COLLOQUIUM PAPER

Methods for causal inference from gene perturbation experiments and validation

Nicolai Meinshausen[a], Alain Hauser[b], Joris M. Mooij[c], Jonas Peters[d], Philip Versteeg[c], and Peter Bühlmann[a,1]

[a]Seminar for Statistics, Eidgenössische Technische Hochschule (ETH) Zurich, CH-8092 Zurich, Switzerland; [b]Department of Engineering and Information Technology, Bern University of Applied Sciences, CH-3400 Burgdorf, Switzerland; [c]Informatics Institute, University of Amsterdam, 1090 GH Amsterdam, The Netherlands; and [d]Max Planck Institute for Intelligent Systems, D-72076 Tuebingen, Germany

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved April 5, 2016 (received for review June 5, 2015)

Inferring causal effects from observational and interventional data is a highly desirable but ambitious goal. Many of the computational and statistical methods are plagued by fundamental identifiability issues, instability, and unreliable performance, especially for large-scale systems with many measured variables. We present software and provide some validation of a recently developed methodology based on an invariance principle, called invariant causal prediction (ICP). The ICP method quantifies confidence probabilities for inferring causal structures and thus leads to more reliable and confirmatory statements for causal relations and predictions of external intervention effects. We validate the ICP method and some other procedures using large-scale genome-wide gene perturbation experiments in *Saccharomyces cerevisiae*. The results suggest that prediction and prioritization of future experimental interventions, such as gene deletions, can be improved by using our statistical inference techniques.

interventional–observational data | invariant causal prediction | genome database validation | graphical models

In this article, we discuss statistical methods for causal inference from perturbation experiments. As this is a rather general topic, we focus on the following problem: based on data from observational and perturbation settings, we want to predict the effect and outcome of an unseen and new intervention or perturbation. Taking applications in genomics as an example, a typical task is as follows: based on observational data from wild-type organisms and interventional data from gene knockout or knockdown experiments, we want to predict the effect of a new gene knockout or knockdown on a phenotype of interest. For example, the organism is the model plant *Arabidopsis thaliana*, the gene knockouts correspond to mutant plants, and the phenotype of interest is the time it takes until the plant is flowering (1).

From a methodological viewpoint, the prediction of unseen future interventions belongs to the area of causal inference where one aims to quantify presence and strength of causal effects among various variables. Loosely speaking, a causal effect is the effect of an external intervention (or say the response to a "What if I do?" question). The corresponding theory, e.g., using Pearl's do-operator (2), provides a link between causal effects and perturbations or randomized experiments. We mostly assume here that all of the variables in the causal model (for inferring causal effects) are observed: the case with hidden variables is mentioned only briefly in a later section, although it is an important theme in causal inference (due to the problem of hidden confounding variables) (cf. refs. 2 and 3).

A popular and powerful route for causal modeling is given by structural equation models (SEMs) (2, 4). We consider a set of random variables $X_1, \ldots, X_p, X_{p+1}$, and we often denote by $Y = X_1$, emphasizing that Y is our response variable of interest (e.g., a phenotype of interest). The main building blocks of a SEM are as follows: (i) an underlying true causal influence diagram for the random variables $X_1, \ldots, X_p, X_{p+1}$, formulated with a directed graph D whose nodes correspond to the variables, most often with a directed acyclic graph (DAG); (ii) each of the random variables is modeled as a function of their parental variables, given by the graph D, and an error term. The system of structural equations is then as follows:

$$X_j \leftarrow f_j(X_{\mathrm{pa}(j)}, \varepsilon_j) \ (j = 1, \ldots, p+1), \qquad [1]$$

where $\mathrm{pa}(j)$ denotes the set of parents of node j in the underlying graph or DAG D and ε_j are error terms that are jointly independent. Furthermore, for $S \subseteq \{1, \ldots, p+1\}$, X_S denotes the variables $\{X_j;\ j \in S\}$, and the arrow "$\leftarrow$" is emphasizing that X_j is caused (or influenced) by $X_{\mathrm{pa}(j)}$, which is a stronger statement than an algebraic equality.

The most commonly used model for an intervention at one or several variables has been pioneered by Pearl (cf. ref. 2): the do-operation $\mathrm{do}(X_j = x)$ is setting the single variable X_j to a deterministic value x, which corresponds to replacing the structural equations [1] for X_j with $X_j \leftarrow x$; analogously, the do-operation can be applied to several variables simultaneously. The distribution $p(y|\mathrm{do}(X_j = x))$ of Y when doing an intervention $\mathrm{do}(X_j = x)$ can be derived via the truncated Markov factorization (2, 3, 5) or by the backdoor adjustment formula, and we can then consider quantities like the expected response Y when having done an intervention at X_j putting its value to x:

$$\mathbb{E}[Y|\mathrm{do}(X_j = x)] = \int y p(y|\mathrm{do}(X_j = x))\, dy. \qquad [2]$$

For more details, we refer to ref. 2. The do-operation has been generalized to probabilistic "soft" interventions where the intervention value (little x in the notation above) becomes a random variable (6). Furthermore, a so-called "mechanism change" with an intervention at variable with index j is replacing the conditional probability distribution $p(X_j|X_{\mathrm{pa}(j)})$, corresponding to the jth equation in the SEM [1], by another distribution $q(X_j|X_{\mathrm{pa}(j)})$ (7). In addition, "fat-hand" interventions (8) (with uncertain intervention targets) and activity interventions (9) (simultaneous mechanism changes of all children of a variable) have been used to model interventions in molecular biology. For the do-interventions, it is sufficient to know the SEM because the intervention $\mathrm{do}(X_j = x)$ itself is fully specified by the known quantities j and x. We will discuss in a section below that do-interventions can be too simple for certain applications.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: N.M. and P.B. designed research; N.M., A.H., J.M.M., J.P., and P.B. performed research; A.H., J.M.M., J.P., and P.V. analyzed data; and N.M. and P.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: buhlmann@stat.math.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510493113/-/DCSupplemental.

Identifiability and Estimation from Data

One of the major challenges is the estimation of the SEM [**1**] from observational or a mix of observational and interventional data. We are particularly interested in high-dimensional settings where the number of variables $p+1$ can be much larger than sample size, as in many applications from, e.g., genomics and genetics.

