On Identifying Significant Edges in Graphical Models

Marco Scutari[1] and Radhakrishnan Nagarajan[2]

[1] Genetics Institute, University College London, London, United Kingdom.
[2] Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA.

Abstract. Graphical models, and in particular Bayesian networks, have been widely used to investigate data in the biological and healthcare domains. This can be attributed to the recent explosion of high-throughput data across these domains and the importance of understanding the causal relationships between the variables of interest. However, classic model validation techniques for identifying significant edges rely on the choice of an ad-hoc threshold, which is non-trivial and can have a pronounced impact on the conclusions of the analysis.
In this paper, we overcome this limitation by proposing simple, statistically-motivated approach based on L_1 approximation for identifying significant edges. The effectiveness of the proposed approach is demonstrated on gene expression data sets across two published experimental studies.

Keywords: graphical models, model averaging, L_1 approximation.

1 Introduction and Background

Graphical models [18, 28] are a class of statistical models which combine the rigour of a probabilistic approach with the intuitive representation of relationships given by graphs. They are composed by a set $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ of *random variables* describing the quantities of interest and a *graph* $\mathcal{G} = (\mathbf{V}, E)$ in which each *vertex* $v \in \mathbf{V}$ is associated with one of the random variables in $\mathbf{X}$. The *edges* $e \in E$ are used to express the dependence relationships among the variables in $\mathbf{X}$. The set of these relationships is often referred to as the *dependence structure* of the graph. Different classes of graphs express these relationships with different semantics, which have in common the principle that graphical separation of two vertices implies the conditional independence of the corresponding random variables [28]. The two examples most commonly found in literature are *Markov networks* [8, 35], which use undirected graphs, and *Bayesian networks* [20, 26], which use directed acyclic graphs.

In principle, there are many possible choices for the joint distribution of $\mathbf{X}$, depending on the nature of the data and the aims of the analysis. However, literature have focused mostly on two cases: the *discrete case* [14, 35], in which both $\mathbf{X}$ and the X_i are multinomial random variables, and the *continuous case* [13, 35], in which $\mathbf{X}$ is multivariate normal and the X_i are univariate normal random variables. In the former, the parameters of interest are the *conditional probabilities* associated with each variable, usually represented as conditional probability tables; in the latter, the parameters of interest are the *partial correlation coefficients* between each variable and its neighbours in $\mathcal{G}$.

The estimation of the structure of the graph $\mathcal{G}$ is called *structure learning* [8, 18], and consists in finding the graph structure that encodes the conditional independencies present in the data. Ideally it should coincide with the dependence structure of $\mathbf{X}$, or it should at least identify a distribution as close as possible to the correct one in the probability space. Several algorithms have been presented in literature for this problem, thanks to the application of many results from probability, information and optimisation theory. Despite differences in theoretical backgrounds and terminology, they can all be traced to only three approaches: *constraint-based* (which are based on conditional independence tests), *score-based* (which are based on goodness-of-fit scores) and *hybrid* (which combine the previous two approaches). For some examples see Bromberg et al. [1], Castelo and Roverato [2], Friedman et al. [12], Larrañaga et al. [21] and Tsamardinos et al. [34].

On the other hand, model validation techniques have not been developed at a similar pace. For example, the characteristics of structure learning algorithms are still studied using a small number of reference data sets [10, 24] as benchmarks, and differences from the true (known) structure are measured with purely descriptive measures such as Hamming distance [17]. This approach is clearly not possible when validating networks learned from real world data sets (because the true structure of their probability distribution is not known) and presents some limits even for synthetic data.

A more systematic approach to model validation, and in particular to the problem of identifying statistically significant features in a network, has been developed by Friedman et al. [11] using bootstrap resampling [9] and model averaging [5]. It can be summarised as follows:

1. For $b = 1, 2, \ldots, m$:
   (a) sample a new data set $\mathbf{X}_b^*$ from the original data $\mathbf{X}$ using either parametric or nonparametric bootstrap;
   (b) learn a the structure of the graphical model $G_b = (\mathbf{V}, E_b)$ from $\mathbf{X}_b^*$.
2. Estimate the probability that each possible edge e_i, $i = 1, \ldots, k$ is present in the true network structure $\mathcal{G}_0 = (\mathbf{V}, E_0)$ as

$$\hat{\mathrm{P}}(e_i) = \frac{1}{m} \sum_{b=1}^{m} \mathbb{1}_{\{e_i \in E_b\}}, \qquad (1)$$

   where $\mathbb{1}_{\{e_i \in E_b\}}$ is the indicator function of the event $\{e_i \in E_b\}$ (i.e., it is equal to 1 if $e_i \in E_b$ and 0 otherwise).

