Marginal integration for fully robust causal inference

Jan Ernest∗ and Peter Bühlmann

e-mail: ernest@stat.math.ethz.ch; buhlmann@stat.math.ethz.ch

Abstract: We consider the problem of inferring the total causal effect of a single variable intervention on a (response) variable of interest. We propose a certain marginal integration regression technique for a very general class of potentially nonlinear structural equation models (SEMs) with known structure, or at least known superset of adjustment variables: we call the procedure $S$-mint regression. We prove that it achieves the convergence rate as for nonparametric regression: for example, single variable intervention effects can be estimated with convergence rate $n^{-2/5}$ assuming smoothness with twice differentiable functions. Our result can also be seen as a major robustness property with respect to model misspecification which goes much beyond the notion of double robustness. Furthermore, when the structure of the SEM is not known, we can estimate (the equivalence class of) the directed acyclic graph corresponding to the SEM, and then proceed by using $S$-mint based on these estimates. We empirically compare the $S$-mint regression method with more classical approaches and argue that the former is indeed more robust, more reliable and substantially simpler.

1. Introduction

Understanding cause-effect relationships between variables is of great interest in many fields of science. An ambitious but highly desirable goal is to infer causal effects from observational data obtained by observing a system of interest without subjecting it to interventions. This would allow to circumvent potentially severe experimental constraints or to substantially lower experimental costs. The words “causal inference” (usually) refer to the problem of inferring effects which are due to (or caused by) interventions: if we make an outside intervention at a variable $X$, say, what is its effect on another response variable of interest $Y$. We describe some examples in Section 1.3. It is well known that “association is not equal to causation”. Hence, the tools for inferring causal effects are different from regression methods, but as we will argue, the regression methods when properly applied remain a useful tool for causal inference. Various fields and concepts have contributed to the understanding and quantification of causal inference: the framework of potential outcomes and counterfactuals (Rubin, 2005, cf.), see also Dawid (2000), structural equation modeling (Bollen, 1998, cf.), and graphical modeling (Lauritzen and Spiegelhalter, 1988; Greenland et al., 1999, cf.); the book by Pearl (2000) provides a nice overview.

∗Partially supported by the Swiss National Science Foundation grant no. 20PA20E-134493.

1More generally, in the presence of both, interventional and observational data, the goal is to infer intervention or causal effects among variables which are not directly targeted by the interventions from interventional data.
We consider aspects of the problem indicated above, namely inferring intervention or causal effects from observational data without external interventions. Thus, we deal (in part) with the question how to infer causal effects without relying on randomized experiments or randomized studies. Besides fundamental conceptual aspects, as treated for example in the books by Pearl (2000), Spirtes et al. (2000) and Koller and Friedman (2009), important issues include statistical tasks such as estimation accuracy and robustness with respect to model misspecification. This paper is focusing on the two latter topics, covering also high-dimensional sparse settings with many variables (parameters) but relatively few observational data points. We make use of a marginal integration regression method which has been proposed for additive regression modeling (Linton and Nielsen, 1995). Its use in causal inference is novel and our main result (Theorem 1) establishes optimal convergence properties and justifies the method as a fully robust procedure against model misspecification, as explained further in Section 1.2.

1.1. Basic concepts and definitions for causal inference

We very briefly introduce some of the basic concepts for causal inference (and the reader who is familiar with them can skip this subsection). We consider \( p \) random variables \( X_1, \ldots, X_p \), where one of them is a response variable \( Y \) of interest and one of them an intervention variable \( X \), that is, the variable where we make an external intervention by setting \( X \) to a certain value \( x \). Such an intervention is denoted by Pearl’s do-operator \( \text{do}(X = x) \) (Pearl, 2000, cf.). We denote the indices corresponding to \( Y \) and \( X \) by \( j_Y \) and \( j_X \), respectively: thus, \( Y = X_{j_Y} \) and \( X = X_{j_X} \). We assume a setting where all relevant variables are observed, i.e., there are no relevant hidden variables.

The system of variables is assumed to be generated from a structural equation model (SEM):

\[
X_j \leftarrow f_j(X_{\text{pa}(j)}, \varepsilon_j), \quad j = 1, \ldots, p. \tag{1}
\]

Thereby, \( \varepsilon_1, \ldots, \varepsilon_p \) are independent noise (or innovation) variables, and there is an underlying structure in terms of a directed acyclic graph (DAG) \( D \), where each node \( j \) corresponds to the random variable \( X_j \): We denote by \( \text{pa}(j) = \text{pa}_D(j) \) the set of parents of node \( j \) in the underlying DAG \( D \),\(^3\) and \( f_j(\cdot) \) are assumed to be real-valued (measurable) functions. For any index set \( U \subseteq \{1, \ldots, p\} \) we write \( X_U := (X_v)_{v \in U} \), for example \( X_{\text{pa}(j)} = (X_v)_{v \in \text{pa}(j)} \).

The causal mechanism we are interested in is the total effect of an intervention at a single variable \( X \) on a response variable \( Y \) of interest.\(^4\) The distribution of \( Y \) when doing an external intervention \( \text{do}(X = x) \) by setting variable \( X \) to \( x \) is denoted with its density (assumed to exist) or discrete probability function

---

\(^2\)It suffices to assume that \( Y, X \) and \( X_{\text{pa}(j_X)} \) (the parents of \( X \)) are observed, see (3).

\(^3\)The set of parents is \( \text{pa}(j) = \{k; \text{there exists a directed edge } k \to j \text{ in DAG } D\} \).

\(^4\)A total effect is the effect of an intervention at a variable \( X \) to another variable \( Y \), taking into account the total of all (directed) paths from \( X \) to \( Y \).
by $p(y|\text{do}(X = x))$. The mathematical definition of $p(y|\text{do}(X = x))$ can be given in terms of a so-called truncated Markov factorization or maybe more intuitively, by direct plug-in of the intervention value $x$ for variable $X$ and propagating this intervention value $x$ to all other random variables including $Y$ in the structural equation model (1); precise definitions are given in e.g. Pearl (2000) or Spirtes et al. (2000). The underlying important assumption in the definition of $p(y|\text{do}(X = x))$ is that the functional forms and error distributions of the structural equations for all the variables $X_j$ which are different from $X$ do not change when making an intervention at $X$.

A very powerful representation of the intervention distribution is given by the well-known backdoor adjustment formula. $^5$ We say that a path in a DAG $D$ is blocked by a set of nodes $S$ if and only if it contains a chain $.. \rightarrow m \rightarrow ..$ or a fork $.. \leftarrow m \rightarrow ..$ with $m \in S$ or a collider $.. \rightarrow m \leftarrow ..$ such that $m \notin S$ and no descendant of $m$ is in $S$. Furthermore, a set of variables $S$ is said to satisfy the backdoor criterion relative to $(X,Y)$ if no node in $S$ is a descendant of $X$ and if $S$ blocks every path between $X$ and $Y$ with an arrow pointing into $X$. For a set $S$ that satisfies the backdoor criterion relative to $(X,Y)$, the backdoor adjustment formula now reads:

$$p(y|\text{do}(X = x)) = \int p(y|X = x, X_{pa(jX)})dP(X_{pa(jX)}),$$

(2)

where $p(\cdot)$ and $P(\cdot)$ are generic notations for the density or distribution (Pearl, 2000, Theorem 3.3.2). An important special case of the backdoor adjustment formula is obtained when considering the adjustment set $S = pa(jX)$: if $jY \notin pa(jX)$, that is, $Y$ is not in the parental set of the variable $X$, then:

$$p(y|\text{do}(X = x)) = \int p(y|X = x, X_{pa(jX)})dP(X_{pa(jX)}).$$

(3)

Thus, if the parental set $pa(jX)$ is known, the intervention distribution can be calculated from the standard observational conditional and marginal distributions. Our main focus is the expectation of $Y$ when doing the intervention $\text{do}(X = x)$, the so-called total effect:

$$E[Y|\text{do}(X = x)] = \int y p(y|\text{do}(X = x))dy.$$ 

A general and often used route for inferring $E[Y|\text{do}(X = x)]$ is as follows: the directed acyclic graph (DAG) corresponding to the structural equation model (SEM) is either known or (its Markov equivalence class) estimated from data; building on this, one can estimate the functions in the SEM (edge functions in the DAG), the error distributions in the SEM, and finally extract an estimate of $E[Y|\text{do}(X = x)]$, or bounds of this quantity if the DAG is not identifiable from the observational distribution. See for example Spirtes et al. (2000); Pearl (2000); Maathuis et al. (2009); Spirtes (2010).

$^5$For a simple version of the formula, skip the text until the second line after formula (2).
1.2. Our contribution

The new results from this paper should be explained for two different scenarios and application areas: one where the structure of the DAG $D$ in the SEM is known, and the other where the structure and the DAG $D$ are unknown and estimated from data. Of course, the second setting is linked to the first by treating the estimated as the true known structure. However, due to estimation errors, a separate discussion is in place.

1.2.1. Structural equation models with known structure

We consider a general SEM as in (1) with known structure in form of a DAG $D$ but unknown functions $f_j$ and unknown error distributions for $\epsilon_j$. As already mentioned before, our focus is on inferring the total effect

$$E[Y|\text{do}(X = x)] = \int y \, p(y|\text{do}(X = x)) \, dy,$$

where $p(y|\text{do}(X = x))$ is the interventional density (or discrete probability function) of $Y$ as loosely described in Section 1.1.

The first approach to infer the total effect in (4) is to estimate the functions $f_j$ and error distributions for $\epsilon_j$, and it is then possible to calculate $E[Y|\text{do}(X = x)]$, typically using some path based method based on the DAG $D$ (see also Section 3.1). This route is essentially impossible without putting further assumptions on the functional form of $f_j$ in the SEM (1). For example, one often makes the assumption of additive errors, and if the cardinality of the parental set $|\text{pa}(j)|$ is large, some additional constraints like additivity of a non-parametric function is in place to avoid the curse of dimensionality. Thus, by keeping the general possibly non-additive structure of the functions $f_j$ in the SEM, we have to reject this approach.

