CAUSAL DANTZIG: FAST INFERENCE IN LINEAR STRUCTURAL EQUATION MODELS WITH HIDDEN VARIABLES UNDER ADDITIVE INTERVENTIONS

BY DOMINIK ROTHENHÄUSLER, PETER BÜHLMANN AND
NICOLAI MEINSHAUSEN

ETH Zürich

Causal inference is known to be very challenging when only observational data are available. Randomized experiments are often costly and impractical and in instrumental variable regression the number of instruments has to exceed the number of causal predictors. It was recently shown in Peters, Bühlmann and Meinshausen (2016) (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012) that causal inference for the full model is possible when data from distinct observational environments are available, exploiting that the conditional distribution of a response variable is invariant under the correct causal model. Two shortcomings of such an approach are the high computational effort for large-scale data and the assumed absence of hidden confounders. Here, we show that these two shortcomings can be addressed if one is willing to make a more restrictive assumption on the type of interventions that generate different environments. Thereby, we look at a different notion of invariance, namely inner-product invariance. By avoiding a computationally cumbersome reverse-engineering approach such as in Peters, Bühlmann and Meinshausen (2016), it allows for large-scale causal inference in linear structural equation models. We discuss identifiability conditions for the causal parameter and derive asymptotic confidence intervals in the low-dimensional setting. In the case of nonidentifiability, we show that the solution set of causal Dantzig has predictive guarantees under certain interventions. We derive finite-sample bounds in the high-dimensional setting and investigate its performance on simulated datasets.

1. Introduction. Using only observational data to infer causal relations is a challenging task and only possible under certain circumstances and assumptions. In the context of structural equation models (Bollen (1989), Pearl (2009), Robins, Hernan and Brumback (2000)), one possibility is to characterize the Markov equivalence class of graphs under the assumption of acyclicity and usually faithfulness (Andersson, Madigan and Perlman (1997), Chickering (2003), Hauser and Bühlmann (2012), Tian and Pearl (2001), Verma and Pearl (1991)). Based on the Markov equivalence class, some causal effects and often only bounds for them can be inferred; see, for example, Maathuis, Kalisch and Bühlmann (2009) and

VanderWeele and Robins (2010). Other approaches exploit non-Gaussianity or nonlinearities, while making suitable assumptions about the causal model (Hoyer et al. (2009), Shimizu et al. (2006)).

If both observational and data under interventions are available and the target and effect of the interventions is perfectly known, the task of inferring causal relationships becomes easier. Hauser and Bühlmann (2015), for example, modify the greedy equivalence search of Chickering (2003) to such a scenario. If an instrumental variable is available, then different forms of instrumental variable regression (Angrist, Imbens and Rubin (1996), Bowden and Turkington (1990), Didelez, Meng and Sheehan (2010), Wright (1928)) can be used to infer the causal effect of a single variable on a target of interest.

Consider a setting where data are recorded in different environments. The environments can have an arbitrary and unknown intervention effect on all predictor variables and the method exploits that the conditional distribution of the target Y of interest, given its causal parents, is invariant across environments under arbitrary interventions on all variables (excluding, just as in instrumental variable regression, direct interventions on the response or target Y). While it was demonstrated in Peters, Bühlmann and Meinshausen (2016) that the method can infer a full causal model, there are two major shortcomings:

(i) It is assumed for invariant causal prediction (ICP) (Peters, Bühlmann and Meinshausen (2016)) that there are no hidden variables that influence Y and its parents simultaneously.

(ii) ICP scans all potential subsets of variables and tests whether the conditional distribution of Y given a subset of variables is invariant across all environments. This makes the method computationally prohibitively expensive as soon as the number of predictor variables starts to exceed one or two dozens.

We will show that both shortcomings can be addressed if we are willing to make a more specific assumption about the type of interventions that generate the different environments.

1.1. *Setting and notation.* Assume we have a $p + 1$ variables $X_1, \ldots, X_{p+1}$ from a linear Structural Equation Model (SEM) (Bollen (1989), Pearl (2009), Robins, Hernan and Brumback (2000)),

$$(1) \qquad X_k \leftarrow \sum_{k' \neq k} A_{k,k'} X_{k'} + \eta_k, \qquad k = 1, \ldots, p + 1,$$

where $\mathrm{pa}(k) := \{k' : A_{k,k'} \neq 0\} \subseteq \{1, \ldots, p + 1\} \setminus k$ is the set of parents of variable k. For notational simplicity, we set $A_{k,k} := 0$ for all k. Deviating from convention, we allow dependence between the components of the noise contribution $\eta = (\eta_1, \ldots, \eta_{p+1})$ which is equivalent to allowing for hidden variables as parents of the observed variables $X_1, \ldots, X_{p+1}$; see Figure 1 for an example. The variables form a directed graph $G = (V, E)$, where the nodes $V = \{1, \ldots, p + 1\}$ are
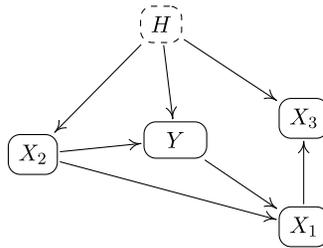
FIG. 1. *An example for $p = 3$. The hidden variable H is suppressed notationally and instead the noise contributions at each variable $\{X_1, X_2, X_3, X_4 = Y\}$ are not assumed to be independent.*

given by the variables themselves and there is an edge from variable k to k' if and only if $k \in \mathrm{pa}(k')$. Furthermore, we allow the underlying graph to be cyclic. The values $(A_{k,k'})$ for $k, k' \in \{1, \ldots, p+1\}$ form a $(p+1) \times (p+1)$-dimensional matrix that we denote by A. We write Id_{p+1} for the $(p+1) \times (p+1)$-dimensional identity matrix. To make the distribution of $X_1, \ldots, X_{p+1}$ well defined in the presence of cycles, we assume that $\mathrm{Id}_{p+1} - A$ is invertible. Note that this is always the case if G is acyclic.

We consider inferring the structural equation for just one of the variables and we take variable X_{p+1} without loss of generality and denote it by Y. Note that Y can be in the parental set of some (or all) of the variables $X_1, \ldots, X_p$, that is, the matrix A is not necessarily lower triangular. With slight abuse of notation, we define $X := (X_1, \ldots, X_p)$, $\beta^0 := A_{p+1,1:p}$ and $\varepsilon := \eta_{p+1}$ such that

$$(2) \qquad Y := X_{p+1} = \sum_{k=1}^{p} \beta_k^0 X_k + \varepsilon.$$

Note that the vector β^0 has a causal interpretation as it is the coefficient vector $A_{p+1,1:p}$ in the structural equation model (2). The goal is to infer β^0.

1.2. *Relation to other work.* We have mentioned already major differences to invariant causal prediction (Peters, Bühlmann and Meinshausen (2016)) and the loose relation to the vast literature on instrumental variable regression (Didelez, Meng and Sheehan (2010)) which will be detailed in Section 3.6. Another method that relies on shift interventions has been published recently (Rothenhäusler et al. (2015)). However, the authors exploit a different type of invariance as *inner-product invariance* does not hold in this setting. Lewbel (2012) uses heteroscedasticity to infer structural equations. While Lewbel (2012) uses cross-products between exogenous variables and error terms to identify structural equations, we directly exploit the covariance structure of endogenous variables and the error terms, resulting in a different method. The comparison in Figure 11 about an application has been published in Meinshausen et al. (2016). The concept of *inner-product invariance*, the *causal Dantzig* method and all its corresponding theory are entirely novel.

1.3. *Overview.* In Section 2, we introduce the notion of *inner-product invariance* and discuss under which assumptions this property is satisfied. In Section 3, we leverage this property to define the unregularized *causal Dantzig* and discuss identifiability, low-dimensional estimation and inference. Furthermore, in the case of nonidentifiability we show that the solution set of *causal Dantzig* has predictive guarantees under certain interventions. We conclude with a comparison to instrumental variable regression and a discussion of *inner-product invariance* from the perspective of potential outcomes. In Section 4, we introduce the regularized *causal Dantzig*, examine its performance in high-dimensional estimation and show how it can achieve consistency under relaxed identifiability assumptions. Practical considerations for both the regularized and unregularized *causal Dantzig* can be found in Section 5. Numerical examples can be found in Section 6.

