Inferring causal models of cancer progression with a shrinkage estimator and probability raising

Loes Olde Loohuis[*]
Department of Computer Science
City University New York, The Graduate Center
New York, USA.

Giulio Caravagna[*] Alex Graudenzi[*]
Daniele Ramazzotti[*] Giancarlo Mauri
Marco Antoniotti
Dipartimento di Informatica Sistemistica e Comunicazione
Università degli Studi Milano-Bicocca
Milano, Italy.

Bud Mishra
Courant Institute of Mathematical Sciences
New York University
New York, USA.

Abstract

Existing techniques to reconstruct tree models of progression for accumulative processes such as cancer, seek to estimate causation by combining correlation and a frequentist notion of temporal priority. In this paper we define a novel theoretical framework to reconstruct such models based on the probabilistic notion of causation defined by Suppes, which differ fundamentally from that based on correlation. We consider a general reconstruction setting complicated by the presence of noise in the data, owing to the intrinsic variability of biological processes as well as experimental or measurement errors. To gain immunity to noise in the reconstruction performance we use a shrinkage estimator. On synthetic data, we show that our approach outperforms the state-of-the-art and, for some real cancer datasets, we highlight biologically significant differences revealed by the reconstructed progressions. Finally, we show that our method is efficient even with a relatively low number of samples and its performance quickly converges to its asymptote as the number of samples increases. Our analysis suggests the applicability of the method on small datasets of real patients.

[*] Equal contributors.

1 Introduction

Cancer is a disease of evolution. Its initiation and progression are caused by dynamic somatic alterations to the genome manifested as point mutations, structural alterations, DNA methylation and histone modification changes [1].

These genomic alterations are generated by random processes, and since individual tumor cells compete for space and resources, the fittest variants are naturally selected for. For example, if through some mutations a cell acquires the ability to ignore anti-growth signals from the body, this cell may thrive and divide, and its progeny may eventually dominate part of the tumor. This *clonal expansion* can be seen as a *discrete state* of the cancer's progression, marked by the acquisition of a set of genetic events. Cancer progression can then be thought of as a sequence of these discrete steps, where the tumor acquires certain distinct properties at each state. Different progression sequences are possible, but some are more common than others, and not every order is viable [2].

In the last two decades, many specific genes and genetic mechanisms that are involved in different types of cancer have been identified (see e.g. [3, 4] for an overview of common cancer genes and [5, 6] for specific genetic analyses of ovarian carcinoma and lung adenocarcinoma, respectively), and *targeted therapies* that aim to affect the activity of these genes are now being developed at a fast pace [2]. However, unfortunately, the *causal and temporal relations* among the genetic events driving cancer progression remain largely elusive.

The main reason for this state of affairs is that information revealed in the data is usually obtained only at one (or a few) points in time, rather than over the course of the disease. Extracting this dynamic information from the available *cross-sectional* data is challenging, and the combination of mathematical, statistical and computational techniques is needed. The results of this research will have important repercussions for disease diagnosis, prognosis, and therapy.

In recent years, several methods to extract progression models from cross-sectional data have been developed; starting from the seminal work on single-path-models by Fearon and Vogelstein [7], up to several models of oncogenetic trees [8, 9, 10], probabilistic networks [11] and conjunctive bayesian networks [12, 13]. In their essence some of these models, e.g. [8, 10, 9], use *correlation* to identify relations among genetic events. These techniques reconstruct tree models of progression as independent acyclic paths with branches and no confluences. More complex models, e.g. [12, 13], extract direct acyclic graphs however, in these cases, other constraints on the joint occurrence of events are imposed. Besides, in a slightly different context, temporal models were reconstructed from time-course gene expression data [14, 15].

In this paper we present a novel theoretical framework to reconstruct cumulative progressive phenomena, such as cancer progression. We base our method on a notion of *probabilistic causation*, more suitable than *correlation* to infer causal structures[1]. More specifically, we adopt the notion of causation pro-

[1] The assumption that correlation proves causation is the well-knonw logical fallacy *cum hoc ergo propter hoc*, similarly to the *post hoc ergo propter hoc* fallacy (i.e. an event that follows another is necessarily its consequence).

posed by Suppes in [16]. Its basic intuition is simple: event a causes event b if (*i*) a occurs *before* b and (*ii*) the occurrence of a *raises the probability* of observing b. Probabilistic causation was used in biomedical applications before (e.g., to find driver genes from CNV data in [17], and to extract causes from biological time series data in [18]), but, to the best of our knowledge, never to infer *progression models* in the *absence* of direct temporal information.

We assume the problem setting of [8] to define a technique to infer *probabilistic progression trees* from cross-sectional data, when the input is a set of pre-selected genetic events such that the presence or the absence of each event is recorded for each sample. Using the notion of probabilistic causation just described, we aim to infer a tree which best describes causal structures implicit in the data.

The problem is complicated by the presence of noise, such as the one provided by the intrinsic variability of biological processes (e.g., *genetic heterogeneity*) and *experimental errors*. To best deal with this issue we adopt a *shrinkage estimator* to measure causation among any pair of events [19]. The intuition of this type of estimators is to improve a *raw estimate* α (here probability raising) with a *correction factor* β (here a measure of temporal distance among events); a generic shrinkage estimator is defined as

$$\hat{\boldsymbol{\theta}} = (1-\lambda)\alpha(\mathbf{x}) + \lambda\beta(\mathbf{x})$$

where $0 \leq \lambda \leq 1$ is the *shrinkage coefficient*, $\mathbf{x}$ is the input data and $\hat{\boldsymbol{\theta}}$ is the estimates that we evaluate. Clearly, $\hat{\boldsymbol{\theta}}$ can be arbitrarily shrank towards α or β by varying λ, i.e. the estimator can be biased. The power of shrinkage lies in the possibility of determining an optimal value for λ to balance the effect of the correction factor on the raw model estimate. This approach is effective to regularize ill-posed inference problems, and sometimes the optimal λ can be determined analytically [20].

In our case, however, the performance we are interested in is that of the reconstruction technique, rather than that of the estimator, usually measured as *mean squared error*. Here we define performance in terms of *structural similarity* among the reconstructed trees, rather than on their induced distribution as done, e.g., in [10]. This measure helps to discriminate the optimal model among those inducing similar distributions, but it can be evaluated only when the target tree to reconstruct is known, as it happen with synthetic data (cfr. Section 5.1). On the one hand, structural equivalence is a stronger result than the distribution analogous, and is also more useful to understand progressive phenomena. On the other hand, it can not be used to compare topologically different models, e.g. [10, 12]. Thus, we compare our algorithm to the state-of-the-art method to reconstruct trees. Here we *numerically* estimate the *global optimal* coefficient value for this performance. Based on synthetic data, we show that our algorithm outperforms the existing tree reconstruction algorithm of [8]. In particular, our shrinkage estimator provides, on average, an increased robustness to noise which ensures it to outperform oncotrees [8]. Finally, we show that the method works in a very efficient way already with a relatively low number of samples and that its performance quickly converges to its asymptote as the number of samples increases. This outcome hints at the applicability of the algorithm with relatively small datasets without compromising its efficiency.

This paper is structured as follows. In Section 2 the reconstruction problem is formally defined. In Section 3 the notion of probability raising, some of its key properties, and some new results are discussed. In Section 4 we propose our novel shrinkage estimator and the algorithm for the reconstruction of tree and forest topologies is presented. In Section 5 we present the numerical estimation of the optimal shrinkage coefficient, and compare of our algorithm to *oncotrees* [8] using synthetic data (in Section 5.3), as well as on actual patient data (in Section 5.4). We conclude with Section 6.