The second approach for inferring the total effect in (4) relies on the powerful backdoor adjustment formula in (2). At first sight, the problem seems ill-posed because of the appearance of $E[Y|X = x, X_S]$ for a set $S$ with possibly large cardinality $|S|$. But since we integrate over the variables $X_S$ in (2), we are not entering the curse of dimensionality. This simple observation is a key idea of this paper. We present an estimation technique for $E[Y|\text{do}(X = x)]$, or other functionals of $p(y|\text{do}(X = x))$, using marginal integration which has been proposed and analyzed for additive regression modeling (Linton and Nielsen, 1995). The idea of marginal integration is to first estimate a fully nonparametric regression of $Y$ versus $X$ and the variables $X_S$ from a valid adjustment set satisfying the backdoor criterion (for example the parents of $j_X$ or a superset thereof) and then average the obtained estimate over the variables $X_S$. We call the procedure “S-mint” standing for marginal integration with adjustment set $S$. Our main result in Theorem 1 establishes that $E[Y|\text{do}(X = x)]$ can be inferred via marginal integration with the same rate of convergence as for one-dimensional nonparametric function estimation for a very large class of structural equation...
models with potentially non-additive functional forms in the equations. Therefore, we achieve a major robustness result against model misspecification, as we assume only some standard smoothness assumptions but no further conditions on the functional form or nonlinearity of the functions $f_j$ in the SEM, not even requiring additive errors. Our main result (Theorem 1) also applies using a superset of the true underlying DAG $D$ (i.e. there might be additional directed edges in the superset), see Section 2.3. For example, such a superset could arise from knowing the order of the variables (e.g. in a time series context), or an approximate superset might be available from estimation of the DAG where one wouldn’t care too much about slight or moderate overfitting.

Inferring $\mathbb{E}[Y|\text{do}(X = x)]$ under model-misspecification is the theme of double robustness in causal inference (van der Laan and Robins, 2003, cf.). There, misspecification of either the regression or the propensity score model is allowed but at least one of them has to be correct: the terminology “double robustness” is intended to reflect this kind of robustness. In contrast to double robustness, we achieve here “full robustness” where essentially any form of “misspecification” is allowed, in the sense that $S$-mint does not require any specification of the functional form of the structural equations in the SEM. More details are given in Section 2.4.

The local nature of parental sets. Our $S$-mint procedure requires the specification of a valid adjustment set $S$: as described in (3), we can always use the parental set $\text{pa}(j_X)$ if $j_Y \notin \text{pa}(j_X)$. The parental variables are often an interesting choice for an adjustment set which corresponds to a local operation. From a computational viewpoint, determining the parental set is very efficient (see Section 4). Furthermore, as discussed below, the local nature of the parental sets can be very beneficial in presence of only approximate knowledge of the true underlying DAG $D$.

1.2.2. Structural equation models with unknown structure

Consider the SEM (1), but now we assume that the DAG $D$ is unknown. For this setting, we propose a two-stage scheme (“est $S$-mint”, Section 3.5). First, we estimate the structure of the DAG (or the Markov equivalence class of DAGs) or the order of the variables from observational data. To do this, all of the current approaches make further assumptions for the SEM in (1). See for example Chickering (2002); Teyssier and Koller (2005); Shimizu et al. (2006); Kalisch and Bühlmann (2007); Schmidt et al. (2007); Hoyer et al. (2009); Shojaie and Michailidis (2010); Bühlmann et al. (2014).

We can then infer $\mathbb{E}[Y|\text{do}(X = x)]$ as before with $S$-mint model fitting, but based on an estimated (instead of the true) adjustment set $S$. This seems often more advisable than using the estimated functions in the SEM, which are readily available from structure estimation, and pursuing a path based method with the estimated DAG. Since estimation of (the Markov equivalence class of) the DAG or of the order of variables is often very difficult and with limited accuracy for finite sample size, the second stage with $S$-mint model fitting seems fairly robust
with respect to errors in order- or structure-estimation and model misspecification, as suggested by our empirical results in Section 5.3. Therefore, such a two-stage procedure with structure- or order-search\(^6\) and subsequent marginal integration leads to reasonably accurate results. We only have empirical results to support this accuracy statement.

As mentioned above, due to their local nature, the parental sets (or subsets thereof) are often a very good choice in presence of estimation errors for inferring the true DAG (or equivalence class thereof): instead of assuming high accuracy for recovering the entire (equivalence class of the) DAG, we only need to have a reasonably accurate estimate of the much smaller and local parental set. Section 5 reports that such a two-stage approach, with S-mint modeling in the second stage, can outperform the direct CAM method (Bühlmann et al., 2014) (which is based on, or assuming, an additive SEM), at least if the sample size is small or moderate only.\(^7\)

### 1.3. The scope of possible applications

Genetic network inference is a prominent example where causal inference methods are used; mainly for estimating an underlying network in terms of a directed graph (Smith et al., 2002; Husmeier, 2003; Friedman, 2004; Yu et al., 2004, cf.). The goal is very ambitious, namely to recover relevant edges in a complex network from observational or a few interventional data. This paper does not address this issue: instead of recovering a network (structure), inferring total causal or intervention effects from observational data is a different, maybe more realistic but still very challenging goal in its full generality. Yet making some progress can be very useful in many areas of applications, notably for prioritizing and designing future randomized experiments which have a large total effect on a response variable of interest, ranging from molecular biology and bioinformatics (Editorial Nat. Methods, 2010) to many other fields including economy, medicine or social sciences. Such model-driven prioritization for gene intervention experiments in molecular biology has been experimentally validated with some success (Maathuis et al., 2010; Stekhoven et al., 2012).

We will discuss an application from molecular biology on a rather “toy-like” level in Section 6. Despite all simplifying considerations, however, we believe that it indicates a broader scope of possible applications. When having approximate knowledge of the parental set of the variables in a potentially large-scale system, one wouldn’t need to worry much about the underlying form of the dependences of (or structural equations linking) the variables: for quantifying the effect of single variable interventions, a specific marginal integration estimator converges with the univariate rate, as stated in (the main result) Theorem 1.

\(^6\)We do not make use of e.g. estimated edge functions, even if they were implicitly estimated for structure-search, as e.g. in Chickering (2002).

\(^7\)The better performance can be seen particularly well if the true SEM is non-additive, but often also for correctly specified models but with incorrectly estimated structure of the DAG due to finite (small or moderate) sample size.
Quantifying single variable interventions from observational data is indeed a useful first step. Further work is needed to address the following issues: (i) inference in settings with additional hidden, unobserved variables (Spirtes et al., 2000; Zhang, 2008; Shpitser et al., 2011; Colombo et al., 2012, cf.); (ii) inference based on both observational and interventional data (He and Geng., 2008; Hauser and Bühlmann, 2012, 2014, 2015, cf.); and finally (iii) developing sound tools and methods towards more confirmatory conclusions. The appropriate modifications and further developments of our new results (mainly Theorem 1) towards these points (i)-(iii) are not straightforward. In view of this, and due to the ambitious goal to draw causal conclusions, all obtained results in applications should be interpreted with care – but we believe that even limited progress within a proper framework of causal inference often leads to better results than sticking to the conceptually wrong framework of marginal correlations or associations from regression.

2. Causal effects for general nonlinear systems via backdoor adjustment: marginal integration suffices

We present here the, maybe surprising, result that marginal integration allows us to infer the causal effect of a single variable intervention with a convergence rate as for one-dimensional nonparametric function estimation in essentially any nonlinear structural equation model.

We assume a structural equation model (as already introduced in Section 1.1)

\[ X_j \leftarrow f^0_j(\text{pa}(j), \varepsilon_j), \quad j = 1, \ldots, p, \]

where \( \varepsilon_1, \ldots, \varepsilon_p \) are independent noise (or innovation) variables, \( \text{pa}(j) \) denotes the set of parents of node \( j \) in the underlying DAG \( D^0 \), and \( f_j^0(\cdot) \) are real-valued (measurable) functions. We emphasize the true underlying quantities with a superscript “\( 0 \)”. We assume in this section that the DAG \( D^0 \), or at least a (super-) DAG \( D^0_{\text{super}} \) which contains \( D^0 \) (see Section 2.3), is known. As mentioned earlier, our goal is to give a representation of the expected value of the intervention distribution \( \mathbb{E}[Y|\text{do}(X = x)] \) for some variables \( Y, X \in \{X_1, \ldots, X_p\} \). That is, \( Y \) and \( X \) are two variables of interest where we do an intervention at \( X \) and want to study its effect on \( Y \). Let \( S \) be a set of variables satisfying the backdoor criterion relative to \( (X,Y) \), see Section 1.1. We repeat to write the backdoor adjustment formula (2)

\[ p(y|\text{do}(X = x)) = \int p(y|X = x, X_S)dP(X_S), \]

where \( p(\cdot) \) and \( P(\cdot) \) are generic notations for the density or distribution. Assuming that we can interchange the order of integration (cf. part 6. of Assumption 1) we obtain

\[ \mathbb{E}[Y|\text{do}(X = x)] = \int \mathbb{E}[Y|X = x, X_S]dP(X_S). \]
This is a function depending on the one-dimensional variable $x$ only and therefore, intuitively, its estimation shouldn’t be much exposed to the curse of dimensionality. We will argue below that this is indeed the case.

### 2.1. Marginal integration

Marginal integration is an estimation method which has been primarily designed for additive and structured regression fitting (Linton and Nielsen, 1995). Without any modifications though, it is also suitable for estimation of $E[Y|\text{do}(X = x)]$ in (6).

Let $S$ be a set of variables satisfying the backdoor criterion relative to $(X,Y)$ (cf. Section 1.1) and denote by $s$ the cardinality of $S$. We use a nonparametric kernel estimator of the multivariate regression function $m(x,x_S) = E[Y|X = x,X_S = x_S]$ minimizing

$$
\sum_{i=1}^{n} (Y_i - \alpha - \beta(X^{(i)} - x))^2 K_{h_1}(X^{(i)} - x)L_{h_2}(X_S^{(i)} - x_S),
$$

where $K$ and $L$ are two kernel functions and $h_1, h_2$ the respective bandwidths, i.e.,

$$
K_{h_1}(t) = \frac{1}{h_1} K \left( \frac{t}{h_1} \right), \quad L_{h_2}(t) = \frac{1}{h_2^2} L \left( \frac{t}{h_2} \right).
$$

We obtain the partial local linear estimator at $(x,x_S)$ as $\hat{m}(x,x_S) = \hat{\alpha}$. We then integrate over the variables $x_S$ with the empirical mean and obtain:

$$
\hat{E}[Y|\text{do}(X = x)] = n^{-1} \sum_{k=1}^{n} \hat{m}(x,X_S^{(k)})
$$

This is a locally weighted average, with localization through the one-dimensional variable $x$. For our main theoretical result to hold, we assume the following.

