

A Log-Linear Graphical Model for Inferring Genetic Networks from High-Throughput Sequencing Data

Genevera I. Allen^{1,2} & Zhandong Liu¹

¹ Department of Pediatrics-Neurology, Baylor College of Medicine

& Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital,

² Department of Statistics, Rice University.

May 30, 2012

Abstract

Gaussian graphical models are often used to infer gene networks based on microarray expression data. Many scientists, however, have begun using high-throughput sequencing technologies to measure gene expression. As the resulting high-dimensional count data consists of counts of sequencing reads for each gene, Gaussian graphical models are not optimal for modeling gene networks based on this discrete data. We develop a novel method for estimating high-dimensional Poisson graphical models, the *Log-Linear Graphical Model*, allowing us to infer networks based on high-throughput sequencing data. Our model assumes a pair-wise Markov property: conditional on all other variables, each variable is Poisson. We estimate our model locally via neighborhood selection by fitting ℓ_1 -norm penalized log-linear models. Additionally, we develop a fast parallel algorithm, an approach we call the *Poisson Graphical Lasso*, permitting us to fit our graphical model to high-dimensional genomic data sets. In simulations, we illustrate the effectiveness of our methods for recovering network structure from count data. A case study on breast cancer microRNAs, a novel application of graphical

models, finds known regulators of breast cancer genes and discovers novel microRNA clusters and hubs that are targets for future research.

Keywords: Markov networks, graphical models, Hammersley-Clifford theorem, next generation sequencing data, microRNAs, regulatory networks

1 Introduction

Graphical models have become a popular technique to depict and explore relationships between genes and estimate genomic pathways (Dobra et al., 2004; Krämer et al., 2009). Undirected graphical models, or Markov Networks, denote conditional dependence relationships between genes (Dempster, 1972). 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 (Meinshausen and Buhlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2007; Ravikumar et al., 2010). Several have used these methods for Gaussian graphical models to infer network structures from microarray gene expression data (Meinshausen and Buhlmann, 2006; Liu et al., 2010). As typical log ratio expression values from microarray data follow approximately a Gaussian distribution, these models are appropriate. Recently, more scientists are using RNA-sequencing technologies to measure gene expression or microRNA levels, as these methods in theory yield less technological variation than that of microarrays (Marioni et al., 2008). 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 (Marioni et al., 2008; Bullard et al., 2010; Li et al., 2011). In this paper, we develop a novel Log-Linear Graphical Model based on the Poisson distribution and

build an algorithm to estimate genomic networks from high-dimensional sequencing data.

High-dimensional methods to estimate Gaussian or binary Markov Networks have been well-studied (Meinshausen and Buhlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2007; Ravikumar et al., 2010). Few, however, have introduced methods for Poisson Markov Networks. Hastie et al. (2009) 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 $p^2 2^p$ 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 (Madigan et al., 1995; Lauritzen, 1996; Bishop et al., 2007; Wainwright and Jordan, 2008) that again, are only computationally feasible for a small number of variables. Our goal in this paper is to develop a model and computationally attractive algorithm to estimate high-dimensional Poisson graphical models that can be used to infer genetic networks based on RNA-sequencing data.

In this paper, we make the several novel contributions: (1) propose a Log-Linear Graphical Model based on a pair-wise Poisson Markov Network; (2) introduce a neighborhood selection approach to infer network structure locally via a series of ℓ_1 penalized log-linear models; and (3) build a fast parallel algorithm to fit our graphical model to high-dimensional genomic data. These methods are developed in detail in Section 2. In Section 3.1, we study the utility of our methods for recovering the underlying graph structure of simulated Poisson networks. We apply our LLGM to infer high-dimensional networks for breast cancer microRNAs, a novel applied contribution, in Section 3.2. In Section 4, we conclude with a discussion of the implications of our work and directions for future research.

2 Methods

We develop a log-linear graphical model and a fast algorithm to fit this model, the Poisson Graphical Lasso, that can be used to estimate genomic networks from high-dimensional sequencing data. We begin with background and considerations for modeling sequencing

data using the Poisson distribution.

2.1 Poisson Models for Sequencing Data

High-throughput RNA-sequencing technologies quantify expression values by mapping short reads of cDNA back to the original genome. The resulting data consist of counts at each genomic location that are non-negative integers (Marioni et al., 2008). Several models have been proposed for this data including Poisson and negative binomial models (Anders and Huber, 2010; Robinson and Oshlack, 2010; Li et al., 2011). In this paper, we will assume that after normalization the data follows a Poisson distribution with a separate mean for each gene.

