.0089)| (.0092)| (.0077)| (.0079)| (.0083)|
| 100   | 3   | 3     | .9342  | .9341 | .928 | .833 | .775 | .734 | .757   | .686 | .628 |
|       |     |       | (.0043)| (.0043)| (.0055)| (.0067)| (.0075)| (.0070)| (.0084)| (.0088)| (.0093)|
| 4     |     |       | .8188  | .8188 | .818 | .725 | .652 | .620 | .658   | .659 | .600 |
|       |     |       | (.0064)| (.0065)| (.0067)| (.0070)| (.0086)| (.0090)| (.0093)| (.0085)| (.0089)|
| 150   | 2   | 3     | .9762  | .9762 | .961 | .954 | .910 | .833 | .900   | .836 | .719 |
|       |     |       | (.0027)| (.0027)| (.0036)| (.0034)| (.0052)| (.0068)| (.0047)| (.0064)| (.0083)|
| 50    | 3   | 3     | .7088  | .709 | .660 | .660 | .631 | .542 | .637   | .618 | .519 |
|       |     |       | (.0041)| (.0041)| (.0047)| (.0044)| (.0045)| (.0047)| (.0039)| (.0043)| (.0046)|
| 20    | 3   | 3     | .8043  | .804 | .766 | .705 | .655 | .573 | .657   | .624 | .527 |
|       |     |       | (.0034)| (.0034)| (.0038)| (.0038)| (.0039)| (.0042)| (.0041)| (.0041)| (.0046)|
| 150   | 2   | 3     | .8230  | .823 | .820 | .682 | .601 | .582 | .621   | .624 | .560 |
|       |     |       | (.0019)| (.0019)| (.0024)| (.0037)| (.0046)| (.0046)| (.0039)| (.0049)| (.0055)|
| 50    | 3   | 3     | .5875  | .587 | .571 | .558 | .544 | .506 | .593   | .565 | .503 |
|       |     |       | (.0029)| (.0029)| (.0031)| (.0038)| (.0041)| (.0041)| (.0039)| (.0043)| (.0043)|
| 20    | 3   | 3     | .5623  | .562 | .542 | .557 | .546 | .501 | .553   | .557 | .509 |
|       |     |       | (.0028)| (.0028)| (.0029)| (.0025)| (.0027)| (.0030)| (.0026)| (.0028)| (.0030)|
errors in parentheses and \( p \)-value from the paired sign rank test on AUC’s over 100 replications are reported.

Let us consider the effects of each factor with the other factors held fixed. As \( \rho \) increased (or the network became dense), AUC’s of all the compared methods have decreased. Similarly, as the dimension \( p \) grew larger, their AUCs became smaller. As the sample size \( N \) grows, AUC’s tends to improve. However the tendency is sometimes not so clear. Now consider the effect of excess zeros. When there is no zero inflation (\( \pi = 0 \)), AUC’s from ZILPGM, LPGM, and NPGM were not significantly different. When we have zero inflations (\( \pi = 0.1, 0.2 \)), ZILPGM seems to significantly outperform LPGM and NPGM. NPGM seems to be outperformed by LPGM. The gaps between AUC’s from ZILPGM and LPGM when \( \pi = 0.2 \) was not necessarily larger than that when \( \pi = 0.1 \). A potential explanation for this phenomenon follows. As \( \pi \) increases, we have more zero counts in the data and thus the estimation accuracy for the mixing parameter will improve. Meanwhile, the estimation accuracy for the Poisson parameters can degrade because Poisson parameters are learned from nonzero counts. The tradeoff between these two estimation errors may occur at a certain level of \( \pi \).

### Chromosome data

To investigate the validity of the proposed method, we applied it to the RNA sequencing data in the form of a count matrix that contains the number of mapped reads for 60 normal individuals in [25]. We selected 899 genes in the sex chromosomes, i.e., X and Y, first. Each of 899 genes has many zero counts. For a gene with almost all the counts equal to zero, its mixing parameter is estimated as one. To reduce the computation, we have reduced the original data to a data of dimension \( n = 60, p = 360 \) by keeping genes with the number of non-zero counts less than or equal to one.

Figure 1 shows the estimate for the network structure from our method. While 49 genes are clustered together, the other genes remain isolated. Top ranked genes are shown in Table 3 according to their degrees. Note that the degree of a gene is the number of edges being incident upon the gene. We further identified the function of genes with large degree. By GO-BP annotation, NDUFA1 and NDUFB11 are involved in mitochondrial electron transport chain (especially complex I), which affects the capacity for the production of ATP through oxidative phosphorylation. GO annotations related to MID1IP1 and PIM2 are protein C-terminus binding and transferase activity, respectively. Proteins with these functions should highly interact with other proteins to control regulation process in cells. Meanwhile, genes with small degrees, SYAP1 and P2RY10, involved in PI3K/Akt signaling and G-protein coupled receptor (GPCR) activity, respectively [26]. GPCR activate the PI3K/Akt signaling pathway involved in the cellular responses including metabolism, proliferation, apoptosis, and survival [27].