**Assumption 1.**

1. The variables $X_S$ have a bounded support $\text{supp}(X_S)$.
2. The regression function $m(u,u_S) = E[Y|X = u, X_S = u_S]$ exists and has bounded partial derivatives up to order 2 with respect to $u$ and up to order $d$ with respect to $u_S$ for $u$ in a neighborhood of $x$ and $u_S \in \text{supp}(X_S)$.
3. The variables $X, X_S$ have a density $p(.,.)$ with respect to Lebesgue measure and $p(u,u_S)$ has bounded partial derivatives up to order 2 with respect to $u$ and up to order $d$ with respect to $u_S$. In addition, it holds that

$$
\inf_{u \in x \pm \delta, x_S \in \text{supp}(X_S)} p(u,x_S) > 0 \text{ for some } \delta > 0.
$$

4. The kernel functions $K$ and $L$ are symmetric with bounded supports and $L$ is an order-$d$ kernel.
For \( \varepsilon = Y - \mathbb{E}[Y|X, X_S] \), it holds that \( \mathbb{E}[\varepsilon^4] \) is finite and \( \mathbb{E}[\varepsilon^2|X = x, X_S = x_S] \) is continuous. Furthermore, for a \( \delta > 0 \), \( \mathbb{E}[|\varepsilon^{2+\delta}| |X = u] \) is bounded for \( u \) in a neighborhood of \( x \).

There exists \( c < \infty \) such that \( \mathbb{E}[|Y||X = x, X_S = x_S] \leq c \) for all \( x_S \).

Note that part 6. of Assumption 1 is only needed for interchanging the order of integration in (6). Due to the bounded support of the variables \( X_S \) it is not overly restrictive.

As a consequence, the following result from Fan et al. (1998) establishes a convergence rate for the estimator as for one-dimensional nonparametric function estimation.

Theorem 1. Suppose that Assumption 1 holds for a set \( S \) satisfying the backdoor criterion relative to \((X, Y)\) in the DAG \( D^0 \) from model (5), and suppose that \( s < d \). Then, if the bandwidths are chosen such that they satisfy \( nh_1h_2^2 / \log^2(n) \to \infty \) and \( h_1, h_2 \to 0 \) such that \( h_2^d / h_1 \to 0 \), we have that

\[
\mathbb{E}[Y|\text{do}(X = x)] - \mathbb{E}[Y|\text{do}(X = x)] = O(h_1^2) + O_P(1 / \sqrt{nh_1}).
\]

Proof. The statement follows from Fan et al. (1998, Theorem 1 and Remark 3).

Theorem 1 establishes the convergence rate \( O(n^{-2/5}) \), choosing \( h_1 \approx n^{-1/5} \); this matches the optimal rate for estimation of one-dimensional smooth functions having second derivatives, and such a smoothness with respect to the variable \( x \) is assumed in part 2. of Assumption 1. Thus, the implication is the important robustness fact that for any potentially nonlinear structural equation model satisfying the regularity conditions in Theorem 1, we can estimate the expected value of the intervention distribution as well as in nonparametric estimation of a smooth function with one-dimensional argument. We note, as mentioned already in Section 1.2.1, that it would be essentially impossible to estimate the functions \( f_j \) in (1) in full generality; interestingly, when focusing on inferring the total effect \( \mathbb{E}[Y|\text{do}(X = x)] \), the problem is much better posed as demonstrated with our concrete \( S\text{-mint} \) procedure. Furthermore, with the (valid) choice \( S = \text{pa}(j_X) \) or an (estimated) superset thereof, one obtains a procedure that is only based on local information in the graph: this turns out to be advantageous, see also Section 1.2.1, particularly when the underlying DAG structure is not correctly specified (cf. Section 5.3). We will report about the performance of such an \( S\text{-mint} \) estimation method in Sections 4 and 5. Note that the rate of Theorem 1 remains valid (for a slightly modified estimator) if we allow for discrete variables in the parental set of \( X \) (cf. Fan et al. (1998)).

It is worthwhile to point out that \( S\text{-mint} \) becomes more challenging for inferring multiple variable interventions such as \( \mathbb{E}[Y|\text{do}(X_1 = x_1, X_2 = x_2)] \); the convergence rate is then of the order \( n^{-1/3} \) for a twice differentiable regression function.

Remark 1. Theorem 1 generalizes to real-valued transformations \( t(\cdot) \) of \( Y \). By using the argument as in (6) and replacing part 6. of Assumption 1 by the
corresponding statement for \( t(Y) \), we obtain
\[
E[t(Y)|\text{do}(X = x)] = \int t(y)p(y|\text{do}(X = x))dy
\]
\[
= \int E[t(Y)|X = x, X_S]dP(X_S),
\]

For example, for \( t(y) = y^2 \) we obtain second moments and we can then estimate
the variance \( \text{Var}(Y|\text{do}(X = x)) = E[Y^2|\text{do}(X = x)] - (E[Y|\text{do}(X = x)])^2 \). Or
with the indicator function \( t(y) = I(y \leq c) \) \((c \in \mathbb{R})\) we obtain a procedure for
estimating \( P[Y \leq c|\text{do}(X = x)] \) with the same convergence rate as for one-
dimensional nonparametric function estimation using marginal integration of
\( t(Y) \) versus \( X, X_S \).

2.2. Implementation of marginal integration

Theorem 1 justifies marginal integration as in (8) asymptotically. One issue is
the choice of the two bandwidths \( h_1 \) and \( h_2 \): we cannot rely on cross-validation
because \( E[Y|\text{do}(X = x)] \) is not a regression function and is not linked to pre-
diction of a new observation \( Y_{\text{new}} \), nor can we use some penalized likelihood
methods with e.g. BIC since \( E[Y|\text{do}(X = x)] \) does not appear in the likeli-
hood. Besides the difficulty to choose the smoothing parameters, we think that
addressing such a smoothing problem will become easier, at least in practice,
when using some iterative boosting approach (Friedman, 2001; Bühlmann and
Yu, 2003, cf.).

We propose here a scheme which we found to be most stable in exten-
sive simulations. The idea is to elaborate on the estimation of the function
\( m(x, x_S) = E[Y|X = x, X_S = x_S] \), from a simple starting point to more com-
plex estimates, while the integration over the variables \( X_S \) is done with the
empirical mean as in (8).

We start with the following simple but useful result.

**Proposition 1.** If \( \text{pa}(j_X) = \emptyset \) or if there are no backdoor paths from \( j_X \) to \( j_Y \)
in the true DAG \( D^0 \) from model (5), then
\[
E[Y|\text{do}(X = x)] = E[Y|X = x].
\]

Proof. If there are no backdoor paths from \( j_X \) to \( j_Y \), the empty set \( S = \emptyset \)
satisfies the backdoor criterion relative to \((X, Y)\). The statement then directly
follows from the backdoor adjustment formula (2). \( \square \)

We learn from Proposition 1 that in simple cases, a standard one-dimensional
regression estimator for \( E[Y|X = x] \) would suffice. On the other hand, we know
from the backdoor adjustment formula in (6), that we should adjust with the
variables \( X_S \). Therefore, it seems natural to use an additive regression approxi-
imation for \( m(x, x_S) \) as a simple starting point. If the assumptions of Proposition 1 hold, such an additive model fit would yield a consistent estimate for the
component of the variable $x$: in fact, it is asymptotically as efficient as when using one-dimensional function estimation for $\mathbb{E}[Y|X = x]$ (Horowitz et al., 2006). If the assumptions of Proposition 1 would not hold, we can still view an additive model fit $\hat{m}_{\text{add}} = \hat{\mu} + \hat{m}_{\text{add},jX}(x) + \sum_{j \in S} \hat{m}_{\text{add},j}(x_j)$ as one of the simplest starting points to approximate the more complex function $m(x, x_S)$.

When integrating out with the empirical mean as in (8), we obtain the estimate $\hat{\mathbb{E}}_{\text{add}}[Y|\text{do}(X = x)] = \hat{m}_{\text{add},jX}(x)$. As motivated above and backed up by simulations, $\hat{m}_{\text{add},jX}(x)$ is quite often already a reasonable estimator for $\mathbb{E}[Y|\text{do}(X = x)]$.

In the presence of strong interactions between the variables, the additive approximation may drastically fail though. Thus, we want to implement marginal integration as follows: starting from $\hat{m}_{\text{add}}$, we implement $L_2$-Boosting with the nonparametric kernel estimator as in (7). More precisely, we compute residuals

$$R_{1(i)} = Y^{(i)} - \hat{m}_{\text{add}}(X^{(i)}, X_S^{(i)}), \ i = 1, \ldots, n.$$  

For simplicity, the residuals are fitted with a locally constant marginal integration estimator similar to the one mentioned in Section 2.1, and the resulting fit is denoted by $\hat{g}_{R_b}(x, x_S)$. We add this new function fit to the previous fit and compute again residuals, and we then iterate the procedure $b_{\text{stop}}$ times. Denoting by $\hat{m}, \hat{g}$ and $Y$ the $n$-dimensional vectors evaluated at the samples $i = 1, \ldots, n$, we have: for $b = 1, \ldots, b_{\text{stop}} - 1$,

$$\hat{m}_1 = \hat{m}_{\text{add}},$$

$$\hat{m}_{b+1} = \hat{m}_b + \hat{g}_{R_b},$$

$$R_{b+1} = Y - \hat{m}_{b+1}.$$  

The final estimate for the total causal effect is obtained by marginally integrating over the variables $X_S$ with the empirical mean as in (8), that is

$$\hat{\mathbb{E}}[Y|\text{do}(X = x)] = \frac{1}{n} \sum_{k=1}^{n} \hat{m}_{b_{\text{stop}}}(x, X_S^{(k)}).$$  

The concrete implementation of the additive model fitting is according to the default from the R-package mgcv, using penalized thin plate splines and choosing the regularization parameter in the penalty by generalized cross-validation, see e.g. Wood (2006, 2003). For the product kernel in (7), we choose $K$ to be a Gaussian kernel and $L$ to be a product of Gaussian kernels. The bandwidths $h_1$ and $h_2$ in the kernel estimator should be chosen “large”, to yield an estimator with low variance but typically high bias. The iterations then reduce the bias. Once we have fixed $h_1$ and $h_2$ (and this choice is not very important as long as the bandwidths are “large”), the only regularization parameter is $b_{\text{stop}}$. It is chosen by the following considerations: for each iteration we approximate the sum of the differences to the previous approximation on the set of intervention values $\mathcal{I}$ (typically the nine deciles, see Section 5), that is

$$\sum_{x \in \mathcal{I}} \left| \frac{1}{n} \sum_{k=1}^{n} \hat{g}_{R_b}(x, X_S^{(k)}) \right|$$  

(9)
When it becomes reasonably “small”, and this needs to be specified depending on the context, we stop the boosting procedure. Such an iterative boosting scheme has the advantage that it is more insensitive to the choice of $b_{\text{stop}}$ than the original estimator in (8) to the specification of the tuning parameters, and in addition, boosting adapts to some extent to different smoothness in different directions (variables). All these ideas are presented at various places in the boosting literature, particularly in Friedman (2001); Bühlmann and Yu (2003); Bühlmann and Hothorn (2007). In Section 4.2 we provide an example of a DAG with backdoor paths, where the additive approximation is incorrect and several boosting iterations are needed to account for interaction effects between the variables. In the following, we summarize the implementation of our method in Algorithm 1.

