Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Gene expression

Analyzing the Performance of High Dimensional Nonparanormal Graphical Models with RNA-Seq Data

Viswesh Periyasamy [1,*]

[1] Department of Computer Science, University of Wisconsin-Madison, Madison, 53706, USA.

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

This study investigates the nonparanormal transformation of data and how it relates to the problem of gene regulatory network inference. Specifically, it investigates several graph estimation methods with an applied nonparanormal transformation to RNA-Seq data as compared to a greedy hill-climbing approach. These experiments are conducted in a high-dimensional setting, where the number of genes highly outweighs the number of samples (a common setting in computational biology). Results show that the nonparanormal approach severely underperforms in this setting as compared to the greedy approach, however further trials with a variety in parameter selection are required for conclusive results.

Contact: vperiyasamy@wisc.edu

Supplementary information: Supplementary data and code available on the biostat servers at `/u/medinfo/handin/bmi776/project_supplement/vperiyasamy/`.

1 Introduction

Graphical models which make a multivariate Gaussian assumption on the relationships between their nodes have been traditionally successful in the domain of gene regulatory network inference problems. Specifically, this normality assumption, although it is not an inherent property, has been shown to be well-suited for microarray gene expression data due to the data's continuous (as opposed to discrete) nature. However, when examining RNA-Seq data, it is normally required to transform the data to conform with this normality assumption - the raw data itself is count-based and incompatible with a Gaussian model. Instead, a network inference method which relaxes this normality constraint might prove to be more appropriate. Previous work has explored using other distributions, such as Poisson distributions, to model the data (Gierliński *et al.*, 2015). However, using a nonparanormal transformation on RNA-Seq data would allow use of graphical models which inherently make a Gaussian assumption to perform graph estimation (Liu *et al.*, 2009). This in turn could lead to more accurate reconstructions in the context of gene regulatory networks, and the aim of this study is to verify this hypothesis.

1.1 RNA-Seq expression data

With traditional microarray expression data, standard normality assumptions are valid for a variety of network inference methods. However, RNA-Sequencing (RNA-Seq) methods for quantifying gene expression have become increasingly widespread to tackle several of the issues that arise with microarray expression data. Specifically, RNA-Seq is a much more attractive approach for measuring gene expression because it requires no reference sequence, its values have a larger dynamic range with low noise, and has high reproducibility (Li *et al.*, 2015).

1.2 Gaussian graphical model estimation

Because RNA-Seq data measures mRNA abundance using discrete counts, it must be transformed in order to be applicable to Gaussian graphical models. The gene regulatory network inference problem can then be framed as a determination of non-zero entires within the inverse covariance matrix, i.e. the precision matrix $\Omega = \Sigma^{-1}$. Such non-zero entries represent edges (conditional dependencies) between genes and conversely zeros represent absent edges (conditional independence) (Meinshausen *et al.*, 2006). Thus, graph estimation from a data set conforming to a multivariate Gaussian distribution can be easily computed by solving for the precision matrix when the number of samples n is sufficiently larger than the dimensionality of the data d. As described in Liu *et al.*, 2009, this can be done using standard maximum likelihood as the covariance matrix Σ remains positive definite under these conditions. Unfortunately in the context of gene regulatory network inference, it is often the case that the

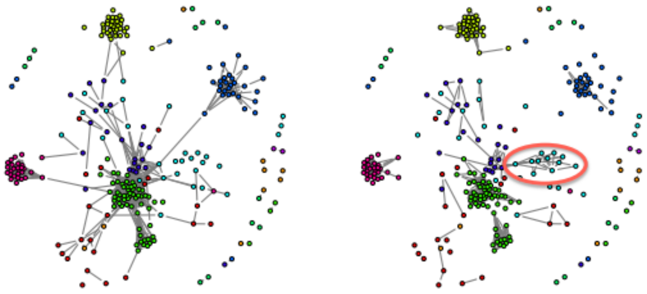

Fig. 1. Two estimated graphs using the glasso method with no transformation (left) and nonparanormal transformation (right) on the data. Figure adapted from Zhao *et al.*, 2012

number of genes is substantially larger than the number of samples (i.e. $d \gg n$) - this is known as a high dimensional setting. Maximum likelihood is no longer suitable as the rank of Σ is now at most n (Liu *et al.*, 2009). Methods such as the graphical lasso proposed by Friedman *et al.*, 2008 and Meinshausen-Bühlmann estimation proposed by Meinshausen *et al.*, 2010 have proven to be effective in this setting but still make a Gaussian assumption on the distribution of the data. This assumption is restrictive and motivates a nonparametric extension of the variables, coined by Liu *et al.*, 2009 as the "nonparanormal" distribution.

