BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–5

Gene network inference by fusing data from diverse distributions Supplementary information

Marinka Žitnik[1] and Blaž Zupan[1,2]

[1]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
[2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

S1 LEARNING THE MODELS IN PRACTICE

Now that we defined the FUSENET model, we explain how to solve related optimization problems. Notice that exact optimization problem one needs to solve depends on a particular data setting, *i.e.*, the particular combination of exponential family distributions that generated a collection of data sets.

There has been a strong line of work on developing fast algorithms to solve sparse regression problems that are similar to Eq. (8) and Eq. (11) including the work by Krishnapuram *et al.* (2005), Meier *et al.* (2008), Jalali *et al.* (2011) and Allen and Liu (2013). Existing algorithms for undirected graphical model selection assume that model parameters are independent of each other. This, however, is not true in FUSENET due to reasons discussed in Sec. 3.7, which ensure data fusion. Consequently, this also means that we cannot use off-the-shelf optimization solvers.

S1.1 Node neighborhood selection

We propose to fit our FUSENET by computing cyclical coordinate descent along the path of regularization parameter λ. Taking derivatives of Eq. (13) and with optimization techniques by Friedman *et al.* (2007a); Yuan (2008); Friedman *et al.* (2010) we can obtain solutions over a range of values for regularization parameter with approximately the same speed as fitting a model at a single value of λ. The technique uses current parameter estimates as warm restarts.

FUSENET employs elastic net penalties (Zou and Hastie, 2005) in their models. Elastic net is a compromise between the ridge penalty ($\lambda = 0$) and the lasso penalty ($\lambda = 1$) and is useful in situations where $p \gg n$ or when many variables are correlated. As λ increases from 0 to 1, for a given α the sparsity of the solution (*i.e.* the number of latent components equal to zero) increases monotonically from 0 to the sparsity of the lasso solution. In each iteration of the coordinate descent we apply soft thresholding to the current FUSENET estimates to care of the lasso contribution to the penalty, and then apply a proportional shrinkage for the ridge penalty (Meinshausen and Bühlmann, 2006; Friedman *et al.*, 2007a; Simon *et al.*, 2013).

S1.2 Selecting regularization parameters ($\boldsymbol{\lambda}$)

The choice of λ is critical since different λ's can lead to different network sparsity patterns, *i.e.* the number and position of edges in the inferred network. We estimate λ in data-dependent way via stability selection (Meinshausen and Bühlmann, 2010), a technique which was shown to lead to better results for the network inference than other parameter selection methods including cross validation, Akaike's information criterion and Bayesian information criterion (Liu *et al.*, 2010; Yu *et al.*, 2012).

For now, we assume that the number of latent components r is given. Here, we choose λ so as to use the least amount of regularization that simultaneously makes the network sparse and stable, *i.e.*, replicable under random sampling. FUSENET employs recently proposed stability selection technique called StARS (Liu *et al.*, 2010). Briefly, StARS repeatedly sub-samples data $\mathcal{D}$ to obtain many data samples $\mathcal{D}_s$. Here, $\mathcal{D}_s$ denotes s-th data sample. It then estimates a separate network $\widehat{E}_s(\lambda, r)$ for each $\mathcal{D}_s$ and each λ from a vector of regularization parameters $\boldsymbol{\lambda}$; the latter being possible due to coordinate descent computed along a regularization path. Selected value for regularization controls the average variance over the edges of the networks inferred from sub-sampled data:

$$\lambda_{\text{opt}}^{(r)} = \arg\min_{\rho}\{\min_{0\leq\lambda\leq\rho}(\sum_{j<k} 2\bar{\mathbf{A}}_{jk}(\lambda, r)(1-\bar{\mathbf{A}}_{jk}(\lambda, r))/\binom{p}{2}) \leq \beta\}$$

where $\bar{\mathbf{A}}_{jk}(\lambda, r) = \frac{1}{S}\sum_{s=1}^{S}\mathcal{I}((j,k) \in \widehat{E}_s(\lambda, r))$. We set β and the size of data samples $\mathcal{D}_s$ to the values recommended in Allen and Liu (2013). We note that we obtain different optimal values of $\lambda_{\text{opt}}^{(r)}$ for different choices of r. Next, we describe how we select r, which in effect determines the exact value of regularization.