**Algorithm 1 S-mint**

1: if $S = \emptyset$ is a valid adjustment set (for example, if $\text{pa}(j_X) = \emptyset$) then
2: Fit an additive regression of $Y$ versus $X$ to obtain $\hat{m}_{\text{add}}$
3: return $\hat{m}_{\text{add}}$
4: else
5: Fit an additive regression of $Y$ versus $X$ and the adjustment set variables $X_S$ to obtain $\hat{m}_1 = \hat{m}_{\text{add}}$
6: for $b = 2, \ldots, b_{\text{stop}} - 1$ do
7: Apply $L_2$-boosting to capture deviations from an additive regression model:
8: (i) Compute residuals $R_b = Y - \hat{m}_b$
9: (ii) Fit the residuals with the kernel estimator defined in Section 2.1 to obtain $\hat{g}_{R_b}$
10: (iii) Set $\hat{m}_{b+1} = \hat{m}_b + \hat{g}_{R_b}$
11: end for
12: return Do marginal integration: output $\frac{1}{n} \sum_{k=1}^{n} \hat{m}_{b_{\text{stop}}} (x, X_S^{(k)})$
13: end if

**2.3. Knowledge of a superset of the DAG**

It is known that a superset of the parental set $\text{pa}(j_X)$ suffices for the backdoor adjustment in (3). To be precise, let

$$S(j_X) \supseteq \text{pa}(j_X)$$

with $S(j_X) \cap \text{de}(j_X) = \emptyset$, \hspace{1cm} (10)

where $\text{de}(j_X)$ are the descendants of $j_X$ (in the true DAG $D^0$). For example, $S(j_X)$ could be the parents of $X$ in a superset of the true underlying DAG (A DAG with additional edges relative to the true DAG). We can then choose the adjustment set $S$ in (8) as $S(j_X)$ and Theorem 1 still holds true, assuming that the cardinality $|S(j_X)| \leq M < \infty$ is bounded. Thus, with the choice $S = S(j_X)$, we can use marginal integration by marginalizing over the variables $X_{S(j_X)}$.

A prime example where we are provided with a superset $S(j_X) \supseteq \text{pa}(j_X)$ with $S(j_X) \cap \text{de}(j_X) = \emptyset$ is when we know the order of the variables and can deduct an approximate superset of the parents from that. For example in a time series context, we might assume a Markovian property and estimate an upper bound of the Markov order, say a maximum time lag $p_{\text{max}}$ which we look back
for determining the conditional distributions. A simple construction for a valid superset is then (when the variables are ordered with $X_j \prec X_k$ for $j < k$):

$$S(j_X) = \{k; j_X - p_{\text{max}} \leq k < j_X\} \supseteq \text{pa}(j_X),$$

where “$\prec$” and $p_{\text{max}}$ denote the order relation among the variables and an upper bound of the lags to ensure that $S(j_X) \supseteq \text{pa}(j_X)$.

**Corollary 1.** Assume the conditions of Theorem 1 for the variables $Y, X$ and $X_{S(j_X)}$ with $S(j_X)$ in (10) or $S(j_X)$ as in (11) for ordered variables. Then,

$$\hat{E}[Y|\text{do}(X = x)] - E[Y|\text{do}(X = x)] = O(h_1^2) + O_P(1/\sqrt{nh_1}).$$

Proof. The statement is an immediate consequence of Theorem 1, as $S(j_X)$ in (10) or in (11) satisfies the backdoor criterion relative to $(X, Y)$.

2.4. Fully robust $S$-mint and connection to double robustness

Theorem 1 establishes that $S$-mint is fully robust against model-misspecification for inferring $E[Y|\text{do}(X = x)]$ or related quantities as mentioned in Remark 1. The existing framework of double robustness is related to this issue and we clarify here the connection.

We adopt for this section the more common notation used in the literature for double robustness. The outcome or response variable of interest is denoted as before by $Y$, the variable $Z$ is the intervention variable, where $Z = z$ was written above as $\text{do}(X = z)$ and $X$ denotes the (potential) confounder variables which were formalized before as $X_S$. One then specifies a regression model for $E[Y|Z, X]$ (appearing above in the backdoor adjustment in (6)) and a propensity score (Rosenbaum and Rubin, 1983) or inverse probability weighting model (IPW; Robins et al. (1994)): for a binary intervention variable where $Z$ encodes “exposure” ($Z = 1$) and “control” ($Z = 0$), the model is a logistic model for $P[Z = 1|X]$.

A double robust (DR) estimator for $E[Y_Z] := E[Y|\text{do}(Z = z)]$ requires that either the regression model or the propensity score model is correctly specified. If both of them are misspecified, the DR estimator is inconsistent. Double robustness of the augmented IPW approach has been proved by Scharfstein et al. (1999) and double robustness in general was further developed by many others, see e.g. Bang and Robins (2005).

Here, we gain full robustness by adopting a nonparametric modeling approach and without using an IPW or propensity score approach. As for DR estimators, we also assume that the potential confounders (in our terminology, the adjustment set $S$ of the intervention variable $X_S$) are known. (Semi-)parametric efficiency issues do not play a role in our procedure since we use the nonparametric marginal integration approach: thus, under correct specification of the regression or propensity score model, the DR estimators might lead to more precise estimators while under misspecification of both models, DR procedures are inconsistent in contrast to $S$-mint. Thus, $S$-mint exhibits a much more general and “full” robustness against model-misspecification.
3. Path-based methods

We assume in the following until Section 3.5 that we know the true DAG and all true functions and error distributions in the general SEM (1). Thus, in contrast to Section 2, we have here also knowledge of the entire structure in form of the DAG $D_0$ (and not only a valid adjustment set $S$ assumed for Theorem 1). This allows to infer $E[Y|do(X = x)]$ in different ways than the generic $S$-mint regression from Section 2. The motivation to look at other methods is driven by potential gains in statistical accuracy when including the additional information of the functional form or of the entire DAG in the structural equation model. We will empirically analyze this issue in Section 5.

3.1. Entire path-based method from root nodes

Based on the true DAG, the variables can always be ordered such that

$$X_{j_1} \prec X_{j_2} \prec \ldots \prec X_{j_p}.$$ 

Denote by $j_X$ and $j_Y$ the indices of the variables $X$ and $Y$, respectively.

If $X$ is not an ancestor of $Y$, we trivially know that $E[Y|do(X = x)] = E[Y]$. If $X$ is an ancestor of $Y$ it must hold that $j_X < j_Y$. We can then generate the intervention distribution of the random variables $X_{j_1} \prec X_{j_2} \prec \ldots \prec Y | do(X = x)$ in the model (1) as follows (Pearl, 2000, Def. 3.2.1.):

**Step 1** Generate $\varepsilon_{j_1}, \ldots, \varepsilon_{j_Y}$.

**Step 2** Based on Step 1, recursively generate:

$$X_{j_1} \leftarrow \varepsilon_{j_1},$$

$$X_{j_2} \leftarrow f_{j_2}^0(X_{pa(j_2)}, \varepsilon_{j_2}),$$

$$\ldots,$$

$$X_{j_X} \leftarrow x,$$

$$\ldots,$$

$$X_{j_Y} \leftarrow f_{j_Y}^0(X_{pa(j_Y)}, \varepsilon_{j_Y}).$$

Instead of an analytic expression for $p(Y|do(X = x))$ by integrating out over the other variables $\{X_k; k \neq j_X, j_Y\}$ we rather rely on simulation. We draw $B$ samples $Y^{(1)} = X_{j_Y}^{(1)}, \ldots, Y^{(B)} = X_{j_Y}^{(B)}$ by $B$ independent simulations of Steps 1-2 above and we then approximate, for $B$ large,

$$E[Y|do(X = x)] \approx B^{-1} \sum_{b=1}^B Y^{(b)}.$$

Furthermore, the simulation technique allows to obtain the distribution of $p(Y|do(X = x))$ via e.g. density estimation or histogram approximation based on $Y^{(1)}, \ldots, Y^{(B)}$. 
The method has an implementation in Algorithm 2 which uses propagation of simulated random variables along directed paths in the DAG. The method exploits the entire paths in the DAG from the root nodes to node $j_Y$ corresponding to the random variable $Y$. Figure 1 provides an illustration.

**Algorithm 2** Entire path-based algorithm for simulating the intervention distribution

1. If there is no directed path from $j_X$ to $j_Y$, the interventional and observational quantities coincide: $p(Y|\text{do}(X = x)) \equiv p(Y)$ and $E[Y|\text{do}(X = x)] \equiv E[Y]$.
2. If there is a directed path from $j_X$ to $j_Y$, proceed with steps 3-9.
3. Set $X = X_{j_X} = x$ and delete all in-going arcs into $X$.
4. Find all directed paths from root nodes (including $j_X$) to $j_Y$, and denote them by $p_1, \ldots, p_q$.
5. \textbf{for} $b = 1, \ldots, B$ \textbf{do}
   6. \hspace{0.5em} for every path, recursively simulate the corresponding random variables according to the order of the variables in the DAG:
      (i) Simulate the random variables of the root nodes of $p_1, \ldots, p_q$;
      (ii) Simulate in each path $p_1, \ldots, p_q$ the random variables following the root nodes; proceed recursively, according to the order of the variables in the DAG.
      (iii) Continue with the recursive simulation of random variables until $Y$ is simulated.
7. Store the simulated variable $Y^{(b)}$.
8. \textbf{end for}
9. Use the simulated sample $Y^{(1)}, \ldots, Y^{(B)}$ to approximate the intervention distribution $p(y|\text{do}(X = x))$ or its expectation $E[Y|\text{do}(X = x)]$.