The empirical probabilities $\hat{\mathrm{P}}(e_i)$ are known as *edge intensities* or *arc strengths*, and can be interpreted as the degree of *confidence* that e_i is present in the network structure $\mathcal{G}_0$ describing the true dependence structure of $\mathbf{X}$ [3]. However, they are difficult to evaluate, because the probability distribution of the networks $\mathcal{G}_b$ in the space of the network structures is unknown. As a result, the value of the confidence threshold (i.e. the minimum degree of confidence for an edge to be significant and therefore accepted as an edge of $\mathcal{G}_0$) is an unknown function of both the data and the structure learning algorithm. This has proved to be a serious limitation in the identification of significant edges and has led to the use of ad-hoc, pre-defined thresholds in spite of the impact on model validation evidenced by several studies [11, 15]. An exception is Nagarajan et al. [25], whose approach will be discussed below.

Apart from this limitation, Friedman's approach is very general and can be used in a wide range of settings. First of all, it can be applied to any kind of graphical model with only minor adjustments (for example, accounting for the direction of the edges in Bayesian networks). Furthermore, it does not require any distributional assumption on the data in addition to the ones needed to by the structure learning algorithm. No assumption is made on the latter, either, so any score-based, constraint-based or hybrid algorithm can be used.

In this paper, we propose a statistically-motivated estimator for the confidence threshold minimising the L_1 norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the confidence levels of the unknown network $\mathcal{G}_0$. Subsequently, we demonstrate the effectiveness of the proposed approach by re-investigating two experimental data sets from Nagarajan et al. [25] and Sachs et al. [30].

2 Selecting Significant Edges

Consider the empirical probabilities $\hat{\mathrm{P}}(e_i)$ defined in Eq. 1, and denote them with $\hat{\mathbf{p}} = \{\hat{p}_i, i = 1, \ldots, k\}$. For a graph of size n, $k = n(n-1)/2$. Furthermore, consider the order statistic

$$\hat{\mathbf{p}}_{(\cdot)} = \{0 \leqslant \hat{p}_{(1)} \leqslant \hat{p}_{(2)} \leqslant \ldots \leqslant \hat{p}_{(k)} \leqslant 1\} \tag{2}$$

derived from $\hat{\mathbf{p}}$. It is intuitively clear that the first elements of $\hat{\mathbf{p}}_{(\cdot)}$ are more likely to be associated with non-significant edges, and that the last elements of $\hat{\mathbf{p}}_{(\cdot)}$ are more likely to be associated with significant edges. The ideal configuration $\tilde{\mathbf{p}}_{(\cdot)}$ of $\hat{\mathbf{p}}_{(\cdot)}$ would be

$$\tilde{p}_{(i)} = \begin{cases} 1 & \text{if } e_{(i)} \in E_0 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

that is the set of probabilities that characterises any edge as either significant or non-significant without any uncertainty. In other words,

$$\tilde{\mathbf{p}}_{(\cdot)} = \{0, \ldots, 0, 1, \ldots, 1\}. \tag{4}$$

[3] The probabilities $\hat{\mathrm{P}}(e_i)$ are in fact an estimator of the expected value of the $\{0, 1\}$ random vector describing the presence of each possible edge in $\mathcal{G}_0$. As such, they do not sum to one and are dependent on one another in a nontrivial way.

