Modeling cumulative biological phenomena with Suppes-Bayes causal networks

Daniele Ramazzotti[1], Alex Graudenzi[1,2], and Marco Antoniotti[1,3]

[1] Department of Informatics, Systems and Communication,
University of Milan-Bicocca, Milan, Italy
[2] Institute of Molecular Bioimaging and Physiology,
Italian National Research Council (IBFM-CNR), Milan, Italy
[3] Milan Center for Neuroscience,
University of Milan-Bicocca, Milan, Italy

Abstract

Several statistical techniques have been recently developed for the inference of cancer progression models from the increasingly available NGS cross-sectional mutational profiles. A particular algorithm, CAPRI, was proven to be the most efficient with respect to sample size and level of noise in the data. The algorithm combines structural constraints based on Suppes' theory of probabilistic causation and maximum likelihood fit with regularization, and defines constrained Bayesian networks, named Suppes-Bayes Causal Networks (SBCNs), which account for the selective advantage relations among genomic events. In general, SBCNs are effective in modeling any phenomenon driven by cumulative dynamical, as long as the modeled events are persistent. We here discuss on the effectiveness of the SBCN theoretical framework and we investigate the influence of: (*i*) the priors based on Suppes' theory and (*ii*) different maximum likelihood regularization parameters on the inference performance estimated on large synthetically generated datasets.

1 Introduction

Cancer development is characterized by the subsequent *accumulation* of specific alterations of the (epi)genome (i.e., *drivers*), which confer a functional *selective advantage* to the mutant clones. In fact, given that during clonal expansions individual tumor cells compete for (limited) space and resources, the fittest variants are naturally selected for and eventually outgrow the competing cells [7]. From the modeling perspective, cancer progression can be considered as a dynamical process in which specific events (i.e., drivers) monotonically accumulate defining definite temporal trajectories.

In the last decades the genomic interplay ruling cancer development has been largely investigated and the huge technological advancements that led, e.g., to the development of *Next Generation Sequencing* (NGS) techniques, produced huge amounts of data, which are increasingly available thanks to projects such as *The Cancer Genome Atlas* (TCGA, [12]). Nevertheless, the understanding of the temporal sequences in which driver alterations fixate, eventually leading to the emergence and development of the disease, is still largely unknown. This is mainly due to the nature of most of the accessible data, which are *cross-sectional*, meaning that the samples (i.e., biopsies) are usually collected at the time of diagnosis (or at most at very few points in time), rather than over the whole course of the disease.

To overcome this hurdle, several statistical techniques to infer cancer progression models from cross-sectional mutational profiles have been developed in recent years, starting from the seminal work on *single-path-models* [18], up to *tree models*, capturing evolutionary branches (see, e.g., [5]), and more complex Bayesian graphical models (i.e., *Bayesian Networks* [9]), which allow for converging evolutionary paths (see, e.g., [2]) Our group has been lately focused on this problem, by developing a novel framework based on the theory of *probabilistic causation* by Patrick Suppes [17]. One first algorithm, CAPRESE (CAncer PRogression with Single Edges, [11]) relies on a *shrinkage*-like estimator for the inference of tree models, whereas a more recent method named CAPRI (CAncer PRogression Inference, [14]) further extends this approach by defining a constrained Bayesian graphical model, and currently represents the state-of-the-art in terms of inference performance with respect to sample size and noise in the data. Both the algorithms are currently implemented in TRONCO, an open source R package for Translational Oncology available in standard repositories [4].

In particular, CAPRI aims at extracting constrained Bayesian networks (also defined as *Suppes-Bayes Causal Network*, SBCN [3]) accounting for the conditional dependencies[1] among genomic events, by combining specific priors based on Suppes' theory with a *maximum likelihood*-fit procedure (with regularization). Hence, CAPRI captures the essential aspects of cancer evolution: branches, confluences and independent progressions, being one of the main novelties the exploitation of (input) groups of exclusive alterations to detect fitness-equivalent routes of progression. We also remark that, even if originally conceived for cancer progression inference, SBCNs can effectively describe the dynamics of any system driven by the monotonic accumulation of events, as long as these events are *persistent*.