When having estimates of the true DAG, all true functions and error distributions in the additive structural equation model (13), we would use the procedure above based on these estimated quantities; for the error distributions, we either use the estimated variances in Gaussian distributions, or we rely on bootstrapping residuals from the structural equation model (typically with residuals centered around zero).

### 3.2. Partially path-based method with short-cuts

Mainly motivated by computational considerations (see also Section 3.3), a modification of the procedure in Algorithm 2 is valid. Instead of considering all paths from root nodes to $j_Y$ (corresponding to variable $Y$), we only consider all paths from $j_X$ (corresponding to variable $X$) to $j_Y$ and simulate the random variables on these paths $p'_1, \ldots, p'_m$. Obviously, in comparison to Algorithm 2, $m \leq q$ and every $p'_k$ corresponds to a path $p_r$ for some $r \in \{1, \ldots, q\}$.

Every path $p'_k$ is of the form

$$j_X = j_{k,1} \rightarrow j_{k,2} \rightarrow \cdots \rightarrow j_{k,\ell_k-1} \rightarrow j_{k,\ell_k} = j_Y,$$

having length $\ell_k$. For recursively simulating the random variables on the paths $p'_1, \ldots, p'_k$ we start with setting

$$X = X_{j_X} \leftarrow x.$$
Then we recursively simulate the random variables corresponding to all the paths \( p'_1, \ldots, p'_m \) according to the order of the variables in the DAG. For each of these random variables \( X_j \) with \( j \in \{ p'_1, \ldots, p'_m \} \) and \( j \neq j_X \), we need the corresponding parental variables and error terms in

\[ X_j \leftarrow f^0_j(X_{pa(j)}, \varepsilon_j), \]

where for every \( k \in pa(j) \) we set

\[ X_k = \begin{cases} 
\text{the previously simulated value,} & \text{if } k \in \{ p'_1, \ldots, p'_m \}, \\
\text{bootstrap resampled } X^*_k, & \text{otherwise,}
\end{cases} \tag{12} \]

where the bootstrap resampling is with replacement from the entire data. The errors are simulated according to the error distribution.

We summarize the procedure in Algorithm 3 and Figure 1 provides an illustration.

**Algorithm 3** Partially path-based algorithm for simulating the intervention distribution

1: If there is no directed path from \( j_X \) to \( j_Y \), the interventional and observational quantities coincide: \( p(Y|\text{do}(X = x)) \equiv p(Y) \) and \( E[Y|\text{do}(X = x)] \equiv E[Y] \).
2: If there is a directed path from \( j_X \) to \( j_Y \), proceed with Steps 3-9.
3: Set \( X = X_{j_X} = x \) and delete all in-going arcs into \( X \).
4: Find all directed paths from \( j_X \) to \( j_Y \), and denote them by \( p'_1, \ldots, p'_m \).
5: \hspace{1em} for \( b = 1, \ldots, B \) do
6: \hspace{2em} for every path, recursively simulate the corresponding random variables according to the order of the variables in the DAG:
7: \hspace{3em} (i) Simulate in each path \( p'_1, \ldots, p'_m \) the random variables following the node \( j_X \);
8: \hspace{3em} proceed recursively as described in (12) according to the order of the variables in the DAG; ;
9: \hspace{3em} (ii) Continue with the recursive simulation of random variables until \( Y \) is simulated.
10: \hspace{1em} end for
11: Store the simulated variable \( Y^{(b)} \).
12: end for
13: Use the simulated sample \( Y^{(1)} , \ldots, Y^{(B)} \) to approximate the intervention distribution \( p(y|\text{do}(X = x)) \) or its expectation \( E[Y|\text{do}(X = x)] \).

**Proposition 2.** Consider the population case where the bootstrap resampling in (12) yields the correct distribution of the random variables \( X_1, \ldots, X_p \). Then, as \( B \to \infty \), the partially path-based Algorithm 3 yields the correct intervention distribution \( p(y|\text{do}(X = x)) \) and its expected value \( E[Y|\text{do}(X = x)] \).

Proof. The statement of Proposition 2 directly follows from the definition of the intervention distribution in a structural equation model. \( \square \)

The same comment applies here as in Section 3.1: when having estimates of the quantities in the additive structural equation model (13), we would use Algorithm 3 based on the plugged-in estimates. The computational benefit of using Algorithm 3 instead of Algorithm 2 is illustrated in Figure 7.
3.3. Degree of localness

We can classify the different methods according to the degree of which the entire or only a small (local) fraction of the DAG is used. Algorithm 2 is a rather global procedure as it uses entire paths from root nodes to $j_Y$. Only when $j_Y$ is close to the relevant root nodes, the method does involve a smaller aspect of the DAG. Algorithm 3 is of semi-local nature as it does not require to consider paths going from root nodes to $j_Y$: it only considers paths from $j_X$ to $j_Y$ and all parental variables along these paths. The S-mint method based on marginal integration described in Section 2 and Theorem 1, is of very local character as it only requires the knowledge of $Y$, $X$ and the parental set $pa(j_X)$ (or a superset thereof) but no further information about paths from $j_X$ to $j_Y$.

In the presence of estimation errors, a local method might be more “reliable” as only a smaller fraction of the DAG needs to be approximately correct; global methods, in contrast, require that entire paths in the DAG are approximately correct. The local versus global issue is illustrated qualitatively in Figure 1, and empirical results about statistical accuracy of the various methods are given in Section 5.
3.4. Estimation of DAG, edge functions and error distributions

With observational data, in general, it is impossible to infer the true underlying DAG $D^0$ in the structural equation model (5), or its parental sets, even as sample size tends to infinity. One can only estimate the Markov equivalence class of the true DAG, assuming faithfulness of the data-generating distribution, see Spirtes et al. (2000); Pearl (2000); Chickering (2002); Kalisch and Bühlmann (2007); van de Geer and Bühlmann (2013); Bühlmann (2013, cf.). The latter three references focus on the high-dimensional Gaussian scenario with the number of random variables $p \gg n$ but assuming a sparsity condition in terms of the maximal degree of the skeleton of the DAG $D^0$. The edge functions and error variances can then be estimated for every DAG member in the Markov equivalence class by pursuing regression of a variable versus its parents.

However, there are interesting exceptions regarding identifiability of the DAG from the observational distribution. For nonlinear structural equation models with additive error terms, it is possible to infer the true underlying DAG from infinitely many observational data (Hoyer et al., 2009; Peters et al., 2014). Various methods have been proposed to infer the true underlying DAG $D^0$ and its corresponding functions $f^0_{jk}(\cdot)$ and error distributions of the $\varepsilon_j$'s: see for example Imoto et al. (2002); Hoyer et al. (2009); Peters et al. (2014); Bühlmann et al. (2014); van de Geer (2014); Nowzohour and Bühlmann (2013) (the fourth and fifth references are considering high-dimensional scenarios). Another interesting class of models where the DAG $D^0$ can be identified from the observational data distribution are linear structural equation models with non-Gaussian noise (Shimizu et al., 2006), or with Gaussian noise but equal or approximately equal error variances (van de Geer and Bühlmann, 2013; Peters and Bühlmann, 2014; Loh and Bühlmann, 2014) (the first and third references are considering the high-dimensional setting).

As an example of a model with identifiable structure (DAG $D^0$) we can specialize (5) to an additive structural equation model of the form

$$X_j \leftarrow \sum_{k \in \text{pa}(j)} f^0_{jk}(X_k) + \varepsilon_j, \ j = 1, \ldots, p,$$

where $\varepsilon_1, \ldots, \varepsilon_p$ are independent with $\varepsilon_j \sim \mathcal{N}(0, (\sigma^0_j)^2)$, and the true underlying DAG is denoted by $D^0$. This model is used for all numerical comparisons of the S-mint procedure and the path-based algorithms in Section 5. Estimation of the unknown quantities $D^0$, $f^0_{jk}$ and error variances $(\sigma^0_j)^2$ can be done with the “CAM” method (Bühlmann et al., 2014). It is consistent, even in the high-dimensional scenario with $p \gg n$ but assuming a sparse underlying true DAG, as shown in the mentioned reference. We will use the “CAM” method for the empirical results in Section 5.4 in connection with the two-stage est S-mint in the following section.
3.5. Two-stage procedure: est S-mint

If the order of the variables or (a superset of) the parental set is unknown, we have to estimate it from observational data; this leads to the following two-stage procedure described here for the case where the parental set $\text{pa}(j_X)$ is identifiable:

**Stage 1** Estimate $S(j_X)$ described in Section 2.3, a superset of the parental set, from observational data.

**Stage 2** Based on the estimate $\hat{S}(j_X)$, run S-mint regression with $S = S(j_X)$.

Even if in Stage 1 one would also obtain estimates of functions in some specified SEM besides an estimate of $S(j_X)$, we would not use the estimated functions in Stage 2. We present empirical results for the est S-mint procedure in connection with the “CAM” method for Stage 1 for estimating a valid adjustment set $S(j_X)$ in Section 5.4.

If the parental set $\text{pa}(j_X)$ is not identifiable (see Section 3.4), one could apply Stage 1 to obtain a set $\{\hat{S}(j_X)^{(1)}, \ldots, \hat{S}(j_X)^{(c_j)}\}$ such that the parental sets from each Markov-equivalent DAG would be contained in at least one of the $\hat{S}(j_X)^{(k)}$ for some $k$. Stage 2 would then be performed for all estimates $\{\hat{S}(j_X)^{(1)}, \ldots, \hat{S}(j_X)^{(c_j)}\}$ and one could then derive bounds of the quantity $\mathbb{E}[Y|\text{do}(X = x)]$ in the spirit of the approach from Maathuis et al. (2009).

We give some intuition why the two stage est S-mint is often leading to better and more reliable results than (at least some) other methods which rely on path-based estimation in Section 5.5.

4. Empirical results: non-additive structural equation models

In this Section we provide simple proof-of-concept examples for the generality of the proposed S-mint estimation method (Algorithm 1). In particular, the robustness of S-mint is experimentally validated for models where the structural equation model is not additive but given in its general form (5). In Section 4.1 we empirically show that the path-based methods based on the wrong additive model assumption in (13) may fail even in the absence of backdoor paths where the S-mint method boils down to estimation of an additive model. In Section 4.2 we add backdoor paths to the graph and a strong interaction term to the corresponding structural equation model. We then empirically show that S-mint manages to approximate the true causal effect, whereas fitting only an additive regression fails. Section 4.3 contains an example that demonstrates a good performance of S-mint even in the presence of non-additive noise in the structural equation model. Finally, Section 4.4 empirically illustrates issues with the fixed choice of the bandwidths in the product kernel in (7) in some cases.

