Study on: Tools for causal inference from cross-sectional innovation surveys with continuous or discrete variables

Dominik Janzing

June 28, 2016

Abstract

The design of effective public policy, in innovation support, and in other areas, has been constrained by the difficulties involved in understanding the effects of interventions. The limited ability of researchers to unpick causal relationships means that our understanding of the likely effects of public policy remains relatively poor. If improved methods could be found to unpick causal connections our understanding of important phenomena and the impact of policy interventions could be significantly improved, with important implications for future social prosperity of the EU.

Recently, a series of new techniques have been developed in the natural and computing sciences that are enabling researchers to unpick causal links that were largely impossible to explore using traditional methods. Such techniques are currently at the cutting edge of research, and remain technical for the non-specialist, but are slowly diffusing to the broader academic community. Economic analysis of innovation data, in particular, stands to gain a lot from applying techniques developed by the machine learning community. Writing recently in the Journal of Economic Perspectives, Varian (2014, p3) explains "my standard advice to graduate students these days is go to the computer science department and take a class in machine learning. There have been very fruitful collaborations between computer scientists and statisticians in the last decade or so, and I expect collaborations between computer scientists and econometricians will also be productive in the future." However, machine learning techniques have not yet been applied to innvovation policy. The aim of this report is to rise to this challenge.

The reader is warned that the discussion is very technical in places, but the report provides the first example of the application of new methods to generate novel insight with potentially very important policy implications. It is, to our knowledge, the first time that these methods have been applied to innovation data in such a comprehensive way. Moreover, the methods have been applied to a body of data that is almost impossible to analyse causally using the existing methods that are commonly applied in policy evaluation in the EU. While the initial findings suggest that causal processes within firms are complex, some novel connections are indicated by the analysis. In turn, some of these suggest important avenues for future policy analysis to follow. For example, to explore the relative impact of EU, national and local funding, given the limited causal impact

of local funding on in-house R&D expenditure found in the data. For the non-technical reader, we would stress that the ability to pull out causal links in complex, relatively poor quality, cross sectional data is a major achievement and suggests that these methods have significant potential in the policy analyst's tool box. On the other hand, drawing *certain* causal conclusions requires data from randomized interventions, while data from passive observations (as used here) can only provide uncertain, but helpful, hints on causal directions. The reader should always be aware of this fundamental limitation and, accordingly, consider all causal conclusions as preliminary.

The report starts with a brief introduction to the theory of data-driven causal search. This report then applies a number of techniques relating to causal discovery to firm-level innovation datasets. The techniques include unconditional and conditional independence tests, search for y-structures, additive noise techniques for continuous and discrete variables, and also linear non-Gaussian additive noise models. The datasets analysed are primarily the Community Innovation Survey (CIS) data for European countries for 2008, as well as Scoreboard data for the world's largest R&D spending companies.

As would be expected given both the complexity of innovation processes within firms and the limited quality of cross sectional survey data, it is not surprising that most of the results are not 'significant', in the sense that we could not conclude, with any certainty about which variable is affecting which. This implies that the relations between economic variables are often complex, with no clear causal ordering being apparent, because the true causality may run in both directions, or because of possible omitted confounding variables. However, in some cases, we could ascertain the direction of causal influence. On the one hand these causal findings are consistent with common sense, but also indicate a degree of novelty that suggests new models and frameworks may be generated by analysis of this kind. These could potentially provide a way to better inform managers and policy makers about the impact of their actions, with important economic implications.

The report presents and briefly discusses the causal relations that have been discovered. Some examples of our results can be quickly mentioned here:

1. Although growth in R&D investment and growth in market capitalization are both forward-looking variables, nevertheless they seem to only weakly influence each other directly. Instead, their causal link seems to be indirect via the intermediate variable sales growth.
2. Regarding public support for innovation, regional level support seems unrelated to national or EU-level funding, which raises questions about their effectiveness or decision-making criteria.
3. Regarding organizational innovation, new methods for organizing external relations (e.g. alliances, partnerships, outsourcing and contracting) has a causal effect on sales growth.
4. Regarding process innovations, it seems that developing process innovations with others, and the introduction of new logistics methods, have causal effects on the introduction of new methods of producing goods or services.
5. Regarding information sources, we the results suggest that conferences / trade fairs caused interest in professional and industrial associations as well as in scientific journals, as sources of information.

Furthermore, sales growth caused interest in professional and industrial associations as sources of information.

6. Regarding innovation expenditure variables, expenditures on machinery / software seem to be causally influenced by i) expenditures on the market introduction of innovations, ii) expenditures on training, and iii) expenditures on acquisition of external knowledge (e.g. licensing).
7. Finally, regarding innovation objectives variables, it seems that the range of goods or services was caused by improved flexibility in production, as well as interest in improving health and safety conditions.

Note that the causal claims 4-7 are particularly uncertain because they rely on the method of *discrete* additive noise models, for which no extensive evaluations on the performance are available yet. Claims 1-3 are slighly more certain because they rely on a more established approach, namely on conditional independence testing. In general, these results should be considered tentative however, and further work is needed before we can be confident about the policy implications. The results we have found mix findings that are entirely consistent with existing analysis, and also novel findings that open up new avenues for future analysis. As a proof-of-concept exercise, the results suggest that these approaches have important implications for future innovation policy analysis. In the first place, one of the goals of this report is to make available a new toolkit for economists and innovation scholars for subsequent work.

Contents

1 Motivation

The main economical challenge for Europe is to create jobs and achieve full employment as well as societal well-being. To this end, the European Commission seeks to provide evidence-based guidance for innovation policies. To predict the effect of policy interventions requires knowledge on *causal* relations between economical variables – as opposed to purely describing *statistical* relations. The latter would be insufficient as basis for predicting the effect of interventions, whereas causal relations are – by definition of causality – make statements on such effects. While purely statistical relations between variables are easy to get, inferring causal relations is still a challenging task. Traditionally, statisticians have widely shied away from drawing causal conclusions from statistical data and left the causal interpretation of statistical relations to the expert in the respective domains. According to this approach, statistical data can only prove that some variables are causally related, but not in which way. Since the 1990s, however, there is an increasing interdisciplinary community from machine learning, philosophy, statistics, and also physics, believing that causal conclusions can be drawn from statistical data provided that appropriate *assumptions* are made. The insight that learning from data always requires strong assumptions – regardless of whether one is concerned with learning causal relations or statistical relations only – is commonly agreed in machine learning anyway. According to this general rule, also this study is based on strong assumptions which can always be questioned. We have therefore tried to be as explicit as possible about the assumptions underlying our causal conclusions. We start with an exposition of the causal inference techniques used in this study, where we discuss the basic assumptions briefly and refer to the literature for more details.

The techniques presented in this report are necessarily advanced. However, where possible, the intuition behind these techniques is explained, and references are given to provide the interested reader with further reading. In addition, to help the reader digest the results, there is a summary of the most important findings, and their implications, at the end of each section of empirical results.

2 Preliminary remarks on the limitations of causal inference methods

2.1 Automated causal discovery?

To avoid potential misunderstandings, it should be emphasized that this study does not advocate using novel causal inference techniques in the form of fully automated causal discovery. By 'fully automated' we mean an approach that blindly feeds data into an algorithm to get causal information out of it. As we will see below, using causal inference techniques is always based on prior judgements. Such judgements include: deciding which are sensible variable groups to look at, or deciding whether or not other, more natural, variables should be generated as a pre-processing step from the raw data. Also the decision to what extent it is reasonable to consider the data points as independently drawn from the same distribution (which is usually referred to as 'the i.i.d. assumption', that is, *independently identically distributed*) depends on judgements that are hard to formalize. We will also see that causal inference relies on assumptions

that are often violated in practice. Yet, it may be justified to apply the methods if they are sufficiently *robust* with respect to these violations – a rationale that is not unique to causal inference because scientific reasoning always depends on oversimplified models. For causal inference, however, it is ongoing research to understand the robustness with respect to violating assumptions, e.g., Mooij et al. (2016). For these reasons, we currently prefer *computer-assisted human causal inference*, where a scientist (who is aware of the limitations of the methods) combines the outputs of different causal inference approaches in a way that results in reasonable causal statements.

2.2 Exploring causal relations in heterogeneous data sets

Our analysis uses data sets includes companies from different countries, regions, sectors, and sizes. Certainly, causal relations may be different within different groups of reference. Fort part of the study we have therefore studied companies from each sector separately. Refining the analysis by further splitting could reduce the sample size in each group to a size where conditional independence testing is no longer feasible. Apart from this purely statistical concern there is, however, a more fundamental issue related to potential refinements: such refinements induce *selection bias*. Assume, for instance, the causal analysis is done after dividing the data sets according to a categorical variable indicating the *size* of the company. Formally, this amounts to using the conditional probability distribution, given the varable `size`. Since `size` can also be an *effect* of other variables (e.g. Net Sales), conditioning on `size` may generate statistical dependences between variables that are a priori independent. This effect is also known as Berkson's paradox (Spirtes et al., 1993). Accordingly, by conditioning on `size` may pretend causal relations between variables that are actually not there. Therefore, inferring causal relations after accounting for specifities like `size` by dividing the data set does not *necessarily* yield more reliable causal statements, although it sometimes does.

These remarks are not meant to say that heterogeneity of the data set would not be a serious issue. After all, variables like `size` may be common causes of other variables under consideration. This way, not controlling for `size` may induce dependences that are actually not there. In other words, `size` may be a *confounder*. Nevertheless, one has to keep in mind that it depends on the causal structure whether controlling for a variable makes sense.

2.3 Sample sizes

For the analysis described in Section 5 we have considered the full dataset of companies since the sample sizes in each sector are moderate or even too low. In Section 6 we have decided to randomly select 2000 companies for the following reasons: first, the computation time for the kernel-based statistical independence tests would become quite large otherwise. Second, and more importantly, 2000 turned out to be a sample size for which most of the conditional and unconditional independences are already rejected. Even worse, a large fraction of p-values are 0 within the given numerical accuracy. One could certainly argue against accepting an independence of which we know that it would be rejected if the sample size were larger. We, however, believe that in reality almost every variable pair influences each other (in at least one direction) when arbitrarily

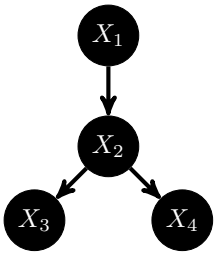

Figure 1: A directed acyclic graph visualizing causal relations between the random variables $X_1, \ldots, X_4$.

weak causal influences are taken into account. Therefore, we need to neglect weak influences in order to get some conditional independences (without which we would not be able to infer any causal relations). Given these issues, our choice of sample size 2000 seemed to be a reasonable compromise.

It is certainly important to check whether the obtained conditional independences can be reproduced by subsampling. We have therefore repeated some of the tests with smaller sample sizes. Some results are briefly discussed in Section 6.12.

To estimate the strength of correlations there is certainly no need to restrict the sample size. We have therefore used the full sample.

For additive noise models, we have taken even smaller subsamples of size 200. Additive noise models is based on a non-linear regression procedure whose computation time grows heavily with larger sample sizes. Further, large sample sizes raise a problem that is related to the one above: the additive noise model is always rejected with astronomically small p-values. These 200 values were chosen independently at random, not as subsamples from the 2000 chosen for the conditional independence tests because the software packages are meant to be used as standalone tools.

3 Formal language: Directed acyclic graphs

Following Pearl (2000); Spirtes et al. (1993), causal relations are formalized as directed acyclic graphs (DAGs) with random variables $X_1, \ldots, X_n$ as nodes, see Figure 1. An arrow from X_i to X_j indicates that interventions on X_i have an effect on X_j given that the remaining variables in the DAG are adjusted to a fixed value.[1] Particularly for economic variables, this definition refers to a highly hypothetical scenario since adjusting one variable will often already be difficult or infeasible (unless the variable refers directly to a policy instrument, such as tax rate or a subsidy). Nevertheless, the hypothetical scenario is the simplest way to define the meaning of the arrows. We can also say that an arrow indicates a 'direct' causal influence, but keep in mind that the distinction between direct and indirect is only meant relative to the set of variables under consideration: 'direct' means that the influence is not mediated by any of the other variables in the DAG. Here we assume that an absolute distinction between

[1] For a previous application of graphical causal models to economical data see Moneta et al. (2013).

'direct' and 'indirect' influence is meaningless. This perspective is motivated by a physical picture of causality where variables may refer to measurements in space and time: If X_i and X_j are variables measured at different locations, then every influence of X_i on X_j requires a physical signal propagating through the space. We can then replace the arrow $X_i \to X_j$ with an arbitrarily long chain of intermediate variables that refer to measurements along the way the signal propagates.

4 Explanation of the causal inference techniques used in this study

The basic assumption relating statistics and causality is Reichenbach's principle (Reichenbach, 1956) stating that every statistical dependence between two observed random variables X and Y indicate that at least one of the following three alternatives is true: 1) X influences Y, 2) there is a common cause Z influencing X and Y, or, 3) Y influences X. In the second case, Reichenbach postulated that X and Y are conditionally independent, given Z, i.e., their probability densities satisfy the equation

$$p(x, y|z) = p(x|z)p(y|z),$$

for all x, y, z. Henceforth, we will denote this by $X \perp\!\!\!\perp Y\,|Z$. The fact that all three cases can also occur together[2] is an additional obstacle for causal inference. For this study, we will mostly assume that only one of the cases occurs and try to distinguish between them subject to this assumption. We are aware of the fact that this oversimplifies many real-life situations. On the other hand, even if the cases interfere, there may often be one of the 3 types of causal links that is more significant than the others. Then it is also more valuable for practical purposes to focus on the *main* causal relations. After all, statements like 'every variable influences every other variable' are inappropriate as guidance for future policies.

To distinguish between the three possible cases, we use two different approaches. First, we describe the 'traditional' technique of looking at X and Y as part of a larger causal network and testing conditional statistical independences, as explained in Section 4.1. Second, we use the 'shape' of the distribution $P_{X,Y}$ to infer the type of causal link, as explained in Section 4.2

4.1 Conditional independence based approach

The cornerstone of causal inference from statistical data is the Causal Markov condition (Spirtes et al., 1993; Pearl, 2000), of which we will use 3 equivalent versions:

1. **local Markov condition:** every node is conditionally independent of its non-descendants, given its parents, see Figure 2.

2. **global Markov condition:** if two sets **X** and **Y** of variables are d-separated by the set **Z** (for the definition of this graphical criterion see

[2]Although 1) and 3) occurring together would no longer be an *acyclic* graph.

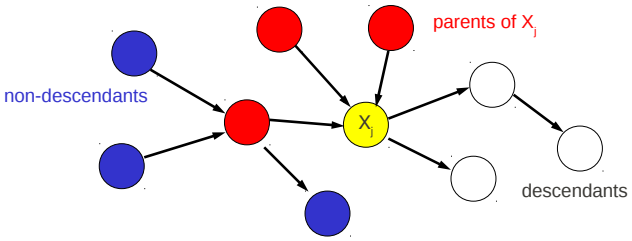

Figure 2: Visualization of the local causal Markov condition: the yellow node is the variable under consideration, the red ones are its parents, the blue ones the non-descendants that remain after conditioning on the parents.

Pearl (2000)), then $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent, given $\mathbf{Z}$. Since the conditional independences stated by the local Markov condition implies further independences (see the semi-graphoid axioms in Pearl (2000)), a rule is required that explicitly states all the conditional independences that are implicitly stated by the local Markov condition. This is done by the global Markov condition.

3. **Factorization:** the density:[3] of the joint distribution $P_{X_1,\ldots,X_n}$ factorizes according to $p(x_1,\ldots,x_n) = \prod_j p(x_j|pa_j)$. Here, every $p(x_j|pa_j)$ describes the causal mechanism according to which each variable X_j is influenced by its parents PA_j.

To consider one of the simplest examples, assume that the DAG reads

$$X \to Y \to Z.$$

Then, applying the local Markov condition to the node Z yields $X \perp\!\!\!\perp Z|Y$ because Y is the only parent of Z and X is a non-descendant. The factorization of the joint density corresponding to this causal structure reads

$$p(x,y,z) = p(x)p(y|x)p(z|y).$$

Note that the factorization

$$p(x,y,z) = p(x)p(y|x)p(z|x,y)$$

holds regardless of the causal structure, but here we have used $p(z|x,y) = p(z|y)$, which is exactly the conditional independence stated above.

It should be noted, however, that applying the causal Markov condition to a set of observed variables implicitly assumes that the set is causally sufficient (Spirtes et al., 1993), i.e., that there are no unobserved variables that have an influence on two or more observed variables. This is a strong assumption that

[3] if it exists w.r.t. product measure, see Lauritzen: Graphical models (1996)

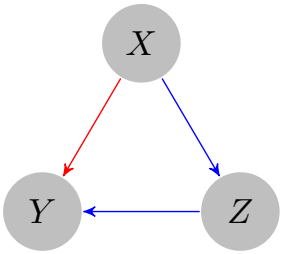

Figure 3: Example of a DAG that does not entail any conditional independences. Therefore, any distribution $P_{X,Y,Z}$ for which conditional or unconditional independences hold true, is unfaithful.

is often a bad approximation of real-life scenarios. It is therefore remarkable that some causal inference methods are able, under appropriate assumptions, to infer causal directions even in the presence of hidden common causes, see, for instance, the discussion of y-structures below.

The second cornerstone of conditional independence based causal inference is an assumption called *faithfulness* (Spirtes et al., 1993) (related to *minimality* by Pearl (2000)), which reads: accept only those causal DAGs as a plausible causal explanation for which all the observed conditional independences follow from the Markov condition. In other words, one rejects all hypotheses allowing for conditional dependences that are not actually true for the observed joint distribution. The idea is that unfaithful distributions are 'not generic' in the following sense: typical choices of the conditional distributions $p(x_j|pa_j)$ induce only the independences entailed by the Markov condition. To provide an intuition about this condition, consider the DAG in Figure 3. For instance, the causal structure does not entail that X and Y are independent, formally denoted by $X \perp\!\!\!\perp Y$. To achieve independence by such a DAG, the direct influence of X on Y and the indirect one intermediated by Z, need to exactly compensate. To see in which sense this requires unlikely coincidences assume, for instance, that the relations between X, Y, Z are described by the linear structure equations

$$\begin{aligned} Z &= \alpha X + N_Z \\ Y &= \beta Z + \gamma X + N_Y, \end{aligned}$$

where X, N_Z, N_Y are independent. Then the total influences of X on Y is described by the structure coefficient $\alpha \cdot \beta + \gamma$. Hence, $X \perp\!\!\!\perp Y$ requires $\alpha \cdot \beta + \gamma = 0$, which is considered unlikely according to the philosophy behind causal faithfulness. An (imperfect) analogy would be that, looking out of my office, it seems unlikely – although I cannot be certain, that there is not a giraffe whose position is perfectly calibrated such that she is hiding directly behind the lamppost, hidden from my line of sight.

Conditional independence based causal inference therefore infers that the true causal DAG is among the set of those DAGs for which the observed conditional independences are exactly those that are entailed by the Markov condition (Spirtes et al., 1993).

A fundamental limitation of the conditional independence based approach is that it is unable to distinguish between causal structures that induce the same

set of conditional independences, so-called *Markov equivalent* DAGs. Verma and Pearl (1990) have shown a graphical criterion for Markov equivalence stating that two DAGs entail the same conditional independences if and only if they have the same skeleton (i.e. the undirected graph obtained by ignoring the directions of the edges) and the same v-structures. v-structures (also called 'unshielded colliders') are defined as triples of nodes A, B, C being linked as $A \to C \leftarrow B$ where A and B are not directly connected.

Testing conditional independences is a non-trivial problem in staistics. For this study, we have assumed that there is no method available that is optimal for all situations. Instead, we have chosen different methods that seemed appropriate for the respective type of variables and the observed type of statistical dependences.

Unconditionally independence tests We start with testing *unconditional* independences. If X and Y attain one-dimensional numeric values (regardless of whether they are continuous or discrete), $X \perp\!\!\!\perp Y$ implies uncorrelatedness $\mathrm{cor}(X, Y) = 0$. In principle, dependences could be only of higher order, i.e., X and Y could be dependent without being correlated. Although this is quite non-generic, we therefore also use a type of independence test that is able to detect higher-order dependences, namely the Hilbert Schmidt Independence Criterion (HSIC) by Gretton et al. (2005a,b). The theory of HSIC involves some functional analysis and uses the method of Reproducing Kernel Hilbert Spaces (RKHS) which is meanwhile widely used in machine learning (Schölkopf and Smola, 2002). We now explain the basic idea behind the HSIC test without using the language of RKHS. HSIC defines a distance measure D on the space of joint distributions P_{XY} such that $D[P_X P_Y, P_{X,Y}]$ is easy to estimate although estimating the distributions P_X, P_Y, P_{XY} themselves is hard:

$$D^2[P_X P_Y, P_{X,Y}] := \int k(x,\tilde{x})k(y,\tilde{y})\Big\{p(x,y) - p(x)p(y)\Big\}\Big\{p(\tilde{x},\tilde{y}) - p(\tilde{x})p(\tilde{y})\Big\}dxdyd\tilde{x}d\tilde{y}.$$

One can show that it vanishes only for $P_{XY} = P_X P_Y$ (i.e. $X \perp\!\!\!\perp Y$) provided that the 'kernel' functions $k(x,\tilde{x}), k(y,\tilde{y})$ satisfy certain positivity conditions. HSIC thus measures dependence of random variables, like a correlation coefficient, with the difference that it accounts also for non-linear dependences. A second difference to correlation is that dependence is always indicated by positive values of HSIC, while negative ones don't exist.

The most popular example is the so-called Gaussian kernel

$$k(x,\tilde{x}) := e^{-\frac{(x-\tilde{x})^2}{2\sigma^2}}, \tag{1}$$

where the 'bandwidth' σ is a free parameter which determines the scale at which dependences can be detected. The following estimator can be shown (Gretton et al., 2005b) to converge to the squared distance D^2:

$$\frac{1}{n^2}\sum_{i,\tilde{i}} k(x_i,x_{\tilde{i}})k(y_i,y_{\tilde{i}}) - 2\frac{1}{n^3}\sum_{i,\tilde{i},j} k(x_i,x_{\tilde{i}})k(y_i,y_j)\frac{1}{n^4}\sum_{i\tilde{i},j,\tilde{j}} k(x_i,x_{\tilde{i}})k(y_j,y_{\tilde{j}}) \quad \longrightarrow$$
$$\int k(x,\tilde{x})k(y,\tilde{y})\Big\{p(x,y) - p(x)p(y)\Big\}\Big\{p(\tilde{x},\tilde{y}) - p(\tilde{x})p(\tilde{y})\Big\}dxdyd\tilde{x}d\tilde{y}$$

Software: The HSIC (Hilbert Schmidt Independence Test) by Gretton et al. (2005b) can be downloaded at `http://people.kyb.tuebingen.mpg.de/arthur/indep.htm`.
Note, however, that the kernel test for *conditional* independence of Zhang et al. (2011), which we need to introduce below anyway, can also be used for *un*conditional independence testing. To avoide using too many different software tools at the same time, we have also based everything on the test in Zhang et al. (2011), which we describe in more detail below.

Although the ability to detect *non-linear* dependences is crucial for causal discovery since the causal relations of interest may not necessarily result in non-zero *correlations*, correlations are particularly interesting for policies since causal dependences that yield correlations are easier to interpret, because desired results are mostly of the form 'if variable X is increased, this also increases the expectation of Y'. Statements like 'increasing X increases the variance of Y' are often less helpful if the goal is to increase some target variable Y.

Conditional independence tests For multi-variate Gaussian distributions, conditional independence can be inferred already from the covariance matrix via computing *partial correlations*. Instead of using the covariance matrix we describe the following more intuitive way to obtain partial correlations: Let $P_{X,Y,Z}$ be Gaussian, then $X \perp\!\!\!\perp Y \,|Z$ is equivalent to

$$\mathrm{cor}(X - \alpha Z, Y - \beta Z) = 0, \tag{2}$$

where α and β are the structure coefficients obtained from least square regression when regressing X on Z and Y on Z, respectively. Explicitly, they are given by

$$\alpha = \mathrm{Cov}(X, Z)/\mathrm{Var}(Z) \quad \beta = \mathrm{Cov}(Y, Z)/\mathrm{Var}(Z).$$

Note, however, that in non-Gaussian distributions, vanishing of the *partial correlation* on the left hand side of (2) is neither necessary nor sufficient for $X \perp\!\!\!\perp Y \,|Z$. On the one hand, there could be higher order dependences not detected by the correlations. On the other hand, the influence of Z on X and Y could be non-linear and in this case it would not entirely be screened off by linear regression on Z. This is why using partial correlations instead of independence tests can introduce two types of errors, namely accepting independence although it does not hold, as well as rejecting it although it holds (even in the limit of infinite sample size). To partly overcome this limitation, we have also used 'partial HSIC' (we are not aware of any example of it in the literature, but it is a straightforward replacement of partial correlation), that is, performing an HSIC test on the residuals $X - \alpha Z$ and $Y - \alpha Z$. If their independence is accepted, then $X \perp\!\!\!\perp Y \,|Z$ necessarily holds. Hence, we have in the infinite sample limit only the risk of rejecting independence although it does not hold, while the second type of error, namely accepting conditional independence although it does not hold, is no longer possible.

To account for the fact, that $X \perp\!\!\!\perp Y \,|Z$ does not necessarily imply that the residuals of X and Y become independent after linearly regressing X and Y on Z, Zhang et al. (2011) propose a kernel test for conditional independence. To describe the theory behind it would go beyond the scope of this report. We only mention that so-called Reproducing Kernel Hilbert Spaces (RKHS) (which

is meanwhile a standard technique in machine learning to account for non-linear dependences (Schölkopf and Smola, 2002)) are used to account for the fact that both variables X and Y may depend on Z in a non-linear way. Nevertheless, conditional independence testing remains a hard problem. Implicitly, testing whether X and Y are independent, given Z infers how X and Y depend on Z in order to check whether the residual uncertainty is independent. We will therefore always trust the results of unconditional tests more than the conditional ones.