1.3 Nonparanormal distributions

To move to a nonparanormal distribution, the Gaussian can be replaced with a semiparametric Gaussian copula. This can be done by transforming the variables (i.e. genes) $X = (X_1, \ldots, X_d)$ with a multivariate Gaussian function $f(X) = (f_1(X_1), \ldots, f_d(X_d))$ which results in a nonparametric extension of the Gaussian (Liu *et al.*, 2009). Thus, this nonparanormal distribution $X \sim NPN(\mu, \Sigma, f)$ has its own mean and covariance which can used to calculate the precision matrix and consequently any conditional dependencies. Figure 1 shows an example of a cluster of nodes with edges predicted by the nonparanormal graphical lasso method that are absent in the standard graphical lasso method. This illustrates how a nonparanormal transformation can reveal dependencies beyond the normality assumption (Zhao *et al.*, 2012). In the context of gene regulatory network inference, a nonparanormal transformation of RNA-Seq expression data could prove to be much more effective when applied as input to standard graphical estimation methods.

2 Methods

In order to adequately evaluate the efficacy of a nonparanormal transformation in the context of gene regulatory network inference, RNA-Seq data is fed through a variety of graph estimation methods for analysis. Specifically, this study considers Meinshausen-Bühlmann graph estimation, *glasso* or graphical lasso estimation, and a greedy hill-climb estimation outlined in the *MERLIN* approach. For Meinshausen-Bühlmann and graphical lasso estimation (both contained within the *huge* package), regularization parameters are used with two parameter selection methods known as Rotation Information Criterion (RIC) and STability Approach to Regularization Selection (STARS) to ensure sparsity (Liu *et al.*, 2010). For MERLIN, a sparsity penalty parameter is encoded into all edge likelihood calculations. The expression data for all experiments comes from an amalgamation of studies across mouse embryonic stem cells. Additionally, the stability selection approach, as outlined by Meinshausen *et al.*, 2010, is used across all experiments to build consensus networks for comparison. Finally, three gold network standards are used as reference for network comparison.

2.1 Data acquisition

The RNA-Seq data used for this study comes from several experiments which profile mouse embryonic stem cells, all contained within the Sequence Read Archive (SRA) from the National Center for Biotechnology Information (NCBI). Each sample has an SRP ID, which corresponds to the study's ID on the SRA. Samples also have an SRX ID that corresponds to the ID of the sample within that study, and finally a GSM ID which also corresponds to the sample within that study. The GSM ID maps to a sample record describing the conditions under which the sample was handled and can be found through the Gene Expression Omnibus (GEO) repository from the NCBI. All expression data has been preprocessed to be log-transformed, zero-meaned, and quantile-normalized. The full list of 1,196 samples that comprises the dataset can be found here:

`http://pages.discovery.wisc.edu/˜asiahpirani/mesc_runs_march16/merlin_tfa_modules/rnaseq.tfa0.010/rnaseq_header.html`

A spreadsheet which describes all datasets and samples along with all SRP, SRX, GSM, and GSE IDs can be found in the supplementary resources of this submission (`mESC-RNASeq-Samples.csv`).

Three gold standard networks were acquired for network evaluation. Two of these were ChIP-based and LOGOF (Loss or Gain of Function) gold standard networks obtained from the Embryonic Stem Cells Atlas of Pluripotency Evidence (ESCAPE) database (Xu *et al.*, 2013, 2014). The LOGOF network is based on gene knockdown expression measurements. An additional network was derived by Alireza Fotuhi Siahpirani from regulatory interactions reported in literature from a multitude of sources (Zhou *et al.*, 2007; Kim *et al.*, 2008; Young, 2011; Buganim *et al.*, 2012; Dunn *et al.*, 2014; Xu *et al.*, 2014; Malleshaiah *et al.*, 2016).

2.2 MERLIN package

The Modular Regulatory Network Learning with per gene INformation (*MERLIN*) package, written in C++, presents a novel approach to gene regulatory network inference using probabilistic graphical models and a greedy hill-climbing structure search (Siahpirani *et al.*, 2016). The actual package used for this study offers a prior-based integrative framework, however that component was ignored to standardize the graph estimation methods compared to what *huge* offers. A sparsity parameter $\rho = -5$ was used as an edge penalty when calculating graph likelihood, with a modularity parameter $r = 4$ and a hierarchical clustering threshold of $h = 0.6$. These parameters were selected based on the recommendations of the original paper. Exact uses of these parameters can be traced through the methods in the paper as reported by Siahpirani *et al.*, 2016. *MERLIN* iteratively searches through all possible regulator and target gene edge pairs and greedily selects the edges that most improve the log-likelihood score. This process is repeated until a convergence threshold is met.