4.1. Causal effects in the absence of backdoor paths

First let us illustrate the sensitivity of the path-based methods with respect to model specification, using a simple example of a 4-node graph with no backdoor
paths between $X_1 = X$ and $Y$ (see Figure 2). We consider a corresponding (non-additive) structural equation model of the form

$$
\begin{align*}
X_1 & \leftarrow \varepsilon_1 \\
X_2 & \leftarrow \varepsilon_2 \\
X_3 & \leftarrow \cos(4 \cdot (X_1 + X_2)) \cdot \exp(X_1/2 + X_2/4) + \varepsilon_3 \\
Y & \leftarrow \cos(X_3) \cdot \exp(X_3/4) + \varepsilon_4
\end{align*}
$$

(14)

where $\varepsilon_j \sim N(0, \sigma_j^2)$ with $\sigma_1 = \sigma_2 = 0.7$ and $\sigma_3 = \sigma_4 = 0.2$. We generate $n$ samples from this model. From Proposition 1 we know that for $j \in \{1, 2, 3\}$, fitting an additive regression of $Y$ versus $X_j$ and $X_{pa}(j)$ suffices to obtain the causal effect $\mathbb{E}[Y|do(X_j = x)]$, that is, all causal effects can be readily estimated with an additive model. Our goal is to infer $\mathbb{E}[Y|do(X_1 = x)]$, based on $n = 500$ and $n = 10^5$ samples of the joint distribution of the 4 nodes. The results are displayed in Figure 2.

We consider the entire path-based Algorithm 2 (and Algorithm 3 as well, not shown) assuming an additive structural equation model as in (13). We impressively see that this approach is exposed to model misspecification while $S$-mint (in this case simply fitting of an additive model, i.e., $b_{stop} = 1$ with the number of additional boosting iterations equaling zero) is not and leads to reliable and correct results. We included two settings; $n = 500$ to be consistent with the settings in the numerical study from Section 5 and $n = 10000$ to demonstrate that the failure of the path-based methods is not a small sample size but an inconsistency phenomenon.
4.2. Causal effects in the presence of backdoor paths

We now consider a slight (but crucial) modification of the above DAG that has been proposed by Linbo Wang and Mathias Drton through private communication. We consider the 4-node graph from Section 4.1 with additional edges $X_1 \to Y$ and $X_2 \to Y$ and corresponding structural equation model

\begin{align*}
X_1 &\leftarrow \varepsilon_1 \\
X_2 &\leftarrow \varepsilon_2 \\
X_3 &\leftarrow X_1 + X_2 + \varepsilon_3 \\
Y &\leftarrow X_1 \cdot X_2 \cdot X_3 + \varepsilon_4
\end{align*}

(15)

where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ with $\sigma_1 = \sigma_2 = 0.7$ and $\sigma_3 = \sigma_4 = 0.2$. Note that this modification introduces two backdoor paths from $X_3$ to $Y$. The goal is to estimate the causal effect $\mathbb{E}[Y|\text{do}(X_3 = x)]$ using the S-mint estimation procedure introduced in Algorithm 1 with different numbers of boosting iterations. In Figure 3 one clearly sees that the additive approximation (with no additional boosting iterations) fails to approximate the total causal effect. It is not able to capture

![Graph showing approximations of the causal effect](image)

**Fig 3.** Approximation of the causal effect $\mathbb{E}[Y|\text{do}(X_3 = x)]$ in model (15) with S-mint regression for additive model fit (starting value) and various boosting iterations (left), absolute differences between consecutive boosting iterations as in (9) (upper right) and integrated squared error for approximating the true effect as a function of boosting iterations (lower right). The boosting iterations in the S-mint procedure account for interactions between the variables. The adjustment set is chosen as the parent set of $X_3$, that is $S(j_X = 3) = \{1, 2\}$. The results are based on one representative sample of size $n = 500$ (top) and $n = 10000$ (bottom).
the full interaction term $X_1 \cdot X_2 \cdot X_3$. However, adding boosting iterations significantly improves the approximation of the true causal effect even for the small sample size $n = 500$.

### 4.3. Causal effects in the presence of non-additive noise

Theorem 1 does not put any explicit restrictions on the noise structure in the structural equation model. In particular, $S$-mint also works well in the case of non-additive noise. As an example, let us consider the causal graph and SEM from Section 4.2, but we replace the structural equation corresponding to $Y$ in (15) with

$$Y \leftarrow \exp(X_1) \cdot \cos(X_2 \cdot X_3 + \varepsilon_4).$$

(16)

The goal is again to estimate the causal effect $E[Y|do(X_3 = x)]$ based on $n = 500$ observed samples of the joint distribution. Figure 4 shows that $S$-mint yields a close approximation to the true causal effect.

![Figure 4](image_url)

**Fig 4.** Approximation of the causal effect $E[Y|do(X_3 = x)]$ in model (16) exhibiting non-additive noise in the structural equation model, with $S$-mint regression for additive model fit (starting value) and various boosting iterations (left). Absolute differences between consecutive boosting iterations as in (9) (upper right) and integrated squared error for approximating the true effect as a function of boosting iterations (lower right). The adjustment set is chosen as the parental set of $X_3$, that is $S(jX = 3) = \{1, 2\}$. The results are based on one representative sample of size $n = 500$.

### 4.4. Choice of the bandwidth

Theorem 1 provides an asymptotic result but does not specify how to choose the bandwidths $h_1$ and $h_2$ in the finite sample case. In particular, the same fixed choice of $h_2$ for all variables in the adjustment set $S$ can be suboptimal in some situations. As an example let us consider the graph and structural equations from Section 4.2 where we replace one equation in (15) by

$$Y \leftarrow X_1 + \sin(X_2 \cdot X_3 + \varepsilon_4).$$

(17)
The goal is to approximate the causal effect $\mathbb{E}[Y|\text{do}(X_3 = x)]$ based on $n = 500$ samples of the joint distribution. Inspecting the scatterplots of $Y$ versus $X_1$, $X_2$ and $X_3$ (see Figure 5) suggests that the bandwidth $h_{2(1)}$ corresponding to $X_1$ should be larger than the bandwidth $h_{2(2)}$ corresponding to $X_2$. Figure 6 depicts the corresponding approximated causal effects using the $S$-mint method for a fixed bandwidth $h_2 = (h_{2(1)}, h_{2(2)}) = (0.4, 0.4)$ and for a variable bandwidth $h_2 = (h_{2(1)}, h_{2(2)}) = (0.8, 0.4)$ respectively. Clearly, the approximation with the variable bandwidth outperforms the approximation with the fixed bandwidth. Adaptive bandwidths choice methods as proposed by Polzehl and Spokoiny (2000) might be suitable, at the price of a more complicated and hence more variable estimation scheme.
5. Empirical results: additive structural equation models

The goal of the numerical experiments in this section is to quantify the estimation accuracy of the total causal effect \(E[Y|\text{do}(X = x)]\) for two variables \(X, Y \in \{X_1, ..., X_p\}\) such that \(Y\) is a descendant of \(X\) (if \(Y\) is an ancestor of \(X\), then the interventional expectation corresponds to the observational expectation \(E[Y]\)). We consider in this section only additive structural equation models as in (13). This allows for a comparison of the \(S\)-mint method and the path-based methods.

For \(S\)-mint regression, we use the implementation described in Section 2.2. The kernel functions \(K\) and \(L\) in the \(S\)-mint procedure are chosen to be a Gaussian kernel with bandwidth \(h_1\) and a product of Gaussian kernels with bandwidth \(h_2\) respectively. For simplicity, in the style of Fan et al. (1998), we choose \(h_1\) and \(h_2\) as 0.5 times the empirical standard deviation of the respective covariables in all of our simulations in this section. We use the following two criteria for \(b_{\text{stop}}\), that is, as an automated stopping criterion for the boosting iterations:

1. Stop if an iteration changes the approximation by less than 1%. That is, the integrated difference (9) to the previous approximation is less than 0.01.
2. Stop if the integrated difference between two consecutive approximations is less than 5% of the initial integrated difference.

When using the path-based methods from Section 3, we estimate the functions \(f^0_{j,k}\) by additive functions using the R-package mgcv with default values (and thus using the knowledge of the form of the nonlinear functions in the SEM).

We test the performance of four different methods: \(S\)-mint with parental sets (Algorithm 1) with the stopping of boosting iterations as described above, additive regression with parental sets (first step of \(S\)-mint, without additional boosting iterations), entire path-based method from root nodes (Algorithm 2) and partially path-based method with short-cuts (Algorithm 3). The reference effect \(E[Y|\text{do}(X = x)]\) is computed using Algorithm 2 with known (true) functions \(f^0_{j,k}\) and error variances \((\sigma^0)^2\) based on 5\(n\) samples.

Since in a nonlinear structural equation model (in contrast to a linear structural equation model) \(E[Y|\text{do}(X = x)]\) is a nonlinear function of the intervention value \(x\), we compute the interventional expectation for several values \(x\): typically, for the nine deciles \(d_1(X), ..., d_9(X)\) of \(X\). To compare the estimation accuracy of the three methods on DAG \(D\), we compute a relative squared error \(e(D)\) over all considered pairs \((X,Y)\) (for details see below), denoted by \(L\), and over all intervention values \(d_1(X), ..., d_9(X)\) as

\[
e(D) = \frac{\sum_{(X,Y) \in L} \sum_{i=1}^9 \left( \frac{E[Y|\text{do}(X = d_i(X))] - E^0[Y|\text{do}(X = d_i(X))]}{\sum_{(X,Y) \in L} \sum_{i=1}^9 (E^0[Y|\text{do}(X = d_i(X))])} \right)^2}{\sum_{(X,Y) \in L} \sum_{i=1}^9 (E^0[Y|\text{do}(X = d_i(X))])^2}. \quad (18)
\]
Typically, we repeat every experiment on $N = 50$ or $N = 100$ random DAGs (described in Section 5.1) and record the relative error $e(D)$ of all methods for each repetition.

