Causal feature selection

Isabelle Guyon, Clopinet, California
Constantin Aliferis, Vanderbilt University, Tennessee
André Elisseeff, IBM Zürich, Switzerland

March 2, 2007

Abstract

This report reviews techniques for learning causal relationships from data, in application to the problem of feature selection. Most feature selection methods do not attempt to uncover causal relationships between feature and target and focus instead on making best predictions. We examine situations in which the knowledge of causal relationships benefits feature selection. Such benefits may include: explaining relevance in terms of causal mechanisms, distinguishing between actual features and experimental artifacts, predicting the consequences of actions performed by external agents, and making predictions in non-stationary environments. Conversely, we highlight the benefits that causal discovery may draw from recent developments in feature selection theory and algorithms.

1 Introduction

The present report makes an argument in favor of understanding and utilizing causal perspectives when designing, characterizing, and using new or existing feature selection algorithms:

- From an algorithm design perspective, bringing causality into play when designing feature selection methods may provide enhanced mechanistic interpretation, robustness to violations of the i.i.d. assumption, and may lead to increased parsimony of selected feature sets.

- From the perspective of characterization of existing and new methods, causal analysis can shed light on whether (and under what conditions) feature selection methods may return superfluous or artifactual variables, whether necessary variables are missed, and when the selected features will not only be predictive but also causally informative.

Determining and exploiting causal relationships is central in human reasoning and decision-making. The goal of determining causal relationships is to predict the consequences of given **actions** or **manipulations**. This is fundamentally different from making predictions from **observations**. Observations imply no **experimentation**, no **interventions** on the system under study, whereas actions introduce a disruption in the natural functioning of the system.

Confusing observational and interventional predictive tasks yields classical paradoxes [28]. Consider for instance that there seems to be a correlation between being in bed and dying.

Should we conclude that we should better not spend time in bed to reduce our risk of dying? No, because arbitrarily *forcing* people to spend time in bed does not normally increase death rate. A plausible causal model is that *sickness* causes both an increase in *time spent in bed* and in *death rate*. This example illustrates that a correlated feature (*time spent in bed*) may be predictive of an outcome (*death rate*), if the system is stationary (no change in the distribution of all the variables) and no interventions are made, yet it does not allow us to make predictions if an intervention is made (*e.g.* forcing people to spend more or less time in bed regardless of their disease condition). This example outlines the fundamental distinction between **correlation** and **causation**.

Policy making in health care, economics, or ecology are examples of interventions, of which it is desirable to know the consequences ahead of time (see our application section, Section 8). The goal of causal modeling is to provide coarse descriptions of **mechanisms**, at a level sufficient to predict the result of interventions. The main concepts are reviewed in Section 5. The most established way of deriving causal models is to carry out *randomized controlled experiments* to test hypothetical causal relationships. Yet such experiments are often costly, unethical or infeasible. This prompted a lot of recent research in **learning causal models from observational data** [11, 28, 32], which we briefly review in Section 7.

Most feature selection algorithms emanating from the machine learning literature (see *e.g.* [22, 14], which we briefly review for comparison in Section 3), do not seek to model mechanisms: they do not attempt to uncover cause-effect relationships between feature and target. This is justified because uncovering mechanisms is unnecessary for making good predictions in a purely observational setting. In our *death rate* prediction example classical feature selection algorithms may include without distinction: features that cause the target (like *sickness*), features that are consequences of a common cause (like *time spent in bed*, which is a consequence of *sickness*, not of *death rate*), or features that are consequences of the target (like *burial rate*). But, while acting on a cause (like *disease*) can influence the outcome (*death rate*), acting on consequences (*burial rate*) or consequences of common causes (*time spent in bed*) cannot. Thus non causality-aware feature selection algorithms do not lend themselves to making predictions of the result of actions or interventions.

Performing an intervention on a system is one way of disturbing the "natural" distribution of variables, which violates the assumption of identically and independently distributed data (i.i.d. data), classically made in machine learning. Under this assumption, training and test data are drawn from the same distribution. In Section 2.4, we explain why the knowledge of causal relationships allows us to build models, which are more predictive under particular distribution changes.

Feature selection algorithms, which ignore the way the data were produced, have an additional weakness. They select features for their effectiveness to predict the target, regardless of whether such predictive power is characteristic of the system under study or the result of **experimental artifacts**. If one considers that features being analyzed in real data are always obtained from measurements, they characterize both the system under study and the measuring instrument. To build robustness against changes in measuring conditions, it is important to separate the effect of measurement error from those of the processus of interest, as outlined in Section 6.2.

On the strong side of feature selection algorithm developed recently [22, 14], relevant features may be spotted among hundreds of thousands of distracters, with less than a hundred examples, in some problem domains. Research in this field has effectively addressed both computational

and statistical problems that related to uncovering significant dependencies in such adverse conditions. In contrast, causal models [11, 28, 32] usually deal with just a few variables and a quasi-perfect knowledge of the variable distribution, which implies an abundance of training examples. Thus there are opportunities for cross-fertilization of the two fields: causal discovery can benefit from feature selection to cut down dimensionality for computational or statistical reasons (albeit with the risk of removing causally relevant features); feature selection can benefit from causal discovery by getting closer to the underlying mechanisms and reveal a more refined notion of relevance (albeit at computational price). This report aims to provide material to stimulate research in both directions.

2 Goals of feature selection and causal discovery

In this section, we compare and contrast the goals of feature selection and causal discovery. We show in which cases these goals converge.

2.1 Goals of feature selection

In the supervised feature selection setting, a set of candidate predictor random variables $\mathbf{X} = [X_1, X_2, ...X_N]$ and a target random variable Y are given. Training samples drawn from $P(\mathbf{X}, Y)$ are provided. The problem of feature selection can be stated in several ways. But rather generally, the goal of feature selection is to **find a subset of X as small as possible, while simultaneously optimizing a primary objective** (typically prediction performance on future data).

The benefits of feature selection may include:

- **Prediction**: Improving prediction performance (or another objective) by eliminating useless noise features and/or alleviating the problem of overfitting via a dimensionality reduction.
- **Effectiveness**: Reducing memory consumption, time of learning, time of processing. Facilitating future data acquisition by reducing the amount and cost of data to be collected.
- **Data understanding**: Identifying factors relevant to the target.

The wide variety of existing feature selection algorithm and the lack of consensus as to which one works best is easily understood. First, there is a wide variety of types of variables, data distributions, learning machines, and objective functions. Second, depending on the application, different emphases are put on the various benefits, privileging prediction performance, feature set compactness, or data understanding.

2.2 Goals of causal discovery

In causal discovery, a set of random variables $\mathbf{X} = [X_1, X_2, ...X_N]$ is given and a joint distribution $P(\mathbf{X})$. There may or may not be a particular variable singled out as a target: every variable may be considered both a predictor and a target. The goal of causal discovery is to **uncover causal relationships between the variables**, with one of several purposes [28]:

- **Prediction**: Predicting future data without disturbing the system (same goal as feature selection).
- **Manipulation**: Predicting the consequence of given actions, which result from manipulations of the system by an external agent.
- **Counterfactual** prediction: Given that a certain outcome was observed, predicting what would have happened if a different action had been taken.
- **Data understanding**: Obtain a model of what the underlying mechanisms of data production are (scientific discovery, reverse engineering). Assign probabilities to causal explanations (causal attribution).

2.3 Comparing and contrasting the goals

An obvious difference between feature selection and causal discovery is the emphasis put on a given target variable Y. However, we can easily focus causal discovery on a particular target variable called Y, or reciprocally have all X_i take turn as the target. In what follows, we will single out a variable Y to make comparisons easier. Feature selection and causal discovery converge on two central goals:

- Making good **predictions**. In causal discovery, a distinction is made between "prediction" and "manipulation". The former assumes that future data are drawn from the same distribution $P(\mathbf{X}, Y)$ as training data, while the latter assumes that an external agent will disturb the system by performing deliberate manipulations, violating the i.i.d. assumption. In machine learning, the i.i.d. assumption is typically assumed, even though a number of non-causal models have been proposed to accommodate particular situations of learning under changes in distribution (see for instance the recent NIPS workshop [6]).
- **Data understanding**. Both causal discovery and feature selection try to uncover "relevant" factors or variables. In Section 6, we develop the idea that feature "relevance" is a notion, which is closely related to that of causality, and that notions of relevance defined in the feature selection community can be refined in light of causal relationships.

Feature selection and causal discovery diverge on other goals:

- Causal discovery is not primarily concerned with **effectiveness**. However, since in practice many causal discovery algorithms can address only problems with few variables, feature selection methods have a role to play by screening the most relevant variables.
- Feature selection is not concerned with **counterfactuals** and we will not address this problem in this report. The interested reader can refer to *e.g.* [28] for an introduction.

In conclusion, causal discovery is more refined than feature selection in that it attempts to uncover mechanisms rather than statistical dependencies. The benefits of incorporating causal discovery in feature selection include understanding more finely the data structure and making prediction possible under manipulations and some distribution changes.

2.4 Two illustrative examples

Before going into any formal details, we wish to give two simple examples to illustrate possible uses of causality to improve predictive modeling. The first example illustrates how a simple distribution shift can be compensated for, if it is known whether the features are causal or consequential with respect to the target. The second example illustrates the process of manipulations and the benefit drawn from inferring causality in the feature selection process.

Example 1: Distribution shift

This first example illustrates that understanding the data generating process can help build better predictors, particularly when the classical i.i.d. assumption is violated.

In machine learning, it is often assumed that data are drawn randomly identically and independently according to a distribution $P(\mathbf{X}, Y)$ (the i.i.d. assumption). Without precluding of any mechanism of data generation, the theorem of Bayes $P(\mathbf{X}, Y) = P(\mathbf{X})P(Y|\mathbf{X}) = P(\mathbf{X}|Y)P(Y)$ gives us two equivalent alternatives to model $P(\mathbf{X}, Y)$. However, those two alternatives are no longer equivalent if we interpret them as generative process models in which the conditioning variables are causes and conditioned variables effects. In one case, denoted as $\mathbf{X} \rightarrow Y$, a sample $\mathbf{x}_i$ is drawn according to $P(\mathbf{X})$ and then $P(Y|\mathbf{X})$ is applied to produce the corresponding y_i from $\mathbf{x}_i$, perhaps by applying a deterministic mechanism or function f and adding random noise ϵ: $y_i = f(\mathbf{x}_i) + \epsilon$. In the other case, denoted as $Y \rightarrow \mathbf{X}$, it is the other way around: a value of y is drawn at random, and then $P(\mathbf{X}|Y)$ is applied to produce a sample $\mathbf{x}_i$.

In the study of causal processes, it is usually assumed that the underlying mechanisms do not change while we are studying them, *e.g.* in the case $\mathbf{X} \rightarrow Y$, $P(Y|\mathbf{X})$ does not change and in the case $Y \rightarrow \mathbf{X}$, $P(\mathbf{X}|Y)$ does not change. This is the mathematical translation of "the same causes produce the same effects". This is a less restrictive assumption than the i.i.d. assumption, which imposes that $P(\mathbf{X}, Y)$ remain unchanged. Thus a causal model can accommodate changes in $P(\mathbf{X})$ in the case of $\mathbf{X} \rightarrow Y$ or $P(Y)$ in the case $Y \rightarrow \mathbf{X}$.

We illustrate how we can take advantage of this in the example of Figure 1. The figure represents scatter plots of 2-dimensional classification problems with class label Y coded by circles (-1) and stars (+1). For concreteness, we can assume that we are studying flowers (*e.g.* Irises) and that X_1 and X_2 represent petal and sepal length. In Figures 1-a and 1-b, we study 2 types of flowers, each type forming a different cluster, labeled by Y. The flower type determines its sepal and petal length, hence $Y \rightarrow \mathbf{X}$. In Figures 1-c and 1-d, we study a population of a single type of flower at different growth stages. When the sum of the petal and sepal length exceeds a given threshold (corresponding to $f(\mathbf{x}) > 0$), we find them large enough to be harvested ($Y = +1$) but we do not cut them otherwise ($Y = -1$). The sepal and petal length determine whether or not to harvest, hence $\mathbf{X} \rightarrow Y$.

In both cases, we show two samples (a training set and a test set) drawn from variants of the distribution: In the $Y \rightarrow \mathbf{X}$ process, only $P(Y)$ is modified, by a change in class proportion; in the $\mathbf{X} \rightarrow Y$ process, only $P(\mathbf{X})$ is modified, by shifting the center of the distribution. To make the two examples resemble one another, we constructed $P(\mathbf{X})$ in the $\mathbf{X} \rightarrow Y$ process such that its two first moments are the same as those of the $Y \rightarrow \mathbf{X}$ process in both the training data and in the test data. Additionally, we chose that the weight and bias of the function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ determining the class label in the $\mathbf{X} \rightarrow Y$ process be the same as the Bayes optimum decision boundary of the training set for the $Y \rightarrow \mathbf{X}$ process.

We argue that, if the process is $\mathbf{X} \rightarrow Y$, then we are better off learning directly $P(Y|\mathbf{X})$, because it will not change under a change of $P(\mathbf{X})$. But if the process is $Y \rightarrow \mathbf{X}$, then we are better off learning $P(\mathbf{X}|Y)$ and making decisions according to $P(Y|\mathbf{X}) \sim P(\mathbf{X}|Y)P(Y)$ (where $\sim$ means "proportional to"). At test time, if $P(Y)$ changes, we can then re-estimate it from the unlabeled test data. In the example, we very simply adjusted the decision boundary with unlabeled test data iteratively, by alternating two steps: (1) label the test examples with $P(\mathbf{X}|Y)P(Y)$; (2) re-estimate $P(Y)$ with $P(Y = +1) = m_+/m$ and $P(Y = -1) = m_-/m$, where m is the number of test examples and m_+ and m_- the number of examples of either class as labeled by the current classifier).

A skilled data analyst may guess by looking at the training data that Figure 1-a corresponds to a $Y \rightarrow \mathbf{X}$ process because the two classes form different clusters and that Figure 1-c corresponds to a $\mathbf{X} \rightarrow Y$ process because the unlabeled data is not clustered and there is a clear-cut decision boundary between the classes. Distributional hints of that sort are used in causal discovery algorithms recently developed [33]. Other clues are obtained from conditional independencies between variables. For instance, the fact that X_1 and X_2 are independent given Y in Figure1-a rules out the possibility that $\mathbf{X} \rightarrow Y$, as we shall see in Sections 6 and 7. However, the data distribution does not always provide sufficient clues to determine causality. In the next example we will see how to exploit manipulations to disambiguate causal relationships.

In summary:

- The Bayes formula $P(\mathbf{X}, Y) = P(\mathbf{X})P(Y|\mathbf{X}) = P(\mathbf{X}|Y)P(Y)$ provides us with the choice of building a predictor by modeling directly $P(Y|\mathbf{X})$ or by modeling $P(\mathbf{X}|Y)$ and performing decisions according to $P(Y|\mathbf{X}) \sim P(\mathbf{X}|Y)P(Y)$. These two choices are not equivalent. If the data generating process is $\mathbf{X} \rightarrow Y$, one should prefer modeling directly $P(Y|\mathbf{X})$, but if $Y \rightarrow \mathbf{X}$, one should prefer modeling $P(\mathbf{X}|Y)$.

