A Local Poisson Graphical Model for Inferring Networks from Sequencing Data

Genevera I. Allen and Zhandong Liu

Abstract—Gaussian graphical models, a class of undirected graphs or Markov Networks, are often used to infer gene networks based on microarray expression data. Many scientists, however, have begun using high-throughput sequencing technologies such as RNA-sequencing or next generation sequencing to measure gene expression. As the resulting data consists of counts of sequencing reads for each gene, Gaussian graphical models are not optimal for this discrete data. In this paper, we propose a novel method for inferring gene networks from sequencing data: the Local Poisson Graphical Model. Our model assumes a Local Markov property where each variable conditional on all other variables is Poisson distributed. We develop a neighborhood selection algorithm to fit our model locally by performing a series of $\ell_1$ penalized Poisson, or log-linear, regressions. This yields a fast parallel algorithm for estimating networks from next generation sequencing data. In simulations, we illustrate the effectiveness of our methods for recovering network structure from count data. A case study on breast cancer microRNAs (miRNAs), a novel application of graphical models, finds known regulators of breast cancer genes and discovers novel miRNA clusters and hubs that are targets for future research.

Index Terms—Markov Networks, undirected graphical models, next generation sequencing data, microRNAs, gene regulatory networks

I. INTRODUCTION

Graphical models have become a popular technique to depict and explore relationships between genes and estimate genomic pathways [1], [2]. Undirected graphical models, or Markov Networks, denote conditional dependence relationships between genes [3]. In other words, genes A and B are linked if given the profiles across the subjects for all other genes, the levels of gene A are still predictive of the levels of gene B. Thus, Markov Networks denote a type of direct dependence that is stronger than merely correlated expression values. Many have developed methods to estimate high-dimensional Markov Networks for Gaussian or binary data by using sparsity to select the edges between genes [4], [5], [6], [7]. Several have used these methods for Gaussian graphical models to infer network structures from microarray gene expression data [4], [8]. As typical log ratio expression values from microarray data follow approximately a Gaussian distribution, these models are appropriate. Recently, however, more scientists are using RNA-sequencing or next generation sequencing technologies to measure gene expression or miRNA expression levels, as these methods yield less technological variation than that of microarrays [9]. Measurements from RNA-sequencing, however, are not approximately Gaussian and are in fact read counts of how many times a transcript has been mapped to a specific genomic location. The RNA sequencing expression values are then integer valued and non-negative; thus, many have advocated to model this count data using the Poisson distribution [9], [10], [11]. In this paper, we develop a novel Local Poisson Graphical Model to infer networks based on the Poisson distribution and build an algorithm to estimate these networks from high-throughput sequencing data.

Next generation sequencing technologies have become a powerful tool for measuring expression levels of a variety of transcripts. RNA-sequencing, for example, (or RNA-seq) refers to the process where mRNA strands are isolated and broken into short fragments. The individual base pairs of these fragments are then sequenced and mapped back to the reference genome where information on the number of fragments associated with each gene location are pooled [9]. Other sequencing procedures like miRNA-seq that measure microRNA expression are performed similarly; microRNAs (miRNAs) are \(22\) base pair RNA fragments thought to be post-transcriptional regulators [12]. Thus, once the sequencing reads have been mapped back to the genome, we are left with a non-negative integer count of mapped reads for each gene that reflects the relative expression levels. Some key features of data produced by next generation sequencing are the presence of exact zeros corresponding to no reads being mapped back to a particular genomic location, left-skewed counts, and the presence of a few extreme values [9], [10]. Because of these key features, methods intended for standardized continuous data, like the Gaussian graphical model, are problematic for RNA-seq data. Several have proposed methods based on the Poisson or negative binomial distribution whose support are non-negative; thus, many have advocated to model this distribution. However, as these methods yield less technological variation than that of microarrays [9]. Measurements from RNA-sequencing, however, are not approximately Gaussian and are in fact read counts of how many times a transcript has been mapped to a specific genomic location. The RNA sequencing expression values are then integer valued and non-negative; thus, many have advocated to model this count data using the Poisson distribution [9], [10], [11]. In this paper, we propose graphical model methods based on the Poisson distribution. While some have argued that due to overdispersion this model is not appropriate for sequencing data [15], [16], we find that with proper normalization, discussed in Section IV-A, our proposed Poisson-based models are appropriate. Additionally, well-studied methods for Poisson regression and the single parameter of the distribution will allow us to simplify models and algorithms for network inference.

High-dimensional methods to estimate Gaussian or binary Markov Networks have been well-studied in the literature (see...
[4], [5], [6], [7], for example). Recently, others have proposed non-parametric extensions of these approaches by first Gaussianizing the data, for example with a copula transform, [17], [18] or by using non-parametric rank-based estimators [19], [20]. These existing approaches, however, are ill-suited to count data, such as that of RNA-seq, with many exact zeros, left skewness, and large counts. A few have suggested more direct Markov network estimation methods for count data. [21] propose a combinatorial approach, augmenting the data matrix and fitting log-linear regression models. The computational complexity of the method, however, grows on the order of \( p^{2p} \) where \( p \) is the number of variables; this is infeasible for data with \( p > 20 \). Others have outlined related approaches for multi-way contingency tables [22], [23], [24], [25] that again, are only computationally feasible for a small number of variables. Thus, existing graphical models and algorithmic approaches are not well suited to the high-dimensional count nature of next generation sequencing data.