In this work we: (i) formally define the theoretical framework underlying SBCNs, assessing their relevance in modeling cumulative phenomena and, (ii) investigate the influence of (a) Suppes' priors and (b) distinct maximum likelihood regularization parameters on CAPRI's inference performance. The performance was assessed by evaluating standard measures such as, e.g., accuracy, sensitivity and specificity, on a large number of synthetic datasets. The datasets were generated from randomly parametrized progression models with distinct key features (such as, e.g., topology and number of nodes), and with distinct sample size and levels of noise.

The paper is structured as follows. In Section 2 the SBCNs are formally defined and discussed. In Section 3 the results of the performance evaluation on synthetic data is presented, whereas in Section 4 a brief discussion on results and further development is provided.

2 Background: Suppes-Bayes causal networks

In [17], Suppes introduced the notion of *prima facie causation*. A prima facie relation between a cause u and its effect v is verified when the following two conditions are true: (i) *temporal priority* (TP), i.e., any cause happens before its effect and (ii) *probability raising* (PR), i.e., the presence of the cause raises the probability of observing its effect.

Definition 1 (Probabilistic causation, [17]). *For any two events u and v, occurring respectively at times t_u and t_v, under the mild assumptions that $0 < P(u), P(v) < 1$, the event u is called a* prima facie cause *of v if it occurs* before *and* raises the probability *of u, i.e.,*

$$\begin{cases} (TP) & t_u < t_v \\ (PR) & P(v \mid u) > P(v \mid \overline{u}) \end{cases} \tag{1}$$

While the notion of prima facie causation has known limitations in the context of the general theories of causality [8], this formulation seems to intuitively characterizes the dynamics of phenomena driven by the monotonic accumulation of events[2] where a temporal order among

[1] Also defined as *selective advantage relations*.

[2] And, specifically, it seems appropriate to describe the notion of selective advantage of somatic alterations that accumulate during tumor progression.

the events is implied and, furthermore, the occurrence of an early event positively correlates to the subsequent occurrence of a later one.

Let us now consider a graphical representation of the aforementioned dynamics in terms of a Bayesian graphical model. A *Bayesian network* (BN) is a statistical model which succinctly represents the conditional dependencies among a set of *random variables* through a *directed acyclic graph* (DAG). Formally, BNs are DAGs, i.e., $G = (V, E)$, where the nodes V represent random variables and the arcs E encode the conditional dependencies among the variables [9].

Therefore we can model the dynamics of cumulative phenomena by means of a specific set of the general BNs where the nodes V represent the accumulating events as *Bernoulli random variables* taking values in $\{0, 1\}$ based on their occurrence: the value of the variable is 1 if the event is observed and 0 otherwise.

Moreover, let us now consider a given node $v_i \in V$ and let us name $Pa(v_i)$ the set of all the nodes in V pointing to (and yet temporally preceding) v_i. Then, the joint probability distribution of the $n = |V|$ variables can be written as:

$$P(v_1, \ldots, v_n) = \prod_{v_i \in V} P(v_i | Pa(v_i)) \tag{2}$$

Furthermore, we can define a class of BNs over Bernoulli random variables named *monotonic progression networks* (MPNs) [6, 10]. The MPNs aim at intuitively representing the progression of events that accumulate monotonically[3] over time, where the conditions for any event to happen is described by probabilistic versions of the canonical boolean operators, i.e., conjunction ($\wedge$), inclusive disjunction ($\vee$), and exclusive disjunction ($\oplus$).

Following [6] and [10], we define 3 types of MPNs given the just mentioned canonical boolean formula: a conjunctive MPN (CMPN), a disjunctive (semi-monotonic) DMPN, and an exclusive disjunction MPN (XMPN). The operator associated with each network type refers to the logical relation among the parents that eventually lead to the common effect to occur. Formally, for each of the MPNs, it holds:

CMPN:

$$\begin{cases} P(v \mid \sum Pa(v) = |Pa(v)|) = \theta \\ P(v \mid \sum Pa(v) < |Pa(v)|) \leq \epsilon \end{cases}$$