- Causal models make less restrictive distribution assumptions than the i.i.d. assumption. If the correct causal model is used, some robustness against distribution changes is built in, or adjustments can be made using unlabeled data.

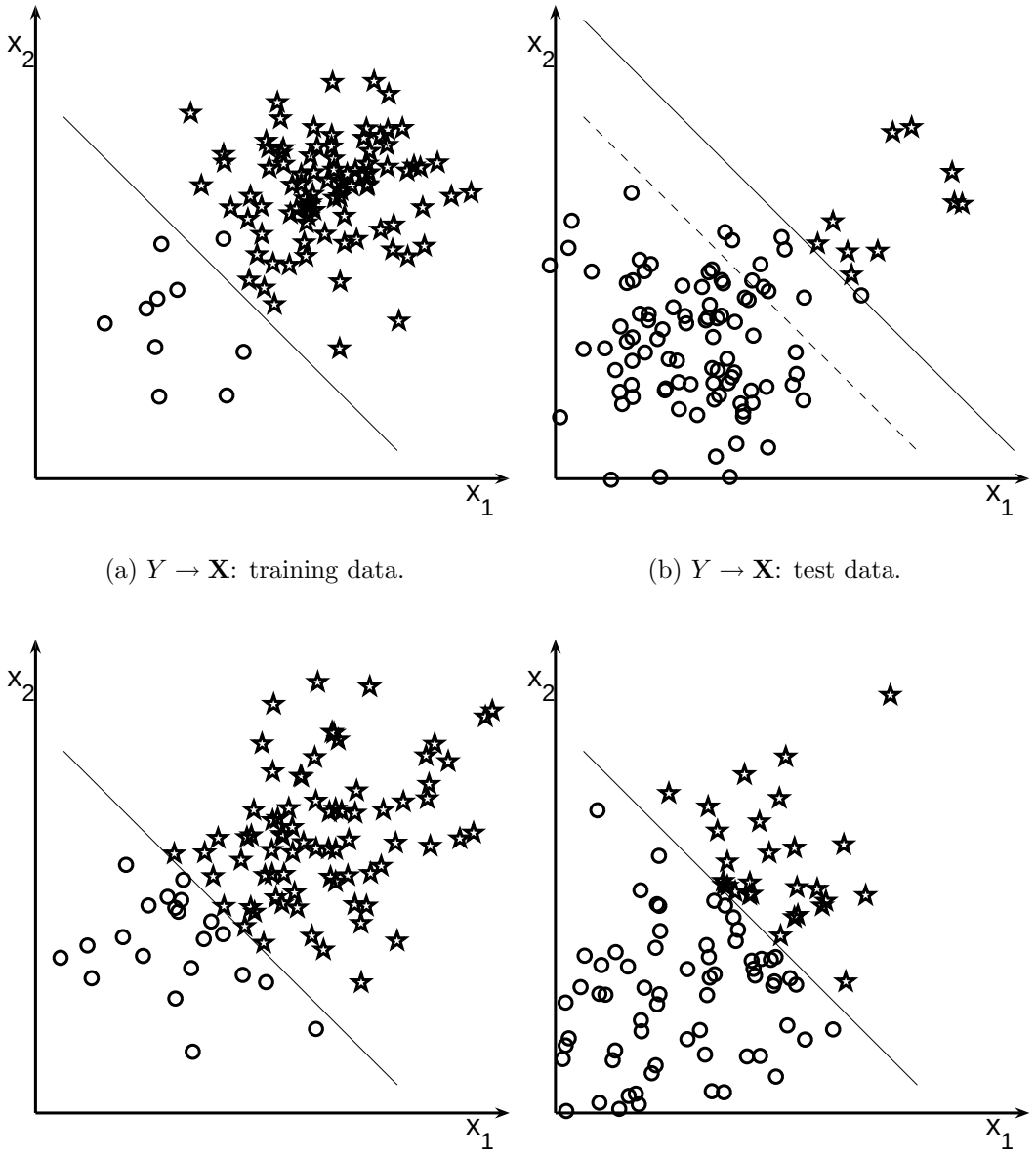
- Causal discovery algorithms exploit properties of the data distribution to infer causal relationships.

Example 2: Manipulation

This second example illustrates the importance of causality in the feature selection process and how causal directions can be inferred from manipulations.

In Figure 2 we depict a process in which two variables X_1 and X_2 are predictive of an outcome Y. Theses two variables are not independent of one another, there is a causal relationship between them: $X_1 \rightarrow X_2$.

To make it easier to understand the example, let us give some meaning to the variables. Suppose that there are two types of workers in a building: those who do clerical work and sit at a desk most of the time and those who do more physical work and run around. Those running around tend to feel warmer and like wearing T-shirts, while the others prefer long-sleeve shirts. Variable X_1 represents the time spent sitting and variable X_2 the length of the sleeves. Imagine now that variable Y represents whether or not one can see the elbows of the person.



(a) $Y \rightarrow \mathbf{X}$: training data.

(b) $Y \rightarrow \mathbf{X}$: test data.

(c) $\mathbf{X} \rightarrow Y$: training data.

(d) $\mathbf{X} \rightarrow Y$: test data.

Figure 1: **Changes in distribution.** Fig. (a) and (b) are generated by a process $Y \rightarrow \mathbf{X}$, in which Y (circle/star coded) is first drawn randomly with $P(Y = +1) = p_+$ and $P(Y = -1) = (1 - p_+)$, then the samples are drawn according to $\mathcal{N}(\mu_\pm, \sigma)$ ($\sigma = 0.75$ for both classes, $\mu_- = [-1, -1]$ and $\mu_+ = [+1, +1]$). (a) Training data: $p_+ = 0.9$; the line represents the Bayes optimal decision boundary $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. (a) Test data, P(Y) changes: $p_+ = 0.1$; the solid line represents the shifted optimal boundary adjusted with unlabeled test data. Fig. (c) and (d) are generated by a process $\mathbf{X} \rightarrow Y$, in which the samples are first drawn randomly according to $\mathcal{N}(\mu, \Sigma)$. The class labels are then obtained deterministically from the sign of $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$. (a) Training data, $\mu = [0.8, 0.8]$: the line represents $f(\mathbf{x})$. (b) Test data, P($\mathbf{X}$) changes: $\mu = [-0.8, -0.8]$; the optimal decision boundary does not change.

In the Figure 2-a, we modeled that situation by drawing X_1 randomly according to a balanced bi-modal distribution (about half of the people are physically active and half sit at a desk) and by drawing X_2 randomly according to a imbalanced bi-modal distribution, whose parameters depend on X_1: 96% people wear long sleeves if they spend a lot of time sitting (that is if x_1 is larger than a threshold) whereas 96% people wear short sleeves if they don't (that is if x_1 is smaller than a threshold). The solid line represents the actual decision boundary determining whether or not one can see the elbow of the person (yes for short sleeve: circle symbol; no for long sleeves: star symbol).

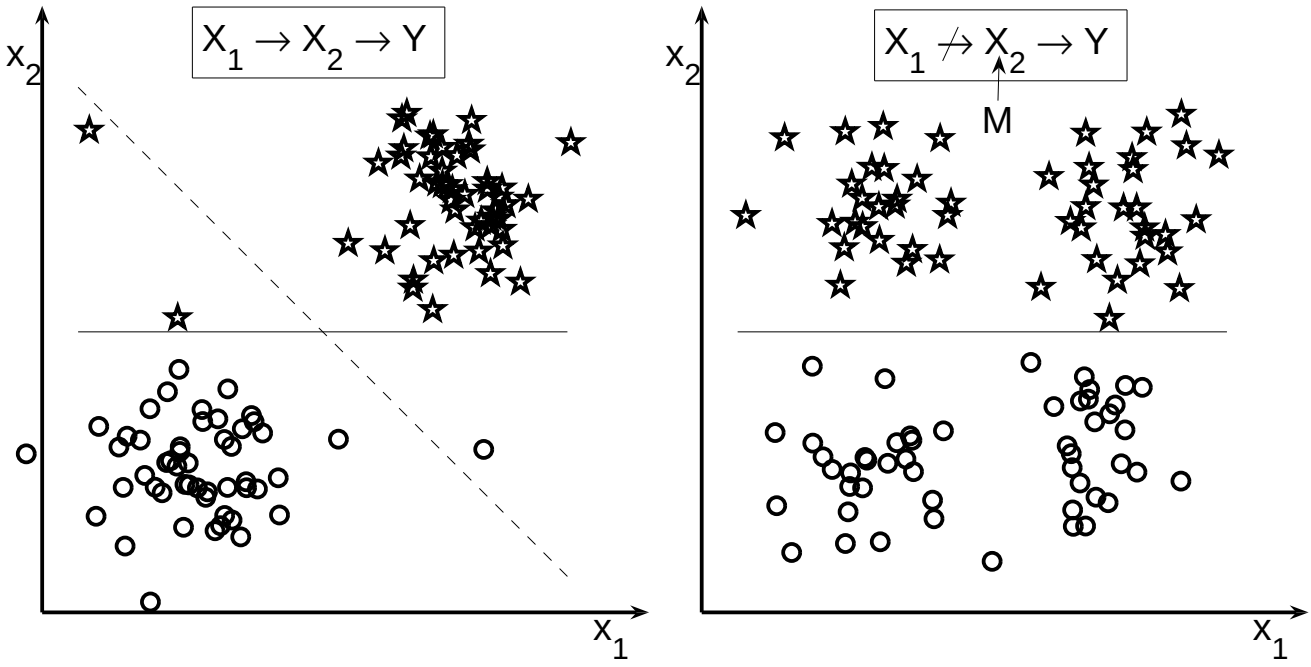
An analyst, who does not know the meaning of the variables and the process, which generated the data, could easily think that both features X_1 and X_2 are informative and that the process is of the type $Y \rightarrow \mathbf{X} = [X_1, X_2]$. According to this hypothesis, the label would be drawn first at random and the samples then drawn according to Gaussian distributions: $P(\mathbf{X}|Y = +1) = \mathcal{N}(\mu_+, \sigma)$, for one class, and $P(\mathbf{X}|Y = -1) = \mathcal{N}(\mu_-, \sigma)$, for the other. In the figure, we represent the Bayes optimum decision boundary corresponding to that model as a dashed line. While, the data distribution does not rule out the causal model $Y \rightarrow \mathbf{X}$, prior knowledge about the meaning of the variables might: the fact that we cannot see the elbow of someone does not determine how long he should be sitting at a desk nor the length of the sleeves he should wear. But things are sometimes subtle. If Y does not represent the *fact* that the elbow is visible, but the *desire* to make the elbow visible, then Y could influence both X_2 (the length of the sleeves) and X_1 (the time spent sitting), because running around maximizes the chances of showing your elbows. This outlines the importance of not mistaking "facts" and "intentions".

Assume that for some reason the meaning of the variables is obscure to the analyst, but he can carry out an experiment: force people at random to wear any length sleeve available, regardless of their work assignments (*e.g.* by asking then to exchange their shirt with someone else chosen at random). We represent this intervention on the system in Figure 2-b by an extra random variable M and an arrow pointing from that variable to X_2. The analyst chose the distribution of variables M to be identical to the marginal distribution $P(X_2)$ in the unmanipulated system. The value of the realization of M drawn at random according to that distribution gets copied to X_2, which is the meaning of the arrow $M \rightarrow X_2$. This effectively disconnects X_2 from any other causal influence and suppresses $X_1 \rightarrow X_2$.

The data obtained after intervention, represented in Figure 2-b, rules out the first hypothesized model $Y \rightarrow \mathbf{X} = [X_1, X_2]$. It is now clear that X_2 influences Y since after manipulation there is a dependency between X_2 and Y: $(X_2 \rightarrow Y)$. The manipulation also indicates that X_2 does not influence X_1, since after manipulation there is no more dependency between X_1 and X_2. This means that the explanation for the dependency between X_1 and X_2 observed in Figure 2-a is $X_1 \rightarrow X_2$.

The causal model determined from the manipulation $X_1 \rightarrow X_2 \rightarrow Y$ indicates that only X_2 is needed to predict Y: given X_2, X_1 and Y are independent. X_2 is the direct cause. Indirect causes like X_1 are not needed to make prediction once the direct causes are known. For this example, the singleton $\{X_2\}$ is an ensemble of variables sufficient to predict Y regardless of the values of the other variables of the system, which we will call in Section 6.1 a Markov blanket of Y.

Determining the causal relationships allows the analyst to build a predictor, which is closer to the optimal predictor (the solid line), because he knows he should use only variable X_2. With this model, no matter what are the changes in fashion and the availability of short/long sleeve



(a) No manipulation.

(b) Manipulation of X_2.

Figure 2: **Effect of a manipulation.** Scatter plots of a 2-dimensional classification problem with class label Y coded by circles (+1) and stars (-1). In both figures, X_1 is drawn from a Gaussian mixture, $\sigma_1 = \sigma_2 = 0.4$, $\mu_1 = 1$, $\mu_2 = 2$, and equal proportions $p_1 = p_2 = 0.5$. X_2 is also drawn from the same Gaussian mixture, but the proportions are different. (a) The mixture proportions for X_2 depend on X_1: if $x_1 > 1.5$, then $p_1 = 0.04$ and $p_2 = 0.96$, if $x_1 \leq 1.5$, the proportions are inverted. (b) The mixture proportions for X_2 do not depend on X_1: $p_1 = p_2 = 0.5$.

shirts (some additional influential variables not taken into account at the time of the analysis), good predictions can be made about whether or not the elbow can be seen. With the wrong model (dashed line), a change in the distribution of X_2 may yield many more errors.

Several lessons can be learned from that example:

- Observational data (Figure 2-a) may be ambiguous with respect to causal relationships.
- Predictors built using wrong causal assumptions ($Y \rightarrow \mathbf{X} = [X_1, X_2]$ instead of $X_1 \rightarrow X_2 \rightarrow Y$) may significantly differ from the optimal predictor (the dashed line instead if the solid line).
- Prior knowledge about the meaning of the variables can considerably help, because if we know the correct causal relationships ($X_1 \rightarrow X_2 \rightarrow Y$), a good predictor can be obtained from observational data only: For the purpose of predicting Y we should use only variable X_2.
- The fact that X_1 and Y are correlated does not make X_1 a useful feature for predicting Y if X_2 is known. In fact, adding it may be *detrimental*, depending on the predictor used.

- Manipulations allow us to disambiguate causal relationships, which cannot be guessed from properties of the distribution alone.