Software: the Kernel Conditional Independence Test (KCI) by Zhang et al. (2011) can be downloaded at `http://people.tuebingen.mpg.de/kzhang/KCI-test.zip`. The matlab function `indtest_new.m` is called by the command

$$[\texttt{pvalue}, \sim] = \texttt{indtest\_new}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, []),$$

where $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ are data matrices of observations (again, all variables can also be vector-valued) when testing $X \perp\!\!\!\perp Y \,|Z$. For unconditional independence tests, $\mathbf{Z}$ is replaced with [].

For the case where Z is a variable attaining only a small number of different values, conditional independence tests can be made that rely on testing independence of X and Y for each possible value and accounting for the multi-hypothesis test in an appropriate way. If X and Y also attain only a small number of values, standard χ^2 tests are possible. Since this report considers relations between discrete and continuous variables we cannot resort to those tests and have chosen kernel tests instead.

For automated causal discovery (which we do not favour for the purpose of this report for reasons explained in the beginning) we refer to the homepage of the TETRAD-project of the Carnegie-Mellon-University (CMU) Pittsburgh `http://www.phil.cmu.edu/tetrad/`.

For the purpose of this study, we have decided to use the following strategy: We start from a set of 5-8 variables $S := \{V_1, \ldots, V_d\}$, The number of variables is kept as small as possible in order to avoid too serious multi-testing issues for the statistical tests. Modern big data approaches construct causal relations among thousands of variables (e.g. gene expression levels in biological informatics) and obtain causal DAGs that are still helpful (although a large number of arrows may be wrong) because they yield, for instance, lists of variables that are good candidates for being causes of some target variable. Our approach, instead, puts more focus on particular causal arrows of interest – this is why we focus on small sets on variables.

For such a small set, we basically implement a reduced version of the usual conditional independence based causal inference algorithms (PC algorithm, named after Peter Spirtes and Clark Glymour (Spirtes et al., 1993) and the IC, 'inductive causation' in Pearl (2000)): We first test all unconditional statistical independences $X \perp\!\!\!\perp Y$ for all pairs (X, Y) of variables in this set. Then we test all conditional independences $X \perp\!\!\!\perp Y \,|Z$ for all possible triples (X, Y, Z) in S. Again, to avoid to get too serious multi-testing issues, we do not perform tests for independences of the form $X \perp\!\!\!\perp Y \,|Z_1, Z_2, ..., Z_n$ with $n > 1$, also because conditioning on more than one variable is a statistically ill-posed problem for

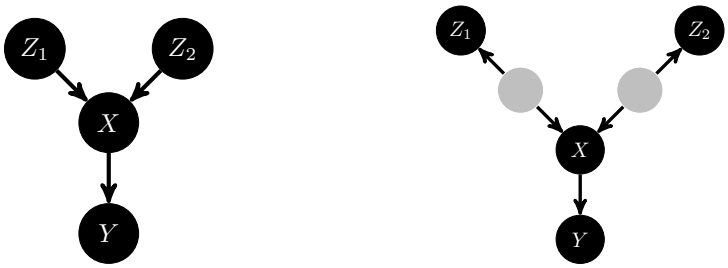

Figure 4: Left: the simplest possible y-structure, see text. Right: A causal structure involving latent variables (these unobserved ones are marked in grey) that entails the same conditional independences on the observed variables as the structure on the left.

limited sample size. We then construct an undirected graph where we connect each pair that is neither unconditionally nor conditionally independent. Whenever the number d of variables is larger then 3, it is possible that we obtain too many edges this way because independence tests conditioning on more variables could render X and Y independent. We take this risk, however, for the above reasons and in the belief that reliable data analysis should avoid methods with too high complexity. In some cases, the pattern of conditional independences also allows for inferring the direction of some of the edges: whenever the resulting undirected graph contains the pattern $X - Z - Y$ where X and Y are non-adjacent. and we observe that X and Y are independent but conditioning on Z renders them dependent, then Z must be common effect of X and Y, i.e., we have a v-structure at Z. For this reason, we perform conditional independence tests also for pairs of variables that have already been verified to be unconditionally independent. From the point of view of constructing the skeleton, i.e., the graph with undirected edges, the conditional test would be redundant, but for orienting edges it can be helpful. This argument, like the whole procedure above, assumes causal sufficiency, i.e., the absence of hidden common causes. It is therefore remarkable that the below method is able to infer causal directions in the presence of common causes.

Identification of genuine causes via y-structures Since it is usually infeasible to reliably infer the whole causal network, it is desirable to look for inference tools that provide insights about parts of the causal structure within the entire network. Statements of the form 'variable X_i influences X_j' (whether this influence is intermediated by other observed variables or not) are extremely helpful. An interesting tool for this purpose is the identification of so-called y-structures which we explain now. A simple case of such a y-structure is displayed in Figure 4. The conditional independence pattern reads $Z_1 \perp\!\!\!\perp Z_2$, $\{Z_1, Z_2\} \not\perp\!\!\!\perp Y \,|X$, and conditioning on X renders Z_1, Z_2 dependent. The reason why this pattern is helpful is that observing it tells us that X influences Y regardless of whether the set $\{Z_1, Z_2, X, Y\}$ is causally sufficient or not (Pearl, 2000). The idea of the argument is that the path[4] connecting Z_j and X needs to have an arrowhead at X, otherwise conditioning on X could not render Z_j

[4]Of course, there can be more than one path, but we only want to provide an intuition, see Pearl (2000) for the full argument.

dependent when they are independent without conditioning. Further, one can argue that the path connecting X and Y needs to be a directed one from X to Y because a path from Y to X or a common cause connection would generate dependences by conditioning on X. Similar arguments hold for the case where Z_1, Z_2 are conditionally independent given some set S of variables where S does not contain X and become dependent when conditioning on $S \cup \{X\}$. In this case, one can still conclude that the path connecting Z_j with X has an arrowhead at X. Thus, one can further argue (as above) that the path generating the dependence between X and Y must be directed from X to Y. To see that causal sufficiency is not needed for these arguments, consider the causal structure in Figure 4, right, which generates the same pattern of conditional independences.

The message of this insight is that it is always helpful to identify quadruples of variables $\{Z_1, Z_2, X, Y\}$ within a larger network for which the following properties hold: (1) $Z_1 \perp\!\!\!\perp Z_2 \,|S$ and (2) $Z_1 \not\perp\!\!\!\perp Z_2 \,|S \cup \{X\}$ where S is some set of variables not containing X, and (3) $\{Z_1, Z_2\} \perp\!\!\!\perp Y \,|X$, because then (subject to causal faithfulness) we have shown that X is a genuine cause of Y, regardless of any further hidden variables.

4.2 Additive noise based causal discovery

Inferring the causal direction between two real-valued variables Additive noised based causal inference is in principle also able to distinguish between DAGs for which the Markov condition entails the same set of conditional independences. In particular, it is able to distinguish between $X \to Y$ and $Y \to X$ from $P_{X,Y}$ alone (Hoyer et al., 2009). Assume Y is a function of X up to a additive noise term that is statistically independent of X, i.e.,

$$Y = f_Y(X) + N_Y,$$

where $N_Y \perp\!\!\!\perp X$. One can then show that there is in the 'generic case' (the precise meaning of 'generic' here is complicated, see Hoyer et al. (2009)) no additive noise model from Y to X, i.e., there is no function f_X such that

$$X = f_X(Y) + N_X,$$

with $N_X \perp\!\!\!\perp Y$. Figure 5 visualizes the idea showing that the noise cannot be independent in both directions.

To see a real example, Figure 6 shows the first example from our data base containing cause-effect variable pairs of which we believe to know the causal direction[5]. Up to some noise, Y is given by a function of X (which is close to linear apart from small altitudes). On the other hand, if we try to describe the altitude as a function of the temperature, the error term has a somehow 'complex' structure, because in the region of the y-axis corresponding to altitude zero (sea level), the distribution of the noise appears as two clusters. Phrased in terms of the language above, writing X as a function of Y yields a residual error term that is highly dependent of Y. On the other hand, writing Y as a function of X yields the noise term that is largely homogeneous along the x-axis. Hence, the noise is almost independent of X. Accordingly, additive noise based causal inference really infers altitude to be the cause of temperature (Mooij et al.,

[5]Database with cause effect pairs `https://webdav.tuebingen.mpg.de/cause-effect/`. Copyright for each pair can be found there.

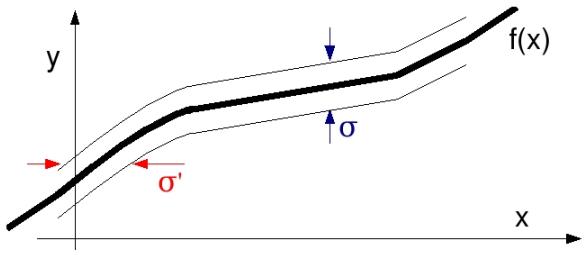

Figure 5: Visualization of the idea of additive noise based inference for the case where the noise has bounded range: an additive noise model from X to Y implies that the width of the distribution in y-direction is constant in x. For non-linear functions, however, this conflicts with the statement that the width in x-direction is constant in y (which would be required by an additive noise model from Y to X).

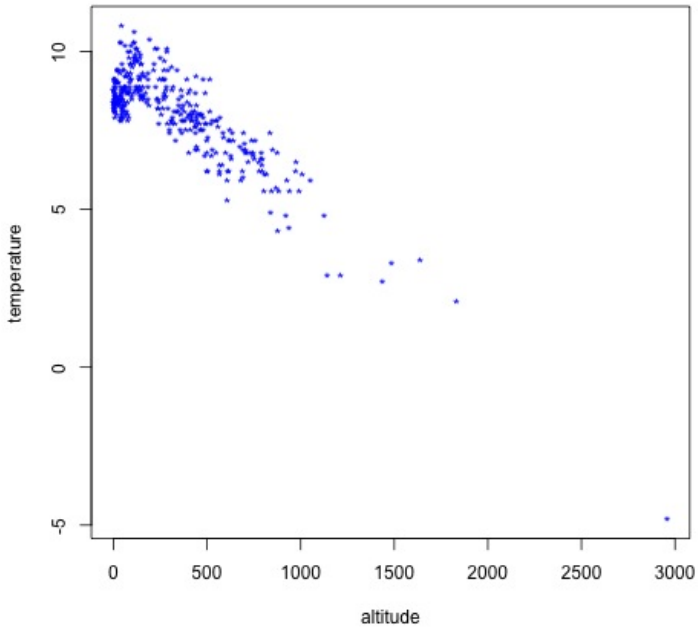

Figure 6: Scatter plot for the relation between altitude (X) and temperature (Y) for places in Germany, see text. Example taken from the database of cause effect pairs at `https://webdav.tuebingen.mpg.de/cause-effect/`, `copyright for each pair can be found there.`

2016), which is certainly true: fixing a thermometer at a balloon would confirm that the temperature changes with the altitude, while heating a place would not change its altitude.

The practical method for inferring causal directions works as follows. (1) Regress Y on X, that is, find the function f_Y with

$$f_Y(x) := \mathbb{E}[Y|x].$$

(2) compute the residual variable $N_Y := Y - f_Y(X)$ and (3) test whether N_Y is independent of X. Then do the same with exchanging the roles of X and Y. If independence of the residual is accepted for one direction but not the other, the former is inferred to be the causal one. If independence is either accepted or rejected for both directions, one does not decide. If a decision is enforced one can just take the direction for which the p-value for the independence is larger. This, however, seems to yield performance that is slightly above chance level only (Mooij et al., 2016). Otherwise, setting the right confidence levels for the independence test is a difficult decision for which there is no general recommendation. Conservative decisions can yield rather reliable causal conclusions, as shown by extensive experiments in Mooij et al. (2016). It should be emphasized that additive noise based causal inference does not assume that every causal relation in real-life can be described by an additive noise model. Instead, it assumes that *if* there is an additive noise model in one direction, this is likely to be the causal one. For a justification of this way of reasoning we refer to Janzing and Steudel (2010). The idea is that a joint distribution $P_{X,Y}$ that admits an additive noise model from X to Y is unlikely to be generated by the causal structure $Y \to X$ because this requires atypical adjustments between P_Y and $P_{X|Y}$. To show his, Janzing and Steudel (2010) derive a differential equation that expresses the second derivative of the logarithm of $p(y)$ in terms of derivatives of $\log p(x|y)$. Therefore, for a given conditional $P_{X|Y}$, only very specific choices of P_Y generate an additive noise model from X to Y.

See Mooij et al. (2016) for a recent extensive evaluation of additive noise based inference on real and simulated data, also in comparison with other causal inference methods that have been proposed during the past two decades. The real data experiments refer to our benchmark data set `http://webdav.tuebingen.mpg.de/cause-effect/`

Software: Pairwise Additive Noise Model based causal inference Hoyer et al. (2009) can be downloaded at
`https://staff.fnwi.uva.nl/j.m.mooij/publications.html`, see 'code' for the aticle Mooij et al. (2016). The matlab function `cep_anm.m` is called by the following command:

$$\texttt{score} = \texttt{cep\_anm}(X, Y, \texttt{methodpars}),$$

where `methodpars` is a structure array containing several parameter values, see `cep_anm.m` for explanations. We have defined it as follows:
`methodpars=struct; methodpars.nrperm=0; methodpars.FITC=0;`
`methodpars.splitdata=0; methodpars.evaluation='pHSIC';`
`methodpars.bandwidths=[0,0]; methodpars.meanf='meanAffine';`
`methodpars.minimize='minimize'; methodpars.entest='Shannon_kNN_k';`.
The output `score` is a value between $-\infty$ and ∞ where large positive values indicate large evidence for $X \to Y$ and large negative ones for $X \leftarrow Y$.

Interestingly, in contrast to standard econometric regression models that assume that relationships are linear, nonlinearity is a helpful property for identifying the causal direction: For linear models, identification of the direction requires non-Gaussian noise, as employed by the inference method LiNGAM explained below, while non-linear relations render the causal direction even identifiable when the noise is Gaussian (Hoyer et al., 2009; Peters et al., 2014; Mooij et al., 2016).

Discrete additive noise models If a variable X attains values in $\{0, 1, 2, \ldots, k_X - 1\}$ for some natural number k_X, and Y attains values in $\{0, 1, 2, \ldots, k_Y - 1\}$ for some number k_Y, one can also define an additive noise model from X to Y by

$$Y = f_Y(X) + N_Y \quad \text{with } N_Y \perp\!\!\!\perp X,$$

where $+$ is now the addition *modulo* k_Y and N_Y is a noise term with values in $\{0, 1, 2, \ldots, k_Y - 1\}$. Peters et al. (2010, 2011) proposed to use such a model for causal inference, in analogy to the continuous case explained above. Explicitly, one prefers the causal direction $X \to Y$ if an additive noise model from X to Y is accepted for a certain given p-value and the analogue additive noise model from Y to X is rejected. To get stronger reliability, one would only decide upon the causal direction if one p-value is significantly higher than the other one. Although Peters et al. (2010, 2011) describe some encouraging results on real and simulated data, there is no extensive evaluation on the performance of discrete additive noise based causal inference comparable to the evaluation of Mooij et al. (2016) for the continuous case. The cyclic structure (i.e., addition modulo k_Y or k_X, respectively) may sound only natural for cases where X and Y are discretized cyclic variables like season or hours. There are, however, also further, less obvious cases where the cyclic structure makes sense: if Y is binary (i.e., a bit), the modulo 2 addition of the noise term corresponds to a bit flip error. It should be emphasized, however, that inferring the causal direction between just two *binary* variables is challenging, if not infeasible. This is because the joint distribution of two binaries is described by 3 parameters, i.e., it is an object that is 'too simple' to contain reliable information on the causal direction. For the purpose of this report, we have applied discrete additive noise

based causal inference only to discrete variables that attain at least 4 different values.

Software: Pairwise causal inference using discrete additive noise models Peters et al. (2011, 2010) can be downloaded at: `http://webdav.tuebingen.mpg.de/causality/online_aistats_arxive_discrete.zip` . The matlab function `fit_both_dir_discrete.m` is called via the following command:

$$[\sim, \texttt{pvalue\_XY}, \sim, \texttt{pvalue\_YX}] = \texttt{fit\_both\_dir\_discrete}(X, 1, Y, 1, \texttt{pvalue}, 0),$$

where `pvalue` defines the p-value for which the algorithm stops looking for a function with a better fit. We have set it to 0.05. `pvalue_XY` and `pvalue_YX` are the outputs indicating the p-values for accepting/rejecting a discrete additive noise model from X to Y and for Y to X, respectively.

Linear non-Gaussian additive noise models For non-Gaussian variables, additive noise models even allow for identification of causal directions even when the functions are linear. That is, whenever

$$Y = \alpha X + N_Y,$$

where $\alpha \in \mathbb{R}$ and N_Y is a noise term that is statistically independent of X, there cannot be a model of the form

$$X = \beta Y + N_X,$$

with $\beta \in \mathbb{R}$, where the noise term N_X is statistically independent of Y (unless the joint distribution $P_{X,Y}$ is a bivariate Gaussian) (Kano and Shimizu, 2003). This way, one can infer the causal direction as the one that admits a linear additive noise model if this there is exactly one direction for which this is the case, as proposed by Kano and Shimizu (2003). This approach is the so-called LiNGAM (Linear Non-Gaussian Additive Noise Models) method. Given powerful statistical independence tests like HSIC, the implementation of LiNGAM is straightforward: Just test whether the residual $Y - (\mathrm{Cov}(X, Y)/\mathrm{Var}(X)) \cdot X$ is independent of X and likewise for the backwards direction.

Software: several implementations of LiNGAM (Kano and Shimizu, 2003) can be found on the internet, we have used our own implementation, a function in R named `pairwise_lingam.R` which is called via

$$\texttt{pvalues} = \texttt{pairwise\_lingam}(\mathbf{X}, \mathbf{Y})$$

where $\mathbf{X}$ and $\mathbf{Y}$ are column vectors of real-valued observations. The output `pvalues` is a vector containing the pvalue for the additive noise models $X \to Y$ and $X \leftarrow Y$, respectively. The program calls the R function `ind_test.R` which, in turn, calls the matlab (conditional) kernel independence test of Zhang et al. (2011), explained above. Both R functions, `pairwise_lingam.R` and `ind_test.R` are provided as part of this project report.

4.3 Non-algorithmic inference by hand

Since the survey contains both continuous and discrete variables, we would actually need software that is able to infer causal directions when one variable is discrete and the other continuous. Unfortunately, there are no off-the-shelves methods available for this case. Sun et al. (2006); Janzing et al. (2009) propose a method that has been tested for a very limited number of data sets. In absence of methods for automated causal discovery, we can try to get hints on the causal direction by our intuition and arguments that rely on the *Principle of Algorithmically Independent Conditionals* (Janzing and Schölkopf, 2010; Lemeire and Janzing, 2012). For the special case of a simple bivariate causal relation with cause and effect it states that the shortest description of the joint distribution $P_{\text{cause,effect}}$ is given by separate descriptions of P_{cause} and $P_{\text{effect|cause}}$. This implies, in particular, that describing $P_{\text{caus,effect}}$ in terms of P_{cause} and $P_{\text{effect|cause}}$ is simpler than describing it in terms of P_{effect} and $P_{\text{cause|effect}}$. To illustrate this principle, Janzing and Schölkopf (2010); Lemeire and Janzing (2012) show the two toy examples shown in Figure 7. In both cases we have a joint distribution of the continuous variable Y and the binary variable X. On the left hand side, P_Y is a mixture of two Gaussians, each of which can be assigned to the cases $X = 0$ and $X = 1$, respectively. This joint distribution $P_{X,Y}$ clearly indicates X causing Y because this naturally explains *why* P_Y is a mixture of two Gaussians and why each component corresponds to a different value of X. When the same distribution is generated via the causal structure $Y \to X$ there is, first, no explanation why P_Y consists of two modes and, second, no explanation why each of the Gaussians corresponds to one value of X.[6] On the other hand, the distribution on the right hand side clearly indicates that Y causes X because the value of X is then obtained by a simple thresholding mechanism, i.e., $P_{X|Y}$ is a 'machine' receiving continuous input y and generating the output $X = 0$ or $X = 1$ depending on whether y is above a certain threshold. To generate the same joint distribution of X and Y when X is the cause and Y is the effect involves a quite unusual mechanism for $P_{Y|X}$. Then, $P_{Y|X}$ would be a 'machine' with binary input X whose output is one of the two sides of a *truncated Gaussian*, depending on the input X.

The examples show that joint distributions of continuous and discrete variables may contain causal information in a particularly obvious manner. There are, however, no algorithms available that employ this kind of information apart from the preliminary tools mentioned above. We therefore rely on human judgements to infer the causal directions in such cases. Below we will therefore visualize some particular bivariate joint distributions of binaries and continuous variables to get some, although quite limited, information on the causal directions. Although we cannot expect to find joint distributions of binaries and continuous variables in our real data for which the causal directions are as obvious as for the cases in Figure 7, we will still try to get some hints.

[6] To understand the last argument the reader may verify that for two overlapping Gaussians it requires quite sophisticated tuning of the conditional $P_{X|Y}$ in order to achieve that both conditional distributions $P_{Y|X=0}$ and $P_{Y|X=1}$ become Gaussians.

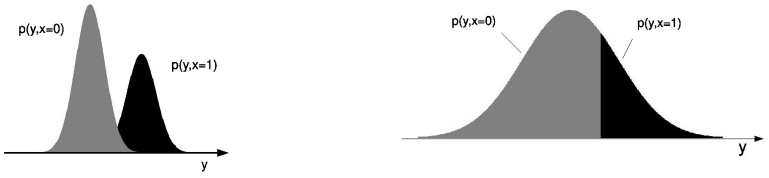

Figure 7: Left: visualization of a joint distribution of a binary variable X and a continuous variable Y for which it is pretty clear that the causal direction reads $X \to Y$. Right: joint distribution for which it is pretty clear that the causal direction is $Y \to X$. Figures are taken from Janzing and Schölkopf (2010), Janzing et al. (2009), and Lemeire and Janzing (2012).

5 Analysis of the Scoreboard data set

Here we focus on the data set 'EU Industrial R&D Investment Scoreboard' compiled by Bureau van Dijk. Together, the firms in this database represent about 90% of the total expenditure by business firms on R&D worldwide. The purpose of this database is to facilitate the monitoring of the world's largest R&D investing companies, and to provide evidence to inform European innovation policy. It focuses on the world's largest R&D investors. Companies are ranked according to their investment in R&D in the past year and the list is cut at 2500. Moreover, only companies with publicly available annual reports and accounts are considered. As a result of these criteria of inclusion, small and young firms are under-represented. It should be emphasized that this selection can already introduce a bias that influence the causal conclusions. To study, for instance, the impact of selecting only companies whose R&D expenditure is above a certain threshold, one may introduce the binary variable `above threshold?`, which is then an effect of R&D. Subject to causal faithfulness, conditiong on `above threshold?` induces dependences between all variables that influence R&D. Causal inference in the presence of selection bias is a hard problem and ongoing research, see e.g., Bareinboim and Pearl (2011); Zhang et al. (2016) for recent developments.

The data set contains short time series of several variables referring to annual values corresponding to the years 2008–2013. We will focus on three of them, namely 'Net Sales', 'R& D Expenditure' and 'Market Capitalization'.[7]

Since these quantities form short and non-stationary time series we are not in the typical domain of time series analysis. We will therefore convert the development of each quantity over the time into *growth rates* and later infer causal relations between these growth rates. This way, the causal inference problem is formally speaking using i.i.d. data (independent identically distributed), where

[7] A caveat concerning the market capitalization variable should be noted. For each firm, we have one observation for market capitalisation per year. However, the level of market capitalization may fluctuate over the course of a year. Furthermore, firms may have different financial reporting periods, which might affect their market capitalization values. Some firms might report their market capitalization for e.g. April, while other firms might report their market capitalization for e.g. October. Hence, firms that report their market capitalization at different points in the year might not be strictly comparable, but this issue is probably less serious than other uncertainties of causal data analysis.

each company is considered an instance of the statistical sample.

5.1 Computing growth rates for companies

Let r_{tc} be the R&D Expenditure of company c in the year t, and, similarly, n_{tc} denote the Net Sales and m_{tc} the Market Capitalization. Then the growth rates r_c, n_c, m_c for company c is given by fitting a linear model on the logarithmic scale:

$$\begin{aligned} \log n_{tc} &= n_c t + u^N_{tc} \\ \log r_{tc} &= r_c t + u^R_{tc} \\ \log m_{tc} &= m_c t + u^M_{tc}, \end{aligned} \tag{3}$$

where n_c, r_c, m_c are chosen such that for each company c, the variance of the residual terms $u^N_{tc}, u^R_{tc}, u^M_{tc}$ over t is minimized.

Below, the values n_c, r_c, m_c are considered the *instances* of the random variables `gr_sales`, `gr_rd`, and `gr_mcap`, indicating 'growth of net sales', 'growth of R&D expenditures', and 'growth of market capitalization' respectively. Moreover, we define the categorical random variable `sector` attaining s_c for company c. When we talk about sector-wise computation of dependences we actually refer to the *conditional* distribution of `gr_sales`, `gr_rd`, and `gr_mcap`, given `sector`.