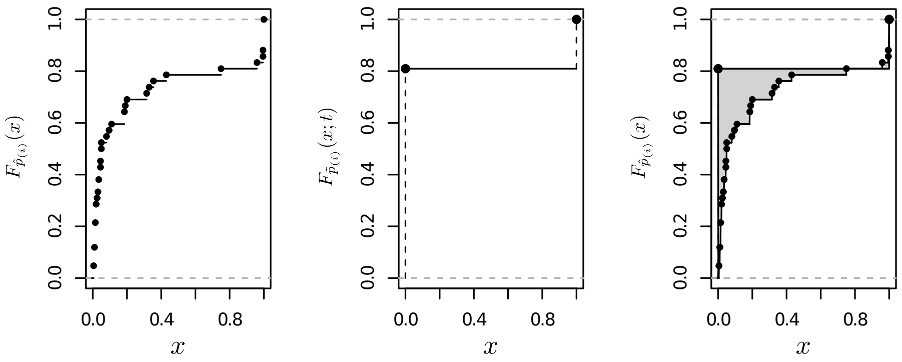

Fig. 1. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$ (left), the cumulative distribution function $F_{\tilde{\mathbf{p}}_{(\cdot)}}$ (centre) and the L_1 norm between the two (right).

Such a configuration arises from the limit case in which all the networks $\mathcal{G}_b$ have exactly the same structure. This may happen in practise with a consistent structure learning algorithm when the sample size is large [4, 22].

A useful characterisation of $\hat{\mathbf{p}}_{(\cdot)}$ and $\tilde{\mathbf{p}}_{(\cdot)}$ can be obtained through the empirical cumulative distribution functions of the respective elements,

$$F_{\hat{\mathbf{p}}_{(\cdot)}}(x) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{\{\hat{p}_{(i)} < x\}} \tag{5}$$

and

$$F_{\tilde{\mathbf{p}}_{(\cdot)}}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0) \\ t & \text{if } x \in [0, 1) \\ 1 & \text{if } x \in [1, +\infty) \end{cases} . \tag{6}$$

In particular, t corresponds to the fraction of elements of $\tilde{\mathbf{p}}_{(\cdot)}$ equal to zero and is a measure of the fraction of non-significant edges. At the same time, t provides a threshold for separating the elements of $\tilde{\mathbf{p}}_{(\cdot)}$, namely

$$e_{(i)} \in E_0 \Longleftrightarrow \hat{p}_{(i)} > F^{-1}_{\tilde{\mathbf{p}}_{(\cdot)}}(t). \tag{7}$$

More importantly, estimating t from data provides a statistically motivated threshold for separating significant edges from non-significant ones. In practise, this amounts to approximating the ideal, asymptotic empirical cumulative distribution function $F_{\tilde{\mathbf{p}}_{(\cdot)}}$ with its finite sample estimate $F_{\hat{\mathbf{p}}_{(\cdot)}}$. Such an approximation can be computed in many different ways, depending on the norm used to measure the distance between $F_{\hat{\mathbf{p}}_{(\cdot)}}$ and $F_{\tilde{\mathbf{p}}_{(\cdot)}}$ as a function of t. Common choices are the L_p family of norms [19], which includes the Euclidean norm, and Csiszar's f-divergences [7], which include Kullback-Leibler divergence.

The L_1 norm

$$L_1\left(t; \hat{\mathbf{p}}_{(\cdot)}\right) = \int \left|F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\tilde{\mathbf{p}}_{(\cdot)}}(x;t)\right| dx \tag{8}$$

appears to be particularly suited to this problem; an example is shown in Fig. 1. First of all, note that $F_{\hat{\mathbf{p}}_{(\cdot)}}$ is piecewise constant, changing value only at the points $\hat{p}_{(i)}$; this descends from the definition of empirical cumulative distribution function. Therefore, for the problem at hand Eq. 8 simplifies to

$$L_1\left(t; \hat{\mathbf{p}}_{(\cdot)}\right) = \sum_{x_i \in \left\{\{0\} \cup \hat{\mathbf{p}}_{(\cdot)} \cup \{1\}\right\}} \left|F_{\hat{\mathbf{p}}_{(\cdot)}}(x_i) - t\right| (x_{i+1} - x_i), \tag{9}$$

which can be computed in linear time from $\hat{\mathbf{p}}_{(\cdot)}$. Its minimisation is also straightforward using linear programming [27]. Furthermore, compared to the more common L_2 norm

$$L_2\left(t; \hat{\mathbf{p}}_{(\cdot)}\right) = \int \left[F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\tilde{\mathbf{p}}_{(\cdot)}}(x;t)\right]^2 dx \tag{10}$$

or the L_∞ norm

$$L_\infty\left(t; \hat{\mathbf{p}}_{(\cdot)}\right) = \max_{x \in [0,1]} \left\{\left|F_{\hat{\mathbf{p}}_{(\cdot)}}(x) - F_{\tilde{\mathbf{p}}_{(\cdot)}}(x;t)\right|\right\}, \tag{11}$$

the L_1 norm does not place as much weight on large deviations, making it robust against a wide variety of configurations of $\hat{\mathbf{p}}_{(\cdot)}$.