2. Conditional and inner-product invariance. In analogy to the setting of Peters, Bühlmann and Meinshausen (2016), we assume that the data are recorded under different discrete environments or experimental conditions $e \in \mathcal{E}$. The random variable X in environment $e \in \mathcal{E}$ is denoted by X^e and the distribution of η by η^e. We observe i.i.d. samples of (X^e, Y^e) from each environment $e \in \mathcal{E}$ and for each sample i we observe from which environment $e_i \in \mathcal{E}$ it was drawn. This variable e_i can be deterministic or random.

The distribution of a variable can be different across environments due to specific or nonspecific interventions. A change in the distribution of X^e, η^e can be caused by different intervention mechanisms such as do-interventions or noise-interventions, which can be randomized or not and known or partially known or unknown.

The type of intervention that generates the environments is arbitrary in Peters, Bühlmann and Meinshausen (2016) with the exception that interventions on the target Y itself are not allowed. The same requirement is also necessary for the instrumental variable approach and we will keep this requirement in the following. For possible relaxations, see Rothenhäusler et al. (2015). Throughout the paper, we assume that the distributions (X^e, Y^e) are nondegenerate and that the Gram matrix of (X^e, Y^e) is well defined and positive definite for all $e \in \mathcal{E}$.

2.1. *Conditional invariance.* The conditional distribution of the target variable Y, given its parents $\mathrm{pa}(Y) = \mathrm{pa}(X_{p+1})$ is denoted by

$$Y^e | X^e_{\mathrm{pa}(Y)} = x.$$

It was assumed in Peters, Bühlmann and Meinshausen (2016) that the conditional distribution is invariant for all $x \in \mathbb{R}^{|\mathrm{pa}(Y)|}$ where it is defined in the absence of hidden confounding (where absence of hidden confounding is fulfilled in (1) if all components of η are independent). It then holds for all environments $e, f \in \mathcal{E}$ and all $x \in \mathbb{R}^{|\mathrm{pa}(Y)|}$ for which the conditional distributions are well defined that

$$(3) \qquad Y^e | X^e_{\mathrm{pa}(Y)} = x \overset{d}{=} Y^f | X^f_{\mathrm{pa}(Y)} = x.$$

This conditional invariance under the true parental set pa(Y) is then exploited for inference by testing for all subsets of $\{1, \ldots, p\}$ whether the invariance of (3) can be rejected. The intersection of all subsets for which invariance cannot be rejected is then automatically a subset of the true parental set with controllable probability.

There are two shortcomings of this invariance approach (Peters, Bühlmann and Meinshausen (2016)) in certain contexts:

(i) The invariance (3) becomes invalid under hidden confounding between Y and the parents of Y as the conditional invariance of (3) can be violated even for the true parental set (Peters, Bühlmann and Meinshausen (2016)).

(ii) Testing each subset of $\{1, \ldots, p\}$ restricts the number of variables to somewhere between $p \leq 20$ in practice.

Both of these shortcomings can be addressed when using a different type of invariance.

2.2. *Inner-product invariance.*   We show in the following that the invariance of the *conditional distribution* (3) can be replaced with an *inner-product invariance* under a more specific assumption on the mechanism that generates the different environments.

DEFINITION 1.   *Inner-product invariance* under $\beta^0 \in \mathbb{R}^p$ is fulfilled iff

$$\mathbb{E}[X_k^e(Y^e - X^e\beta^0)] = \mathbb{E}[X_k^f(Y^f - X^f\beta^0)]$$

for all $e, f \in \mathcal{E}$ and $k \in \{1, \ldots, p\}$.

We will show that *inner-product invariance* is true for the causal vector β^0 under the assumption of additive interventions made precise in the following. A derivation of this result from potential outcome assumptions is discussed in Section 3.7. The concept of *inner-product invariance* will then later be exploited for computationally fast causal inference for both low- and high-dimensional data.

2.3. *Additive interventions.*   We assume here that the structural equations (1) are constant across all environments and that the change in the distribution of X^e between environments is caused by a shift in the distribution of η^e between different environments.

ASSUMPTION 1.   Assume that the distributions of $(X_1^e, \ldots, X_{p+1}^e)$, $e \in \mathcal{E}$, are generated by the linear SEM

$$X_k^e \leftarrow \sum_{k' \neq k} A_{k,k'} X_{k'}^e + \eta_k^e, \qquad \text{for } k = 1, \ldots, p + 1 \text{ and } e \in \mathcal{E}.$$

Assume that there exist random variables $\eta^0, \delta^e \in \mathbb{R}^p$ with $\mathrm{Cov}(\eta^0, \delta^e) = 0$ for all $e \in \mathcal{E}$ such that η^e can be written as

$$\eta^e \overset{d}{=} \eta^0 + \delta^e \qquad \text{for all } e \in \mathcal{E}.$$

We assume that $\delta_{p+1}^e \equiv 0$ for all $e \in \mathcal{E}$ and $\mathbb{E}[\eta^0] = 0$.

Note that the components of η^0 and of each vector δ^e, $e \in \mathcal{E}$ are allowed to be dependent to allow for hidden confounding. We call the random variables δ^e, $e \in \mathcal{E}$, *additive interventions* as they are additive and specific to the environment $e \in \mathcal{E}$. δ^e can for example be an additive contribution if $\mathbb{E}(\delta_k^e) \neq 0$ for some variable $k \in \{1, \ldots, p\}$ or a noise contribution if $\mathrm{Var}(\delta_k^e) \neq 0$ or both. If $\delta_k^e \equiv 0$ for some $e \in \mathcal{E}$ and $k \in \{1, \ldots, p\}$, we say that there is *no intervention on variable k in environment $e \in \mathcal{E}$*. The last part of the assumption ensures that the noise part δ^e that is specific to environment $e \in \mathcal{E}$ does not include an intervention on the target variable Y itself and is a type of exclusion restriction (Pearl (2009)). Mathematically, the crucial property of Assumption 1 is that the covariance between the error of covariates and target variable is constant, that is, that $\mathrm{Cov}(\eta_{1:p}^e, \eta_{p+1}^e)$ is constant across environments $e \in \mathcal{E}$. This allows us to obtain the following result.

PROPOSITION 1. *Under Assumption 1, we have inner-product invariance under the true causal coefficients $\beta^0 = (A_{p+1,k})_{k=1,\ldots,p}$:*

$$\mathbb{E}[X_k^e(Y^e - X^e \beta^0)] = \mathbb{E}[X_k^f(Y^f - X^f \beta^0)]$$

for all $e, f \in \mathcal{E}$ and $k \in \{1, \ldots, p\}$.

The proof of this result can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). A derivation of this result from potential outcome assumptions is discussed in Section 3.7. We will exploit inner-product invariance to infer the causal effects in linear SEMs in the following.

2.4. *Errors-in-variables.* In many real-world applications, we cannot directly observe $X_1, \ldots, X_p, Y$, but make a measurement error ζ when observing it. In other words, we measure

$$(4) \qquad \tilde{Y}^e = Y^e + \zeta_y^e \quad \text{and} \quad \tilde{X}_k^e = X_k^e + \zeta_k^e, \qquad e \in \mathcal{E}, k = 1, \ldots, p,$$

where $\zeta_y^e, \zeta_k^e, e \in \mathcal{E}, k = 1, \ldots, p$ are centered, jointly independent and independent of $X^e, Y^e, e \in \mathcal{E}$ with finite variance. Furthermore, we make the assumption that the distributions of $\zeta_k^e, k = 1, \ldots, p$ are invariant for different settings $e \in \mathcal{E}$. Note that we do not assume that the distribution of ζ_y^e is invariant for different settings $e \in \mathcal{E}$. Errors-in-variables exhibit an effect called "regression dilution" or

"attenuation." As an example, consider a structural equation model of the following form:

$$\text{latent variables } X_1 \text{ and } Y \text{ with } Y = 2X_1 + \varepsilon,$$

$$\text{observed variables } \tilde{X}_1 \text{ and } \tilde{Y} \text{ with } \tilde{X}_1 = X_1 + \zeta_1,$$

$$\text{and} \quad \tilde{Y} = Y + \zeta_y.$$

For now, let us assume that there is no confounding between X_1 and Y and X_1. When regressing $\tilde{Y}$ on $\tilde{X}_1$, we obtain a smaller regression coefficient than when regressing Y on X_1 due to higher variance of $\tilde{X}_1$. The smaller regression coefficient is by definition the best linear prediction of $\tilde{Y}$ given $\tilde{X}_1$. In this sense, attenuation can be ignored if one wants to make predictions based on $\tilde{X}_1$. However, in causal inference we are interested in knowing what happens when *intervening on X_1*, and this effect would be underestimated by the regressing $\tilde{Y}$ on $\tilde{X}_1$. The following proposition shows that if inner-product invariance holds for $X_1, \ldots, X_p, Y$ then it also holds for proxy variables $\tilde{X}_1, \ldots, \tilde{X}_p, \tilde{Y}$.