There are several items one must consider when normalizing high-throughput sequencing data. First, the samples may contain vastly different numbers of total read counts reflecting technological variation in sequencing depths with no biological relevance (Marioni et al., 2008; Mortazavi et al., 2008). Some have suggested to normalize samples by the total counts (Marioni et al., 2008), the RPKM (reads per KB per million) (Mortazavi et al., 2008; Jiang and Wong, 2009), or more robust methods such as normalizing via the geometric mean (Anders and Huber, 2010) or quantiles (Bullard et al., 2010). Another characteristic of sequencing data is that the read counts for some genomic locations may have zero or nearly zero expression values (Mortazavi et al., 2008). As these genes and others that are constant across the samples will not be meaningful genes to study via network models, we filter out these genes. Finally, many have noted that even after adjusting for sequencing depth and filtering genes, the data can still be overdispersed compared to a Poisson distribution (Robinson and Oshlack, 2010; Anders and Huber, 2010; Li et al., 2011). We choose to adjust for this by transforming the data via a power $\alpha \in (0, 1]$ where α is chosen to yield approximately Poisson data (Li et al., 2011). While others have advocated using the negative binomial distribution in this context (Robinson and Oshlack, 2010; Anders and Huber, 2010), the Poisson distribution has a number of advantages for graphical models. These include

a simple one-parameter form and well established methods for fitting penalized log-linear models.

In summary we employ three common steps to normalizing high-throughput sequencing data: (i) adjust for sequencing depth, (ii) filter out genes with low variance across samples, and (iii) adjust for possible overdispersion. After this normalization, we assume that the data follows a Poisson distribution with a separate mean for each gene. Specifically, if we have sequencing data X_{ij} for $i = 1, \dots, n$ samples and $j = 1, \dots, p$ genes, then we assume that for each gene j : $X_{1,j}, \dots, X_{n,j} \stackrel{iid}{\sim} \text{Poisson}(\lambda_j)$.

2.2 A Log-Linear Graphical Model

We develop the mathematical framework for our Log-Linear Graphical Model by assuming pair-wise conditional Poisson relationships between variables and defining a Poisson Markov Network. Interestingly, this joint distribution on the nodes places severe constraints on the types of pair-wise conditional relationships between variables that have little meaning in the context of genetic networks. Thus, we propose to estimate unrestricted relationships between pairs of variables locally and then put together this set of local networks. While this procedure does not fit a joint model on all the nodes, it will allow us to infer a more general graph structure. To denote the difference between this set of local models and the joint Poisson Markov Network, we use *llgm* and *LLGM* respectively.

First, let us define the undirected network structure as $\mathcal{G} = \{V, E\}$; that is, the graph, \mathcal{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 \mathbf{X} be a matrix of n samples measured on p genes with entries taking on integer values, $X_{ij} \in \{0, 1, 2, \dots, \infty\}$.

Our model is characterized by conditional Poisson relationships between pairs of nodes: $X_j | X_k \sim \text{Poisson} \ \forall j \neq k \in V$, where X_j is the vector of n samples measured for variable j . Then, both the *llgm* and *LLGM* models are defined in terms of this conditional density

for each node:

$$p(X_j | X_k = x_k \ \forall k \neq j, \Theta) \sim \text{Poisson} \left(e^{\theta_j + \sum_{k \neq j} \theta_{jk} x_k} \right), \quad (1)$$

with parameters $\Theta = (\theta_j, \theta_{jk}, \forall j \neq k \in V)$ where $\theta_j, \theta_{jk} \in \mathfrak{R}$. Here, the parameter θ_j is an intercept, adjusting the conditional mean of X_j , and the parameter θ_{jk} gives the conditional relationship between nodes j and k . Note that (1) denotes pair-wise relationships or the local Markov property (Lauritzen, 1996).

Recall that for the local, pair-wise Markov property to hold jointly for all nodes to form a Poisson Markov Random field, the conditions of the Hammersley-Clifford theorem must be satisfied (Hammersley and Clifford, 1971). This result states that the probability density of the graph may be factored into a product of potentials on the set of maximal cliques. Besag (1974) went on to define the conditions under which Markov random fields can be defined for exponential families. From this, the probability density of the Poisson Markov Network, we call this our global *LLGM*, is given by the following:

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

Here, $\Psi(\Theta)$ is the log-partition function: $\Psi(\Theta) = \log \left[\sum_{x_j \in \{0,1,\dots,\infty\}, x_k \in \{0,1,\dots,\infty\}} \exp \left(\sum_{j \in V} (\theta_j x_j - \log(x_j!)) + \sum_{j,k \in E} \theta_{jk} x_j x_k \right) \right]$. This term acts as a normalizing constant ensuring that (2) is a proper probability density function; that is, it sums to one. This requires that $\Psi(\Theta) < \infty \ \forall \ \mathbf{X}$. As the term $\theta_{jk} X_j X_k$ dominates the above summation, this implies that $\theta_{jk} \leq 0 \ \forall \ j \neq k$ (Besag, 1974). Thus, the Poisson Markov Network, *LLGM*, is only defined for conditionally negatively correlated relationships between variables.