Then the identification of significant edges can be thought of either as a *least absolute deviations estimation* or an L_1 *approximation* of the form

$$\hat{t} = \operatorname*{argmin}_{t \in [0,1]} L_1\left(t; \hat{\mathbf{p}}_{(\cdot)}\right) \tag{12}$$

followed by the application of the following rule:

$$e_{(i)} \in E_0 \Longleftrightarrow \hat{p}_{(i)} > F^{-1}_{\tilde{\mathbf{p}}_{(\cdot)}}(\hat{t}). \tag{13}$$

A simple example of its use is illustrated below.

Example 1. Consider a graphical model based on an undirected graph $\mathcal{G}$ with vertex set $\mathbf{V} = \{A, B, C, D\}$. The set of possible edges of $\mathcal{G}$ contains 6 elements: (A, B), (A, C), (A, D), (B, C), (B, D) and (C, D). Suppose that that we have estimated the following confidence values:

$$\hat{p}_{AB} = 0.2242, \qquad \hat{p}_{AC} = 0.0460, \qquad \hat{p}_{AD} = 0.8935, \tag{14}$$

$$\hat{p}_{BC} = 0.3921, \qquad \hat{p}_{BD} = 0.7689, \qquad \hat{p}_{CD} = 0.9439. \tag{15}$$

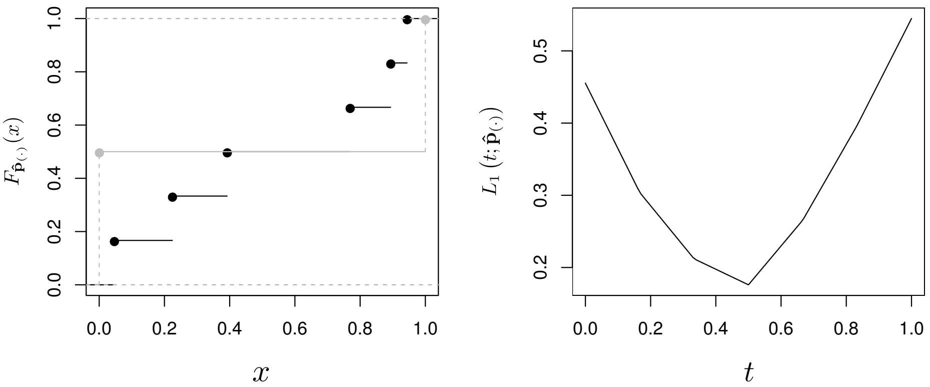

Fig. 2. The cumulative distribution functions $F_{\hat{\mathbf{p}}_{(\cdot)}}$ and $F_{\tilde{\mathbf{p}}_{(\cdot)}}(\hat{t})$, respectively in black and grey (left), and the $L_1\left(t;\hat{\mathbf{p}}_{(\cdot)}\right)$ norm (right) from Example 1.

Then $\hat{\mathbf{p}}_{(\cdot)} = \{0.0460, 0.2242, 0.3921, 0.7689, 0.8935, 0.9439\}$ and

$$F_{\hat{\mathbf{p}}_{(\cdot)}}(x) = \begin{cases} 0 & \text{if } x \in (-\infty, 0.0460) \\ \frac{1}{6} & \text{if } x \in [0.0460, 0.2242) \\ \frac{2}{6} & \text{if } x \in [0.2242, 0.3921) \\ \frac{3}{6} & \text{if } x \in [0.3921, 0.7689) \\ \frac{4}{6} & \text{if } x \in [0.7689, 0.8935) \\ \frac{5}{6} & \text{if } x \in [0.8935, 0.9439) \\ 1 & \text{if } x \in [0.9439, +\infty) \end{cases} . \tag{16}$$