A first complication concerns identifiability: the data-generating probability distribution(s) might be represented by different structures (acyclic directed graphs) D and corresponding different functions in the SEM [**1**]. The different graph structures D that can generate the data-generating distribution(s) build an equivalence class $\mathcal{D}$. Situations where this equivalence class is large (and hence the degree of identifiability is low) occur when the data-generating distribution is observational and the SEM in [**1**] is either fully nonparametric with no specified additional structure or when the functions $f_j(\cdot)$ are linear and the error terms are Gaussian (cf. ref. 2). More identifiability is possible when the data-generating distribution corresponds to a mix of observational and interventional data (7–12) or when the SEM has additional structure. Regarding the latter, it is possible to identify from the observational data distribution P the single underlying causal DAG: the most prominent examples are linear, non-Gaussian, acyclic models (LiNGAM) where the functions f_j are linear but all error terms are non-Gaussian (13), the functions are nonlinear and the error terms are additive (14, 15), or the functions f_j are linear with Gaussian error terms that all have the same variances (16).

Algorithms and Methods. Given data, we want to estimate the SEM in [**1**] (its equivalence class if it is not identifiable), and based on this, we often aim to estimate the total causal average effect $d/dx(\mathbb{E}[Y|\mathrm{do}(X_j=x)])$ (see also [**2**]) or its generalization when intervening at more than one variable. If the underlying causal DAG D is not identifiable from the distributions, we can only obtain bounds for $d/dx(\mathbb{E}[Y|\mathrm{do}(X_j=x)])$.

For linear Gaussian SEMs, estimation of the Markov equivalence class based on observational data can be done by penalized maximum-likelihood estimation (17) or by constraint-based methods with the PC-algorithm using conditional independence testing (3, 18). Based on the estimated Markov equivalence class, lower bounds for the absolute value of the parameter $\alpha_{j(x)}=d/dx(\mathbb{E}[Y|\mathrm{do}(X_j=x)])$ can be derived using a computationally efficient strategy (19). The setting with a mix of observational and interventional data and estimation of the corresponding smaller Markov equivalence class is discussed in ref. 20. Bayesian methods for structure and parameter estimation include those in refs. 8, 9, 21, and 22. Theoretical performance guarantees in the high-dimensional setting with underlying sparse DAGs have been given in refs. 19, 23, and 24.

For identifiable models, some other estimation strategies have been proposed. For linear SEMs with non-Gaussian errors (LiNGAM), one can make use of independent component analysis (13), and for additive SEMs, proposals include independence testing of residuals (15) or penalized nonparametric maximum-likelihood estimation (25). Based on an estimated causal graph, the quantity $\mathbb{E}[Y|\mathrm{do}(X_j=x)]$ in [**2**] can be nonparametrically inferred using marginal integration for the backdoor formula adjustment (26).

Sometimes, the direct (instead of total) causal effects are of interest. They are typically given by the parameters of a graphical or SEM. For example, the edge function f_j in [**1**] encodes all of the direct effects from $X_{\mathrm{pa}(j)}$ to X_j, or the parameter γ_j^* in [**4**] describes the direct effect of X_j to Y.

Challenges and Validation. There are a number of difficulties that occur, implying that estimation of a causal graph or influence diagram or of a total causal effect is a very ambitious goal, particularly in the high-dimensional context where p is much larger than sample size. Even with simulated data from a specific model under consideration, one often needs a substantial sample size to ensure that the estimated (equivalence class of) graphs or causal effects are fairly accurate. The supporting mathematical theory is often of crude asymptotic nature as sample size tends to infinity and does not advance more detailed understanding; for an exception, presenting some more refined convergence rates, see ref. 24. Furthermore, the faithfulness assumption can become rather severe for moderate dimensions already (27, 28). (A distribution P is faithful with respect to a DAG D if all conditional and marginal dependencies among the variables can be derived from the DAG D.)

Quantifying uncertainty. When using methods based on estimating first a causal graph or an equivalence class thereof and subsequently inferring the direct or total causal average effect, it seems difficult to accurately quantify uncertainty in terms of confidence bounds. Confidence regions based on sample-splitting procedures are rather unreliable and much worse than for inferring association in regression models (29). Thus, in absence of "error bars," one cannot draw reliable confirmatory conclusions. We will discuss later a recently proposed methodology that provides confidence intervals for direct causal effects, in the setting with not only observational but also additional, rather general kind of interventional data.

Validation. Because (asymptotic as sample size tend to infinity) correctness of causal inference relies on strong assumptions, with some of them being uncheckable in practice, a central point is their empirical validation with new interventional data. Perhaps best posed is the validation of the "total causal effect." [A total causal effect of X to Y measures the overall effect on Y when doing a perturbation at X. Its expected value (as a function of x) is defined in ref. 2.] This can be done by holding out some interventional test data. Mathematically, we aim to predict the expected value $\mathbb{E}[Y|\mathrm{do}(X_j=x)]$ in [**2**]. The prediction error is then a measure between the estimated quantity of $\mathbb{E}[Y|\mathrm{do}(X_j=x)]$, based on the training data, and the actual value of the variable Y under the perturbation $\mathrm{do}(X_j=x)$ in a new (test data) intervention experiment. Instead of this qualitative measure, we can also test for the existence of a total causal effect. Mathematically, this is the case if $\mathbb{E}[Y|\mathrm{do}(X_j=x)]$ is different from $\mathbb{E}[Y]$ for some x; see [**2**]. To validate this in empirical data, we define below a notion of (total) "strong intervention effect" (SIE); it is a binary event and allows to validate whether a method (based on training data) was successful in correctly predicting the binary outcome of a SIE being present or absent in new test data.

We note that a procedure that is estimating direct causal effects, as most causal prediction methods do, can also be used to predict total (strong) intervention effects; for more details, see *SI Appendix*.

Validation of estimated lower bounds of the total causal effects in a large-scale gene deletion experiment for yeast has been performed in ref. 30: however, the result depends in a rather sensitive way on how a true positive finding is defined. We thus suggest next a conservative notion for validation of total causal effects.

We propose here the criterion of SIE, which is well suited for validation with large-scale interventions having one measurement each, such as gene knockdowns or deletions. It conservatively classifies total intervention (causal) effects as being strong or not. Consider a variable X_j that is intervened on and a response variable Y of interest.

SIE. The intervention of variable X_j on the response Y is strong if both of the following events occur:

i) the intervened variable X_j has a value that is below or above all values of variable X_j seen in other interventional/observational data (with no intervention on X_j);
ii) the response variable Y has a value below or above the range of values of Y in all other interventional/observational data (with no intervention on Y).

Thus, the definition of an SIE is based on a given dataset. An SIE of X_j to Y corresponds to an event with corresponding extreme behavior of the realized values of the two variables. It is an estimate of presence or absence of a strong total causal effect. We note that, for a given Y, there can only be at most one X_j that fulfills the

criterion. Therefore, the criterion is conservative, meaning that not all non-SIE effects are noncausal.