2.3 huge package

The High-Dimensional Undirected Graph Estimation (*huge*) package, written natively in C and externally in R, offers a variety of functions for estimating network paths. For this study, three functions were used for all experiments: nonparanormal transformation of the data, graph estimation, and regularization parameter selection (Zhao *et al.*, 2012). Specific parameters for the graph estimation and regularization selection are outlined below. Additional functionality was necessary to compare outputs, in the form of an undirected adjacency matrix, to the directed network output of MERLIN and the directed gold standard networks.

2.3.1 Graph estimation methods

The two methods of graph estimation used from *huge* are Meinshausen-Bühlmann and *glasso* (Meinshausen *et al.*, 2006; Friedman *et al.*, 2008).

For both methods, the nonparanormal data is input and regularization parameters $\lambda = (\lambda_1, \ldots, \lambda_10)$ are automatically generated as a sequence of decreasing positive numbers which corresponds to an evenly-spaced sequence of graph sparsity levels from 0 to 0.1. The number of regularization parameters and sparsity level threshold are set to the default values as this study could not find sufficient evidence for changing them. For *glasso*, a lossy screening rule is applied to preselect the neighborhood for a large performance boost (Zhao *et al.*, 2012).

2.3.2 Regularization selection

The two methods of regularization parameter selection used are RIC and STARS (Liu *et al.*, 2010). For RIC, all experiments use 20 rotations. For STARS, a subsampling ratio of $10\frac{\sqrt{n}}{n}$ is used, where n is the number of samples, for a total of 20 subsamples, and the variability threshold is set to 0.1. Again, parameters were set to default values as the author could not find a significant reason to modify them. These methods respectively select the optimal graph from the 10 generated adjacency matrices corresponding to λ.

2.3.3 Converting to directed networks

Because *huge* inherently uses undirected graph estimation, further processing was required in order to compare the output networks to those of *MERLIN* as well as the gold standard networks. The list of possible regulators and target genes was given along with the data, and the following convention was adopted for converting the undirected network:

Given a set of regulators R and target genes T, for each edge between genes i and j:

- if $i \in R \wedge j \in R$, create a directed edge in both directions
- if $(i \in R \wedge j \in T) \vee (j \in R \wedge i \in T)$, create a directed edge from regulator to target gene
- if $i \in T \wedge j \in T$, do not add a directed edge in either direction

This convention ensured that any edges that connected to a regulator gene would have an outgoing edge from that gene, and edges connecting to a target gene would have an incoming edge to that gene as long as its source was not another target gene.

2.4 Stability selection for consensus networks

Stability selection was incorporated because the adjacency matrices from *huge* had no measure of confidence on the edges. By using 100 subsamples, each containing half of the full number of samples, 100 networks were generated from 100 respective runs. From these, a consensus network was built by scoring each edge e with a probability $P(e) = \frac{C_e}{N}$ where C_e is the count of networks which contained edge e and N is the total number of networks (in this case 100). This was necessary in order to create precision-recall curves which require confidence measures for each edge.

Additionally, subsampling is well-suited for high-dimensional data to provide sample control of error rates such as the False Discovery Rate (FDR) (Meinshausen *et al.*, 2010). It inherently supports the notion that true edges will appear in multiple subsamples as opposed to one appearing by chance.

3 Results

3.1 Runtime

Experiments were carried out using computing resources acquired through the Center for High Throughput Computing (CHTC) at UW-Madison. *MERLIN* runs took over a week to complete, while run times for *huge* varied by estimation method. For Meinshausen-Bühlmann, a single run typically took 30 minutes with the RIC selection method and 90 minutes with the STARS selection method. For *glasso*, a single run typically took over a day to complete. Due to technical issues with loading a custom R installation to the CHTC servers, consensus networks for *glasso* runs and Meinshausen-Bühlmann estimation with the STARS selection method were unable to created. A single run using the full dataset was run locally for these methods and evaluated, however because edge confidence scores were created using stability selection, precision-recall curves were unable to be drawn for these respective runs.