The L_1 norm takes the form

$$\begin{aligned} L_1\left(t;\hat{\mathbf{p}}_{(\cdot)}\right) = |0-t|(0.0460-0) + \left|\frac{1}{6}-t\right|(0.2242-0.0460) + \\ \left|\frac{2}{6}-t\right|(0.3921-0.2242) + \left|\frac{3}{6}-t\right|(0.7689-0.3921) + \\ \left|\frac{4}{6}-t\right|(0.8935-0.7689) + \left|\frac{5}{6}-t\right|(0.9439-0.8935) + \\ |1-t|\,(1-0.9439) \end{aligned} \tag{17}$$

and is minimised for $\hat{t} = 0.4999816$. Therefore, an edge is deemed significant if its confidence is strictly greater than $F^{-1}_{\tilde{\mathbf{p}}_{(\cdot)}}(0.4999816) = 0.3921$, or, equivalently, if it has confidence of at least 0.7689; only (A, D), (B, D) and (C, D) satisfy this condition.

3 Applications to Gene Expression Profiles

We will now examine the effectiveness of the proposed estimator for the significance threshold on two gene expression data sets from Nagarajan et al. [25] and Sachs et al. [30]. All the analyses will be performed with the bnlearn package [31, 32] for R [29], which implements several methods for structure learning, parameter estimation and inference on Bayesian networks. Following Imoto et al. [16], we will consider the edges of the Bayesian networks disregarding their direction. Edges identified as significant will be oriented according to the direction observed with the highest frequency in the bootstrapped networks $\mathcal{G}_b$. This combined approach allows the proposed estimator to handle the edges whose direction cannot be determined by the structure learning algorithm (which are called *score equivalent edges* [3]), because directions are completely ignored in the estimation. At the same time, it can be observed that in practise the two possible orientations of such edges usually appear with comparable frequencies in the networks $\mathcal{G}_b$. Therefore, proper interpretation of their meaning in the network structure resulting from the application of the approach outlined in Sec. 2 is possible.

3.1 Differentiation Potential of Aged Myogenic Progenitors

In a recent study [25] the interplay between crucial myogenic (Myogenin, Myf-5, Myo-D1), adipogenic (C/EBPα, DDIT3, FoxC2, PPARγ), and Wnt-related genes (Lrp5, Wnt5a) orchestrating aged myogenic progenitor differentiation was investigated by Nagarajan et al. using clonal gene expression profiles in conjunction with Bayesian network structure learning techniques. The objective was to investigate possible functional relationships between these diverse differentiation programs reflected by the edges in the resulting networks. The clonal expression profiles were generated from RNA isolated across 34 clones of myogenic progenitors obtained across 24-month-old mice and real-time RT-PCR was used to quantify the gene expression. Such an approach implicitly accommodates inherent uncertainty in gene expression profiles and justified the choice of probabilistic models.

In the same study, the authors proposed a non-parametric resampling approach to identify significant functional relationships. Starting from Friedman's definition of confidence levels (Eq. 1), they computed the *noise floor distribution* $\hat{\mathbf{f}} = \{\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_k\}$ of the edges by randomly permuting the expression of each gene and performing Bayesian network structure learning on the resulting data sets. An edge e_i was deemed significant if $\hat{p}_i > \max(\hat{\mathbf{f}})$. In addition to revealing several functional relationships documented in literature, the study also revealed new relationships that were immune to the choice of the structure learning techniques. These results were established across clonal expression data normalised using three different housekeeping genes and networks learned with three different structure learning algorithms.

The approach presented in [25] has two important limitations. First, the computational cost of generating the noise floor distribution may discourage its