3 Classical "non-causal" feature selection

In this section, we give formal definitions of irrelevance and relevance from the point of view of classical (non-causal) feature selection. We review examples of feature selection algorithms. We show the limitation of univariate methods. In what follows, the feature set is a random vector $\mathbf{X} = [X_1, X_2, X_N]$ and the target is a random variable Y. Training and test data are drawn according to a distribution $P(\mathbf{X}, Y)$. We use the following definitions and notations for independence and conditional independence between random variables:

Conditional independence Two random variables A and B are *conditionally independent* given a set of random variables $\mathbf{C}$, denoted as $A \perp B|\mathbf{C}$, *iff* $P(A, B|\mathbf{C}) = P(A|\mathbf{C})P(B|\mathbf{C})$, for all assignment of values to A, B, and $\mathbf{C}$. If $\mathbf{C}$ is the empty set, then A and B are *independent*, denoted as $A \perp B$.

3.1 Individual feature relevance

A simple notion of relevance can be defined by considering only the dependencies between the target and individual variables.

Individual feature irrelevance. The feature X_i is individually irrelevant to the target Y *iff* X_i is independent of Y (denoted as $X_i \perp Y$): $P(X_i, Y) = P(X_i)P(Y)$.

From that definition it should simply follow that all non individually irrelevant features are individually relevant (denoted as $X_i \not\perp Y$). However, when a finite training sample is provided, the statistical significance of the relevance must be assessed. This is done by carrying out a statistical test with null hypothesis "H_0: the feature is individually irrelevant" (that is X_i and Y are statistically independent). A wide variety of test statistics have been proposed, providing various tradeoffs between type I errors (features falsely considered relevant) and type II errors (features falsely considered irrelevant). Such statistics are used to discard features called "irrelevant", which are above a certain *pvalue*, or simply to provide a feature ranking. For a review, see *e.g.* [14], chapters 2 and 3.

3.2 When univariate feature selection fails

Feature selection based on individual feature relevance, as defined in the previous section, is called "univariate". In this section, we provide a justification for the development and use of "multivariate" techniques: In the context of other variables $\mathbf{X}^{\setminus i} = [X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_N]$, a variable X_i individually relevant to Y may become irrelevant, or vice versa.

We illustrate various cases with scatter plots, using problems with two continuous variables and one binary variable. The two values of the binary variable are represented by star and circle symbols. The first two examples are *classification* problems in which y is the binary variable. The last example is a *regression* problem: y is continuous and one of the predictor variables is binary:

- **Falsely irrelevant variables.** In Figure 3 (a), variable X_2 and Y are independent ($X_2 \perp Y$), yet, in the context of variable X_1 there are dependent ($X_2 \not\perp Y|X_1$). The separation between the two classes by the individually relevant variable X_1 can be improved by adding the individually **irrelevant** variable X_2. Univariate feature selection methods would fail to discover the usefulness of variable X_2. In Figure 3 (b), we show a trickier example in which both variables X_1 and X_2 are independent of Y ($X_1 \perp Y$ and $X_2 \perp Y$). Each variable taken separately does not separate the target at all, while taken jointly, they provide a perfect non-linear separation ($\{X_1, X_2\} \not\perp Y$). This problem is known in machine learning as the "chessboard problem" and bears resemblance with the XOR and parity problems. Univariate feature selection methods fail to discover the usefulness of variables for such problems. Also notice that in this example, X_1 and X_2 are independent of one another (this can be seen if you ignore nature of the markers star or circle). Therefore independence between features does not provide a useful justification for using univariate feature selection methods.

- **Falsely relevant variables.** In Figure 3(c) and (d), we show an example of the opposite effect. X_1 and Y are dependent when taken out of the context of X_2. However, conditioned on any value of X_2, they are independent ($X_1 \perp Y|X_2$). This problem is know in statistics as Simpson's paradox. In this case, univariate feature selection methods might find feature X_1 relevant, even though it is redundant with X_2. If X_2 were unknown (unobserved), the observed dependency between X_1 and Y may be spurious, as it vanishes when the "confounding factor" X_2 is discovered (see Section 2.4).

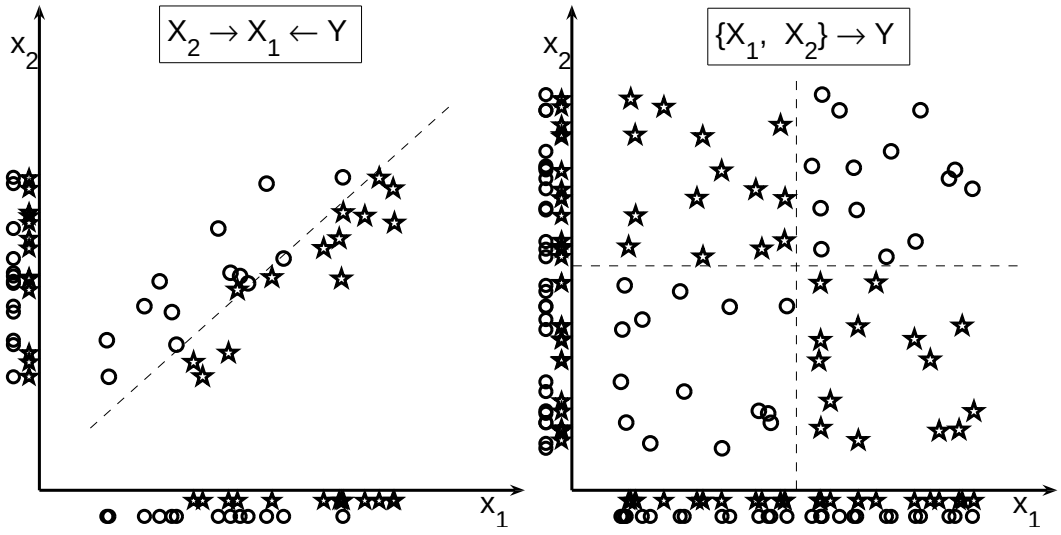
The limitations of univariate feature selection fostered research in multivariate techniques, which are briefly reviewed in the next section.

3.3 Classical feature selection strategies

Multivariate feature selection may be performed by searching in the space of possible feature subsets for an optimal subset. Figure 4 depicts the feature selection space for four features according to [19]. Search algorithms explore this space by making elementary moves, re-evaluating at each step an objective function. Therefore, the ingredients of feature selection algorithms include:

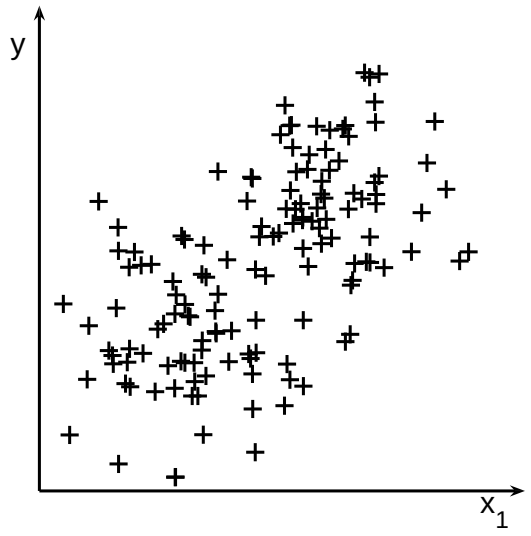
- a search algorithm,
- an objective function,
- a stopping criterion.

Feature selection techniques have been classified into **filters**, **wrappers** and **embedded methods** [19, 5]. They differ in the choice of the three basic ingredients previously outlined: **Wrappers** use the actual risk functional of the machine learning problem at hand to evaluate feature subsets. They must train one learning machine for each feature subset investigated. **Filters** use another evaluation function than the actual risk functional. Often no learning machine is involved in the feature selection process.

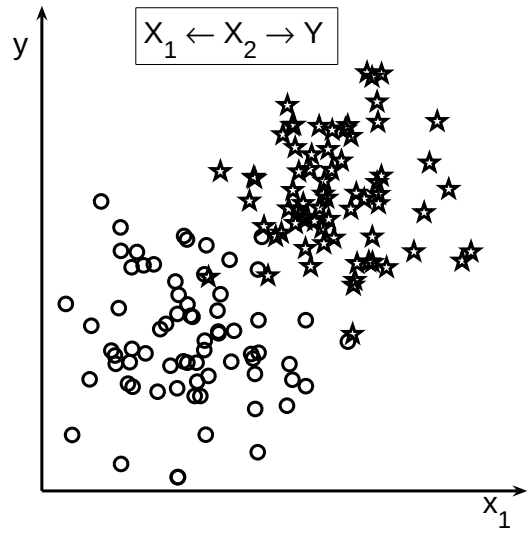


(a) $X_2 \perp Y, X_2 \not\perp Y|X_1$

(b) $X_1 \perp Y, X_2 \perp Y, \{X_1, X_2\} \not\perp Y$



(c) $X_1 \not\perp Y$



(d) $X_1 \perp Y|X_2$

Figure 3: **Multivariate dependencies.** (a) **Spouse problem:** Feature X_2 (a spouse of Y having the common child X_1) is individually irrelevant to Y ($X_2 \perp Y$), but it becomes relevant in the context of feature X_1 ($X_2 \not\perp Y|X_1$). (b) **Chessboard problem:** Two individually irrelevant features ($X_1 \perp Y$ and $X_2 \perp Y$) become relevant when taken jointly ($\{X_1, X_2\} \not\perp Y$). (c-d) **Simpson's paradox and the confounder problem:** (c) X_1 is correlated to Y, it not independent of Y ($X_1 \not\perp Y$). It is individually relevant, but it may become irrelevant in the context of another feature, see case d. (d) For any value of X_2 (star or circle) X_1 is independent of Y ($X_1 \perp Y|X_2$). Note: we show at the top of each scatter plot the causal structure of the models, which generated the data. In some cases, the same data can be explained by several alternative causal models (see Section 5).

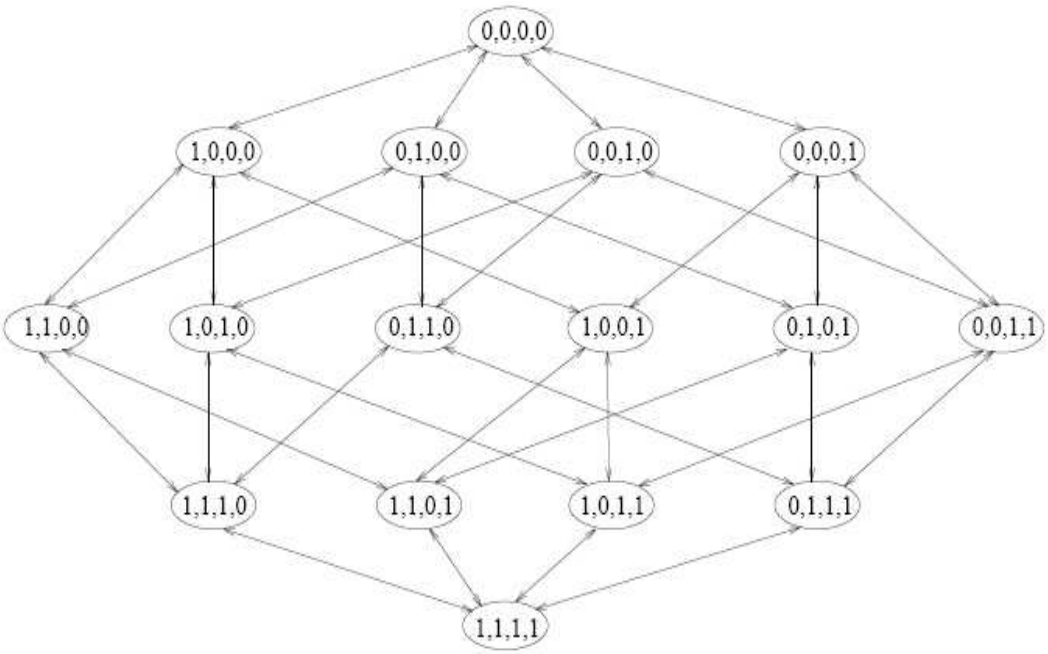


Figure 4: **Exploring feature space.** This is an example of a four-dimensional feature selection state space. Each node represents one feature selection subset (0/1 indicates whether the feature is selected). Arrows indicate transitions by adding or removing a single feature.

The size of the search space for N features is 2^N. This exponential growth in the number of features causes both computational problems and statistical problems (overfitting). Various proposals have been made to reduce both computational and statistical complexity. For wrappers, some greedy search methods, which require only of the order of $O(N^2)$ moves in feature selection space, are popular. They include forward selection methods, which progressively incorporate features starting from an empty set, and backward elimination methods which progressively remove features starting from the full set. Mixed forward/backward search have been developed to make the search more flexible, while incurring moderate additional search time. In addition to using cheap search methods, the designers of filter methods also proposed many cheap evaluation function not requiring to train a learning machine. The cheapest of all are univariate filters requiring only $O(N)$ evaluations. Finally, many learning machines lend themselves to the incorporation of an embedded feature selection method. For **embedded methods**, the feature selection space and the learning machine parameter space are searched simultaneously. For a review of filter, wrapper and embedded methods, see [14].

Notably, in spite of the limitations of univariate feature selection, which we have outlined in Section 3.2, and despite the existence of many good multivariate feature selection techniques, it has been observed in practical applications and in competitions (see the results of the feature selection challenge [13]) that univariate feature selection or even no feature selection at all often yields better results.

Multivariate feature selection techniques have been obtaining success on "parsimony" by discovering smaller feature subsets with moderate degradation in performance. However, this is often accompanied with a great sensitivity to changes in the distribution (covariate shift). In what follows, we will see how a finer analysis using causal discovery may help strengthening

multivariate feature selection. Before that, we give a more general definition of feature relevance, which will come handy in our discussion.

4 Non causal feature relevance

In Section 3.1, we have defined individual feature relevance. We now generalize this definition to the relevance of features in the context of others. We first introduce irrelevance as a consequence of random variable independence and then define relevance by contrast. For simplicity, we provide only asymptotic definitions, which assume the full knowledge of the data distribution. For a discussion of the finite sample case, see the introduction chapter of [14]. In what follows, $\mathbf{X} = [X_1, X_2, ..., X_i, ..., X_N]$ denotes the set of all features, $\mathbf{X}^{\setminus i}$ is the set of all features except X_i, and $\mathbf{V}^{\setminus i}$ any subset of $\mathbf{X}^{\setminus i}$ (including $\mathbf{X}^{\setminus i}$ itself).

Feature irrelevance A feature X_i is irrelevant to the target Y *iff* for all subset of features $\mathbf{V}^{\setminus i}$, and for all assignments of values

$$P(X_i, Y|\mathbf{V}^{\setminus i}) = P(X_i|\mathbf{V}^{\setminus i})P(Y|\mathbf{V}^{\setminus i}).$$

Kohavi and John define a notion of strong and weak relevance. Intuitively and in many practical cases (but not always as we will show below), a strongly relevant feature is needed on its own and cannot be removed without degrading prediction performance, while a weakly relevant feature is redundant with other relevant features and can therefore be omitted if similar features are retained.