2 Problem setting

The set-up of the reconstruction problem is as follows. Assuming that we have a set G of n mutations (*events*, in probabilistic terminology) and m samples, we can represent a cross-sectional dataset as an $m \times n$ binary matrix. In this matrix, an entry $(k, l) = 1$ if the mutation l was observed in sample k, and 0 otherwise. We reemphasize that such a dataset does not provide explicit information of time. The problem we solve is to extract a set of edges E yielding a progression *tree* $\mathcal{T} = (G \cup \{\diamond\}, E, \diamond)$ from this matrix. More precisely, we aim at reconstructing a *rooted tree* that satisfies: (i) each node has at most one incoming edge, (ii) the root has no incoming edges (iii) there are no *cycles*. The root of $\mathcal{T}$ is modeled using a (special) event $\diamond \notin G$ to extract, in principle, *heterogenous progression paths*, i.e. *forests*.

Each progression tree subsumes a distribution of observing a subset of mutations in a cancer sample.

Definition 1 (Tree-induced distribution)**.** *Let $\mathcal{T}$ be a tree and $\alpha : E \to [0, 1]$ a labeling function denoting the independent probability of each edge, $\mathcal{T}$ generates a distribution where the probability of observing a sample with the set of alterations $G^* \subseteq G$ is*

$$\mathcal{P}(G^*) = \prod_{e \in E'} \alpha(e) \cdot \prod_{\substack{(u,v) \in E \\ u \in G^*, v \notin G}} \Big[1 - \alpha(u, v)\Big] \tag{1}$$

where $E' \subseteq E$ is the set of edges connecting the root $\diamond$ to the events in G^*.

The *temporal priority* principle states that all causes must precede their effects [21]. This distribution subsumes that, for any oriented edge $(a \to b)$, a sample contains alteration b with probability $\mathcal{P}(a)\mathcal{P}(b)$, that is the probability of observing a is greater than the probability of observing b.

The notion of tree-induced distribution can be used to state an important aspect which hardens the reconstruction problem. The input data is a set of samples generated, ideally, from an unknown distribution induced by an unknown tree that we aim to reconstruct. However, in some cases it could be that no tree exists whose induced distribution generates *exactly* those data. When this happens, the set of observed samples slightly diverges from any tree-induced distribution. To model these situations a notion of *noise* can be introduced, which depends on the context in which data are gathered, as we discuss in Section 5.

2.1 The *oncotree* approach

In [8] Desper *et al.* developed a method to extract progression trees, named *"oncotrees"*, from static CNV data. In these trees, nodes represent CNV events and edges correspond to possible progressions from one event to the next.

The reconstruction problem is exactly as described above, and each tree is rooted in the special event $\diamond$. The choice of which edge to include in a tree is based on the estimator

$$w_{a\to b} = \log\left[\frac{\mathcal{P}(a)}{\mathcal{P}(a)+\mathcal{P}(b)} \cdot \frac{\mathcal{P}(a,b)}{\mathcal{P}(a)\mathcal{P}(b)}\right], \tag{2}$$

which assigns, to each edge $a \to b$, a weight accounting for both the relative and joint frequencies of the events – thus measuring *correlation.* The estimator is evaluated after including $\diamond$ to each sample of the dataset. In this definition the rightmost term is the (symmetric) *likelihood ratio* for a and b occurring together, while the leftmost is the asymmetric *temporal priority* measured by rate of occurrence. This implicit form of timing assumes that, if a occurs *more often* than b, then it likely occurs *earlier*, thus satisfying

$$\frac{\mathcal{P}(a)}{\mathcal{P}(a)+\mathcal{P}(b)} > \frac{\mathcal{P}(b)}{\mathcal{P}(a)+\mathcal{P}(b)}\,.$$

An oncotree is the rooted tree whose total weight (i.e. sum of all the weights of the edges) is maximized, and can be reconstructed in $O(|G|^2)$ steps using Edmond's algorithm [22]. By construction, the resulting graph is a proper tree rooted in $\diamond$: each event occurs only once, *confluences* are absent, i.e. any event is caused by at most one other event. The branching trees method has been used to derive progressions for various cancer datasets e.g., [23, 24, 25]), and even though several extensions of the method exist (e.g.[9, 10]), to the best of our knowledge, it is currently the most used method to reconstruct trees and forests.

3 A probabilistic approach to causation

Before introducing the notion of *causation*, upon which our algorithm is based, we briefly review the approach to probabilistic causation. For an extensive discussion on this topic we refer to [26].

In his seminal work [16], Suppes proposed the following notion.

Definition 2 (Probabilistic causation, [16]). *For any two events c and e, occurring respectively at times t_c and t_e, under the mild assumptions that* $0 <$

$\mathcal{P}(c), \mathcal{P}(e) < 1$, the event c causes *the event e if it occurs* before *the effect and the cause* raises the probability *of the effect, i.e.*

$$t_c < t_e \quad \textit{and} \quad \mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})\,. \tag{3}$$

We remark that we consider cross-sectional data where no information about t_c and t_e is available, so we are restricted to consider solely the *probability raising* (PR) property, i.e. $\mathcal{P}(e \mid c) > \mathcal{P}(e \mid \overline{c})$. Now we review some its properties.

Proposition 1 (Dependency)**.** *Whenever the* PR *holds between two events a and b, then the events are* statistically dependent *in a positive sense, i.e.*

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \iff \mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b)\,. \tag{4}$$

This and the next proposition are well-known facts of the PR; their derivation as well as the proofs of all the results we present is in the Supplementary Material. Notice that the opposite implication holds as well: when the events a and b are still dependent but in a negative sense, i.e. $\mathcal{P}(a, b) < \mathcal{P}(a)\mathcal{P}(b)$, the PR does not hold, i.e. $\mathcal{P}(b \mid a) < \mathcal{P}(b \mid \overline{a})$.

We would like to use the asymmetry of the PR to determine whether a pair of events a and b satisfy a causation relation so to place a before b in the progression tree but, unfortunately, the PR satisfies the following property.

Proposition 2 (Mutual PR)**.** *$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \iff \mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$.*

That is, if a raises the probability of observing b, then b raises the probability of observing a too.

Nevertheless, in order to determine causes and effects among the genetic events, we can use the *confidence degree* of probability raising to decide the direction of the causation relationship between pairs of events. In other words, if a raises the probability of b *more* than the other way around, then a is a more likely cause of b than b of a[2]. As mentioned, the PR is not symmetric, and the *direction* of probability raising depends on the relative frequencies of the events. We make this asymmetry precise in the following proposition.

Proposition 3 (Probability raising and temporal priority)**.** *For any two events a and b such that the probability raising $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$ holds, we have*

$$\mathcal{P}(a) > \mathcal{P}(b) \iff \frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})}\,. \tag{5}$$

That is, given that the PR holds between two events, a raises the probability of b *more* than b raises the probability of a, if and only if a is observed more frequently than b. Notice that we use the ratio to assess the PR inequality. The proof of this proposition is technical and can be found in the Supplementary

[2] This is sound as long as each event has *at most* one cause. Otherwise, *frequent late events* with more than one cause, which are rather common in biological progressive phenomena, should be treated differently.

Material. From this result it follows that if we measure the timing of an event by the rate of its occurrence (that is, $\mathcal{P}(a) > \mathcal{P}(b)$ implies that a happens before b), this notion of PR subsumes the same notion of temporal priority induced by a tree (cfr. Section 2). We also remark that this is the temporal priority made explicit in the coefficients of Desper's method (cfr. Section 2.1). Given these results, we define the following notion of causation.

Definition 3. *We state that a* causes *b if a is a probability raiser of b, and it occurs more frequently: $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$ and $\mathcal{P}(a) > \mathcal{P}(b)$.*

Finally, we recall the conditions for the PR to be computable: every mutation a should be observed with probability strictly $0 < \mathcal{P}(a) < 1$. Moreover, we need each pair of mutations (a, b) to be *distinguishable* in terms of PR, that is $\mathcal{P}(a \mid \overline{b}) < 1$ or $\mathcal{P}(b \mid \overline{a}) < 1$ similarly to the above condition. Any non-distinguishable pair of events can be merged as a single composite event. From now on, we will assume these conditions to be verified.

In the next section we will use PR to define a shrinkage estimator and, in turn, to extract progression trees.

4 Extracting progression trees with probability raising and a shrinkage estimator

Our reconstruction method is described in Algorithm 1. The algorithm is very similar in spirit to Desper's algorithm, with the main difference being an alternative weight function based on this shrinkage estimator.