3.2 Total predictions

Table 1 shows total edge predictions for each of the evaluated methods against the three previously described gold standard networks. As evident by the tables, all of the nonparanormal methods have a very small amount of edge predictions as compared to the true networks. Contrastingly, the *MERLIN* approach appears to predict much more edges than the true networks have. This leads to recall scores that are extremely high when compared to the nonparanormal approach, but precision scores that are lower than the Meinshausen-Bühlmann with RIC (mb-ric) selection method for the most part. The literature-derived gold standard network has significantly smaller amount of edges than any predicted or gold standard network, leading to extremely low precision by the all approaches (and low recall for the glasso method which had a small amount of predictions to begin with).

Table 1. Total edge predictions for each method as compared to three gold standard networks.

Method	Gold	Predicted	Retrieved	Recall	Precision
mb-ric	ChIP	9792	96/178820	0.00054	0.00980
glasso-ric	ChIP	974	9/178820	5.033E-5	0.00924
merlin	ChIP	2873098	25517/178820	0.14270	0.00888
mb-ric	LOGOF	9792	100/146564	0.0.00068	0.01021
glasso-ric	LOGOF	974	13/146564	8.870E-5	0.01335
merlin	LOGOF	2873098	22051/146564	0.15045	0.00767
mb-ric	Literature	9792	1/267	0.00375	0.00010
glasso-ric	Literature	974	0/267	0.0	0.0
merlin	Literature	2873098	182/267	0.68165	6.335E-5

3.3 Precision-recall curves

Figure 2, Figure 3 and Figure 4 depicts precision-recall curves of the mb-ric consensus network versus the *MERLIN* consensus network over the three gold standard networks. All three curves show a very poor performance by the nonparanormal methods as compared to the *MERLIN* approach. Area under the precision-recall (AUPR) and receiver operating characteristic (AUROC) curves are given below in Table 2. These values reinforce the strength that *MERLIN* shows as opposed to the nonparanormal methods. Specifically, it can be observed that the AUROC values for the nonparanormal method are all close to 0.5 (which can be obtained by a random guessing method). *MERLIN* is not much higher than this, but still there is a notable difference.

4 Discussion

4.1 Performance of huge compared to MERLIN

As the results show, *MERLIN* clearly outperforms the nonparanormal methods in almost every category. When inspecting the total predicted edges, *MERLIN* predicts more edges (on the order of ten times or greater)

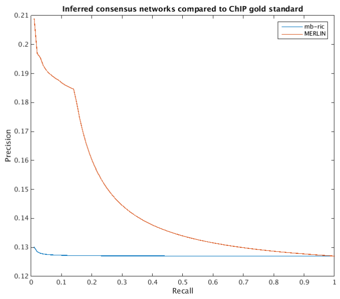

Fig. 2. Meinshausen-Bühlmann and MERLIN compared to ChIP gold standard.

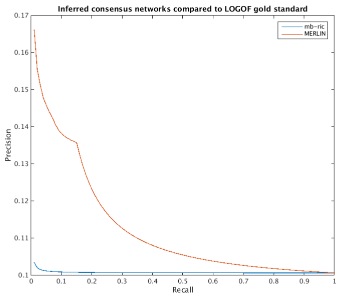

Fig. 3. Meinshausen-Bühlmann and MERLIN compared to LOGOF gold standard.

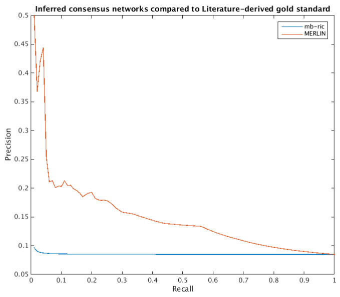

Fig. 4. Meinshausen-Bühlmann and MERLIN compared to literature-derived gold standard.

Table 2. AUPR and AUROC values for each predicted network compared to three gold standard networks.

Method	Gold	AUPR	AUROC
mb-ric	ChIP	0.12730	0.50013
merlin	ChIP	0.14477	0.52561
mb-ric	LOGOF	0.10092	0.50015
merlin	LOGOF	0.11306	0.52188
mb-ric	Literature	0.08591	0.50066
merlin	Literature	0.15193	0.62838

than each network, while the nonparanormal methods tend to underpredict. One possible reason for this is that the graph estimation methods used with the nonparanormal methods use parameters that enforce sparsity more heavily than *MERLIN*. Specifically, the RIC regularization parameter selection method may be choosing the regularization parameter λ which results in the sparsest possible network. For example, drilling down on the *glasso* method shows an extremely sparse derived network which results in 0 precision and recall when compared to the literature-derived network. Conversely, the sparsity penalty parameter of *MERLIN* may not enforce sparsity as strongly thus resulting in more predicted edges. Another reason that the nonparanormal methods might be underpredicting is that only a subset of the genes were used (5,573 of 20,551 genes) in order to make computation feasible. While *MERLIN* performs a greedy hill-climb of all possible edges, the graph estimation methods from *huge* are unable to retrieve many of these high-scoring edges in a high-dimensional setting.