Due to this restriction on the parameter space of the *LLGM*, this model has limited applicability. 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 have a positively correlated relationship. These positive

dependencies cannot be captured or estimated via the *LLGM* model. Hence, this model is not ideal for estimating network structures from high-throughput genomic data.

Therefore, we propose to estimate the local model, *llgm*, for each node and then combine each of these estimated local models to infer a network structure. Returning to (1), we will assume that the intercept term, θ_j , is zero as we assume each genomic variable has been adjusted for sequencing depth as described in the previous section. We are then left with the following log-linear model, the inspiration for the name *llgm*:

$$\log [\mathbb{E}(X_j | X_k = x_k \ \forall k \neq j)] = \sum_{k \neq j} \theta_{jk} x_k. \quad (3)$$

Estimating the parameters of this model, $\theta_{jk} \ \forall (j, k) \in V$, is sufficient for inferring the local network structure of the graph. Notice, for example, that if $\theta_{jk} = 0$, then this implies that $X_j \perp X_k | X_{\neq j, k}$. In other words, variables j and k are conditionally independent given all other variables. In our graph structure, \mathcal{G} , this local conditional independence implies that there is no edge between X_j and X_k . This approach is closely related to the neighborhood selection problem proposed for Gaussian graphical models and Ising models in Meinshausen and Buhlmann (2006) and Ravikumar et al. (2010) respectively. The major difference between our approach for Poisson graphical models and these existing methods is that estimating the pair-wise conditional dependencies for the Ising and Gaussian graphical model are sufficient for determining the joint dependence structure of the random field. While this does not hold for our models, our local approximation will allow us to estimate a richer set of dependence structures.

2.3 Poisson Graphical Lasso

We describe our method for fitting the local Log-Linear Graphical Model, *llgm*, an algorithm we call the *Poisson Graphical Lasso* named after the Graphical Lasso algorithm of Friedman et al. (2007) for Gaussian Graphical Models.

2.3.1 Neighborhood Selection

We propose to fit the local Log-Linear Graphical Model, *llgm*, by for each node, estimating the set of edges extending out from the node, of the node's *neighborhood*. Meinshausen and Buhlmann (2006) first proposed to automatically select the neighborhood of node j by placing an ℓ_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 ℓ_1 -norm penalized linear regression (Meinshausen and Buhlmann, 2006) and logistic regression (Ravikumar et al., 2010). We extend this to ℓ_1 -norm penalized log-linear regression for neighborhood selection for the local *llgm*.

Mathematically, we can write the neighborhood selection problem for node j as the solution to the following penalized log-linear regression problem:

$$\underset{\Theta_{\neq j,j}}{\text{maximize}} \quad \frac{1}{n} \sum_{i=1}^n [X_{ij} (X_{i,\neq j} \Theta_{\neq j,j}) - \exp (X_{i,\neq j} \Theta_{\neq j,j})] - \rho \|\Theta_{\neq j,j}\|_1. \quad (4)$$

Here, $\rho \geq 0$ is a regularization parameter controlling the amount of sparsity in the neighborhood and the notation $X_{i,\neq j}$ denotes the i^{th} row of \mathbf{X} and all columns other than column j , and analogously for $\Theta_{\neq j,j}$. Thus, we estimate the zero elements in one column of our parameter matrix, Θ , at a time by regressing the j^{th} variables, X_j onto all other variables $X_{\neq j}$. For ease of notation, we denote this estimated column as $\hat{\Theta}_j(\rho)$ to make explicit the dependency on the regularization parameter, ρ ; note that there is no j^{th} element to this column vector. We denote the estimated graph structure as the adjacency matrix $\hat{\mathbf{A}}(\rho)$ implied by the zero elements in $\hat{\Theta}(\rho) : \hat{\mathbf{A}}(\rho) = |\text{sign}(\hat{\Theta}(\rho))|$. There are many fast computational approaches to fitting these ℓ_1 -norm penalized log-linear models (Friedman et al., 2010) 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 regressor (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010). Thus, we define our estimated graph, $\hat{\mathbf{A}}(\rho)$, as the union over the set of these edges, noting that the intersection is also appropriate:

$$\hat{A}_{jk}(\rho) = \max \left\{ |\text{sign}(\hat{\Theta}(\rho)_{jk})|, |\text{sign}(\hat{\Theta}(\rho)_{kj})| \right\} \quad \forall j \neq k. \quad (5)$$

In other words, an edge connecting nodes j and k is estimated if either solving (4) with X_j or X_k as regressors yields a non-zero coefficient in the other.