PROPOSITION 2. *Assume inner-product invariance holds for variables X_1^e, $\ldots, X_p^e, Y^e$, $e \in \mathcal{E}$, under β^0. Assume we have an errors-in-variables model as defined in equation* (4). *Then inner-product invariance holds for $\tilde{X}_1^e, \ldots, \tilde{X}_p^e, \tilde{Y}^e$, $e \in \mathcal{E}$ under β^0:*

$$\mathbb{E}\big[\tilde{X}_k^e(\tilde{Y}^e - \tilde{X}^e \beta^0)\big] = \mathbb{E}\big[\tilde{X}_k^f(\tilde{Y}^f - \tilde{X}^f \beta^0)\big]$$

for all e, $f \in \mathcal{E}$ and $k \in \{1, \ldots, p\}$.

The proof of this result can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). As a result, methods based on inner-product invariance will be robust with respect to errors-in-variables. Note that the analogous statement is true for instrumental variable regression. Now let us turn to the definition of the unregularized *causal Dantzig*.

3. Causal Dantzig without regularization. In this section, we introduce the unregularized *causal Dantzig*, discuss its basic properties and an example. We introduce the unregularized *causal Dantzig* in Section 3.1. Asymptotic confidence intervals for low-dimensional estimation are discussed in Section 3.3. Section 3.4 provides an example and explains basic usage of the method `causalDantzig` in the R-package `InvariantCausalPrediction` (R Core Team (2017)). Identifiability and consistency issues are discussed in Section 3.5. We conclude with a comparison to instrumental variable regression in Section 3.6.

3.1. *The estimator.* Assume that we observe i.i.d. samples of (X^e, Y^e) in two environments $e \in \mathcal{E} = \{1, 2\}$ with n_1, n_2 samples in each environment. Let $\mathbf{X}^1$ and $\mathbf{X}^2$ be the $n_1 \times p$ and $n_2 \times p$-dimensional matrices that contain the realized values of the random variables X^e in environment $e = 1$ and $e = 2$, respectively, and let $\mathbf{Y}^1 \in \mathbb{R}^{n_1}$ and $\mathbf{Y}^2 \in \mathbb{R}^{n_2}$ be the respective measurements of the response variables. Define the differences between the two environments in inner-product and Gram matrices, the so-called *Gram-shift matrices*

(5)
$$\hat{\mathbf{Z}} := \frac{1}{n_1}(\mathbf{X}^1)^t \mathbf{Y}^1 - \frac{1}{n_2}(\mathbf{X}^2)^t \mathbf{Y}^2 \in \mathbb{R}^p$$

$$\hat{\mathbf{G}} := \frac{1}{n_1}(\mathbf{X}^1)^t \mathbf{X}^1 - \frac{1}{n_2}(\mathbf{X}^2)^t \mathbf{X}^2 \in \mathbb{R}^{p \times p}.$$

Assuming inner-product invariance holds under β^0,

$$\mathbb{E}[\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0] = 0.$$

A simple estimator of β^0 is the empirical minimizer of the ℓ_∞-norm of the differences between $\hat{\mathbf{Z}}$ and $\hat{\mathbf{G}}\beta$.

DEFINITION 2 (Unregularized causal Dantzig). The causal Dantzig estimator $\hat{\beta}$ is defined as a solution to the optimization problem

(6)
$$\min_{\beta \in \mathbb{R}^p} \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta\|_\infty.$$

The choice of how to center and scale variables deserves some attention. We will discuss this in Section 5.1. Causal Dantzig is uniquely defined if and only if $\hat{\mathbf{G}}$ is invertible and can in this case be written as

(7)
$$\hat{\beta} = \hat{\mathbf{G}}^{-1}\hat{\mathbf{Z}}.$$

Note that by equation (5) this estimator is closely related to least squares in linear regression. Recall that for observations $\mathbf{Y} \in \mathbb{R}^n$ and design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the least squares estimator is defined as

$$\hat{\beta}_{LS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}.$$

Causal Dantzig is strikingly similar, with the Gram matrices replaced by differences of Gram matrices in different settings. As such, it is straightforward to derive asymptotic confidence intervals for this estimator. Many properties from linear regression do not carry over. For example, the causal Dantzig is only asymptotically unbiased.

3.2. *More than two environments.* There are two straightforward extensions to more than two environments $|\mathcal{E}| > 2$. Pooling data from different environments preserves inner-product invariance. If some of the environments are "observational" and the others are "interventional," one option for splitting the data into two environments $\mathcal{E}' = \{1, 2\}$ is pooling all observational data ($e' = 1$) and pooling all interventional data ($e' = 2$). Instead of splitting the data into two environments one can change the definition of the estimator to accommodate for more than two settings, for example, by defining $\hat{\beta}$ as a solution to the optimization problem

$$(8) \qquad \min_{\beta \in \mathbb{R}^p} \max_{e \in \mathcal{E}} \|\hat{\mathbf{Z}}^e - \hat{\mathbf{G}}^e \beta\|_\infty,$$

where

$$
\hat{\mathbf{Z}}^e := \frac{1}{n_e}(\mathbf{X}^e)^t \mathbf{Y}^e - \frac{1}{|\mathcal{E}| - 1} \sum_{\tilde{e} \neq e} \frac{1}{n_{\tilde{e}}} (\mathbf{X}^{\tilde{e}})^t \mathbf{Y}^{\tilde{e}} \in \mathbb{R}^p,
$$

$$(9)$$

$$
\hat{\mathbf{G}}^e := \frac{1}{n_e}(\mathbf{X}^e)^t \mathbf{X}^e - \frac{1}{|\mathcal{E}| - 1} \sum_{\tilde{e} \neq e} \frac{1}{n_{\tilde{e}}} (\mathbf{X}^{\tilde{e}})^t \mathbf{X}^{\tilde{e}} \in \mathbb{R}^{p \times p}.
$$

Note that for two environments, solutions of equation (8) coincide with equation (6). It depends on the type of interventions and the signal strength which of the two options mentioned above is better. If the data can be split into two environments $\mathcal{E}' = \{1, 2\}$ that are homogeneous, doing so is preferable as the estimators of $\mathbf{G}^{e'}$ and $\mathbf{Z}^{e'}$, $e' \in \mathcal{E}'$ have low variance. If the environments $\mathcal{E}$ have different (strong) interventions, solving equation (8) can be preferable as the effect of several strong interventions might get "washed out" when averaging over many environments. We will return later to the case of more than two environments. For the following discussion, we assume that there are two environments $\mathcal{E} = \{1, 2\}$.

3.3. *Confidence intervals.* In the settings described above, $\hat{\beta}$ is in general only asymptotically unbiased. This bias is unknown as it depends on the unknown amount of confounding between X^e and Y^e. Hence we will only pursue asymptotic confidence intervals. We will show that the estimator (7) is under certain conditions asymptotically normally distributed, that is, for $n_1, n_2 \to \infty$,

$$(10) \qquad \left(\frac{V^1}{n_1} + \frac{V^2}{n_2}\right)^{-\frac{1}{2}} (\hat{\beta} - \beta^0) \rightharpoonup \mathcal{N}_p(0, \mathrm{Id}_p).$$

The matrices V^1 and V^2 are positive definite under suitable assumptions and can be consistently estimated from the data as $\hat{V}^1$ and $\hat{V}^2$ as we will discuss later. We can then define asymptotically valid confidence intervals for β_k^0 as

$$(11) \qquad I_k = [\hat{\beta}_k - q\sqrt{\hat{V}_{kk}}, \hat{\beta}_k + q\sqrt{\hat{V}_{kk}}],$$

where $\hat{V}_{kk}$ is the kth diagonal element of $\hat{V} = \hat{V}^1/n_1 + \hat{V}^2/n_2$ and $q = \Phi^{-1}(1 - \alpha/2)$. Here, Φ denotes the distribution function of a standard Gaussian random variable. The interval I_k has asymptotic coverage

$$\mathbb{P}[\beta_k^0 \in I_k] \to 1 - \alpha \qquad \text{for } n_1, n_2 \to \infty.$$

The conditions for asymptotic normality (10) are fourth-moment conditions on the observed random variables as well as conditions that guarantee that V^1 and V^2 are invertible and that *causal Dantzig* is unique.