Our model and algorithmic approach will be based on a proposal of [26], also discussed recently in [27]. This method assumes that node-conditional distributions, or the distribution of each variable conditional on all other variables, is distributed according to a Poisson distribution. While this provides a natural formulation of a Poisson Markov Random Field, [26], [27] note that restrictions on the parameter space, namely negative edge weights, are required to ensure a proper joint distribution on the nodes. As we review in Section II-A, these imply that only negative conditional dependencies can be estimated, an assumption which would not hold for complex genomic data where genes often work cooperatively in genomic pathways. Thus, we consider a relaxation of this model in which we do not restrict the parameter space of the node-conditional distributions. While this may provide a more realistic model in practice, this relaxation unfortunately gives a model that is not a consistent Markov Random Field distribution where the Hammersley-Clifford Theorem [28] and global Markov Property hold. We will show, however, that the Local Markov Property holds, and hence the name Local Poisson Graphical Model. This local property is defined in terms of node-neighborhoods, or the set of nodes connected via an edge to the node of interest; the Local Markov Property specifies independence conditional on the node-neighborhoods [23]. For example, if there is no edge linking genes A and B then this implies that gene A is independent of gene B given the set of other genes connected to gene A. Thus, our model is in the vein of other conditional independence graphical models defined by a series of seemingly unrelated regressions [29], [30], [31], and is still able to capture important direct connections between genes.

Estimating conditional dependencies based on the Local Markov Property specified by neighborhoods leads to a simple algorithmic strategy: estimate the neighborhood of each gene separately. This so-called neighborhood selection algorithm was first proposed by [4] to estimate Gaussian graphical models via a series of \( \ell_1 \)-penalized regression problems, and later in [7] to estimating Ising models via a series of \( \ell_1 \)-penalized logistic regression models. Our algorithmic strategy will follow in similar vein: we propose to estimate each neighborhood via a series of \( \ell_1 \)-penalized Poisson regression models and then combine each of these estimated node-neighborhoods to achieve our network estimate. Thus, our algorithmic approach is based on well-studied neighborhood selection methods which many have shown to yield consistent network estimates for Gaussian, Ising, and multinomial graphical models [4], [7], [32].

This paper is organized as follows. In Section II, we propose our Local Poisson Graphical Model and develop a numerical algorithm to fit this model for high-dimensional data. We examine the network recovery properties of this method in Section III through a simulation study. In Section IV-B, we discuss considerations for applying the LPGM method to next generation sequencing data and present a case study on inferring a miRNA network from miRNA-Seq data, a novel application of graphical models. We conclude with a discussion of our work in Section V.

II. METHODS

We develop the technical framework of our Local Poisson Graphical Model and an algorithm to fit this model in high-dimensional settings. First, however, we review the underlying Poisson Markov Random Field Distribution on which our model is based.

A. Background: The Poisson Markov Random Field

The Poisson Markov Random Field (MRF), first introduced in [26], is a natural extension of the univariate Poisson distribution to the multivariate graphical model setting. Let us define this undirected graph structure as \( G = \{ V, E \} \); that is, the graph, \( G \), consists of the set of vertices (variables), \( V \), and the set of edges (links), \( E \). In the context of genetic networks, each of the vertices or nodes corresponds to a specific gene, and edges denote links or important relationships between genes. Let \( X = \{ X_1, \ldots, X_p \} \) be the \( p \)-dimensional random vector that is associated with the graph structure \( G \). Then, [26] and later [27], define MRFs by specifying a distribution where all node-conditional distributions, or the conditional distribution of one variable given all other variables, follow a univariate exponential family. Taking these node-conditional distributions to be Poisson, the Poisson MRF is specified by the following:

\[
P(X_j | X_k \forall k \neq j; \Theta) \sim \text{Poisson}\left( \exp \left( \theta_j + \sum_{k \neq j} \theta_{jk} X_k \right) \right),
\]

with parameters \( \Theta = (\theta_j, \theta_{jk}, \forall j \neq k \in V) \) where \( \theta_j, \theta_{jk} \in \mathbb{R} \). Here, the parameter \( \theta_j \) is an intercept, adjusting the conditional mean of \( X_j \), and the parameter \( \theta_{jk} \) gives the conditional relationship between nodes \( j \) and \( k \). [26], [27] show via the Hammersley-Clifford theorem that these node-conditional distributions combine to yield the following joint
Poisson MRF distribution:

\[ P(X; \Theta) = \exp \left[ \sum_{j \in V} (\theta_j X_j - \log(X_j!)) + \sum_{(j,k) \in E} \theta_{jk} X_j X_k - \Psi(\Theta) \right]. \tag{2} \]

Here, \( \Psi(\Theta) \) is the log-normalization term that ensures this is a proper distribution.