Software: We used the R program `compute_growth_rates_scoreboard.R` provided as part of this project. It is called via the command

$$M = \texttt{compute\_growth\_rates\_scoreboard}(\texttt{sector}),$$

where `sector` is a string describing the sector under consideration. Output: $n \times 3$ data matrix M:
1st column: growth rates of Net Sales for all n companies in the respective sector
2nd column: growth rates of R&D for all companies
3nd column: growth rates of Market Capitalization

5.2 Including prior knowledge and simple common sense arguments

Due to the uncertainties of causal discovery methods (Mooij et al., 2016) it is recommended to restrict the search for possible causal structures by prior knowledge up to an extent where the latter can be considered more reliable than purely data driven causal discovery. By common sense, R cannot influence N without delay. It is likely that this delay is even larger than the time scale under consideration. On the other hand, by visual inspection of time series for Net Sales and R & D expenditures, it is obvious that there is a strong correlation in time of these two quantities for most companies, as shown in Figure 8.

Software: To create the plots in Figure 8 we used the R program `plot_NS_RD_for_random_companies.R`, which will be provided as part of this project. It has no input and no output and selects randomly 9 companies and generates the time series plots as pdf files whose names read `company_name.pdf`.

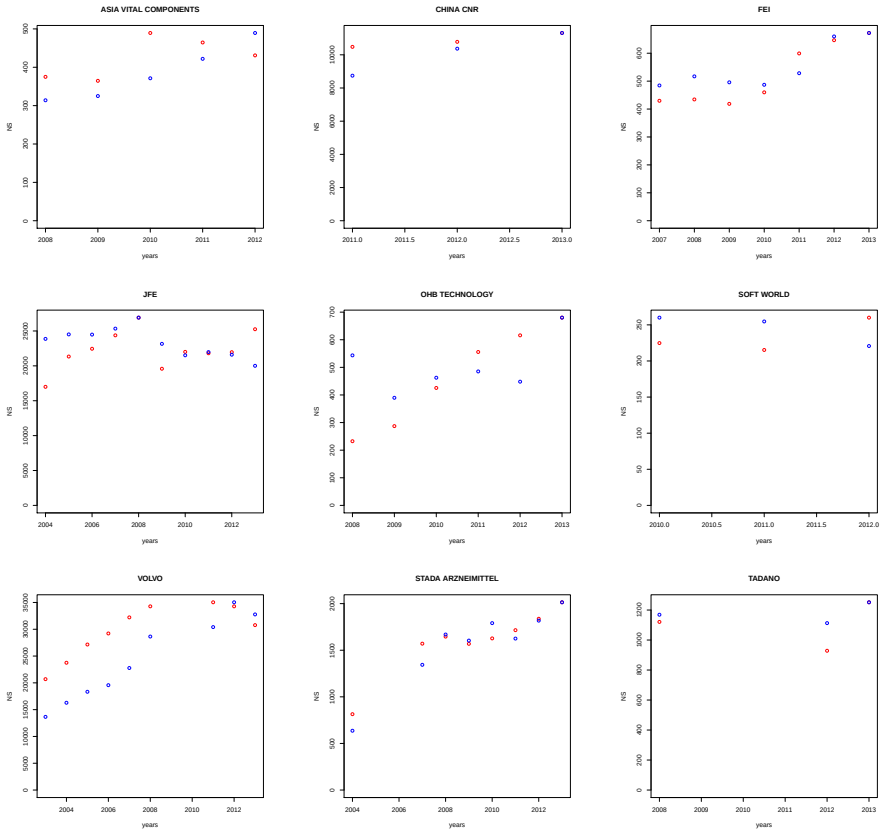

Figure 8: Time evolution of Net Sales (red) and R&D expenditure (blue) of 9 randomly selected companies from the Scoreboard data set. The y-values are normalized to the same maximum. The figures show that increase and decrease tend to occur almost simultaneously.

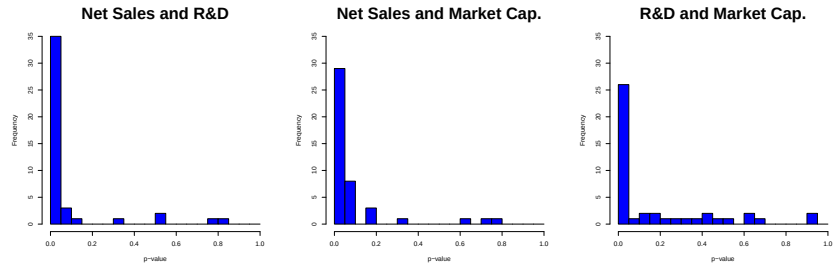

Figure 9: Histogram of p-values for pairwise correlation tests.

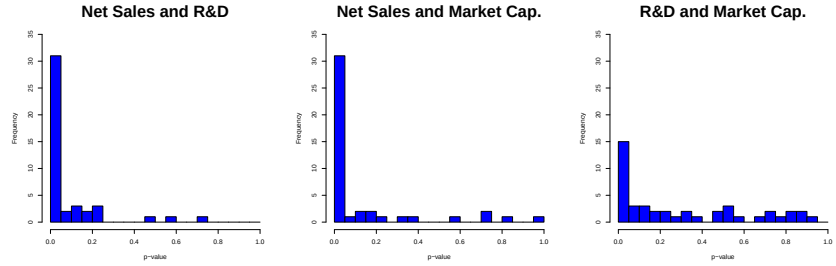

Figure 10: Histogram of p-values for pairwise partial correlation tests given the third variable.

In agreement with Thompson (1999) and by common sense we thus assume that the strong similarities between the above time series are just due to the fact that companies tend to increase $R\&D$ spending when Net Sales increases and, likewise, decrease it with decreasing Net Sales. This suggests a quite strong arrow $\texttt{gr\_sales} \to \texttt{gr\_rd}$.

5.3 Conditional independence based approach

We first check the causal statements that follow from conditional independence based causal discovery methods (Spirtes et al., 1993; Pearl, 2000) which assumes the causal Markov condition and causal faithfulness. Since conditional independence testing is difficult with limited data, we first resort to correlations and partial correlations, which is actually only justified for multi-variate Gaussian data. Since we believe that dependences depend on sectors, we perform the independence testing for each sector separately and display the histogram of the p-values for all sectors.

Figure 9 shows histograms of the p-values of a two-sided correlation test over the sectors. Figure 10 shows histograms of the p-values of a two-sided partial correlation test over the sectors after regressing on the third variable. Identifying dependence with correlatedness and conditional dependence with partial correlatedness (both on the 5% significance level) we observe that the only (conditional) independence that is not rejected for the majority of sectors reads

$$\texttt{gr\_rd} \perp\!\!\!\perp \texttt{gr\_mcap} \,|\texttt{gr\_sales}. \tag{4}$$

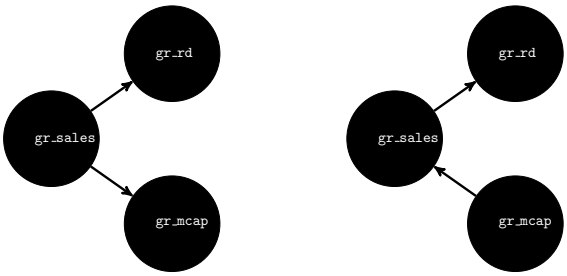

Figure 11: Left: DAGs consistent with the observed conditional independence (4) together with the prior knowledge `gr_sales` → `gr_rd`.

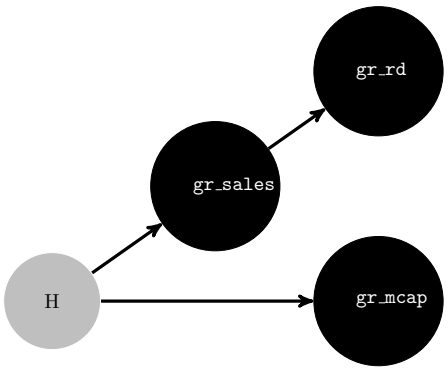

Figure 12: Latent structure that is also compatible with the observed conditional independence pattern.

The only DAGs on the three nodes `gr_rd`, `gr_mcap`, and `gr_sales` that are consistent with such a pattern of conditional independences (according to the causal Markov condition and causal faithfulness (Spirtes et al., 1993; Pearl, 2000)) are the ones depicted in Figure 11.

We cannot necessarily assume that `gr_sales`, `gr_rd`, and `gr_mcap` is a causally sufficient (Spirtes et al., 1993) set of variables, i.e., that there are no unobserved common causes. Therefore, the arrows between `gr_sales`, `gr_rd` and between `gr_sales`, `gr_mcap` in the two DAGs in Figure 11 could also be replaced or complemented by a hidden common cause H, as shown in Figure 12.

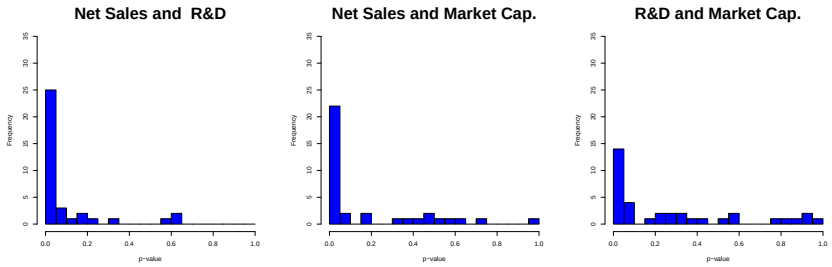

Figure 13: frequencies of sector-wise p-values for pairwise statistical independence via HSIC.

Software: The histograms of p-values for all pairwise correlation and partial correlation tests has been created by the R function `print_correlation_histograms_scoreboard.R` which is provided as part of this project. It has no input and not output and is called by the command

`print_correlation_histograms_scoreboard`().

The histograms are stored as pdf files with the names `correlations_variable1_variable2.pdf` and `partial_correlations_variable1_variable2.pdf`, respectively, where `variable1` and `variable2` run over all pairs out of the three variables `gr_sales`, `gr_rd`, and `gr_mcap`.

To get stronger credibility for the above causal statements we now perform independence tests that are capable of detecting also higher-order statistical dependences beyond correlations using Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2005b). In the infinite sample limit, vanishing HSIC is known to be necessary and sufficient for statistical independence. Figure 13 shows the histograms of p-values for all pairs of variables. `gr_sales` $\perp\!\!\!\perp$ `rd_rd` and `gr_sales` $\perp\!\!\!\perp$ `gr_mcap` are rejected for the majority of sectors, while `gr_rd` $\perp\!\!\!\perp$ `gr_mcap` is rejected for only about half of the cases. However, setting the confidence level to 0.1 instead of 0.05 yielded rejection also in the majority of cases. The fact that the dependence of `gr_rd` and `gr_mcap` is weaker is consistent with the hypothesis that it is mainly mediated by `gr_sales` as in the causal DAGs in Figure 11 and the latent structure in Figure 12.

To test conditional independence for each variable pair, given the third variable, we test HSIC after linearly regressing on the third variable. We use the following simple result:

Lemma 1 *Let X, Y, Z be variables with some arbitrary joint distribution. If there are functions f, g such that $X - f(Z)$ and $Y - g(Z)$ are statistically independent then X and Y are conditionally independent, given Z.*

The proof is immediate: if $X - f(Z)$ and $Y - g(Z)$ are independent then $X - f(z)$ and $Y - g(z)$ are independent, with respect to the conditional distribution $P_{X,Y|Z=z}$ for any possible value z of Z, hence X and Y are conditionally independent, given Z. We thus conclude for the special case of linear functions f, g: if $X - \alpha Z$ and $Y - \beta Z$ are independent for some α, β, then X and Y are

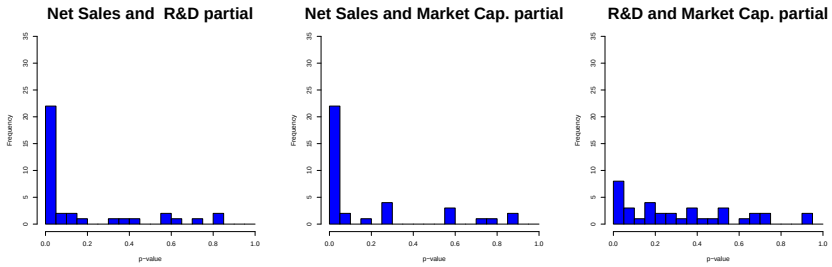

Figure 14: frequencies of sector-wise p-values for pairwise statistical independence via HSIC after linearly regressing on the third variable.

conditionally independent, given Z. This is in particular the case if α, β are chosen according to standard least square regression, that is, if they minimize the variances of $X - \alpha Z$ and $Y - \beta Z$, respectively. To test the independence of $X - \alpha Z$ and $Y - \beta Z$ we use again the HSIC test. The resulting procedure will be called 'partial HSIC test' henceforth. Later we will apply the kernel conditional independence (KCI) test instead to the same data and the compare the results of partial correlations, partial HSIC, and KCI.

The resulting p-values of partial HSIC are shown in Figure 14. We then reject `gr_sales` $\perp\!\!\!\perp$ `gr_rd` $|$`gr_mcap`, `gr_sales` $\perp\!\!\!\perp$ `gr_mcap` $|$`gr_rd` for the majority of cases, while `gr_rd` $\perp\!\!\!\perp$ `gr_mcap` $|$`gr_mcap` is accepted for the majority of sectors. We thus have obtained further evidence for the DAGs in Figure 11 and the latent structure in Figure 12.

Software: The histograms of p-values for all pairwise HSIC and partial HSIC tests have been created by the R function `print_hsic_histograms_scoreboard.R` which is provided as part of this project. It has no input and no output and is called by the command

`print_hsic_histograms_scoreboard`().

The function calls the R function `ind_test.R`, which is also provided. The histograms are stored as pdf files with the names `hsic_variable1__variable2.pdf` and `partial_correlations_variable1__variable2.pdf`, respectively, where `variable1` and `variable2` run over all pairs out of the three variables `gr_sales`, `gr_rd`, and `gr_mcap`.

Note, however, that partial independence is *sufficient* for conditional independence, but *not necessary*. Thus, our particular HSIC test cannot exclude that the variables `gr_sales` and `gr_rd` are conditionally independent, given `gr_mcap` or that `gr_sales` and `gr_mcap` are conditionally independent, given `gr_rd`. The former case is, however, extremely unlikely given that it is generally accepted that `gr_sales` directly influences `gr_rd` as discussed above. Also `gr_sales` and `gr_mcap` are conditionally independent, given `gr_rd`, is quite unlikely given the fact that the relations between the variables do not look very non-linear from visually inspecting scatter plots. Figure 15 shows, for instance the relations between the different growth rates for the sector 'Chemicals'.

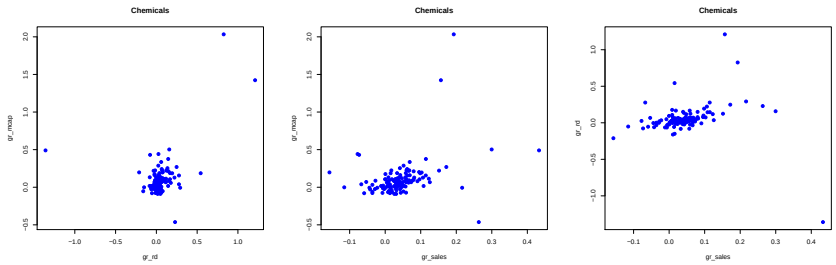

Figure 15: Relations between all pairs out of the variables `gr_sales`, `gr_rd`, and `gr_mcap` for the sector 'Chemicals'.

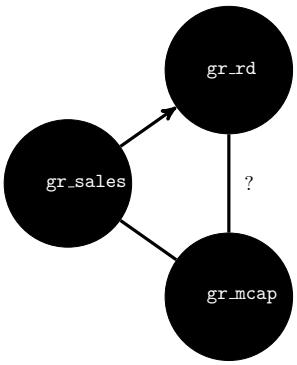

Figure 16: Since there is a weak conditional dependence of `gr_rd` and `gr_mcap` after regressing on `gr_sales`, there may be a weak direct causal link between `gr_mcap` and `gr_rd` that remains to be explored. There could, however, also be a hidden common cause.

Software: To generate the scatter plots in Figure 15 we have used the R function `generate_scatter_plots_growth_rates.R`. It is called via the command

`generate_scatter_plots_growth_rates(sector)`

where `sector` is a string specifying the sector of interest according to the sector names used in the file `2015_12_10_data_panel.csv`. The function calls the R function `compute_growth_rates_scoreboard.R`, which is also provided.

5.4 Discussion of additional weak arrows

Although (4) is rejected only for the minority of sectors on the 5% confidence level (see Figure 10, right), we have to reject the hypothesis that independence holds for *all* sectors because the distribution of p-values significantly deviates from the uniform distribution. Excluding the possibility of confounding (for sake of simplicity), we have to assume – on this fine-grained level – that there is an additional weak arrow with (unknown direction) that directly links `gr_rd` and `gr_mcap`, as depicted in Figure 16.

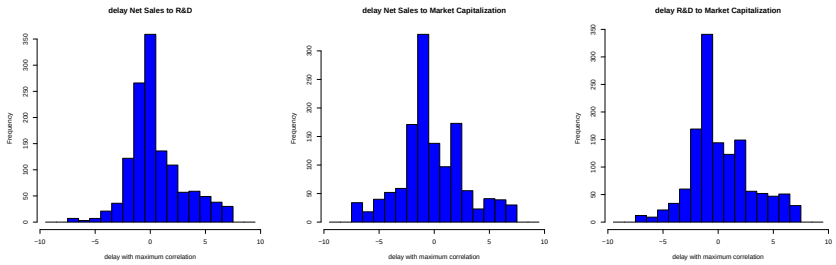

Figure 17: frequencies of company-wise delays. Left: the average delay reads 0.49. Middle: the average delay reads -0.19. Right: the average delay reads 0.23.

5.5 Testing the delays company-wise

Exploring only growth rates blurs the information contained in time delays of statistical dependences. On the one hand, possible time delays in the correlations may exclude some causal directions because the influence cannot go to the past. On the other hand, some causal arrows can be excluded to be instantaneous by common sense. It would be unusual, for instance, to assume that R&D expenditure could have an *immediate* effect on Net Sales, e.g., an effect without delay.

To explore possible delays in the response of one variable on changes of the other, we have done the following analysis. First, we compute for each company in the respective sector 3 vectors containing the logarithms of 'Net Sales', 'R&D Expenditures', and 'Market Capitalization', respectively, over the reported years in the time period of $2008-2013$ (provided that at least 5 years are reported). For each of these vectors we compute a vector containing the differences of adjacent values (i.e., the logarithmic growth factor). For each of the three possible pairs of variables, e.g., 'Net Sales' and 'Market Capitalization', we compute the correlations between the logarithmic growth factors when the difference vectors are shifted against each other. If Δ_N, and Δ_M, for instance, describe vectors of length $l-1$ of logarithmic growth rates of 'Net Sales' and 'Market Capitalization', for a company for which l years in the period $2008-2013$ have been reported, then shifting one of the vectors by k against the other one generates an overlap of $l-1-k$. Whenever this overlap is at least 3, we compute the correlation over the resulting overlap. (for simplicity, this procedure neglects the fact that the reported years are not equidistant). For each company, we picked the k that maximizes the correlation and call it the 'delay' of the correlation. Figure 17 shows histograms that display the number of occurrences of delays over the companies.

Software: To generate the histograms of delays we have used the R function `plot_histogram_correlation_delay.R`. It is called via the command

`plot_histogram_correlation_delay(variable1, variable2)`,

where `variable1` and `variable2` are the names of the variables of interest, i.e., 'Net Sales', as used in the data set `2015_12_10_data_panel.csv`.

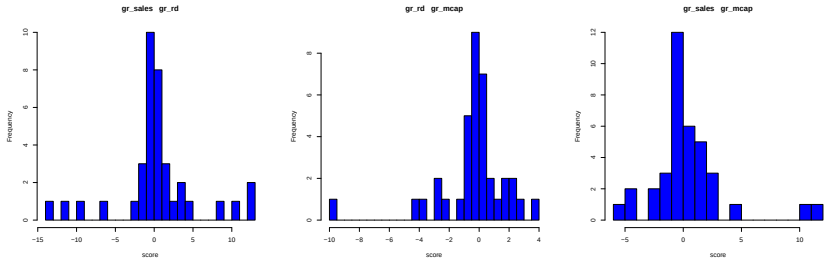

Figure 18: Occurences of scores for LiNGAM over the 42 different sectors. Positive score means that the first variable is inferred to be the cause of the other, negative score indicates the opposite.

The relation between Net Sales and R&D shows no significant delay, see Figure 17, left. This is plausible given a model where R&D expenditures are just a constant fraction of Net Sales. The relation between Net Sales and Market Capitalization is less clear, see Figure 17, middle, although Market Capitalization seems to slightly preceed Net Sales. For the relation between R&D and Market Capitalization, see Figure 17, right, the latter seems to preceed the former by about 1 year. This is plausible, e.g., because the Market Capitalization at the end of last year may influence the decision on R&D expenditure this year.

5.6 LiNGAM test for causal directions

Since conditional independences have provided too tittle information about causal *directions*, we need to include other approaches to causal discovery now. To infer the causal direction between all variable pairs out of , `gr_mcap`, and `gr_rd` we used our own implementation of LiNGAM (Kano and Shimizu, 2003) as explained in Section 4.2 for each sector separately. To this end, computed a score given by the logarithm of the quotient of the p-values for the two possible causal directions. Then we generate a histogram of occurrences of scores over the sectors. The outputs are shown in Figure 18.

Despite the progress regarding new causal inference tools, inferring causal directions from purely observational data (without intervening on a system) remains a hard problem. Accordingly, there is no *certain* method for inferring pairwise causal relations, but only methods that work above change level (Mooij et al., 2016) (that is, the fraction of correct decision compared to fraction of wrong ones is above 1/2). Therefore, it is always insightful to have an addition way of double-checking whether the respective method works for the type of data under consideration. To this end, we consider the arrow `gr_sales` $\rightarrow$ `gr_rd` as ground truth and discuss to what extent it is reproduced by LiNGAM. This will guide us in designing an appropriate decision rule for the other cases. To this end, consider Figure 18, left. For scores whose absolute value is below 1 we should not decide since the p-values differ by less than the factor e. Neglecting these cases, we see that the cases with positive score are slightly more than the ones with negative score (11 versus 8 cases), which slightly supports the causal hypothesis `gr_sales` $\rightarrow$ `gr_rd`, although this evidence is not really convincing.

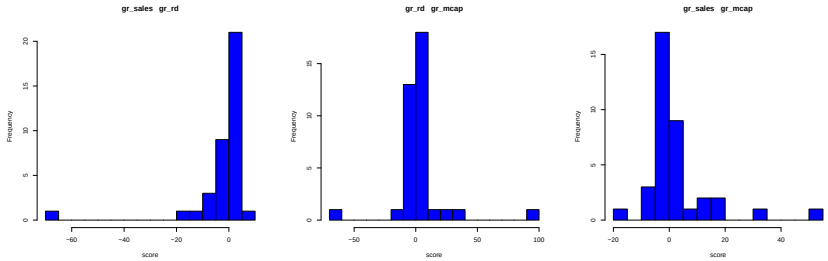

Figure 19: Occurences of scores for non-linear additive noise models over the 42 different sectors. Positive score means that the first variable is inferred to be the cause of the other, negative score indicates the opposite.

Software: To generate the histograms in Figure 18 we have used the R function `plot_histogram_lingam_scores.R` which is called via the command

`plot_histogram_lingam_scores(variable1,variable2)`,

where `variable1,variable2` is a pair out of the three variables `gr_sales`, `gr_mcap`, and `gr_rd`. The histogram is printed to a pdf file called `histogram_lingam_scores_,variable1,__,variable2,.pdf`.

5.7 Non-linear additive noise test for pairwise causal directions

To get further hints on the causal directions we use additive noise models with non-linear functions, as suggested in Hoyer et al. (2009), further explored in Peters et al. (2014) and extensively evaluated in Mooij et al. (2016) and generate the same type of histograms of scores as for LiNGAM. The output is shown in Figure 19. Here, ANM seems to slightly support the hypothesis `gr_rd` $\rightarrow$ `gr_sales`, in contradiction to what we discussed earlier. However, the values of the absolute scores are quite small in most cases. Therefore, we should actually abstain instead of deciding on either of the directions.

Software: To generate the histograms in Figure 19 we have used the R function `plot_histogram_anm_scores.R` which is called via the command

`plot_histogram_anm_scores(variable1,variable2)`,

where `variable1,variable2` is a pair out of the three variables `gr_sales`, `gr_mcap`, and `gr_rd`. The histogram is printed to a pdf file called `histogram_anm_scores_variable1__variable2.pdf`.

5.8 Analysis for selected sectors

It may be insightful to analyze the causal relations for some specific sectors that depend particularly strongly on R&D. One candidate is, for instance, biotechnology. The results for the unconditional independence tests are:

gr_rd ⫫ gr_mcap	gr_sales ⫫ gr_mcap	gr_sales ⫫ gr_rd
0.725	0.201	0.003

For the conditional independence tests we obtain:

gr_rd ⫫ gr_mcap |∗	gr_sales ⫫ gr_mcap |∗	gr_sales ⫫ gr_rd |∗
0.109	0.038	0.000

Here, the symbol $*$ is a placeholder for the third variable which we omitted to save space. The results are, however, subjects to high statistical uncertainty since the dataset contains only 26 samples which is actually to few for reasonable conditional independence testing. Nevertheless, the results are consistent with the results above in the sense that, again, the dependence between `gr_rd` and `gr_mcap` is weaker than for the other two pairs because the causal relation seems to be indirect via the intermediate vaiable `gr_sales`. Accordingly, the conditional independence `gr_rd` ⫫ `gr_mcap` |`gr_sales` is accepted at a 10% significance level.