The existence of a "direct causal effect" is difficult to extract from hold-out interventional validation data, unless we adopt a model that we want to avoid for validation purposes. (In an SEM as in [1], there is a direct effect from X to Y if X is a parent of Y.) In *Validation: Gene Perturbation Experiments*, we consider scores measuring direct effects using external information from a genome database and transcription factor (TF) binding based on ChIP-on-chip data as a source of indirect evidence for a direct effect. With such approaches, we have to keep in mind that the external source of validation might be very noisy and error-prone. In particular, it turns out to be difficult to predict SIEs from external information alone.

Causal Inference Based on Invariance Across Experiments

We outline here a recently published method (31) that exploits the fact that the data arise from different experimental conditions or perturbations. In the advent of big-data scenarios, the latter setting with heterogeneous data sources becomes more common. The method has a few crucial benefits addressing some of the difficulties mentioned in the previous section: (*i*) an "automatic identifiability" property (see the discussion after [5]); (*ii*) some confidence bounds for inferring causal variables (see [5]); (*iii*) the flexibility that the interventions and perturbations do not need to be exactly specified; and (*iv*) avoiding some typically unstable and complicated estimation of a graph (or an equivalence class of graphs) from data.

The method is based on invariance of conditional distributions across intervention experiments from a rather general type. The role of invariance in causal inference has received some attention in the literature (2, 32–34). To the best of our knowledge, however, the work in ref. 31 is the first of its kind that exploits invariance of conditional distributions for statistical estimation and confidence statements.

As before, we consider a response or target variable of interest, denoted by Y, and a p-dimensional predictor variable $X=(X_1,\ldots,X_p)$. We assume a setting with data (Y^e,X^e) for different experimental settings $e\in\mathcal{E}$. For example, with $\mathcal{E}=\{1,2\}$ in the context of gene perturbation experiments, the experimental settings could correspond to observational data ($e=1$) and data from unspecified interventions ($e=2$). We could also consider a larger set of experimental settings $\mathcal{E}=\{1,2,3,4\}$ when having in addition data from say two gene-specific interventions, encoded in addition by $e=3$ and $e=4$. Thereby, Y^e is a $n_e\times 1$ vector of the response variable and X^e an $n_e\times p$ design matrix, containing the n_e different data points in the setting e. It is important to point out that we have more than say observational data only (assuming that $\mathcal{E}$ does contain more than one element of experimental settings). We use a linear model for the response or target variable:

$$Y^e=X^e\gamma^e+\varepsilon^e, \qquad [3]$$

where the error or noise term ε^e has mean zero. We note that the regression vector γ^e and the noise term ε^e are unknown or unobservable, respectively. An intercept that is constant across environments could be added, but we will not do so here for notational simplicity. We refer to *SI Appendix* for some potential violations of the assumed linearity in [3].

The response variables in [3] are assumed to correspond to a linear SEM:

$$Y\leftarrow\sum_{k\in S^*}\gamma_k^*X_k+\varepsilon_Y, \qquad [4]$$

where ε_Y is a noise term that is independent from X_{S^*}, and $S^*=\mathrm{pa}(Y)$ is unique and equals the parental set of Y and γ^* corresponds to the coefficients (edge weights) in such a SEM. The variables Y^e and X^e are generated from a rather general class of interventions on X. We require that these interventions, or the corresponding experimental settings, are such that the following invariance assumption holds. (The invariance assumption will be exploited in the next section: the main idea will be to look for components of the regression vector that are invariant among experimental settings.)

Invariance Assumption. For S^* and γ^* from [4], define a $p\times 1$ vector γ such that $\gamma_j=\gamma_j^*(j\in S^*)$ and $\gamma_j=0\ (j\notin S^*)$. Then:

$$\begin{aligned}&\text{for all } e\in\mathcal{E}: Y^e=X^e\gamma+\varepsilon^e,\\ &\varepsilon^e \text{ has the same distribution for all } e\in\mathcal{E} \text{ and}\\ &\varepsilon^e \text{ is independent of } X^e_{S^*} \text{ for all } e\in\mathcal{E}.\end{aligned}$$

We note that $S^*=\mathrm{pa}(Y)$ are the causal variables for Y (sometimes called the direct causes of Y) and the distribution for ε^e (for all $e\in\mathcal{E}$) is equal to the one of ε_Y in [4].

As an example, consider experimental settings $\mathcal{E}$ that arise from do-interventions (2) at variables different from Y in a SEM as in [4]: then the invariance assumption holds.

We give in the following a simple example under noise interventions. Assume the SEM for a target of interest Y and two potentially causal variables X_1,X_2 is given by the following (to improve readability, we omit the superscript "e" and write X_1,X_2,Y instead of X_1^e,X_2^e,Y^e):

$$\begin{aligned}Y&\leftarrow X_2+\varepsilon_y\\ X_1&\leftarrow 2Y+\sigma(e)\varepsilon_1\\ X_2&\leftarrow\sigma(e)\varepsilon_2,\end{aligned}$$

where $\varepsilon_y,\varepsilon_1,\varepsilon_2$ are independent with mean zero and unit variance and the strength of the noise $\sigma(e)$ is a function of the environment. The true causal parent of Y is just the second variable $S^*=\{2\}$. A simple regression from Y on the two variables X_1,X_2 will put a nonzero regression coefficient on both variables (even though X_1 is a child of Y in the causal graph and hence not causal for Y, it has predictive power for the outcome Y). For the causal discovery, we propose, intuitively speaking, to look through all possible subsets S of $\{1,2\}$. For each subset S, we ask whether it is possibly a parental set of the outcome of interest by checking an invariance property across different environments. For the true causal parents, the regression coefficients when regressing Y on X_S and the residual variance will be identical across environments. Let $\gamma_S(e)$ be the optimal regression coefficient when regressing Y onto X_S for a given subset S of predictor variables (a function of the environment e) and let $V_S(e)$ be the residual variance of $\mathrm{Var}(Y-X_S\gamma_S(e))$. The two environments $\mathcal{E}=\{1,2\}$ are defined in this example by a change in the noise level so that $\sigma(e=1)=1$ and $\sigma(e=2)=2$ (this fact does not have to be known; the change could also consist of do-interventions or other types of interventions on X_1,X_2, and we do not require knowledge of the precise location or type of intervention). The regression coefficients and residual variances in the two environments are then given for all possible subsets of variables by Table 1.