THEOREM 1 (Asymptotic normality).   *Let (X^1, Y^1) and (X^2, Y^2) have finite fourth moments and assume that inner product invariance holds under β^0. Assume that $(\mathbf{X}^1, \mathbf{Y}^1)$ and $(\mathbf{X}^2, \mathbf{Y}^2)$ are independent. Define $\mathbf{G} := \mathbb{E}[\hat{\mathbf{G}}]$ and $\mathbf{Z} := \mathbb{E}[\hat{\mathbf{Z}}]$ and let $\mathbf{G}$ and the covariance matrix of $X^e \eta_{p+1}^e$, $e \in \mathcal{E}$ be invertible. For $n_1, n_2 \to \infty$,*

$$\left(\frac{V^1}{n_1} + \frac{V^2}{n_2}\right)^{-\frac{1}{2}} (\hat{\beta} - \beta^0) \rightharpoonup \mathcal{N}(0, \mathrm{Id}_p),$$

where $V^e := \mathrm{Cov}(\mathbf{G}^{-1}(X^e)^t \eta_{p+1}^e)$, $e \in \{1, 2\}$ are invertible. Note that we allow n_1 and n_2 to have different asymptotic growth rates.

REMARK 1 (Estimation of V^1 and V^2).   The empirical covariance matrix of

$$-\hat{\mathbf{G}}^{-1}(\mathbf{X}_{i\cdot}^1)^t \mathbf{X}_{i\cdot}^1 \hat{\mathbf{G}}^{-1}\hat{\mathbf{Z}} + \hat{\mathbf{G}}^{-1}(\mathbf{X}_{i\cdot}^1)^t \mathbf{Y}_i^1, \qquad i = 1, \dots, n_1,$$

is a consistent estimator of V^1. V^2 can be estimated analogously.

The proof of this result can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). The assumption that $\mathbf{G}$ is invertible will be discussed further in Section 3.5. In Section 4, we will discuss how the regularized *causal Dantzig* can be consistent in some situations where population $\mathbf{G}$ is not invertible. Asymptotic efficiency is discussed in Section A.5 in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

3.4. *Implementation and example.*   We use data generated according to a SEM with the structure given by Figure 1 as an example. Suppose the data are generated in two environments $\{1, 2\} = \mathcal{E}$ according to

(12)
$$\begin{cases} X_2^e & \leftarrow & \eta^0 + \sigma^e \eta_2 \\ Y^e & \leftarrow & X_2^e + \eta^0 + \eta_y \\ X_1^e & \leftarrow & Y^e + X_2^e + \sigma^e \eta_1 \\ X_3^e & \leftarrow & X_1^e + \eta^0 + \sigma^e \eta_3 \end{cases},$$

where $(\eta^0, \eta_y, \eta_1, \eta_2, \eta_3)$ is assumed to be drawn from $\mathcal{N}_5(0, \mathrm{Id}_5)$ and the noise variances are $\sigma^e = 1$ for environment $e = 1$ and $\sigma^e = 4$ for environment $e = 2$. We
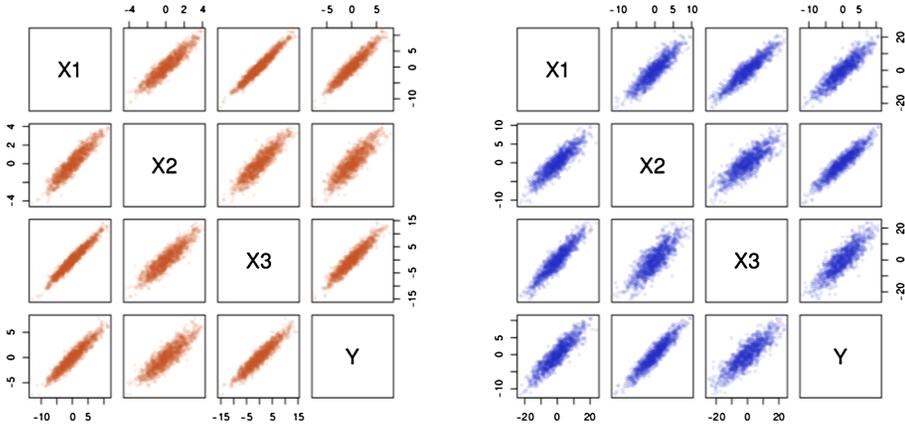
FIG. 2.   *The scatterplot of the variables in the graph of Figure* 1 *and SEM* (12) *for environment* 1 (*red, left panel*) *and environment* 2 (*blue, right panel*). *The estimate is based on the difference in the two Gram matrices.*

draw 1000 i.i.d. samples from each environment and the corresponding pairwise scatterplots are shown in Figure 2. For one realization, we obtain the estimate $\hat{\beta}$ via the difference in Gram matrices $\mathbf{G}$ and inner products with the target $\mathbf{Z}$ as

$$\hat{\mathbf{G}} = \begin{pmatrix} 15.9 & 6.5 & 16.1 \\ 6.5 & 3.2 & 6.5 \\ 16.1 & 6.5 & 19.1 \end{pmatrix}, \qquad \hat{\mathbf{Z}} = \begin{pmatrix} 6.4 \\ 3.2 \\ 6.5 \end{pmatrix} \quad \Rightarrow \quad \hat{\beta} = \hat{\mathbf{G}}^{-1}\hat{\mathbf{Z}} = \begin{pmatrix} -0.04 \\ 1.00 \\ 0.03 \end{pmatrix},$$

where the correct vector of causal coefficients in this problem is

$$\beta^0 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Asymptotic confidence intervals can be computed via (11).