We also analyzed the sector 'automobiles and parts' and obtained the following results for the unconditional independence tests:

gr_rd ⫫ gr_mcap	gr_sales ⫫ gr_mcap	gr_sales ⫫ gr_rd
0	0	0

For the conditional tests we obtain:

gr_rd ⫫ gr_mcap |∗	gr_sales ⫫ gr_mcap |∗	gr_sales ⫫ gr_rd |∗
0.051	0.000	0.000

Here, all the conditional independences are rejected if we set the significance level to 10%, but `gr_rd` ⫫ `gr_mcap` |`gr_sales` is accepted if we set it to 5%. Since the sample size is 126 and thus larger than for biotechnology, this does not imply that the conditional dependence is stronger than for the sector 'biotechology'. Qualitatively, the result, again, suppoprts the general claim that the causal relation between `gr_rd` and `gr_mcap` is indirect via the intermediate variable `gr_sales`.

Software: conditional independences of growth rates for a selected sector can be obtained by the R function `analyze_scoreboard_for_one_sector.R`.
The command reads:

`analyze_scoreboard_for_one_sector.R()`,

without input. The program then prints a menue of sectors from which the user can choose by typing the respective number and pressing 'enter'. The program then prints the p-vaues of all 3 possible unconditional and 3 conditional independence tests.

5.9 Summary of Scoreboard results

Our analysis of the scoreboard data suggested that the statistical relation between growth of R&D expenditure and growth of market capitalization is indirect via the intermediate variable growth of Net Sales. The causal directions are not clear, Net Sales can be a common cause of both R&D and Market Capitalization or Market Capitalization could influence R&D via Net Sales as an

intermediate variable. Although the observed statistical dependences could also be explained by a model where R&D influences Market Capitalization via Net Sales, we exclude this as unlikely because an influence of R&D on Net Sales is implausible on the time scale of consideration. Possible policy implications would be that, if firms are to be encouraged to increase their R&D investments, this could be stimulated by supporting their sales growth (e.g. by supporting their entry into export market).

One possible objection against the above analysis would be that we don't include many control variables. This objection can be countered by stating that saying that the problem is less important, given that the analysis focuses on differences (i.e. growth rates) rather than levels (i.e. size levels).

We should also mention novel methods for inferring causal directions among time series (Shajarisales et al., 2015), which we have not used here because our time series are probably too short for them. Future work could, however, also try to apply the so-called 'Trace Method' (Janzing et al., 2010; Zscheischler et al., 2011), a novel technique to infer causal directions for pairs of multi-variate variables. Then, one variable would be, for instance, the full time series of Net Sales for each each company and the other the full time series of R&D expenditure. For the Scoreboard data, however, one would be faced with the problem that the time instances do not coincide across companies, which requires some pragmatic preprocessing steps.

6 Analysis of the CIS dataset

6.1 Description of the data set

We analyse data taken from the Community Innovation Surveys (CIS), which are based on the OECD's Oslo Manual, and were administered in several European countries to gather information on the innovative activities of firms. The CIS questionnaire can be found online.[8]

CIS data has been extensively analysed and mined by economists and innovation scholars. While previous datasets on firm-level innovation focused on R&D expenditures and patent counts, CIS data has shed valuable light on other aspects of firm-level innovative activity, although it also has a number of drawbacks, such as being cross-sectional in nature (thus impeding the investigation of lagged effects, or controlling for time-invariant firm-specific heterogeneity), and also having few variables that can serve as valid instrumental variables.

Mairesse and Mohnen (2010), p1138, write: "Basically innovation survey data are of a cross-sectional nature, and it is always problematic to address econometric endogeneity issues and make statements about directions of causality with cross-sectional data. ... we have very few exogenous or environmental variables that can serve as relevant and valid instruments."

Moreover, data confidentiality restrictions often prevent CIS data from being matched to other datasets, or from matching the same firms across different CIS waves. In addition, at time of writing, the 2008 wave is already rather dated.

Given these strengths and limitations, we consider the CIS data to be ideal for our current application, for several reasons:

[8] See `http://ec.europa.eu/eurostat/web/microdata/community-innovation-survey`

- It is a very well-known dataset – hence the performance of our analytical tools will be widely appreciated
- It has been extensively analysed in previous work, but our new tools have the potential to provide new results, therefore enhancing our contribution over and above what was previously reported
- Standard methods for estimating causal effects (e.g. instrumental variables, regression discontinuity design, panel econometrics) are difficult or impossible to apply
- Most variables are not continuous but categorical or binary, which can be problematic for some estimators but not necessarily for our techniques
- Causal estimates based on CIS data will be valuable for innovation policy

To be precise, we focus on the 2008 wave of the CIS, with our raw data covering 16 countries: Bulgaria (BG), Cyprus (CY), Czech Republic (CZ), Germany (DE), Estonia (EE), Spain (ES), Hungary (HU), Ireland (IE), Italy (IT), Lithuania (LT), Latvia (LV), Norway (NO), Portugal (PT), Romania (RO), Slovenia (SI), and Slovakia (SK).

Our data have been deliberately noise-contaminated to anonymise the firms (Mairesse and Mohnen, 2010), p1148; see also (Eurostat, 2009). This is done by capping the continuous variables relating to sales and R&D expenditure, and for the largest values, the true values are not reported, but instead the largest values are approximated.

These countries are pooled together to create a pan-European database. Note however that these countries' databases differ in terms of number of firms, the hence representativeness of the country's overall economy (in terms of representativeness of firms of different sizes, and firms in manufacturing vs services sectors, etc). There is slight variation across countries regarding which questions are asked, and the order in which they appear in the questionnaire (Mairesse and Mohnen, 2010). Furthermore, the data does not accurately represent the proportions of innovative vs non-innovative firms across European countries. Hence, we are not interested in international comparisons.[9] Nevertheless, we argue that this data is sufficient for our purposes of analyzing causal relations between innovation variables in a sample of innovative firms.

6.2 CIS data: methodological choices and remarks on the variables

Appendix A shows a list of the variables in the CIS data set including some that are derived from those one by post-procesing. Previous warnings found in the literature, as well as theoretical considerations, emphasize that endogeneity is a serious concern for many (if not all) of these variables. Some variables are continuous, while others are categorical or binary. Some variables are relatively objective quantities (assuming that respondents accurately provide information), while others are subjective.

[9] See Mairesse and Mohnen (2010), p1140:"it is heroic to make international comparisons when the questionnaires differ in their content, the order of the questions and their formulations, and when the sampling of respondents differs across countries."

Some remarks on variables to be includeds: *Sales growth* is defined by taking log-differences of size (Tornqvist et al., 1985). While most of the literature on firm growth focuses on annual growth rates (see e.g. Coad and Binder (2014) for a survey), our CIS 2008 data only has information on sales for the years 2008 and 2006. Hence we calculate our indicator of biennial sales growth as follows:

$$\texttt{gr\_sales} = \log(\text{sales}(t = 2008)) - \log(\text{sales}(t = 2006)).$$

R&D intensity is calculated as follows:

$$\texttt{rdint} = \text{R\&D/sales}$$

For some of the analysis, we take the intensity, rather than the total amount of R&D expenditure, to avoid scale effects. Focusing on R&D amounts would mean that only the largest firms can have the highest levels of total R&D amounts, with the ranking changing little from year to year because of sheer size effects, while the R&D intensity variable is less sensitive to sheer size effects. Ideally we would have preferred to calculate R&D growth rates, but this is not possible because we don't have information on R&D expenditure in 2006 (only 2008).

Government support: these binary variables refer to whether the firm received public funds to support its innovation activities. Naturally, the literature is interested in the causal effects of government support on firm performance, to evaluate whether or not the intervention was effective. Government support is of course endogenous, because receiving support is not independent of performance, but it could be related either positively ('picking winners') or negatively (i.e. targeting firms that are otherwise experiencing difficulties). Mairesse and Mohnen (2010) explain:"the proper way to estimate the effect of government support is to treat it as an endogenous variable."

Innovation output variables are endogenous: Mairesse and Mohnen (2010), p1144: "the innovation output statistics are much noisier than R&D statistics (probably because they are subjective measures) and need to be instrumented to correct for errors in variables. The endogeneity of innovation outputs in the production function are due to errors of measurement rather than to simultaneity."

Previous work on CIS data has analysed labour productivity, calculated as sales per employee (e.g. Coad, Pellegrino and Savona, 2016). However, in our database we don't have a continuous variable for number of employees (instead, we only have categories such as 50-249 employees, 250+ employees), therefore we don't use it.

Note that the other variables (growth of Net sales, R&D expenditures, capital expenditures, employment and operating profits) have already been analyzed in Coad and Grassano (2015).

Control variables? We focus on the causal relations between these endogenous variables, without pre-processing our variables (a la Coad and Binder (2014)) to adjust for the possibly confounding effect of control variables. There are several reasons why we don't pre-process our variables:

- If Y is binary, then a pre-processed variable $\tilde{Y} := Y - f(X)$ would generally not be binary. However, we want to demonstrate that our techniques can work with binary dependent variables.

- CIS data only contains only a limited set of control variables that are not directly related to innovation (i.e. exporting status, sector and region dummies, and potential group affiliation).
- The predominance of unexplained variance can be interpreted as a limit on how much omitted variable bias (OVB) can be reduced by including the available control variables, because innovative activity is fundamentally difficult to predict. Mairesse and Mohnen (2010), p1142, write: "the unexplained residual, that is, the measure of our ignorance in matters of innovation, is larger than the explained part of the share of total sales due to new products, even more in low tech than in high tech sectors."
- At this initial exploratory stage, we wish to keep our analysis as simple as possible. Further work may refine our baseline model to explore our initial results in more detail.
- Section 2.2 describes a principled concern about including control variables: this can both correct or spoil causal analysis, depending on their causal role.

As reported abovem the CIS data contains only companies whose R&D spending was above a certain threshold. We have further selected the ones with nonzero in-house R&D spending (i.e., `rrdinx` $\neq 0$), which is a criterion that seems a bit more apparent than the arbitrary cut of the list according to ranks in R&D spending. Nevertheless, also this decision can be questioned. To explore whether this generates selection bias which influences the causal analysis, we have replicated some of the crucial parts of the analysis without the conditioning on nonzero `rrdinx`. The results of the replication are discuused in Section 6.12.

6.3 Specific software tools

Since this data set plays the main role for this study, we have developed software tools that allow for convenient data analysis by high-level commands. We found it particularly helpful to call variables by the names used in the headlines of the data file (which will be, at the same time, used in our text and figures).

First, we have the following command that loads selected variables specified by their names:

Software: The R function `load_cis_data.R` loads a set of variables (specified by their names as in the headline of the file `cis2008_MASTER.csv`) into a matrix. It is called via the command

$$M = \texttt{load\_cis\_data}(\texttt{variables}),$$

where `variables` is a vector of strings containing the names of the desired variables. It is defined by the command

$$\texttt{variables} = c(\texttt{var1}, ..., \texttt{var2})).$$

One can either type the two commands

$$\begin{aligned} \texttt{variables} &= c('\texttt{rrdinx}','\texttt{turn06m}') \\ M &= \texttt{load\_cis\_data}(\texttt{variables}), \end{aligned}$$

or simply

$$M = \texttt{load\_cis\_data}(c('\texttt{rrdinx},'\texttt{turn06m}'))$$

to load the variables `rrdinx` and `turn06m`. Note that $\texttt{c}(.,.,\ldots,.)$ is the standard R command to create a vector. The matrix M the contains the values of these variables as its two columns.

To perform systematic unconditional and conditional independence tests we have the following tool:

Software: A systematic exploration of all possible conditional and unconditional independence tests with the set of variables in the CIS data can be performed via the the R function `cis_arbitrary_subsets.R` which is called by the following command:

$$(*) \quad \texttt{cis\_arbitrary\_subsets}(\texttt{variables}), \tag{5}$$

where `variables` is a vector of strings containing the set of variables under investigation. Example: first type

$$\texttt{variables} = c(\texttt{'turn06m'},\texttt{'turn08m'},\texttt{'rrdinx'},\texttt{'orgbup'}).$$

Then type the command $(*)$ and the program performs all possible $\binom{4}{2}$ unconditional independence tests as well as all $3 \cdot \binom{4}{2}$ possible conditional independence tests (note that our software only allows for conditioning on 1 variable). The program generates two files, one containing the results of all *unconditional* independence tests, called, for instance `test_results_turn06m_turn08m_rrdinx_orgbup.txt` for the above example. The second file contains the results of all *conditional* independence tests, called, for instance, `cond_test_results_turn06m_turn08m_rrdinx_orgbup.txt`. Each row of the first file consists of the following three entries:

`X` & `Y` & `p-value for X` $\perp\!\!\!\perp$ `Y`

Likewise, each row of the second file consists of the following four entries:

`X` & `Y` & `Z` & `p-value for X` $\perp\!\!\!\perp$ `Y` $|$`Z`

To infer pairwise causal directions between continuous variables via additive noise models, we have the following tool:

Software: To run ANM on arbitrary variables in the CIS data set, we have used the R function `pairwise_anm_cis.R` which is called via the command

$$\texttt{score} = \texttt{pairwise\_anm\_cis}(\texttt{variable1},\texttt{variable2}),$$

where `variable1` and `variable2` may be any variable names among the ones from the header of file `cis2008_MASTER.csv`. The output `score` between $-\infty$ and ∞ indicates the inferred causal direction. A positive value means that the method infers `variable1` $\rightarrow$ `variable2` and a negative one the opposite causal direction. The absolute value indicates the strength of the evidence. The R function calls the matlab functions `cep_anm.m` described in Section 4.2.

For the same purpose of inferring directions between continuous variables we also have linear non-Gaussian models (LiNGAM) which can be used via the following tool:

Software: To run LiNGAM on arbitrary pairs of variables in the CIS data set, we have used the R function `pairwise_lingam_cis.R` which is called via the command

`pairwise_lingam_cis(variable1,variable2)`,

where `variable1` and `variable2` may be any variable names among the ones from the header of file `cis2008_MASTER.csv`. The output is written to the command line and consists of the p-values for both possible causal directions. The R function calls `ind_test.R` (which, in turn calls the KCI test implemented in matlab), as described earlier.

The causal direction between a pair of discrete variables from the CIS data set can be inferred via the following tool:

Software: To infer the causal direction for one pair of variables from the CIS data set we use the R function `apply_discrete_anm.R`. It is called via the command

$$\texttt{pvalues} = \texttt{apply\_discrete\_anm}(\texttt{variable1},\texttt{variable2}),$$

where `variable1` and `variable2` is the variable name used in `cis2008_MASTER.csv`, using '...' to indicate that the input is read as string. This function calls the matlab function `fit_both_dir_discrete.m` used in Peters et al. (2011). The output `pvalues` is a vector containing the p-value for a discrete additive noise model from `variable1` to `variable2` and the p-value for an additive noise model in the converse direction.

The following tool infers the directions for all pairs of a set of variables from the CIS data set specified by the input:

Software: To infer all pairwise causal directions within a subset of discrete variables contained in the CIS dataset we have used the R function `infer_causal_direction_for_all_pairs_discrete.R`. It is called via the command

$$(*) \quad \texttt{infer\_causal\_direction\_for\_all\_pairs\_discrete}(\texttt{variables}),$$

where `variables` is a vector of strings containing the set of variables under investigation. Example: first type

$$\texttt{variables} = c('\texttt{orange}','\texttt{orepl}','\texttt{oenmk}','\texttt{oimks}').$$

Then type the command $(*)$ and the program infers all pairwise directions for the possible $\binom{4}{2}$ pairs. Note that $\texttt{c}(.,.,\ldots,.)$ is the standard R command to create a vector. The function generates a text file of the form `discrete_anm_orange_orepl_oenmk_oimks.txt` It contains rows of the following form:

$$X \quad \& \quad Y \quad \& \quad \text{p-value for } X \to Y \quad \& \quad \text{p-value for } X \leftarrow Y$$

6.4 Classification of variables

For this study, it is important to emphasize that the questionaire contains quite different types of variables in several respects. On the one hand, they contain continuous, discrete, and categorical variables. By the latter we mean variables whose values cannot be interpreted as numbers in any straightforward way – as opposed to variables like company size that are given by discretizing the number of employees. For causal inference methods, this distinction is important because different types require different inference methods. The second aspect with respect to which the variables differ are whether they describe 'hard facts' or whether they rely on the opinion of the person who answered the questionaire, for instance whether there were significant innovations regarding organisational structures. – Mostly, the questionaire consists of the latter. In this case, there is a serious limitation for causal inference methods because statistical dependences between the variables can be expect to be due to the fact that they were answered by the same person. Even if 100 different persons are asked to answer the questionaire for the same company the answers will have statistical spread and, probably, the answers for different type of innovations will correlate. For instance, one person may consider all innovations significant while another one considers them insignificant. These kind of statistical dependences are due to the common cause 'attitude of the subject who answered the question' and further causal discovery becomes pointless. It is hard to assess to what extent statistical dependences between different answers are due to such a 'confouding by subjectivity'. At least, we should not be surprised to obtain heavily connected causal graphs for variables that are heavily subjective and, for the same reason, we cannot necessarily expect that causal directions can be inferred.

We first list the variables that we consider more or less 'hard facts', although they may also contain some uncertainty:

- Market: `marloc, marnat, mareur, maroth, larmar`
- Turnover: `turn06m, turn08m`
- R&D Expenditure: `rrdinx`
- Funding: `funloc, fungmt, funeu`
- Company size: `emp06, emp08`

6.5 Sanity check of the causal inference tools

To what extent causal inference tools generate reasonable results may in principle depend on the domain under consideration. This is because every causal inference method relies on assumptions and whether they hold or not can only be judged from *empirical* evidence and not from theoretical insights alone. To explain this, we briefly review the crucial assumptions. First, the *causal Markov condition* can only be applied to a set of observed variables if the set is *causally sufficient*, i.e., there are no unobserved common causes of the observed variables. Second, the limitations of *causal faithfulness* are still an issue of debates. On the one hand, random choices of parameters are not unlikely to generate distributions that are close to unfaithful. In other words, although accidental conditional independences that are not implied by the causal structure occur

X	Y	Z	p-value
turn06m	rrdinx	turn08m	0.099790
turn08m	rrdinx	turn06m	0.152495
turn06m	turn08m	rrdinx	0.000000

Figure 20: p-values of the conditional independence test $X \perp\!\!\!\perp Y \,|Z$ for the three possible choices of the conditioning variable Z in turn06m, turn08m, rrdinx. Here, we reject independence if the p-values are below 0.1. Much lower p-values indicate highly significant dependences.

with probability zero, it often happen that the dependences are so weak that they are not detected with limited data. On the other hand, it has been argued that, for instance in biology, causal relations may violate faithfulness (Spirtes et al., 1993) because evolution may 'intentionally' have fine-tuned parameters in a way that generates additional (conditional) independences. Also technical systems show this kind of behavior because control mechanisms are constructed in a way that one causal path compensates the influence of an undesired factor. Moreover, testing conditional independences from finite data relies on implicit assumptions like *smoothness* of distributions. After all, for every given independence test and every sample size it is possible to construct probability distributions whose statistical dependences are invisible for the respective test and sample size. The novel method of causal inference via *additive noise* depends (on top of the additivity of noise) on *non-linearity* or *non-Gaussianity* – again, assumptions that may hold for one domain but not the other.

For the above reasons, we have tried to identify variables for which causal relations are more or less known in order to check whether the believed causal structure is consistent with the results obtained by causal inference tools. To this end, we consider `turn06m`, `turn08m`, `rrdinx`, which describe the total turnover for 2006, the same for 2008, and the R&D expenditures in 2008, respectively.

In Section 5 we have already argued that an influence of R&D expenditure on Net Sales can be excluded on the given time scale. Moreover, no causal influence can go backwards in time. Therefore, we exclude arrows `turn08m` $\rightarrow$ `rrdinx`, `rrdinx` $\rightarrow$ `turn06m`, and `rrdinx` $\rightarrow$ `turn08m`. We first check *unconditional* independences via the KCI test described in Section 4.1 (which is actually a conditional independence test which can also be used for unconditional independence testing). We have randomly selected 2000 companies in each run to reduce the computational load since the latter strongly increases with sample size. The p-values are 0.000000 for all three variable pairs, which shows (not surprising) that all three show strong pairwise statistical dependences. Figure 20 shows the p-values of all three possible conditional independence tests. Based on the significance level 0.1, we accept turn08m $\perp\!\!\!\perp$ turn06m $|$rrdinx. This suggests a DAG whose skeleton (that is, the corresponding undirected graph obtained by ignoring directions) is shown in Figure 21.

To infer the directions of the direct causal links, we apply pairwise ANM and obtained the following scores:

X	Y	score
`turn06m`	`turn08m`	0.007098
`turn06m`	`rrdinx`	-0.047896

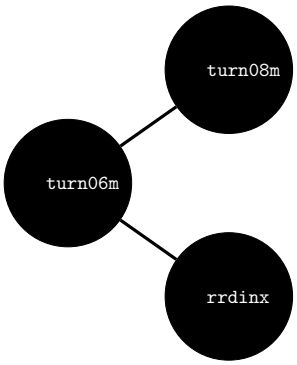

Figure 21: Undirected graph suggested by the independence patterns in Fig. 20. The pattern does not allow a v-structure at `turn06m`, in agreement with common sense.

X	Y	p-value
rrdinx	funloc	0.055210
rrdinx	fungmt	0.000000
rrdinx	funeu	0.000000
funloc	fungmt	0.039232
funloc	funeu	0.000000
fungmt	funeu	0.000000

Figure 22: p-values of the independence test $X \perp\!\!\!\perp Y$ for all six variable pairs in `rrdinx`, `funloc`, `fungmt`, `fungmt`.

The first score is extremely small, but has the right sign, while the second one suggests an arrow to the past, which is obviously impossible. The example has to be taken as a warning that reminds us of the fact that causal inference rules are not certain, but only provide hints. Mooij et al. (2016) report a fraction of correct decisions (if one decides in all cases) that is between 60% and 70%, the ratio gets higher, fortunately, when decisions are only made for higher absolute values of the scores.

6.6 R&D and public funding

To start with some variables that we count as describing hard facts, we first explore the relation between in house R&D expenditure `rrdinx` on the one hand and, on the other hand, the binary variables `funloc`, `fungmt`, `funeu` describing whether the company received funding from local, governmental, or European funding agencies, respectively. The result of the unconditional tests are shown in Figure 22. Based on our 0.1 confidence level, we reject all independences. However, remarkably, `funloc` is not tightly connected to the other variables. Maybe local funding agencies have criteria for funding that differ significantly from governmental and European ones. Figure 23 shows the p-values of all conditional independences. Based on a confidence level of 0.1, we accept the following conditional independences:

X	Y	Z	p-value
rrdinx	funloc	fungmt	0.261488
rrdinx	funloc	funeu	0.396313
rrdinx	fungmt	funloc	0.000000
rrdinx	fungmt	funeu	0.000000
rrdinx	funeu	funloc	0.000011
rrdinx	funeu	fungmt	0.000036
funloc	fungmt	rrdinx	0.538130
funloc	fungmt	funeu	0.412634
funloc	funeu	rrdinx	0.000052
funloc	funeu	fungmt	0.000031
fungmt	funeu	rrdinx	0.219991
fungmt	funeu	funloc	0.000126

Figure 23: p-values of the independence test $X \perp\!\!\!\perp Y \,|Z$ for all 12 possible choices of X, Y, Z (accounting for the equivalence of $X \perp\!\!\!\perp Y \,|Z$ and $Y \perp\!\!\!\perp X \,|Z$) out of `rrdinx`, `funloc`, `fungmt`, `fungmt`.

`rrdinx`	$\perp\!\!\!\perp$	`funloc`	| `fungmt`
`rrdinx`	$\perp\!\!\!\perp$	`funloc`	| `funeu`
`funloc`	$\perp\!\!\!\perp$	`fungmt`	| `rrdinx`
`funloc`	$\perp\!\!\!\perp$	`fungmt`	| `funeu`
`fungmt`	$\perp\!\!\!\perp$	`funeu`	| `rrdinx`

The above pattern of (conditional and unconditional) independences yields the skeleton depicted in Figure 24. R&D expenditure `rrdinx` screens off the dependences between governmental and European funding, `fungmt` and `funeu`. According to faithfulness, this is only possible if at most one of the funding variables have an arrow pointing towards `rrdinx` since `rrdinx` would otherwise be a collider and conditioning on `rrdinx` would render them dependent.

Given that we can rule out the case where both arrowheads point to `rrdinx`, this implies that we can rule out the possibility that R&D has no causal effect on receipt of funding. Instead, our results suggest that R&D has a causal effect on receipt of funding (but it could also be the case that funding has feedback effects on R&D). This has interesting implications for the evaluation of R&D funding schemes.

We now discuss some causal directions. The causal relation between funding and in house R&D expenditure may in principle be in both directions (apart from hidden common causes, which can never be excluded): first, companies that already spend a significant amount of money on R&D may have better chances to receive public funding when they apply for it. On the other hand, funding could trigger a process of innovation that influences the company to spend more own money on R&D. It could, however, also be the case that companies use public funding to reduce their own expenditures. All these different causal explanations for the statistical relation between `rrdinx` and `funeu` or `fungmt` could be simultaneously true. As example for a common cause relation, on may think of a company's decision to increase R&D, both by increasing in house R&D and by applying for external funding. In this case, neither of the variables `rrdinx` and `funeu` (or `fungmt`, respectively) can be considered as the

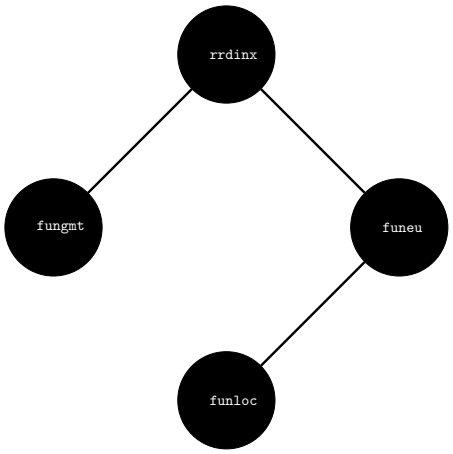

Figure 24: Undirected graph resulting from the independence pattern of `rrdinx`, `fungmt`, `funloc`, `funeu`.

cause of the other. Instead the company's R&D-friendly attitude is the common cause of both.