2.3.2 Selecting Regularization Parameters

The regularization parameter ρ 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 local algorithm based on the *llgm* does not maximize the joint likelihood of (2). We then, propose to estimate the regularization parameter via stability selection, an approach which seeks the ρ leading to the most stable set of edges (Liu et al., 2010). In brief, stability selection sub-samples the data $\mathbf{X}^{(b)}$ and estimates a separate graph $\mathbf{A}^{(b)}(\rho)$ for each sub-sample and vector of regularization parameters, ρ . The optimal value of ρ controls the average variance over the edges of the sub-sampled graphs (Liu et al., 2010) (reproduced using our notation for completeness):

$$\rho_{opt} = \underset{\rho}{\operatorname{argmin}} \left\{ \underset{0 \leq \lambda \leq \rho}{\operatorname{minimize}} \left\{ \sum_{j < k} 2\bar{A}_{jk}(\lambda)(1 - \bar{A}_{jk}(\lambda)) / \binom{p}{2} \right\} \leq \beta \right\}, \quad (6)$$

where $\bar{A}_{jk}(\rho) = \frac{1}{B} \sum_{b=1}^B A_{jk}^{(b)}(\rho)$. We note that default values for β , $\beta = .05$, and the number of sub-samples, $m = \lfloor 10\sqrt{n} \rfloor$, from (Liu et al., 2010) are used.

2.3.3 Algorithm

We are interested in developing a fast algorithm to fit our *llgm* model 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, (4), 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 ℓ_1 penalized models via path-wise coordinate methods over a range of regularization parameters, $\boldsymbol{\rho}$, allow us to compute the entire neighborhood solution path for each node with approximately the same speed as fitting at a single value of ρ (Friedman et al., 2010). 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 $\boldsymbol{\rho}$ we consider must be fixed in advance for each node. This means we must know the value of ρ_{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 (4), the minimum value of ρ at which no edges are selected for X_j is $\max_{k \neq j} |X_k^T X_j|$. Hence, $\rho_{max} = \max_{j, k \neq j} |X_k^T X_j|$, the maximum over all the j regression problems.

Algorithm 1 summarizes these items and the steps of our Poisson Graphical Lasso 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 $O((1+B)p^3)$ for each node (Friedman et al., 2010). 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 Poisson Graphical Lasso Algorithm is a computationally efficient method for inferring the high-dimensional *llgm*.

Algorithm 1 Poisson Graphical Lasso for *llgm*

1. Normalize the data \mathbf{X} as described in Section 2.1. Set $\rho_{max} = \max_{k,j} |X_{k,\neq j}^T X_j|$. Fix $\rho_{min} \approx 1.0 \times 10^{-4}$. Define 100 log-spaced values $\boldsymbol{\rho} = [\rho_{max} \ \dots \ \rho_{min}]^T$.
 2. For each X_j , $j = 1, \dots, p$, do in parallel:
 - (a) Solve (4) with regressor X_j and predictors $X_{\neq j}$ path-wise for $\boldsymbol{\rho}$ yielding $\hat{\Theta}_j(\boldsymbol{\rho})$.
 - (b) For $b = 1, \dots, B$:
 - i. Sample $m = \lfloor 10\sqrt{n} \rfloor$ observations, yielding the sub-sampled data, $\mathbf{X}^{(b)}$.
 - ii. Solve (4) with regressor $X_j^{(b)}$ and predictors $X_{\neq j}^{(b)}$ path-wise for $\boldsymbol{\rho}$ yielding $\hat{\Theta}_j^{(b)}(\boldsymbol{\rho})$.
 3. Determine the graphs $\hat{\mathbf{A}}(\boldsymbol{\rho})$ from $\hat{\Theta}(\boldsymbol{\rho})$ and $\hat{\mathbf{A}}^{(b)}(\boldsymbol{\rho})$ from $\hat{\Theta}^{(b)}(\boldsymbol{\rho})$ via (5).
 4. Determine ρ_{opt} via stability selection, (6).
 5. Return the graph, $\hat{\mathbf{A}}(\rho_{opt})$.
-

3 Results

We evaluate the performance of our local Log-Linear Graphical Model for recovering network structure via experiments on simulated data and through a novel application to microRNA sequencing data.