The procedure is implemented as method `causalDantzig` in R-package `InvariantCausalPrediction` (R Core Team (2017)). The output for the example above is shown below, where X is the matrix with predictor variables, Y the outcome of interest and E is an n-dimensional vector with entries 1 for samples from environment $e = 1$ and entries 2 for samples from environment $e = 2$.

```
> fit <- causalDantzig(X,Y,E,regularization=FALSE)
> print(fit)
Unregularized causal Dantzig
Call:
causalDantzig(X = X, Y = Y, E = E, regularization = FALSE)

    Estimate StdErr p.value
X1   -0.042  0.059   0.481
X2    0.999  0.106  <2e-16 ***
```

```
X3     0.035  0.042   0.403
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
Only the direct causal effect of the second variable turns out to be statistically significant. Note that in this setting, instrumental variables regression would fail. One problem is that the number of covariates exceeds the number of "instruments." Additionally, the expectation of X^1 and X^2 are equal, implying that there is no mean shift due to the two environments. We will discuss these issues in more detail in Section 3.6.

3.5. *Identifiability of β^0 and practical implications.* In the simplest setting, the number of samples greatly exceeds the number of parameters, and the interventions δ^e, $e \in \mathcal{E}$ are sufficiently different to make the parameter β^0 identifiable. Theorem 2 gives conditions under which this is the case.

THEOREM 2. *Consider a SEM that satisfies Assumption 1. Assume that there exists an "observational" environment, that is, an environment $e \in \mathcal{E}$ with $\delta^e \equiv 0$. Furthermore assume that all interventions δ^e are full-rank on its support, that is, that the Gram matrix of $\delta_{S^e}^e$ is positive definite for $S^e = \{k : \delta_k^e \not\equiv 0\}$:*

1. *The causal coefficient is identifiable in the population case if and only if for each $k = 1, \ldots, p$ there exists $e \in \mathcal{E}$ such that $\delta_k^e \not\equiv 0$.*
2. *If the condition in 1. holds then the solution of* causal Dantzig *as defined in equation* (8) *is unique in the population case and equal to β^0.*

The proof of this result can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). Usually there are many different SEMs satisfying Assumption 1 that can generate a given observed distribution of (X^e, Y^e), $e \in \mathcal{E}$. Theorem 2 gives a condition under which these SEMs all share the same direct causal effect β^0 from $X_1, \ldots, X_p$ to Y. If said condition is satisfied, the *causal Dantzig* has a unique solution in the population case that is equal to β^0. Furthermore, it tells us that if this condition is not satisfied, there exist at least two SEMs satisfying Assumption 1 with different direct causal effects from $X_1, \ldots, X_p$ to Y that generate the given distribution. Without further assumptions, it is then not possible to consistently estimate the direct causal effects, but only a set of potential causal effects. We will characterize this set later.

Note that Theorem 2 describes a rather strong condition for identifiability. Especially if p is large it might be unrealistic to have nonzero interventions δ_k^e on each of the variables X_k, $k = 1, \ldots, p$. However, making additional assumptions can help resolve these identifiability issues. If the interventions δ_k^e only act on a subset of the variables $X_1, \ldots, X_p$ or when the number of covariates exceeds the sample size $p > n$, the regularized *causal Dantzig* can be consistent under the additional assumption of sparsity. We discuss consistency of the regularized *causal Dantzig*

1700 OCR

in such scenarios in Section 4.2 and Section 4.3. Alternatively, it can be advisable to first run LASSO on the pooled dataset to select a subset of the variables. Under the assumption of faithfulness, it is sufficient to have nonzero interventions on the selected subset. Some justification for this approach can be found in Section 5.3.

If the assumptions for identifiability of β^0 are not fulfilled, it should still be possible to guarantee predictive performance under certain new environments. The following theorem makes this intuition more precise. The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

THEOREM 3. *Consider a SEM that satisfies Assumption 1. Assume that there exists an "observational" environment, that is, an environment $e \in \mathcal{E}$ with $\delta^e \equiv 0$. Furthermore, assume that all interventions δ^e are full-rank on its support, that is, that the Gram matrix of $\delta^e_{S^e}$ is positive definite for $S^e = \{k : \delta^e_k \not\equiv 0\}$. Let β be a solution of causal Dantzig as defined in equation (8) in the population case.*

1. *Then the distribution of the residuals is invariant, that is,*

$$Y^e - X^e \beta \overset{d}{=} Y^f - X^f \beta \qquad \text{for all } e, f \in \mathcal{E}.$$

2. *For a new environment $\tilde{e} \notin \mathcal{E}$ that satisfies Assumption 1 for (X^e, Y^e), $e \in \mathcal{E} \cup \{\tilde{e}\}$ with $\{k : \delta^{\tilde{e}}_k \not\equiv 0\} \subset \bigcup_{e \in \mathcal{E}} S^e$, we have*

$$Y^e - X^e \beta \overset{d}{=} Y^{\tilde{e}} - X^{\tilde{e}} \beta \qquad \text{for all } e \in \mathcal{E}.$$

In words, solutions of *causal Dantzig* guarantee that the residuals have the same distribution across all environments $e \in \mathcal{E}$. Perhaps more importantly, solutions of *causal Dantzig* are guaranteed to have the same predictive performance on new environments $\tilde{e} \notin \mathcal{E}$ with arbitrary large additive perturbations $\delta^{\tilde{e}}_k$ as long as these perturbations act on a subset of the variables $\bigcup_{e \in \mathcal{E}} S^e$.

3.6. *Comparison with instrumental variables.* Consider a setting where the underlying DAG takes the following form:
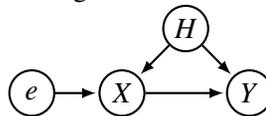
We assume that H is not observed and that e takes values in $\{1, 2\}$. To be able to use the *causal Dantzig*, we have to define settings $\mathcal{E}$. It is rather straightforward to write (X^1, Y^1) for the variables (X, Y) conditioned on $e = 1$ and (X^2, Y^2) for the variables (X, Y) conditioned on $e = 2$. As e is binary, the method of instrumental variables (IV) coincides with the Wald estimator (Wald (1940)). In the population case, it can be written as

$$(13) \qquad \lim_{n \to \infty} \hat{\beta}_{\text{IV}} = \frac{\mathbb{E}[Y|e=1] - \mathbb{E}[Y|e=2]}{\mathbb{E}[X|e=1] - \mathbb{E}[X|e=2]} = \frac{\mathbb{E}[Y^1] - \mathbb{E}[Y^2]}{\mathbb{E}[X^1] - \mathbb{E}[X^2]}.$$

Consistency of the causal Dantzig *and the instrumental variables approach. Consider a model* $Y = \beta X + H + \eta_y$ *and a structural equation model for X as depicted in the table. The case on the left is a mean-shift, whereas on the right-hand side the error variance changes between setting* $e = 1$ *and* $e = 2$. *We assume* $\alpha \neq 0$, *and that the random variables* e, η_y, η_x, H *are independent and nondegenerate with* $\mathbb{E}[H] = 0$

Consistency	$X = \alpha e + H + \eta_x$ (mean-shift)	$X = H + (1 + \alpha e)\eta_x$ (change in error distribution)
Instrumental variable regression	Yes	No
Unregularized causal Dantzig	Yes	Yes

Causal Dantzig leads to

$$(14) \qquad \lim_{n \to \infty} \hat{\beta} = \frac{\mathbb{E}[X^1 \cdot Y^1] - \mathbb{E}[X^2 \cdot Y^2]}{\mathbb{E}[(X^1)^2] - \mathbb{E}[(X^2)^2]}.$$

Both the IV approach and the *causal Dantzig* have different strengths and weaknesses in this setting. For example, equation (13) is based on means, whereas equation (14) is based on covariances. If, say, $X = e \cdot \eta_x + H$, $Y = \beta X + H + \eta_y$, with centered noise η_x, η_y independent of the centered confounder H, then $\mathbb{E}[X|e = 1] = \mathbb{E}[X|e = 2]$. Hence the IV estimator is not well defined in the population case and one should use the causal Dantzig. If the instrument is weak, causal Dantzig can exhibit efficiency gains. An example of this can be found in Section 6.2. A more general comparison can be found in Table 1. It is also possible to construct examples where equation (14) is not well defined. For this to happen, the second moments of X^1 and X^2 have to be equal.