Now let us investigate the effect of gender. In order to compare the networks for different gender groups with 27 males and 33 females, we applied our method to each of gender groups separately. The estimated networks for male and female groups are shown in Figs. 2 and 3. The differentially expressed genes in each group are listed in Table 4. Originally, a differentially expressed gene in a treatment and control groups is a gene with mean expression levels in those groups are significantly different. Although our method is not explicitly related to a hypothesis testing for comparing mean levels, it is implicitly related to a hypothesis testing for conditional independence of counts between genes through a regularized
Table 2 Comparison of ZILPGM, LPGM, and NPGM on simulated data

| \( \pi \) | \( 0\% \) | \( 10\% \) | \( 20\% \) | \( 0\% \) | \( 10\% \) | \( 20\% \) | \( 0\% \) | \( 10\% \) | \( 20\% \) |
|---|---|---|---|---|---|---|---|---|---|
| \( p \) | \( N \) | \( \rho \) | \( \text{vs.} \) | \( \text{vs.} \) | \( \text{vs.} \) | \( \text{vs.} \) | \( \text{vs.} \) | \( \text{vs.} \) | \( \text{vs.} \) |
| 2 | 0.518 | 0.000 | 0.000 | 0.055 | 0.000 | 0.001 | 0.004 | 0.000 | 0.000 |
| 50 | 3 | 0.509 | 0.156 | 0.152 | 0.001 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 |
| 150 | 3 | 0.493 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 10 | 100 | 3 | 0.499 | 0.306 | 0.308 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4 | 0.503 | 0.473 | 0.462 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 |
| 2 | 0.497 | 0.481 | 0.483 | 0.079 | 0.011 | 0.186 | 0.011 | 0.000 | 0.012 |
| 100 | 3 | 0.499 | 0.306 | 0.308 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4 | 0.503 | 0.473 | 0.462 | 0.000 | 0.000 | 0.004 | 0.000 | 0.000 | 0.000 |
| 2 | 0.497 | 0.481 | 0.483 | 0.079 | 0.011 | 0.186 | 0.011 | 0.000 | 0.012 |

The \( p \)-value has been obtained from the sign rank test on AUC’s from ZILPGM, LPGM, and NPGM over 100 replications.

The graph learning method on count data. So, in this sense, we call a gene differentially expressed in male and female groups if it appears only one of the the estimated networks for male or female groups. For example, ARMCX1 was selected as a node in the network of the male group and CLIC2 was selected in the network of the female group.

The ARMCX1 gene encodes a member of the ALEX family of proteins and is located on the X chromosome. It was reported that downregulated ARMCX1 transcripts have been found to be significantly reduced prostate cancer and may play a role of tumor suppressor gene [28, 29]. CLIC2, a member of the glutathione S-transferase structural family and a suppressor of cardiac ryanodine receptor (RyR2) Ca2+ channels located in the membrane of the sarcoplasmic reticulum, is controlled by redox-dependent processes and would allow to limit cellular damage in terms of oxidative stress [30]. Above mentioned cellular oxidant detoxification and glutathione metabolic process could inhibit age-related deterioration, protect the human neuronal cells, and regulate the expression of many genes primarily involved during immune system activities and inflammatory responses [31].

Following GO functional enrichment analysis, genes differentially expressed in the male group included SLC9A7, PLP2, MAGT1, COX7B, STK26, CYBB, MMGT1, BCAP31, and SLC9A6, whereas genes differentially expressed in the female group were RAB33A and UBQLN2. The differentially expressed genes in the male group
It has been implicated that ion transport pathways may play a key role in the male reproductive potential, such as capacitation and the acrosome reaction, which are critical steps in sperm physiology preparing for fertilization [32]. On the other hand, it has been investigated on the formation of an autophagosome stimulated by oxidative or metabolic stress taking into account the sex/gender disparities in terms of immunity and inflammation [33–35]. Furthermore, these advantages of women in immunity and inflammation have been well known and these phenotypic differences in immune responses from males result from direct genetic differences [34, 36].

**Discussion**

In this paper, we propose a penalized version of zero inflated spatial Poisson regression and derive an efficient EM algorithm built on coordinate descent. On simulated data, our method was shown to yield competitive performances in terms of AUC. Particularly, in the presence of excess zeros, our method outperformed LPGM, which is a state of art method in learning graph structures for count data. Note that one may apply the likelihood ratio test for non-nested hypotheses in [37] in order to test for excess zeros on each node. Also we have applied our method to the chromosome data to infer its network structure. Constructing the networks for different genders, we identified the genes differentially expressed in the male and female groups.