DMPN:

$$\begin{cases} P(v \mid \sum Pa(v) > 0) = \theta \\ P(v \mid \sum Pa(v) = 0) \leq \epsilon \end{cases}$$

XMPN:

$$\begin{cases} P(v \mid \sum Pa(v) = 1) = \theta \\ P(v \mid \sum Pa(v) \neq 1) \leq \epsilon \end{cases}$$

where $\theta, \epsilon \in [0, 1]$ and $\theta \gg \epsilon$. Specifically, θ represents the conditional probability of any effect to follow its preceding cause and ϵ models the probability of any noisy observation. Also note that the above inequalities define for each type of MPN the specific constraints to the induced distributions namely in terms of the probabilistic logical relations of *noisy-AND*, *noisy-OR* and *noisy-XOR* networks [13, 10].

[3] The events accumulate over time and when later events occurs earlier events are observed as well.

Given these premises, in [14] the authors describe an efficient algorithm (see Algorithm 1) to learn a constrained Bayesian network which accounts for Suppes' criteria and which are later on dubbed *Suppes-Bayes Causal Networks* (SBCNs) in [3].

We next present in details the characteristics of the SBCNs and we discuss to which extent they are capable of modeling cumulative phenomena and, in particular MPNs.

Temporal priority. The first constraint assumes an underlying temporal (partial) order among the events depicted in a SBCN. However, unfortunately, when we are dealing with cross-sectional data, we are not provided with an explicit measure of time hence the temporal order needs to be imputed[4]. To do so, in [14] the temporal priority constraint is assessed in terms of marginal frequencies: more frequent events are occurring earlier than rare ones, which is sound when we assume the accumulating events to be irreversible[5].

The impact of the assessment of the temporal priority needs a further consideration. First of all we observe that the general problem of learning the structure of a Bayesian Network is NP-hard [9]. But, as a result of the assessment of temporal priority, we constrain our search space to the networks with a given order: because of time irreversibility, temporal priority produces a partially order set (poset) which induces a set of total orders to the considered variables. Learning networks given a fixed order may become tractable subject to the cardinality of the parent set of any node, and, in general, it turns out to be easier than the general case [9].

Probability raising. As a second constraint, we further reduce the network of the temporal priority by removing the arcs which are not consistent to the condition of probability raising. We recall that the probability raising condition is equivalent to constraining for positive statistical dependence [11]: we model all and only the positive dependant relations among the nodes ordered by temporal priority, consistently with Suppes' characterization of causation in terms of relevance. If the previous condition reduces the search space of the possible valid structures for the network by setting a specific partial order to the nodes, the probability raising condition instead reduces the search space of the possible valid parameters of the network by requiring that the related conditional probability tables account only for positive statistical dependencies.

In particular, for any pair of nodes of a SBCN learned by Algorithm 1 for which we have a directed edge, say, from node u to node v, it holds[6]:

$$\begin{cases} P(v \mid u) = \theta \\ P(v \mid \overline{u}) \leq \epsilon \end{cases}$$

where once again $\theta, \epsilon \in [0, 1]$ and $\theta \gg \epsilon$.

Network simplification. Suppes' criteria are known to be necessary but not sufficient to evaluate any causal claim [14]. Especially when dealing with small sample sized datasets, the prima facie partially order set may contain *false positive* (spurious causes) arcs, in spite of very few *false negatives*. Consequently, although we expect all the statistically relevant causal relations to be modelled in the prima facie network, we also expect some spurious arcs in it.

The last step of the learning procedure of a SBCN aims at overcoming this issue by learning the network by likelihood fit estimation within the search space of the valid partially order

[4] We notice that in the case we were provided with explicit observations of time, the temporal priority would be directly and, yet, more efficently assessed.

[5] This assumption holds, e.g., for cancer driver alterations as discussed in [14].

[6] One should notice that Algorithm 1 adopts an efficient implementation of Suppes' constraints where each condition is evaluated only on pairs of nodes and not for the whole parent set with the respective limitations.

sets induced by the network of the prima facie relations. However, one must recall that, due to statistical noise and sample size, exact independence between pair of variables is never (or unlikely) observed. Hence, the likelihood score is known to overfit the data unless in the situation of infinite sample size. For this reason, in this final step, we also add to the likelihood fit a regularization term based on a prior probability that penalizes complex models [9].