We can see from Table 1 that both the optimal regression coefficient γ_S and the residual variance V_S stay constant in all environments for the true set of causal parents (here, if S is equal to $S^*=\{2\}$), whereas the residual variance changes for all other subsets of variables, including the empty set $S=\emptyset$ that would correspond to Y being a root node in the SEM. The invariant causal prediction

Table 1. Invariance example

Set	$\gamma_S(e=1)$	$\gamma_S(e=2)$	Invariant?	$V_S(e=1)$	$V_S(e=2)$	Invariant?
$S=\emptyset$	—	—	Yes	2	5	No
$S=\{1\}$	4/9	5/12	No	18/81	5/6	No
$S=\{2\}$	1	1	Yes	1	1	Yes
$S=\{1,2\}$	(2/5, 1/5)	(1/4,1/2)	No	1/5	1/2	No

Table 2. Timing comparisons in minutes

Method	No. genes		
	50	500	5,000
ICP	0.233	2.64	27.7
hiddenICP	0.012	0.12	1.4
pc	0.004	0.10	2.4
rfci	0.004	0.12	3.6
ges	0.002	0.80	1,002.4
gies	0.010	4.06	842.8
Regression	0.069	0.70	7.5

(ICP) method works by collecting all subsets S of variables for which we cannot statistically reject the hypothesis that $\gamma(e)$ is identical in all environments $e \in \mathcal{E}$ and can also not reject the hypothesis that $V(e)$ is constant in all environments $e \in \mathcal{E}$. In the population example above, only $S = \{2\}$ satisfies this invariance. Once we have collected all subsets S of variables for which we cannot reject invariance, we take the empirical estimate $\hat{S}$ to be the intersection across all of the subsets with the invariance property (where we cannot statistically reject the hypothesis of invariance); that is, we look for variables that are common among all invariant subsets. For sufficiently many data points, in the example this yields the answer $\hat{S} = \{2\}$, and we thus detect that variable 2 has to be causal for the outcome Y. Note that, if we just observed a single environment, we would see invariance for all subsets S, including the empty set (because invariance across one environment always holds). The intersection across all invariant sets would thus be the empty set, and we could not determine that one or more of the variables is causal for the outcome of interest from observational data alone. The same phenomenon would occur if the strengths $\sigma(e = 1)$ and $\sigma(e = 2)$ of the noise in both environments had the same value.

Invariant Prediction Method. Having data as in [**3**] from various experimental settings, the main idea (as outlined in the example above) is to look for sets of predictor variables that leave the corresponding regression vectors and noise terms invariant across experimental settings. This is the basis of the invariant prediction method (31) (more details in *SI Appendix*). Here, we simply present the main result.

As with any causal inference method, one might face identifiability problems. This is a fundamental and unavoidable issue. However, assuming the invariance assumption, the ICP approach will lead to a set $\hat{S}(\mathcal{E}) \subseteq \{1, \ldots, p\}$ (*SI Appendix*, formula **S.4**), which has the following confidence property:

$$\mathbb{P}\left[\hat{S}(\mathcal{E}) \subseteq S^* = \mathrm{pa}(Y)\right] \geq 1 - \alpha, \qquad [5]$$

for some prespecified confidence level $1 - \alpha$ such as 0.95 or 0.99. [The set $\hat{S}(\mathcal{E})$ depends on the data, and hence is random when interpreting the data as usual as realizations of random variables.] Thus, when applying the method many times to different datasets, we expect that in approximately $(1 - \alpha) \times 100\%$ of the cases, all of the selected variables are causal (i.e., direct causes). The main assumption for deriving the statement in [**5**] is the invariance assumption. As described above, it holds if the experimental settings $\mathcal{E}$ come from a rather broad class of interventions that do not directly act on the response variable Y (*SI Appendix*). As an example where this assumption is plausible in practice, consider gene perturbation experiments and assume a phenotypic response Y and predictor variables corresponding to expression values of all of the genes in the genome (the columns of the matrix X). Suppose that the interventions act on some of the (possibly vaguely specified) genes, such as gene deletions or gene knockdowns. Then, these interventions do not directly target the response Y, and thus, the main assumption outlined above is satisfied.

The elegance of the method and of the statement is that we do not need to know or specify whether a causal variable is identifiable from the data-generating probability distribution: the method automatically takes care about potential identifiability problems. (We do not know whether the method is complete: that is, whether all direct causal effects that are identifiable from the data-generating distribution would be correctly detected by the method.)

Heterogeneity in Big Data. The ICP method outlined above crucially depends on the fact that we have access to different experimental conditions from $\mathcal{E}$. In presence of say observational data alone ($\mathcal{E}$ consisting of one experimental setting only, i.e., $|\mathcal{E}| = 1$), we would not detect any causal variable. [Causal inference from observational data would require approaches based on, e.g., fitting SEMs and graphical modeling (*Identifiability and Estimation from Data*).]

With the ICP method, the degree of identifiability increases as the space of experimental settings $\mathcal{E}$ becomes larger. Denote the causal variables that are identifiable from the invariance assumption by $S(\mathcal{E}) \subseteq \{1, \ldots, p\}$ (*SI Appendix*, formula **S.1**). We then have that

$$S(\mathcal{E}) \nearrow \text{ as } \mathcal{E} \nearrow, \qquad [6]$$

meaning that for $\mathcal{E}_2 \supseteq \mathcal{E}_1$ we have $S(\mathcal{E}_2) \supseteq S(\mathcal{E}_1)$. Sufficient conditions under which $S(\mathcal{E}) = S^*$, that is the causal variables are uniquely identifiable, have been worked out for linear Gaussian SEMs, requiring that $\mathcal{E}$ is sufficiently "rich" and form a certain class of interventions (31).

From [**6**], we conclude that with a larger amount of experimental settings ("more heterogeneity") we have higher degree of identifiability of causal effects. Thus, in the setting of big data with a large space $\mathcal{E}$ of experimental settings, we have an advantage to exploit invariance across a large $\mathcal{E}$. For example, with gene perturbation experiments discussed below, it can be of interest to consider experimental settings not only from different gene interventions but also arising from change of environments from potentially different datasets.

The ICP method can deal with a general variety of experimental settings, including observational data, known interventions of a certain type at a known variable, random interventions at an unknown variable, or observational data in a changed environment. We emphasize the importance that one does not need to know what the working experimental conditions from $\mathcal{E}$ actually mean. In practice, it is often difficult to know whether an intervention has been done at, e.g., one specified variable only, or to specify the kind of intervention that has been done, e.g., a do-intervention (2) or a "soft" intervention (6). The fact that one does not need to specify the nature of an experimental setting in $\mathcal{E}$ contributes to robustness and generality of the procedure. The only necessary background knowledge regarding the types of interventions is that they do not target the response variable Y itself.

It has been assumed so far that $\mathcal{E}$ is the set of the true available experimental settings, but this is not necessary. In principle, we can construct the working experimental settings $\mathcal{E}$ as we like and still obtain [**5**], as long as the invariance assumption is satisfied with respect to $\mathcal{E}$. (The data in such a constructed experimental setting has then a mixture distribution. For an experiment $e \in \mathcal{E}$, the mixture distribution is $\sum_{j \in \mathcal{A}} w_j^e F_j$, where $\mathcal{A}$ is the entire space of all possible experimental settings, w_j^e are positive weights summing up to 1, and F_j are probability distributions.)