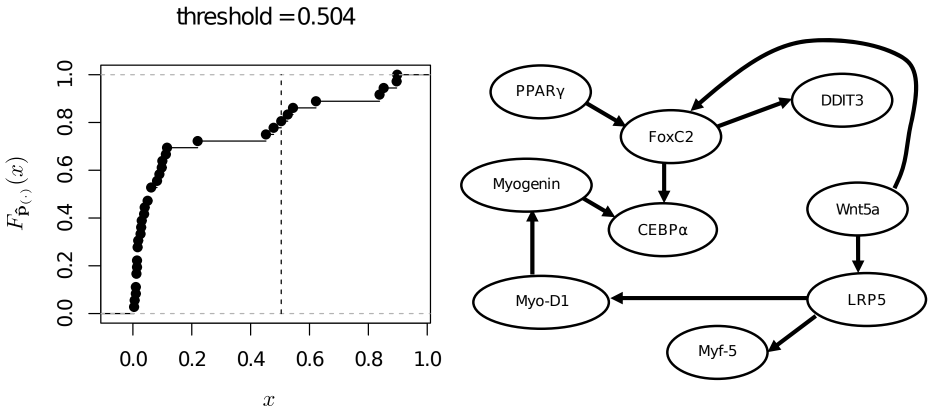

Fig. 3. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$ for the myogenic progenitors data from Nagarajan et al. [25] (on the left), and the network structure resulting from the selection of the significant edges (on the right). The vertical dashed line in the plot of $F_{\hat{\mathbf{p}}_{(\cdot)}}$ represents the threshold $F^{-1}_{\hat{\mathbf{p}}_{(\cdot)}}(\hat{t})$.

application to large data sets. In fact, the generation of the required permutations of the data and the subsequent structure learning (in addition to the bootstrap resampling and the subsequent learning required for the estimation of $\hat{\mathbf{p}}$) essentially doubles the computational complexity of Friedman's approach. Second, a large sample size may result in an extremely low value of $\max(\hat{\mathbf{f}})$, and therefore in a large number of false positives.

In the present study, we re-investigate the myogenic progenitor clonal expression data normalised using housekeeping gene GAPDH with the approach outlined in Sec. 2 and a constraint-based learning strategy based on the Incremental Association Markov Blanket (IAMB) algorithm [33]. The latter is used to learn the Markov blanket of each vertex as a preliminary step to reduce the number of its candidate parents and children; a network structure satisfying these constraints is then identified as in the Grow-Shrink algorithm [23]. It is important to note that this strategy was also used in the original study [25], hence its choice. The order statistic $\hat{\mathbf{p}}_{(\cdot)}$ was computed from 500 bootstrap samples. The empirical cumulative distribution function $F_{\hat{\mathbf{p}}_{(\cdot)}}$, the estimated threshold and the network with the significant edges are shown in Fig. 3.

All edges identified as significant in the earlier study [25] across the various structure learning techniques and normalisations techniques were also identified by the proposed approach (see Fig. 3D in [25]). In contrast to Fig. 3, the original study using IAMB and normalisations with respect to GAPDH alone detected a considerable number of additional edges (see Fig. 3A in [25]). Thus it is quite possible that the approach proposed in this paper reduces the number of false positives and spurious functional relationships between the genes. Furthermore, the application of the proposed approach in conjunction with the algorithm from

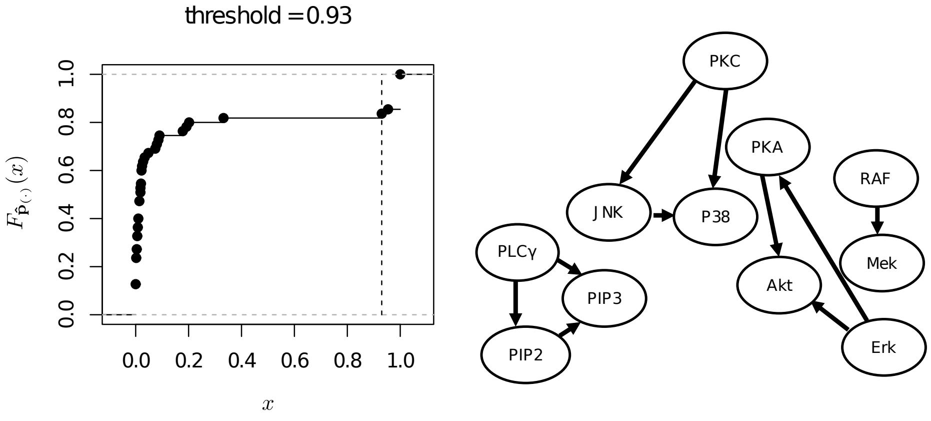

Fig. 4. The empirical cumulative distribution function of $\hat{\mathbf{p}}_{(\cdot)}$ for the flow cytometry data from Sachs et al. [30] (on the left), and the network structure resulting from the selection of the significant edges (on the right). The vertical dashed line in the plot of $F_{\hat{\mathbf{p}}_{(\cdot)}}$ represents the threshold $F^{-1}_{\tilde{\mathbf{p}}_{(\cdot)}}(\hat{t})$.