Taking an in-depth look at *MERLIN*'s recall scores show that they heavily outweigh those of the nonparanormal methods - it is able to retrieve many of the true edges for each gold standard network. However, its high number of edge predictions also leads to lower precision scores than the mb-ric method. This is a natural outcome, as over-predicting edges tends to drive recall up and precision down. Whether this performance is desired is context dependent.

Another consideration for the poor performance of methods from *huge* is that parameter selection was not varied. Due to time constraints and lack of intuition about the graph estimation methods, default parameters were used across all settings. These parameters need to be fine tuned in order to achieve better performance, and it may be necessary to explore a variety of estimation methods, parameter settings and regularization selection methods to get a better grasp on the predictive power of nonparanormal methods. This lies outside the scope of this report but is a necessity for conclusive results.

When looking at nonparanormal methods as a whole, another consideration might be to try using the count-based data instead of gene expression data. Since the nonparanormal transformation can be applied to any type of data set, this could possibly lead to a better representation of the data as input to methods in *huge* as compared to transforming the gene expression data which has been calculated from relative abundances.

4.2 Future work

Future work in this comparison of nonparanormal methods to standard methods requires that a variety of parameter selections are experimented with. It may also prove to be useful to test across multiple datasets. This set contained an amalgamation of studies which may introduce a batch effect (normalization measures were used to combat this but it can still remain).

Additionally, it may be useful to establish one graph estimation method (as a control) and only vary the transformation of the data - this work compared multiple different graph estimation methods which may have had a stronger hand in the performance difference than the nonparanormal transformation. For example, only using Meinshausen-Bühlmann graph estimation methods and comparing between untouched data and data that has been transformed using the nonparanormal method.

Finally, using count-based data as input to the nonparanormal transformation could provide more conclusive results as it is inherently non-Gaussian. This study used the same data set across all experiments to see if the nonparanormal transformation could still outperform the greedy approach, however it may be the case that it outperforms only in certain environments.

Acknowledgements

Acknowledgements to Alireza Fotuhi Siahpirani for providing the RNA-Seq data set, the SRA ID mappings, and the derived gold standard network from the literature.

References

Buganim,Y. *et al*. (2012) Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, **150**, 1209-1222.

Dunn,S.J. *et al*. (2014) Defining an essential transcription factor program for naïve pluripotency. *Science*, **344**, 1156-1160.

Friedman,J., Hastie,T., Tibshirani,R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Gierliński,M. *et al*. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics*, **31**, 3625-3630.

Kim,J. *et al*. (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049-1061.

Li,P. *et al*. (2015) Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, **16**, 347.

Liu,H., Lafferty,J., Wasserman,L. (2009) The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, **10**, 2295-2328.

Liu,H., Roeder,K., Wasserman,L. (2010) Stability approach to regularization selection (StARS) for high dimensional graphical models. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, **2**, 1432-1440.

Malleshaiah,M. *et al*. (2016) Nac1 coordinates a sub-network of pluripotency factors to regulate embryonic stem cell differentiation. *Cell Reports*, **14**, 1181-1194.

Meinshausen,N., Bühlmann,P. (2006) High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, **34**, 1436-1462.

Meinshausen,N., Bühlmann,P. (2010) Stability selection. *Journal of the Royal Statistical Society*, **72**, 417-473.

Siahpirani,A.F., Roy,S. (2016) A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Research*, **45**, e21.

Xu,H. *et al*. (2013) ESCAPE: database for integrating high-content published data collected from human and mouse embryonic stem cells. *Database (Oxford)*, **2013**, bat045.

Xu,H. *et al*. (2014) Construction and validation of a regulatory network for pluripotency and self-renewal of mouse embryonic stem cells. *PLoS Computational Biology.*, **10(8)**, e1003777.

Young,R.A. (2011) Control of the embryonic stem cell state. *Cell*, **144**, 940-954.

Zhao, T. *et al*. (2012) The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, **13**, 1059-1062.

Zhou,Q. *et al*. (2007) A gene regulatory network in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, **104(42)**, 16438-16443.