Invariance in Presence of Hidden Variables. The ICP method from the previous section implicitly assumes that there are no hidden confounding variables. The method can be generalized to situations where invariant effects correspond to causal effects in SEMs where the intervention or perturbation does not have an effect on the hidden variables (35). Further explanation is beyond the scope of this paper.

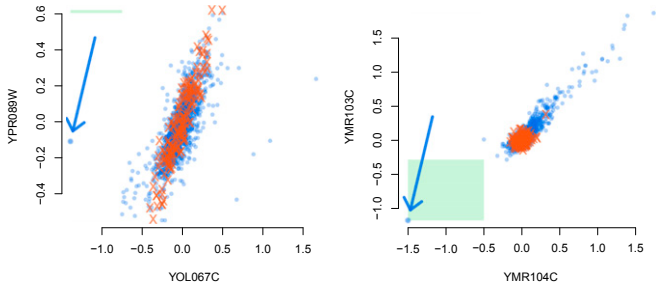

Fig. 1. The activities of two pairs of genes. Observational data are shown as red crosses and interventional data as blue circles. A blue arrow marks the experiment where an intervention on gene YOL067C occurs (activity shown on the *x* axis in the left panel) and analogously in the right panel. The intervention on the *Left* is deemed not "strongly successful" [not fulfilling (*i*) in the definition of an SIE] as the activity of gene YPR089W (*y* axis) under the intervention is well within its usual range. The intervention on the *Right* is called strongly successful as the activity of YMR103C (*y* axis) under the intervention is outside of the previously seen range and likewise for the gene YMR104C on which the intervention occurs (strongly successful area is marked by the green box).

Software

Comparing different approaches for detection of causal effects is in practice often cumbersome as they use very different implementations. To address this issue, we provide a software package CompareCausalNetworks (36) for the R language (37). It provides a unified interface to the following methods: GES [Greedy Equivalence Search (17)], CAM [Causal Additive Model (25)], Lingam [Linear, Non-Gaussian, Acyclic Models (13)], rfci (38) (really fast causal inference), and pc [PC-algorithm (3)], which are all classically applied to observational data only, as well as GIES [Greedy Interventional Equivalence Search (12)], which is making explicit use of the knowledge where interventions took place. The ICP methods are implemented as ICP and hiddenICP (allowing for hidden variables) in the R package InvariantCausalPrediction. They use the knowledge of the environment where an observation took place (for example, whether it is part of the observational or interventional data) but are not requiring knowledge about the precise nature of the interventions. As a further benchmark, we also implement cross-validated sparse regression as a method regression and offer the option of using stability selection (39) on all implemented methods.

Validation: Gene Perturbation Experiments

We consider large-scale gene deletion experiments in yeast (*Saccharomyces cerevisiae*) (40). Genome-wide mRNA expression levels are measured for 6,170 genes: 160 observational data points from wild-type individuals and 1,479 interventional data arising from single gene deletions (1,479 perturbation/deletion experiments where a single gene has been deleted from a strain). The goal is to predict the expression levels of all (except the deleted) genes of a new and unseen single gene deletion intervention. More precisely, denoting the expression levels of the genes by variables $X_1, \ldots, X_{p+1}$ with $p+1=6{,}170$, we want to predict whether $Y \in \{X_1, \ldots, X_{p+1}\} \backslash X_j$ significantly changes when deleting gene j (for each of the 6,170 different responses Y).

For this task, we aim to obtain a confidence statement from the ICP method for $\gamma_j^* \neq 0$ in [**4**], or to estimate γ_j^* using other methods, describing the direct causal effect of X_j on Y. We then use the strong direct effects [ranked according to the proportion of times the effect gets selected/top-ranked when running each procedure on 50 random subsamples of the data (39)] as a proxy for the strong total effect of X_j to Y, because we expect strong direct and strong total effects to be very similar to each other (*SI Appendix*).

We use the following separation into training and validation data. We divide the 1,479 interventional data into five blocks B_1, B_2, B_3, B_4, B_5. We use as training data all 160 observational data points and four blocks $B_r, (r \in T)$ with $T \subset \{1, \ldots, 5\}$ and $|T|=4$, of interventional data, and the validation data are the remaining block $B_t, (t \in \{1, \ldots, 5\} \backslash T)$. Therefore, we can predict the effects from the interventions from block B_t without having used these interventional data in the training set. By repeating the separation into training and validation data five times, each gene perturbation is held out once and we can use it to validate the predictions.

For the ICP method, we use a very simple labeling of experimental settings: $\mathcal{E}=\{1,2\}$, where $e=1$ corresponds to observational data and $e=2$ to all interventional data (regardless which gene has been targeted to be knocked down). As discussed above, when choosing a small set $\mathcal{E}$, we might pay a price in terms of statistical power to detect significant causal variables. On the other hand, due to potential off-target effects of interventions, pooling all interventions into one experimental setting is more robust against "noisy interventions" that potentially affect many genes. As significance level, we use a level of $\alpha=0.01$ corresponding to a probability greater or equal to 0.99 in the confidence statement in [**5**]. We repeat the experiment on subsamples of the data in the spirit of stability selection (39, 41) and rank the edges in order of decreasing selection frequency. [For the ICP method, there is a directed edge from gene j to k if: gene k is the response variable and gene j is an element of the selected causal variables $\hat{S}(\mathcal{E})$. For methods using directed graphical modeling, the meaning of "directed edge" is given by the corresponding DAG.] Going down this list, we check whether the prediction of a directed edge from gene j to gene k is "successful" in the sense of an SIE as defined before (where gene j corresponds to the X variable and gene k to the response variable Y). Note that the SIE is an estimate of the total causal effect. Among 9,125,430 possible edges, there are 10,757 interventions (about 0.1% of all edges) that we classify as strongly successful with this criterion.