Given of these considerations, following [3] we now report a formal definition a SBCN.

Definition 2 (Suppes-Bayes Causal Network)**.** Given an input cross-sectional dataset D of n Bernoulli variables and m samples, the Suppes-Bayes Causal Network $SBCN = (V, E)$ subsumed by D is a directed acyclic graph such that the following requirements hold:

[**Suppes' constraints**] for each arc $(u \rightarrow v) \in E$ involving the selective advantage relation between nodes $u, v \in V$, under the mild assumptions that $0 < P(u), P(v) < 1$*:*

$$P(u) > P(v) \quad and \quad P(v \mid u) > P(v \mid \neg u) \, .$$

[**Simplification**] let E' be the set of arcs satisfying the Suppes' constraints as before; among all the subsets of E', the set of arcs E is the one whose corresponding graph maximizes the likelihood of the data and of a certain regularization function $R(f)$:

$$E = \underset{E \subseteq E', G=(V,E)}{\arg\max} \left(LL(D|G) - R(f) \right) .$$

Logical patterns. We now observe that the efficient implementation of Suppes' constraints of Algorithm 1 does not, in general, guarantee to converge to the monotonic progression networks depicted before. In fact, the probabilistic relations of the MPNs are defined on the $Pa(v_i)$, while Algorithm 1 only consider pair of nodes rather than any pair v_i and $Pa(v_i)$. To overcome this limitation, one could extend Algorithm 1 in order to learn together both the network structure and the parent sets with the respective relations among the parents with an obvious blow in the complexity of the algorithm.

In [14], the authors introduce the notion of *progression pattern* to underline this situation. Also, they claim that Algorithm 1 can infer any MPN provided that the dataset given as input is *lifted* ([14]) with a Bernoulli variable per causal relation representing the logical formula involving any $Pa(v_i)$, but, it can also infer any CMPN without any further input[7].

We conclude by reporting the pseudocode (see Algorithm 1) of the efficient learning algorithm to infer SBCNs presented in [14], which will be used for the assesment of the performance by the simulations of the next Section.

3 Results and discussion

We now evaluate the performance of Algorithm 1 on simulated data with specific attention on how the constraints based on Suppes' theory of probabilistic causation impact the performance. All the simulations are performed with the following settings.

We consider 6 different topological structures as depicted in Figure 1: the first two where any node has at the most one predecessor, i.e., (i) trees, (ii) forests, and the others where we set a limit of 3 predecessors and, hence, we consider (iii) directed acyclic graphs with a single source and conjunctive parents, (iv) directed acyclic graphs with multiple sources and conjunctive parents, (v) directed acyclic graphs with a single source and disjunctive parents,

[7] This is happening because of the properties of the noisy-AND logical operator.

Algorithm 1: CAPRI

```
Input: a dataset $D$ of $n$ Bernoulli variables, e.g., genomic alterations or patterns, and $m$ samples.
Result: a graphical model $G = (V, E)$ representing all the relations of "probabilistic causation".
Let $\mathcal{G} \leftarrow$ a directed graph over the vertices $n$;
forall the arcs $(u, v) \in \mathcal{G}$ do
    Compute a score $S(\cdot)$ for the nodes $u$ and $v$ in terms of Suppes' criteria;
    Remove the arc $(u, v)$ if Suppes' criteria are not met;
end
Let $E \leftarrow$ the subset of the remaining arcs $E' \in \mathcal{G}$, that maximize the log-likelihood of the model, computed as: $LL(D|\mathcal{G}) - R(f)$;
return The resulting graphical model $G = (V, E)$.
```

(vi) directed acyclic graphs with multiple sources and disjunctive parents. For each of these configurations, we generate 100 random structures.