Definition 4 (Shrinkage estimator)**.** *We define the* shrinkage estimator *$m_{a\to b}$ of the confidence in the causation relationship from a to b as*

$$m_{a\to b} = (1-\lambda)\alpha_{a\to b} + \lambda\beta_{a\to b}\,, \tag{6}$$

where $0 \leq \lambda \leq 1$ and

$$\alpha_{a\to b} = \frac{\mathcal{P}(b \mid a) - \mathcal{P}(b \mid \overline{a})}{\mathcal{P}(b \mid a) + \mathcal{P}(b \mid \overline{a})} \qquad \beta_{a\to b} = \frac{\mathcal{P}(a,b) - \mathcal{P}(a)\mathcal{P}(b)}{\mathcal{P}(a,b) + \mathcal{P}(a)\mathcal{P}(b)}\,. \tag{7}$$

This estimator combines a normalized version of the PR, the raw model estimate α, with the correction factor β. The shrinkage aims at improving the performance of the *overall* reconstruction process, not limited to the performance of the estimator itself. In other words, m induces an ordering to the events reflecting our confidence for their causation. However, this framework does not imply any performance bound for the, e.g., mean squared error of m. In Section 5 we show that the shrinkage estimator is an effective way to get such an ordering when data is noisy. In Algorithm 1 we use a pairwise matrix version of the estimator.

We now comment on our reconstruction technique by first explaining the role of the components α and β in m, and then by discussing the use of the correlation filter.

Algorithm 1 Tree-alike reconstruction with shrinkage estimator

```
consider a set of genetic events $G = \{g_1, \ldots, g_n\}$ plus a special event $\diamond$, added to each sample of the dataset;
define a $n \times n$ matrix $M$ where each entry contains the shrinkage estimator
    $m_{i \to j} = (1 - \lambda) \cdot \frac{\mathcal{P}(j \mid i) - \mathcal{P}(j \mid \bar{i})}{\mathcal{P}(j \mid i) + \mathcal{P}(j \mid \bar{i})} + \lambda \cdot \frac{\mathcal{P}(i,j) - \mathcal{P}(i)\mathcal{P}(j)}{\mathcal{P}(i,j) + \mathcal{P}(i)\mathcal{P}(j)}$
  according to the observed probability of the events $i$ and $j$;
[PR causation] define a tree $\mathcal{T} = (G \cup \{\diamond\}, E, \diamond)$ where $(i \to j) \in E$ for $i, j \in G$ if and only if:
    $m_{i \to j} > 0 \quad \text{and} \quad m_{i \to j} > m_{j \to i} \quad \text{and} \quad \forall i' \in G,\ m_{i,j} > m_{i',j}\,.$
[Correlation filter] define $G_j = \{g_i \in G \mid \mathcal{P}(i) > \mathcal{P}(j)\}$, replace edge $(i \to j) \in E$ with edge $(\diamond \to j)$ if, for all $g_w \in G_j$, it holds
    $\frac{1}{1 + \mathcal{P}(j)} > \frac{\mathcal{P}(w)}{\mathcal{P}(w) + \mathcal{P}(j)} \frac{\mathcal{P}(w,j)}{\mathcal{P}(w)\mathcal{P}(j)}\,.$
```

The raw estimator and the correction factor. By considering only the raw estimator α, we would include an edge $(a \to b)$ in the tree consistently in terms of (i) Definition 3 and (ii) if a is the best probability raiser for b [3]. Notice that this formulation of α is a monotonic normalized version of the PR ratio.

Proposition 4 (Monotonic normalization)**.** *For any two events a and b we have*

$$\mathcal{P}(a) > \mathcal{P}(b) \iff \frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \bar{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \bar{b})} \iff \alpha_{a \to b} > \alpha_{b \to a}\,. \tag{8}$$

This raw model estimator satisfies $-1 \leq \alpha_{a \to b}, \alpha_{b \to a} \leq 1$: when it tends to -1 the pair of events appear disjointly (i.e. they show an anti-causation pattern), when it tends to 0 no causation or anti-causation can be inferred and the two events are statistically independent and, when it tends to 1, the causation relationship between the two events is robust. Therefore, α provides a quantification of the degree of confidence for a given PR causation relationship.

However, α does not provide a general criterion to disambiguate among groups of candidate parents of a given node. We show a specific case in which α is not a sufficient estimator. Let us consider, for instance, a causal linear path: $a \to b \to c$. In this case, when evaluating the candidate parents a and b for c we have: $\alpha_{a \to c} = \alpha_{b \to c} = 1$. Accordingly, we can only infer that $t_a < t_c$ and

[3] When $\mathcal{P}(a) = \mathcal{P}(b)$ the events a and b are indistinguishable in terms of temporal priority, thus α is not sufficient to decide their causal relation, if any. This intrinsic ambiguity becomes unlikely when we introduce β even if, in principle, it is still possible.

$t_b < t_c$, i.e. a partial ordering, which does not help to disentangle the relation among a and b with respect to c.

In this case, the β coefficient can be used to determine which of the two candidate parents occurs earlier. In general, such a correction factor provides information on the *temporal distance* between events, in terms of statistical dependency. In other words, the higher the β coefficient, the closer two events are. The shrinkage estimator m then results in a shrinkable combination of the raw PR estimator α and of the β correction factor, which respects the temporal priority induced by α.

Proposition 5 (Coherence in dependency and temporal priority)**.** *The β correction factor is* symmetrical *and subsumes the same notion of dependency of the raw estimator α, that is*

$$\mathcal{P}(a,b) > \mathcal{P}(a)\mathcal{P}(b) \; \Leftrightarrow \alpha_{a\to b} > 0 \Leftrightarrow \beta_{a\to b} > 0 \quad \textit{and} \quad \beta_{a\to b} = \beta_{b\to a}\,. \tag{9}$$

The correlation filter. Following Desper's approach, we add a *root* $\diamond$ with $\mathcal{P}(\diamond) = 1$ so to separate different progression paths, i.e. the different sub-trees rooted in $\diamond$. Algorithm 1 initially builds a unique tree by using m. Then the correlation-alike weight between any node j and $\diamond$ is computed as

$$\frac{\mathcal{P}(\diamond)}{\mathcal{P}(\diamond)+\mathcal{P}(j)}\frac{\mathcal{P}(\diamond,j)}{\mathcal{P}(\diamond)\mathcal{P}(j)} = \frac{1}{1+\mathcal{P}(j)}\,.$$

If this quantity is greater than the weight of j with each upstream connected element i, we substitute the edge $(i \to j)$ with the edge $(\diamond \to j)$. We remark that here we use a correlation filter because it would make no sense to ask whether $\diamond$ was a probability raiser for j, besides the technical fact that α is not defined for events of probability 1 (see Section 3).

Notice that this filter is indeed implying a non-negative threshold for the shrinkage estimator, when a cause is valid.

Theorem 1 (Independent progressions)**.** *Let $G^* = \{a_1, \dots, a_k\} \subset G$ a set of k events which are* candidate causes *of some $b \notin G^*$, i.e. $\mathcal{P}(a_i) > \mathcal{P}(b)$ and $m_{a_i\to b} > 0$ for any a_i. There exist $1 < \gamma < 1/\mathcal{P}(a_i)$ and $\delta > 0$ such that b determines an* independent progression tree *in the reconstructed forest, i.e. the edge $\diamond \to b$ is picked by Algorithm 1, if, for any a_i,*

$$\mathcal{P}(a_i, b) < \gamma\mathcal{P}(a_i)\mathcal{P}(b) + \delta\,. \tag{10}$$

The proof of this theorem can be found in the Supplementary Material. What this theorem suggests is that, in principle, by examining the level of statistical dependency of each pair of events, it would be possible to determine how many trees compose the reconstructed forest. Furthermore, this suggests that Algorithm 1 could be defined by first processing the correlation filter, and then using m to build the independent progression trees in the forest.

To conclude, the algorithm reconstructs well-defined trees in this sense.