S1.3 Selecting the number of latent components (r)

Our FUSENET has another parameter, the number of latent components r, which otherwise does not appear in current Markov models. The latent dimensionality is selected from a set of predefined candidate values $\{0.05n, 0.1n \ldots, 0.5n\}$, where n is the mean number of observations across all considered data sets. We seek to use the fewest number of latent components that produce stable and sparse network:

$$r_{\text{opt}} = \arg\min_{\tau} \lambda_{\text{opt}}^{(\tau)}.$$

As a consequence, the optimal regularization value is $\lambda_{\text{opt}} = \lambda_{\text{opt}}^{(r_{\text{opt}})}$. Notice that the entire set of computations including path-wise coordinate descent and selection of regularization via stability selection can be performed in parallel for each candidate value of r.

S2 MULTIVARIATE DATA SIMULATION

Four network structures are simulated: (1) the Erdős Rényi random network, where an edge between each pair of nodes is set with equal probability and independently of other edges; (2) a hub network, where each node is connected to one of three hub nodes; (3) a scale-free network, in which node degree distribution follows a power-law; and (4) a small-world network, in which most nodes are not neighbors of each other but most nodes can be reached from every other by a small number of hops.

In simulations involving the Poisson model we closely follow the approach described by Karlis (2003) and Allen and Liu (2013). We generate n independent observations with p nodes, $\mathcal{D} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$, where $\mathbf{x}^{(i)}$ is a p-dimensional count data vector, $\mathbf{x}^{(i)} \in \{0, 1, \dots, \infty\}^p$. A matrix of observations $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}]^T$ is obtained from the model $\mathbf{X} = \mathbf{Y}\mathbf{B} + \mathbf{E}$. Here, $\mathbf{Y}$ is a $n \times (p + p(p-1)/2)$ matrix with each entry $\mathbf{Y}_{ij} \overset{iid}{\sim}$ Poisson(λ_{true}) and $\mathbf{E}$ is a $n \times p$ matrix with $\mathbf{E} \overset{iid}{\sim}$Poisson($\lambda_{\text{noise}}$). Let $\mathbf{A}^*$ denote the adjacency matrix of a given true network structure E^*. The adjacency matrix is encoded by matrix $\mathbf{B}$ as $\mathbf{B} = [\mathbf{I}_p; \mathbf{P} \odot (\mathbf{1}_p \text{tri}(\mathbf{A}^*)^T)]^T$. Here, $\mathbf{P}$ is a $p \times (p(p-1)/2)$ permutation matrix, $\odot$ represents the entry-wise product and $\text{tri}(\mathbf{A}^*)$ is the $(p(p-1)/2) \times 1$ vectorized upper triangular part of $\mathbf{A}^*$. As done by Allen and Liu (2013) we simulate data at two signal-to-noise ratio (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.

In simulations involving the multinomial model we fix the alphabet size to $m = 3$. For a given true network structure E^*, we pick the parameter set $\theta_{st;jk} \in \{\theta_{st;jk} : s, t \in V; (s,t) \in E^*; j, k \in \{1, 2\}\}$ as follows. If $(s,t) \in E^*$ then each nonzero entry $\theta_{st;jk}$ for $j, k \in \{1, 2\}$ is set to $\theta_{st;jk} \in [-0.5, 0.5]$ uniformly at random; there are $4 = (3-1)^2$ such entries. We then generate n observations to construct a data set according to the probability distribution corresponding to $\theta_{st;jk}$. We solve the problem in Eq. (12) and compare the inferred network $\widehat{E}$ with the true network E^*.

S3 CANCER GENOMIC DATA

We apply network inference algorithms to two examples of non-Gaussian high-throughput genomic data to learn (1) an mRNA expression network, (2) a somatic mutation network and (3) a collectively inferred gene network from both data types.

We download breast cancer (BRCA-US) gene expression data measured by next generation sequencing and breast cancer (BRCA-US) simple somatic mutation data from the International Cancer Genome Consortium (ICGC) (Hudson *et al.*, 2010) portal (release 17). We follow the steps in Allen and Liu (2013) and process the data to be approximately Poisson as is shown in Suppl. Fig. 1. Genes with little variation across samples, the bottom 50%, are filtered out, and the data is adjusted for possible overdispersion by transforming them via a power $\alpha \in (0, 1]$ where α is chosen to yield approximately Poisson data as assessed via Kolmogorov-Smirnov tests (Li *et al.*, 2011). The power transformation has another advantage. When neighboring genes have extremely large counts, the exponential in Eq. (6) causes the conditional Poisson mean to become large. The transformation limits the extreme counts and subsequently improves the fit of the model. Data preprocessing results in a matrix with rows as the subjects ($n_{\text{exp}} = 1{,}012$) and columns as genes ($p_{\text{exp}} = 657$). These genes form the nodes of our Poisson breast cancer mRNA network.