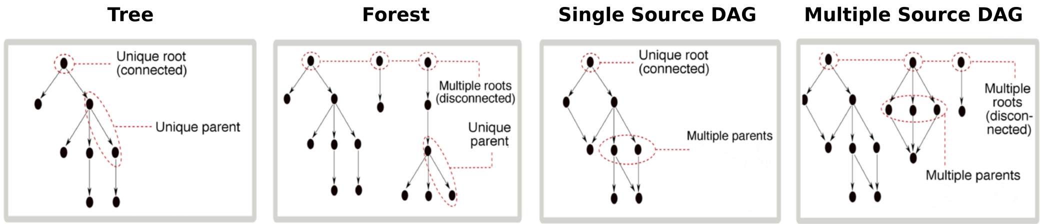

Figure 1: We consider the topological structures shown in the Figure. From left to right an example of tree, forest, directed acyclic graphs with a single source and directed acyclic graphs with multiple sources. For directed acyclic graphs we considered either conjunctive or disjunctive relations among set of parents of the same common effect.

Moreover, we consider 4 different sample sizes (50, 100, 150 and 200 samples) and 9 noise levels (i.e., probability of a random entry for the observation of any node in a sample) from 0% to 20% with step 2.5%. Furthermore, we repeat the above settings for networks of 10 and 15 nodes. Any configuration is then sampled 10 times independently for a total of more than 4 millions different simulated datasets.

Finally, the performance is assessed in terms of: $accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$, $sensitivity = \frac{TP}{(TP+FN)}$ and $specificity = \frac{TN}{(FP+TN)}$ with TP and FP being the true and false positive (we define positive any arc that is present in the network) and TN and FN being the true and false negative (we define negative any arc that is not present in the network). All these measures are values in $[0, 1]$ with results close to 1 indicators of good performance.

Let n be the number of nodes we want to include in the network and let $p_{\min} = 0.05$ and $p_{\max} = 0.95$ be the minimum and maximum probabilities of any node. A *directed acyclic graph without disconnected components* (i.e. an instance of types (iii) and (v) topologies) with maximum depth $\log n$ and where each node has at most $w^* = 3$ parents is generated as shown in Algorithm 2.

Algorithm 2: Data generation: single source directed acyclic graphs

```
Input: $n$, the number of nodes of the graph, $p_{\min} = 0.05$ and $p_{\max} = 0.95$ be the
       minimum and maximum probabilities of any node and $w^* = 3$ the maximum
       incoming edges per node.
Result: a randomly generated single source directed acyclic graph.
Pick an event $r \in G$ as the root of the directed acyclic graph;
Assign to each node $u \neq r$ an integer in the interval $[2, \lceil \log n \rceil]$ representing its depth in
the graph (1 is reserved for $r$), ensuring that each level has at least one node;
forall the nodes $u \neq r$ do
    Let $l$ be the level assigned to the node;
    Pick $|P(u)|$ uniformly over $(0, w^*]$, and accordingly define the parents of $u$ with
    events selected among those at which level $l - 1$ was assigned;
end
Assign $P(r)$, a random value in the interval $[p_{\min}, p_{\max}]$;
forall the events $u \neq r$ do
    Let $\alpha$ be a random value in the interval $[p_{\min}, p_{\max}]$;
    Let $Pa(u)$ be the direct predecessor of $u$;
    Then assign:
        $$P(u) = \alpha P(x \in Pa(u))\,;$$
end
return The generated single source directed acyclic graph.
```

In Figures 2, 3 and 4 we show the performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of accurancy, sensitivity and specificity for different settings which we discuss in details in the next Paragraphs.

Suppes' prima facie conditions are necessary but not sufficient. We first discuss the performance by applying *only* the prima facie criteria and we evaluate the obtained prima facie network in terms of accurancy, sensitivity and specificity on simulated datasets of 100 samples and networks of 15 nodes (see Figures 2, 3 and 4). As expected, the sensitivity is much higher than the specificity implying the significant impact of false positives rather than false negatives for the networks of the prima facie arcs. This result is indeed expected being Suppes' criteria mostly capable of removing some of the arcs which do not represent valid causal relations rather than asses the exact set of valid arcs. Interestingly, the false negatives are still limited even when we consider DMPN, i.e., when we do not have guarantees for Algorithm 1 to converge. The same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

