Genetic Networks

Korbinian Strimmer
IMISE, Universität Leipzig

Seminar: Statistical Analysis of RNA-Seq Data

19 June 2012

Paper

G. I. Allen and Z. Liu. 2012.
A log-linear graphical model for inferring genetic networks from high-throughput sequencing data.
ArXiv 1204.3941.

Overview

❶ Background

❷ Network Reconstruction (GGM and Poisson)

❸ Simulation Study

❹ Analysis of microRNA Data

❺ Discussion

Genetic Networks

- Most analysis of RNA-Seq data (e.g. differential expression, clustering, classification) ignores the dependencies among genes.
- In contrast, in genetics networks one is specifically interested in these dependencies.

Questions:

1. What are suitable statistical models for dependency networks for count data?
2. How do we learn these networks from actual biological data?

Graphical Models

For *microarray* data a common way to model dependencies are graphical models, e.g., **Gaussian Graphical Models** (GGMs).

For sequence data a similar approach is needed: **Poisson graphical model**.

Allen and Liu propose a to use a **log-linear graphical model** (llgm) similar to regression-based GGMs and develop a fast algorithm based on lasso regression suitable for estimation from high-dimensional data.

Previous Work on Poisson Graphical Models

There exist some literature on graphical models for count data and contingency tables, for example:

- Whittaker 1990
- Madigan et al 1995
- Lauritzen 1996
- Hastie et al 2009

However, all these algorithms do not work well for large number of variables. Inference for dimension $d > 20$ is infeasible.

Allen and Liu (2012) address this issue by introducing the llgm algorithm.

Poisson vs. Negative Binomial Model and Preprocessing

Allen and Liu use the Poisson distribution rather than the Negative Binomial.

But overdispersion is accounted for in preprocessing:

1. genes with zero counts, that are constant or with low variance are filtered out.
2. adjustment for sequence depth via scale factors (e.g. Anders and Huber 2012).
3. power transform X^α with $\alpha \in [0; 1]$ to correct overdispersion.

Log-Linear Model

Conventional linear model:

$$\mu = E(Y|X_i = x_i) = \sum \beta_i x_i$$

with normal error.

log linear model:

$$\log \mu = \log E(Y|X_i = x_i) = \sum \beta_i x_i$$

with Poisson error
(in GLM speak: Poisson regression with natural log link function)

Log-Linear Model: Properties

- automatically ensures that $\mu > 0$
- the predictors x_i need not be integers (preprocessing!)
- effects of predictor are multiplicative, as

$$\mu = \prod e^{\beta_i x_i}$$

- the regression coefficients can be estimated by ML, penalized ML (e.g. lasso or elastic net) or Bayesian approaches.

Gaussian Graphical Model: Basics

Starting point:

- genes $X_1, \ldots, X_d$ are jointly normal distributed with mean μ and covariance Σ and corresponding correlation matrix $P = (\rho_{ij})$

From P we compute partial correlations $\tilde{P}$:

- $\Omega = P^{-1} = (\omega_{ij})$
- $\tilde{\rho}_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$

A vanishing partial correlation coefficient $\tilde{\rho}_{ij} = 0$ implies (for normal data) conditional independence of gene i and j given all other genes.

Non-zero coefficients are represented by edges → GGM network.

GGM: Regression View

Partial correlation between X_1 and X_2 can also be computed by linear regression:

$$E(X_1|X_j = x_j)_{j\neq 1} = \sum_{j\neq 1} \beta_j^1 x_j$$

$$E(X_2|X_j = x_j)_{j\neq 2} = \sum_{j\neq 2} \beta_j^2 x_j$$

and

$$\tilde{r}_{ij}^2 = \hat{\beta}_2^1 \hat{\beta}_1^2$$

Partial correlation is the geometric mean of the two regression coefficients (one for each direction of an edge in a network).

GGM: Neighborhood Selection

Meinshausen and Bühlmann (2006) propose the *neighborhood selection* approach to inference of GGM networks:

- for each potential edge between X_i and X_j estimate the corresponding regression coefficients $\hat{\beta}_i^j$ and $\hat{\beta}_j^i$ using L1-penalized regression ("lasso").
- lasso has built-in variable selection: coefficients can be exactly zero.
- include an edge in the graph if both coefficients are non-zero (alternative: if at least one of them is non-zero).

Advantages: very fast and can be applied to very high dimensions.
Drawback: this procedure does not always produce a consistent global joint distribution (e.g. the resulting implied covariance is not guaranteed to be positive definite.

llgm Algorithm

Inspired by GGM neighborhood selection Allen and Liu propose their local llgm (log-linear graphical model) algorithm:

- use L1-penalized log-linear regression to estimate regression coefficients.
- optimal regularization parameter is chosen via *stability selection* (Meinshausen and Bühlmann 2010).
- construct a llgm network by including an edge between if at least of the two regression coefficients corresponding to an edge is non-zero (union). Alternatively, include an edge only if both coefficients are non-zero (intersection).

Advantages: very fast and can be applied to very high dimensions.
Drawback: this procedure does not necessarily produce a consistent global Poisson graphical model.