While constructing a distribution for Poisson graphical models in this manner is a natural extension of the univariate Poisson distribution, there is a major caveat: the parameters of the Poisson MRF are restricted to be non-positive [26], [27]. To see this, note that the log-normalization term \( \Psi(\Theta) \) must be finite to ensure that this distribution sums to one over the set of \( X \in \{0, 1, 2, \ldots, \infty\}^p \). This term is given by

\[ \Psi(\Theta) = \log \left[ \sum_{x_j, x_k \in X} \exp \left( \sum_{j \in V} (\theta_j x_j - \log(x_j!)) + \sum_{j,k \in E} \theta_{jk} x_j x_k \right) \right], \]

where \( X = \{0, 1, \ldots, \infty\} \) is the domain of \( X \). Notice that the term \( \theta_{jk} X_j X_k \) dominates the above summation and must be finite for infinite values of \( X_j \). To ensure this, the parameters denoting the edge weights must be non-positive: \( \theta_{jk} \leq 0 \quad \forall \ j \neq k \) [26], [27]. This implies that the Poisson MRF only permits negative conditional dependencies or only competitive relationships between variables, a severe limitation in practice. Consider for example, that graphical models are often used to estimate regulatory pathways from gene expression data. Thus, conditional on other genes, genes belonging to the same regulatory pathway would be positively correlated. These positive dependencies cannot be captured or estimated via the Poisson MRF and hence, this model is not appropriate for estimating network structures from high-throughput genomic data.

### B. A Local Poisson Graphical Model

We introduce a model based on the Poisson Markov Random Field that will allow us to capture the rich dependence structure expected in genomic networks. Since the joint distribution over all nodes of the Poisson MRF has problematic restrictions, we define our model locally by assuming the same node-conditional Poisson distributions without specifying a joint model. Our Local Poisson Graphical Model (LPGM) is then a series of local models defined for each gene, \( X_j \), and given by the following node-conditional distribution:

\[ P(X_j | X_k \forall k \neq j; \Theta_j) = \exp \left[ \theta_j X_j - \log(X_j!) + \sum_{k \neq j} \theta_{jk} X_j X_k - A(\theta_j, \theta_{jk}) \right]. \tag{3} \]

Here, \( A(\theta_j, \theta_{jk}) \) is the log-partition term of the Poisson distribution. We then propose to define the network structure of our model as the union of the local conditional Poisson models for each gene. These local models can also be written in the form of the log-linear model as the conditional mean of \( X_j \) is given by:

\[ \log \left[ E(X_j | X_k = x_k \forall k \neq j) \right] = \theta_j + \sum_{k \neq j} \theta_{jk} x_k. \tag{4} \]

In our LPGM, if the edge-weight \( \theta_{jk} = 0 \), then this implies that \( X_j \) is conditionally independent of \( X_k \) given all other variables. Thus, since our model is defined in terms of multivariate log-linear regression, the pair-wise Markov property is preserved. Notice also that we can relax the summation in (4) to be over the neighborhood of \( X_j \), denoted as \( N(j) \):

\[ \sum_{k \neq j} \theta_{jk} X_j X_k = \sum_{k \in N(j)} \theta_{jk} X_j X_k \]

because \( \theta_{jk} = 0 \) for nodes \( k \) that are not in the node-neighborhood of node \( j \). Hence, the Local Markov Property, which defines independence between two variables conditional on their neighbors [23], also holds for our model. While the LPGM does not imply a joint density and hence the global Markov Property, it still preserves important conditional independence relationships and allows us to estimate more complex dependence structures expected in genomics data.

### C. LPGM Algorithm

Our main contribution is the development of an algorithm to estimate our Local Poisson Graphical Model from high-throughput genomic sequencing data. Specifically, let us assume we observe data \( X_{ij} \) for \( i = 1, \ldots, n \) independent samples and \( j = 1, \ldots, p \) variables of interest, or genes. Our goal is to infer the network structure, \( G \), that denotes the conditional relationships between pairs of variables. To accomplish this, we propose to fit our model locally by performing neighborhood selection and give a fast algorithm that can be fit in a distributed manner.

1) **Neighborhood Selection:** We propose to locally fit our LPGM by for each node, estimating the set of edges extending out from the node, or the node’s neighborhood. [4] first proposed to automatically select the neighborhood of node \( j \) by placing an \( \ell_1 \)-norm penalty on linear regression coefficients to encourage sparsity. The regression coefficients of variables with weak relationships to variables \( j \) will be shrunk to zero, and there is no edge between the nodes in the graph. Variables with strong relationships with gene \( j \) will have non-zero regression coefficients, and these will be connected to node \( j \) in the graph. Neighborhood selection methods have been developed for high-dimensional graph estimation using \( \ell_1 \)-norm penalized linear regression [4] and logistic regression [7]. We extend this to \( \ell_1 \)-norm penalized log-linear regression for neighborhood selection for our LPGM.