To see what the data says, we consider the joint distribution of `rrdin` and `funeu` or `fungmt` or `funloc`. First, we observe that the correlation between `rrdinx` and `funeu` reads 0.10 with a 0.95 confidence interval being $[0.08, 0.13]$, i.e., we obtain a significant positive correlation. If the relation between `rrdinx` and `funeu` was completely due to the fact that companies reduce their own R&D spending because they receive funding, it would be negative. Yet, we cannot exclude that this happens, but then an additional explanation for the positive correlation is required. Likewise, we obtain a slightly larger correlation of 0.13 for `rrdinx` and `fungmt`, with the confidence interval $[0.10, 0.15]$ (while the correlation between `rrdinx` and `funloc` is only 0.019, which is not significant because the confidence interval reads $[-0.0039, 0.041]$).

The conjecture that companies that also spend own money on R&D are more likely to get accepted when they apply for funding seems, a priori, a plausible explanation for the positive correlations. Moreover, companies that spend much money on R&D may, after running their projects, may get the demand for funding. Both types of explanations for the relation between R&D and funding describe it as the causal direction `rrdinx` $\rightarrow$ `funding` where `funding` stands for `funeu`, `fungmt`, `funloc`, respectively. To check the plausibility of this causal hypothesis, we check whether the conditionals $P(\texttt{funeu}|\texttt{rdinx})$ are smooth because this provides some support for this causal direction. This is because conditionals of the cause, given the effect tend to be less smooth and simple than conditionals of the effect, given the cause (although there is no formal criterion known apart from the ideas explained and cited in Section 4.3). Figure 25 visualizes the conditional distributions for the case where the binary is given by European, governmental, and local funding, respectively. As a rough tendency, we would assign smooth curves rather with causal influence in the corresponding direction, while volatile curves suggest that the conditional does not describe the correct causal direction or that the relation is due to common

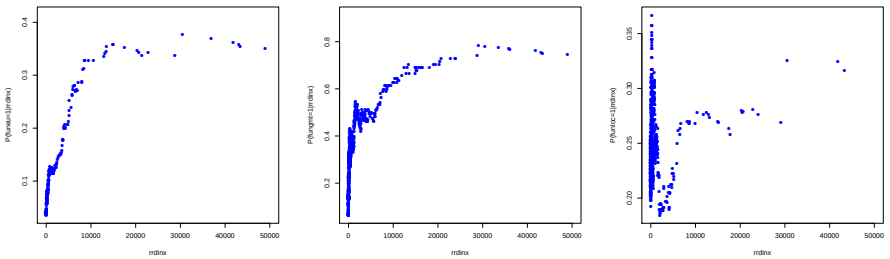

Figure 25: Visualization of the conditional distributions $P(\texttt{bin} = \texttt{1}|\texttt{rddinx})$ for the binary variables `bin`= `funeu`, `fungmt`, `funloc`, respectively. For `funeu` and `fungmt` the dependence is quite smooth, apart from the region of small values of `rrdinx`. As comparison, for `funloc` the relation is quite random, which is not surprising given the result that the HSIC independence tests have accepted `funloc` $\perp\!\!\!\perp$ `rrdinx`.

X	Y	p-values
`rrdinx`	`turn08m`	0.000000
`rrdinx`	`funeu`	0.000000
`rrdinx`	`fungmt`	0.000000
`rrdinx`	`funloc`	0.212637
`turn08m`	`funeu`	0.000065
`turn08m`	`fungmt`	0.000000
`turn08m`	`funloc`	0.468859
`funeu`	`fungmt`	0.000000
`funeu`	`funloc`	0.000000
`fungmt`	`funloc`	0.000000

Figure 26: pvalues of the unconditional independence tests for the variables `rrdinx`, `turn08m`, `funeu`, `fungmt`, `funloc`.

causes. This is, however, a very preliminary statement which actually refers to ongoing research.

Including Net Sales Since we have no algorithms to infer the causal direction between continuous and discrete variables (apart from some preliminary proposals like (Janzing et al., 2009)), it may be helpful to include additional continuous variables because inferring causal directions between them could help in inferring the causal directions between continuous and discrete ones (given the pattern of conditional independences). For this reason, we now consider the variables `rrdinx`, `turn08m`, `funeu`, `fungmt`, `funloc`. The p-values for the unconditional independence tests are shown in Figure 26. We thus accept the following independences:

$$\begin{array}{lcl} \texttt{rrdinx} & \perp\!\!\!\perp & \texttt{funloc} \\ \texttt{turn08m} & \perp\!\!\!\perp & \texttt{funloc} \end{array}$$

The p-values for the conditional independence tests are shown in Figure 27.

X	Y	Z	p-values
rrdinx	turn08m	funeu	0.000000
rrdinx	turn08m	fungmt	0.000000
rrdinx	turn08m	funloc	0.000000
rrdinx	funeu	turn08m	0.007939
rrdinx	funeu	fungmt	0.000137
rrdinx	funeu	funloc	0.000032
rrdinx	fungmt	turn08m	0.000000
rrdinx	fungmt	funeu	0.000000
rrdinx	fungmt	funloc	0.000000
rrdinx	funloc	turn08m	0.168283
rrdinx	funloc	funeu	0.743933
rrdinx	funloc	fungmt	0.266900
turn08m	funeu	rrdinx	0.010503
turn08m	funeu	fungmt	0.506355
turn08m	funeu	funloc	0.017367
turn08m	fungmt	rrdinx	0.003421
turn08m	fungmt	funeu	0.004939
turn08m	fungmt	funloc	0.000672
turn08m	funloc	rrdinx	0.000000
turn08m	funloc	funeu	0.509219
turn08m	funloc	fungmt	0.661349
funeu	fungmt	rrdinx	0.680689
funeu	fungmt	turn08m	0.000000
funeu	fungmt	funloc	0.000083
funeu	funloc	rrdinx	0.000210
funeu	funloc	turn08m	0.004008
funeu	funloc	fungmt	0.229628
fungmt	funloc	rrdinx	0.335725
fungmt	funloc	turn08m	0.000000
fungmt	funloc	funeu	0.000000

Figure 27: p-values for all conditional independence tests for the variables rrdinx, turn08m, funeu, fungmt, funloc.

We thus accept the following conditional independences:

$$\begin{array}{lcl|l} \texttt{rrdinx} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{turn08m} \\ \texttt{rrdinx} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{funeu} \\ \texttt{rrdinx} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{fungmt} \\ \texttt{turn08m} & \perp\!\!\!\perp & \texttt{funeu} & \texttt{fungmt} \\ \texttt{turn08m} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{funeu} \\ \texttt{turn08m} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{fungmt} \\ \texttt{funeu} & \perp\!\!\!\perp & \texttt{fungmt} & \texttt{rrdinx} \\ \texttt{funeu} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{fungmt} \\ \texttt{fungmt} & \perp\!\!\!\perp & \texttt{funloc} & \texttt{rrdinx} \end{array}$$

Figure 28 shows the edges resulting from the remaining dependences.

Here we see a violation of faithfulness because funloc is disconnected to all other variables although it gets dependent to some of them when conditioning on

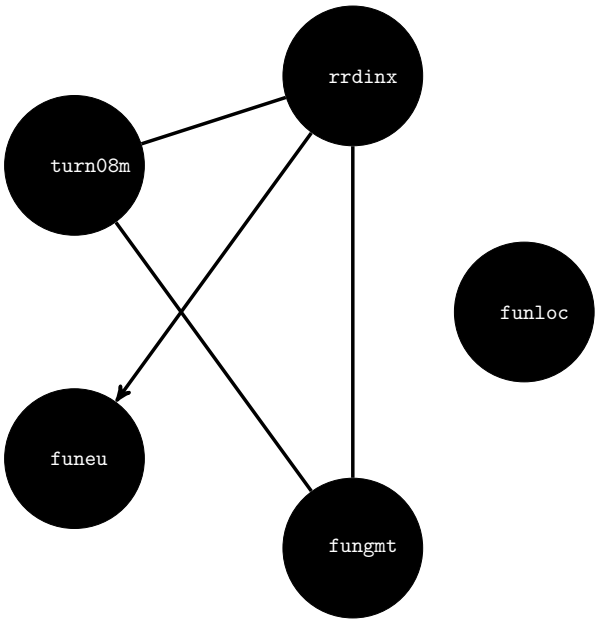

Figure 28: Partially directed graph resulting from the independence pattern of `rrdinx`, `turn08m`, `funeu`, `fungmt`, `funloc`.

others. For instance, we have rejected the independence of `fungmt` and `funloc`, given `funeu`, with high significance. For this reason, we will now focus on the remaining variables instead of `funloc` because its causal role remains unclear for the moment. Moreover, the pattern of independences tells us something about the causal directions, namely that the edge `funeu` – `rrdinx` has its arrowhead at `funeu`, otherwise `funeu` would be independent of all other variables. Given our previous discussion on the discontinuity of the acceptance rate, it is, however, likely that this arrow is actually bidirectional. Unfortunately, bidirectional causal influence goes beyond the formalism of Spirtes et al. (1993); Pearl (2000) and most of the other available causal inference tools as well, except for some very special scenarios as the one with additive Gaussian noise studied by Mooij et al. (2011). We are therefore not able to explore this relation any further.

To discuss directions of the remaining arrows, note that we have argued earlier that Net Sales influences R&D, not vice versa. This statement, however, referred actually to the growth rates of these quantities on the particular short time scale considered in the Scoreboard data set. Long terms influence of R&D on Net Sales cannot be excluded. On the contrary, they are hoped to exist. In cross sectional data, they could in principle be apparent since the statistical relation could be from such a long-term effect. Therefore, we test the causal direction using both tools from our tool box, namely ANM and LiNGAM. For LiNGAM, the p-values are zero for both directions. ANM prefers the direction `rrdinx` $\rightarrow$ `turn08m`, but with the insignificantly low score -0.04.

Sales growth and R&D intensity instead of absolute values Instead of `turn08m`, i.e., the absolute value of Net Sales, we now consider Sales growth

X	Y	p-values
rdint	gr_sales	0.000431
rdint	funeu	0.007419
rdint	fungmt	0.000000
rdint	funloc	0.000018
gr_sales	funeu	0.638588
gr_sales	fungmt	0.291060
gr_sales	funloc	0.004793
funeu	fungmt	0.000548
funeu	funloc	0.000000
fungmt	funloc	0.336662

Figure 29: p-values for unconditional independence tests for all pairs in the set rdint, gr_sales, funeu, fungmt, funloc.

gr_sales defined by log(turn08m/turn06m) and R&D intensity rdint defined by the quotient rrdinx/turn08m. Thus, we now consider the variables rdint, gr_sales, funeu, fungmt, funloc. The p-values for the unconditional independence tests are shown in Figure 29. Accordingly, we accept the following independences:

$$\begin{aligned} \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{funeu} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{fungmt} \\ \texttt{fungmt} &\perp\!\!\!\perp \texttt{funloc} \end{aligned}$$

Figure 30 shows the p-values for all conditional independence tests. We thus accept the following conditional independences:

$$\begin{aligned} \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{funeu} \mid \texttt{rdint} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{funeu} \mid \texttt{fungmt} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{funeu} \mid \texttt{funloc} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{fungmt} \mid \texttt{rdint} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{fungmt} \mid \texttt{funeu} \\ \texttt{gr\_sales} &\perp\!\!\!\perp \texttt{fungmt} \mid \texttt{funloc} \\ \texttt{fungmt} &\perp\!\!\!\perp \texttt{funloc} \mid \texttt{rdint} \\ \texttt{fungmt} &\perp\!\!\!\perp \texttt{funloc} \mid \texttt{gr\_sales} \\ \texttt{fungmt} &\perp\!\!\!\perp \texttt{funloc} \mid \texttt{funeu} \end{aligned}$$

From the above pattern, we have constructed the skeleton shown in Figure 31.

Summary on R&D and public funding In this section, we looked at the relationships between funding variables and in-house R&D expenditure (and sometimes including net sales). Conditional independence tests suggested that local/regional funding was not tightly connected to the other variables. One suggestion for policy might be to investigate why support for innovation, when given by regional funds, seems to have different characteristics. Our results also suggest that R&D spending has a causal influence on receipt of funding (although there could also be a feedback effect of funding on further R&D spending). This highlights how public support for innovation should always pay

X	Y	Z	p-values
rdint	gr_sales	funeu	0.000008
rdint	gr_sales	fungmt	0.000263
rdint	gr_sales	funloc	0.002595
rdint	funeu	gr_sales	0.000986
rdint	funeu	fungmt	0.028242
rdint	funeu	funloc	0.093280
rdint	fungmt	gr_sales	0.000000
rdint	fungmt	funeu	0.000000
rdint	fungmt	funloc	0.000000
rdint	funloc	gr_sales	0.000019
rdint	funloc	funeu	0.072072
rdint	funloc	fungmt	0.000001
gr_sales	funeu	rdint	0.126534
gr_sales	funeu	fungmt	0.414491
gr_sales	funeu	funloc	0.370240
gr_sales	fungmt	rdint	0.261253
gr_sales	fungmt	funeu	0.286464
gr_sales	fungmt	funloc	0.145403
gr_sales	funloc	rdint	0.032227
gr_sales	funloc	funeu	0.008407
gr_sales	funloc	fungmt	0.043809
funeu	fungmt	rdint	0.021619
funeu	fungmt	gr_sales	0.021313
funeu	fungmt	funloc	0.000008
funeu	funloc	rdint	0.000000
funeu	funloc	gr_sales	0.000000
funeu	funloc	fungmt	0.000001
fungmt	funloc	rdint	0.488108
fungmt	funloc	gr_sales	0.201702
fungmt	funloc	funeu	0.305139

Figure 30: p-values for all conditional independence tests in the set `rdint`, `gr_sales`, `funeu`, `fungmt`, `funloc`.

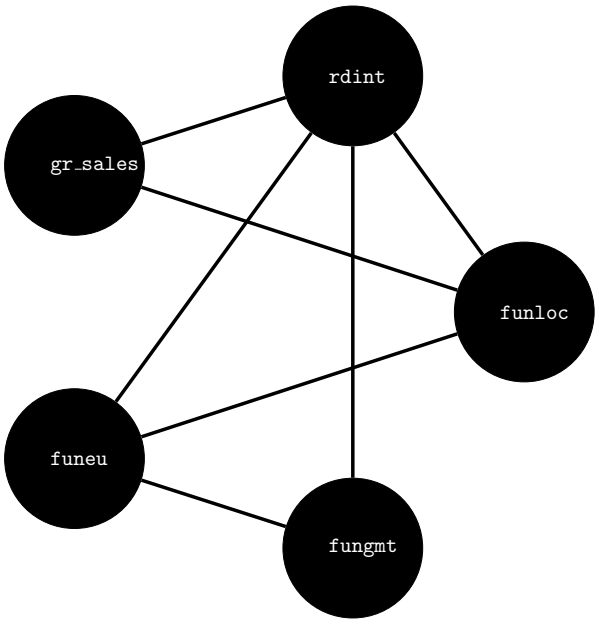

Figure 31: Undirected graph resulting from the independence pattern of `rdint`, `gr_sales`, `funeu`, `fungmt`, `funloc`.

heed to the additionality of innovation support funds given.

6.7 Relating R&D with variables describing organizational innovations

We now consider the variables `orgbup` (new business practice), `orgwkp` (new methods of organising work responsibilities and decision making), and `orgexr` (new methods of organising external relations) describing organizational innovations with the variables `gr_sales` (defined by log `turn08m`/`turn06m` and rdint (defined by `rrdinx`/`turn08m`). The correlations read:

	orgbup	orgwkp	orgexr	gr_sales
orgwkp	0.48			
orgexr	0.34	0.44		
gr_sales	0.062	0.070	0.048	
rdint	-0.07	-0.024	-0.032	-0.024

The different organisational innovation variables are strongly correlated, but their correlation to sales growth is weaker than their correlation to R&D intensity. It is therefore interesting to understand the causal relation between organisational innovations and R&D intensity. It is remarkable that the strength of the correlation does not coincide with the significance of the HSIC dependence: for instance, sales growth is significantly dependent on several organisational innovation variables.

The p-values of the unconditional independence tests for all 10 possible pairs of variables are shown in Figure 32. On the 0.1 confidence level, we thus accept

X		p-value
orgbup	orgwkp	0.000000
orgbup	orgexr	0.000000
orgbup	rdint	0.000001
orgbup	gr_sales	0.169267
orgwkp	orgexr	0.000000
orgwkp	rdint	0.097314
orgwkp	gr_sales	0.000033
orgexr	rdint	0.227833
orgexr	gr_sales	0.000036
rdint	gr_sales	0.000001

Figure 32: p-values for $X \perp\!\!\!\perp Y$ for all pairs (X, Y) out of the set `orgbup`, `orgwkp`, `orgexr`, `gr_sales`, and `rdint`.

the following unconditional independences:

$$\begin{array}{rcl} \texttt{orgbup} & \perp\!\!\!\perp & \texttt{gr\_sales} \\ \texttt{orgex} & \perp\!\!\!\perp & \texttt{rdint} \end{array}$$

The hypothesis orgwkp $\perp\!\!\!\perp$ `rdint` is only close to being accepted. Hence, innovation with respect to new business practice is unrelated to sales growth. Likewise, innovations regarding organising work responsibilities and decision making is not related to R&D intensity.

The results of the conditional independence tests are shown in Figure 33. We thus accept the following conditional independences

$$\begin{array}{rcl|l} \texttt{orgbup} & \perp\!\!\!\perp & \texttt{gr\_sales} & \texttt{orgwkp} \\ \texttt{orgbup} & \perp\!\!\!\perp & \texttt{gr\_sales} & \texttt{orgexr} \\ \texttt{orgbup} & \perp\!\!\!\perp & \texttt{gr\_sales} & \texttt{rdint} \\ \texttt{orgwkp} & \perp\!\!\!\perp & \texttt{rdint} & \texttt{orgbup} \\ \texttt{orgwkp} & \perp\!\!\!\perp & \texttt{rdint} & \texttt{orgexr} \\ \texttt{orgexr} & \perp\!\!\!\perp & \texttt{rdint} & \texttt{orgbup} \\ \texttt{orgexr} & \perp\!\!\!\perp & \texttt{rdint} & \texttt{orgwkp} \end{array}$$

The independence `orgbup` $\perp\!\!\!\perp$ `gr_sales` suggests that causal link between new business practice and sales growth is *indirect* if it exists. This is surprising because business practice seems to be one of the innovations that is, on the one hand, most likely to have a direct impact on sales, and, on the other hand, an innovation that could be directly enforced by the growth of sales.

New methods of organising external relations seem to be strongly related to sales since there is no variable that screens off the dependences between `orgexr` and `gr_sales`.

We first draw an undirected graph containing edges between all pairs that are either independent or conditionally independent, given any of the other variables. This graph may contain too many edges, because further edges may be removed after conditioning on more than one variable (which we don't consider reliable, however).

The undirected graph resulting from such a procedure for the variables `gr_sales`, `orgbup`, `orgwkp`, `orgexr)`, `rdint`, is the skeleton of the DAG shown in Figure 34. There, we have moreover inferred to have a collider at `gr_sales` for

X	Y	Z	p-value
orgbup	orgwkp	orgexr	0.000000
orgbup	orgwkp	rdint	0.000000
orgbup	orgwkp	gr_sales	0.000000
orgbup	orgexr	orgwkp	0.000000
orgbup	orgexr	rdint	0.000000
orgbup	orgexr	gr_sales	0.000000
orgbup	rdint	orgwkp	0.000007
orgbup	rdint	orgexr	0.000002
orgbup	rdint	gr_sales	0.000000
orgbup	gr_sales	orgwkp	0.836837
orgbup	gr_sales	orgexr	0.752863
orgbup	gr_sales	rdint	0.127398
orgwkp	orgexr	orgbup	0.000000
orgwkp	orgexr	rdint	0.000000
orgwkp	orgexr	gr_sales	0.000000
orgwkp	rdint	orgbup	0.456098
orgwkp	rdint	orgexr	0.306136
orgwkp	rdint	gr_sales	0.168441
orgwkp	gr_sales	orgbup	0.001250
orgwkp	gr_sales	orgexr	0.087381
orgwkp	gr_sales	rdint	0.000040
orgexr	rdint	orgbup	0.907024
orgexr	rdint	orgwkp	0.634590
orgexr	rdint	gr_sales	0.044482
orgexr	gr_sales	orgbup	0.000472
orgexr	gr_sales	orgwkp	0.019390
orgexr	gr_sales	rdint	0.000007
rdint	gr_sales	orgbup	0.000005
rdint	gr_sales	orgwkp	0.000000
rdint	gr_sales	orgexr	0.000025

Figure 33: p-values for $X \perp\!\!\!\perp Y \,|Z$ for all possible choices of X, Y, Z out of the set `orgbup`, `orgwkp`, `orgexr`, `gr_sales`, and `rdint`.

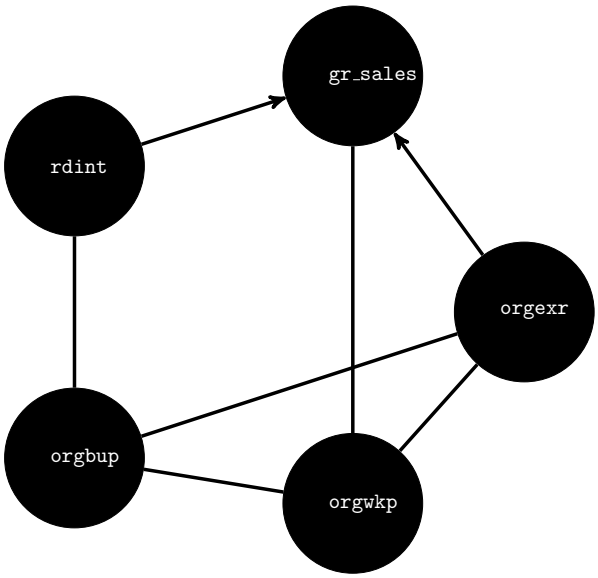

Figure 34: (Mostly undirected) graph suggested by the independence patterns in Figures 32 and 33.

two reasons: first, conditioning on this node renders `rdint` and `orgexr` dependent (although they were unconditionally independent), and second, it renders `rdint` and `orgwkp` dependent, which are also unconditionally independent.

To further check whether we trust the discovered v-structure at `gr_sales` we apply the pairwise additive noise method to the two continuous variables `rdint` and `gr_sales` and obtain `rdint` $\rightarrow$ `gr_sales` with score 0.141234 (via the R function `pairwise_anm_cis.R`). With pairwise LiNGAM we only obtained a very small preference for `rdint` $\rightarrow$ `gr_sales` with p-values 0.318 versus 0.316. Given our earlier discussion suggesting the causal direction `gr_sales` $\rightarrow$ `gr_rd`, the arrow `rdint` $\rightarrow$ `gr_sales` is surprising. However, in our previous discussion the argument was based on the assumption that the influence of R&D on sales cannot occur on the time scale the time series in the Scoreboard data set refer to. Here, we consider cross sectional data, where also long-term influences could appear on the statistical level. Yet, we have to be careful with trusting this arrow.

The implications of these findings depend heavily on the signs of the causal influences. To explore this for the arrows `rdint` $\rightarrow$ `gr_sales` and `orgexr` $\rightarrow$ `gr_sales` we have computed the regression coeffients for jointly regressing `gr_sales` on its two direct causes via standard linear regression. The result read:

`rdint`	-0.08676412
`orgexr`	0.03253082.

Thus, `rdint` seems to have a small negative influence on `gr_sales`, while `orgexr` influences `gr_sales` in a positive way. Note that this technique for identifying the sign of the influence relies on two assumptions: first, that there are no direct causes apart from those we have used in the regression (which is, indeed, sug-

gested by our partially directed graph in Figure 34). Second, we have assumed that the influence is linear (otherwise there is no obvious notion of the 'sign' of the causal influence, also defining the 'strength' would be more sophisticated otherwise (Janzing et al., 2013)). Since the graph in Figure 34 is the only example within our analysis of the cis data set where the first condition was met, we have performed the analysis of sign and strength only for this case.

Summary on R&D and organizational innovations In this section on organizational innovation variables, a first result was that sales growth was related to many organizational innovation variables (when looking at the unconditional dependences). Indeed, one might expect that organizational restructuring could be related to sales growth, either as a cause or as a consequence (or both). However, we could not find any conclusive evidence for most of the possible causal relations. Our DAG (directed acyclic graph) suggested that sales growth was slightly influenced by R&D intensity in a negative way (a somewhat surprising result given our earlier findings for Scoreboard data), and also that sales growth is caused by new methods for organizing external relations (with other firms or public institutions, such as first use of alliances, partnerships, outsourcing and subcontracting, etc) which is interesting given the current debates about 'Open Innovation'.

6.8 Process innovations, sales growth, R&D intensity

We now consider the following variables measuring process innovations: `inpspd` (new or significantly improved methods of manufacturing or producing goods or services), `inpslg` (new or significantly improved logistics, delivery or distribution methods for inputs, goods or services) `inpssu` (new or significantly improved supporting activities for your processes, such as maintenance systems or operations for purchasing, accounting, or computing), `inpcsw` (a categorical variable indicating who developed these process innovations), `inpsnm` (a variable indicating whether any of these process innovations introduced between 2006 and 2008 were new to the respective market). Moreover, we look at their relation to the variables `gr_sales` and `rdint`. The categorical variable `inpcsw` is encoded as values $1, 2, 3$, which we process as usual numbers. Since the second category is indeed a mixture of the first and the second, this should be a reasonable approach, while this approach would not necessarily be sensible for completely different categories without natural ordering. The variable `inpsnm` attains the values $1, 2, 3$ for 'yes', 'no', and 'don't know'. To obtain a more natural ordering we defined a corrected variable `inpsnmc` attaining $+1, 0, -1$ for 'yes', 'don't know', and 'no' (the procedure 'load_data.R' is also able to output `inpsnmc` if desired). We could have also excluded data points with 'don't know', but this could introduce selection bias.