Fig. 1 shows two examples. On the left is a pair of genes YOL067C,YPR089W that was selected due to the high correlation between the two genes on observational data. When intervening on YOL067C (the gene on the *x* axis), the activity of gene YPR089W is

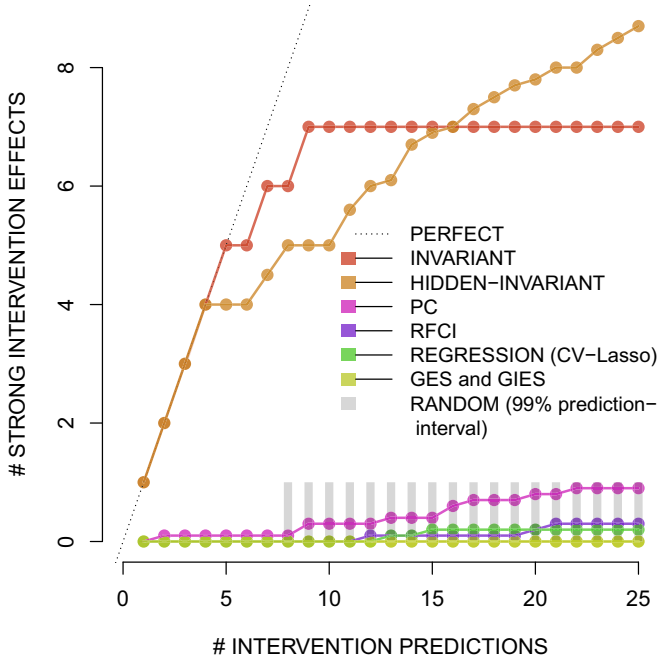

Fig. 2. The average number of successful interventions (*y* axis) against the number of selected edges (*x* axis) for various methods.

Table 3. Overlap between SIEs and scores from yeastgenome.org and TF bindings

	A	B	C	D	E	F	TF
No. gene pairs with both SIE and score	327	92	267	105	60	61	117
Expected no. under independence	128.2	44.9	86.2	20.8	10.0	12.1	9.5

still well within its usual range and the intervention is deemed not successful according to our criterion. The right side shows the pair YMR104C,YMR103C. The edge from the first to the second one of these genes is the most frequently selected edge by our method of ICP. It is selected because the model for observational and interventional data fits equally well to the data (and there is no other explanation using other genes that would achieve this). This is in contrast to the gene pair on the left, where the interventional data have a much broader variance around a regression fit than the observational data.

For the K most often selected edges, we can check how many of the corresponding interventions were strongly successful on the test data. Fig. 2 shows results for several methods, where gene pairs are ordered by the selection frequency for each method when fitting models on 100 random bootstrap samples of the data. In case of equal selection frequency, we compute the expected number of SIEs when breaking the ties randomly. The number of SIEs among the most frequently selected gene pairs can thus take a noninteger value in the presence of ties. As a crude benchmark, we can first look at the success of random guessing. If choosing randomly, we will with probability of at least 95% not predict a single SIE for $K < 40$ and the expected number of selected SIE with random guessing is just $0.029 \ll 1$ for $K = 25$. With ICP (allowing for hidden variables or not), the first four top edges correspond to an SIE.

As a comparison, we show results for GES, GIES, PC, RFCI, and a regression-based estimate, where we always choose the first 25 ranked candidate genes with stability selection. These methods are fitted with the default values once on just observational data and then also (as shown here) on both observational and interventional data, but the difference between the two is very small, with the results that use interventional data (the ones shown) slightly superior. All of the considered methods yield estimates of direct effects, but we validate with SIE which is estimating the total effect (justification in *SI Appendix*).

To give an impression of the computational complexity of the algorithms, Table 2 shows the runtime of used methods (with the same settings as used to produce the other results on the whole dataset) on a single core of a 2.8-GHz processor when estimating the causal graph for either the first 50, 500, or 5,000 genes. The runtimes reported here do not include the prescreening (*SI Appendix*), as this was a common preprocessing step for all methods.

One question is whether a mechanism is already known that links the top-scoring gene pairs. If so, this can be viewed as a validation of the SIEs one can see in the data. The question can be turned around, however. We can ask whether we could predict the SIEs just using biological background knowledge.

To this end, we extracted six scores that measure interactions between gene pairs from a bioinformatics source, the *Saccharomyces* Genome Database (SGD) based on ref. 42 at yeastmine.yeastgenome.org. This database contains over 3×10^6 gene interactions collected by the BioGRID public resource from the scientific literature. Interactions are categorized as either "physical" or "genetic," and a subset of each is labeled as manually curated. For all gene pairs, we queried the database for an interaction and interpret the "bait" to "hit" directionality of the interaction as a direct causal effect from bait to hit. [One has to keep in mind that this interpretation is questionable. The presence of an edge in the database indeed may suggest the presence of a direct causal relation, but the directionality of the edge in the database need not coincide with the directionality of the causal relation but could point in the opposite direction. Furthermore, the fact that two proteins bind (physical interaction) or that two genes have a nonlinear interaction on some phenotype (genetic interaction) does not yet imply that there is a causal relation between the gene expression levels.] This resulted in six binary scores for gene pairs. Scores A, B, and C use the following, respectively: (A) both physical and genetic interactions, (B) only physical interactions, and (C) only genetic interactions. Scores D, E, and F are similar but use only the subsets of manually curated interactions.

In addition, we used ChIP-on-chip data from ref. 43 as another source of indirect evidence for validation. The dataset contains binding activity of a subset of genes that function as TFs and regulate the expression levels of other genes. A binary score was constructed by matching the binding activity at 10^{-4} confidence level for 118 TFs with the SGD naming scheme, resulting in 8,073 nonzero entries.

Table 3 shows the number of gene pairs that have both an SIE and also a nonzero value in the scores from the yeastgenome.org and TF datasets among all of more than 9 million possible gene pairs. There are 10,679 gene pairs with an SIE, 8,073 with a score on TF, and the number of gene pairs with a score on A–F are given by 109,549, 38,377, 73,688, 17,778, 8,581, 10,324, respectively. Table 3 shows that the number of pairs that have both an SIE and score is substantially higher compared with a situation where these measures would be independent. Using Fisher's exact test, these associations are significant at level less than 10^{-9} for all scores. Still, the overlap is rather small, as just a few hundred out of all 10,679 gene pairs with an SIE have a nonvanishing score. We note that, although SIE measures the total effect and the six yeastgenome.org scores A–F and TF are more related to direct effects, some joint occurrences are interesting to look at; see also the discussion on validation in *Challenges and Validation*. [A strong direct effect is expected to result in a strong total effect (*SI Appendix*).]

We also show the 20 gene pairs that are most frequently selected with the ICP method in Table 4, starting with the most frequently selected pair "YMR104C → YMR103C." The columns show whether the pairs have an SIE and whether an interaction is indicated by the nonzero score A–F from yeastgenome.org dataset and by a nonzero score TF from the ChIP-on-chip data.