Mathematically, we can write the neighborhood selection problem for node \( j \) as an \( \ell_1 \)-penalized log-linear regression problem. Let \( \Theta \) denote the \( p \times p \) matrix of edge weights \( \{\theta_{jk}\} \) from (3). As in many neighborhood selection algorithms, we are interested in recovering the network structure, and hence ignore the intercept terms, \( \{\theta_j\} \) in (3). Then, a penalized log-linear model is fit with response \( Y = X_j \), predictors \( W = X_{-j} \) (defined as all columns other than \( j \)), and coefficients \( \beta = \Theta_{j, -j} \), or the \( j^{th} \) row (minus diagonals) of our edge
weight matrix:

\[
\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} [Y_i(W_i\beta) - \exp(W_i\beta)] - \rho \|\beta\|_1. \tag{5}
\]

Here, \(\rho \geq 0\) is a regularization parameter controlling the amount of sparsity in the neighborhood. Thus, we estimate the zero elements in one column of our parameter matrix, \(\Theta\), at a time by regressing the \(j^{th}\) variables, \(Y = X_j\) onto all other variables \(W = X_{\setminus j}\). For ease of notation, we denote this estimated column as \(\hat{\Theta}_j(\rho)\) to make explicit the dependency on the regularization parameter, \(\rho\); note that there is no \(j^{th}\) element to this column vector. We denote the estimated graph structure as the adjacency matrix \(\hat{A}(\rho)\) implied by the zero elements in \(\hat{\Theta}(\rho) : \hat{A}(\rho) = [\text{sign}(\hat{\Theta}(\rho))]\). There are many fast computational approaches to fitting these \(\ell_1\)-norm penalized log-linear models [33] that we will discuss further when we introduce our algorithm subsequently.

Finally, notice that neighborhood selection is not symmetric. In other words, while nodes \(j\) and \(k\) may be estimated to have an edge when node \(j\) is regressed on all others, this edge may not be present when node \(k\) is the regresor [4], [7]. Thus, we define our estimated graph, \(\hat{A}(\rho)\), as the union over the set of these edges, noting that the intersection is also appropriate:

\[
\hat{A}_{jk}(\rho) = \max \left\{ |\text{sign}(\hat{\Theta}(\rho)_{jk})|, |\text{sign}(\hat{\Theta}(\rho)_{kj})| \right\}, \quad \forall \, j \neq k. \tag{6}
\]

In other words, an edge connecting nodes \(j\) and \(k\) is estimated if either solving (5) with \(X_j\) or \(X_k\) as regressors yields a non-zero coefficient in the other.

2) Selecting Regularization Parameters: The regularization parameter \(\rho\) controls the sparsity of the graph structure, or in other words, the number of estimated links between nodes. We seek a data-driven method for estimating this parameter. In the Gaussian graphical model literature, many data-driven methods such as cross-validation, BIC, AIC, and stability selection have been proposed. The former three approaches, however, require calculating the log-likelihood. Recall that our model is defined in terms of a series of node-conditional likelihoods that do not yield a joint likelihood. Thus, likelihood-based approaches to selecting the regularization parameter will not work in our settings. Instead, we propose to estimate the regularization parameter via stability selection, an approach which seeks the \(\rho\) leading to the most stable set of edges [34], [8]. In brief, stability selection sub-samples the data \(X^{(b)}\) and estimates a separate graph \(\hat{A}^{(b)}(\rho)\) for each sub-sample and vector of regularization parameters, \(\rho\). The optimal value of \(\rho_{opt}\) controls the average variance over the edges of the sub-sampled graphs [8] (reproduced using our notation for completeness):

\[
\rho_{opt} = \arg\min_{\rho} \left\{ \min_{\lambda \leq \lambda \leq \rho} \frac{1}{2} \sum_{j < k} 2\hat{A}_{jk}(\lambda)(1 - \hat{A}_{jk}(\lambda)) / \begin{pmatrix} p \\ 2 \end{pmatrix} \right\}, \tag{7}
\]

where \(\hat{A}_{jk}(\rho) = \frac{1}{B} \sum_{b=1}^{B} A_{jk}^{(b)}(\rho)\). We note that default values for \(\beta, \beta = 0.05\), and the number of sub-samples, \(m = \lfloor 10\sqrt{n} \rfloor\), from [8] are used throughout this paper.

3) Algorithm: We are interested in developing a fast algorithm to fit our LPGM method to high-throughput genomic data. We accomplish this by incorporating fast path-wise algorithms and stability selection into a parallel computing framework.