The likelihood function overfits the data. In Figures 2, 3 and 4 we also show the performance of the inference by likelihood fit (without any regularizator) on the prima facie network in terms of accurancy, sensitivity and specificity on simulated datasets of 100 samples and networks of 15 nodes. Once again, in general the sensitivity is much higher than the specificity implying also in this case a significant impact of false positives rather than false negatives for the inferred networks. This results make explicit the need for a regularization heuristic when dealing with real (not infinite) sample sized datasets as discussed in the next Paragraph. Another interesting consideration comes from the observation that the prima facie networks and

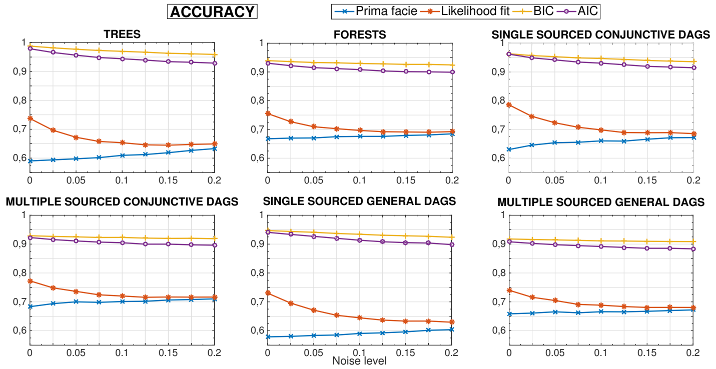

Figure 2: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of accurancy for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

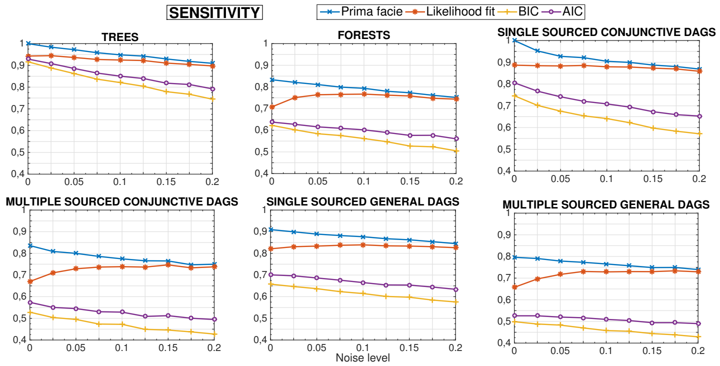

Figure 3: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of sensitivity for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

the ones by likelihood fit without regularization seems to converge to the same performance as the noise level increases. This is due to the fact that, in general, the prima facie constraints are very conservative in the sense that false positives are admitted as long as false negatives are limited. When the noise level increases, the positive dependencies among nodes are generally reduced and, hence, less arcs pass the prima facie cut for positive dependency. Also in this case, the same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

Performance with respect to different regularizators. We now investigate the role of different regularizations on the performance. In particular, we consider two commonly used regularizations: (i) the *Bayesian information criterion* (BIC) [15] and (ii) the *Akaike information criterion* (AIC) [1].

Although BIC and AIC are both scores based on maximum likelihood estimation and a penalization term to reduce overfitting, yet with distinct approaches, they produce significantly different behaviors. More specifically, BIC assumes the existence of one *true* statistical model

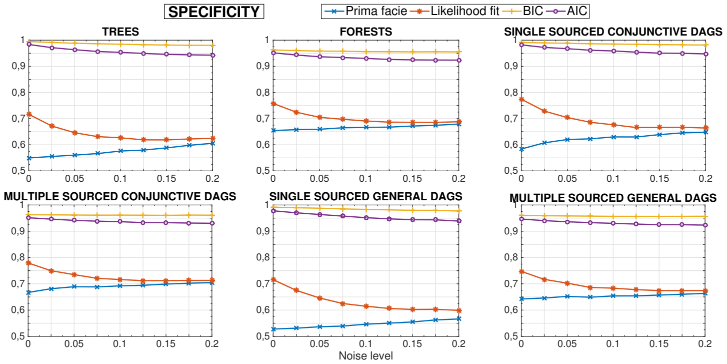

Figure 4: Performance of the inference on simulated datasets of 100 samples and networks of 15 nodes in terms of specificity for the 6 considered topological structures. The y axis refers to the performance while the x axis represent the different noise levels.

which is generating the data while AIC aims at finding the best approximating model to the unknown data generating process. As such, BIC presents the danger that it might underfit, whereas AIC presents the danger that it might overfit[8].