A drawback of the IV approach is that the number of instruments has to equal or exceed the number of endogenous variables. However, this is not necessary for the *causal Dantzig*. Two settings $|\mathcal{E}| = 2$ in our framework correspond to a single binary exogenous variable. In that case, the number of endogenous variables p can be arbitrarily large as long as $\mathbf{G}$, the difference of Gram matrices, is invertible. On the other hand, for $p > 2$ the number of endogenous variables exceeds the number of exogenous variables and the IV approach is bound to fail. We compare the performance of the IV approach and *causal Dantzig* on simulated datasets in Section 6.2.

3.7. *Inner-product invariance in the potential outcome framework.* In this section, we will investigate the notion of *inner-product invariance* under potential outcome assumptions (Neyman (1923), Rubin (1974)). Note that here, as in the rest of the paper, we consider a continuous exposure $X \in \mathbb{R}^p$. In the following, we use a slightly different notation compared to the rest of the paper. We write $X(e) \in \mathbb{R}^p$ for the potential outcome of a continuous exposure if the environment

E takes value $e \in \mathcal{E}$. Equivalently, we write $Y(x, e) \in \mathbb{R}$ for the potential outcome of the response of a unit if the exposure takes level $X = x$ and environment E takes value $e \in \mathcal{E}$. We assume that these quantities are well defined. We make the following additional assumptions:

A1. Exclusion restriction:

$$Y(x, e) = Y(x) \qquad \text{holds for all } x \in \text{range}(X) \text{ and } e \in \mathcal{E}.$$

A2. Independence:

$$\big(X(e), Y(0)\big) \perp\!\!\!\perp E \qquad \text{for all } e \in \mathcal{E}.$$

A3. Constant confounding across environments $\mathcal{E}$:

$$\text{Cov}\big(X(e), Y(0)\big) = \text{Cov}\big(X(f), Y(0)\big) \qquad \text{for all } e, f \in \mathcal{E}.$$

A4. Treatment effect homogeneity and linearity:

$$\mathbb{E}\big[Y(x) - Y(0)|X = x, E = e\big] = \mathbb{E}\big[Y(x) - Y(0)\big]$$
$$= x\beta^0 \qquad \text{for all } x \in \text{range}(X) \text{ and } e \in \mathcal{E}.$$

A5. The variables are normalized:

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{E}[Y] = 0.$$

Note that we did not make any cross-world assumptions (Richardson and Robins (2013)), that is, we made no assumptions on the joint distribution of $Y(x)$, $x \in \text{range}(X)$ or on the joint distribution of $X(e)$, $e \in \mathcal{E}$. Condition (A2) can be relaxed to an assumption on the cross-product between $X(e)$ and $Y(0)$. Details can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)) in the proof of Proposition 3. Condition (A3) is crucial: we allow for confounding (nonzero covariance of $X(e)$ and $Y(0)$), but we assume that the covariance is constant across environments. Loosely speaking, this can be seen as a noninteraction-assumption of environment and confounding. Condition (A4) ensures that the average treatment effect is the same within strata defined by X and E and allows the usage of a linear model. For a discussion of similar assumptions in the context of the IV framework, see Wang and Tchetgen Tchetgen (2018).

If these assumptions are fulfilled, then we have *inner-product invariance* under the average treatment effect β^0.

PROPOSITION 3.   *Under assumptions* (A1)–(A5), *we have* inner-product invariance *under the vector $\beta^0 \in \mathbb{R}^p$ which satisfies $\mathbb{E}[Y(x) - Y(0)] = x\beta^0$, that is,*

$$\mathbb{E}\big[X^t(Y - X\beta^0)|E = e\big] = \mathbb{E}\big[X^t(Y - X\beta^0)|E = f\big] \qquad \text{for all } e, f \in \mathcal{E}.$$

The proof of this result can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). Using inner-product invariance for estimating the average treatment effect β^0, it is possible to consistently estimate the average treatment effect in cases in which two-stage least squares (or the Wald estimand) is degenerate. For example, in settings where the dimension of exposure variables X exceeds the number of environments $|E|$ or when $\mathbb{E}[Y - Xb|E = 1] = \mathbb{E}[Y - Xb|E = 0]$ for $\mathcal{E} = \{0, 1\}$. In the presence of weak instruments, *causal Dantzig* can exhibit efficiency gains compared to estimators based on conditional means of X and Y. This is investigated further in Section 3.6 and Section 6.

4. Causal Dantzig with regularization.

In this section, we introduce the regularized *causal Dantzig*, and discuss its theoretical properties. The estimator is motivated and introduced in Section 4.1. Section 4.2 contains finite sample bounds. The bounds presented in this section involve a quantity that we call the "causal cone invertibility factor." The behavior of this quantity is discussed in Section 4.3.

4.1. *The estimator.* Weak interventions on some of the variables (i.e., $\mathbb{E}[(\delta_k^e)^2]$ small) may lead to coefficient estimates with high variance in equation (7). Furthermore, if the number of predictors p exceeds the total sample size n, the matrix $\hat{\mathbf{G}}$ is not invertible and the solution to equation (6) is not unique. In such settings, regularization and shrinkage is desirable and can outperform unpenalized estimation procedures; see, for example, Bühlmann and van de Geer (2011). In particular, ℓ_1-penalized estimation procedures have attracted much interest in high-dimensional models. For linear models, Candes and Tao (2007) proposed an ℓ_1-minimization method called the Dantzig selector. Consider $\mathbf{Y} = \mathbf{X}\beta^* + \epsilon$ with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$. For a tuning parameter $\lambda \geq 0$, the Dantzig selector is defined as a solution to the regularization problem

$$
(15) \quad \min \|\beta\|_1 \quad \text{subject to} \quad \|\tilde{\mathbf{Z}} - \tilde{\mathbf{G}}\beta\|_\infty \leq \lambda,
$$
$$
\text{where} \quad \tilde{\mathbf{Z}} = \mathbf{X}^t \mathbf{Y}/n \quad \text{and} \quad \tilde{\mathbf{G}} = \mathbf{X}^t \mathbf{X}/n.
$$

The geometry of the Dantzig selector is depicted in Figure 3. The ℓ_1-minimization favors sparse solutions, that is, vectors in which many coefficients are exactly zero. This facilitates interpretation. Furthermore, if λ gets larger, the Dantzig selector shrinks toward the zero vector. Choosing λ is a trade off: small values will generally result in larger variance of the estimator, but smaller bias. We propose the regularized *causal Dantzig* $\hat{\beta}^\lambda$, which in analogy to equation (6) is defined as a
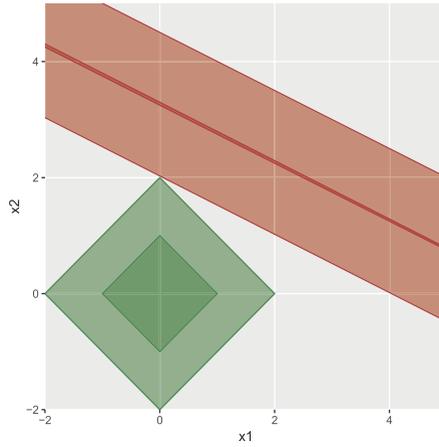
FIG. 3. *A visualization of the Dantzig selector. The red region is the feasible set* $\{\beta : \|\tilde{\mathbf{Z}} - \tilde{\mathbf{G}}\beta)\|_\infty \leq \lambda\}$ *for* $\lambda = \sqrt{1.25}$. *The two green regions are the level sets* $\{\beta : \|\beta\|_1 \leq 1\}$ *and* $\{\beta : \|\beta\|_1 \leq 2\}$. *The Dantzig selector for* $\lambda = \sqrt{1.25}$ *is at the intersection of the light green and the red region.*

solution to

$$\min \|\beta\|_1 \quad \text{subject to} \quad \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta\|_\infty \leq \lambda$$

(16)
$$\text{where } \hat{\mathbf{Z}} = \frac{(\mathbf{X}^1)^t \mathbf{Y}^1}{n_1} - \frac{(\mathbf{X}^2)^t \mathbf{Y}^2}{n_2},$$

$$\text{and } \hat{\mathbf{G}} = \frac{(\mathbf{X}^1)^t \mathbf{X}^1}{n_1} - \frac{(\mathbf{X}^2)^t \mathbf{X}^2}{n_2}.$$

On a superficial level, the difference to the Dantzig selector is merely that $\mathbf{X}^t \mathbf{Y}/n$ is replaced by $(\mathbf{X}^1)^t \mathbf{Y}^1/n_1 - (\mathbf{X}^2)^t \mathbf{Y}^2/n_2$ and $\mathbf{X}^t \mathbf{X}/n$ is replaced by $(\mathbf{X}^1)^t \mathbf{X}^1/n_1 - (\mathbf{X}^2)^t \mathbf{X}^2/n_2$. Hence the geometry of the optimization problem is akin to the Dantzig selector and the *causal Dantzig* inherits its variable selection, shrinkage and regularization properties. Furthermore, the *causal Dantzig* can be cast as a linear program for fixed λ. Details can be found in the Supplementary Material, Section A.3.6 (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