Strong relevance A feature X_i is strongly relevant to the target Y *iff* there exist some values x, y and $\mathbf{v}$ with $P(X_i = x, \mathbf{X}^{\setminus i} = \mathbf{v}) > 0$ such that: $P(Y = y|X_i = x, \mathbf{X}^{\setminus i} = \mathbf{v}) \neq P(Y = y|\mathbf{X}^{\setminus i} = \mathbf{v})$.

Weak relevance A feature X_i is weakly relevant to the target Y *iff* it is not strongly relevant and if there exist a subset of features $\mathbf{V}^{\setminus i}$ for which there exist some values x, y and $\mathbf{v}$ with $P(X_i = x, \mathbf{V}^{\setminus i} = \mathbf{v}) > 0$ such that: $P(Y = y|X_i = x, \mathbf{V}^{\setminus i} = \mathbf{v}) \neq P(Y = y|\mathbf{V}^{\setminus i} = \mathbf{v})$.

From the above definitions, and noting that $P(Y|X_i, \mathbf{V}^{\setminus i}) = P(Y|\mathbf{V}^{\setminus i})$ implies that $P(X_i, Y|\mathbf{V}^{\setminus i}) = P(X_i|\mathbf{V}^{\setminus i})P(Y|\mathbf{V}^{\setminus i})$ (by applying Bayes' rule), one can easily see that a feature is either irrelevant, strongly relevant or weakly relevant.

Kohavi and John also make a distinction between relevance and usefulness. The feature selection problem, which we described in Section 3.3, seeks feature subsets that are optimal with respect to a given objective. The resulting features are "useful" to achieve the desired objective as well as possible. Relevant features (according to the definitions of weak and strong relevance) are not always useful to improving predictions according to a given objective. This is easily understood for *weakly relevant* features because their redundancy with other features may render them unnecessary to improve prediction. Perhaps more surprisingly, eliminating a *strongly relevant* features may yield no performance degradation. In the example of Figure 5 we show a classification problem. Feature X_2 is strongly relevant, since clearly there are values of X_2 such that $P(Y|X_2, X_1) \neq P(Y|X_1)$. Yet, the classification performance achieved by the best separator based on both X_1 and X_2 is no better than the one based on X_1 alone.

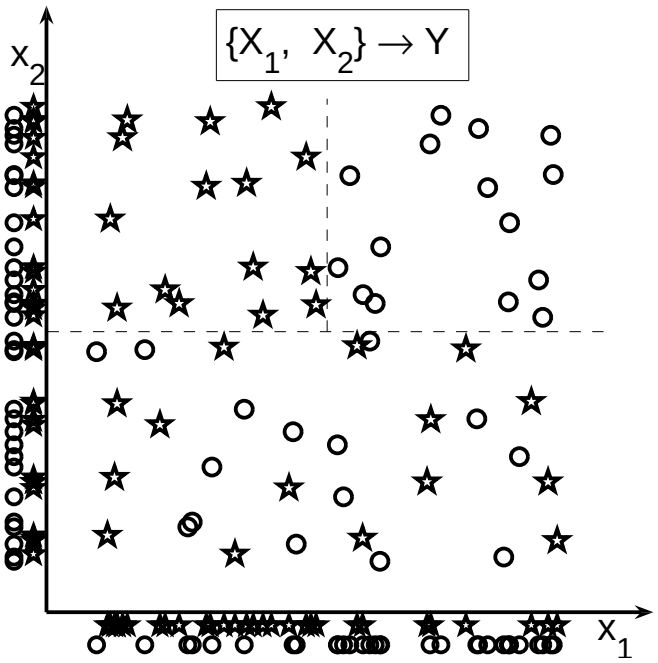


Figure 5: **Relevance does not imply usefulness.** In this classification example (the target Y is represented by the star and circle symbols), the samples are uniformly drawn at random in the unit square. Y is randomly assigned in the bottom part, $Y = +1$ is the top left corner and $Y = +1$ in the top right corner. Feature X_2 is strongly relevant in the Kohavi-John sense since $P(Y|X_2, X_1) \neq P(Y|X_1)$ for any value of X_2, yet it is not useful to improve classification performance, compared to using X_1 alone (the error rate of 25% cannot be improved by adding X_2).

Reciprocally Kohavi and John argue that a useful feature may be irrelevant. They give the examples of a constant input to a linear predictor, which transforms it into a more powerful affine predictor and increases its performance; yet a constant value cannot be thought of as "relevant".

The issue of relevance has been the subject of much debate (see *e.g.* the special issue of Artificial Intelligence on relevance [1]) and the recent discussion of Tsamardinos and Aliferis [34], which challenge the universality of any particular notion of relevance or usefulness of features). Although much remains to be said about such non causal relevance, we wish now to introduce the concept of causality and show how it will shed light on the notion of feature relevance.

5 The concept of causality

Formal, widely-accepted definitions of causality have eluded philosophers of science for centuries. However the notion of causality is at the core of the scientific endeavor and also a universally accepted and intuitive notion of everyday life. The lack of broadly acceptable definitions of causality has not prevented the development of successful and mature mathematical and algorithmic frameworks for inducing causal relationships.

From an engineering point of view, causality is a very goal-oriented notion, which can simply be defined as finding modes of action on a system, which will result in a desired outcome. For example, taking a drug to cure illness. Thus, even though causality may not find a perfect definition regrouping all the notions it encompasses in philosophy, psychology, history, law, religion, statistics, physics, and engineering, we can devise tests of causality that satisfy our engineering-oriented goal by assessing the effect of actual or hypothetical manipulations performed on the system.

In this section we introduce some simple models of causal relationships based on Markov models and Bayesian networks, which prove to be useful tools to understand and use the notion of causality.

5.1 Systemic causality, events as causes and effects

Following frameworks developed in the past decades in various application domains [11, 28, 32, 25], our approach to causality is systemic: causes and effects are limited to the study of well defined systems, which may be isolated from their environment and subjected to experimentation or **manipulation** by external agents. This makes a fundamental distinction between "inside" and "outside" factors and makes possible the definition of input, state, and output. In some cases, experimentation is impossible for practical or ethical reasons. Still, it is important to the concept of causality, which we adopted, that the system could in principle be subject to experimentation, even though it may be infeasible and will not happen.

Mathematically, cause-effect relationships may be defined either between **events** or between **random variables**. In this report, we are mostly interested in causality between random variables because of its deep connection to the problem of feature selection. So we briefly explain the notion of causality between events (defined as sets of system states) because of its intuitive nature and then move to the problem of causality between random variables, which we fully develop.

The following widely accepted premises characterize causality:

1. **Causes precede their effects** and therefore causality is both irreflexive (A cannot cause A) and antisymmetric (if A causes B, B cannot cause A).

2. **The same causes produce the same effects** and therefore experiments can be conducted to validate causal conjectures.

3. **Causality is transitive**: if A causes B and B causes C, than A causes C.

For instance, if our system of interest is a fully observable deterministic finite state automaton, any event C (*e.g.* a single state of the model) happening before another event E is a cause of E and E is its effect. In what follows, we do not specify the exact timing, but we will always assume that C happens at a given time step n and E at another time step m, with $m > n$.

If our automaton is only partially observable, defining causality becomes more subtle. An event is now a set of states having in common the same visible part (the value of a subset of the state variables). Now, observing that C precedes E is no-longer sufficient to determine a cause-effect relationship between C and E. The reason is that there may be a **common ancestor state**, preceding C and E, which triggered both C and E. The information from that ancestor state used to trigger E does not need to be conveyed to E by C, it may be conveyed to E via the hidden state variables. To establish causal relationships, we must decouple the effect of ancestor states from the alleged effect of C by conducting experiments.

5.2 Probabilistic causality

To accommodate the fact that in non-deterministic systems such as Markov processes multiple causal events can produce the same consequential event and that and a same causal event can produce multiple consequential events, we must relax the condition "the same causes produce the same effects" and define causality in a probabilistic manner. If we perform an experiment or **manipulation** consisting in **setting** visible state variables to desired values C, which we call do(C) after the nomenclature of Pearl [28], then C is called a cause of E, and E its effect, if $P(E|do(C) \neq P(E)$, that is if the manipulation results in a change in the distribution of E compared to letting the system evolve according to its own dynamics.

It is important to understand the difference between $P(E|do(C)$ and $P(E|C)$. The former corresponds to the probability of reaching E when the visible part of the state is deliberately set by an external agent to C, the other variables assuming random values. The latter corresponds to observing when E is reached after C was *observed*, which corresponds to a different distribution of the hidden variables. If $P(E|C) \neq P(E)$, we can only say that there is a statistical dependency (correlation) between C and E, not a causal relationship. This important distinction is referred to as "correlation *vs.* causation".

5.3 Time dependency of causality

Our everyday-life concept of causality is very much linked to the time dependency of events. However, such temporal notion of causality is not always necessary nor convenient. In particular, many machine learning problem are concerned with "cross-sectional studies", which are studies where many samples of the state of a system were drawn at a given point in time. Thus, we will drop altogether the reference to time and replace it by the notion of "causal ordering".

Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t + \Delta t$, where Δt can be made as small as we want.

5.4 Causality between random variables

Events provided us with a convenient way of thinking about causal relationships, but to make connections to variable/feature selection and causality, we must now move to the notion of causality between random variables. A simple way to make this transition is to introduce indicator variables, whose ± 1 values indicate whether an event is present or absent. With some abuse of notation, we will use the uppercase notation C end E for the indicator variables representing presence or absence of events C and E.

Test of causality between indicator variables. We define a manipulation "do(C)" for indicator variables as the action of an external agent imposing with equal probability that $C = -1$ or $C = +1$. A random variable C is then identified as a cause of E if $P(E|do(C) \neq P(E)$ (or equivalently if $P(E|do(C = -1)) \neq P(E|do(C = +1))$).

We notice the similarity between this test and our definition of causality between events in Section 5.2. It is inspired by the the operational criterion of causality based on randomized controlled experiments (RCE) of Glymour and Cooper [11].

For example, C may be the choice of one of two available treatments for a patient with Lung Cancer and E may represent 5-year survival. If we randomly assign patients to the two treatments by flipping a fair coin and observe their survival status after 5 years and the probability distribution for survival differs between the two treatment groups, then according to the operational criterion we conclude that the choice of treatment causally determines survival in patients with Lung cancer.

There is also a parallel between the test of Definition 5.4 and the notion of individual feature relevance of Definition 3.1. A feature X is individually irrelevant to the target Y *iff* $P(X, Y) = P(X)P(Y)$, that is, assuming that $P(X) > 0$, if $P(Y|X) = P(Y)$. Hence, individual relevance defined by contrast occurs if for some assignment of values $P(Y|X) \neq P(Y)$. This is to be compared with $P(Y|do(X)) \neq P(Y)$ in the test of causality.

The test of causality between indicator variables can then be generalized to any random variable, binary, categorical or continuous. We use it to define a notion of "individual causal relevance". We consider a system described by a set of random variables $\{\mathbf{X}, Y\}$ including $X_i \in \mathbf{X}$.

Manipulation. We call "manipulation" of X_i and denote it as "do(X_i)" an intervention of an external agent imposing values to X_i, not drawn according to its natural distribution, but according to an arbitrary distribution independent of the system, usually (but not necessarily) uniform over a set of admissible values.

Individual causal relevance. A random variable (feature) X_i is individually causally relevant to a random variable (target) Y *iff* there exits a manipulation "do(.)" and there exist some values x and y with $P(do(X_i = x)) > 0$ such that $P(Y = y|do(X_i = x)) \neq P(Y = y)$.

Individual causal relevance does not fully characterize causal relevance. This criterion is incomplete because it examines one variable at a time. Hence if, for example, Y is the parity of $X_1, X_2, ...X_n$, the individual causal relevance criterion will not verify that X_1 causes Y (and similarly for the chessboard problem, see Figure 3-b). If one however conducts a joint manipulation on $\mathbf{X}$ the criterion becomes complete.

To go beyond individual causal relevance, using manipulations as tools to unravel causal dependencies, we can also define notions of weak and strong *causal* relevance analogously to the Kohavi-John sense *non causal* relevance, (see for comparison Definitions 4 and 4). We call $\mathbf{X}^{\setminus i}$ the subset of all variables not including X_i.

Strong causal relevance A random variable (feature) X_i is strongly causally relevant to a random variable (target) Y *iff* there exits a manipulation "do(.)" and there exist some values x, y and $\mathbf{v}$ with $P(do(X_i = x), do(\mathbf{X}^{\setminus i}) = \mathbf{v}) > 0$ such that: $P(Y = y|do(X_i) = x, do(\mathbf{X}^{\setminus i}) = \mathbf{v}) \neq P(Y = y|do(\mathbf{X}^{\setminus i}) = \mathbf{v})$.

Weak causal relevance A random variable (feature) X_i i s weakly causally relevant to a random variable (target) Y *iff* it is not strongly causally relevant and if there exit a manipulation "do(.)" and a subset $\mathbf{V}^{\setminus i} \in \mathbf{X}^{\setminus i}$ for which there exist some values x, y and $\mathbf{v}$ with $P(X_i = x, do(\mathbf{V}^{\setminus i}) = \mathbf{v}) > 0$ such that: $P(Y = y|do(X_i) = x, do(\mathbf{V}^{\setminus i}) = \mathbf{v}) \neq P(Y = y|do(\mathbf{V}^{\setminus i}) = \mathbf{v})$.

Although these definitions are formally interesting in that they establish a parallel with the feature selection framework, they have little practical value. First, they have the same drawback as their non-causal feature relevance counterpart that they require exploring all possible subsets of features and assignment of values to features. Second, the fact that they require exploring all possible manipulations to establish the absence of causal relationship with certainty, is also unrealistic. When we may establish a causal relationship using a manipulation on X_i, thereafter any other manipulation that affects X_i will potentially affect Y. But the converse is not true. We must in principle try "all possible" manipulations to establish with certainty that there is no causal relationship. Practically however, planned experiments have been devised as canonical manipulations and are commonly relied upon to rule out causal relationships (see *e.g.* [30]). Nonetheless, they require conducting experiments, which may be costly, impractical, unethical or even infeasible. The purpose of the following sections is to introduce the reader to the discovery of causality in the absence of experimentation. Experimentation will be performed punctually, when absolutely needed.

5.5 Causal Bayesian networks

Causal Bayesian networks provide a convenient framework for reasoning about causality between random variables. Causal Bayesian networks implement a notion of causal ordering and do not model causal time dependencies in details (although they can be extended to do so if desired). Even though other frameworks exist (like structural equation modeling [17, 18]), we limit ourselves in this report to Bayesian networks, because they allow us to illustrate simply the connections between feature selection and causality we are interested in.

Recall that in a directed acyclic graph (DAG), a node A is the parent of B (B is the child of A) if there is a direct edge from A to B, A is the ancestor of B (B is the descendant of A) if there is a direct path from A to B. "Nodes" and "variables" will be used interchangeably.

As in previous sections, we denote random variables with uppercase letters X, Y, Z, realizations (values) with lowercase letters, x, y, z, and sets of variables or values with boldface uppercase $\mathbf{X} = [X_1, X_2, ..., X_N]$ or lowercase $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ respectively. A "target" variable is denoted as Y.