Imoto et al. [16] reveals directionality of the edges, in contrast to the undirected network reported by Nagarajan et al. [25].

3.2 Protein Signalling in Flow Cytometry Data

In a recent study, Sachs et al. [30] used Bayesian networks as a tool for identifying causal influences in cellular signalling networks from simultaneous measurement of multiple phosphorylated proteins and phospholipids across single cells. The authors used a battery of perturbations in addition to the unperturbed data to arrive at the final network representation. A greedy search score-based algorithm that maximises the posterior probability of the network [14] and accommodates for variations in the joint probability distribution across the unperturbed and perturbed data sets was used to identify the edges [6]. More importantly, significant edges were selected using an arbitrary significance threshold of 0.85 (see Fig. 3, [30]). A detailed comparison between the learned network and functional relationships documented in literature was presented in the same study.

We investigate the performance of the proposed approach in identifying significant functional relationships from the same experimental data. However, we limit ourselves to the data recorded without applying any molecular intervention, which amount to 854 observations for 11 variables. We compare and contrast our results to those obtained using an arbitrary threshold of 0.85. The combination of perturbed and non-perturbed observations studied in Sachs et al. [30] cannot be analysed with our approach, because each subset of the data follows a different probability distribution and therefore there is no single "true" network $\mathcal{G}_0$. Analysis of the unperturbed data using the approach presented in Sec. 2 reveals the edges reported in the original study. The resulting network is shown in Fig.

4 along with $F_{\hat{\mathbf{p}}_{(\cdot)}}$ and the estimated threshold. From the plot of $F_{\hat{\mathbf{p}}_{(\cdot)}}$ we can clearly see that significant and non-significant edges present widely different levels of confidence, to the point that any threshold between 0.4 and 0.9 results in the same network structure. This, along with the value of the estimated threshold ($\hat{p}_{(i)} \geqslant 0.93$), shows that the noisiness of the data relative to the sample size is low. In other words, the sample is big enough for the structure learning algorithm to reliably select the significant edges. The edges identified by the proposed method were the same as those identified by [30] using general stimulatory cues excluding the data with interventions (see Fig. 4A in [30], Supplementary Information). In contrast to [30], using Imoto et al. [16] approach in conjunction with the proposed thresholding method we were able to identify the direction of the edges in the network. The directionality correlated with functional relationships documented in literature (Tab. 3, [30], Supplementary Information) as well as with the directionality of the network learned from both perturbed and unperturbed data (Fig. 3, [30]).

4 Conclusions

Network abstractions provided by graphical models have enjoyed considerable attention across the biological and medical communities, where they are used to represent the concerted working as a system as opposed to independent entities. For example, these networks may represent the underlying signalling mechanisms and pathways within the context of biological data. Classic model validation techniques identify significant edges using an ad-hoc threshold across multiple realisations of networks learned from the given data. Such ad-hoc approaches can have pronounced effect on the resulting networks and biological conclusions. The present study overcomes this critical caveat by proposing a more straightforward and statistically-motivated approach for identifying significant edges in a graphical model. The proposed estimator minimises the L_1 norm between the cumulative distribution function of the observed confidence levels and the cumulative distribution function of the "edge confidence" determined from the given data. The effectiveness of the proposed approach is demonstrated on gene expression data sets across two different studies [25, 30]. However, the approach is defined in a more general setting and can be applied to many classes of graphical models learned from any kind of data. A more detailed investigation is underway in elucidating the various aspects of the proposed approach.

Acknowledgements

This work was supported by the National Library of Medicine grant R03LM008853 (Radhakrishnan Nagarajan) and BBSRC & Technology Board grant TS/I002170/1 (Marco Scutari). Marco Scutari would also like to thank Adriana Brogini for proofreading the paper and providing useful suggestions.