Out of the top 20 pairs from the ICP method, 7 show an SIE (measuring a total effect), as already seen in Fig. 2. Four pairs

Table 4. Top 20 stable results from ICP

Rank	Cause	Effect	SIE	A	B	C	D	E	F	TF
1	YMR104C	YMR103C	✓							
2	YPL273W	YMR321C	✓							
3	YCL040W	YCL042W	✓							
4	YLL019C	YLL020C	✓							
5	YMR186W	YPL240C	✓	✓	✓	✓	✓		✓	
6	YDR074W	YBR126C		✓	✓	✓	✓	✓	✓	
7	YMR173W	YMR173W-A	✓							
8	YGR162W	YGR264C								
9	YOR027W	YJL077C	✓							
10	YJL115W	YLR170C								
11	YOR153W	YDR011W		✓	✓					
12	YLR270W	YLR345W								
13	YOR153W	YBL005W								
14	YJL141C	YNR007C								
15	YAL059W	YPL211W								
16	YLR263W	YKL098W								
17	YGR271C-A	YDR339C								
18	YLL019C	YGR130C								
19	YCL040W	YML100W								
20	YMR310C	YOR224C								

correspond to an indication of an interaction based on the scores A–F derived from the yeastgenome.org database. Just a single pair shows both an SIE and a positive interaction score. Even though the overlap is small, both results on their own are significant, if compared with random sampling of gene pairs, as shown in Table 5. As the causes from these top 20 pairs include no TFs found in the TF dataset, there is no overlap with the corresponding score.

Application to Flow Cytometry Data

As another example of the possible applications of invariant prediction, we consider the flow cytometry data of ref. 44. The abundance of 11 biochemical agents is measured in several different environments. One of the environments can be considered as observational data without external interventions. In the other environments, different reagents have been added (or stimulants have been removed) that modify the behavior of some of the biochemical agents (see refs. 44 and 9 for a more detailed description of the experiments). [The description in ref. 9 is not entirely accurate: conditions 7 and 8 (activation by PMA and β2cAMP) are abundance interventions in combination with a global intervention, as the α-CD3 and α-CD28 activators were not applied in those conditions (in contrast with all other conditions, including the baseline).] Each environment contains between 700 and 1,000 samples. Here, we only use a subset of eight environments, the same ones that ref. 9 used, to allow comparison of the results.

A naive implementation of invariant prediction would try to find an invariant regression model for each of the 11 agents in turn, where invariance is measured across all eight environments simultaneously. The ICP method does not allow interventions on the target variable itself (see the discussion after [**5**] and also *SI Appendix*). We thus follow a slightly adapted strategy by comparing all 28 possible pairs of environments. For each pair of environments, we estimate the set of causal predictors using ICP. Taking the union of all of the estimated parental sets across the 28 pairs of environments leads to the graph of edges shown in Fig. 3 when controlling the overall familywise error rate at 10^{-3}. We give a more detailed description of the application of the ICP method to the flow cytometry data in the *SI Appendix*.

Allowing for hidden variables with hiddenICP allows correlation between the noise input at different variables. Under a shift mechanism for the interventions described in more detail in ref. 35, the presence of feedback loops will not invalidate the output of hiddenICP (for details, see refs. 31 and 35). Some two-cycles can be seen to be selected by the method in Fig. 3. Of the 15 edges found by invariant prediction (with or without hidden variables), 12 have been previously reported in the literature. Each of the three previously unreported edges adds a reverse link to a previously reported edge. In general, one would expect feedback loops to be present in this system. However, there seems to be no consensus yet on the exact nature of these feedback loops. Therefore, we should consider what Sachs et al. (44) call the "consensus" network to be an incomplete description of a more complicated biological reality. *SI Appendix*, Table S1, gives a list of all of the edges that have been found by different causal discovery methods. Our estimated network in Fig. 3 also confirms quite well with the point estimate in ref. 35, which allows for interventions on the target variables but cannot produce confidence intervals and significance testing.

Table 5. Significance of top 20 ICP results

Score	Strong effects among top 20 ICP results	Expected no. under random guessing	*P* value
SIE	7	≤ 0.03	$\leq 10^{-18}$
A	3	≤ 0.3	$\leq 2 \cdot 10^{-3}$
B	3	≤ 0.09	$\leq 8 \cdot 10^{-5}$
C	2	≤ 0.2	$\leq 2 \cdot 10^{-2}$
D	2	≤ 0.04	$\leq 8 \cdot 10^{-4}$
E	1	≤ 0.02	$\leq 2 \cdot 10^{-2}$
F	2	≤ 0.03	$\leq 3 \cdot 10^{-4}$
TF	0	≤ 0.02	1

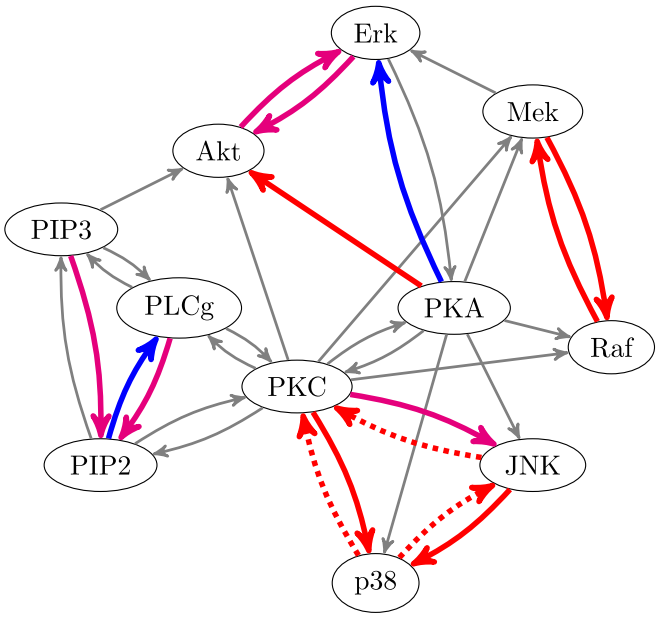

Fig. 3. The graph of estimated causal relations between the biochemical agents in the ref. 44 data. Blue edges are found by an invariant prediction approach, whereas red edges are found if allowing hidden variables and feedback with invariant prediction. Purple edges are found with invariant prediction whether allowing for hidden variables or not. The solid edges (including the gray edges) are all relations that have been reported in either the consensus network according to ref. 44, or the newly reported edges in ref. 44, 9, or 8. See also *SI Appendix*, Table S1.

Conclusions

The recently developed ICP method (31) for causal inference is equipped with confidence bounds for inferential statements, without the need of prespecifying whether causal effects are identifiable or not. A notable feature of the approach is that with increased heterogeneity, in terms of more experimental settings, it automatically achieves better identifiability. The underlying invariance principle (invariance assumption) can be used if data from different experimental conditions (such as observational–interventional data) are available. We provide open-source software in the R language (36), which enables an easy use of the ICP method and comparing it to some other, well-known causal inference algorithms.