Theorem 2 (Algorithm correctness)**.** *Algorithm 1 reconstructs a well defined tree $\mathcal{T}$ without disconnected components, transitive connections and cycles.*

The proof of this Theorem follows immediately from Proposition 3 (see the Supplementary Material).

5 Performance of the algorithm and estimation of the optimal shrinkage coefficient

We made substantial use of *synthetic data* to evaluate the performance of Algorithm 1 as a function of the shrinkage coefficient λ. Many distinct synthetic datasets were created on this purpose, as explained in Section 5.1. The algorithm performance was measured in terms of *Tree Edit Distance* (TED, [27]), i.e. the minimum-cost sequence of node edit operations (relabeling, deletion and insertion) that transforms the reconstructed trees into the ones generating the data.

In Section 5.2 we show the empirical estimation of an optimal λ which ensures the best reconstruction performance, on average. We show that Algorithm 1 has always better performance than branching trees in Section 5.3 and conclude the comparison on real, albeit dated, cancer datasets by showing that our technique predicts highly-confident progression trees unlike those reconstructed by branching trees (Section 5.4).

5.1 Synthetic data generation

Synthetic datasets were generated by sampling from various random trees, constrained to have depth $\log(|G|)$, since wide branches are hard to reconstruct than straight paths.

Unless differently specified, in all the experiments we used 100 distinct random trees (or forests, accordingly to the test to perform) of 20 events each. This seems a fairly reasonable number of events and is in line with the usual size of reconstructed trees, e.g. [28, 29, 30, 31]. The *scalability* of the reconstruction performance was tested against the number of samples by ranging $|G|$ from 50 to 250, with a step of 50, and by replicating 10 independent datasets for each parameters setting (see the caption of the figures for details).

We included a form of *noise* in generating the datasets, in order to account for (i) the realistic presence of *biological noise* (such as the one provided by bystander mutations, genetic heterogeneity, etc.) and (ii) *experimental errors.* A noise parameter $0 \leq \nu < 1$ denotes the probability that any event assumes a random value (with uniform probability), after sampling from the tree-induced distribution[4]. Clearly, this introduces both false negatives and false positives

[4] The assumption that noise is uniformly distributed may appear simplistic since some events may be more robust, or easy to measure, than others. In future works more sophisticated noise distributions could be considered.

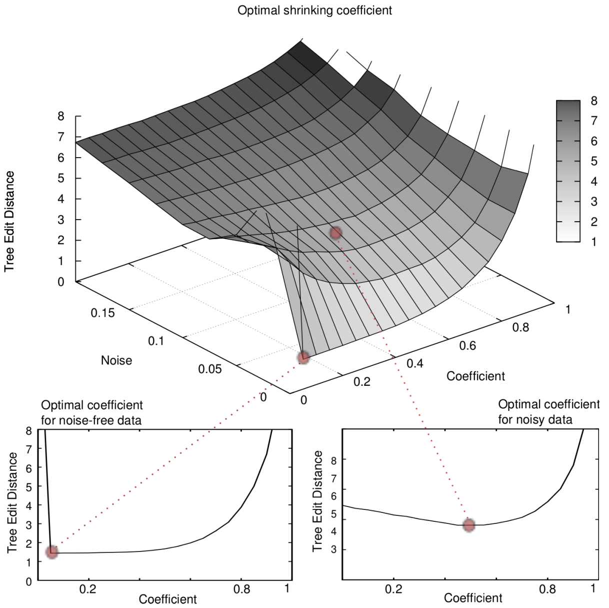

Figure 1: **Optimal shrinkage coefficient for reconstruction performance.** We show here the performance in the reconstruction of trees (TED surface) with 150 samples as a function of the shrinkage coefficient λ. Notice the global optimal performance for $\lambda \to 0$ when $\nu \to 0$ and for $\lambda \approx 1/2$ when $\nu > 0$.

in the datasets. Algorithmically this process generates, on average, $|G|\nu/2$ random entries in each sample (e.g. with $\nu = 0.1$ we have, on average, one error per sample). We wish to assess whether these noisy samples can mislead the reconstruction process, even for low values of ν.

In what follows, we will refer to datasets generated with $\nu > 0$ as noisy synthetic dataset. In the experiments, usually ν is discretized by 0.025, (i.e. 2.5% noise).

5.2 Optimal shrinkage coefficient

Given that our events are dependent on the topology to reconstruct, we cannot determine an optimal value for λ in an analytical way, e.g., by using the standard results in shrinkage statistics [19]. Therefore, we opted for an empirical estimation of its optimal value, both in the case of trees and forests.

In Figure 1, we show the variation of the performance of Algorithm 1 as a function of λ, for datasets with 150 samples generated from tree topologies. The optimal value (i.e. lowest TED) for noise-free datasets (i.e. $\nu = 0$) is obtained for $\lambda \to 0$, whereas for the noisy datasets a series of U-shaped curves suggests

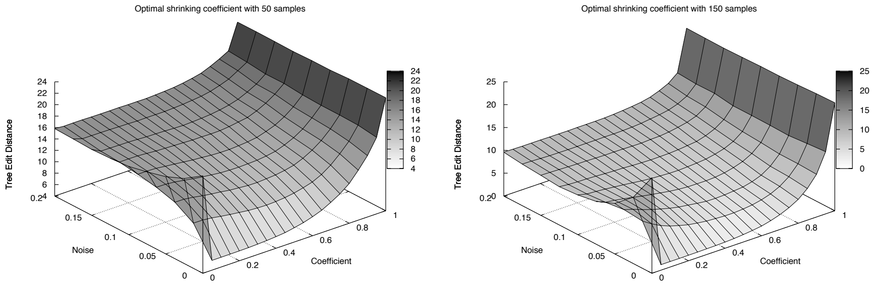

Figure 2: **Optimal λ with datasets of different size.** We show the analogous of Figure 1 with 50 and 250 samples. The estimation of the optimal shrinkage coefficient λ appears to be irrespective of the sample size.

a unique optimum value for $\lambda \to 1/2$, regardless of ν. Identical results are obtained when dealing with forests (not shown here). Besides, further experiments show that the estimation of the optimal λ is not dependent on the number of samples in the datasets (see Figure 2). We here remark that we limited our analysis to datasets with the typical sample size that is characteristic of data currently available.

In other words, if we consider the noise-free case the best performance is obtained by shrinking m to the PR raw estimate α, i.e.

$$m_{a\to b} \overset{\lambda\to 0}{\approx} \alpha_{a\to b} \tag{11}$$

which is obtained by setting λ to a very small value, e.g. 10^{-2}, in order to consider the contribution of the correction factor too. Conversely, when $\nu > 0$, the best performance is obtained by averaging the shrinkage effect, i.e.

$$m_{a\to b} \overset{\lambda=1/2}{=} \frac{\alpha_{a\to b}}{2} + \frac{\beta_{a\to b}}{2}\,. \tag{12}$$

These results suggest that, in general, a unique optimal value for the shrinkage coefficient can be determined.

5.3 Performance of the algorithm compared to *oncotrees*

In Figure 3 we compare the performance of Algorithm 1 with oncotrees, for the case of noise-free synthetic data. In this case, we used the optimal shrinkage coefficient in equation (11): $\lambda \to 0$. In Figure 4 we show an example of reconstructed tree where, for the noise-free case, Algorithm 1 infers the correct tree while oncotrees mislead a causation relation.