3.1 Experiments on Simulated Networks

We assess the performance of our *llgm* for selecting the correct underlying network 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 each edge has equal probability. The hub and scale-free networks are known to mimic the behavior of biological networks. Our *llgm* is compared to the Graphical Lasso algorithm (Friedman et al., 2007) and the Graphical Lasso after applying a log transform to the data plus one. Unlike for Gaussian graphical models and Ising models, simulating Poisson networks is not a trivial task; we employ an approach based on Karlis (2003). In brief, n independent observations from our

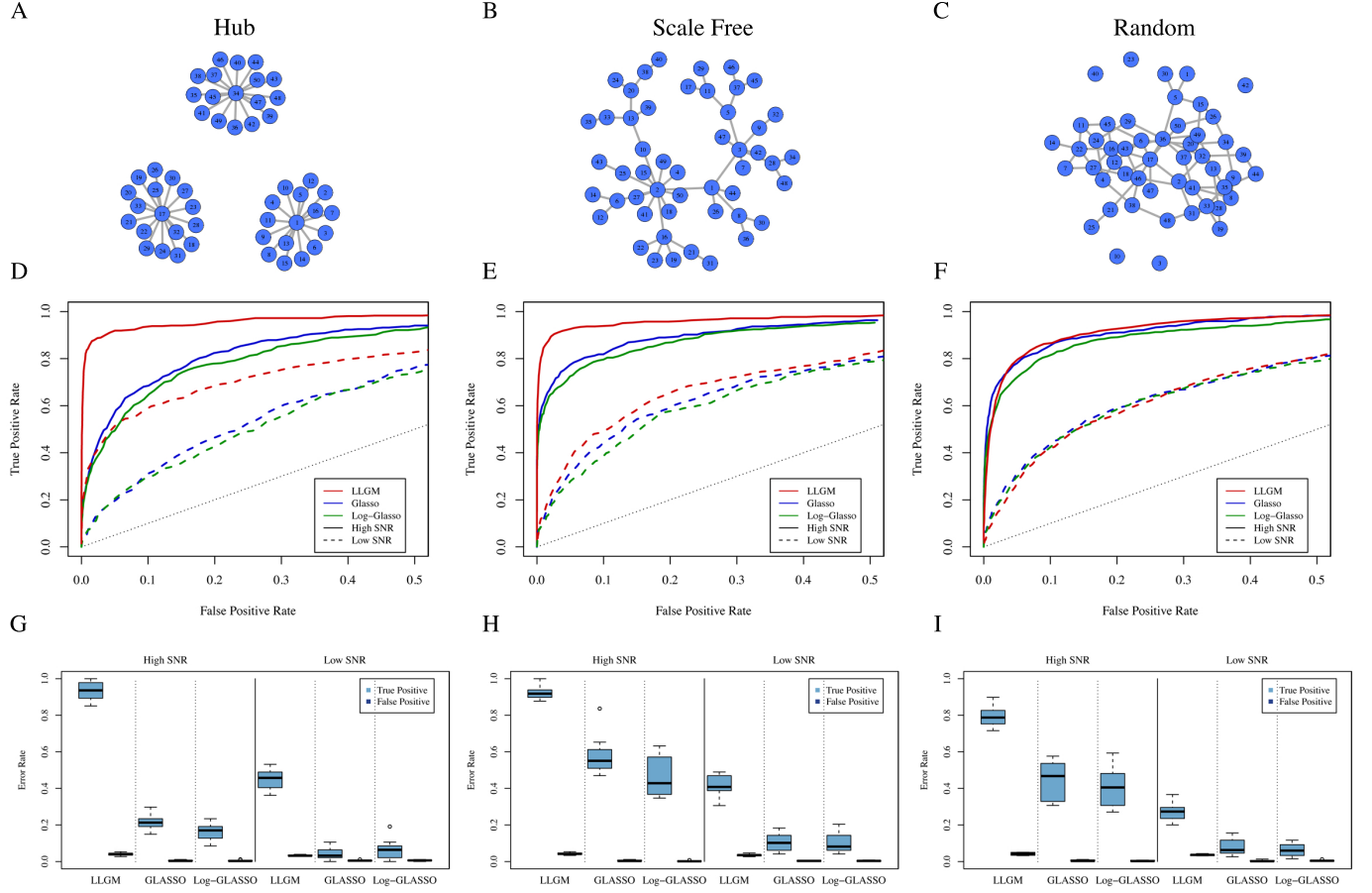


Figure 1: Experimental simulation study for three network structures: hub (A), scale-free (B), and random (C). For each type of graph, Poisson networks are generated with 200 observations and 50 nodes at a high and low signal-to-noise ratio (SNR). Our *llgm* is compared to the Graphical Lasso and the Graphical Lasso on the log-transformed data through Receiver Operator Curves (D-F) obtained by varying the regularization parameter, ρ , and boxplots (G - I) of true and false positive rates (G-I) for fixed ρ estimated by stability selection. Our *llgm* outperforms competing methods for all three simulated network structures.