Breast cancer simple somatic mutation data from the ICGC portal include single base substitutions, multiple base substitutions and short indels. Mutation data are converted into a matrix with rows as subjects ($n_{\text{mut}} = 954$) and columns as genes containing mutations or variations (25,834 genes). Each matrix entry is categorized into one of three groups based on the type of mutation: no mutation, single base substitution, insertion/deletion of < 200 base pairs. Differentially mutated genes, *i.e.* genes containing mutations relative to the corresponding normal sample data, are ordered by their percentage of mutations across all samples and the top $p = 500$ genes were used in our analysis. These genes form the nodes of our multinomial breast cancer somatic mutation network.

For the collectively inferred network, we consider both gene expression profiles and somatic mutation data provided by the ICGC assuming the Poisson model for the RNA-seq data and the multinomial model for the mutation data. The genes that form the nodes of this network are taken as the union of sets of genes from the respective gene expression and somatic mutation matrices ($p = |V_{\text{exp}} \cup V_{\text{mut}}|$). Mutational and expression profiles from both matrices are matched by the subjects.

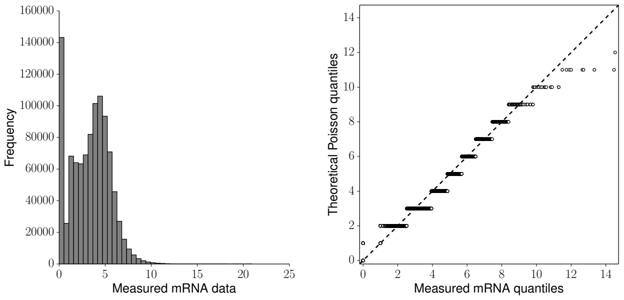

Fig. S1. A histogram of the overall breast cancer RNA-seq data from the ICGC (Hudson *et al.*, 2010) (left) and a comparison of these data to the quantiles of the Poisson distribution via a q-q plot (right). A q-q plot shows that breast cancer RNA-seq data approximately follow the Poisson distribution. The multivariate count data arising from the measurements of gene expression with the next generation sequencing technology is only an example of recent high-throughput technologies that produce non-Gaussian distributed data.

S4 QUANTIFYING THE FUNCTIONAL CONTENT OF INFERRED NETWORKS

We employ two approaches to evaluate “functional correctness” of the networks inferred from cancer data.

First, we use SANTA (Cornish and Markowetz, 2014) to quantify the strength of association between sets of functionally related genes and the inferred network. The input to SANTA are a gene network and a gene set and the output is a score representing statistical significance of their association. We obtain gene sets from the Gene Ontology (GO) (Ashburner *et al.*, 2000) and test only GO terms associated with between 20 and 100 network genes to ensure that the functional sets are not too thinly or thickly spread.

Second, we overlay the inferred network with gene information from the GO and for every GO term assess how community-like a subnetwork of genes that belong to a particular GO term is. Four different structural notions of network communities exist in networks and we report the values of their representative scoring functions (Yang and Leskovec, 2012). Given is the inferred network

$G(V, \widehat{E})$, where $p = |V|$. Let $T \subseteq V$ be genes that belong to a specific GO term and let p_T be their number, $p_T = |T|$. We also need m_T, which is the number of edges in G whose both endpoints are annotated with a given GO term, $m_T = |\{(s,t) \in \widehat{E} : s \in T, t \in T\}|$, and c_T, which counts how many edges are on the boundary of set T, $c_T = |\{(s,t) \in \widehat{E} : s \in T, t \notin T\}|$. We denote degree of gene s with $d(s)$. Scoring functions build on the intuition that communities are sets of genes with many connections between the members and few connections to the rest of the network. We consider the following four scoring functions:

- **triangle participation ratio (TPR)** is the fraction of genes in T that belong to a triad, $|\{s : s \in T, \{(t,u) : t,u \in T, (s,t) \in \widehat{E}, (s,u) \in \widehat{E}, (t,u) \in \widehat{E}\} \neq \emptyset\}|/p_T$;
- **cut ratio** is the fraction of all possible edges in T that connect T to the remainder of the network, $\frac{c_T}{p_T(p-p_T)}$;
- **conductance** is the fraction of total edge volume that points outside the GO term T, $\frac{c_T}{2m_T+c_T}$;
- **flake-over-median-degree (flake-ODF)** is the fraction of genes in T with fewer edges linking inside than outside of T, $|\{s : s \in T, |\{(s,t) \in \widehat{E} : t \in T\}| < d(s)/2\}|/p_T$.