First, notice that each of the penalized log-linear models, (5), can be fit independently as the results of each do not depend on others. Thus, the neighborhood of each node can be estimated in parallel. In addition, recent advances in computing \(\ell_1\) penalized models via path-wise coordinate methods over a range of regularization parameters, \(\rho\), allow us to compute the entire neighborhood solution path for each node with approximately the same speed as fitting at a single value of \(\rho\) [33]. Thus, we seek to fit the penalized log-linear models path-wise over a range of regularization parameters in parallel for each node. To accomplish this, the vector of regularization parameters \(\rho\) we consider must be fixed in advance for each node. This means we must know the value of \(\rho_{max}\) at which all coefficients are zero for all nodes, or in other words, no edges are estimated in the graph. Examining the Karush-Kuhn-Tucker conditions of (5), the minimum value of \(\rho\) at which no edges are selected for \(X_j\) is \(\max_{k \neq j} |X_j^T X_j|\). Hence, \(\rho_{max} = \max_{k \neq j} |X_j^T X_j|\), the maximum over all the \(j\) regression problems.

**Algorithm 1** Local Poisson Graphical Model Algorithm

1) Set \(\rho_{max} = \max_{k \neq j} |X_j^T X_j|\). Fix \(\rho_{min} \approx 1.0 \times 10^{-4}\). Define 100 log-spaced values \(\rho = [\rho_{max} \ldots \rho_{min}]^T\).

2) For each \(X_j\), \(j = 1, \ldots, p\), do in parallel:
   a) Solve (5) with regressor \(Y = X_j\) and predictors \(W = X_{\setminus j}\) path-wise for \(\rho\) yielding \(\hat{\Theta}(\rho)\).
   b) For \(b = 1, \ldots, B\):
      i. Sample \(m = \lfloor 10\sqrt{n} \rfloor\) observations, yielding the sub-sampled data, \(X^{(b)}\).
      ii. Solve (5) with regressor \(Y = X_j^{(b)}\) and predictors \(W = X_{\setminus j}^{(b)}\) path-wise for \(\rho\) yielding \(\hat{\Theta}^{(b)}(\rho)\).

3) Determine the graphs \(\hat{A}(\rho)\) from \(\hat{\Theta}(\rho)\) and \(\hat{A}^{(b)}(\rho)\) from \(\hat{\Theta}^{(b)}(\rho)\) via (6).

4) Determine \(\rho_{opt}\) via stability selection, (7).

5) Return the graph, \(\hat{A}(\rho_{opt})\).

Algorithm 1 summarizes these items and the steps of our LPGM estimation method. Notice that the entire set of computations including path-wise log-linear models and stability selection are performed in parallel for each node. This dramatically reduces the computational complexity to approximately \(O(1 + B)p^2\) for each node [33]. After this, stability selection results are combined to estimate the optimal regularization parameter and the final graph is determined via maximum edge agreement. Thus, our algorithm is a computationally efficient method for inferring networks from high-dimensional count data.
We assess the performance of our LPGM for selecting the correct underlying network structure based on simulated count data. Three graph structures are simulated: (i) a hub network, where each node is connected to one of three hub nodes, (ii) a scale-free network, in which the number of nodes of a certain degree follow a power law, and (iii) a random network, in which the presence of each edge has equal probability. The hub and scale-free networks are known to mimic the behavior of biological networks. Our LPGM is compared to the Graphical Lasso algorithm [6], the Graphical Lasso after applying a log transform to the data plus one, and two non-parametric extensions of the Graphical Lasso, the Non-paramormal rank-based skeptic estimator [19]. Unlike for Gaussian graphical models and Ising models, simulating Poisson networks is not a trivial task; we employ an approach based on [35]. In brief, $n$ independent observations from our simulated Poisson network with $p$ nodes, $X \in \{0, 1, \ldots, \infty\}^{n \times p}$, are generated from the following model: $X = Y B + E$. Here, $Y$ is an $n \times p(p-1)/2$ matrix with each element $y_{ij} \sim \text{Poisson}(\lambda_{\text{true}})$ and $E$ is $n \times p$ with $e_{ij} \sim \text{Poisson}(\lambda_{\text{noise}})$. The matrix $B$ encodes the true underlying graph structure denoted by the adjacency matrix $A \in \{0, 1\}^{p \times p}$; $B = [I_p] \odot (1_p \text{tri}(A))^T$. Here, $P$ is the $p \times p(p-1)/2$ pair-wise permutation matrix, $\odot$ denotes the Hadamard or element-wise product, and $\text{tri}(A)$ denotes the $p(p-1)/2 \times 1$ vectorized upper triangular portion of the adjacency matrix $A$. We simulate $n = 200$ observations for $p = 50$ nodes at two signal-to-noise (SNR) levels. We set $\lambda_{\text{true}} = 1$ with $\lambda_{\text{noise}} = 0.5$ for the high SNR level and $\lambda_{\text{noise}} = 5$ for the low SNR level.

Results of our experiments conducted over fifty replicates are given in Figure 1. Both receiver operator curves (ROC) computed by varying the regularization parameter $\rho$ and boxplots of true and false positive rates for fixed $\rho$ estimated via stability selection are given. True positives are estimated as the fraction of edges found by LPGM that are in the true simulated network structure $A$; false positives are estimated analogously. These results indicate that LPGM uniformly outperforms Gaussian-based competitors for the hub and scale-free graphs. The improved statistical power of our LPGM for recovering the hub graph structure is particularly striking. The ROC curves of all methods on the random graph structure are approximately equal. When stability selection is used to estimate the sparsity level, however, we see that LPGM retains its advantage over the competitors. This behavior is not surprising as employing the correct statistical model, in this case the LPGM, often leads to improved model selection.