simulated Poisson network with p nodes, $\mathbf{X} \in \mathbb{R}^{n \times p}$, are generated from the following model: $\mathbf{X} = \mathbf{Y}\mathbf{B} + \mathbf{E}$. Here, \mathbf{Y} is a $n \times p + p(p-1)/2$ matrix with each element $Y_{ij} \stackrel{iid}{\sim} \text{Poisson}(\lambda_{true})$ and \mathbf{E} is $n \times p$ with $E_{ij} \stackrel{iid}{\sim} \text{Poisson}(\lambda_{noise})$. The matrix \mathbf{B} encodes the true underlying graph structure denoted by the adjacency matrix $\mathbf{A} \in \{0, 1\}^{p \times p}$: $\mathbf{B} = [\mathbf{I}_{(p)}; \mathbf{P} \odot (\mathbf{1}_{(p)} \text{tri}(\mathbf{A})^T)]^T$. Here, \mathbf{P} is the $p \times p(p-1)/2$ pair-wise permutation matrix, \odot denotes the Hadamard or element-wise product, and $\text{tri}(\mathbf{A})$ denotes the $p(p-1)/2 \times 1$ vectorized upper triangular portion of the adjacency matrix \mathbf{A} . We simulate $n = 200$ observations for $p = 50$ nodes at

two signal-to-noise (SNR) levels. We set $\lambda_{true} = 1$ with $\lambda_{noise} = 0.5$ for the high SNR level and $\lambda_{noise} = 5$ for the low SNR level.

Results of our experiments conducted over ten replicates are given in Figure 1. Both receiver operator curves (ROC) computed by varying the regularization parameter ρ and boxplots of true and false positive rates for fixed ρ as estimated via stability selection are given at the high and low SNR levels. True positives are estimated as the fraction of edges found by *llgm* that are in the true simulated network structure \mathbf{A} ; false positives are estimated analogously. These results indicate that *llgm* uniformly outperforms the Gaussian graphical models for the hub and scale-free graphs. The improved statistical power of our *llgm* 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 *llgm* retains its advantage over the competitors. This behavior is not surprising as employing the correct statistical model, in this case the *llgm*, often leads to improved model selection. Overall, these simulation results demonstrate the strong performance of our *llgm* for recovering network structures based on Poisson distributed data.

3.2 Discovering microRNA Networks

We apply our *llgm* to discover relationships among microRNAs 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 microRNAs 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/>) (Collins and Barker, 2007). This data set consists of 544 patients and 524 microRNAs. The sequencing data was normalized as described in Section 2.1 with 50% of the microRNAs that varied the least across the samples filtered out, giving us 262 microRNA nodes.

In Figure 2, we present the results of our *llgm* applied to the breast cancer microRNA