The functions take values from $[0, 1]$ interval. To make the higher the better, we report $(1 -$ Conductance$)$, $(1 -$ Cut ratio$)$ and $(1 -$ flake-ODF$)$ for conductance, cut ratio and flake-ODF, respectively.

S5 NETWORK RECOVERY WITH SIMULATED DATA

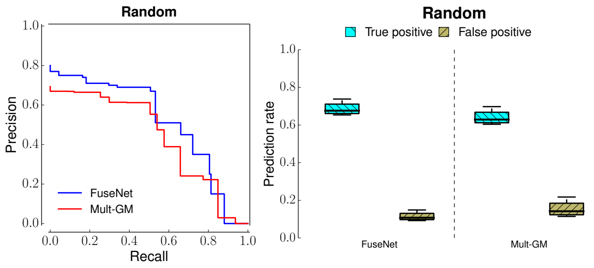

Fig. S2. Application of gene network inference algorithms to multinomial-distributed simulated data. Simulation studies on four network types were performed. Shown are results for Erdős Rényi random network, see main text for other network types. We generated $n = 300$ observations at a high signal-to-noise ratio (SNR) with $p = 50$ variables (nodes) taking values from an alphabet of size $m = 3$. Receiver operating curves and boxplots are shown for the multinomial FuseNet (proposed here) and the multinomial graphical model (Mult-GM) (Jalali *et al.*, 2011).

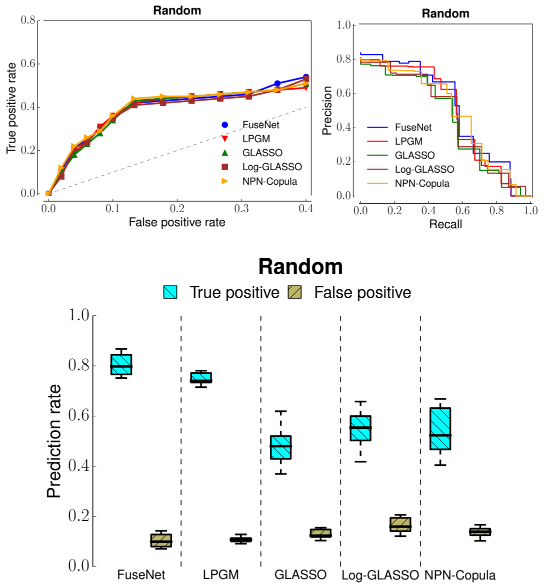

Fig. S4. Application of gene network inference algorithms to Poisson-distributed simulated data. Simulation studies on four network types were performed. Shown are results for Erdős Rényi random network, see main text for other network types. We generated data with $n = 200$ observations with $p = 100$ variables (nodes) at a low (first row; left) and high (first row; right and second row) signal-to-noise ratio (SNR). Receiver operating curves, precision-recall curves and boxplots are shown for the Poisson FuseNet (proposed here), the Local Poisson Graphical Model (LPGM) (Allen and Liu, 2013), the Graphical Lasso (GLASSO) (Friedman *et al.*, 2007b), the GLASSO on log-transformed data (Log-GLASSO) (*e.g.* cf. Gallopin *et al.*, 2013) and the GLASSO on data transformed through nonparanormal Gaussian copula (NPN-Copula) (Liu *et al.*, 2009)