4.2. *Finite-sample bound.* The regularized *causal Dantzig* is related to the Dantzig selector and enjoys similar properties. Notably, it attains the same rates of convergence under comparable regularity conditions. To this end, we introduce the quantity "causal cone invertibility factor," similar to the "cone invertibility factor" for the Dantzig selector as defined in Ye and Zhang (2010). For ease of exposition, we will first treat the case $\mathcal{E} = \{1, 2\}$. The treatment of the general case is sketched in Remark 2.

4.2.1. *Causal cone invertibility factor.* Let $\hat{\boldsymbol{\Sigma}}$ denote the empirical covariance matrix of $\mathbf{X}$ and consider a set $S \subset \{1, \ldots, p\}$. Later we will mainly be interested in the case where S is the active set of β^0. Ye and Zhang (2010) proved bounds for the Dantzig selector that involve the so-called cone invertibility factor (CIF). For the upper bound, the relevant quantity in Ye and Zhang (2010) is $\text{CIF}_q(S)$. Roughly speaking, the cone invertibility factor is a lower bound on the ℓ_∞-norm of $\hat{\boldsymbol{\Sigma}}u$, given that u lies in the cone $\{u : \|u_{S^c}\|_1 \leq \|u_S\|_1\}$ and has unit norm $\|u\|_q = 1$. To make the quantity comparable across different norms, it is scaled by a factor $|S|^{1/q}$. To be more precise,

$$\text{CIF}_q(S) = \inf_u \left\{ \frac{|S|^{1/q} \|\hat{\boldsymbol{\Sigma}}u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}.$$

Now we are ready to define the *causal cone invertibility factor* $\text{CCIF}_q(S, \hat{\mathbf{G}})$:

$$(17) \qquad \text{CCIF}_q(S, \hat{\mathbf{G}}) := \inf_u \left\{ \frac{|S|^{1/q} \|\hat{\mathbf{G}}u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}.$$

Analogously define $\text{CCIF}_q(S, \mathbf{G})$ for $\mathbf{G} := \mathbb{E}[\hat{\mathbf{G}}]$. Here and in the following, notationally we do not treat the case $q = \infty$ separately. Instead, with small abuse of notation we set $|S|^{1/q} := 1$ for $q = \infty$. In the new definition, the positive semidefinite matrix $\hat{\boldsymbol{\Sigma}}$ is replaced by the symmetric matrix $\hat{\mathbf{G}}$. As $\hat{\boldsymbol{\Sigma}}$, the matrix $\hat{\mathbf{G}}$ is not positive definite in high-dimensional settings and even indefinite in general. However, it can be shown that the CCIF behaves similarly to the CIF in several ways. This is further discussed in Section 4.3. For now, let us turn to the finite-sample bound of the *causal Dantzig*.

4.2.2. *Finite sample bound.* The finite-sample results of the *causal Dantzig* are analogous to the Dantzig selector while the issue of identifiability is now addressed by the causal cone invertibility factor $\text{CCIF}_q(S, \hat{\mathbf{G}})$. Similarly, as in Ye and Zhang (2010), define $z^* := \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0\|_\infty$ and let S denote the active set of β^0. The first result is purely algebraic and follows from the definitions of $\text{CCIF}_q(S, \hat{\mathbf{G}})$ and the *causal Dantzig*.

LEMMA 1. *On the event $z^* \leq \lambda$, we have*

$$(18) \qquad \|\hat{\beta}^\lambda - \beta^0\|_q \leq \frac{|S|^{1/q}(\lambda + z^*)}{\text{CCIF}_q(S, \hat{\mathbf{G}})} \leq \frac{2|S|^{1/q}\lambda}{\text{CCIF}_q(S, \hat{\mathbf{G}})} \qquad \text{for all } q \geq 1.$$

The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). There are two terms on the right-hand side in equation (18) that deserve further attention. First, $\text{CCIF}_q(S, \hat{\mathbf{G}})$ is bounded away from zero under certain assumptions, as discussed in Section 4.3. Second, it is

crucial to understand the behavior of $z^* := \|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0\|_\infty$. Using a union bound over the p entries, it can be shown that with high probability, z^* is of the order $\max_{e\in\mathcal{E}} \max(\log(p)/n_e, \sqrt{\log(p)/n_e})$.

LEMMA 2. *Assume that inner-product invariance holds for* $(X^e, Y^e), e \in \{1, 2\}$ *under* β^0. *Assume* $X^1, X^2, \eta^1_{p+1}, \eta^2_{p+1}$ *are centered and multivariate Gaussian. Let* $t \geq 0$. *Then, with probability exceeding* $1 - 4\exp(-t)$,

$$z^* \leq \sigma_\varepsilon \sum_{e\in\{1,2\}} \sigma^e_{\max}\left(\sqrt{\frac{4t + 4\log(p)}{n_e}} + \frac{4t + 4\log(p)}{n_e}\right),$$

where

$$\sigma_\varepsilon := \sqrt{\operatorname{Var}(\eta^e_{p+1})} \quad \text{and} \quad \sigma^e_{\max} := \max_k \sqrt{\operatorname{Var}(X^e_k)}.$$

The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). This result can be extended to situations where (X^1, η^1_{p+1}) and (X^2, η^2_{p+1}) have sub-Gaussian tails; see, for example, exercise 14.3 in Bühlmann and van de Geer (2011). By combining Lemma 1 and Lemma 2 we obtain the following theorem. The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

THEOREM 4. *Let* $\lambda \asymp 5C\sqrt{\log(p)/\min_{e\in\{1,2\}} n_e} \to 0$ *for a constant* $C > 0$ *that satisfies* $\sigma_\varepsilon \cdot \sigma^e_{\max} \leq C < \infty$ *for* $e \in \{1, 2\}$. *Under the assumptions mentioned in Lemma 2,*

$$\|\hat{\beta}^\lambda - \beta^0\|_q \leq \frac{10C}{\operatorname{CCIF}_q(S, \hat{\mathbf{G}})} |S|^{1/q} \sqrt{\frac{\log(p)}{\min_{e\in\{1,2\}} n_e}}$$

with $\mathbb{P} \to 1$ *for* $n_1, n_2, p \to \infty$.

Another consequence of these two lemmata is the screening property of the *causal Dantzig* under a so-called betamin-condition. The short proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

PROPOSITION 4. *Let* $\hat{S}$ *denote the active set of* $\hat{\beta}^\lambda$. *Using the notation of Theorem* 4, *assume that*

$$\min_{k\in S} |\beta^0_k| > \frac{10C}{\operatorname{CCIF}_\infty(S, \hat{\mathbf{G}})} \sqrt{\frac{\log(p)}{\min_{e\in\{1,2\}} n_e}}.$$

Then under the assumptions mentioned in Theorem 4 *for* $q = \infty$, *we have*

$$\mathbb{P}[\hat{S} \supseteq S] \to 1 \qquad \text{for } n_1, n_2, p \to \infty.$$

Note that the convergence rate in Theorem 4 coincides with the usual rate of convergence in high-dimensional linear regression (Ye and Zhang (2010)) under comparable assumptions. For consistency in the ℓ_2 norm in the regression setting, it is required that $|S| \log(p)/n \to 0$, that $\lambda \asymp C\sqrt{\log(p)/n}$ for constant $C > 0$ large enough and that the population quantity $\mathrm{CIF}_2(S)$ is bounded away from zero. In our framework, if $n_1 \asymp n_2$, the assumptions on the asymptotic behavior of $n = n_1 + n_2$, p, $|S|$ and λ stay essentially the same, but $\mathrm{CCIF}_2(S, \hat{\mathbf{G}})$ plays the role of $\mathrm{CIF}_2(S)$. The next section aims to shed some light on the behavior of this quantity.

REMARK 2. The results of this section can be extended to more than two settings $|\mathcal{E}| > 2$. To be more precise, in the general case one can define the regularized *causal Dantzig* as a solution to

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{subject to} \quad \max_{e \in \mathcal{E}} \|\hat{\mathbf{Z}}^e - \hat{\mathbf{G}}^e \beta\|_\infty \le \lambda,$$

where $\hat{\mathbf{Z}}^e, \hat{\mathbf{G}}^e, e \in \mathcal{E}$ are defined as in equation (9). The causal cone invertibility factor is then defined as

$$\mathrm{CCIF}_q\big(S, \{\hat{\mathbf{G}}^e, e \in \mathcal{E}\}\big) := \inf_u \max_{e \in \mathcal{E}} \left\{ \frac{|S|^{1/q} \|\hat{\mathbf{G}}^e u\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \le \|u_S\|_1 \right\}.$$

With this notation, it is straightforward to obtain analogous results to Lemma 1–3, Theorem 4 and Proposition 4.

4.3. *Behavior of the causal cone invertibility factor.* In the preceding section, we showed that the causal cone invertibility factor $\mathrm{CCIF}_q(S, \hat{\mathbf{G}})$ is a crucial quantity to understand the behavior of the regularized *causal Dantzig*. How do we guarantee that this quantity is bounded away from zero? There are two issues that we will treat separately. First, for $p > n = n_1 + n_2$, $\hat{\mathbf{G}}$ is not invertible. Second, the environments might not be sufficiently different to make population version $\mathbf{G}$ invertible. In Section 4.3.1, we will discuss how to relate the empirical causal cone invertibility factor to the population causal cone invertibility factor. In Section 4.3.2, we consider the case where the environments are sufficiently different to make the population version $\mathbf{G}$ invertible. In Section 4.3.3, we examine a setting where the environments are not sufficiently different, that is, where $\mathbf{G}$ is not invertible.