There are several issues we have not addressed in this paper. First, one may study the properties our estimators. For Gaussian graphical models, asymptotic properties of the estimators are rather well studied in the literature. For example, [38] study asymptotic normality and optimalities in the estimation of Gaussian graphical models. Monod

**Table 3** Top ranked genes with their degrees for chromosome data

| ID   | Gene   | Degree | ID   | Gene   | Degree | ID   | Gene   | Degree |
|------|--------|--------|------|--------|--------|------|--------|--------|
| g144 | MID1P1 | 44     | g634 | PBDC1  | 14     | g163 | OTUD5  | 8      |
| g149 | NDUFA1 | 31     | g714 | STS    | 14     | g148 | NSDHL  | 7      |
| g580 | MSN    | 31     | g180 | RAB33A | 12     | g292 | APEX2  | 7      |
| g150 | NDUFB11| 30     | g471 | HSD17B10| 12    | g512 | MAGT1  | 7      |
| g721 | SYN1   | 29     | g520 | MMGT1  | 12     | g673 | RNF113A| 7      |
| g766 | UBQLN2 | 29     | g799 | ZBTB33 | 12     | g390 | COX7B  | 6      |
| g460 | GPC4   | 28     | g314 | BEX3   | 11     | g614 | PNPLA4 | 6      |
| g179 | PIM2   | 22     | g451 | GLUD2  | 11     | g693 | SLC10A3| 6      |
| g677 | SEPT6  | 22     | g472 | HPRT1  | 11     | g87  | GPKOW  | 5      |
| g853 | RPS5Y1 | 22     | g642 | PGRMC1 | 11     | g407 | EBP    | 5      |
| g196 | ARHGEF6| 20     | g654 | PLP2   | 11     | g468 | HCCS   | 5      |
| g241 | TSR2   | 20     | g701 | SLC9A6 | 11     | g658 | PJRY10 | 5      |
| g28  | CXCR3  | 17     | g205 | SASH3  | 10     | g139 | MAGEH1| 4       |
| g207 | SH3BGR1| 17     | g53  | ELK1   | 9      | g21  | BCAP31 | 3      |
| g351 | XCorf21| 17     | g105 | LAGE3  | 9      | g720 | SYAP1  | 3      |
| g185 | RAB9A  | 16     | g315 | BEX4   | 9      |       |        |        |
| g338 | CHST7  | 14     | g54  | ERCC6L | 8      |       |        |        |
[14] provides sufficient conditions for the consistency of the MLE for ZISP model and discusses some properties such as asymptotic normality and efficiency of the MLE. Because our model is based on a penalization of ZISP model, the results in [14] will provide a starting point for studying properties of the estimators. Particularly, in our case, the properties of the estimators for the incidence matrix rather than the coefficients are of interest. Second, our method can be applied to construct biological networks as well as other networks for count data with excess zeros. Examples include user-ratings, spatial incidence of a disease or crime, word-document counts, and others. Third, one may also extend our model to Poisson graphical models with multiple-inflations as in [39]. Still another direction is to generalize our model to other distributions such as negative binomial and gamma distributions.

**Conclusions**

In the present study, expression of ARMCX1 and CLIC2 turned out to be different according to gender. Very little is known about the functional properties of these two genes, this could make ARMCX1 and CLIC2 the possible candidates of medical relevance, such as prostate cancer in male [28, 29] and oxidative stress-related diseases for female [40]. Therefore, further evidences seem to be necessary for identifying gene expression patterns and validating its diagnostic potential that differentiated patients with relevant diseases from healthy controls in each sex in the population-based cohorts and, afterwards, it will be translated to clinical practice with its diagnostic impact.

**Abbreviations**

AUC: Area under the curve; EM: Expectation-maximization; FPR: false positive rate; GPCR: G-protein coupled receptor; LLA: Local linear approximation; LPGM: Local Poisson graphical model; MM: Majorization minimization; PGM: Poisson graphical model; ROC: Receiver operating characteristic; TPR: True positive rate; ZILPGM: Zero-inflated local Poisson graphical model; ZIP: Zero-inflated Poisson; ZISP: Zero-inflated spatial Poisson.

**Acknowledgements**

Not applicable.

**Funding**

The research of S. Won was supported by the Bio-Synergy Research Project (NRF-2017M3A9C4065964) of the Ministry of Science, ICT and Future Planning through the National Research Foundation. The research of Y.J. Kim was supported by the Ministry of Education, Science and Technology of Korea (No. 2012M3A9C4048761). The research of C. Park was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2015R1D1A1A01059984).

**Availability of data and materials**

The chromosomal data can be downloaded from http://jungle.unige.ch/rnaseq_CEU60/ and all the source codes for the current study are available from the corresponding author.

**Authors’ contributions**

Statistical modeling: CP. Algorithm and software development: HC. Data processing and bioinformatics: JG. Biological interpretation: SW, YJK. Simulation and data analysis: SW, SK. Manuscript drafting: CP, HC, YJK. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable.