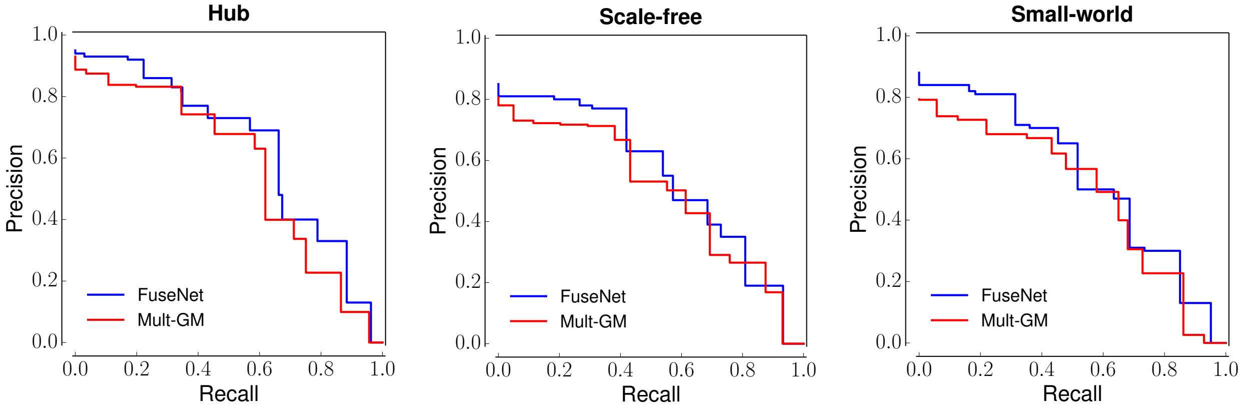

Fig. S3. Application of gene network inference algorithms to multinomial-distributed simulated data. Simulation studies on four network types were performed: random (See Suppl. Fig. 2), hub, scale-free and small-world. These graph structures appear in many real biological networks. For each graph type, we generated data with $n = 300$ observations at a high signal-to-noise ratio (SNR) with $p = 50$ variables (nodes) taking values from an alphabet of size $m = 3$. Precision-recall curves are shown for the multinomial FUSENET (proposed here) and the multinomial graphical model (Mult-GM) (Jalali *et al.*, 2011).

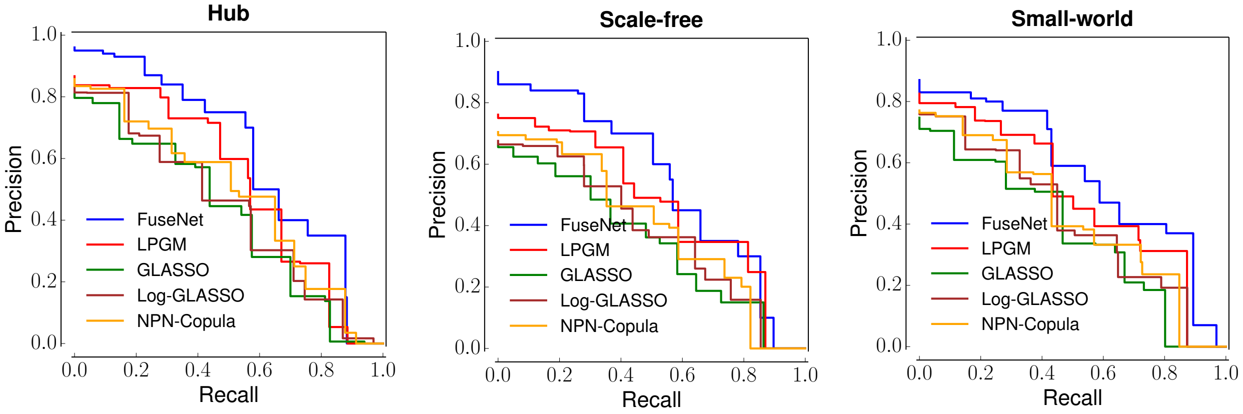

Fig. S5. Application of gene network inference algorithms to Poisson-distributed simulated data. Simulation studies on four network types were performed: random (see Suppl. Fig. 3), hub, scale-free and small-world. These graph structures appear in many real biological networks. For each graph type, we generated data with $n = 200$ observations with $p = 100$ variables (nodes) at a high signal-to-noise ratio (SNR). Precision-recall curves are shown for the Poisson FUSENET (proposed here), the Local Poisson Graphical Model (LPGM) (Allen and Liu, 2013), the Graphical Lasso (GLASSO) (Friedman *et al.*, 2007b), the GLASSO on log-transformed data (Log-GLASSO) (*e.g.* cf. Gallopin *et al.*, 2013) and the GLASSO on data transformed through nonparanormal Gaussian copula (NPN-Copula) (Liu *et al.*, 2009)