We validate the statistical ICP method for *Saccharomyces cerevisiae*. We consider new interventional gene deletion experiments that have not been used for training the method, and we also look at additional information from the SGD at yeastgenome.org and from TF binding based on ChIP-on-chip data. The validation itself has to be set up carefully to avoid validating spurious effects: we propose here the notion of an SIE. To increase the range of validation to other applications and datasets, we also considered flow cytometry data: the validation is on less rigorous grounds without hold-out intervention experiments and SIE, but it nevertheless allows to compare with existing results in the literature. The best validation is, in our opinion, successful prediction of the effects of previously unseen interventions, as demonstrated here for the ICP method and gene knockout data.

ACKNOWLEDGMENTS. We thank Patrick Kemmeren for generously providing the gene perturbation data. J.M.M. and P.V. were supported by The Netherlands Organization for Scientific Research (VIDI Grant 639.072.410).

1. Stekhoven DJ, et al. (2012) Causal stability ranking. *Bioinformatics* 28(21):2819–2823.
2. Pearl J (2000) *Causality: Models, Reasoning and Inference* (Cambridge Univ Press, New York).
3. Spirtes P, Glymour C, Scheines R (2000) *Causation, Prediction, and Search* (MIT Press, Cambridge, MA), 2nd Ed.
4. Bollen KA (1989) *Structural Equations with Latent Variables* (Wiley, New York).
5. Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Math Model* 7(9):1393–1512.
6. Markowetz F, Grossmann S, Spang R (2005) Probabilistic soft interventions in conditional Gaussian networks. *Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)* (Society for Artificial Intelligence and Statistics, NJ), pp 214–221.
7. Tian J, Pearl J (2001) Causal discovery from changes. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI)* (Morgan Kaufmann, San Francisco), pp 512–521.
8. Eaton D, Murphy K (2007) Exact Bayesian structure learning from uncertain interventions. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)* 2:107–114.
9. Mooij JM, Heskes T (2013) Cyclic causal discovery from continuous equilibrium data. *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI)* (AUAI Press, Corvallis, OR), pp 431–439.
10. Meek C (1995) Causal inference and causal explanation with background knowledge. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI)* (Morgan Kaufmann, San Francisco), pp 403–418.
11. Cooper G, Yoo C (1999) Causal discovery from a mixture of experimental and observational data. *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI)* (Morgan Kaufmann, San Francisco), pp 116–125.
12. Hauser A, Bühlmann P (2012) Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J Mach Learn Res* 13:2409–2464.
13. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen AJ (2006) A linear non-Gaussian acyclic model for causal discovery. *J Mach Learn Res* 7:2003–2030.
14. Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2009) Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems 21 (NIPS)* (Curran Associates, Red Hook, NY), pp 689–696.
15. Peters J, Mooij JM, Janzing D, Schölkopf B (2014) Causal discovery with continuous additive noise models. *J Mach Learn Res* 15:2009–2053.
16. Peters J, Bühlmann P (2014) Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* 101:219–228.
17. Chickering DM (2002) Optimal structure identification with greedy search. *J Mach Learn Res* 3:507–554.
18. Harris N, Drton M (2013) PC algorithm for nonparanormal graphical models. *J Mach Learn Res* 14:3365–3383.
19. Maathuis MH, Kalisch M, Bühlmann P (2009) Estimating high-dimensional intervention effects from observational data. *Ann Stat* 37:3133–3164.
20. Hauser A, Bühlmann P (2015) Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J R Stat Soc B* 77:291–318.
21. Heckerman D, Geiger D, Chickering DM (1995) Learning Bayesian networks: The combination of knowledge and statistical data. *Mach Learn* 20:197–243.
22. Koller D, Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA).
23. Kalisch M, Bühlmann P (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J Mach Learn Res* 8:613–636.
24. van de Geer S, Bühlmann P (2013) ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann Stat* 41:536–567.
25. Bühlmann P, Peters J, Ernest J (2014) CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann Stat* 42:2526–2556.
26. Ernest J, Bühlmann P (2015) Marginal integration for nonparametric causal inference. *Electron J Stat* 9:3155–3194.
27. Uhler C, Raskutti G, Bühlmann P, Yu B (2013) Geometry of the faithfulness assumption in causal inference. *Ann Stat* 41:436–463.
28. Mooij JM, Cremers J (2015) An empirical study of one of the simplest causal prediction algorithms. *Proceedings of the UAI 2015 Workshop on Advances in Causal Inference, CEUR Workshop Proceedings* (CEUR-WS.org, Aachen, Germany), Vol 1504, pp 30–39.
29. Bühlmann P, Rütimann P, Kalisch M (2013) Controlling false positive selections in high-dimensional regression and causal inference. *Stat Methods Med Res* 22(5): 466–492.
30. Maathuis MH, Colombo D, Kalisch M, Bühlmann P (2010) Predicting causal effects in large-scale systems from observational data. *Nat Methods* 7(4):247–248.
31. Peters J, Bühlmann P, Meinshausen N (2016) Causal inference using invariant prediction: Identification and confidence intervals (with discussion). arXiv:1501.01332.
32. Aldrich J (1989) Autonomy. *Oxf Econ Pap* 41:15–34.
33. Schölkopf B, et al. (2012) On causal and anticausal learning. *Proceedings of the 29th International Conference on Machine Learning (ICML)* (Omnipress, New York), pp 1255–1262.
34. Bareinboim E, Pearl J (2014) Transportability from multiple environments with limited experiments: Completeness results. *Advances in Neural Information Processing Systems 27 (NIPS)* (Curran Associates, Red Hook, NY), pp 280–288.
35. Rothenhäusler D, Heinze C, Peters J, Meinshausen N (2015) Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems 28 (NIPS)* (Curran Associates, Red Hook, NY), pp 1513–1521.
36. Heinze C, Meinshausen N (2015) CompareCausalNetworks. R package, version 0.1.1. Available at https://cran.r-project.org/web/packages/CompareCausalNetworks/CompareCausalNetworks.pdf. Accessed April 29, 2016.
37. R Development Core Team (2005) *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
38. Colombo D, Maathuis MH, Kalisch M, Richardson TS (2012) Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann Stat* 40:294–321.
39. Meinshausen N, Bühlmann P (2010) Stability selection (with discussion). *J R Stat Soc B* 72:417–473.
40. Kemmeren P, et al. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* 157(3):740–752.
41. Breiman L (1996) Heuristics of instability and stabilization in model selection. *Ann Stat* 24:2350–2383.
42. Cherry JM, et al. (2012) *Saccharomyces* Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res* 40(Database issue):D700–D705.
43. MacIsaac KD, et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7:113.
44. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721): 523–529.