Experimental results for high-dimensional scenarios with the number of variables, $p = 100$, greater than the number of samples, $n = 50$, are given under the same simulation settings in Figure 2. The high-dimensional results reveal the same trend as observed in our first simulation study; namely, the LPGM outperforms all competing methods, and especially so...
distributed data. LPGM for recovering network structures based on Poisson simulation results demonstrate the strong performance of our method for discovering relationships between miRNAs as sparse graph estimation methods are known to break down for hub networks in high-dimensional settings. Overall, these simulation results demonstrate the strong performance of our LPGM for recovering network structures based on Poisson distributed data.

Fig. 3. Normalization results for the breast cancer miRNA-Seq data. In part a, histograms of four example miRNAs are shown of the raw data (top, blue) and after normalization (bottom, red). Normalization preserves the exact zeros and extreme examples of skewness. A histogram of the overall normalized data is shown in part b, and a comparison to the quantiles of the Poisson distribution via a q-q plot in part c.

for hub and scale-free networks. This is particularly important as sparse graph estimation methods are known to break down for hub networks in high-dimensional settings. Overall, these simulation results demonstrate the strong performance of our LPGM for recovering network structures based on Poisson distributed data.

IV. INFERRING NETWORKS FROM NEXT GENERATION SEQUENCING DATA

Now that we have presented our Local Poisson Graphical Model and have studied the performance of this method for network recovery with simulated data, we discuss considerations for applying our model to infer genetic networks from sequencing data, Section IV-A, and study the performance of our method for discovering relationships between miRNAs based on breast cancer miRNA-Seq data, Section IV-B.

A. Normalization and Poisson Models for Sequencing Data

High-throughput RNA-sequencing technologies quantify expression values by mapping short reads of cDNA back to the reference genome. The resulting data consist of counts at each genomic location that are non-negative integers [9]. Given this, many have sought to model sequencing data via the Poisson [13], [10], [11] or negative binomial [15], [16] distributions whose support is on the set of non-negative integers. Many have proposed pre-processing and normalization procedures specific to these distributions. Indeed, normalization to a parametric distribution is especially important for large-scale inference to find differential gene or isoform expression. For our Local Poisson graphical model, however, we do not require strict adherence to a Poisson distribution for each gene and/or sample, and instead require that the node-conditional distributions are approximately Poisson. Here, we discuss normalization procedures for RNA-seq or microRNA-seq data, paying particular attention to how this pre-processing affects network inference via our algorithm.

First, there are several general items one must consider when pre-processing high-throughput sequencing data. The samples may contain vastly different numbers of total read counts reflecting technological variation in sequencing depths with no biological relevance [9], [36]. Some have suggested to normalize samples by the total counts [9], the RPKM (reads per KB per million) [36], [13], or more robust methods such as normalizing via the geometric mean [16] or quantiles [10]. Another characteristic of sequencing data is that the read counts for some genomic locations may have zero or nearly zero expression values across all the samples acquired [36]. As these genes and others that are constant across the samples will not be meaningful genes to study via network models, we suggest to filter out these genes. Next for many studies, batch effects may be present that can confound the results. One can adjust for these effects via global methods such as principal components analysis or univariate methods tested via ANOVA-like models [37]. We choose to detect and possibly adjust for batch effects using Poisson ANOVA models after we have normalized the sequencing data to more closely follow a Poisson distribution as described below.

After these general pre-processing steps for sequencing data have been completed, we consider the modeling assumptions of our Local Poisson Graphical Model. Recall from (3), our model assumes that each gene, \( X_j \) conditional on its gene neighbors is Poisson distributed for all genes \( j = 1, \ldots, p \):

\[
(X_j | X_k \; \forall k \in N(j)) \sim \text{Poisson} \left( \exp \left( \sum_{k \in N(j)} \theta_{jk} X_k \right) \right)
\]