Simulations: Setup

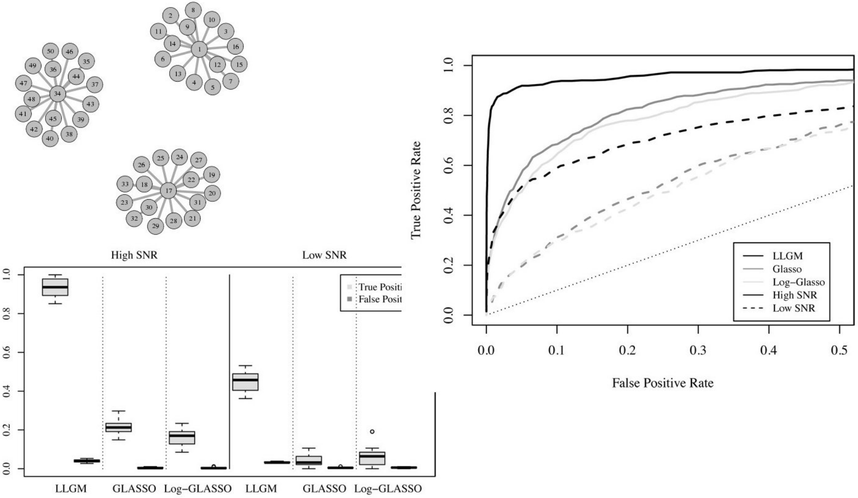
Three graphs structures are simulated (50 nodes):

1. hub network
2. scale-free network
3. random network

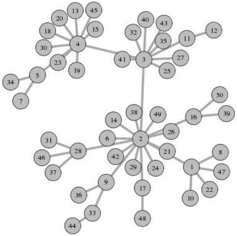
Poisson data with sample size $n = 200$ for these networks are simulated using an algorithm by Karlis (2003).

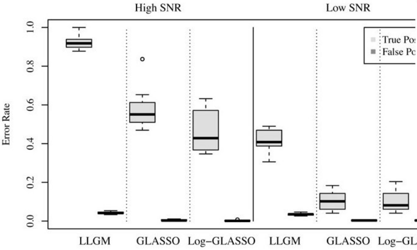
Comparison with GGM lasso algorithm (directly on count data or on log-transformed count data).

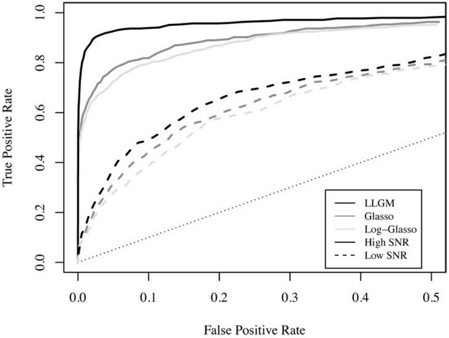
Simulations: Hub Network



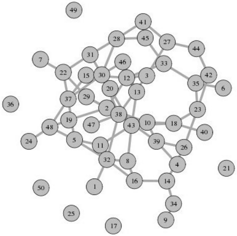
Simulations: Scale-Free Network

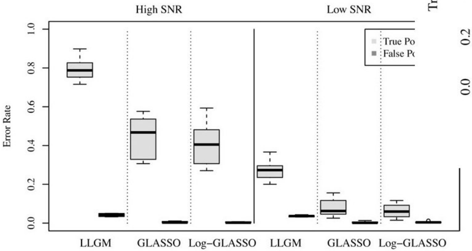


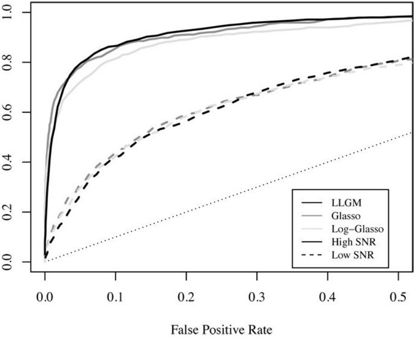




Simulations: Random Network







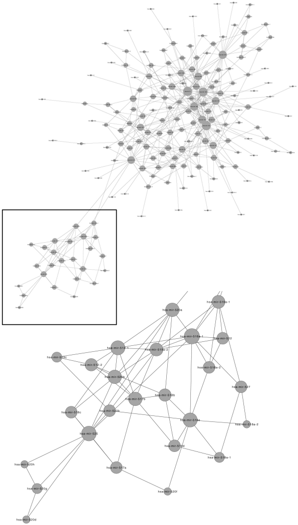
Simulations: Results

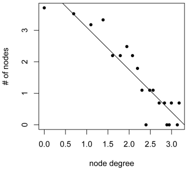
- the llgm algorithm greatly outperforms GGM-based algorithms on hub and scale-free networks.
- for GGM graphs it does not matter whether the data are log-transformed.
- for random networks the ROC curves of all methods are approximately equal.

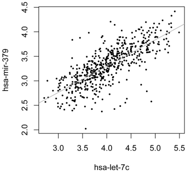
microRNA Data Set

- aim: infer network to discover relationship among microRNAs (breast cancer samples).
- data set: 544 patients and 524 microRNAs.
- after preprocessing and filtering 262 microRNAs remained for analysis ($n = 544$, $d = 262$).

Inferred microRNA Network







microRNA Network Details

- Node-degrees follows a power-law (scale free network).
- many well-known hub genes are recovered.
- plus additional potentially interesting hub genes.
- micoRNA cluster identified without transcript location

Biological hypothesis obtained from network reconstruction:
mir-379 is a regulatory microRNA for breast cancer progression.

Discussion

- A framework for inferring Poisson graphical networks from count data was developed.
- Based on Poisson L1-penalized regression combined with neighborhood selection.
- Applicable to much higher dimension than previous algorithms.
- The proposed approach clearly outperforms in simulations GGM networks inferred from the same data.
- Using a microRNA data set previously known facts were recovered and new biological hypotheses were generated for further validation.