The performance on simulated datasets are shown in Figure 2, 3, 4. In general, the performance are improved in all the settings with both the regularizators and we observe that the fact that the purpose of regularizator is to shrinks the results toward sparse networks is made explicit by the observed drop in the performance in terms of sensitivity together with a big improvement in terms of specificity. Overall, the accurancy of the inference is improved.

Furthermore, we observe that the performance of Algorithm 1 are still good even when we consider simulated data generated by DMPN. Although in this case we do not have any guarantee of convergence, in practice the algorithms seems capable of approximate the generative model. In conclusion, without any further input, SBCNs inferred by Algorithm 1 can model CMPNs and, yet, depict the more significant arcs of DMPNs.

The same simulations with different sample sizes (50, 150 and 200 samples) and on networks of 10 nodes present a similar trend (results not shown here).

4 Conclusions

In this work we investigated the properties of a constrained version of Bayesian network, named SBCN, which is particularly sound in modeling the dynamics of system driven by the monotonic accumulation of events, thanks to encoded priors based on Suppes' theory of probabilistic causation. In particular, we showed how SBCNs can, in general, describe different types of MPN, which makes them capable of characterizing a broad range of cumulative phenomena not limited to cancer evolution.

Besides, we investigated the influence of Suppes' priors on the inference performance with

[8] Thus, BIC tends to make a trade off between the likelihood and model complexity with the aim of inferring the statistical model which generates the data. This makes it useful when the purpose is to detect the best model which is describing the data. Instead, asymptotically, minimizing AIC is equivalent to minimizing the cross validation value [16]. It is this property that makes the AIC score useful in model selection when the purpose is prediction. Overall, the choise of the regularizator tunes the level of sparsity of the reulated SBCN and, yet, the confidence of the inferred arcs.

cross-sectional synthetic datasets. In particular, we showed that Suppes' constraints can be effective in building a partially order set of the accumulating events with very few false negatives, but many false positives. To overcome this limitation, we explored the role of two maximum likelihood regularization parameters, i.e., BIC and AIC, being the former more suitable to test previously conjectured hypotheses and the latter to predict novel hypotheses.

References

[1] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[2] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive bayesian networks. *Bernoulli*, pages 893–909, 2007.

[3] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *Submitted. Available on arXiv.org.*, 2015.

[4] Luca De Sano, Giulio Caravagna, Daniele Ramazzotti, Alex Graudenzi, Giancarlo Mauri, Bud Mishra, and Marco Antoniotti. TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics, in print, 2015. Archived at http://dx.doi.org/10.1101/027474 http://dx.doi.org/10.1101/027474*, 2015.

[5] Richard Desper, Feng Jiang, Olli-P Kallioniemi, Holger Moch, Christos H Papadimitriou, and Alejandro A Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology*, 6(1):37–51, 1999.

[6] Hossein Shahrabi Farahani and Jens Lagergren. Learning oncogenetic networks by reducing to mixed integer linear programming. 2013.

[7] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[8] Christopher Hitchcock. Probabilistic causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

[9] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[10] Ilya Korsunsky, Daniele Ramazzotti, Giulio Caravagna, and Bud Mishra. Inference of cancer progression models with biological noise. *arXiv preprint arXiv:1408.6032*, 2014.

[11] Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PLoS One*, 9(10), 2014.

[12] NCI and the NHGRI. The cancer genome atlas. http://cancergenome.nih.gov/, 2005.

[13] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[14] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

[15] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[16] Mervyn Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47, 1977.

[17] Patrick Suppes. *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam, 1970.

[18] Bert Vogelstein, Eric R Fearon, Stanley R Hamilton, Scott E Kern, Ann C Preisinger, Mark Leppert, Alida MM Smits, and Johannes L Bos. Genetic alterations during colorectal-tumor development. *New England Journal of Medicine*, 319(9):525–532, 1988.