In general, one can observe that the TED of Algorithm 1 is, on average, always bounded above by the TED of the oncotrees, both in the case of trees and forests. For trees, with 50 samples the average TED of Algorithm 1 is

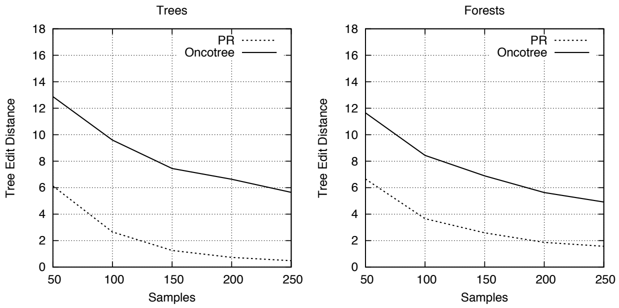

Figure 3: **Comparison on noise-free synthetic data.** Performance of Algorithm 1 (dashed line) and oncotrees (full line) in average TED when data are generated by random trees (left) and forests (right). In this case $\nu = 0$, and the estimator m is shrank by $\lambda \to 0$.

around 7, whereas for Desper's technique is around 13. The performance of both algorithms improves as long as the number of samples is increased: Algorithm 1 has the best performance (i.e. TED ≈ 0) with 250 samples, while oncotrees have TED around 6. When forests are considered, the difference between the performance of the algorithms slightly reduces, but also in this case Algorithm 1 clearly outperforms branching trees.

Notice that the improvement due to the increase in the sample set size seems to reach a *plateau*, and the initial TED for our estimator seems rather close to the plateau value. Thus, this suggests that Algorithm 1 has already good performances with few samples. This is an indeed important result, particularly considering the scarcity of available biological data.

In Figure 5 we extend the comparison to *noisy* datasets. In this case, we used the optimal shrinkage coefficient in equation (12): $\lambda \to 1/2$. The results confirm what observed in the case of noise-free data, as Algorithm 1 outperforms Desper's branching trees up to $\nu = 0.15$, for all the sizes of the sample sets. In the supplementary material we show similar plots for the noise-free case.

5.4 Performance on cancer datasets

The results in the previous sections indicate that our method outperforms oncotrees. We test now our algorithm on a real dataset of cancer patients.

To test our reconstruction approach on a real dataset we applied it to the *ovarian cancer* dataset made available within the oncotree package [8]. The data was collected through the public platform SKY/M-FISH [32], used to allow investigators to share molecular cytogenetic data. The data was obtained by using the *Comparative Genomic Hybridization* technique (CGH) on samples from *papillary serous cystadenocarcinoma* of the ovary. This technique uses fluorescent staining to detect CNV data at the resolution of chromosome arms.

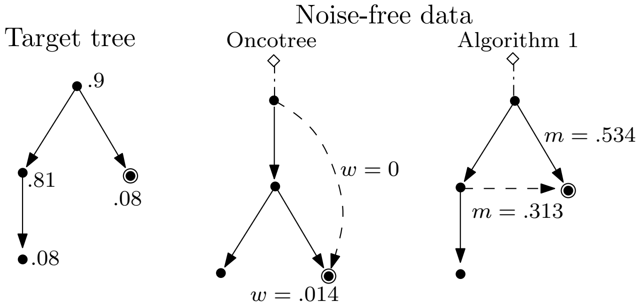

Figure 4: **Example of reconstructed trees** Example of reconstruction from a dataset sampled by the left tree (in there, numbers represent the probability of observing a mutation while generating samples), with $\nu = 0$. The oncotree misleads the correct causal relation for the double-circled mutation (it evaluates $w = 0$ for the real causal edge and $w = 0.014$ for the wrong one). Algorithm 1 infers the correct tree, the numbers represent the values of the estimator m.

Nowadays this kind of analysis can be done at a higher resolution, making this dataset rather outdated. Nevertheless, it can still serve as a perfectly good test-case for our approach. The seven most commonly occurring events are selected from the 87 samples, and the set of events are the following gains and losses on chromosomes arms $G = \{8q+, 3q+, 1q+, 5q-, 4q-, 8p-, Xp-\}$ (e.g., $4q-$ denotes a deletion of the q arm of the 4^{th} chromosome).

In Figure 6 we compare the trees reconstructed by the two approaches. Our technique differs from Desper's by predicting the causal sequence of alterations

$$8q+ \;\rightarrow\; 8p- \;\rightarrow\; Xp- \;.$$

Notice that all the samples in the dataset are generated by the distribution induced by the recovered tree, thus allowing to consider this dataset as noise-free (algorithmically, this allows us to use the estimator for $\lambda \rightarrow 0$).

At this point, we do not have a biological interpretation for this result. However, we do know that common cancer genes reside in these regions, e.g. the tumor suppressor gene PDGFR on $5q$ and the oncogene MYC on $8q$), and loss of heterozygosity on the short arm of chromosome 8 is very common[5]. Recently, evidence has been reported that $8p$ contains many cooperating cancer genes [33].

In order to assign a confidence level to these inferences we applied both parametric and non-parametric *bootstrapping methods* to our results. Essentially, these tests consists of using the reconstructed trees (in the parametric case), or the probability observed in the dataset (in the non-parametric case) to generate new synthetic datasets, and then reconstructs again the progressions (see, e.g., [34] for an overview of these methods). The confidence is given by the number of times the trees in Figure 6 are reconstructed from the gener-

[5] See e.g., `http://www.genome.jp/kegg/`

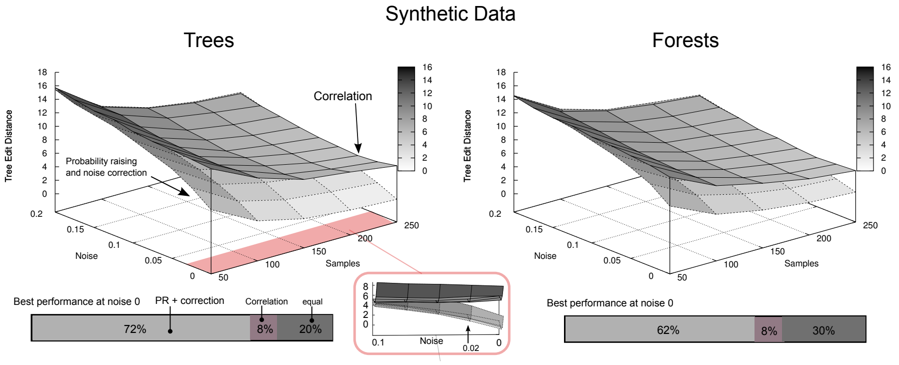

Figure 5: **Reconstruction with noisy synthetic data and $\lambda = 1/2$.** Performance of Algorithm 1 and oncotrees as a function of the number of samples and noise ν. According to Figure 1 the shrinkage coefficient is set to $\lambda = 1/2$. The magnified image shows the convergence to Desper's performance for $\nu \approx 0.1$. The barplot represents the percentage of times the best performance is achieved at $\nu = 0$.

ated data. A similar approach can be used to estimate the confidence of every edge separately. For oncotrees the *exact tree* is obtained 83 times out of 1000 non-parametric resamples, so its estimated confidence is 8.3%. For our algorithm the confidence is 8.6%. In the non-parametric case, the confidence of oncotrees is 17% while ours is much higher: 32%. For the non-parametric case, edges confidence is shown in Table 7. Most notably, our algorithm reconstructs the inference $8q+ \rightarrow 8p-$ with high confidence (confidence 62%, and 26% for $5q- \rightarrow 8p-$), while the confidence of the edge $8q+ \rightarrow 5q-$ is only 39%, almost the same as $8p- \rightarrow 8q+$ (confidence 40%).