We first study and discuss the correlations:

X	Y	p-value
inpspd	inpslg	0.000000
inpspd	inpssu	0.000000
inpspd	inpcsw	0.000890
inpspd	inpsnm	0.000000
inpspd	gr_sales	0.213366
inpspd	rdint	0.000084
inpslg	inpssu	0.000000
inpslg	inpcsw	0.107283
inpslg	inpsnm	0.000000
inpslg	gr_sales	0.101224
inpslg	rdint	0.000000
inpssu	inpcsw	0.000000
inpssu	inpsnm	0.173906
inpssu	gr_sales	0.003678
inpssu	rdint	0.000000
inpcsw	inpsnm	0.000596
inpcsw	gr_sales	0.435742
inpcsw	rdint	0.000677
inpsnm	gr_sales	0.299800
inpsnm	rdint	0.338664
gr_sales	rdint	0.000262

Figure 35: p-values for $X \perp\!\!\!\perp Y$ for all pairs out of inpspd, inpslg, inpssu, inpcsw, inpsnm, gr_sales, rdint.

	inpspd	inpslg	inpssu	inpcsw	inpsnm	gr_sales
inpslg	0.32					
inpssu	0.33	0.40				
inpcsw	-0.12	-0.00050	0.097			
inpsnm	0.13	0.10	0.041	0.0063		
gr_sales	0.0012	0.050	0.050	-0.0030	0.0012	
rdint	-0.013	-0.070	-0.086	-0.060	-0.013	-0.024

The variables inpspd, inpslg, inpssu, and inpsnm are strongly positively correlated, while their correlation to the other ones are weak (with partly slighly negative correlations). The only further correlation that is worthwhile mentioning is the one between inpspd and rdint. – This makes sense since new methods for manufactoring are naturally associated with R&D. The results of the unconditional independence tests are shown in Figure 35.
We thuis accept the following unconditional independences:

$$\begin{aligned}
\texttt{inpspd} &\perp\!\!\!\perp \texttt{gr_sales} \\
\texttt{inpslg} &\perp\!\!\!\perp \texttt{inpcsw} \\
\texttt{inpslg} &\perp\!\!\!\perp \texttt{gr_sales} \\
\texttt{inpssu} &\perp\!\!\!\perp \texttt{inpsnm} \\
\texttt{inpcsw} &\perp\!\!\!\perp \texttt{gr_sales} \\
\texttt{inpsnm} &\perp\!\!\!\perp \texttt{gr_sales} \\
\texttt{inpsnm} &\perp\!\!\!\perp \texttt{rdint}
\end{aligned}$$

It is evident that the variable sales growth is independent of the other ones except for `rdint`. Moreover, we have `inpsnm` $\perp\!\!\!\perp$ `rdint`, i.e, whether these process innovations were new in the respective market is not associated with R&D intensity. The results of the conditional independence tests are shown in Figure 36.

We thus accept the following conditional independences:

inpspd	$\perp\!\!\!\perp$	gr_sales	inpslg
inpspd	$\perp\!\!\!\perp$	gr_sales	inpssu
inpspd	$\perp\!\!\!\perp$	gr_sales	inpcsw
inpspd	$\perp\!\!\!\perp$	gr_sales	inpsnm
inpspd	$\perp\!\!\!\perp$	gr_sales	rdint
inpspd	$\perp\!\!\!\perp$	rdint	inpcsw
inpslg	$\perp\!\!\!\perp$	inpcsw	inpssu
inpslg	$\perp\!\!\!\perp$	gr_sales	inpspd
inpslg	$\perp\!\!\!\perp$	gr_sales	inpsnm
inpslg	$\perp\!\!\!\perp$	gr_sales	rdint
inpssu	$\perp\!\!\!\perp$	inpsnmc	inpcsw
inpssu	$\perp\!\!\!\perp$	inpsnmc	gr_sales
inpssu	$\perp\!\!\!\perp$	gr_sales	inpspd
inpssu	$\perp\!\!\!\perp$	gr_sales	inpcsw
inpssu	$\perp\!\!\!\perp$	gr_sales	rdint
inpcsw	$\perp\!\!\!\perp$	gr_sales	inpspd
inpcsw	$\perp\!\!\!\perp$	gr_sales	inpslg
inpcsw	$\perp\!\!\!\perp$	gr_sales	inpssu
inpcsw	$\perp\!\!\!\perp$	gr_sales	inpsnmc
inpcsw	$\perp\!\!\!\perp$	gr_sales	rdint
inpsnm	$\perp\!\!\!\perp$	gr_sales	inpspd
inpsnm	$\perp\!\!\!\perp$	gr_sales	inpslg
inpsnm	$\perp\!\!\!\perp$	gr_sales	inpcsw
inpsnm	$\perp\!\!\!\perp$	gr_sales	rdint
inpsnm	$\perp\!\!\!\perp$	rdint	inpspd
inpsnm	$\perp\!\!\!\perp$	rdint	inpslg
inpsnm	$\perp\!\!\!\perp$	rdint	inpssu
inpsnm	$\perp\!\!\!\perp$	rdint	inpcsw
inpsnmc	$\perp\!\!\!\perp$	rdint	gr_sales

The undirected graph resulting from this independence pattern is the skeleton of the DAG shown in Figure 38.

To direct some of the edges between the discrete variables, we use discrete additive noise models and obtain the following p-values for the respective models:

X	Y	$X \to Y$	$X \leftarrow Y$
inpspd	inpslg	0.000000	0.087043
inpcsw	inpslg	0.273026	0.273026
inpcsw	inpspd	0.423351	0.000488
inpsnm	inpslg	0.007715	0.007715
inpsnm	inpspd	0.000074	0.000000
inpsnm	inpcsw	0.142503	0.166889

inpspd	inpslg	inpssu	0.000000
inpspd	inpslg	inpcsw	0.000000
inpspd	inpslg	inpsnmc	0.000037
inpspd	inpslg	gr_sales	0.000000
inpspd	inpslg	rdint	0.000000
inpspd	inpssu	inpslg	0.000000
inpspd	inpssu	inpcsw	0.000000
inpspd	inpssu	inpsnmc	0.000000
inpspd	inpssu	gr_sales	0.000000
inpspd	inpssu	rdint	0.000000
inpspd	inpcsw	inpslg	0.000048
inpspd	inpcsw	inpssu	0.000514
inpspd	inpcsw	inpsnmc	0.033005
inpspd	inpcsw	gr_sales	0.000246
inpspd	inpcsw	rdint	0.000207
inpspd	inpsnmc	inpslg	0.000000
inpspd	inpsnmc	inpssu	0.000000
inpspd	inpsnmc	inpcsw	0.000000
inpspd	inpsnmc	gr_sales	0.000000
inpspd	inpsnmc	rdint	0.000000
inpspd	gr_sales	inpslg	0.560503
inpspd	gr_sales	inpssu	0.134279
inpspd	gr_sales	inpcsw	0.409504
inpspd	gr_sales	inpsnmc	0.202086
inpspd	gr_sales	rdint	0.618911
inpspd	rdint	inpslg	0.000465
inpspd	rdint	inpssu	0.010798
inpspd	rdint	inpcsw	0.147705
inpspd	rdint	inpsnmc	0.074942
inpspd	rdint	gr_sales	0.000426
inpslg	inpssu	inpspd	0.000000
inpslg	inpssu	inpcsw	0.000000
inpslg	inpssu	inpsnmc	0.000000
inpslg	inpssu	gr_sales	0.000000
inpslg	inpssu	rdint	0.000000
inpslg	inpcsw	inpspd	0.002089
inpslg	inpcsw	inpssu	0.369611
inpslg	inpcsw	inpsnmc	0.041159
inpslg	inpcsw	gr_sales	0.041683
inpslg	inpcsw	rdint	0.070603
inpslg	inpsnmc	inpspd	0.000000
inpslg	inpsnmc	inpssu	0.000000
inpslg	inpsnmc	inpcsw	0.000000
inpslg	inpsnmc	gr_sales	0.000000
inpslg	inpsnmc	rdint	0.000000
inpslg	gr_sales	inpspd	0.188785
inpslg	gr_sales	inpssu	0.081347
inpslg	gr_sales	inpcsw	0.086256
inpslg	gr_sales	inpsnmc	0.193268
inpslg	gr_sales	rdint	0.129602
inpslg	rdint	inpspd	0.000385
inpslg	rdint	inpssu	0.008461
inpslg	rdint	inpcsw	0.001568
inpslg	rdint	inpsnmc	0.000058
inpslg	rdint	gr_sales	0.000000
inpssu	inpcsw	inpspd	0.000000
inpssu	inpcsw	inpslg	0.000000
inpssu	inpcsw	inpsnmc	0.000000
inpssu	inpcsw	gr_sales	0.000000
inpssu	inpcsw	rdint	0.000000
inpssu	inpsnmc	inpspd	0.000050
inpssu	inpsnmc	inpslg	0.017112
inpssu	inpsnmc	inpcsw	0.275708
inpssu	inpsnmc	gr_sales	0.135320
inpssu	inpsnmc	rdint	0.016074
inpssu	gr_sales	inpspd	0.251001
inpssu	gr_sales	inpslg	0.057639
inpssu	gr_sales	inpcsw	0.106455
inpssu	gr_sales	inpsnmc	0.076437
inpssu	gr_sales	rdint	0.118541
inpssu	rdint	inpspd	0.000000
inpssu	rdint	inpslg	0.000314
inpssu	rdint	inpcsw	0.086885
inpssu	rdint	inpsnmc	0.019061
inpssu	rdint	gr_sales	0.000000
inpcsw	inpsnmc	inpspd	0.027007
inpcsw	inpsnmc	inpslg	0.000002
inpcsw	inpsnmc	inpssu	0.000645
inpcsw	inpsnmc	gr_sales	0.000252
inpcsw	inpsnmc	rdint	0.000925
inpcsw	gr_sales	inpspd	0.444775
inpcsw	gr_sales	inpslg	0.649188
inpcsw	gr_sales	inpssu	0.220738
inpcsw	gr_sales	inpsnmc	0.891050
inpcsw	gr_sales	rdint	0.459642

Figure 36: p-values for $X \perp\!\!\!\perp Y \,|Z$ for of X, Y, Z taken from the set `inpspd`, `inpslg`, `inpssu`, `inpcsw`, `inpsnm`, `gr_sales`, `rdint`, first part. The second part is shown in Figure 37.

inpcsw	rdint	inpspd	0.000045
inpcsw	rdint	inpslg	0.000033
inpcsw	rdint	inpssu	0.012104
inpcsw	rdint	inpsnmc	0.000080
inpcsw	rdint	gr_sales	0.077681
inpsnmc	gr_sales	inpspd	0.721066
inpsnmc	gr_sales	inpslg	0.920148
inpsnmc	gr_sales	inpssu	0.077860
inpsnmc	gr_sales	inpcsw	0.636981
inpsnmc	gr_sales	rdint	0.509830
inpsnmc	rdint	inpspd	0.181084
inpsnmc	rdint	inpslg	0.537842
inpsnmc	rdint	inpssu	0.839522
inpsnmc	rdint	inpcsw	0.673609
inpsnmc	rdint	gr_sales	0.361830
gr_sales	rdint	inpspd	0.002559
gr_sales	rdint	inpslg	0.000647
gr_sales	rdint	inpssu	0.000009
gr_sales	rdint	inpcsw	0.000000
gr_sales	rdint	inpsnmc	0.001225

Figure 37: p-values for $X \perp\!\!\!\perp Y \,|Z$ for of X, Y, Z taken from the set `inpspd`, `inpslg`, `inpssu`, `inpcsw`, `inpsnm`, `gr_sales`, `rdint`, second part of the table in Figure 36.

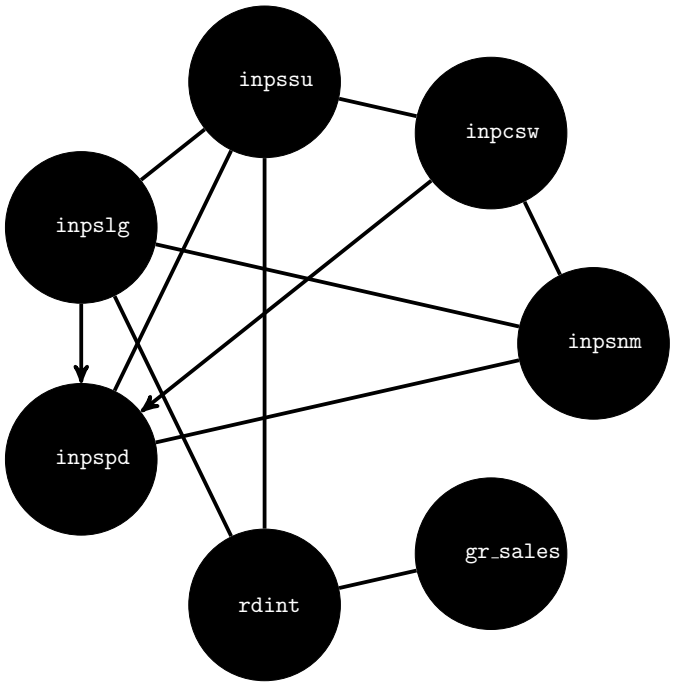

Figure 38: Undirected graph resulting from the independence pattern of `inpcsw`, `inpssu`, `inpslg`, `inpspd`, `rdint`, `gr_sales`, `inpsnm`.

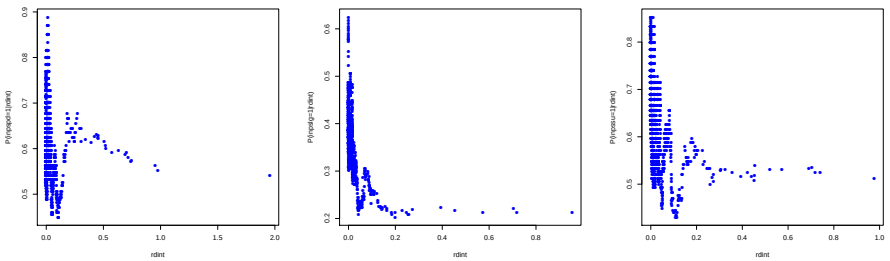

Figure 39: $P(\texttt{bin} = 1|\texttt{rdint})$ for the binary variables `bin`= `inpspd`, `inpslg`, `inpssu`. The conditionals don't look smooth, which suggests that the statistical dependence of the respective variables is not due to some 'simple' causal relation.

If we only infer causal directions when one p-value is at least 0.05 and the other significantly less, we only obtain the results `inpslg` $\rightarrow$ `inpspd` and `inpcsw` $\rightarrow$ `inpspd`. We have added these directions to Figure 38. The obtained collider at `inpspd` is consistent with the fact that `inpslg` and `inpcsw` are not conditionally independent, given `inpspd`.

For the potential causal relation between the continuous variable `rdint` and the three binary variables `inpspd`, `inpslg`, `inpssu`, we visualize the conditional distribution by the same type of diagrams as in Figure 25. The results are shown in Figure 39.

The curve for `inpspd`in Figure 39, left, looks quite complex, which renders the causal hypothesis `rdint` $\rightarrow$ `inpspd` less plausible than it would be a priori. The other two curves would correspond to more natural mechanisms. However, according to our pattern of conditional independences, only `inpspd` is not directly linked with `rdint`.

Software: For the relation between an arbitrary real-valued variable `var_continuous` and an arbitrary binary variable `var_binary` from the CIS data set the R function `plot_conditional_continuous_to_binary.R` plots diagrams of the type in Figure 25 and 39. It is called by the command

`plot_conditional_continuous_to_binary`(`var_continuous`, `var_binary`),

where `var_continuous` and `var_binary` are the variable names in the header of the CIS data file `2015_12_10_data_panel.csv`. The command plots the function $P(\texttt{var\_binary} = 1|\texttt{var\_contiuous}$. The plot is printed to the file `conditional_var_continuous__var_binary.pdf`.

We perform the same kind of investigation to the relations between the continuous variable `gr_sales` and the binary variables `inpspd`, `inpslg`, `inpssu`. The plots are shown in Figure 40. In all three cases the relation between these variables seems rather chaotic (even when accounting for the fact that the number of companies with negative growth is small and therefore statistical fluctuations get stronger). Only in the regime of strong growth there seems to more regular.

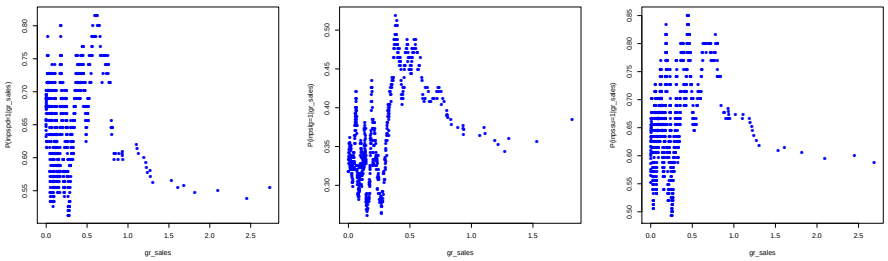

Figure 40: Visualization of the conditional distributions $P(\texttt{bin} = \texttt{1}|\texttt{rddinx})$ for the binary variables `bin`= `inspsd`, `inpslg`, and `inpslu`, see text.

Summary on process innovations Our analysis of process innovations variables suggested, first, that sales growth was only weakly correlated with the process innovations variables. This might seem surprising at first, although a large literature that focuses on the determinants of firm growth has generally emphasized the quasi-random nature of firm growth, and difficulties in predicting or explaining firm growth, see e.g. Coad (2009).

Our DAG (directed acyclic graph) showed, as before, that in most cases we could not conclude for a causal direction between variable-pairs, although the evidence suggested that introducing new methods of manufacturing or producing goods or services was caused by: i) who developed the process innovations (own firm or other), and ii) the introduction of new logistics, delivery or distribution methods. Some tentative policy implications would be that cooperation activity for process innovations, and also the introduction of new logistics, can cause the appearance of new process innovations.

6.9 Information sources and sales growth

We now consider the variables scon (conferences, trade fairs, exhibitions were important for the company), `sjou` (scientific journals and trade/technical publications were important), `spro` (professional and industry associations were important), `gr_sales`, and `rdint`. To see that these variables are clearly associated, we first look at the correlations:

	scon	sjou	spro	gr_sales
sjou	0.55			
spro	0.38	0.41		
gr_sales	0.0099	-0.0050	0.018	
rdint	0.019	0.099	-0.021	-0.024

The information source variables `scon`, `sjou`, and `spro` are strongly correlated. Two of them, namely, `scon` and `sjou`, show correlations to R&D intensity that are strong enough to require explanations. As expected, all their correlations are positive. Negative correlations would probably be harder to explain. From the point of view of a scientist in academia, `scon` and `sjou` are more directly related to what one would call 'research' than the variable `spro`. We would therefore consider it natural that the former are stronger correlated with R&D intensity. The results of the unconditional independence tests are shown

X	Y	p-value
scon	sjou	0.000000
scon	spro	0.000000
scon	gr_sales	0.143829
scon	rdint	0.312474
sjou	spro	0.000000
sjou	gr_sales	0.682140
sjou	rdint	0.001163
spro	gr_sales	0.544035
spro	rdint	0.000000
gr_sales	rdint	0.000000

Figure 41: p-values for $X \perp\!\!\!\perp Y$ for all pairs out of scon, sjou, spro, gr_sales, and rdint.

in Figure 41. We thus accept the following unconditional independences:

$$\begin{aligned} \texttt{scon} &\perp\!\!\!\perp \texttt{gr\_sales} \\ \texttt{scon} &\perp\!\!\!\perp \texttt{rdint} \\ \texttt{sjou} &\perp\!\!\!\perp \texttt{gr\_sales} \end{aligned}$$

It is not surprising that the importance of conferences is not related to sales. It is also plausible that the importance of scientific journals is not associated with sales growth. More remarkable, however, is the fact that the importance of conferences is independent of R&D intensity since one would expect that companies that spend a large fraction for R&D are more interested in scientific conferences. For the unconditional tests we obtained the results shown in Figure 42. We thus accept the following conditional independences:

$$\begin{aligned} \texttt{scon} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{sjou} \\ \texttt{scon} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rdint} \\ \texttt{sjou} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{scon} \\ \texttt{sjou} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{spro} \\ \texttt{sjou} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rdint} \\ \texttt{scon} &\perp\!\!\!\perp \texttt{rdint} \mid \texttt{sjou} \end{aligned}$$

The last one of the conditional independences suggests that the importance of scientific journals intermediates the causal link between R&D intensity and the importance of scientific conferences. This is plausible since scientific conferences are hard to follow without the respective background obtained from scientific publications. The partially directed graph resulting from the above pattern of independences is shown in Figure 43. To direct some edges, we observe that `scon` and `gr_sales` get dependent when conditioning on `spro` although they are independent otherwise. Assuming causal faithfulness, this implies that the edges `scon` – `spro` and `gr_sales` – `spro` have arrowheads at `spro`, as shown in Figure 43, although one has to be careful with this conclusion particularly because the corresponding p-value for the conditional test is not far from our confidence level 0.1, namely 0.071.

In trying to direct further edges, we again use discrete additive noise models among the discrete variables `scon`, `sjou`, `spro` and obtain the following p-values for the corresponding models:

X	Y	Z	p-value
scon	sjou	spro	0.000000
scon	sjou	gr_sales	0.000000
scon	sjou	rdint	0.000000
scon	spro	sjou	0.000000
scon	spro	gr_sales	0.000000
scon	spro	rdint	0.000000
scon	gr_sales	sjou	0.317289
scon	gr_sales	spro	0.071358
scon	gr_sales	rdint	0.372389
scon	rdint	sjou	0.460608
scon	rdint	spro	0.045403
scon	rdint	gr_sales	0.090862
sjou	spro	scon	0.000000
sjou	spro	gr_sales	0.000000
sjou	spro	rdint	0.000000
sjou	gr_sales	scon	0.518410
sjou	gr_sales	spro	0.871700
sjou	gr_sales	rdint	0.843280
sjou	rdint	scon	0.017065
sjou	rdint	spro	0.000008
sjou	rdint	gr_sales	0.005561
spro	gr_sales	scon	0.426570
spro	gr_sales	sjou	0.661447
spro	gr_sales	rdint	0.650719
spro	rdint	scon	0.000000
spro	rdint	sjou	0.000000
spro	rdint	gr_sales	0.000010
gr_sales	rdint	scon	0.000000
gr_sales	rdint	sjou	0.000000
gr_sales	rdint	spro	0.000000

Figure 42: p-values for $X \perp\!\!\!\perp Y \,|Z$ for all possible choices of X, Y, Z out of the set `scon`, `sjou`, `spro`, `gr_sales`, and `rdint`.

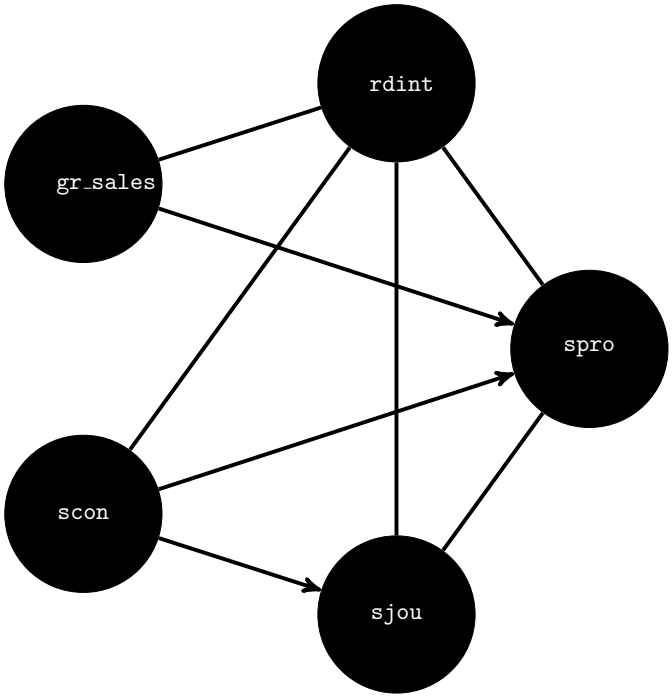

Figure 43: Undirected graph resulting from the independence pattern of `rdint`, `gr_sales`, `scon`, `sjou`, `spro`, the edge `scon` – `sjou` has been directed via discrete additive noise models, see text.

X	Y	$X \to Y$	$X \leftarrow Y$
sjou	scon	0.074151	0.000000
spro	scon	0.000337	0.000000
spro	sjou	0.007182	0.000000

For `sjou` $\to$ `scon` the p-value is signifiantly above 0.05 while the one for the converse direction is 0.000000. We thus accept this arrow as shown in Figure 43.