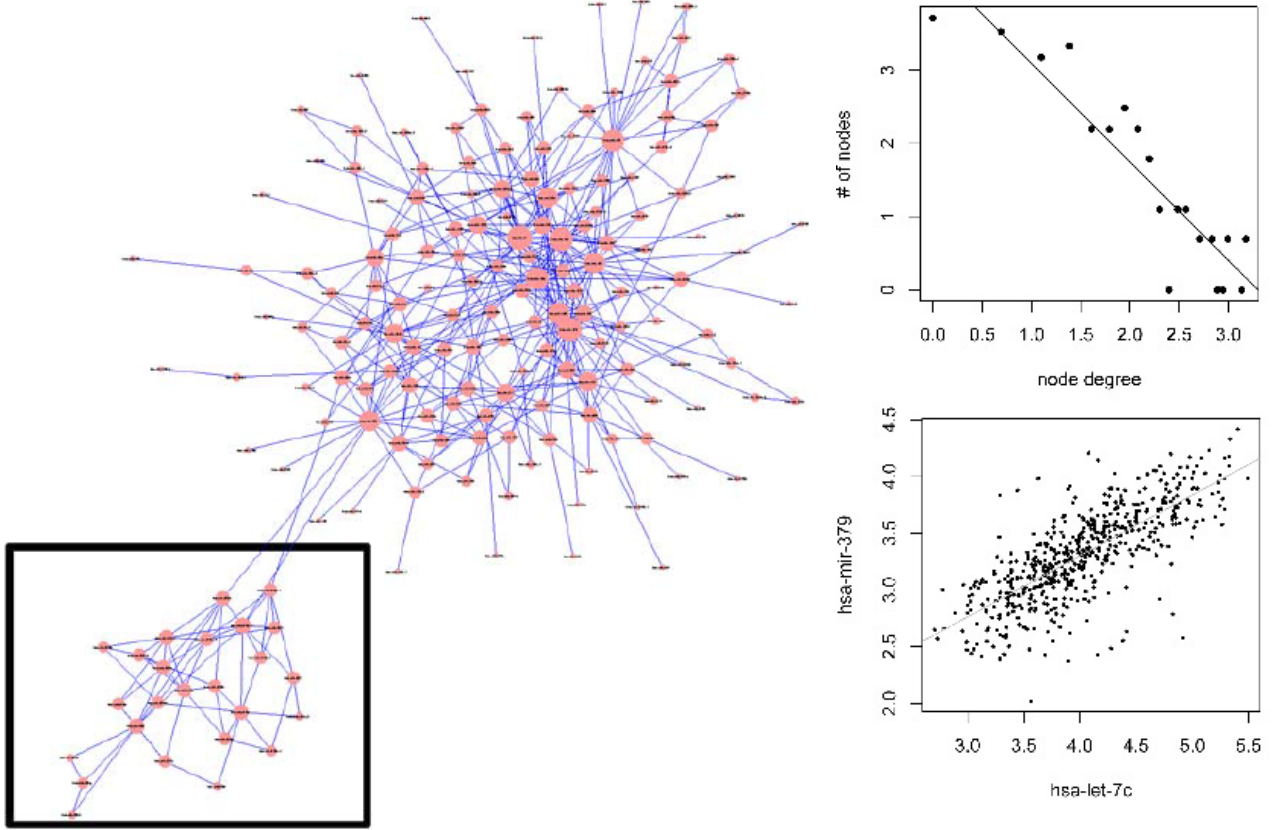


Figure 2: Breast cancer microRNA network estimated by *llgm* (left). This network is scale-free as demonstrated by the power-law plot (top right) of node degree on the log-scale verses the number of nodes for such degree. Our *llgm* found many hub genes previously associated with breast cancer such as let-7c, and identified new potential regulators of breast cancer such as mir-379 which is tightly correlated with let-7c (bottom right). A microRNA cluster (left, boxed) was also identified by our *llgm* in an unsupervised manner without using transcript location.

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 the top right panel of Figure 2 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 (Barabási and Albert, 1999); thus, we can add microRNA expression networks to this list.

Many of the hub microRNAs identified in our *llgm* 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 (Yu et al., 2007). High level expression of mir-10b has been observed in triple negative, ER negative, PR negative, Her2 negative, breast cancer patients (Radojicic et al., 2011). Silencing mir-10b has been proposed as potential therapeutic target and tested in mouse mammary gland tumor models (Ma et al., 2010). Blocking mir-375 in ER-positive cancer cells can slow down the cancer cell growth (de Souza Rocha Simonini et al., 2010). Other hub microRNAs identified in our *llgm* 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 2, with another hub microRNA, let-7c. There have been no studies on the functionality of mir-379, but based on its hub status in the *llgm* and its connections with other studied microRNAs, we hypothesize that mir-379 is a regulatory microRNA for breast cancer progression and metastasis.

Our results also indicate interesting sub-network modules related to microRNA clusters and functional regulatory pathways. We identified a large microRNA cluster in the right, boxed portion of Figure 2 which contains has-mir-516a-1, has-mir-521-1, hsa-mir-522, has-mir-519a-1, and has-mir-527 from chromosome 19:54251890-54265684 [+]. Many have established that microRNAs appear in clusters on a single polycistronic transcript (Bentwich et al., 2005). The expression levels of precursor microRNAs in the same cluster are synchronized and coordinated by similar transcription factors. Mature microRNAs levels, however, are regulated independently. As sequencing technologies measure mature microRNA levels and we did not incorporate any outside information such as transcript location, we would not necessarily expect to find these microRNA clusters. Interestingly, our *llgm* identifies a major microRNA cluster, indicating that perhaps these microRNAs are functionally related, regulating similar biological processes in breast cancer.

Overall, the novel application of our local Log-Linear Graphical Model to understand breast cancer microRNA networks has yielded results consistent with the known literature and identified potential biomarkers and pathways for future research.

4 Discussion

We have developed a novel framework for estimating high-dimensional graphical models with Poisson distributed data. While we have defined the global Poisson Markov Network, we have chosen to estimate network structure locally, thus permitting a richer set of dependencies among variables. This is achieved through fitting ℓ_1 penalized log-linear models to select the neighborhood of each node. Through simulations and a microRNA case study, we have demonstrated the effectiveness of our *llgm* for estimating network structure from high-throughput data.

Our work leads to many areas of new methodological work. These include further studying the global *LLGM* and establishing theoretical properties such as consistent graph recovery and estimation of model parameters. Also, a joint density defined on all the nodes that does not severely restrict conditional relationships as in the *LLGM* would be an important contribution. Extensions of graphical models to encompass other distributions is another direction for future research.