We begin by formally defining a discrete Bayesian network:

Discrete Bayesian network. Let $\mathbf{X}$ be a set of discrete random variables and P be a joint probability distribution over all possible realizations of $\mathbf{X}$. Let $\mathcal{G}$ be a directed acyclic graph (DAG) and let all nodes of $\mathcal{G}$ correspond one-to-one to members of $\mathbf{X}$. We require that for every node $A \in \mathbf{X}$, A is probabilistically independent of all non-descendants of A, given the parents of A (Markov Condition). Then we call the triplet $\{\mathbf{X}, \mathcal{G}, P\}$ a (discrete) Bayesian Network, or equivalently a Belief Network or Probabilistic Network (see *e.g.* [26]).

Discrete Bayesian networks can be generalized to networks of continuous random variables and distributions are then replaced by densities. To simplify our presentation, we limit ourselves to discrete Bayesian networks. From the definition of Bayesian networks, a formal definition of causal Bayesian networks can be given:

Causal Bayesian network. A causal Bayesian network is a Bayesian Network $\{\mathbf{X}, \mathcal{G}, P\}$ with the additional semantics that $(\forall A \in \mathbf{X})$ and $(\forall B \in \mathbf{X})$, if there is an edge from A to B in $\mathcal{G}$ then A directly causes B (see *e.g.* [32]).

Using the notion of *d-separation* (see section 5.6 *e.g.* [28]), it is possible to read from a graph $\mathcal{G}$ if two sets of nodes $\mathbf{A}$ and $\mathbf{B}$ are independent, conditioned on a third set $\mathbf{C}$: $\mathbf{A} \perp_{\mathcal{G}} \mathbf{B}|\mathbf{C}$. Furthermore, in causal Bayesian Network, the existence of a directed path between two nodes indicates a causal relationship.

It is usually assumed that in addition to the Markov condition, which is part of the definition of Bayesian networks, another condition called "faithfulness" is also fulfilled. The faithfulness condition entails *dependencies* in the distribution from the graph while the Markov condition entails that *independencies* in the distribution are represented in the graph. Together the Markov and faithfulness conditions guarantee that the Bayesian network will be an accurate map of dependencies and independencies of the represented distribution.

Faithfulness of a DAG to a distribution A Directed Acyclic Graph $\mathcal{G}$ is faithful to a joint probability distribution P over a set of variables $\mathbf{X}$ *iff* every independence present in P is entailed by $\mathcal{G}$ and the Markov condition, that is $(\forall A \in \mathbf{X}, \forall B \in \mathbf{X}$ and $\forall \mathbf{C} \subset \mathbf{X})$, $A \not\perp_{\mathcal{G}} B|\mathbf{C} \Rightarrow A \not\perp_P B|\mathbf{C}$.

Both the Markov condition and the faithfulness conditions can be trivially specialized to causal Bayesian networks as follows:

Causal Markov condition (CMC) In a causal Bayesian networks satisfying the CMC, every node is independent of is non-effects given its immediate causes.

Causal faithfulness condition (CFC) A causal Bayesian network is faithful if it satisfies the faithfulness condition of Definition 5.5.

As a consequence of the CMC, the joint probability can be factorized simply as: $P(X_1, X_2, ..., X_N) = \sum_i P(X_i|DirectCauses(X_i))$. Another consequence is that $A \perp_{\mathcal{G}} B|\mathbf{C} \Rightarrow A \perp_P B|\mathbf{C}$ meaning that independence relationships read on the graph with the notion of *d-separation* reflect true independence relationships according to the distribution P. In a faithful causal Bayesian networks, *d-separation* captures all conditional dependence and independence relations that are encoded in the graph (see *e.g.* [32]).

It is also uselful to define the faithfulness of a distribution:

Faithfulness of a distribution A distribution P over a set of variables $\mathbf{X}$ is said to be *faithful iff* there exists a DAG $\mathcal{G}$ satisfying the faithfulness condition of Definition 5.5.

The chessboard/XOR problem is a typical example of unfaithfulness (Figure 3-b). In that case, we have independency between all pairs of variables, which could only be represented by a graph with no connection if the CFC held. But in reality, Y was produced from the joint distribution of X_1 and X_2. The arrows in the causal graph describing the data generating process $X_1 \leftarrow Y \rightarrow X_2$ should not mean that X_1 or X_2 are individually dependent on Y. This violates the CFC. It can be easily verified that no graph can faithfully represent the given set of dependencies and independencies.

Causal Bayesian networks are fully defined by their graph and the conditional probabilities $P(X_i|DirectCauses(X_i))$. Those may be given by experts or trained from data (or a combination of both). Once trained, a Bayesian network may be used to compute any joint probability or conditional probability involving a subset of variables, using the Bayesian **chain rule** $P(A, B, C, D, E) = P(A|B, C, D, E)P(B|C, D, E)P(C|D, E)P(E)$ and **marginalization** $P(A, B) = \sum_{C,D,E} P(A, B, C, D, E)$. Such calculations are referred to as **inference** in Bayesian networks. In the worst cases, inference in Bayesian networks is intractable. However many very efficient algorithms have been described for exact and approximate inference [27, 26].

5.6 Learning the network causal structure

The structure of a causal graph can, to some extent, be determined from observational data (*i.e.* without manipulation). One method consists in making statistical **tests of conditional independence** between variables (see Section 6).

Let us take the simple example of **a system of only 3 variables** A, B, C. We can enumerate all the possibilities of DAGS (directed acyclic graphs), up to a permutation of the variable names:

1. Completely unconnected graph: A, B, C.
2. Single arrow chain: $A \rightarrow C, B$ or $A \leftarrow C, B$.
3. **Chain**: $A \rightarrow C \rightarrow B$ or $A \leftarrow C \leftarrow B$.
4. **Fork**: $A \leftarrow C \rightarrow B$.
5. **Collider** (also called V-structure): $A \rightarrow C \leftarrow B$
6. Fully connected graph, with two alternative paths: $A \rightarrow C \rightarrow B$ and $A \rightarrow B$ (and all permitted permutations of arrow directions, avoiding cycles).

In faithful causal Bayesian networks, these cases correspond to the following conditional independencies (also called Markov properties), excluding other independencies:

1. Completely unconnected graph: $A \perp B$, $A \perp C$, and $C \perp B$.
2. Single arrow chain: $A \perp B$, and $C \perp B$.
3. **Chain**: $A \perp B|C$.
4. **Fork**: $A \perp B|C$.
5. **Collider**: $A \perp B$ (but $A \not\perp B|C$).
6. Graph with two alternative paths: No independencies.

It can be verified for this simple three node DAG that:

- In the case of the chains, the direction of the arrows can be reverted without changing the conditional independencies.
- The two-arrow fork and chain have the same conditional independencies.
- When there are no independencies, the direction of the arrows can be anything.

Consequently, only the unconnected graph and the collider are unambiguously determined by the Markov properties. This property of colliders is key in the learning of causal structure in Bayesian networks with so-called **constraint-based methods** (see Section 7).

Having defined **chains**, **forks** and **colliders** we can simply define **d-separation**, which is a useful operational criterion that can be applied to a Bayesian network graph to obtain all the independencies entailed by the Markov condition as proven by [28].

d-separation, d-connection. A path Π between two variables A and B is blocked by a set of nodes $\mathbf{C}$ if (1) Π contains a chain $I \rightarrow C \rightarrow J$ or a fork $I \leftarrow C \rightarrow J$ such that C is in $\mathbf{C}$, or (2) Π does not contain a collider $I \rightarrow C \leftarrow J$ such that C or any of its descendants are in $\mathbf{C}$. A set $\mathbf{C}$ is said to d-separate A from B if $\mathbf{C}$ blocks every path between A and B. If A and B are not d-separated, then they are d-connected.

Two variables A and B are d-separated given a conditioning set $\mathbf{C}$ in a faithful Bayesian networks (or causal Bayesian networks) if and only if $A \perp_P B|\mathbf{C}$ [32]. It follows, that if they are d-connected, they are not conditionally independent. Thus, in a faithful Bayesian networks, d-separation captures all conditional dependence and independence relations that are encoded in the graph. This property is algorithmically exploited to propose network structures, which are consistent with a set of dependencies and independencies between variables determined from data (see Section 7). Unfortunately, many alternative structures may satisfy the same set of dependencies and independencies, giving rise to so-called Markov equivalence classes. Further structure disambiguation to unravel causal relationships may require devising proper experiments ("manipulations").

6 Feature relevance in Bayesian networks

In this section we relate notions of non-causal feature relevance introduced in Section 4 with Bayesian networks introduced in Section 5.5. Strongly relevant features in the Kohavi-John sense are found in the Bayesian network DAG in the immediate neighborhood of the target, but are not necessarily strongly causally relevant. These considerations will allow us in Section 6.3 to characterize various cases of features called relevant according to different definitions.

6.1 Markov blanket

Pearl [28] introduced the notion of Markov blanket in a Bayesian network as the set of nodes shielding a given node from the influence of the other nodes (see Figure 6). In other words, a Markov blanket *d-separates* a node from the other nodes, which are not in the Markov blanket.

Formally, let $\mathbf{X} \cup Y$ ($Y \notin \mathbf{X}$) be the set of all variables under consideration and $\mathbf{V}$ a subset of $\mathbf{X}$. We denote by "\" the set difference.

Markov Blanket A subset $\mathbf{M}$ of $\mathbf{X}$ is a Markov blanket of Y *iff* for any subset $\mathbf{V}$ of $\mathbf{X}$, Y is independent of $\mathbf{V} \backslash \mathbf{M}$ given $\mathbf{M}$ (*i.e.* $Y \perp \mathbf{V} \backslash \mathbf{M} | \mathbf{M}$, that is $P(Y, \mathbf{V} \backslash \mathbf{M} | \mathbf{M}) = P(Y|\mathbf{M})P(\mathbf{V} \backslash \mathbf{M}|\mathbf{M})$ or for $P(\mathbf{V} \backslash \mathbf{M}|\mathbf{M}) > 0$, $P(Y|\mathbf{V} \backslash \mathbf{M}, \mathbf{M}) = P(Y|\mathbf{M})$) [26].

Markov blankets are not unique in general and may vary in size. But, importantly, any given causal Bayesian network satisfying the CMC and CFC (see Section 5.5), has a unique Markov blanket, which includes its direct causes (parents), direct effects (children), and direct causes of direct effects (spouses) (see *e.g.* [27, 26]). Note that the Markov blanket does not include direct consequences of direct causes (siblings), and direct causes of direct causes (grand-parents). To understand the intuition behind Markov blankets, consider the example of Figure 6 in which we are looking at the Markov blanket of the central node "lung cancer":

- **Direct causes (parents):** Once all the direct causes have been given, an indirect cause (e.g. "anxiety") does not bring any additional information. In Figure 6-d for instance, increased "anxiety" will eventually increase "smoking" but not influence directly "lung cancer", so it suffices to use "smoking" as a predictor, we do not need to know about "anxiety". Similarly, any consequence of a direct cause (like "other cancers" in Figure 6-c, which is a consequence of a "genetic factor") brings only indirect evidence, but no additional information once the direct cause "genetic factor" is known. Direct causes in faithful distributions are individually predictive, but may otherwise need to be known jointly to become predictive (see *e.g.* the example of figure 3-b: the chessboard/XOR problem).
- **Direct effects (children)** and **direct causes of direct effects (spouses)**: In faithful distributions, direct effects are always predictive of the target. But their predictive power can be enhanced by knowing other possible causes of these direct effects. For instance, in Figure 6-a "allergy" may cause "coughing" independently of whether we have "lung cancer". It is important to know of any "allergy" problem, which would eventually *explain away* that "coughing" might be the result of "lung cancer". Spouses, which do not have a direct connecting path to the target (like "allergy"), are not individually predictive of

the target ("lung cancer"), they need a common child ("coughing") to become predictive. However, in unfaithful distributions, children are not necessarily predictive without the help of spouses, for example in the case of the chessborad/XOR problem.

Following [34], we interpret the notion of Markov blanket in faithful distributions in terms Kohavi-John feature relevance as follows:

1. **Irrelevance:** A feature is irrelevant if it is disconnected from Y in the graph.
2. **Strong relevance:** Strongly relevant features form Markov blanket $\mathbf{M}$ of Y.
3. **Weak relevance:** Features having a connecting path to Y, but not belonging to $\mathbf{M}$, are weakly relevant.

These properties do not require that the Bayesian network be causal. The first statement interprets Definition4 (irrelevance) in terms of disconnection to Y in the graph. It follows directly from the Markov properties of the graph. The second statement casts Definition4 (strong relevance) into the Markov blanket framework. Only strongly relevant features cannot be omitted without changing the predictive power of $\mathbf{X}$. Therefore, non strongly relevant features can be omitted without changing the predictive power of $\mathbf{X}$. Hence the set $\mathbf{M}$ of all strongly relevant features should be sufficient to predict Y, regardless of the values $\mathbf{v}$ assumed by the other features in $\mathbf{X}\backslash\mathbf{M}$: $P(Y|\mathbf{M}) = P(Y|\mathbf{M}, \mathbf{X}\backslash\mathbf{M} = \mathbf{v})$. Therefore, following Definition 6.1, $\mathbf{M}$ is a Markov blanket. Markov blankets are unique for faithful distributions (see [27]), which ensures the uniqueness of the set of strongly relevant features for faithful distributions.

The interpretation of the Markov blanket as the set of strongly relevant variables, which is valid for all Bayesian networks, extends to *causal* Bayesian networks. This means that strongly relevant in the Kohavi-John sense include direct causes (parents), direct effects (chidren), and direct causes of the direct effects (spouses). Yet only direct causes are *strongly causally relevant* according to our Definition 5.4. Consequently, from Definition 5.4 *weakly causally relevant* features coincide with indirect causes, which are ancestors in the graph (excluding the parents).

6.2 The problem of making measurements

When we presented the notion of systemic causality 5.1, and later the framework of Bayesian networks 5.5, we always assumed that there is a clear notion of "inside" and "outside" of the system under study. Practically, the random variables characterizing our system of interest are the result of measurements made by instruments. Ideally, these instruments should not introduce additional "inside" undesired variables, but they sometimes do. We call those artifacts. They are not part of the system of interest, but cannot be separated from it for practical reasons. Hence, even though our feature selection algorithm finds them "relevant", we cannot call them "relevant". In fact, they are often referred to in the statistics literature as "nuisance" variables.

Fortunately, an analysis in terms of causal relationships between features found "relevant" may help us identify potential artifacts, as we will see in the following section.