Sources of information and co-operation For the variables `sentg` (use of internal information sources), `ssup` (use of information from suppliers of equipment, materials, components, or software), `scli` (use of information from clients or customers), `scom` (use of information from competitors or other enterprises in the respective sector), `sins` (use of information from consultants, commercial labs, or private R&D institutes), `suni` (use of information from universities or other higher education institutions), and `sgmt` (use of information from government or public research institutes) we obtained the following results: p-value 0.000000 for all unconditional independences. For the conditional independences, all p-values were also orders of magnitude below our threshold 0.1. The fist part of the correlation matrix reads:

	sentg	ssup	scli	scom	sins	suni	sgmt
ssup	0.058						
scli	0.11	0.17					
scom	0.043	0.20	0.43				
sins	0.024	0.17	0.025	0.19			
suni	0.093	0.072	0.11	0.17	0.34		
sgmt	0.069	0.10	0.14	0.27	0.32	0.68	
scon	0.032	0.18	0.24	0.31	0.20	0.25	0.28
sjou	0.078	0.15	0.15	0.25	0.25	0.34	0.35
spro	-0.017	0.17	0.146	0.23	0.33	0.27	0.34
gr_sales	-0.017	-0.0067	0.00	0.018	0.023	0.03	0.054
rdint	0.033	-0.075	0.00	-0.022	0.0060	0.12	0.13

The second part of this table reads:

	scon	sjou	spro	gr_sales
sjou	0.55			
spro	0.37	0.41		
gr_sales	0.0099	-0.0050	0.018	
rdint	0.019	0.099	-0.021	-0.024

Remarkably, the correlation of the information source variables to sales growth are rather weak, while the information source variables themselves strongly correlate. It is interesting to see that some of the information source variable have some reasonably strong correlations to R&D intensity: `sentg` (internal sources) and `scli` (clients and customers) as well as non-commercial information sources like universities `suni` and `gmt` show strong correlations to R&D intensity. This, again, suggests that the answers of the questionaire correlate with 'hard facts' such as R&D intensity. The unconditional independence tests resulted in p-values 0.000000 for all cases and also the conditional independences were all rejected, see Figure 44.

sentg	ssup	scli	0.000000
sentg	ssup	scom	0.000000
sentg	ssup	sins	0.000000
sentg	ssup	suni	0.000000
sentg	ssup	sgmt	0.000000
sentg	scli	ssup	0.000000
sentg	scli	scom	0.000000
sentg	scli	sins	0.000000
sentg	scli	suni	0.000000
sentg	scli	sgmt	0.000000
sentg	scom	ssup	0.000000
sentg	scom	scli	0.006701
sentg	scom	sins	0.000000
sentg	scom	suni	0.000000
sentg	scom	sgmt	0.000000
sentg	sins	ssup	0.000000
sentg	sins	scli	0.000000
sentg	sins	scom	0.000000
sentg	sins	suni	0.000000
sentg	sins	sgmt	0.000000
sentg	suni	ssup	0.000000
sentg	suni	scli	0.001985
sentg	suni	scom	0.000050
sentg	suni	sins	0.000050
sentg	suni	sgmt	0.000000
sentg	sgmt	ssup	0.000000
sentg	sgmt	scli	0.000813
sentg	sgmt	scom	0.000357
sentg	sgmt	sins	0.004584
sentg	sgmt	suni	0.000020
ssup	scli	sentg	0.000000
ssup	scli	scom	0.000000
ssup	scli	sins	0.000000
ssup	scli	suni	0.000000
ssup	scli	sgmt	0.000000
ssup	scom	sentg	0.000000
ssup	scom	scli	0.000000
ssup	scom	sins	0.000000
ssup	scom	suni	0.000000
ssup	scom	sgmt	0.000000
ssup	sins	sentg	0.000000
ssup	sins	scli	0.000000
ssup	sins	scom	0.000000
ssup	sins	suni	0.000000
ssup	sins	sgmt	0.000000
ssup	suni	sentg	0.000000
ssup	suni	scli	0.003324
ssup	suni	scom	0.047740
ssup	suni	sins	0.000904
ssup	suni	sgmt	0.010265
ssup	sgmt	sentg	0.000000
ssup	sgmt	scli	0.000044
ssup	sgmt	scom	0.001387
ssup	sgmt	sins	0.000058
ssup	sgmt	suni	0.000951
scli	scom	sentg	0.000000
scli	scom	ssup	0.000000
scli	scom	sins	0.000000
scli	scom	suni	0.000000
scli	scom	sgmt	0.000000
scli	sins	sentg	0.000000
scli	sins	ssup	0.000000
scli	sins	scom	0.000000
scli	sins	suni	0.000000
scli	sins	sgmt	0.000000
scli	suni	sentg	0.000000
scli	suni	ssup	0.000000
scli	suni	scom	0.000000
scli	suni	sins	0.000000
scli	suni	sgmt	0.000000
scli	sgmt	sentg	0.000000
scli	sgmt	ssup	0.000000
scli	sgmt	scom	0.000035
scli	sgmt	sins	0.000000
scli	sgmt	suni	0.001911
scom	sins	sentg	0.000000
scom	sins	ssup	0.000000
scom	sins	scli	0.000000
scom	sins	suni	0.000000
scom	sins	sgmt	0.000000
scom	suni	sentg	0.000000

Figure 44: p-values for all conditional indepences in the set *sentg*, `ssup`, `scli`, `scom`, `sins`, `suni`, `sgmt`. First part of the table. Figure 45 shows the second part.

scom	suni	ssup	0.000000
scom	suni	scli	0.000000
scom	suni	sins	0.000000
scom	suni	sgmt	0.000000
scom	sgmt	sentg	0.000000
scom	sgmt	ssup	0.000000
scom	sgmt	scli	0.000000
scom	sgmt	sins	0.000000
scom	sgmt	suni	0.000002
sins	suni	sentg	0.000000
sins	suni	ssup	0.000000
sins	suni	scli	0.000000
sins	suni	scom	0.000000
sins	suni	sgmt	0.000000
sins	sgmt	sentg	0.000000
sins	sgmt	ssup	0.000000
sins	sgmt	scli	0.000000
sins	sgmt	scom	0.000000
sins	sgmt	suni	0.000000
suni	sgmt	sentg	0.000000
suni	sgmt	ssup	0.000000
suni	sgmt	scli	0.000000
suni	sgmt	scom	0.000000
suni	sgmt	sins	0.000000

Figure 45: Second part of the table in Figure 44, listing the conditional independences in the set `sentg`, `ssup`, `scli`, `scom`, `sins`, `suni`, `sgmt`.

To get additional information on the causal directions, we used discrete additive noise models and obtained the results shown in Figure 46.

We tend to infer `suni` → `scli`, `suni` → `scom`, `sgmt` → `scli`, `sgmt` → `scom`, `sgmt` → `sins`, `sjou` → `scli`, `spro` → `scom` and thus obtain the partially directed graph in Figure 47.

Summary on information sources The variables relating to information sources were highly correlated between them, but were only weakly correlated (if at all) with sales growth. Constructing a DAG (directed acyclic graph) with the help of discrete additive noise analysis, we observed a few causal relations: i) that conferences/trade fairs as a source of information caused interest in journals as a source of information, ii) that conferences/trade fairs caused interest in professional and industrial associations as a source of information, and also iii) that growth of sales caused interest in professional and industrial associations as a source of information (which makes good sense if growth brings new challenges to the firm). A possible (and of course tentative) policy implication would be that it makes little sense to spend public funds towards developing professional and industrial associations, because these will be grown 'from the bottom up' by microeconomic agents' activity, and in any case these associations seem to be an outcome rather than an input in the innovation process. Subsequent analysis on a larger set of information-source variables showed that competitors/firms in the same industry (as an information source) is caused by several other variables (universities, government/public research, and professional & industrial associations). Universities also seemed to play a role in stimulating firms to look to their clients and customers as a source of information. There may be policy implications regarding the role of universities in causing firms to become aware of the value of new sources of information.

6.10 Innovation expenditures & sales growth

We consider the variables `rrdex` (external R&D) , `rmac` (acquisition of machinery, equipment and software), `roek` (acquisition of external knowledge), `rtr` (training for innovative activities), `rmar` (market introduction of innovations), `gr_sales`. We first discuss the correlations.

	rrdex	rmac	roek	rtr	rmar
rmac	0.20				
roek	0.23	0.22			
rtr	0.19	0.33	0.26		
rmar	0.23	0.23	0.27	0.33	
gr_sales	0.06	0.094	0.05	0.054	0.016

The fact that the presence of external R&D activity correlates with the R&D intensity is a useful check for the consistency of the 'soft' answers with the reported 'hard' variables like R&D intensity. All the innovation expenditures are strongly positively correlated, but the correlations to sales growth are quite weak. The only remarkable correlation relating sales growth with the innovation expenditure variables is `rmac`, which is plausible because the acquisition of machinery could be required by sales growth. On the other hand, the acquisition of machinery may have enabled an increasing production (e.g., by first

ssup	sentg	0.000002	0.426429
scli	sentg	0.005978	0.000450
scli	ssup	0.005435	0.004302
scom	sentg	0.144877	0.102829
scom	ssup	0.000012	0.000000
scom	scli	0.000164	0.000062
sins	sentg	0.018601	0.018601
sins	ssup	0.001929	0.000011
sins	scli	0.018404	0.000000
sins	scom	0.030256	0.001750
suni	sentg	0.000127	0.000007
suni	ssup	0.111572	0.111572
suni	scli	0.153438	0.000471
suni	scom	0.208594	0.000000
suni	sins	0.000760	0.000000
sgmt	sentg	0.036118	0.036118
sgmt	ssup	0.036577	0.029988
sgmt	scli	0.875848	0.009585
sgmt	scom	0.758114	0.000005
sgmt	sins	0.081725	0.000000
sgmt	suni	0.000004	0.000000
scon	sentg	0.007180	0.007245
scon	ssup	0.066757	0.002744
scon	scli	0.001981	0.000001
scon	scom	0.000662	0.000931
scon	sins	0.000073	0.004425
scon	suni	0.000000	0.055260
scon	sgmt	0.000620	0.104172
sjou	sentg	0.002727	0.002456
sjou	ssup	0.000675	0.000001
sjou	scli	0.140888	0.000098
sjou	scom	0.023016	0.000002
sjou	sins	0.000115	0.005604
sjou	suni	0.000001	0.015291
sjou	sgmt	0.000018	0.000146
sjou	scon	0.000070	0.000001
spro	sentg	0.005561	0.000122
spro	ssup	0.042978	0.000133
spro	scli	0.002425	0.000000
spro	scom	0.084544	0.000007
spro	sins	0.006054	0.000001
spro	suni	0.001889	0.000607
spro	sgmt	0.000000	0.000186
spro	scon	0.000117	0.000000
spro	sjou	0.000050	0.000000

Figure 46: p-values for pairwise discrete additive noise models for all pairs out of the variables `sentg`, `ssup`, `scli`, `scom`, `sins`, `suni`, `sgmt`, `scon`, `sjou`, `spro`.

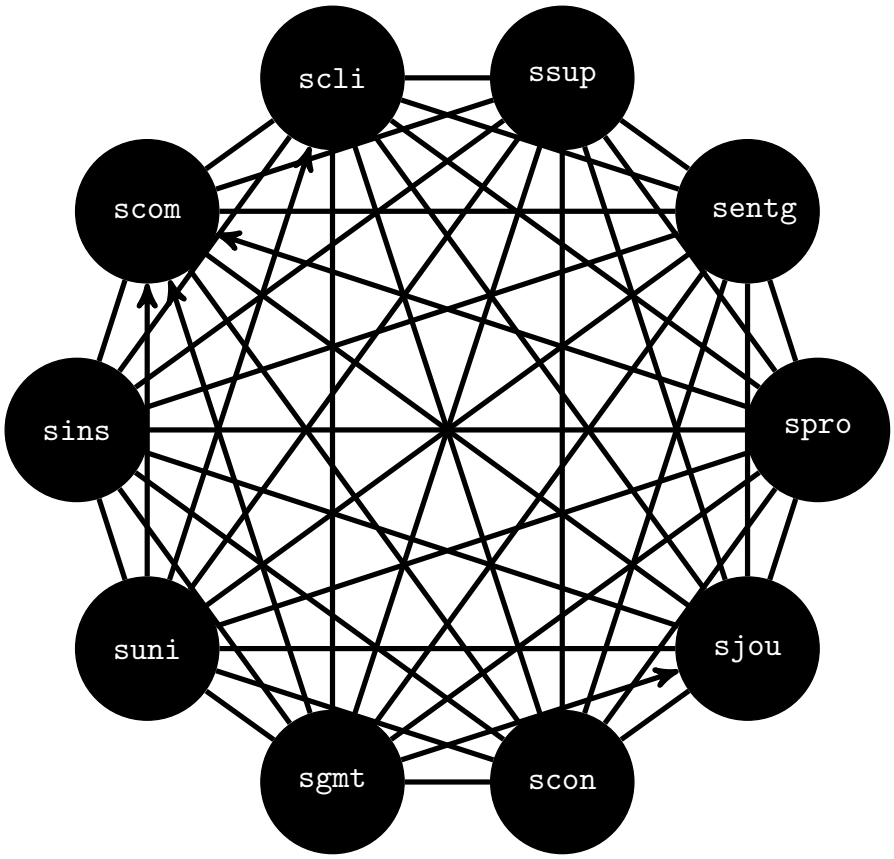

Figure 47: Partially directed graph for the information source and co-operation variables `sentg`, `ssup`, `scli`, `scom`, `sins`, `suni`, `sgmt`, `scon`, `sjou`, `spro`. Since there are no conditional independences, all pairs of nodes are connected and the directions are inferred via discrete additive noise models.

rrdex	rmac	0.000000
rrdex	roek	0.000000
rrdex	rtr	0.000000
rrdex	rmar	0.000000
rrdex	gr_sales	0.422264
rmac	roek	0.000000
rmac	rtr	0.000000
rmac	rmar	0.000000
rmac	gr_sales	0.100560
roek	rtr	0.000000
roek	rmar	0.000000
roek	gr_sales	0.056163
rtr	rmar	0.000000
rtr	gr_sales	0.275638
rmar	gr_sales	0.745748

Figure 48: p-values for all unconditional independence tests for the innovation expenditure variables `rrdex`, `rmac`, `roek`, `rtr`, `rmar`, `gr_sales`.

allowing for a cheaper production). It may be worthwhile to test the direction of the latter causal relation. The results of the unconditional tests are shown in Figure 48. It is evident that sales growth is independent of most variables:

$$\begin{aligned} \texttt{rrdex} &\perp\!\!\!\perp \texttt{gr\_sales} \\ \texttt{rmac} &\perp\!\!\!\perp \texttt{gr\_sales} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \end{aligned}$$

Note also that the independence of `roek` and `gr_sales` is almost accepted (for instance if we set the significance level to 0.05 instead of 0.1. The results of the conditional tests are shown in Figure 49. The list of conditional independences reads:

$$\begin{aligned} \texttt{rrdex} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmar} \\ \texttt{rrdex} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmac} \\ \texttt{rrdex} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{roek} \\ \texttt{rrdex} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rtr} \\ \texttt{rmac} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rmac} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{roek} \\ \texttt{rmac} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rtr} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmax} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{roek} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmax} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rrdex} \end{aligned}$$

Again, we observe that `gr_sales` is independent of most of the other vari-

rrdex	rmac	roek	0.000000	rmac	rmar	rtr	0.000000
rrdex	rmac	rtr	0.000000	rmac	rmar	gr_sales	0.000000
rrdex	rmac	rmar	0.000000	rmac	gr_sales	rrdex	0.193435
rrdex	rmac	gr_sales	0.000000	rmac	gr_sales	roek	0.149505
rrdex	roek	rmac	0.000000	rmac	gr_sales	rtr	0.141779
rrdex	roek	rtr	0.000000	rmac	gr_sales	rmar	0.083774
rrdex	roek	rmar	0.000000	roek	rtr	rrdex	0.000000
rrdex	roek	gr_sales	0.000000	roek	rtr	rmac	0.000000
rrdex	rtr	rmac	0.000000	roek	rtr	rmar	0.000000
rrdex	rtr	roek	0.000000	roek	rtr	gr_sales	0.000000
rrdex	rtr	rmar	0.000000	roek	rmar	rrdex	0.000000
rrdex	rtr	gr_sales	0.000000	roek	rmar	rmac	0.000000
rrdex	rmar	rmac	0.000000	roek	rmar	rtr	0.000000
rrdex	rmar	roek	0.000000	roek	rmar	gr_sales	0.000000
rrdex	rmar	rtr	0.000000	roek	gr_sales	rrdex	0.245901
rrdex	rmar	gr_sales	0.000000	roek	gr_sales	rmac	0.094863
rrdex	gr_sales	rmac	0.603512	roek	gr_sales	rtr	0.051298
rrdex	gr_sales	roek	0.817179	roek	gr_sales	rmar	0.039962
rrdex	gr_sales	rtr	0.740415	rtr	rmar	rrdex	0.000000
rrdex	gr_sales	rmar	0.736662	rtr	rmar	rmac	0.000000
rmac	roek	rrdex	0.000000	rtr	rmar	roek	0.000000
rmac	roek	rtr	0.000000	rtr	rmar	gr_sales	0.000000
rmac	roek	rmar	0.000000	rtr	gr_sales	rrdex	0.571532
rmac	roek	gr_sales	0.000000	rtr	gr_sales	rmac	0.601842
rmac	rtr	rrdex	0.000000	rtr	gr_sales	roek	0.294568
rmac	rtr	roek	0.000000	rtr	gr_sales	rmar	0.343553
rmac	rtr	rmar	0.000000	rmar	gr_sales	rrdex	0.901735
rmac	rtr	gr_sales	0.000000	rmar	gr_sales	rmac	0.905659
rmac	rmar	rrdex	0.000000	rmar	gr_sales	roek	0.325692
rmac	rmar	roek	0.000000	rmar	gr_sales	rtr	0.911551

Figure 49: p-values of conditional independence tests for the variables `rrdex`, `rmac`, `roek`, `rtr`, `rmar`, `gr_sales`.

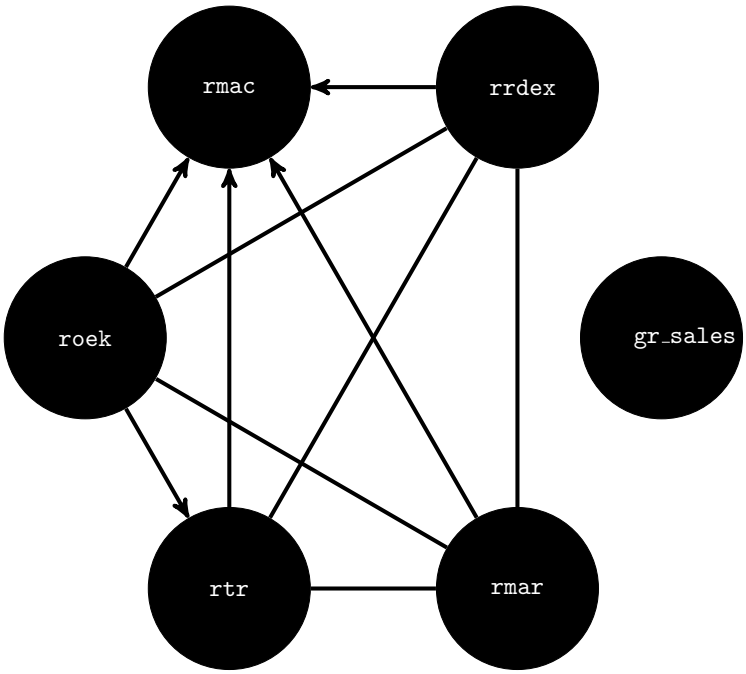

Figure 50: Partially directed graph resulting from the independence pattern of `rrdex`, `rmac`, `roek`, `rtr`, `rmar`, `gr_sales` and discrete additive noise models.

ables, no matter on which variable we condition. Drawing the corresponding undirected graph would be pointless, since it looks trivial: `gr_sales` is disconnected with the other variables, while all the other variables form a complete graph. This pattern is, of course, completely uninformative about causal directions. To infer some causal directions, we use discrete additive noise models and obtain the following p-values:

X	Y	$X \to Y$	$X \leftarrow Y$
rmac	rrdex	0.000771	0.247597
roek	rrdex	0.000460	0.000460
roek	rmac	0.856951	0.000042
rtr	rrdex	0.000007	0.001341
rtr	rmac	0.827332	0.000000
rtr	roek	0.000000	0.233057
rmar	rrdex	0.002336	0.002336
rmar	rmac	0.242554	0.000000
rmar	roek	0.000000	0.023299
rmar	rtr	0.001623	0.000000

This clearly favors the causal directions `rrdex` $\to$ `rmac`, `roek` $\to$ `rmac`, `rmar` $\to$ `rmac`, and `roek` $\to$ `rtr`. We have visualized the inferred causal relations in Figure 50.

X	Y	p-values
roek	rtr	0.000000
roek	rmar	0.000000
roek	gr_sales	0.040619
roek	rdint	0.142440
rtr	rmar	0.000000
rtr	gr_sales	0.121377
rtr	rdint	0.002308
rmar	gr_sales	0.363790
rmar	rdint	0.000146
gr_sales	rdint	0.000000

Figure 51: p-values for the unconditional tests for the variabales `roek`, `rtr`, `rmar`, `gr_sales`, `rdint`.

Including R&D intensity To include also R&D intensity and to see to whet extent the conditional independences are reproduced we now consider a set that strongly overlaps with the former, we now consider the variables `roek`, `rtr`, `rmar`, `gr_sales`, `rdint`. Figure 51 shows the p-values for the unconditional tests. We thus obtain the following independences:

$$\begin{aligned} \texttt{roek} &\perp\!\!\!\perp \texttt{rdint} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \end{aligned}$$

Figure 52 shows the p-values for all conditional independence tests. Accordingly, we accept the following conditional independences:

$$\begin{aligned} \texttt{roek} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rtr} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmar} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rdint} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{rdint} \mid \texttt{rtr} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{rdint} \mid \texttt{rmar} \\ \texttt{roek} &\perp\!\!\!\perp \texttt{rdint} \mid \texttt{gr\_sales} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{roek} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rmar} \\ \texttt{rtr} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rdint} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{roek} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rtr} \\ \texttt{rmar} &\perp\!\!\!\perp \texttt{gr\_sales} \mid \texttt{rdint} \end{aligned}$$

The undirected graph resulting from this pattern is shown in Figure 53.

Summary on innvovation expenditures In this section we focused on a variety of innovation expenditure variables, and how they related to sales growth (their correlations with the latter were weak at best). Constructing a DAG (directed acyclic graph) from the discrete additive noise model results suggested that expenditures on the acquisition of machinery, equipment and software were very much an output, rather than a causal determinant, of innovation expenditures: expenditures on machinery (etc) were causally influenced by i) expen-

X	Y	Z	p-values
roek	rtr	rmar	0.000000
roek	rtr	gr_sales	0.000000
roek	rtr	rdint	0.000000
roek	rmar	rtr	0.000000
roek	rmar	gr_sales	0.000000
roek	rmar	rdint	0.000000
roek	gr_sales	rtr	0.216564
roek	gr_sales	rmar	0.218049
roek	gr_sales	rdint	0.113354
roek	rdint	rtr	0.346965
roek	rdint	rmar	0.343331
roek	rdint	gr_sales	0.386870
rtr	rmar	roek	0.000000
rtr	rmar	gr_sales	0.000000
rtr	rmar	rdint	0.000000
rtr	gr_sales	roek	0.460852
rtr	gr_sales	rmar	0.550643
rtr	gr_sales	rdint	0.451425
rtr	rdint	roek	0.013849
rtr	rdint	rmar	0.031608
rtr	rdint	gr_sales	0.001067
rmar	gr_sales	roek	0.287705
rmar	gr_sales	rtr	0.742195
rmar	gr_sales	rdint	0.832266
rmar	rdint	roek	0.002653
rmar	rdint	rtr	0.020051
rmar	rdint	gr_sales	0.000071
gr_sales	rdint	roek	0.000209
gr_sales	rdint	rtr	0.001680
gr_sales	rdint	rmar	0.000170

Figure 52: p-values for the conditional tests for the variabales `roek`, `rtr`, `rmar`, `gr_sales`, `rdint`.

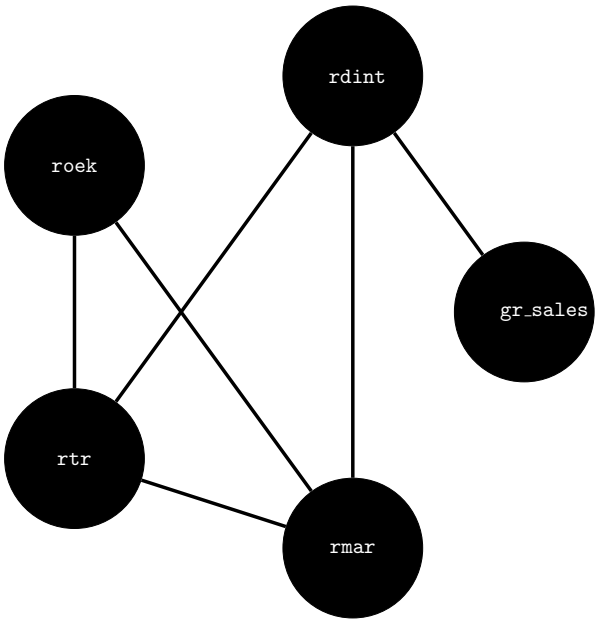

Figure 53: Undirected graph resulting from the independence pattern of `rdint`, `roek`, `rtr`, `rmar`, `gr_sales`.

ditures on the market introduction of innovations, ii) expenditures on training, and iii) expenditures on acquisition of external knowledge (e.g. licensing). A tentative policy implication would be that there is little point in targeting policy towards investments in machinery, because these will be causally influences by expenditures on other innovation variables, and will have no causal knock-on effects on other dimensions of the firm-level innovation process.

6.11 Innovation objectives

We now consider the innovation objectives variables `orange` (increase range of goods or services) `orepl` (replace outdated products or processes), `oenmk` (enter new markets) `oimks` (increase market share) `oqua` (improve quality of goods or services) `oflex` (improve flexibility for producing goods or services) `ocap` (increase capacity for producing goods or services) `ohes` (improve health and safety) `olbr` (reduce labour costs per unit output). In agreement to what is already suggested by earlier results on groups of similar variables there are no (conditional) independences within the group. Here, we obtained even p-values of 0.000000 for all unconditional and conditional tests. Accordingly, we obtain a complete graph and have to focus only on the causal *directions* using discrete additive noise models. The p-values of pairwise models are shown in Figure 54. For the confidence level 0.05, for instance, there are some examples where an additive noise model is accepted for exactly one direction. This way, we infer `oflex` $\leftarrow$ `orange`, `ohes` $\rightarrow$ `orange`. These directions are visualized in Figure 55.