5.1. Data simulation

To simulate data we first fix a causal order $\pi^0$ of the variables, that is $X_{\pi^0(1)} \prec X_{\pi^0(2)} \prec \cdots \prec X_{\pi^0(p)}$ and include each of the $\binom{p}{2}$ possible directed edges, independently of each other, with probability $p_c$. In the sparse setting we typically choose $p_c = \frac{2}{p-1}$ which yields an expected number of $p$ edges in the resulting DAG. Based on the causal structure of the graph we then build the structural equation model. We simulate from the additive structural equation model (13), where every edge $k \to j$ in the DAG is associated with a nonlinear function $f_{j,k}^0$ in the structural equation model. We use two function types:

1. edge functions $f_{j,k}^0$ drawn from a Gaussian process with a Gaussian kernel with bandwidth one
2. sigmoid-type edge functions of the form $f_{j,k}^0(x) = a \cdot \frac{k(x+c)}{1 + |b \cdot (x+c)|}$ with $a \sim \text{Exp}(4) + 1$, $b \sim \text{Unif}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \text{Unif}([-2, 2])$.

All variables with empty parental set (root nodes in the DAG) follow a Gaussian distribution with mean zero and standard deviation which is uniformly distributed in the interval $[1, \sqrt{2}]$. To all remaining variables we add Gaussian noise with standard deviation uniformly distributed in $[1/5, \sqrt{2}/5]$. Note that both simulation settings correspond to the ones used by Bühlmann et al. (2014).

5.2. Estimation of causal effects with known graphs

In this section we compare the different methods in terms of estimation accuracy and CPU time consumption for known underlying DAGs $D^0$. To that end we generate random DAGs with $p = 10$ variables and simulate $n = 500$ samples of the joint distribution applying the simulation procedure introduced in Section 5.1. We then select all index pairs $(k, j)$ such that there exists a directed path from $X_k$ to $X_j$ and estimate the causal effect $\mathbb{E}[X_j|\text{do}(X_k)]$ for all $k, j$ on the nine deciles of $X_k$.

The experiment is done for two different levels of sparsity, a sparse graph with an expected number of $p$ edges and a non-sparse graph with an expected number of $4p$ edges. We record the relative squared error (18) and the CPU time consumption, both averaged over all index pairs, for $N = 100$ ($N = 20$ in the dense setting, respectively) different DAGs $D^0$. The results are displayed in Figure 7 for the sigmoid-type edge functions and in Figure 8 for the Gaussian process-type edge functions.

The method based on the entire paths (Algorithm 2) yields the smallest errors followed by the path-based methods with short-cuts (Algorithm 3). The $S$-mint and additive regression exhibit a slightly worse performance. This finding can be explained by the fact that the path-based methods benefit from the
full (and correct) structural information of the DAG whereas the \( S\)-mint and additive regression methods only use local information (cf. Section 3.3). For the monotone sigmoid-type function class, additive regression seems to be a very good approximation to the true causal effect even in dense settings. For both settings we observe that the boosting iterations in \( S\)-mint do not improve the additive approximation substantially.

In terms of CPU time consumption, \( S\)-mint and additive regression outperform the path-based methods. Additive regression is particularly fast as it only requires the fit of one nonparametric additive regression of \( X_j \) versus \( X_k \) and \( X_{\text{pa}(k)} \) whereas the path-based methods each require one nonparametric additive model fit for every node on all the traversed paths. As the set of paths in the partially path-based method is a subset of the one in the entire path-based method (cf. Section 3.2 and Figure 1), the partially path-based method needs less model fits which explains the reduction of time consumption. In particular, both \( S\)-mint and additive regression are computationally feasible for computing \( \mathbb{E}[X_j|\text{do}(X_k)] \) for all pairs \((k,j)\), even when \( p \) is large and in the thousands assuming that the cardinality of the corresponding adjustment sets are reasonably small.
5.3. Estimation of causal effects on perturbed graphs

In the previous section we demonstrated that the two path-based methods exhibit a better performance than S-mint and the additive regression approximation if causal effects are estimated based on the underlying true DAG $D^0$. We will now focus on the more realistic situation in which we are only provided with a partially correct DAG $\tilde{D}$. We model this by constructing a set of modified DAGs $\{\tilde{D}_{hr}\}_{r \in K}$ with pre-specified (fixed) structural Hamming distances $\{h_r\}_{r \in K}$ to the true DAG $D_0$, where $K = \{1, 2, \ldots, 6\}$ and the corresponding $\{h_r\}_{r \in K}$ are described in Figures 9 and 10. To do so, we use the following rule: starting from $D^0$ with $p = 50$ nodes, for each $r \in K$, we randomly remove and add $\frac{h_r}{2}$ edges each to obtain $\tilde{D}_{hr}$. The structural Hamming distance between $D_0$ and the perturbed graph $\tilde{D}_{hr}$ is then equal to $h_r$, and a percentage of $1 - \frac{h_r}{2|E|}$ edges in $\tilde{D}_{hr}$ are still correct, where $|E|$ denotes the number of edges in the DAG $D^0$. Note that this modification may change the order of the variables (especially for large values of $h_r$).

We randomly choose 20 = $|L|$ index pairs $(k, j)$ such that there exists a directed path from $X_k$ to $X_j$ in $D^0$, but now consider the problem of estimating the total causal effect $E[X_j|\text{do}(X_k)]$ based on the perturbed graph $\tilde{D}_{hr}$ for the adjustment sets or the paths, respectively (and based on sample size $n = 500$ as in Section 5.2). For every $r \in K$, this is repeated $N = 100$ times and in each repetition, we record the relative squared error $e(D)$ in (18). As before, we distinguish between a sparse graph with an expected number of 50 edges and a non-sparse graph with an expected number of 200 edges and we use both simulation settings described in Section 5.1 for generating the edge functions $f^0$. The results are shown in Figures 9 and 10.

For both, the sparse and non-sparse settings, one observes that the larger the structural Hamming distance (or equivalently, the smaller the percentage of correctly specified edges in $D^0$), the better is the performance of S-mint and additive regression in comparison with the path-based methods. That is, both methods are substantially more robust with respect to possible misspecifications of edges in the graph. This may be explained by the different degrees of localness (cf. Section 3.3) of the respective methods. For the two local methods we can hope to have approximately correct information in the parental set of $X_k$ even if the modified DAG is far away from the true DAG $D^0$ in terms of the structural Hamming distance. For the path-based methods however, randomly removing edges may break one or several of the traversed paths which results in causal information being partially or fully lost. This effect is most evident in the two sparse settings. A similar behavior is also observed in Figure 11.

Note that except for the true DAG $D^0$, the performance of the partially path-based method is at least as good as for the entire path-based method. The shortcut introduced in Algorithm 3 does not only seem to yield computational savings but also improve (relative to the full path-based Algorithm 2) statistical estimation accuracy of causal effects in incorrect DAGs. Again, a possible explanation for this observation is that the partially path-based method acts
Fig 9. The plots compare the relative squared error performance of the three methods on a set of modified DAGs \{\tilde{D}_{h_r}\}_{r \in K} with given structural Hamming distances \{h_r\}_{r \in K} to the true DAG \hat{D}^0 (or equivalently, with a given percentage of correct edges) for the sigmoid-type additive structural equation model. The top and bottom panels show the relative squared error error \(e(D)\) in a sparse and dense setting, respectively. The larger the structural Hamming distance \(h_r\) between the modified DAG \(\tilde{D}_{h_r}\) and the true DAG \(D^0\), the better is the performance of \(S\)-mint with parental sets in comparison with the two path-based methods. Number of variables \(p = 50\) and sample size \(n = 500\).

5.4. Estimation of causal effects in estimated graphs

We now turn our attention to the case where the goal is to compute causal effects on a DAG \(\hat{D}\) that has been estimated by a structure learning algorithm (while still relying on a correct model specification). In conjunction with \(S\)-mint regression, this is then the method \(est\ S\)-mint described in Section 3.5.

We generate \(N = 50\) random DAGs with \(p = 20\) nodes for different numbers \(n\) of observational data, which are simulated according to the procedure in Section 5.1.

Using the knowledge that the structural equation model is additive, we apply the recently proposed CAM-algorithm (Bühlmann et al., 2014) for estimation of the true underlying DAG \(D^0\) (which is identifiable from the observational distribution). The CAM-algorithm involves the following three steps:

1. Preliminary neighborhood selection to restrict the number of potential parents per node (set to a maximum of 10 by default);
2. Estimation of the correct order by greedy search (we use 6 basis functions per parent to fit the generalized additive model);
3. Optional: Pruning of the DAG by feature selection to keep only the significant edges (on level \(\alpha = 0.001\) by default).
After having estimated a DAG $\hat{D}$ with the above procedure, we randomly select $10 = |\mathcal{L}|$ index pairs $(k, j)$ such that there exists a directed path from $X_k$ to $X_j$ in the true DAG $D^0$ and approximate the total causal effect $\mathbb{E}[X_j|\text{do}(X_k)]$ based on the estimated graph $\hat{D}$. Figure 11 displays the relative squared errors as defined in (18).

All four methods show a similar performance with respect to relative squared error on the DAGs that are obtained applying the CAM-algorithm without feature selection. These DAGs mainly represent the causal order of the variables but otherwise are densely connected. An incorrectly specified order of the variables (e.g. for small sample sizes $n$) seems to comparably affect the $S$-$\text{mint}$ and additive regression with parental sets and the path-based methods. If the sample size increases, the estimated graph $\hat{D}$ is closer to the true graph $D^0$ which improves the estimation accuracy of causal effects for all the four methods.

The two path-based methods approximate the causal effects more accurately on the DAGs that are obtained without feature selection, that is, pruning the DAG does not seem to be advantageous for the estimation accuracy of causal effects, at least for a small number of observations. However, the pruning step yields vast computational savings for the two path-based methods as demonstrated in Figure 12. The $S$-$\text{mint}$ regression is very fast in both settings and pruning the DAG before estimating the causal effects only has a minor effect on the time consumption and estimation accuracy.
Fig 11. Sigmoid-type additive structural equation models. Relative squared error performance as in (18), for different numbers of observations (n), computed on graphs that have been estimated using the CAM-algorithm (Bühlmann et al., 2014). The algorithm has been applied without the pruning step (left) and with the pruning step (right). We use the estimated parental sets as adjustment sets and the number of variables is \( p = 20 \). The S-mint regression corresponds to \( \text{est S-mint} \) as described in Section 3.5.

Fig 12. Sigmoid-type additive structural equation models. CPU time performance for \( n = 500 \) for \( N = 50 \) graphs of \( p = 20 \) variables that have been estimated using the CAM-algorithm (Bühlmann et al., 2014) with and without pruning step. Pruning the DAG yields vast computational savings for the two path-based methods. S-mint and additive regression are barely affected by the pruning step and are considerably faster than the two path-based methods in both scenarios.