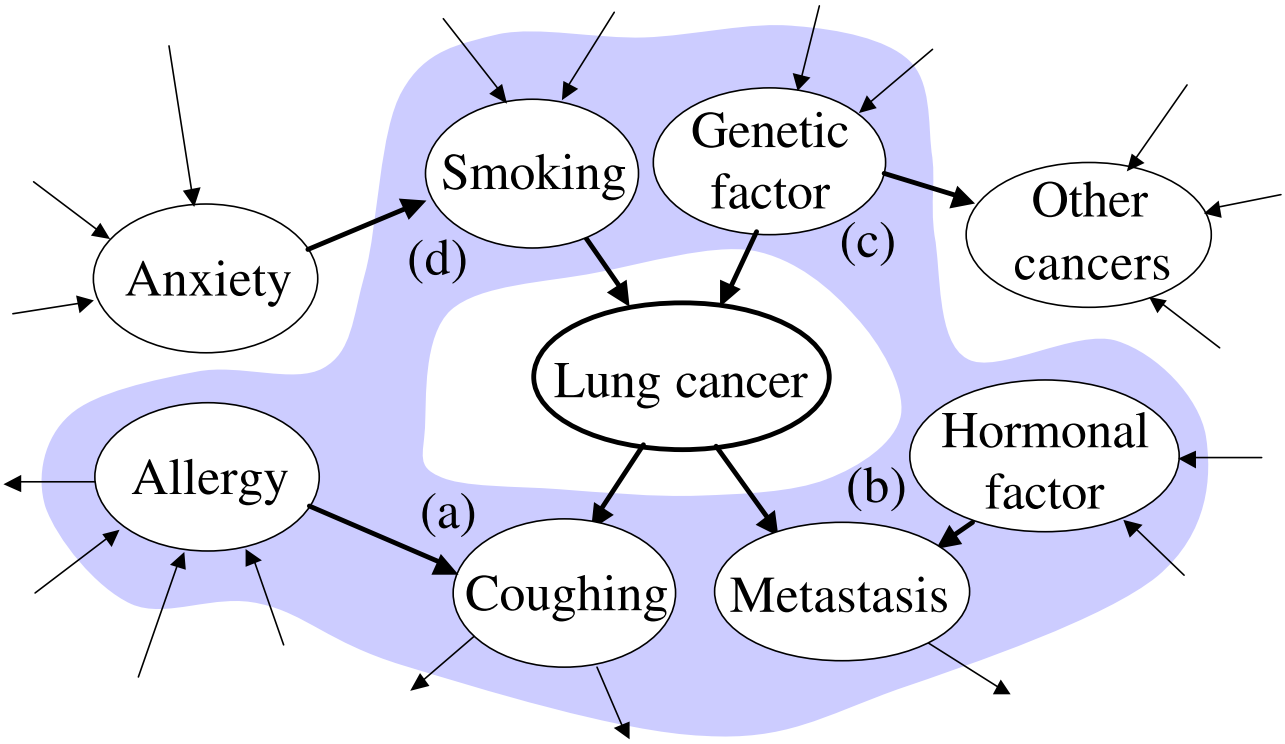


Figure 6: **Markov blanket.** The central node represents a disease of interest, which is our target of prediction. The nodes in the shaded area include members of the Markov blanket. Given these nodes, the target is independent of the other nodes in the network. The letters identify local three-variable causal templates: (a) and (b): colliders, (c) fork, and (d) chain.

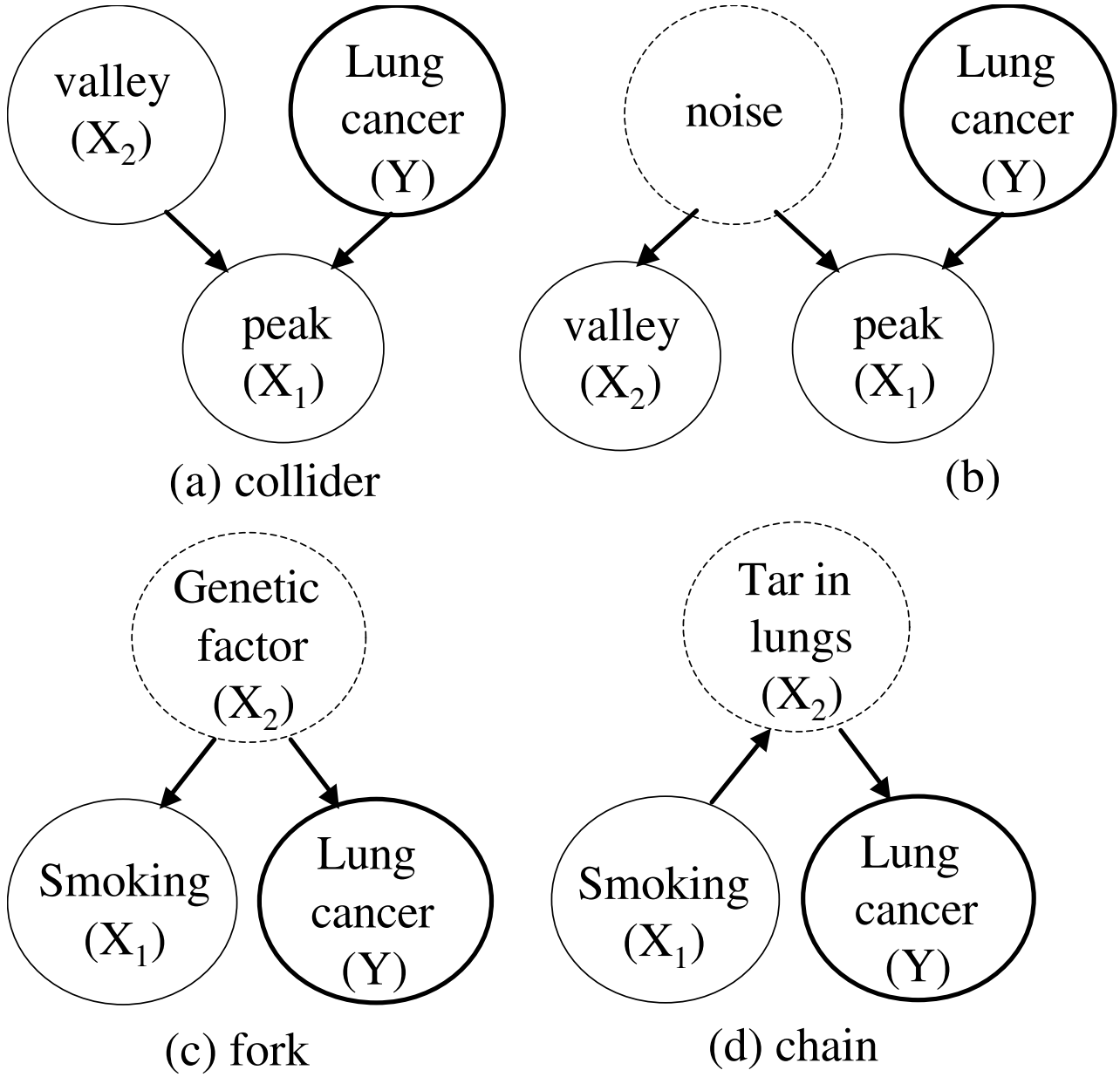


Figure 7: **Feature relevance scenarios.** In all examples, X_1 and X_2 are two potentially relevant features and Y is the target. **Artifacts:** In (a) and (b) variable X_2 represents an experimental artifact. Those are spurious causes of the desired effect X_1 of Y due to measurement errors. **Confounder:** In (c), if variable X_2 (genetic factor) is not known, observing a correlation between X_1 (smoking) and Y (lung cancer) might yield us to draw a wrong conclusion about a possible causal relationship between X_1 and Y. **Indirect effect:** in (d) Y (lung cancer) is an indirect effect of X_1 (smoking) via the intermediate variable X_2 (tar in lungs). (c) and (d) may provide two alternative explanations of the same data (see text).

6.3 Characterizing features selected with classical methods

In this section, we review the small artificial examples we used previously in Section 3 to illustrate multivariate cases of feature-target dependencies. We used these examples to caution against univariate filter methods, and made an argument in favor of multi-variate methods. We now show that a finer analysis in terms of causal dependencies may shed a new light on the notion of feature relevancy. We limit ourself to the analysis of variables, which are in the immediate proximity of the target, namely members of the Markov blanket (MB) and some features in the immediate proximity of the MB (Figure 6). We analyze several scenarios illustrating some basic three-variable causal templates: chain, forks, and colliders. This allows us to refine the notion of feature relevance into:

- Direct cause (parent).
- Unknown direct cause (absent parent called *confounder*, which may result in mistaking a sibling for a parent).
- Direct effect (child).
- Unknown direct effect (which may cause *sampling bias* and result in mistaking a spouse for a parent).
- Other truly relevant MB members (spouses).
- Nuisance variables member of the MB (also spouses).

6.3.1 Upstream chain and fork patterns

Let us first examine the roles played by variables directly connected to parents of the target, including grand-parents and siblings. Those are involved in *chain* and *fork* patterns, featuring the type of dependencies: $Y \not\perp X_1$ and $Y \perp X_1|X_2$. Only parents are part of the Markov blanket and should, in principle, be considered "strongly relevant". The study of some examples allow us to understand the relevance of grand-parents and siblings, as well as potential confusions between siblings or grand-parents and parents.

Since grand-parents and siblings are not in the Markov blanket, we have noted before that they are in principle not useful to make predictions. In our "lung cancer" example of Figure 6 (c) and (d) the direct causes ("smoking" and "genetic factor") are *strongly relevant* (in the Markov blanket). Indirect causes and consequences of causes are only *weakly relevant* (outside the Markov blanket). We argue that siblings and grand-parents are nevertheless worthy of attention. In particular, if the direct causes are not controllable (cannot be acted upon), it may be interesting to look at indirect causes (*e.g.* reducing "anxiety" might indirectly reduce the risk of "lung cancer"). Consequences of direct causes are also interesting for a different reason: they might weaken the relevance of strongly relevant features. In our example, the fact that the "genetic factor" of interest causes not only "lung cancer" but also "other cancers" makes it a non specific maker of "lung cancer".

An example of distribution that would satisfy the type of independence relations of chains and forks ($Y \not\perp X_1$ and $Y \perp X_1|X_2$) is depicted in Figure 3(d). We have already examined this example in Section 3.2: An apparent dependency between X_1 and Y may vanish if we introduce

a new variable X_2 (Simpson's paradox). Importantly, the pattern of dependencies does not allow us to determine whether we have a chain or a fork, which prevents us from distinguishing grand-parents and siblings. This can sometime be resolved using dependencies with other variables, higher order moments of the distribution, experiments, or prior knowledge (see Section 7).

The ambiguity between forks and chains is at the heart of the correlation *vs.* causation problem. If the "true" parents are not known, grand-parents become the most direct identifiable causes. However, if one cannot distinguish between sibling and grand-parents, we may falsely think that siblings are causes. This is illustrated by hypothetical scenarios in Figure 7 (c) and (d). The question is whether "smoking" (variable X_1) is a cause of "lung cancer" (target Y), given that there may be other unknown factors. Health officials issued new restrictions on smoking in public places based on the correlation between smoking and lung cancer. Tobacco companies argued that there may be a common genetic factor causing both craving for nicotine and therefore smoking and predispositions to get lung cancer (Figure 7-c). Such confounding factor has not been identified to date.

Direct causation is always relative to the set of variables considered. A parent may become a grand-parent if another variable is added (Figure 7-d). Indirect causes have sometime more practical importance than direct causes, because direct causes are often not controllable (like the amount of tar in lungs).

Downstream colliders

Patterns of dependencies ($X_2 \perp Y$, $X_2 \not\perp Y|X_1$), are characteristic of *colliders*. Both children and spouses are involved in such patterns, which are found downstream of the target. As explained before, both children and spouses are members of the Markov blanket, and as such they are "strongly relevants" in the Kohavi-John sense for faithful distributions.

We need to emphasize that children and spouses are not "causally" relevant, in the sense that manipulating them does not affect the target. Yet, they may be used to make predictions for stationary systems, or as predictors of the effect of manipulations of the target (*e.g.* the effect of a treatment of "lung cancer"). We previously noted that "allergy" is a useful complement of "coughing" to predict "lung cancer", because knowing about allergy problems allows the doctor to "explain away" the fact that coughing may be the symptom of "lung cancer". Now, after a patient receives a treatment for "lung cancer" (manipulation), a reduction in "coughing" may be an indication of success of the treatment.

Two cases of distributions corresponding to colliders are shown in Figures 3 (a) and (b). One corresponds to a faithful case and the other to an unfaithful case (chessboard). In either case, spouses can be useful complements of children to improve prediction power. Nonetheless, we now want to caution against two types of problems that may be encountered: sampling bias and artifacts.

In Figure 6 (b) we show a scenario of "sampling bias" in the subgraph: $Lungcancer \rightarrow Metastasis \leftarrow Hormonalfactor$. The presence of metastases may be unknown. It may turn out that all the patients showing up in the doctor's office are more likely to have late stage cancer with metastases because only then do they experiment alarming fatigue. In this situation, the sample of patients seen by the doctor is biased. If from that sample a correlation between a certain hormonal factor and lung cancer is observed, it may be misinterpreted as causal. In reality, the dependency may only be due to the sampling bias. "Hormonal factor" (playing the

role of X_2) cannot be used as a predictive factor without knowing about the "Metastasis" factor (playing the role of X_1) , and wrong results could be inferred if that factor is unknown.

To illustrate the problem of artifacts, we take an example, which happened in real world situations we were confronted with, illustrated here with artificial data for increased clarity. The example corresponds to the scatter plot of Figure 3 (a) and the causal graph of Figure 7 (a). The target Y is the health status of a patient "lung cancer". We are using a spectrometer to analyse the patient's serum. Each position in the spectrum is one of several tens of thousands of variables. Our feature selection algorithm identifies two most promising complementary features: X_1 positioned near a peak in the spectrum and X_2 positioned near a valley (Figure 3-a). Further analysis indicates that the peak (feature X_1) might correspond to the abundance of a protein in serum; it is individually predictive of the health status. The valley (feature X_2) is uncorrelated with the target, but taken together with X_1, it provides better predictions. We realize that a plausible explanation is that there is an unknown "noise" variable due to the measuring instrument (Figure 3-b). Feature X_2 is an indirect measurement of that noise, which increases or lowers the baseline of the signal, *i.e.* variables corresponding to neighboring values in the signal have a similar offset. Subtracting the baseline helps improving the signal. Therefore feature X_2 proves to be useful for this particular instrument because it provides a good local estimate. However it can hardly be qualified as "relevant", since its relevance is a consequence of a measurement error and is not related to the target.

7 Causal discovery algorithms

In previous sections, we have motivated the introduction of the concept of causality in feature selection. It has long been thought that causal relationships can only be evidenced by "manipulations", as summarized by the motto commonly attributed to Paul Holland and Don Rubin: "no causation without manipulation". For an introduction on manipulation methods of inferring causation, see for instance [30]. Yet, in the recent years much fruitful research has been devoted to inferring causal relationships from "observational data", that is data collected on a system of interest, without planned experiments or intervention. Current textbooks reviewing these techniques include [11, 28, 32]. We collectively refer to the algorithms as "causal discovery machine learning methods".