X	Y	$X \to Y$	$X \leftarrow Y$
orepl	orange	0.064476	0.131941
oenmk	orange	0.000010	0.000000
oenmk	orepl	0.070170	0.040637
oimks	orange	0.000183	0.040540
oimks	orepl	0.000880	0.001953
oimks	oenmk	0.000244	0.000001
oqua	orange	0.000000	0.000000
oqua	orepl	0.000080	0.000153
oqua	oenmk	0.001051	0.000000
oqua	oimks	0.030645	0.000012
oflex	orange	0.004620	0.061415
oflex	orepl	0.041603	0.066835
oflex	oenmk	0.000016	0.001901
oflex	oimks	0.002458	0.191999
oflex	oqua	0.000212	0.001102
ocap	orange	0.003321	0.641278
ocap	orepl	0.011002	0.001409
ocap	oenmk	0.001025	0.037521
ocap	oimks	0.000081	0.000174
ocap	oqua	0.000823	0.229418
ocap	oflex	0.083761	0.004832
ohes	orange	0.125850	0.000167
ohes	orepl	0.049671	0.000023
ohes	oenmk	0.103147	0.015557
ohes	oimks	0.039759	0.013521
ohes	oqua	0.000005	0.012408
ohes	oflex	0.000001	0.002528
ohes	ocap	0.000020	0.000115
olbr	orange	0.191239	0.056070
olbr	orepl	0.057458	0.055585
olbr	oenmk	0.001657	0.001124
olbr	oimks	0.258359	0.738078
olbr	oqua	0.000146	0.000004
olbr	oflex	0.001897	0.002903
olbr	ocap	0.000000	0.000042
olbr	ohes	0.000000	0.018026

Figure 54: Pairwise inference of causal directions for all pairs (X, Y) out of the innovation objective variables orange, orepl , oimks, oflex, ocap, ohes, olbr. The two columns on the right hand side show the p-values for the respective additive noise model.

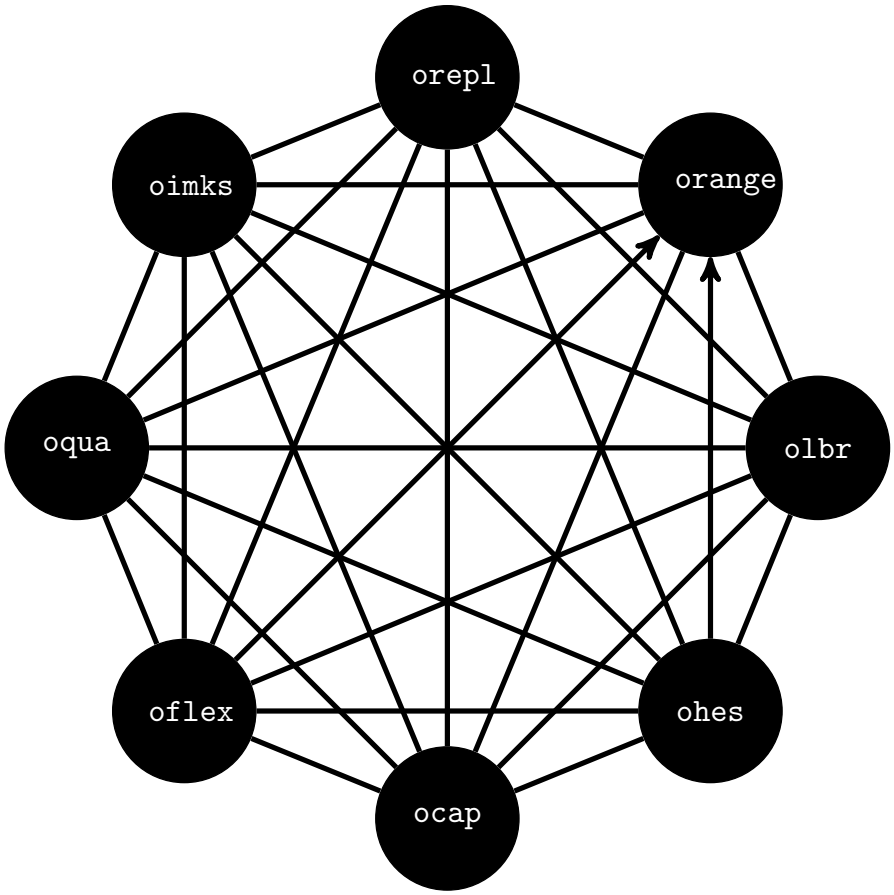

Figure 55: Causal directions between the innovation objectives variables `orange`, `orepl`, `oimks`, `oqua`, `oflex`, `ocap`, `ohes`, and `olbr`, inferred via discrete additive noise models.

Summary on innovation objectives This section investigated the causal relations between innovation objectives variables. As before, most of the relationships between variable-pairs were undirected. However, we could observe two directed edges (subject to trusting discrete additive noise models, which is questionable): i) a causal effect of improved flexibility in production on the objective to increase the range of goods or services, and ii) a causal effect of interest in improving health and safety conditions on the objective to increase the range of goods or services. Increasing either flexibility or health & safety will therefore be expected to have positive knock-on effects on increasing the range of goods and services.

6.12 Robustness of results

Replication with small sample sizes To check staibility of the conditional independence statements with respect to subsampling, we repeated the experiments of Section 6.5 regarding the variables `rrdinx`, `turn06m`, `turn08m`., this time with sample size 200 instead of 2000. Again, we obtained no unconditional independence since the p-values for the unconditional tests read:

X	Y	p-values
rrdinx	turn06m	0.000000
rrdinx	turn08m	0.000000
turn06m	turn08m	0.000000

For the conditional tests, we obtained the following p-values:

X	Y	Z	p-values
rrdinx	turn06m	turn08m	0.163928
rrdinx	turn08m	turn06m	0.187153
turn06m	turn08m	rrdinx	0.000000

Accordingly, we accept this time the following conditional independences:

$$\texttt{rrdinx} \perp\!\!\!\perp \texttt{turn06m} \mid \texttt{turn08m}$$
$$\texttt{rrdinx} \perp\!\!\!\perp \texttt{turn08m} \mid \texttt{turn06m}$$

The second independence has been rejected previously, but was also close to being accepted. In agreement with the experiment for smaple size 2000, the p-value is larger than for the first one, although only slightly.

We now repeat another experiment regarding 'hard' variables (as opposed to 'subjective' ones), namely `gr_sales`, `rdint`, `funloc`, `fungmt`, `funeu`, as considered in Section 6.6. The corresponding p-values are shown in Figure 56.

We thus accept the following independences:

$$\texttt{gr\_sales} \perp\!\!\!\perp \texttt{rdint}$$
$$\texttt{gr\_sales} \perp\!\!\!\perp \texttt{fungmt}$$
$$\texttt{gr\_sales} \perp\!\!\!\perp \texttt{funeu}$$
$$\texttt{rdint} \perp\!\!\!\perp \texttt{fungmt}$$
$$\texttt{funloc} \perp\!\!\!\perp \texttt{fungmt}$$
$$\texttt{fungmt} \perp\!\!\!\perp \texttt{funeu}$$

These independences contain all independences found in Section 6.6, but also some additional ones – which is not surprising given the smaller sample size.

X	Y	p-values
gr_sales	rdint	0.625000
gr_sales	funloc	0.080000
gr_sales	fungmt	0.319000
gr_sales	funeu	0.393000
rdint	funloc	0.005000
rdint	fungmt	0.454000
rdint	funeu	0.013000
funloc	fungmt	0.933000
funloc	funeu	0.069000
fungmt	funeu	0.820000

Figure 56: p-values for the unconditional independence tests for the variables gr_sales, rdint, funloc, fungmt, funeu, repetition of the experiment in Section 6.6, now with sample size 200.

The p-values for the conditional tests are shown in Figure 57. Accordingly, the following conditional independences are accepted:

gr_sales	$\perp\!\!\!\perp$	fungmt	| rdint
gr_sales	$\perp\!\!\!\perp$	fungmt	| funloc
gr_sales	$\perp\!\!\!\perp$	fungmt	| funeu
gr_sales	$\perp\!\!\!\perp$	funeu	| rdint
gr_sales	$\perp\!\!\!\perp$	funeu	| funloc
gr_sales	$\perp\!\!\!\perp$	funeu	| fungmt
rdint	$\perp\!\!\!\perp$	funeu	| gr_sales
rdint	$\perp\!\!\!\perp$	funeu	| fungmt
funloc	$\perp\!\!\!\perp$	fungmt	| gr_sales
funloc	$\perp\!\!\!\perp$	fungmt	| rdint
funloc	$\perp\!\!\!\perp$	fungmt	| funeu

As for unconditional tests, we obtain all the independences from Section 6.6 but also additional ones, namely the two conditional independences rdint $\perp\!\!\!\perp$ funeu, given gr_sales and fungmt, respectively.

The corresponding undirected graph representing the above pattern would contain fewer edges compared to Figure 31, but we should rather trust the ones with larger sample size.

Selection criteria Our approach of selecting companies for which rrdinx $\neq 0$ can be questioned since every selection can induce undesired bias. We have therefore replicated some experiments without applying this selection.

To this end, we consider the variables We now consider R&D and its relation to the funding variables rrdinx, funeu, fungmt, funloc, as in Section 6.6. In agreement with the results there, we obtained no unconditional independence. The following conditional independences were accepted:

funeu	$\perp\!\!\!\perp$	fungmt	| rrdinx
rrdinx	$\perp\!\!\!\perp$	funloc	| funeu
rrdinx	$\perp\!\!\!\perp$	funloc	| fungmt

All of them were also found in Section 6.6, but there we found two addi-

X	Y	Z	p-values
gr_sales	rdint	funloc	0.000020
gr_sales	rdint	fungmt	0.000433
gr_sales	rdint	funeu	0.000005
gr_sales	funloc	rdint	0.004973
gr_sales	funloc	fungmt	0.002369
gr_sales	funloc	funeu	0.000845
gr_sales	fungmt	rdint	0.247066
gr_sales	fungmt	funloc	0.118432
gr_sales	fungmt	funeu	0.176619
gr_sales	funeu	rdint	0.409844
gr_sales	funeu	funloc	0.597440
gr_sales	funeu	fungmt	0.889458
rdint	funloc	gr_sales	0.000024
rdint	funloc	fungmt	0.000006
rdint	funloc	funeu	0.000360
rdint	fungmt	gr_sales	0.000000
rdint	fungmt	funloc	0.000000
rdint	fungmt	funeu	0.000054
rdint	funeu	gr_sales	0.413939
rdint	funeu	funloc	0.071069
rdint	funeu	fungmt	0.147744
funloc	fungmt	gr_sales	0.112112
funloc	fungmt	rdint	0.328345
funloc	fungmt	funeu	0.451224
funloc	funeu	gr_sales	0.000000
funloc	funeu	rdint	0.000001
funloc	funeu	fungmt	0.000000
fungmt	funeu	gr_sales	0.002736
fungmt	funeu	rdint	0.001276
fungmt	funeu	funloc	0.000048

Figure 57: p-values for the unconditional independence tests for the variables gr_sales, rdint, funloc, fungmt, funeu, repetition of the experiment in Section 6.6, now with sample size 200.

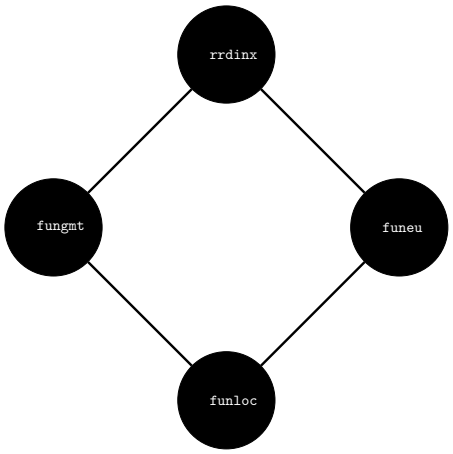

Figure 58: Undirected graph resulting from the independence pattern of `rrdinx`, `fungmt`, `funloc`, `funeu` as in Figure 24 with modified selection criterion, see text.

tional ones, namely `funloc` $\perp\!\!\!\perp$ `fungmt` $|$ `rrdinx` and `funloc` $\perp\!\!\!\perp$ `fungmt` $|$ `funeu`. This suggests that for the full data set (without conditioning on non-zero in-house R&D spending) local funding is more tightly connected to the other two funding variables than for the analysis with conditioning, which corresponds to the additional edge `fungmt` – `funloc` in Figure 58 compared to Figure 24. One should add, however, that `fungmt` $\perp\!\!\!\perp$ `funloc` $|$ `rrdinx` has been rejected based on a p-value of 0.074 which is not far from being accepted. Hence the results of the conditional tests for the sets with and without data selection differ only up to an extent that could be due to usual statistical fluctuations.

Let us also discuss the variables `rdint`, `gr_sales`, `orgbup`, `orgwkp`, `orgexr`, as in Section 6.7, since this has been an example where interesting statements on causal directions have been derived from the conditional independences alone without resorting to discrete additive noise models. We don't display all the p-values and only mention that we only accept the following unconditional independence:

$$\texttt{rdint} \quad \perp\!\!\!\perp \quad \texttt{orgexr}$$

This independence is consistent with what we obtained in Section 6.7, but there we also had `orgbup` $\perp\!\!\!\perp$ `gr_sales` in addition.

We accept the following conditional independences:

$$\begin{array}{rcll}
\texttt{gr_sales} & \perp\!\!\!\perp & \texttt{orgbup} & | \ \texttt{orgwkp} \\
\texttt{gr_sales} & \perp\!\!\!\perp & \texttt{orgbup} & | \ \texttt{orgexr} \\
\texttt{rdint} & \perp\!\!\!\perp & \texttt{orgwkp} & | \ \texttt{orgbup} \\
\texttt{rdint} & \perp\!\!\!\perp & \texttt{orgwkp} & | \ \texttt{orgexr} \\
\texttt{rdint} & \perp\!\!\!\perp & \texttt{orgexr} & | \ \texttt{orgbup} \\
\texttt{rdint} & \perp\!\!\!\perp & \texttt{orgexr} & | \ \texttt{orgwkp} \\
\texttt{rdint} & \perp\!\!\!\perp & \texttt{orgwkp} & | \ \texttt{gr_sales} \\
\texttt{gr_sales} & \perp\!\!\!\perp & \texttt{orgwkp} & | \ \texttt{orgexr}
\end{array}$$

The first six conditional independences have also been obtained in Sec-

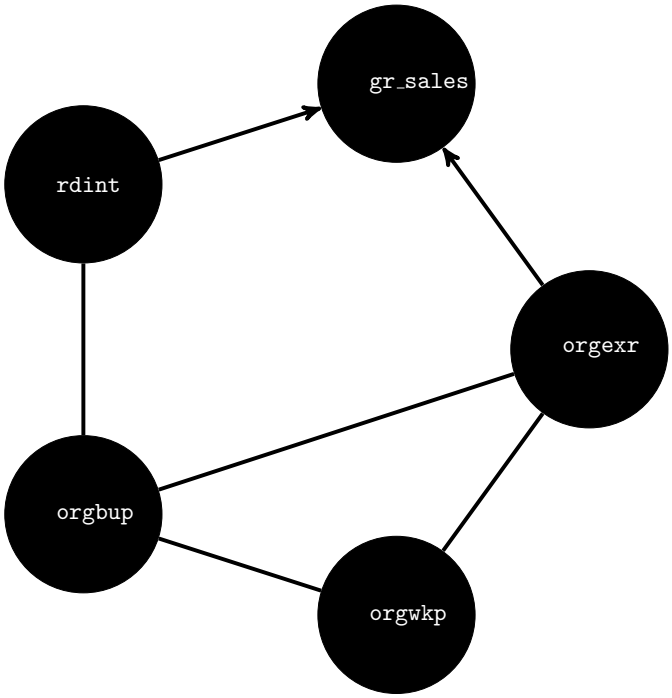

Figure 59: Partially directed graph resulting from the independence pattern of `gr_sales`, `rdint`, `orgbup`, `orgwkp`, `orgexr`, as in Figure 34, but with a modified data selection rule, see text.

tion 6.7, but the last two are new. Moreover, the independence `orgbup` $\perp\!\!\!\perp$ `gr_sales` $|$`rdint`, swhich we obtained earlier is missing in the list above. We now discuss how these differences influence the causal conclusions. The result is shwon in Figure 59, which coincides with Figure 32 apart from the undirected edge `orgwkp` – `gr_sales` which is missing now. Most importantly, again `rdint` and `orgexr` have an arrow to `gr_sales`, i.e., the crucial finding remains the same.

7 Conclusions

This report sought to apply a new range of techniques to some well-known innovation databases, to discover the causal relations between key innovation variables. Groups of variables were analyzed together, and by examining the conditional and unconditional dependence relations (using correlations, HSIC statistics, and analysis of conditionals), we could infer the direction of causality. Indeed, the ability to get causal estimates, rather than just statistical relationships, is of fundamental importance for policymakers.

Our results can complement the existing literature to show some new results and derive some new policy implications. For example, we observed that firm sales (not market value) is a driver of R&D expenditures (rather than vice versa), which suggests that policies to increase firm's R&D could first seek to help

firms boost their sales. Local government funding for innovation is not closely related to national or European funding - which might raise concerns that local funding might not be as effective. R&D sometimes influences receipt of funding, rather than vice versa, which provides new insights into debates about the additionality of innovation funding schemes. Organizational innovations that specifically concern organizing external relations with other firms/institutions were seen to contribute directly to sales growth. New logistics methods are seen to cause new methods of process innovations for producing goods or services. Acquisition of external knowledge, and market introduction of innovations, both cause increases in acquisition of machinery, equipment and software.

Future work could replicate our analysis on more recent CIS data, to investigate the robustness of our findings, or also apply these techniques to other datasets.

Acknowledgements

The author would like to thank Alexander Coad and Paul Nightingale for helpful and detailed comments on an earlier version of this manuscript.

A Description of the variables in the CIS data set

emp06	Number of employees in 2006 (categorical)
emp08	Number of employees in 2008 (categorical)
funeu	Receipt of public financial support for innovation: EU
fungmt	Receipt of public financial support for innovation: central government
funloc	Receipt of public financial support for innovation: local or regional authorities
gr_sales	defined as log(turn08m) - log(06m)
inpcsw	Who developed the process innovations: mainly your enterprise, collaboration, mainly others
inpslg	You introduced new logistics, delivery or distribution methods
inpsnm	Were any process innovations new to your market? Yes, no, don't know
inpspd	You introduced new methods of manufacturing or producing goods or services
inpssu	You introduced new supporting activities for your processes, such as maintenance systems
larmar	Which is your largest market: local/regional, national, other EU, all other countries
mareur	Market where goods sold: EU
marloc	Market where goods sold: local/regional
marnat	Market where goods sold: national
maroth	Market where goods sold: other non-EU
ocap	Innovation objectives: increase capacity for producing goods or services
oenmk	Innovation objectives: enter new markets
oflex	Innovation objectives: improve flexibility for producing goods or services
ohes	Innovation objectives: improve health and safety
oimks	Innovation objectives: increase market share
olbr	Innovation objectives: reduce labour costs per unit output
oqua	Innovation objectives: improve quality of goods or services
orange	Innovation objectives: increase range of goods or services
orepl	Innovation objectives: replace outdated products or processes

orgbup	Organisational innovation: introduced new business practices
orgexr	Organisational innovation: new methods of organizing external relations
orgwkp	Organisational innovation: new methods of organizing work responsibilities & decision making
rmac	Expenditure on innovation activities: acquisition of machinery, equipment and software
rmar	Expenditure on innovation activities: market introduction of innovations
roek	Expenditure on innovation activities: acquisition of external knowledge
rrdex	Expenditure on innovation activities: purchase of external R&D
rrdinx	Expenditure on innovation activities: inhouse R&D
rtr	Expenditure on innovation activities: training for innovative activities
scli	Sources of information and cooperation: clients or customers
scom	Sources of information and cooperation: competitors
scon	Sources of information and cooperation: conferences, trade fairs, exhibitions
sentg	Sources of information and cooperation: within your enterprise
sgmt	Sources of information and cooperation: govt or public research institutes
sins	Sources of information and cooperation: consultants, commercial labs, private R&D
sjou	Sources of information and cooperation: scientific journals and trade/technical publications
spro	Sources of information and cooperation: professional and industry associations
ssup	Sources of information and cooperation: suppliers of equipment, materials, components, or software
suni	Sources of information and cooperation: universities or higher education institutions
turn06m	Total turnover in 2006
turn08m	Total turnover in 2008
rdint	R&D intensity, defined as rrdinx/turn08m

B List of software packages

Software that has been provided to the European Commission in form of a single zip-file (sent to Alexander Coad):

Routines for general data sets, described in the grey boxes in Section 4:

ind_test.R
kci_test.R
pairwise_lingam.R
pairwise_anm.R

Routines for analyzing the Scoreboard data (described in the grey boxes in Section 5:

compute_growth_rates_scoreboard.R
plot_NS_RD_for_random_companies.R
print_correlation_histograms_scoreboard.R
print_hsic_histograms_scoreboard.R
generate_scatter_plots_growth_rates.R
plot_histogram_correlation_delay.R
plot_histogram_lingam_scores.R
plot_histogram_anm_scores.R
analyze_scoreboard_for_one_sector.R

Routines for analyzing the CIS data set, described in the grey boxes in Section 6:

load_cis_data.R
cis_arbitrary_subsets.R
pairwise_lingam_cis.R
apply_discrete_anm.R
infer_causal_direction_for_all_pairs_discrete.R
plot_conditional_continuous_to_binary.R

Weblinks to publicly available software to be downloaded in addition:

1. Kernel Conditional Independence Test:

   `http://people.tuebingen.mpg.de/kzhang/KCI-test.zip`

2. Continuous additive noise models:

   `https://staff.fnwi.uva.nl/j.m.mooij/publications.html`, at the link 'code' for the article Mooij et al. (2016)

3. Discrete additive noise models:

   `http://webdav.tuebingen.mpg.de/causality/online_aistats_arxive_discrete.zip`

If you have problems downloading the software for this project, please write to Alexander.Coadec.europa.eu for assistance.

References

E. Bareinboim and J. Pearl. Controlling selection bias in causal inference. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7 –11*, pages 247–254, 2011.

A. Coad. *The Growth of Firms: a Survey of Theories and Empirical Evidence.* Edward Elgar, Cheltenham, UK, 2009.

A. Coad and M. Binder. Causal linkages between work and life satisfaction and their determinants in a structural var approach. *Economics Letters*, pages 263–268, 2014.

A. Coad and N. Grassano. Causal relations between sales growth, employment growth, profits growth, assets growth and r&d growth: Svar evidence from sb companies. 6th IRIMA workshop, December 2015.

Eurostat. Work Session on Statistical Data Confidentiality, Manchester, 17-19 December 20. Office for Official Publications of the European Communities, Luxembourg, Retrieved April 12th, 2016. `http://ec.europa.eu/eurostat/en/web/products-statistical-working-papers/-/KS-78-09-723`, 2009.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of the 16th Conference on Algorithmic Learning Theory*, pages 63–77, Berlin, 2005a. Springer-Verlag.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Proceedings of the conference Neural Information Processing Systems (NIPS) 2008*, Vancouver, Canada, 2009. MIT Press.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

D. Janzing and B. Steudel. Justifying additive-noise-based causal discovery via algorithmic information theory. *Open Systems and Information Dynamics*, 17(2):189–212, 2010.

D. Janzing, X. Sun, and B. Schölkopf. Distinguishing cause and effect via second order exponential models. `http://arxiv.org/abs/0910.5561`, 2009.

D. Janzing, P. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel*, 06:479–486, 2010.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Annals of Statistics*, 41(5):2324–2358, 2013.

Y. Kano and S. Shimizu. Causal inference using nonnormality. In *Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion*, pages 261–270, Tokyo, Japan, 2003.

J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, pages 1–23, 7 2012.

J. Mairesse and P. Mohnen. Using innovation surveys for econometric analysi. In B. Hall and N. Rosenberg, editors, *Handbook of the Economics of Innovation*, chapter 26. 2010.

A. Moneta, D. Entner, P. Hoyer, and A. Coad. Causal inference by independent component analysis: Theory and applications*. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730, 2013.

J. Mooij, D. Janzing, B. Schölkopf, and T. Heskes. Causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems 24, Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS 2011), Curran*, pages 639–647, NY, USA, 2011. Red Hook.

J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

J. Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2000.

J. Peters, D. Janzing, and B. Schölkopf. Identifying cause and effect on discrete data using additive noise models. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP 9,*, Chia Laguna, Sardinia, Italy, 2010.

J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2436–2450, 2011.

J. Peters, JM. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15: 2009–2053, 2014.

H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.

B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32th International Conference on Machine Learning (ICML)*, pages 285–294. Journal of Machine Learning Rearch, 2015.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, prediction, and search (Lecture notes in statistics)*. Springer-Verlag, New York, NY, 1993.

X. Sun, D. Janzing, and B. Schölkopf. Causal inference by choosing graphs with most plausible Markov kernels. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pages 1–11, Fort Lauderdale, FL, 2006.

P. Thompson. Rationality, rules of thumb, and r&d. *Structural Change and Economic Dynamics*, 10:321–340, 1999.

L. Tornqvist, P. Vartia, and Y. Vartia. How should relative changes be measured? *American Statistician*, 39, 1985.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pages 255–270, New York, NY, 1990. Elsevier Science Publishers.

K. Zhang, P. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011. `http://uai.sis.pitt.edu/papers/11/p804-zhang.pdf`.

K. Zhang, J. Zhang, B. Huang, B. Schölkopf, and C. Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *Proceedings of the 32th Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, New York City, 2016.

J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011. `http://uai.sis.pitt.edu/papers/11/p839-zscheischler.pdf`.