In contrast to other Poisson models for sequencing data [13], [10], [11], our model assumes a conditional Poisson distribution. Practically, it is difficult to assess the appropriateness of this assumption with real sequencing data as the true gene neighbors are unknown. We can however, assess the appropriateness of the null model where a gene has no neighbors; this model implies that each gene should follow a Poisson distribution with constant mean. Many have suggested, however, that sequencing data appears overdispersed compared to this constant Poisson model [15], [16], [11]. We choose to adjust for this by transforming the data via a power \( \alpha \in (0, 1) \) where \( \alpha \) is chosen to yield approximately Poisson data [11] as assessed via goodness-of-fit or Kolmogorov-Smirnov tests. This power transformation to correct for overdispersion has another advantage in the context of our LPGM. Notice that if neighboring genes \( X_k \) have extremely large, or overdispersed counts, then the exponential causes the conditional Poisson mean to become large. Thus, correcting for overdispersion which subsequently limits the extreme counts as seen in Figure 3, also improves the fit of our model. Also, as noted in [11], this power transform results in non-integer data, but this does not pose a problem for our LPGM algorithm which
Fig. 2. Experimental high-dimensional simulation study for three network structures: hub, scale-free, and random graphs. For each graph type, high-dimensional Poisson networks are generated with \( n = 50 \) observations and \( p = 100 \) nodes at a high and low signal-to-noise ratio (SNR). Our LPGM is compared to the Graphical Lasso (Glasso) [6], the Glasso on the log-transformed data, the Non-Paranormal Gaussianized through a copula transform (NPN-Copula) [17], and the Non-paranormal rank-based Skeptic method (NPN-Skeptic) [19]. Parts a, b, and c, show receiver-operator curves at high and low SNRs obtained by varying the regularization parameter, \( \rho \). Parts d, e, and f, show boxplots of true and false positive rates for fixed \( \rho \) estimated by stability selection for the High SNR simulation. Our LPGM method outperforms competing approaches and is still able to recover some network structure in difficult high-dimensional regimes.

**Algorithm 2 Normalization Steps for Sequencing Data**

**General Normalization Steps:**

1. Adjust for sequencing depth.
2. Filter out genes with low variance across the samples.
3. Adjust for possible overdispersion.
4. Detect and possibly adjust for batch effects.

**Detailed Steps for miRNA-Seq Case Study:**

1. Used 75% quantile matching [10].
2. Selected top 25% most variable miRNAs.
3. Used a power transform \( X^\alpha \) for \( \alpha \in [0,1] \) [11] with \( \alpha \) selected via the minimum Kolmogorov-Smirnov statistic.
4. Fitted a Poisson ANOVA model for each miRNA with batch labels as predictors, and adjusted for multiple testing. No significant batch association was detected.

In summary, we employ four common steps to normalizing high-throughput sequencing data: (i) adjust for sequencing depth, (ii) filter out genes with low variance across samples, (iii) adjust for possible overdispersion, and (iv) detect and possibly adjust for batch effects. These steps are outlined in Algorithm 2, as well as the specific processing details employed for the miRNA-Seq data used in the case study presented subsequently. The results of these normalization procedures can be seen for the case study miRNA-Seq data in Figure 3. Notice that normalization preserves exact zeros and left-skewness for particular miRNAs as seen in the top portion of Figure 3, and overall, yields data that approximately follows the Poisson distribution as seen by the q-q plot in the bottom portion of Figure 3.

**B. Case Study: Networks via miRNA-Seq Data**

We apply our LPGM algorithm to discover relationships among microRNAs (miRNAs) based on sequencing data from breast cancer patients. There is a long record of applying Markov Networks to understand gene expression data, but inferring networks based on miRNAs is a novel application of graphical models. Level III breast cancer data was obtained from the Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/) [38]. This data set consists of 544 patients and 524 miRNAs. The sequencing data was normalized as described in Section IV-A with 50% of the miRNAs that varied the least across the samples filtered out, giving us 262 miRNA nodes.

In Figure 4, we present the results of our LPGM algorithm applied to the breast cancer miRNA sequencing data. Analysis of this network reveals results consistent with the breast cancer genomics literature as well as novel biomarkers and clusters to investigate further. First, however, notice from
Fig. 4. Breast cancer miRNA network estimated by Local Poisson Graphical Models (left). This network is scale-free as demonstrated by the power-law plot (top right) of node degree on the log-scale versus the number of nodes for such degree. Our LGPM found many hub genes previously associated with breast cancer such as let-7c, and identified new potential regulators of breast cancer such as miR-379 which is tightly correlated with let-7c (bottom right). A miRNA cluster (left, boxed) was also identified by our LPGM in an unsupervised manner without using transcript location.

the top right panel of Figure 4 that the estimated network closely follows a power law in the number of nodes at each node degree, and thus appears to be a scale-free network. Many biological networks, such as gene expression networks, protein-protein interaction networks, and metabolic networks, have been observed to be scale-free [39]; thus, we can add miRNA expression networks to this list.

Many of the hub miRNAs identified in our LPGM network such as let-7c, miR-10b, and miR-375 have been previously associated with breast cancer progression and metastasis. For example, let-7c has been shown to regulate the breast cancer metastatic [40]. High level expression of miR-10b has been observed in triple negative, ER negative, PR negative, Her2 negative, breast cancer patients [41]. Silencing miR-10b has been proposed as potential therapeutic target and tested in mouse mammary gland tumor models [42]. Blocking miR-375 in ER-positive cancer cells can slow down the cancer cell growth [43]. Other hub miRNAs identified in our network are novel biomarkers that need to be validated further for associations in breast cancer. Consider, for example, miR-379 which forms an edge and is tightly correlated, bottom right panel of Figure 4, with another hub miRNA, let-7c. There have been no studies on the functionality of miR-379, but based on its hub status in the LPGM network and its connections with other studied miRNAs, we hypothesize that miR-379 is a regulatory miRNA for breast cancer progression and metastasis.