Analysis of other datasets. We report the differences between the reconstructed trees also based on datasets of gastrointestinal and oral cancer ([29, 31] respectively). In the case of gastrointestinal stromal cancer, among the 13 CGH events considered in [29] (gains on $5p$, $5q$ and $8q$, losses on $14q$, $1p$, $15q$, $13q$, $21q$, $22q$, $9p$, $9q$, $10q$ and $6q$), the branching trees identify the path progression

$$1p- \rightarrow 15q- \rightarrow 13q- \rightarrow 21q-$$

while Algorithm 1 reconstructs the branch

$$1p- \rightarrow 15q- \qquad\qquad 1p- \rightarrow 13q- \rightarrow 21q- \;.$$

In the case of oral cancer, among the 12 CGH events considered in [31] (gains on $8q$, $9q$, $11q$, $20q$, $17p$, $7p$, $5p$, $20p$ and $18p$, losses on $3p$, $8p$ and $18q$), the

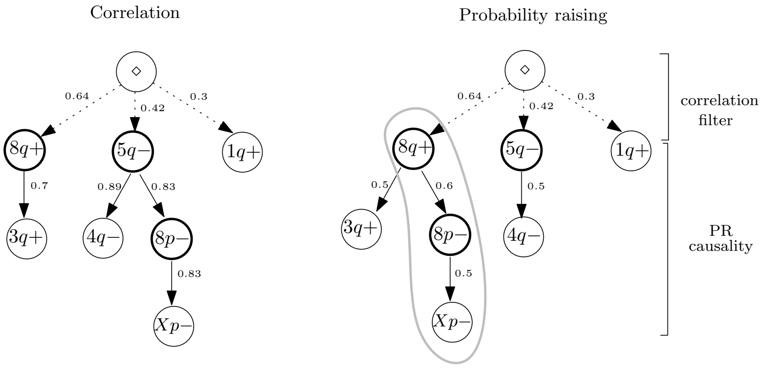

Figure 6: **Oncotree reconstruction of ovarian cancer progression.** Trees reconstructed by branching trees and with Algorithm 1 (for $\lambda \to 0$). The set of CGH events considered are gains on $8q$, $3q$ and $1q$ and losses on $5q$, $4q$, $8p$ and Xp. Events on chromosomes arms containing the key genes for ovarian cancer are in bolded circles. In the left tree all edge weights are the observed probabilities of events. In the right the full edges are the causation inferred with the PR and the weights represent the normalizes coefficients of Algorithm 1. Weights on dashed lines are as in the left tree.

reconstructed trees differ since oncotrees identifies the path

$$8q+ \to 20q+ \to 20p+$$

while our algorithm reconstructs the path

$$3p- \to 7p+ \to 20q+ \to 20p+ \ .$$

These examples show that Algorithm 1 provides important differences in the reconstruction compared to the branching trees.

6 Discussion and future works

In this work we have introduced a novel theoretical framework for the reconstruction of the causal topologies underlying cumulative progressive phenomena, based on the *probability raising* notion of causation. Besides such a probabilistic notion, we also introduced the use of a *shrinkage estimator* to efficiently unravel ambiguous causal relations, often present when data are noisy. As a first step towards the definition of our new framework, we have here presented an effective novel technique for the reconstruction of tree or, more in general, forest models of progression which combines probabilistic causation and shrinkage estimation.

We compared this technique with a standard approach based on correlation, and show that our method outperforms the state-of-the-art on synthetic data, Oncotrees (overall confidence 8.3%)

→	$8q+$	$3q+$	$5q-$	$4q-$	$8p-$	$1q+$	$Xp-$
$\diamond$	**.99**	.06	**.51**	.22	.004	**.8**	.06
$8q+$	0	**.092**	.08	0.16	0.4	.02	.007
$3q+$	.002	0	.04	0	0	.09	.04
$5q-$	.001	.002	0	**.52**	**.39**	.009	.16
$4q-$	0	0	.27	0	.14	.05	.11
$8p-$	0	0	.07	.08	0	.004	**.59**
$1q+$	0	0	0	.004	0	0	0
$Xp-$	0	0	.003	.003	.04	.01	0

Algorithm 1 (overall confidence 8.6%)

→	$8q+$	$3q+$	$5q-$	$4q-$	$8p-$	$1q+$	$Xp-$
$\diamond$	**.99**	.06	**.51**	.22	.004	**.8**	.06
$8q+$	0	**.92**	.06	.16	**.62**	.01	.008
$3q+$	.002	0	.03	.002	0	.09	.04
$5q-$	.001	.002	0	**.5**	.26	.009	.17
$4q-$	0	0	.29	0	.09	.05	.12
$8p-$	0	0	.07	.08	0	.004	**.59**
$1q+$	0	0	0	.004	0	0	0
$Xp-$	0	.001	.003	.004	.01	.01	0

Figure 7: **Estimated confidence for ovarian progression.** Frequency of edge occurrences in the non-parametric bootstrap test, for the trees shown in Figure 6. Colors represent confidence: light gray is $< .4\%$, mid gray is $.4\% \div .8\%$ and dark gray is $> .8\%$. Bold entries are the edges recovered by the algorithms.

also exhibiting a noteworthy efficiency with relatively small datasets. Furthermore, we tested our technique on low-resolution CNV cancer data[6]. This analysis suggested that our approach can infer, with high confidence, novel causal relationships which would remain unpredictable by correlation-based techniques. Even if the cancer data that we used is coarse-grained and does not account for, e.g. small-scale mutations and epigenetic information, we remark that this technique can be applied to data at any resolution. In fact, it requires an input set of samples containing some alterations (e.g. cancer mutations), supposed to be involved in a certain causal process. The results of our technique can be used not only to describe the *progression* of the process, but also to *classify*. In the case of cancer, for instance, this genome-level classifier could be used to group patients and to set up a genome-specific *therapy design*.

Several future research directions are possible. Firstly, more complex models of progression, e.g. directed acyclic graphs, could be inferred with probability raising and compared to the standard approaches of [12, 13, 36], as we explained in the introduction. These models, rather than trees, could explain the common phenomenon of *preferential progression paths* in the target process via, e.g., *confluence* among events. In the case of cancer, for instance, these models

[6] This data is rather dated at this point and we are working to apply our work to more recent *Next Generation Sequencing* (NGS) data from which we are extracting CNVs using publicly available tools in the Galaxy pipeline [35].

would be certainly more suitable than trees to describe the accumulation of mutations.

Secondly, the shrinkage estimator itself could be improved by introducing, for instance, different correction factors. In addiction, an analytical formulation of the optimal shrinkage coefficient could be investigated by starting from the hypotheses which apply to our problem setting, along the lines of [20].

Besides, advanced statistical techniques such as *bootstrapping* [34] could be used to account for more sophisticated models of noise within data, so to decipher complex causal dependencies. Finally, a further development of the framework could involve the introduction of *timed data*, so to extend our techniques to settings where a temporal information on the samples is available.

Acknowledgments. This research was funded by the NSF grants CCF-0836649 and CCF-0926166 and by Regione Lombardia (Italy) under the research projects RetroNet through the ASTIL [12-4-5148000-40]; U.A 053 and NEDD Project [ID14546A Rif SAL-7] Fondo Accordi Istituzionali 2009.

References

[1] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: The next generation,” *Cell*, vol. 144, pp. 646–674, 2011.

[2] J. Luo, N. L. Solimini, and S. J. Elledge, “Principles of cancer therapy: Oncogene and non-oncogene addiction,” *Cell*, vol. 136, pp. 823–837, Mar. 2009.

[3] B. Vogelstein and K. Kinzler, “Cancer genes and the pathways they control,” *Nature medicine*, vol. 10, no. 8, pp. 789–799, 2004.

[4] S. A. Frank, *Dynamics of Cancer*. Princeton University Press, 2007.

[5] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, and P. Glenn, “Integrated genomic analyses of ovarian carcinoma,” 2011.

[6] M. Imielinski *et al.*, “Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing,” *Cell*, vol. 150, no. 6, pp. 1107–1120, 2012.

[7] B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, A. M. Smits, and J. L. Bos, “Genetic alterations during colorectal-tumor development,” *New England Journal of Medicine*, vol. 319, no. 9, pp. 525–532, 1988.

[8] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, “Inferring tree models for oncogenesis from comparative genome hybridization data,” *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, 1999.

[9] R. Desper, F. Jiang, O. Kallioniemi, H. Moch, C. Papadimitriou, and A. Schäffer, "Distance-based reconstruction of tree models for oncogenesis," *Journal of Computational Biology*, vol. 7, no. 6, pp. 789–803, 2000.

[10] N. Beerenwinkel, J. Rahnenführer, M. Däumer, D. Hoffmann, R. Kaiser, J. Selbig, and T. Lengauer, "Learning multiple evolutionary pathways from cross-sectional data," *Journal of Computational Biology*, vol. 12, no. 6, pp. 584–598, 2005.