There are many potential applications of our Poisson graphical model and algorithm. We have presented a case study on microRNA networks, but clearly *llgm* 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) independently and identically Poisson distributed. In particular, as our model is parametric, it is sensitive to zero-inflation and overdispersion, both of which commonly occur with RNA-sequencing data. Thus, it is our strong recommendation to follow the normalization steps described in Section 2.1; that is, to filter out non-variable genes and adjust for overdispersion. Additionally, our novel application using undirected graphs to study microRNA networks yields a new method to examine microRNAs in groups instead of the more common approaches of studying the genomic targets of a single microRNA. Further work to biologically validate our predicted microRNA 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 *llgm* 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 Log-Linear Graphical Models 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 microRNA 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.

References

- Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biol* 11(10), R106.
- Barabási, A. and R. Albert (1999). Emergence of scaling in random networks. *science* 286(5439), 509.
- Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, et al. (2005). Identification of hundreds of conserved and nonconserved human micrornas. *Nature genetics* 37(7), 766–770.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2), 192–236.
- Bishop, Y., S. Fienberg, and P. Holland (2007). *Discrete multivariate analysis*. Springer Verlag.
- Bullard, J., E. Purdom, K. Hansen, and S. Dudoit (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics* 11(1), 94.

- Collins, F. and A. Barker (2007). Mapping the cancer genome. *Scientific American Magazine* 296(3), 50–57.
- de Souza Rocha Simonini, P., A. Breiling, N. Gupta, M. Malekpour, M. Youns, R. Omranipour, F. Malekpour, S. Volinia, C. M. Croce, H. Najmabadi, S. Diederichs, O. Sahin, D. Mayer, F. Lyko, J. D. Hoheisel, and Y. Riazalhosseini (2010, November). Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor alpha in breast cancer cells. *Cancer research* 70(22), 9175–9184.
- Dempster, A. P. (1972). Covariance selection. *Biometrics* 28(1), 157–175.
- Dobra, A., C. Hans, B. Jones, J. Nevins, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* 90(1), 196–212.
- Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the lasso. *Biostatistics* 9(3), 432–441.
- Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1.
- Hammersley, J. and P. Clifford (1971). Markov fields on finite graphs and lattices.
- Hastie, T., R. Tibshirani, and J. J. H. Friedman (2009). *The elements of statistical learning* (2 ed.). Springer.
- Jiang, H. and W. Wong (2009). Statistical inferences for isoform expression in rna-seq. *Bioinformatics* 25(8), 1026–1032.
- Karlis, D. (2003). An em algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics* 30(1), 63–77.
- Krämer, N., J. Schäfer, and A. Boulesteix (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC bioinformatics* 10(1), 384.
- Lauritzen, S. (1996). *Graphical models*, Volume 17. Oxford University Press, USA.
- Li, J., D. Witten, I. Johnstone, and R. Tibshirani (2011). Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*.
- Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Arxiv preprint arXiv:1006.3316*.
- Ma, L., F. Reinhardt, E. Pan, J. Soutschek, B. Bhat, E. G. Marcusson, J. Teruya-Feldstein, G. W. Bell, and R. A. Weinberg (2010, April). Therapeutic silencing of miR-10b inhibits metastasis in a mouse mammary tumor model. *Nature biotechnology* 28(4), 341–347.
- Madigan, D., J. York, and D. Allard (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 63(2), 215–232.
- Marioni, J., C. Mason, S. Mane, M. Stephens, and Y. Gilad (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* 18(9), 1509–1517.

- Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3), 1436–1462.
- Mortazavi, A., B. Williams, K. McCue, L. Schaeffer, and B. Wold (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods* 5(7), 621–628.
- Radojicic, J., A. Zaravinos, T. Vrekoussis, M. Kafousi, D. A. Spandidos, and E. N. Stathopoulos (2011, February). MicroRNA expression analysis in triple-negative (ER, PR and Her2/neu) breast cancer. *Cell cycle (Georgetown, Tex.)* 10(3), 507–517.
- Ravikumar, P., M. Wainwright, and J. Lafferty (2010). High-dimensional l1 model selection using l1-regularized logistic regression. *The Annals of Statistics* 38(3), 1287–1319.
- Robinson, M. and A. Oshlack (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol* 11(3), R25.
- Wainwright, M. and M. Jordan (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1-2), 1–305.
- Yu, F., H. Yao, P. Zhu, X. Zhang, Q. Pan, C. Gong, Y. Huang, X. Hu, F. Su, J. Lieberman, and E. Song (2007, December). let-7 regulates self renewal and tumorigenicity of breast cancer cells. *Cell* 131(6), 1109–1123.
- Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* 94(1), 19.