Our results also indicate interesting sub-network modules related to miRNA clusters and functional regulatory pathways. We identified a large miRNA cluster in the right, boxed portion of Figure 4 which contains miR-516a-1, miR-521-1, miR-522, miR-519a-1, and miR-527 from chromosome 19:54251890-54265684 [+]. Many have established that miRNAs appear in clusters on a single polycistronic transcript [44]. The expression levels of precursor miRNAs in the same cluster are synchronized and coordinated by similar transcription factors. Mature miRNAs levels, however, are regulated independently. As sequencing technologies measure mature miRNA levels and we did not incorporate any outside information such as transcript location, we would not necessarily expect to find these miRNA clusters. Interestingly, our LPGM network identifies a major miRNA cluster, indicating that perhaps these miRNAs are functionally related, regulating similar biological processes in breast cancer.

We also compare the structure of our estimated LPGM network on this breast cancer miRNA-Seq data to that of existing techniques, namely the graphical lasso algorithm [6] for Gaussian graphical models. A structural comparison in terms of a comparative adjacency matrix and circle graphs is given in Figure 5. To ensure a fair comparison, the sparsity
level of both the LPGM and graphical lasso algorithms were selected via stability selection with a fixed parameter $\beta = 0.05$ [8]. First, notice that the graphical lasso estimates many fewer edges than the LPGM algorithm. As stability selection has been shown to control the false positives [34], [8], this reveals that the LPGM algorithm likely has more statistical power to discover edges for sequencing data than the graphical lasso, results consistent with our findings in the simulated networks. Also, the graphical lasso algorithm identifies some of the same structure and features in the miRNA network, but misses some hub miRNAs that have important roles in breast cancer such as miR-10b, identified via our LPGM algorithm.

Overall, the novel application of our Local Poisson Graphical Model to understand breast cancer miRNA networks has yielded results consistent with the known literature and identified potential biomarkers and pathways for future research. Furthermore, our algorithm demonstrates numerous advantages over existing Gaussian-based methods for inferring networks from sequencing data.

V. DISCUSSION

We have developed a novel framework for estimating high-dimensional graphical models from count data. While our Local Poisson Graphical Model does not jointly define a Markov Random Field, it retains the pair-wise and local Markov properties, and permits a rich set of dependencies among variables. Network inference is achieved locally through fitting $\ell_1$ penalized log-linear models to select the neighborhood of each node. Through simulations and a miRNA case study, we have demonstrated the effectiveness of our LPGM for estimating network structure from high-throughput count data.

As graphical models for count data have received relatively little attention compared to those for Gaussian and discrete data, our work leads to many potential areas of new methodological work. [27] study the theoretical properties of graphical models defined by exponential families, which encompasses the Poisson MRF, showing that consistent network recovery can be achieved. Further work could also include studying the network recovery properties of the LPGM algorithm as well as determining whether our set of local models determine any global distribution, proper or improper, on the joint set of nodes. Also, a proper joint MRF distribution for count data that does not severely restrict the conditional dependencies as with the Poisson MRF would be an important contribution.

There are many potential applications of our model and algorithm. We have presented a case study on miRNA networks, but clearly LPGM will be useful for constructing gene expression networks from RNA-sequencing data as well. A major consideration when applying our model to sequencing data is proper normalization to ensure that the samples are approximately Poisson distributed. In particular, as our model is parametric, it is sensitive to overdispersion, which commonly occurs with RNA-sequencing data. Thus, it is our strong recommendation to follow the normalization steps described in Section IV-A. Furthermore, a statistical test or diagnostic for understanding whether the LPGM is an appropriate model in practice would help to ensure proper model fit. Additionally, our novel application using undirected graphs to study miRNA networks yields a new method to examine miRNAs in groups instead of the more common approaches of studying the genomic targets of a single miRNA. Further work to biologically validate our predicted miRNA biomarkers and clusters in breast cancer is needed to gain a more complete picture of the regulatory process in this disease. Beyond genomics, there are many potential applications of LPGM to multivariate Poisson distributed data such as that from user-ratings, web site visits, advertising clicks, bibliometrics, and social networks.

In conclusion, our work developing the Local Poisson Graphical Model and algorithm for high-throughput sequencing data has many implications and has opened new directions for research both in the area of high-dimensional graphical models and in the application of these to gene expression and miRNA expression networks.
ACKNOWLEDGMENTS

The authors thank Pradeep Ravikumar for thoughtful insights and helpful discussion related to this work. This work is supported in part through the Collaborative Advances in Biomedical Computing seed funding program at the Ken Kennedy Institute for Information Technology at Rice University supported by the John and Ann Doerr Fund for Computational Biomedicine and through the Center for Computational and Integrative Biomedical Research Seed Funding Program at Baylor College of Medicine. GA is also partially supported by NSF DMS-1209017.

REFERENCES