[11] M. Hjelm, "New probabilistic network models and algorithms for oncogenesis," *Journal of Computational Biology*, vol. 13, pp. 853–865, 2006.

[12] N. Beerenwinkel, N. Eriksson, and B. Sturmfels, "Conjunctive bayesian networks," *Bernoulli*, pp. 893–909, 2007.

[13] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive bayesian networks," *Bioinformatics*, vol. 25, no. 21, pp. 2809–2815, 2009.

[14] A. Gupta and Z. Bar-Joseph, "Extracting dynamics from static cancer expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, pp. 172–182, June 2008.

[15] N. Ramakrishnan, S. Tadepalli, L. T. Watson, R. F. Helm, M. Antoniotti, and B. Mishra, "Reverse engineering dynamic temporal models of biological processes and their relationships.," *PNAS*, vol. 107, pp. 12511–12516, July 2010.

[16] P. Suppes, *A probabilistic theory of causality.* North Holland Publishing Company, 1970.

[17] I. Ionita, R. Daruwala, and B. Mishra, "Mapping Tumor-Suppressor genes with multipoint statistics from Copy-Number–Variation data," *American Journal of Human Genetics*, vol. 79, pp. 13–22, July 2006. PMID: 16773561 PMCID: 1474131.

[18] S. Kleinberg, *Causality, Probability, and Time.* Cambridge University Press, 2012.

[19] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press, 2013.

[20] B. Efron and C. Morris, "Stein's estimation rule and its competitors–an empirical bayes approach," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 117–130, 1973.

[21] H. Reichenbach, *The Direction of Time.* University of California Press, 1956.

[22] J. Edmonds, "Optimum branchings," *Journal of Research of the National Bureau of Standards B*, vol. 71, pp. 233–240, 1967.

[23] T. Kainu *et al.*, "Somatic deletions in hereditary breast cancers implicate 13q21 as a putative novel breast cancer susceptibility locus," *Proceedings of the National Academy of Sciences*, vol. 97, no. 17, pp. 9603–9608, 2000.

[24] Q. Huang, G. Yu, S. McCormick, J. Mo, B. Datta, M. Mahimkar, P. Lazarus, A. A. Schäffer, R. Desper, and S. Schantz, "Genetic differences detected by comparative genomic hybridization in head and neck squamous cell carcinomas from different tumor sites: construction of oncogenetic trees for tumor progression," *Genes, Chromosomes and Cancer*, vol. 34, no. 2, pp. 224–233, 2002.

[25] M. Radmacher, R. Simon, R. Desper, R. Taetle, A. Schäffer, and M. Nelson, "Graph models of oncogenesis with an application to melanoma," *Journal of theoretical biology*, vol. 212, no. 4, pp. 535–548, 2001.

[26] C. Hitchcock, "Probabilistic causation," in *The Stanford Encyclopedia of Philosophy* (E. Zalta, ed.), winter 2012 ed., 2012.

[27] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM journal on computing*, vol. 18, no. 6, pp. 1245–1262, 1989.

[28] E. Samuelson, S. Karlsson, K. Partheen, S. Nilsson, C. Szpirer, and A. Behboudi, "Bac cgh-array identified specific small-scale genomic imbalances in diploid dmba-induced rat mammary tumors," *BMC cancer*, vol. 12, no. 1, p. 352, 2012.

[29] B. Gunawan *et al.*, "An oncogenetic tree model in gastrointestinal stromal tumours (gists) identifies different pathways of cytogenetic evolution with prognostic implications," *The Journal of pathology*, vol. 211, no. 4, pp. 463–470, 2007.

[30] T. Longerich, M. Mueller, K. Breuhahn, P. Schirmacher, A. Benner, and C. Heiss, "Oncogenetic tree modeling of human hepatocarcinogenesis," *International Journal of Cancer*, vol. 130, no. 3, pp. 575–583, 2012.

[31] S. Pathare, A. Schäffer, N. Beerenwinkel, and M. Mahimkar, "Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression," *International journal of cancer*, vol. 124, no. 12, pp. 2864–2871, 2009.

[32] T. Knutsen, V. Gobu, R. Knaus, H. Padilla-Nash, M. Augustus, R. Strausberg, I. Kirsch, K. Sirotkin, and T. Ried, "The interactive online sky/m-fish & cgh database and the entrez cancer chromosomes search database: Linkage of chromosomal aberrations with the genome sequence," *Genes, Chromosomes and Cancer*, vol. 44, no. 1, pp. 52–64, 2005.

[33] W. Xue *et al.*, "A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions," *Proceedings of the National Academy of Sciences*, vol. 109, no. 21, pp. 8212–8217, 2012.

[34] B. Efron, *The jackknife, the bootstrap and other resampling plans*, vol. 38. SIAM, 1982.

[35] "The galaxy project." `http://galaxyproject.org`.

[36] M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein, and N. Beerenwinkel, "The temporal order of genetic and pathway alterations in tumorigenesis," *PloS one*, vol. 6, no. 11, p. e27136, 2011.

A Supplementary Materials

A.1 Proofs

Here the proofs of all the propositions and theorems follow.

Proof of Proposition 1 (Dependency).

Proof. For $\Rightarrow$ write $\mathcal{P}(\overline{a}, b) = \mathcal{P}(b) - \mathcal{P}(a, b)$, then write the PR as

$$\frac{\mathcal{P}(a,b)}{\mathcal{P}(a)} > \frac{\mathcal{P}(b) - \mathcal{P}(a,b)}{1 - \mathcal{P}(a)}$$

and, since $0 < \mathcal{P}(a) < 1$, the proposition follows by simple algebraic arrangements of $\mathcal{P}(a, b) \cdot [1 - \mathcal{P}(a)] > \mathcal{P}(a)\mathcal{P}(b) - \mathcal{P}(a, b) \cdot \mathcal{P}(a)$. The derivations are analogous but in reverse order for the implication $\Leftarrow$. □

Proof of Proposition 2 (Mutual probability raising).

Proof. The proof follows by Property 1 and the subsequent implication:

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \Leftrightarrow \mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b) \Leftrightarrow \mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b}) \,.$$

□

Proof of Proposition 3 (Probability raising and temporal priority).

Proof. We first prove the left-to-right direction $\Rightarrow$. Let $x = \mathcal{P}(\overline{a}, b)$, $y = \mathcal{P}(a, b)$ and $z = \mathcal{P}(a, \overline{b})$. We have two assumptions we will use later on:

1. $\mathcal{P}(a) > \mathcal{P}(b)$ which implies $\mathcal{P}(\overline{a}, b) < \mathcal{P}(a, \overline{b})$, i.e. $x < z$.
2. $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$ which, when $0 < x + y < 1$, implies by simple algebraic rearrangements the inequality

$$y[1 - x - y - z] > xz \,. \tag{13}$$

We proceed by rewriting $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) > \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$ as

$$\frac{\mathcal{P}(a,b)\mathcal{P}(\overline{a})}{\mathcal{P}(\overline{a},b)\mathcal{P}(a)} > \frac{\mathcal{P}(a,b)\mathcal{P}(\overline{b})}{\mathcal{P}(a,\overline{b})\mathcal{P}(b)}$$

which means that

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})} \iff \frac{\mathcal{P}(\overline{a})}{\mathcal{P}(\overline{a},b)\mathcal{P}(a)} > \frac{\mathcal{P}(\overline{b})}{\mathcal{P}(a,\overline{b})\mathcal{P}(b)} \tag{14}$$

We can rewrite the right side of (14) by using x, y, z where $\mathcal{P}(a) = \mathcal{P}(a, b) + \mathcal{P}(a, \overline{b}) = y + z$ and $\mathcal{P}(b) = \mathcal{P}(a, b) + \mathcal{P}(\overline{a}, b) = x + y$, and then do some algebraic manipulations. We have

$$\frac{1-y-z}{x(y+z)} > \frac{1-x-y}{z(x+y)} \iff yz - y^2z - xz^2 - yz^2 > xy - x^2y - x^2z - xy^2 \tag{15}$$

when $x(y+z) \neq 0$ and $z(x+y) \neq 0$. To check that the right side of (15) holds we show that

$$(xy - x^2y - x^2z - xy^2) - (yz - y^2z - xz^2 - yz^2) < 0\,.$$

First, we rearrange it to $(x-z)[y - y^2 - xz - y(x+z)] < 0$ so to show that

$$(x-z)[y(1-y-x-z) - zx] < 0 \tag{16}$$

is always negative. By observing that, by assumption 1 we have $z > x$ and thus $(x-z) < 0$, and, by equation (13) we have $y(1-y-x-z) - zx > 0$, we derive

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})}$$

which concludes the ⇒ direction.