All causal discovery machine learning methods have two fundamental components: a language for expressing causal models, and an inductive bias that prefers certain kinds of causal models over others. Our brief presentation of such algorithms uses discrete causal Bayesian networks as the language in which to represent causal relationships, and a special kind of distribution (i.e., Faithful distributions) as the main inductive bias of these algorithms. We rely on the definitions given in previous sections, including those of conditional independence (Definition 3), Bayesian network (Definition 5.5), causal Bayesian network (Definition 5.5), faithful Bayesian network (Definition 5.5), faithful distribution (Definition 5.5), d-separation and d-connection (Definition 5.6), and Markov blanket (Definition 6.1).

7.1 Task definition and assumptions made

Learning a Bayesian network $\{\mathbf{X}, \mathcal{G}, P\}$ from data consists in two subtasks, sometimes performed jointly, sometimes in sequence: learning the structure of the graph G and learning the probability

distribution P. From the point of view of causal discovery and feature selection, learning the structure of the graph is the subtask of interest.

For the purpose of this short introduction to algorithms of causal discovery, we will assume that the set $\mathbf{X}$ of variables is self *sufficient* to characterize all causal dependencies of interest. The definition of sufficiency is often vague, but here we can give it a precise meaning, if we assume that the variables may be selected among a set $\mathbf{S}$ of indicator variables representing all possible events in our system under study (see Section 5.4). Our definition follows closely [28].

Causal sufficiency. A set of indicator random variables $\mathbf{X} = [X_1, X_2, ...X_N] \in \mathbf{S}$ is said to be causally self sufficient if no set of two or more variables in $\mathbf{X}$ possess a common cause in $\mathbf{S}\backslash\mathbf{X}$.

To infer causal relationships in a set of variables $\mathbf{X}$, which is not causally sufficient is more difficult since there may be dependencies between variables, which cannot causally be explained from variables in $\mathbf{X}$, but from "confounders" in $\mathbf{S}\backslash\mathbf{X}$. This can be modeled with extra "latent" variables modeling the unknown factors. Readers interested in these more advanced problems are referred to textbooks, including [28].

In what follows, we will make the following set of "causal discovery assumptions":

- The set of variables considered $\mathbf{X}$ is causally self sufficient.
- The learner has access to a sufficiently large training set and reliable statistical tests for determining conditional dependencies and independencies in the original distribution where the data is sampled from.
- The process that generated the data having the distribution $P(\mathbf{X}, Y)$ can be faithfully represented by the family of models under consideration (for this report, we limit ourselves to causal Bayesian networks).
- The hypothesis space is the space of all possible models under consideration (here, causal Bayesian networks $\{\mathbf{X}, \mathcal{G}, P\}$).

7.2 A prototypical causal discovery algorithm

We outline here the fundamental operation of the PC algorithm (barring speed-up techniques and implementation details in order to simplify the presentation; see [32] for a complete description). Under the causal discovery assumptions stated above, this algorithm is provably sound in the large sample limit [32], in the sense that it can recover the structure of a Bayesian network that generated the data, up to a Markov equivalence class.

The algorithm begins with a fully connected un-oriented graph and has three phases:

Algorithm: PC

Let A, B, and C be variables in $\mathbf{X}$ and $\mathbf{V}$ any subset of $\mathbf{X}$.
Initialize with a fully connected un-oriented graph.

1. Find direct edges by using the criterion that variable A shares a direct edge with variable B *iff* no subset of other variables $\mathbf{V}$ can render them conditionally independent ($A \perp B|\mathbf{V}$).

2. Orient edges in "collider" triplets (i.e., of the type: $A \rightarrow C \leftarrow A$) using the criterion that if there are direct edges between A, C and between C, B, but not between A and B, then $A \rightarrow C \leftarrow B$, *iff* there is no subset $\mathbf{V}$ containing C such that $A \perp B|\mathbf{V}$.

3. Further orient edges with a constraint-propagation method by adding orientations until no further orientation can be produced, using the two following criteria:

   - If $A \rightarrow B \rightarrow ... \rightarrow C$, and $A - C$ (*i.e.* there is an undirected edge between them A and C) $A \rightarrow C$.
   - If $A \rightarrow B - C$ then $B \rightarrow C$.

Without going into details we note that all of the causal discovery assumptions can be relaxed via a variety of approaches. For example, if the causal sufficiency property does not hold for a pair of variables A and B, and there is at least one common parent C of the pair that is not measured, the PC algorithm might wrongly infer a direct edge between A and B. The FCI algorithm addresses this issue by considering all possible graphs including hidden nodes (latent variables) representing potential unmeasured "confounders", which are consistent with the data. It returns which causal relationships are guaranteed to be unconfounded and which ones cannot be determined by the observed data alone. The FCI algorithm is described in detail in [32].

7.3 Markov Blanket induction algorithms

From our previous discussion it follows that one can apply the PC algorithm (or other algorithm that can learn high-quality causal Bayesian networks) and extract the Markov Blanket of a target variable of interest Y. However when the dataset has tens or hundreds of thousands of variables, or when at least some of them are highly interconnected, applying standard causal discovery algorithms that learn the full network becomes impractical. In those cases, local causal discovery algorithms can be used, which focus on learning the structure of the network only in the immediate neighborhood of Y.

Two efficient Markov blanket discovery algorithms have been recently proposed. Aliferis et al. have introduced HITON [3] and Tsamardinos et al. have described MMMB [34]. Both algorithms find direct edges to Y (without direction), making a distinctions between parents or children and spouses, therefore providing more information that a mere list of strongly relevant features.

The same induction step than PC is used to find edges (*i.e.* , X_i shares a direct edge with Y *iff* there is no subset $\mathbf{V}$ of the variables set $\mathbf{X}$ such that: $X_i \perp Y|\mathbf{V}$). One difference is that PC starts with a fully connected graph, while HITON and MMPB start with an empty graph. Aside from limiting the search to finding edges to Y and searching for "spouses" of Y, which already represents a significant computational speedup compared building an entire Bayesian network, HITON and MMPB use a number of heuristics to accelerate the search, which prove to be very efficient in practice. These include limiting the search to conditioning sets of sizes permitting the sound estimation of conditional probabilities, and heuristics of sorting the variables that are candidate parent/chidren of Y. The two algorithms use otherwise different heuristics to prune the search an perform sanity checks. These algorithms scale well for large datasets with up to 10^5 thousands of variables. In published experiments, HITON and MMMB have been

compared to Koller-Sahami, GS, and IAMB family algorithms (see Section 9 for more details on computational considerations). Both algorithms work well, but HITON eliminates some false positives that may be output by MMMB and has been tested specifically for feature selection for classification with very promising results (see Section 8 for examples of application).

A major novelty of local methods is circumventing non-uniform graph connectivity. A network may be non-uniformly dense (or sparse). In a global learning framework, if a region is particularly dense that region cannot be discovered fast and when learning with small sample it will produce many errors. These errors propagate to remote regions in the network (including those that are learnable accurately and fast with local methods). On the contrary, local methods will be both fast and accurate in the less dense regions. Thus HITON and MMMB compare well to the standard PC algorithm for discovering full Bayesian networks.

Localizing the search for direct edges is desirable according to the previous explanation, but far from obvious algorithmically. A high-level explanation is that when building the parents/children sets around Y in a localized manner we do not incur false negatives and we will occasionally omit variables X_i not connected to Y but connected to other variables X_j which are not parents or children of Y. Variables such as X_i act as "hidden variables" insofar the localized criterion for independence is concerned, thus leaving false positives. It turns out however that (i) the configuration in which this problem can occur is very specific and in practical data the number of false positives minimal; (ii) the false positives can be detected by running the localized criterion in the opposite direction (*i.e.* , seeking the parents/children of X_j in a local fashion). This constitutes the symmetry correction of localized learning of direct edges.

The Causal Explorer software package, including HITON, MMMB, and many other useful causal discovery algorithms, is available from the Internet (see Appendix A).

8 Examples of applications

Causality and feature selection as described in this paper have been used to achieve various objectives in different areas such as bio-informatics, econometrics and engineering. We present below one example from each of these fields that illustrates the use of causal and feature techniques in practice.

With the advent of the DNA microarrays technology [31], biologists have collected the expression of thousands of genes under several conditions. Xing et al. were among the first ones to use a Markov blanket discovery algorithm for feature selection [36] in DNA microarray data, to diagnose disease (two kinds of Leukemia). Friedman and colleagues [9] applied a causal discovery technique on microarray data to build a causal network representing the potential dependencies between the regulations of the genes. If the expression level of one gene causes the up or down regulation of another gene, an edge should link them. A simple feature selection technique based on correlation is first applied to select a set of potential causes for each gene. A causal discovery method is then run on the reduced set of potential causes to refine the causal structure [10]. In this example, feature selection is used to reduce the set of variables that will feed a causal discovery algorithm, the latter being generally computationally expensive. Feature selection in that context can be understood as feature rejection: only irrelevant features should be rejected.

More recently, the Markov blanket discovery algorithm HITON [3] have been applied with sucess to clinical, genomic, structural and proteomic data, and mining the medical literature,

achieving significantly better reduction in feature set size without classification degradation compared to a wide range of alternative feature selection methods. The applications of HITON also include understanding physician decisions and guideline compliance in the diagnosis of melanomas, discovering biomarkers in human cancer data using microarrays and mass spectrometry, and selecting features in the domain of early graft failure in patients with liver transplantations. For a review of the applications of HITON and comparison with other methods, see [4].

In biology and medicine, causal discovery aims at guiding scientific discovery, but the causal relationships must then be validated by experiments. The original problem, *e.g.* the infeasibility of an exhaustive experimental approach to detect and model gene interactions, is addressed using causality by defining a limited number of experiments that should be sufficient to extract the gene regulatory processes. This use of causality is in contrast with our second example where experiments in a closed environment are usually not possible, *i.e.* there is no possible laboratory validation before using the treatment in real situations.

Causality has been used by economists for more than 40 years. Some years before artificial intelligence started to address the topic, Clive Granger [12] defined a notion of temporal causality that is still in use today. In 1921, Wright introduced Structure Equation Modeling (SEM) [35], a model widely known by sociologists and economists. It is therefore singular to see that Marketing Research – a field close to economy and sociology – does not contain much work involving causality. In his review [29], Rigdon explains it as follows: 'At least as far back as Robert Ling's (1982) scathing review of David A. Kenny's (1979) book, *Correlation and Causality*, users of SEM methods have found themselves on the defensive, careful not to claim too much'. The defense of SEM by Pearl [28] might change the status though and causality appears slowly as to be a subject of interest in marketing. From a practical perspective, causality can be directly used to addresses one of the key question that marketers ask: how to assess the impact of promotions on sales? It is known that many potential factors come into play when computing the effect of promotions: weather, word of mouth, availability, special days (e.q. valentine's day), etc. Understanding how these factors influence the sales is interesting from a theoretical point of view but is not the primary objective: what practically matters is what to do next, that is, what will be the effect of promotions versus no promotions next month. This is typically a problem of causal discovery and parameter estimation. Finding the causal link is not enough. It is necessary to know whether the promotion will have a positive effect and how positive it will be in order to compute the expected profit. A promotion that has a small positive effect but costs a lot to implement might not be worth launching. Extracting the true causal structure is also less critical than estimating $P(\text{sales}|\text{do}(\text{promotions}))$. The role of feature selection in such application is not clear as the number of variables is usually rather small and the features are built based on expert knowledge directly. The amount of marketing data is indeed usually restricted because of space and time constraints or because of legal issues.

Failure diagnosis is the last application we shall consider. In diagnosing a failure, engineers are interested in detecting the cause of defect as early as possible to save cost and to reduce the duration of service breach. Bayesian networks and their diagnostic capabilities which are of particular relevance when the links are causal, have been used to quickly perform a root cause analysis and to design a series of tests minimizing the overall cost of diagnosis and repair. Kraaijeveld et al. [21] present an approach that relies on a user-defined causal structure. The latter is simplified to a bi-partite graph having only links from causes on one side to symptoms

on the other. The parameters are either estimated from data or defined by hand using Noisy-OR nodes [8] to ease the entry. Causal inference is then used to infer the most probable causes based on a description of the symptoms. This structure is widely inspired from the Quick Medical Reference tool [24] that was designed in the mid eighties for medical diagnosis. In such application, the use of feature selection and causal discovery is left to the user. It is a seldom process but in such situation, it cannot be avoided. Consider performing a diagnosis of a failure in a pharmaceutical manufacturing processes, one of the failure or symptom is the lack of documentation. The FDA requires that each error in the production line be recorded and documented. When analyzing a particular site, data only is not reliable: lack of data can mean that everything works well and there is no error, or it can be a symptom of the lack of documentation and therefore should be counted as a failure. In such situations, expert knowledge and manual entry are unavoidable and causality techniques are run for the inference part, i.e. finding the most probable cause. They are also used to derive the most efficient treatment or repair to the failure.

These three applications show that causality techniques can be used in different settings with completely different requirements. The interaction with feature selection seems weaker for human driven causal structure discovery but appears to be mandatory to handle the large amount of variables of microarray data. As storage devices get cheaper and more efficient, one might expect to collect more and more variables making the use of feature selection a requirement to apply causal structure discovery techniques in many areas.

9 Discussion, advanced issues and open problems

To our knowledge, Koller and Sahami were the first to argue convincingly about using the Markov blanket for feature selection [20]. Their focus was on optimal large-sample feature selection for classification, and not causality. Kohavi and John showed why many definitions of relevance were undesirable and provided new ones [19]. They also emphasized the dichotomy of filters *vs.* wrappers, arguing in favor of wrappers. Tsamardinos and Aliferis elucidated the link between local causality and feature selection in faithful distributions and identified strongly relevant features in the Kohavi-John sense to members of the Markov Blanket [34]. These authors also proposed that filter algorithms are just implementations of definitions of relevance and using this perspective proved that there cannot be a universally preferred definition of relevance, optimal filter algorithm or universally optimal or superior wrapper algorithm. In this report, we reviewed these concepts of relevance and causal relevance and illustrated them with small examples.

The first two algorithms for Markov Blanket induction by Koller and Sahami and Cooper et al. [20, 7] contained many promising ideas and the latter was successfully applied in real data, however they were not guaranteed to find the actual Markov blanket and nor could they be scaled to thousands of variables. Margaritis [23] subsequently invented a sound algorithm, GS, however it required sample at least exponential to the size of the Markov Blanket and would not scale to thousands of variables in most real datasets with limited samples sizes. Tsamardinos [34] introduced several improvements of GS, that led to the IAMB family of algorithms. The IAMB algorithms are guaranteed to find the actual Markov blanket given enough training data and are more sample efficient than GS, however they still require a sample size exponential in the size of the Markov Blanket. They are highly usable however with up to hundreds of thousands of

variables when the Markov blanket is small relative to the available sample. Finally [3]and [34] introduced algorithms (HITON and MMMB) that return the local causal neighborhood and the Markov Blanket without requiring sample exponential to the size of the local neighborhood or the Markov blanket.