5.5. Summary of the empirical results, and the advantage of the proposed two-stage \( \text{est S-mint} \) method

With respect to statistical accuracy, measured with the relative squared error as in (18), we find that \( S\text{-mint} \) and additive regression are substantially more robust against incorrectness of the true underlying DAG (or against a wrong order of the variables) and against model misspecification, in comparison to the alternative path-based methods. The latter robustness of \( S\text{-mint} \) is rigorously backed-up by our presented theory in Theorem 1 and Corollary 1 whereas the former seems to be due to the higher degree of localness as described in Section 3.3. Therefore, the proposed two-stage \( \text{est S-mint} \) (Section 3.5) where we first estimate the order of the variables or the structure of the DAG (or the Markov equivalence class of DAGs) and subsequently perform \( S\text{-mint} \) is expected in general to lead to reasonably accurate results (and empirically quantified above for some settings). Only when the DAG is perfectly known and the model correctly specified (here by an additive structural equation model), which
seems a rather unrealistic assumption for practical applications, the path-based methods were found to have a slight advantage. Thus, we recover here a typical robustness phenomenon against model misspecification of our nonparametric and more “model-free” S-mint regression procedure.

Regarding computational efficiency, S-mint and in particular also the additive regression approximation are massively faster than the path-based procedures making them feasible for larger scales where the number of variables is in the thousands.

6. Real data application

In this section we want to provide two examples for the application of our methodology to real data. We use gene expression data from the isoprenoid biosynthesis in *Arabidopsis thaliana* (Wille et al., 2004). The data consists of $n = 118$ gene expression measurements from $p = 39$ genes. In the original work the authors try to infer connections between the individual genes in the network using Gaussian graphical modeling. Our goal is to find the strongest causal connections between the individual genes. We do not standardize the original data but adjust the bandwidths in S-mint by scaling with the standard deviations of the corresponding variables.

6.1. Estimation and error control for causal connections between and within the pathways

We first turn our attention to the whole isoprenoid biosynthesis data set and want to find the causal effects within and between the different pathways, with an error control for false positive selections. To be able to compute the causal effects we have to estimate a causal network. In order to do that we use the CAM-algorithm (Bühlmann et al., 2014).

We estimate a DAG using CAM with the default settings. We then apply the S-mint procedure with parental sets obtained from the estimated DAG (which corresponds to the est S-mint procedure from Section 3.5) to rank the total causal effects according to their strength. We define the relative causal strength $CS_{k \rightarrow j}$ of an intervention $X_j|\text{do}(X_k)$ as a sum of relative distances of observational and interventional expectation for different intervention values divided by the range of the intervention values, i.e.

$$CS_{k \rightarrow j}^{rel} = \frac{1}{R_k(d)} \sum_{i=1}^{9} \frac{\left| E[X_j] - E[X_j|\text{do}(X_k = d_i)] \right|}{|E[X_j]|},$$

where we choose $d_1(X_k), ..., d_9(X_k)$ to be the nine deciles of $X_k$ and we denote their range by $R_k(d) = d_9(X_k) - d_1(X_k)$.

To control the number of false positives (i.e. falsely selected strong causal effects) we use stability selection (Meinshausen and Bühlmann, 2010) which
provides (conservative) error control under a so-called (and uncheckable) exchangeability condition. We randomly select 100 subsamples of size \( n/2 = 59 \) and repeat the procedure above 100 times. For each run, we record the indices of the top 30 ranked causal strengths. At the end we keep all index pairs that have been selected at least 66 times in the 100 runs as this leads to an expected number of falsely selected edges (false positives) which is less or equal to 2 (Meinshausen and Bühlmann, 2010). The graphical representation of the network in Figure 13 is based on Wille et al. (2004). The dotted arcs represent the underlying metabolic network (known from biology), the six red solid arcs correspond to the stable index pairs found by \( \text{est S-mint} \) with stability selection.

None of the stable edges are opposite to the causal direction of the metabolic network. In particular, there seems to be a strong total causal effect between GGPPS variables in the MEP pathway, MVA pathway and mitochondrion. Note that in this section we heavily rely on model assumption (13) as the CAM-algorithm for estimating a DAG assumes additivity of the parents. Therefore we can not fully exploit the advantage of the \( S\text{-mint} \) method that it works for arbitrary non-additive models (5) (but we would hope to be somewhat less sensitive to model misspecification than with path-based methods, see for example Figures 9 and 10).
6.2. Estimation and error control of strong causal connections within the MEP pathway

We now want to present a possible way of exploiting the very general model assumptions of S-mint. If the underlying order and an approximate graph structure are known a priori, we can use this information to proceed with S-mint using the order information as described in Corollary 1. This relieves us from any model assumptions on the functional connections between two variables (e.g. linearity, additivity, etc.).

To give an example, let us focus on the genes in the MEP pathway (black box in Figure 13). The goal is to find the strongest total causal effects within this pathway. The metabolic network (dotted arcs) is providing us with an order of the variables which we use for S-mint regression as follows: we choose the adjustment set $S(j|x)$ in (11) by going three levels back ($p_{\max} = 3$) in the causal order (to achieve a reasonably sized set), for example, the adjustment set for CMK is DXPS1, DXPS2, DXPS3, DXR, MCT, whereas the adjustment set for GPPS is HDS, HDR, IPP1. We cannot use the full set of all ancestors because there are only $n/2 = 59$ data points to fit the nonparametric additive regression and marginal integration, as we again use stability selection based on subsampling for controlling false positive selections as described in the previous section. For each among the 100 subsampling runs we record the top 10 ranked index pairs and keep the ones that are selected at least 65 times out of 100 repetitions. This results in an expected number of false positives being less than 1 (Meinshausen and Bühlmann, 2010). The stable edges are shown in Figure 14. One of the four edges corresponds to an edge in the metabolic pathway. We find that the upper part of the pathway contains the strongest total causal effects and it might be an interesting target for intervention experiments.

7. Conclusions

We considered the problem of inferring expected values of intervention distributions from observational data. A first main result (Theorem 1 and Corollary 1) says that if we know the local parental variables or a superset thereof (e.g., from the order of the variables), there is no need to base estimation and computations on a causal graph as we can directly infer the expected values of single-intervention distributions via marginal integration: we call the procedure S-mint. This result holds for any nonlinear and non-additive structural equation model apart from some mild smoothness and regularity conditions. Thus, from another point of view, S-mint estimation of expected values of single intervention distributions is robust against model misspecification of the functional form of the structural equations.

We complement the robustness view-point by empirical results indicating that S-mint also works reasonably well when the DAG- or order-structure is misspecified to a certain extent, as will be the case when we estimate these quantities from data; in fact, S-mint regression is substantially more robust
Fig 14. Stable edges (with stability selection) for the MEP pathway in the Arabidopsis thaliana dataset. The dotted arcs represent the metabolic network whereas the red solid arcs denote the top ranked causal effects found by \textit{S-mint} with adjustment sets chosen from the order of the metabolic network structure by considering all ancestors up to three levels back.

than methods which follow all directed paths in the DAG to infer causal effects. This suggests that the two-stage \textit{est S-mint} procedure is most reliable for causal inference from observational data: first estimate the DAG- or order-structure (or equivalence classes thereof) and second, subsequently pursue \textit{S-mint} regression. In addition, such a procedure is computationally much faster than methods which exploit directed paths in (estimated) DAGs.

Acknowledgments. We sincerely thank Linbo Wang and Mathias Drton for having pointed out a major error in an earlier and very different version of the paper. We also thank anonymous reviewers for constructive comments.

References

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. \textit{Biometrics}, 61:962–972.
Bollen, K. A. (1998). *Structural equation models*. Wiley Online Library.

Bühlmann, P. (2013). Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research*, 77:357–370.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science*, 22:477–505.

Bühlmann, P., Peters, J., and Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42(6):2526–2556.

Bühlmann, P. and Yu, B. (2003). Boosting with the $L_2$ loss: regression and classification. *Journal of the American Statistical Association*, 98:324–339.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Colombo, D., Maathuis, M., Kalisch, M., and Richardson, T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40:294–321.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Editorial (2010). Cause and effect. *Nature Methods*, 7:243.

Fan, J., Härdle, W., Mammen, E., et al. (1998). Direct estimation of low-dimensional components in additive models. *Annals of Statistics*, 26(3):943–971.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13:2409–2464.

Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55:926–939.

Hauser, A. and Bühlmann, P. (2015). Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society, Series B*, 77:291–318.

He, Y.-B. and Geng., Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research*, 9:2523–2547.

Horowitz, J., Klemelä, J., and Mannen, E. (2006). Optimal estimation in additive regression models. *Bernoulli*, 12:271–298.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pages 689–696.
Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19:2271–2282.

Imoto, S., Goto, T., and Miyano, S. (2002). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. In *Proceedings of the Pacific Symposium on Biocomputing (PSB-2002)*, volume 7, pages 175–186.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, pages 157–224.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100.

Loh, P. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15:3065–3105.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164.

Meinshausen, N. and Bühlmann, P. (2010). Stability Selection (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:417–473.

Nowzohour, C. and Bühlmann, P. (2013). Score-based causal learning in additive noise models. arXiv:1311.6359.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ. Press.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101:219–228.

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

Polzehl, J. and Spokoiny, V. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society, Series B*, 62:335–354.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some of the regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469).

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94:1096–1146.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). Learning graphical model structure using 11-regularization paths. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1278. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Shimizu, S., Hoyer, P., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.

Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97:519–538.

Shpitser, I., Richardson, T. S., and Robins, J. M. (2011). An efficient algorithm for computing interventional distributions in latent variable causal models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 661–670.

Smith, V. A., Jarvis, E. D., and Hartemink, A. J. (2002). Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, 18(suppl 1):S216–S224.

Spirtes, P. (2010). Introduction to causal inference. *The Journal of Machine Learning Research*, 11:1643–1662.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press, second edition.

Stekhoven, D., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M., and Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28:2819–2823.

Teyssier, M. and Koller, D. (2005). Ordering-based search: a simple and effective algorithm for learning Bayesian networks. In *Proceedings of the 21th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 584–590, Edinburgh, Scotland, UK.

van de Geer, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electronic Journal of Statistics*, 8(1):543–574.

van de Geer, S. and Bühlmann, P. (2013). $l_0$-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41:536–567.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., et al. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biol.*, 5(11):R92.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statis-
Yu, J., Smith, V. A., Wang, P. P., Hartemink, A. J., and Jarvis, E. D. (2004). Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20:3594–3603.

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896.