The other direction ⇐ follows immediately by contraposition: assume that $\mathcal{P}(a \mid b) > \mathcal{P}(a \mid \overline{b})$, $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) > \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$ and $\mathcal{P}(b) \leq \mathcal{P}(a)$. We distinguish two cases:

1. $\mathcal{P}(b) = \mathcal{P}(a)$, then $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) = \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$.
2. $\mathcal{P}(b) < \mathcal{P}(a)$, then by symmetry $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$, and by the ⇒ direction of the proposition it follows that $\mathcal{P}(b \mid a)/\mathcal{P}(b \mid \overline{a}) < \mathcal{P}(a \mid b)/\mathcal{P}(a \mid \overline{b})$.

In both cases we have a contradiction. This finishes the proof. □

Proof of Proposition 4 (Monotonic normalization).

Proof. We prove the left-to-right direction ⇒, the other direction follows by a similar argument. Let us assume

$$\frac{\mathcal{P}(b \mid a)}{\mathcal{P}(b \mid \overline{a})} > \frac{\mathcal{P}(a \mid b)}{\mathcal{P}(a \mid \overline{b})} \tag{17}$$

then $\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) > \mathcal{P}(a \mid b)\mathcal{P}(b \mid \overline{a})$. Now, to show the righthand side of the implication, we will show that

$$\Big[\mathcal{P}(b \mid a) - \mathcal{P}(b \mid \overline{a})\Big]\Big[\mathcal{P}(a \mid b) + \mathcal{P}(a \mid \overline{b})\Big] > \Big[\mathcal{P}(b \mid a) + \mathcal{P}(b \mid \overline{a})\Big]\Big[\mathcal{P}(a \mid b) - \mathcal{P}(a \mid \overline{b})\Big]$$

which reduces to show

$$\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) - \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) > \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) - \mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b})\,.$$

By (17), two equivalent inequalities hold

$$\begin{gathered}\mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) - \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) > 0\\ \mathcal{P}(b \mid \overline{a})\mathcal{P}(a \mid b) - \mathcal{P}(b \mid a)\mathcal{P}(a \mid \overline{b}) < 0\end{gathered}$$

and hence the implication holds. □

Proof of Proposition 5 (Coherence in dependency and temporal priority).

Proof. We make two assumptions:

1. $\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a})$ which implies $\alpha_{a\to b} > 0$.

2. $\mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b)$ which implies $\beta_{a\to b} > 0$.

The proof regarding dependency follows by Property 1 and its implication:

$$\mathcal{P}(b \mid a) > \mathcal{P}(b \mid \overline{a}) \Leftrightarrow \mathcal{P}(a, b) > \mathcal{P}(a)\mathcal{P}(b) \Leftrightarrow \alpha_{a\to b} > 0 \Leftrightarrow \beta_{a\to b} > 0.$$

Moreover, being β symmetric by definition, the proof regarding temporal priority follows directly by Property 2 □

Proof of Theorem 1 (Independent progressions).

Proof. For each candidate cause, $\alpha_{a_i\to b} > 0$ by definition. According to the correlation filter, the connection $\diamond \to b$ is picked in favor of $a_{k^*} \to b$, where $k^* = \max_i\{m_{a_i\to b} \mid a_i \in G^*\}$ if, for any a_i, it holds

$$\frac{1}{1 + \mathcal{P}(b)} > \frac{\mathcal{P}(a_i)}{\mathcal{P}(a_i) + \mathcal{P}(b)} \frac{\mathcal{P}(a_i, b)}{\mathcal{P}(a_i)\mathcal{P}(b)}$$

which, with some algebraic manipulations, rewrites as

$$\mathcal{P}(a_i, b) < \frac{\mathcal{P}(a_i) + \mathcal{P}(b)}{\mathcal{P}(a_i)[1 + \mathcal{P}(b)]}\mathcal{P}(a_i)\mathcal{P}(b)\,.$$

In other words, it is required that at least one of the candidate causes a_i has a minimum level of positive statistical dependency with b. Let us define

$$\gamma = \frac{\mathcal{P}(a_i) + \mathcal{P}(b)}{\mathcal{P}(a_i)[1 + \mathcal{P}(b)]}$$

so to have $\mathcal{P}(a_i, b) = \gamma[\mathcal{P}(a_i)\mathcal{P}(b)] + \delta$, with $\delta \geq 0$. We remark that $\alpha_{a_i\to b} > 0$ implies $\gamma > 1$ by Proposition 1 and that this condition is the same implied by the correlation filter. Also, since $\mathcal{P}(a_i, b)$ is bounded above by $\mathcal{P}(b)$, by substituting in γ we find the maximum γ in the limit of $\delta = 0$ to be $1/\mathcal{P}(a_i)$. Therefore, the correlation filter is implying a non-negative threshold to the shrinkage estimator. □

Proof of Theorem 2 (Algorithm correctness).

Proof. It is clear that Algorithm 1 does not create disconnected components since, to each node in G, a unique parent is attached (either from G or $\diamond$). For the same reason, no transitive connections can appear.

The absence of cycles results from Properties 3, 4 and 5. Indeed, suppose for contradiction that there is a cycle $(a_1, a_2), (a_2, a_3), \ldots, (a_n, a_1)$ in E, then by the three propositions we have

$$\mathcal{P}(a_1) > \mathcal{P}(a_2) > \ldots > \mathcal{P}(a_n) > \mathcal{P}(a_1)$$

which is a contradiction. □

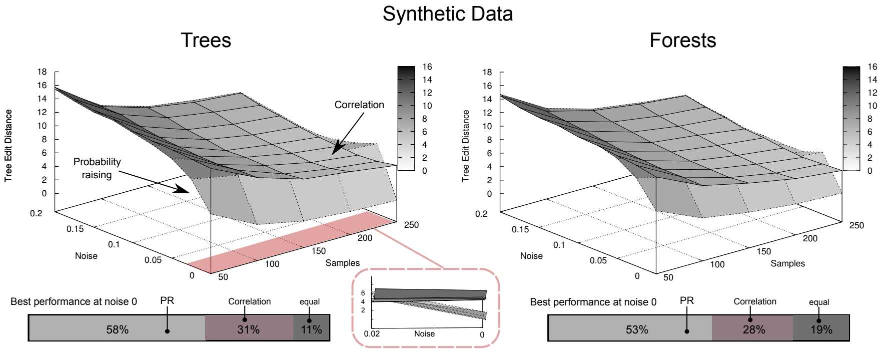

Figure 8: **Reconstruction with noisy synthetic data and $\lambda \to 0$.** The settings of the experiments are the same as those used in Figure 5, but in this case the estimator is shrank to α by $\lambda \to 0$, i.e. $\lambda = 0.01$. In the magnified image one can sees that the performance of Algorithm 1 converges to Desper's one already for $\nu \approx 0.01$, hence largely faster than in the case of $\lambda \approx 1/2$ (Fig. 5).

A.2 Further results

We show here the results of the experiments discussed but not presented in the main text.

Reconstruction of noisy synthetic data with $\lambda \to 0$. Although we know that $\lambda \to 0$ is not the optimal value of the shrinkage coefficient for noisy data, we show in Figure 8 the analogue of Figure 5 when the estimator is shrank to α by $\lambda \to 0$, i.e. $\lambda = 0.01$. When compared to Figure 5 it is clear that a best performance of Algorithm 1 is obtained with $\lambda \approx 1/2$, as suggested by Figure 1.