Bibliography

[1] Bromberg, F., Margaritis, D., Honavar, V.: Efficient Markov Network Structure Discovery using Independence Tests. Journal of Artificial Intelligence Research 35, 449–485 (2009)

[2] Castelo, R., Roverato, A.: A Robust Procedure For Gaussian Graphical Model Search From Microarray Data With p Larger Than n. Journal of Machine Learning Research 7, 2621–2650 (2006)

[3] Chickering, D.M.: A Transformational Characterization of Equivalent Bayesian Network Structures. In: Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI95). pp. 87–98 (1995)

[4] Chickering, D.M.: Optimal Structure Identification with Greedy Search. Journal of Machine Learning Resesearch 3, 507–554 (2002)

[5] Claeskens, G., Hjort, N.L.: Model Selection and Model Averaging. Cambridge University Press (2008)

[6] Cooper, G.F., Yoo, C.: Causal Discovery from a Mixture of Experimental and Observational Data. In: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI). pp. 116–125. Morgan Kaufmann (1999)

[7] Csiszár, I., Shields, P.: Information Theory and Statistics: A Tutorial. Now Publishers Inc. (2004)

[8] Edwards, D.I.: Introduction to Graphical Modelling. Springer, 2nd edn. (2000)

[9] Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman & Hall (1993)

[10] Elidan, G.: Bayesian Network Repository (2001), http://www.cs.huji.ac.il/site/labs/compbio/Repository

[11] Friedman, N., Goldszmidt, M., Wyner, A.: Data Analysis with Bayesian Networks: A Bootstrap Approach. In: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99). pp. 206 – 215. Morgan Kaufmann (1999)

[12] Friedman, N., Pe'er, D., Nachman, I.: Learning Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm. In: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence (UAI). pp. 206–221. Morgan Kaufmann (1999)

[13] Geiger, D., Heckerman, D.: Learning Gaussian Networks. Tech. rep., Microsoft Research, Redmond, Washington (1994), Available as Technical Report MSR-TR-94-10

[14] Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Machine Learning 20(3), 197–243 (September 1995), Available as Technical Report MSR-TR-94-09

[15] Husmeier, D.: Sensitivity and Specificity of Inferring Genetic Regulatory Interactions from Microarray Experiments with Dynamic Bayesian Networks, vol. 19 (2003)

[16] Imoto, S., Kim, S.Y., Shimodaira, H., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S.: Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression. Genome Informatics 13, 369–370 (2002)

[17] Jungnickel, D.: Graphs, Networks and Algorithms. Springer-Verlag, 3rd edn. (2008)

[18] Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)

[19] Kolmogorov, A.N., Fomin, S.V.: Elements of the Theory of Functions and Functional Analysis. Graylock Press (1957)

[20] Korb, K., Nicholson, A.: Bayesian Artificial Intelligence. Chapman and Hall (2004)

[21] Larrañaga, P., Sierra, B., Gallego, M.J., Michelena, M.J., Picaza, J.M.: Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma. In: Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe (AIME '97). pp. 261–272. Springer (1997)

[22] Lauritzen, S.L.: Graphical Models. Oxford University Press (1996)

[23] Margaritis, D.: Learning Bayesian Network Model Structure from Data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA (May 2003), Available as Technical Report CMU-CS-03-153

[24] Murphy, P., Aha, D.: UCI Machine Learning Repository (1995), http://archive.ics.uci.edu/ml

[25] Nagarajan, R., Datta, S., Scutari, M., Beggs, M.L., Nolen, G.T., Peterson, C.A.: Functional Relationships Between Genes Associated with Differentiation Potential of Aged Myogenic Progenitors. Frontiers in Physiology 1(21), 1–8 (2010)

[26] Neapolitan, R.E.: Learning Bayesian Networks. Prentice Hall (2003)

[27] Nocedal, J., Wright, S.J.: Numerical Optimization. Springer-Verlag (1999)

[28] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988)

[29] R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010), http://www.R-project.org

[30] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science 308(5721), 523–529 (2005)

[31] Scutari, M.: Learning Bayesian Networks with the bnlearn R Package. Journal of Statistical Software 35(3), 1–22 (2010)

[32] Scutari, M.: bnlearn: Bayesian Network Structure Learning (2011), http://www.bnlearn.com/, R package version 2.4

[33] Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Algorithms for Large Scale Markov Blanket Discovery. In: Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference. pp. 376–381. AAAI Press (2003)

[34] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning 65(1), 31–78 (2006)

[35] Whittaker, J.: Graphical Models in Applied Multivariate Statistics. Wiley (1990)