In writing this report we touched upon a number of issues connecting feature selection and causal discovery, with the purpose of triggering the interest of researchers in machine learning. The apparent coherence of this report should not be deceiving. A lot of complex issues were only touched upon and some were not addressed at all. In spite of enormous advances made recently there is to date no unified framework to address all the aspects of causal discovery. The interested reader is therefore encouraged to pursue his reading, starting perhaps with [11, 28, 32, 25]. We emphasize in particular that Bayesian networks provide a convenient way to reason about causal relationships, but have several limitations, including that they are not well fit to describing unfaithful distributions (or which the XOR problem is the archetypical example), they require estimating probability distributions or densities (which is notoriously difficult with small samples and large numbers of features), cannot represent certain counterfactual situations (see *e.g.* [28] p. 57), and, in their usual form, do not represent time. Bayesian models have been successful in biomedical applications [16, 9] and are now making their way in risk assessment and failure diagnosis. But, in econometrics and the social sciences, structural equation models have been most prominently used [17, 18]. Granger's causality is another type of approach geared towards the analysis of time series [12].

Despite its limitations, the framework of Bayesian networks is powerful and well developed [11, 28, 32, 25]. Tools like the do-calculus open new doors by allowing to predict the consequence of actions or manipulations, without requiring any actual experimentation.

Concerning the definition of causality, we provided a definition based on manipulations, which is convenient from an engineering viewpoint. However, philosophically, this type of definition can be challenged because it requires defining an "inside" and an "outside" of a system, the notion of agent, which can carry out the manipulation, and the notion of "manipulation" as primitive concepts. However, the "manipulation" ought to be itself a postulated causal mechanism, and therefore the definition of causality is recursive. We have seen its limits already when we pointed out the problem of the interference of measuring instruments. We can always push the limits of the system and include more mechanisms in the system (the measuring instruments, the agents), but in the limit, when the system includes the whole universe, who will test it? As an additional challenge of the definition of causality via manipulations, the criterion admits hypothetical and counter-factual situations in order to accommodate inducing causal models from observations of unique phenomena or phenomena non-manipulatable by available mechanisms. Such phenomena for example include evolutionary selection, causal attribution in deceased patients, non-manipulatable entities such as celestial body positions, etc.

Regarding the definition of *causal feature relevance*, we have extended the Kohavi-John notions of strong and weak relevance, which conducted us to identify direct causes as strongly causally relevant and indirect causes as weakly causally relevant. Obviously, depending on the application, the strength of relevancy may have different meanings. For instance, the root cause or a bottleneck cause may be more relevant than any direct cause.

We have briefly mentioned the problem of unknown variables, confounders and artifact. In practical applications, one often does not know which variables of actual interest should be measured and which nuisance variable should be monitored. We believe that the analysis of

observational data with causal feature selection tools can help guiding the collection of new data and designing proper experiments.

With respect to causal discovery algorithms, we only glossed over one particular type of approach to learn the structure of causal networks, which are constructive methods using conditional independencies. Other approaches include methods performing a search in structure space to globally optimize a fitness function (see *e.g.* [15]). When only two variables are under study, conditional independence criteria cannot be used to infer causal direction. Recently, a new method has been proposed to solve the two-variable problem using information-theoretic criteria [33].

The definitions of causality, which we proposed, assume that the probability distributions are known and are therefore only valid in the infinite sample size limit. When dealing with finite samples, decisions about *e.g.* $P(Y = y|do(X_i) = x, do(\mathbf{V}^{\setminus i}) = \mathbf{v}) \neq P(Y = y|do(\mathbf{V}^{\setminus i}) = \mathbf{v})$ will occasionally be wrong. Much remains to be done to study the statistical complexity of the problem of causal feature selection.

Given faithful distributions and reliable statistical tests of conditional independence for the given sample size and data, the IAMB algorithm (and all members of the IAMB family [34]) returns the exact Markov Blanket (for a proof, see [34]). With similar assumptions, HITON-PC [3] returns exactly direct causes and direct effects of the target variables. MMMB [34] returns all members of the Markov Blanket of the target and some false positives, which may be eliminated with various techniques, if needed. The worst case asymptotic complexity of IAMB is $O(N^2)$, where N is the number of variables/features, and in published experiments the average case complexity is $O(MB\ N)$ conditional independence tests, where MB is the size of the Markov blanket [34]. Conditional independence tests have complexities that vary from test to test, however linear-time tests have been presented in the literature [34]. parallelizing IAMB reduces computation time by a factor of k, where k is the number of available processors. MMPC and HITON-PC have complexity $O(NCPC^{\ell+1})$ tests, where CPC is the largest set of parents and children built during the operation of the algorithm and ℓ bounds the conditioning set size as a result of finite sample size. In practice the average time complexity is $O(NPC^{\ell+1})$, where PC is the number of number of edges in the graph. The complexity of MMMB and HITON-MB in practice is of the order of PC more expensive than HITON-PC, with the typical optimizations and heuristics employed in these algorithms. In general we note that Markov Blanket and direct causes/direct effects algorithms are worst-case exponential, but in practice in both benchmark and real-life datasets with up to $100,000$ variables the algorithms run very fast (from minutes to a few hours on a single CPU).

Feature selection algorithms are often used as filters before causal discovery algorithms are applied, as a heuristic to reduce the computational and/or statistical complexity of causal discovery. Practitioners who employ such hybrid methods capitalize on the power of recent feature selection algorithms, which are capable of handling very large space dimensions, often in excess of 10^5 variables [14]. This may make practical sense and may yield better results than the straight application of causal discovery algorithms in particular cases. For example, strong univariate predictors are typically selected and reported in relation to disease phenotype in analysis of gene expression microarray data. It is hoped that such strong predictors are members of the gene regulatory pathways that are primarily responsible for the corresponding disease (phenotype) and that these highly associated genes can be subsequently used as promising targets for developing new drugs for the disease. As explained in [2], this heuristic is often harmful to causal discovery;

nevertheless we have to recognize that feature selection is a kind of de facto standard for inducing causal hypotheses from massive datasets. Recent theoretical results connect causal to predictive feature selection and provide an emerging framework for studying causality and feature selection [34]. This framework enables us to understand why popular and successful feature selection methods will, in some cases, introduce redundant features. In addition, recent algorithmic developments allow discovery of the causal features, and causal graphs in very high-dimensional spaces in a computationally and sample-efficient manner [3, 34].

These exciting new directions open many new roads for future research, including:

1. Characterizing all major existing and novel causally and non-causally-motivated feature selection methods in terms of causal validity using theoretical and empirical approaches.

2. Developing appropriate metrics, research designs, benchmarks etc. to empirically study the performance and *pros* and *cons* of causal *vs.* non-causal feature selection methods.

3. Studying the concept of relevancy and its relationship with causality beyond faithful distributions and beyond Kohavi-John relevancy.

4. Further improving the computational performance and accuracy of causal feature selection methods for large dimensional problems and small samples sizes.

5. Developing a theory of the statistical complexity of learning causal relationships.

6. Developing powerful and versatile software environments for causally-oriented feature selection.

7. Examining the validity of and relaxing assumptions motivated by efficiency or convenience (*e.g.* faithfulness, causal sufficiency, normality of distributions, linearity of relationships) when applied to real world feature selection situations.

10 Summary and conclusion

Feature selection focuses on uncovering subsets of variables $X_1, X_2, ...$ predictive of a target Y. In light of causal relationships, the notion of variable relevance can be refined. In particular, causes are better targets of action of external agents than effects: if X_i is a cause of Y, manipulating it will have an effect on Y, not if X_i is a consequence (or effect). In the language of Bayesian networks, direct causes (parents), direct effects (children), and other direct causes of the direct effects (spouses) are all members of the Markov blanket. The members of the Markov blanket are strongly relevant in the Kohavi-John sense, for faithful distributions. Direct causes are strongly causally relevant. Spouses are not individually relevant, but both parents and children are, in faithful distributions. Both causes and consequences of Y are predictive of Y, but consequences can sometimes be "explained away" by other causes of the consequences of Y. So the full predictive power of children cannot be harvested without the help of spouses. Causes and consequences have different predictive power when the data distribution changes between training and utilization time, depending on the type of change. In particular, causal features should be more predictive than consequential features, if new unknown "noise" is added to the variables $X_1, X_2, ...$ (the co-variate shift problem). If new unknown noise is added to

Y however, consequential variables are a better choice. Unknown features, including possible artifacts or confounders, may cause the whole scaffold of causal feature discovery to fall apart if their possible existence is ignored. Causal feature selection method can assist the design of new experiments to disambiguate feature relevance.

A Code availability

The Causal Explorer library [4] provides an integrated API callable from C and Matlab that can be used for research experimentation as well as data analysis with causal discovery and causally-motivated feature selection algorithms. Full causal graph learning algorithms in Causal Explorer include the Greedy Search-and-Score, IC*, OR, TPDA, MMHC, PC, and Sparse Candidate algorithms. Markov Blanket and Local neighborhood learning algorithms include HITON, MMMB, IAMB, Grow-Shrink, Koller-Sahami, and LCD2. Causal Explorer also provides a tiling algorithm that enables construction of "tiled" Bayesian Networks so that experimenters can test feature selection and causal discovery algorithms for their scalability. HITON is also implemented in the GEMS and FAST-AIMS systems for the automated analysis of microarray and Mass Spectrometry data respectively.

Causal Explorer can be downloaded from: `www.dsl-lab.org`.

References

[1] *Special issue on relevance*, Artificial Intelligence **97** (1997), no. 1-2.

[2] C. F. Aliferis, A. Statnikov, and I. Tsamardinos, *Challenges in the analysis of mass-throughput data: A technical commentary from the statistical machine learning perspective*, Cancer Informatics **2** (2006), 133–162.

[3] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, *HITON, a novel Markov blanket algorithm for optimal variable selection*, 2003 American Medical Informatics Association (AMIA) Annual Symposium, 2003, pp. 21–25.

[4] C. F. Aliferis, I. Tsamardinos, A. Statnikov, and L.E. Brown, *Causal explorer: A probabilistic network learning toolkit for biomedical discovery*, 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS) (Las Vegas, Nevada, USA), CSREA Press, June 23-26 2003, `http://discover1.mc.vanderbilt.edu/discover/public/causal_explorer/`.

[5] A. Blum and P. Langley, *Selection of relevant features and examples in machine learning*, Artificial Intelligence **97** (1997), no. 1-2, 245–271.

[6] J. Quiñonero Candela, N. Lawrence, A. Schwaighofer, and M. Sugiyama, Organizers, Learning when test and training inputs have different distributions (Whistler, Canada), NIPS workshop, December 2006, `http://ida.first.fraunhofer.de/projects/different06/`.

[7] G.F. Cooper and E. Herskovits, *A bayesian method for the induction of probabilistic networks from data*, Machine Learning **9** (1992), no. 4, 309–347.

[8] Francisco J. Dez, *Parameter adjustment in bayes networks. the generalized noisy or-gate*, Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (San Mateo, CA, USA), Morgan Kaufmann, 1993, pp. 99–105.

[9] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, *Using bayesian networks to analyze expression data*, RECOMB, 2000, pp. 127–135.

[10] N. Friedman, I. Nachman, and D. Pe'er, *Learning bayesian network structure from massive datasets: The sparse candidate algorithm*, Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI '99), 1999, pp. 196–205.

[11] C. Glymour and G.F. Cooper, Editors, *Computation, causation, and discovery*, AAAI Press/The MIT Press, Menlo Park, California, Cambridge, Massachusetts, London, England, 1999.

[12] C.W.J. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica **37** (1969), 424–438.

[13] I. Guyon, S. Gunn, A. Ben Hur, and G. Dror, *Result analysis of the nips 2003 feature selection challenge*, Advances in Neural Information Processing Systems 17 (Cambridge, MA), MIT Press, 2005, to appear.

[14] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Editors, *Feature extraction, foundations and applications*, Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.

[15] D. Heckerman, *A tutorial on learning with bayesian networks*, 1995.

[16] Edward H. Herskovits and Azar P. Dagher, *Application of bayesian networks to health care.*

[17] D. Kaplan, *Structural equation modeling: Foundations and extensions*, Advanced Quantitative Techniques in the Social Sciences series, vol. 10, SAGE, 2000.

[18] R. B. Kline, *Principles and practice of structural equation modeling*, The Guilford Press, 2005.

[19] R. Kohavi and G. John, *Wrappers for feature selection*, Artificial Intelligence **97** (1997), no. 1-2, 273–324.

[20] D. Koller and M. Sahami, *Toward optimal feature selection*, 13th International Conference on Machine Learning, July 1996, pp. 284–292.

[21] Pieter C. Kraaijeveld and Marek J. Druzdzel, *Genierate: An interactive generator of diagnostic bayesian network models*, 16th International Workshop on Principles of Diagnosis (Monterey, California, USA), 2005.

[22] H. Liu and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*, Kluwer Academic, 1998.

[23] D. Margaritis and S. Thrun, *Bayesian network induction via local neighborhoods*, Technical Report CMU-CS-99-134, Carnegie Mellon University, August 1999.

[24] Randolph A. Miller, Melissa A. McNeil, Sue M. Challinor, Jr Fred E. Masarie, and Jack D. Myers, *The internist-1/quick medical reference projectstatus report*, The Western Journal of Medicine **145** (1986), no. 6, 816–822.

[25] R. E. Neapolitan, *Learning bayesian networks*, Prentice Hall series in Artificial Intelligence, Prentice Hall, 2003.

[26] R.E. Neapolitan, *Probabilistic reasoning in expert systems: Theory and algorithms*, John Wiley and Sons, 1990.

[27] J. Pearl, *Probabilistic reasoning in intelligent systems*, Morgan Kaufman, San Mateo, California, 1988.

[28] Judea Pearl, *Causality: models, reasoning and inference*, Cambridge University Press, March 2000.

[29] E. E. Rigdon, *New books in review. causality: models, reasoning and inference from judea pearl, and causation prediction and search from peter spirtes, clark glymour and richard scheines*, Journal of Marketing Research **39** (2002), 137–140.

[30] Donald Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies*, Journal of Educational Psychology **66** (1974), no. 5, 688–701.

[31] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, *Quantitative monitoring of gene expression patterns with a complementary dna microarray*, Science **270** (1995), no. 5235, 467–470.

[32] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction, and search*, The MIT Press, Cambridge, Massachusetts, London, England, 2000.

[33] X. Sun, D. Janzing, and B. Schlkopf, *Causal inference by choosing graphs with most plausible Markov kernels*, Ninth International Symposium on Artificial Intelligence and Mathematics, 2006.

[34] I. Tsamardinos, C.F. Aliferis, and A. Statnikov, *Algorithms for large scale Markov blanket discovery*, 16th International Florida Artificial Intelligence Research Society (FLAIRS) Conference (St. Augustine, Florida, USA), AAAI Press, May 12-14 2003, pp. 376–380.

[35] S. Wright, *Correlation and causation*, Journal of Agricultural Research **20** (1921), 557–585.

[36] E. P. Xing, M. I. Jordan, and R. M. Karp, *Feature selection for high-dimensional genomic microarray data*, Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2001, pp. 601–608.