S6 FUNCTIONAL CONTENT OF THE INFERRED NETWORKS

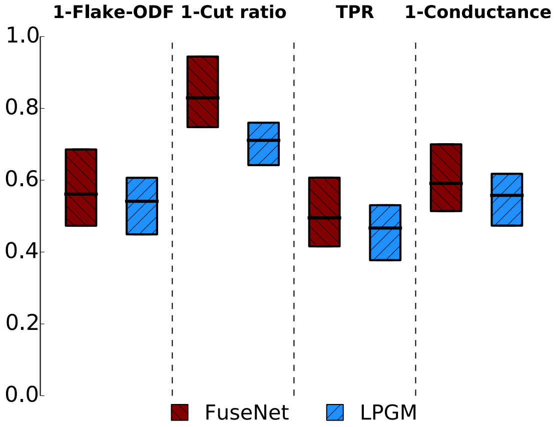

Fig. S6. The strength of association between gene sets from the Gene Ontology (GO) and networks inferred with Poisson FUSENET (proposed here) and LPGM (Allen and Liu, 2013). Inferred networks were overlaid with GO terms and subnetworks induced by each GO term were assessed for how well they corresponded to network communities. Four different scoring functions were used to quantify the presence of different structural notions of communities (Yang and Leskovec, 2012) that can appear in biological networks: flake-over-median-degree (flake-ODF), cut ratio, triangle participation ratio (TPR) and conductance. Results are shown for breast cancer RNA-sequencing data because LPGM method was designed for Poisson distributed data.

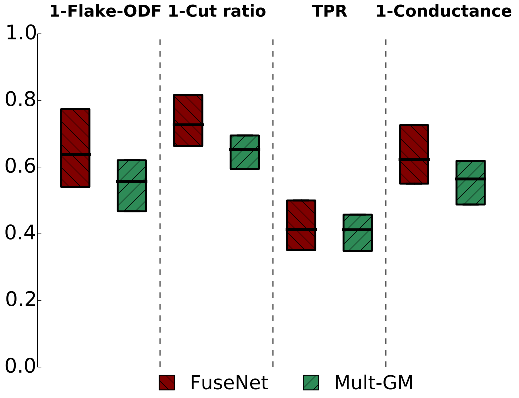

Fig. S7. The strength of association between gene sets from the Gene Ontology (GO) and networks inferred with multinomial FUSENET (proposed here) and multinomial graphical model (Mult-GM) (Jalali *et al.*, 2011). Inferred networks were overlaid with GO terms and subnetworks induced by each GO term were assessed for how well they corresponded to network communities. Four different scoring functions were used to quantify the presence of different structural notions of communities (Yang and Leskovec, 2012) that can appear in biological networks: flake-over-median-degree (flake-ODF), cut ratio, triangle participation ratio (TPR) and conductance. Results are shown for breast cancer somatic mutation data because Mult-GM method was designed for multinomial distributed data.

REFERENCES

Allen, G. I. and Liu, Z. (2013). A local poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on NanoBioscience*, **12**(3), 189–198.

Ashburner, M. *et al.* (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.

Cornish, A. J. and Markowetz, F. (2014). SANTA: quantifying the functional content of molecular networks. *PLoS Computational Biology*, **10**(9), e1003808.

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., *et al.* (2007a). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**(2), 302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2007b). Sparse inverse covariance estimation with the lasso. *Biostatistics*, **9**, 432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1.

Gallopin, M., Rau, A., and Jaffrézic, F. (2013). A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS One*, **8**(10), e77503.

Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M., Calvo, F., Eerola, I., Gerhard, D. S., *et al.* (2010). International network of cancer genome projects. *Nature*, **464**(7291), 993–998.

Jalali, A., Ravikumar, P. D., Vasuki, V., and Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. In *AISTATS*, pages 378–387.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, **30**(1), 63–77.

Krishnapuram, B. *et al.* (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **27**(6), 957–968.

Li, J. *et al.* (2011). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, pages 1–16.

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

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In *NIPS*, pages 1432–1440.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**(2), 231–245.

Yang, J. and Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In *ACM MDS*.

Yu, H., Dauwels, J., and Wang, X. (2012). Copula Gaussian graphical models with hidden variables. In *IEEE ICASSP*, pages 2177–2180.

Yuan, M. (2008). Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, **17**(4), 809–826.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.