4.3.1. *General properties.* In this section, we discuss how to relate the empirical causal cone invertibility factor $\mathrm{CCIF}_q(S, \hat{\mathbf{G}})$ to the population quantity $\mathrm{CCIF}_q(S, \mathbf{G})$. The following lemma gives a deterministic bound for these quantities. The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)).

LEMMA 3. *Let $q \geq 1$. Then*

$$\left| \text{CCIF}_q(S, \hat{\mathbf{G}}) - \text{CCIF}_q(S, \mathbf{G}) \right| \leq 2|S| \|\hat{\mathbf{G}} - \mathbf{G}\|_\infty,$$

where $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ denotes the matrix max norm.

Hence the problem is reduced to understanding the behavior of $\|\hat{\mathbf{G}} - \mathbf{G}\|_\infty$. Let the rows of $\mathbf{X}^e$ consist of i.i.d. centered multivariate Gaussian random variables for $e \in \{1, 2\}$. It can be shown that with probability at least $1 - 4\exp(-t)$,

$$(19) \qquad \|\hat{\mathbf{G}} - \mathbf{G}\|_\infty \leq \sum_{e \in \{1,2\}} (\sigma_{\max}^e)^2 \left(\sqrt{\frac{4t + 8\log(p)}{n_e}} + \frac{4t + 8\log(p)}{n_e} \right).$$

This result can be extended to situations where X^1 and X^2 have sub-Gaussian tails; see, for example, exercise 14.3 in Bühlmann and van de Geer (2011). Hence by Lemma 3, even if $\hat{\mathbf{G}}$ is not invertible, the quantity in equation (17) is well behaved for $\sqrt{\min(n_1, n_2)} \gg |S|\sqrt{\log(p)}$, in the sense that it is strictly bounded away from zero if the same is true for the population quantity. The latter assumption is nontrivial and depends on the distribution of the interventions $\delta^e, e \in \{1, 2\}$.

4.3.2. *Population $\mathbf{G}$ invertible*. Under the assumptions discussed in Section 4.3.1, $\text{CCIF}_q(S, \hat{\mathbf{G}})$ is bounded away from zero if $\text{CCIF}_q(S, \mathbf{G})$ is bounded away from zero. Hence, the problem is reduced to understanding the population quantity $\text{CCIF}_q(S, \mathbf{G})$. If $\mathbf{G}$ is invertible, then

$$(20) \qquad \begin{aligned} \text{CCIF}_q(S, \mathbf{G}) &\geq \min_u \frac{|S|^{1/q} \|\mathbf{G}u\|_\infty}{\|u\|_q} \\ &= \min_{\|u\|_q = |S|^{1/q}} \|\mathbf{G}u\|_\infty > 0. \end{aligned}$$

As

$$\begin{aligned} \mathbf{G} &= \mathbb{E}\left[(X_{1:p}^1)^t X_{1:p}^1 - (X_{1:p}^2)^t X_{1:p}^2 \right] \\ &= ((\text{Id} - A)^{-1})_{1:p,1:p} \mathbb{E}\left[(\delta_{1:p}^1)^t \delta_{1:p}^1 - (\delta_{1:p}^2)^t \delta_{1:p}^2 \right] ((\text{Id} - A)^{-t})_{1:p,1:p}, \end{aligned}$$

this is a measure of the difference in the intervention strength δ^e between the two settings $e = 1$ and $e = 2$. In this sense, this bound is similar to the discussion in Section 4.3.3. However, the bound fails to capture appropriately what happens if the interventions only act on a subset of the variables $X_i, i = 1, \ldots, p$. In that case, the bound in equation (20) is not useful as $\mathbf{G}$ is not invertible. The next section shows that in some of these settings it is still true that $\text{CCIF}_q(S, \mathbf{G}) > 0$.

4.3.3. *Population **G** not invertible.*   The setting of Section 4.3.2 and the bound in equation (20) are rather restrictive. Consider a situation with a block structure in the Gram matrix, that is, where $\mathbb{E}[X_k^e X_{k'}^e] = 0$ for all $k \leq k_0 < k'$ and $e \in \mathcal{E}$. In this case, there might be no interventions on the variables $\{X_{k'}, k' > k_0\}$, that is, $\delta_{k'}^e \equiv 0$ for all $k' > k_0$. As a result, **G** might not be invertible. However, if $\mathbf{G}_{1:k_0,1:k_0}$ is invertible and $S \subset \{1, \ldots, k_0\}$, then

$$\mathrm{CCIF}_q(S, \mathbf{G}) \geq \inf\left\{ \frac{|S|^{1/q}\|\mathbf{G}_{1:k_0,1:k_0}u_{1:k_0}\|_\infty}{\|u\|_q} : \|u_{S^c}\|_1 \leq \|u_S\|_1 \right\}$$

$$\geq \inf\left\{ \frac{|S|^{1/q}\|\mathbf{G}_{1:k_0,1:k_0}u_{1:k_0}\|_\infty}{2\|u_{1:k_0}\|_q} \right\} > 0.$$

Hence, under the assumptions discussed in Section 4.2.2, the *causal Dantzig* is a consistent estimator for β^0. Generally speaking, the *causal Dantzig* tends to screen out variables that have not been affected by the intervention. In this light it is crucial that the interventions act on the variables in the active set of β^0 directly or indirectly.

5. Practical considerations.   In this section, we discuss practical considerations for the *causal Dantzig*. Recommendations are given for centering and scaling of the variables, choice of the regularization parameter λ and a procedure for pre-selection.

5.1. *Centering and scaling.*   Centering and scaling in the *causal Dantzig* setting is a bit more intricate than in a regression setting. Let $\hat{\mu}^e \in \mathbb{R}^{p+1}$ denote the empirical mean of $(\mathbf{X}^e, \mathbf{Y}^e)$. For centering, we recommend subtracting $\frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}\hat{\mu}^e$ from each sample. By mean-centering globally (and not with an environment-specific intercept), the estimator is able to leverage changes in mean between environments. For scaling, define

(21)
$$c_{k,e} = \frac{\mathbb{E}[(X_k^e)^2]}{n_e}$$
$$+ \frac{1}{(|\mathcal{E}| - 1)^2}\sum_{e' \neq e}\frac{\mathbb{E}[(X_k^{e'})^2]}{n_{e'}} \qquad \text{for } e \in \mathcal{E} \text{ and } k = 1, \ldots, p.$$

We recommend to scale the kth row of $\hat{\mathbf{Z}}^e$ and $\hat{\mathbf{G}}^e$ by approximately $1/\sqrt{c_{k,e}}$ for all $k = 1, \ldots, p$ and $e \in \mathcal{E}$. What is the motivation behind this scaling? In the following, we will discuss the special case $\mathcal{E} = \{1, 2\}$. In absence of noise in equation (16), $\|\mathbf{Z} - \mathbf{G}\beta^0\|_\infty = 0$. By allowing for $\|\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0\|_\infty \leq \lambda$, we account for the variance of $\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0$. Since we work with a supremum bound and the same λ for all components, we want all scaled components to have roughly the same

variance. To be more precise, we want

$$\text{(22)} \qquad \text{Var}\left(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_k}{\sqrt{c_{k,1}}}\right) = \text{Var}\left(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_l}{\sqrt{c_{l,1}}}\right) \qquad \text{for all } k, l = 1, \dots, p.$$

It can be challenging to scale according to equation (22) as the correlation between X_k^e and $\eta_{p+1}^e = Y^e - X^e\beta^0$ is unknown and changes for different k. In the absence of confounding however and if X_k and X_l are not descendants of Y in the graph G, $\varepsilon = \eta_{p+1}^e$ is independent of X_k^e and X_l^e and the scaling of equation (21) implies

$$\text{Var}\left(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_k}{\sqrt{c_{k,1}}}\right) = \sigma_\varepsilon^2 = \text{Var}\left(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_l}{\sqrt{c_{l,1}}}\right),$$

where σ_ε denotes the standard deviation of $\varepsilon = \eta_{p+1}^e$. The scaling of equation (21) still has some theoretical justification in more general cases. In the presence of confounding and for general k, l it depends on the joint distribution of X_k^e, X_l^e and $\varepsilon = \eta_{p+1}^e$ whether $\text{Var}(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_k}{\sqrt{c_{k,1}}})$ and $\text{Var}(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_l}{\sqrt{c_{l,1}}})$ are of the same order. Notably, if equation (21) holds with equality and if the variables X_k^e, $k = 1, \dots, p$ and $\varepsilon = \eta_{p+1}^e$, $e \in \{1, 2\}$ are centered multivariate Gaussian, using moment inequalities,

$$\mathbb{E}[(X_k^e)^2]\sigma_\varepsilon^2 \leq \text{Var}(X_k^e \eta_{p+1}^e) \leq 2\mathbb{E}[(X_k^e)^2]\sigma_\varepsilon^2$$

for $e \in \{1, 2\}$, $k \in \{1, \dots, p\}$. Using independence of samples from different environments $e \in \{1, 2\}$,

$$\sum_{e \in \{1,2\}} \frac{\mathbb{E}[(X_k^e)^2]}{n_e}\sigma_\varepsilon^2 \leq \text{Var}((\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_k) \leq 2 \sum_{e \in \{1,2\}} \frac{\mathbb{E}[(X_k^e)^2]}{n_e}\sigma_\varepsilon^2$$

for all $k = 1, \dots, p$. Using equation (21),

$$\sigma_\varepsilon^2 \leq \text{Var}\left(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_l}{\sqrt{c_{l,1}}}\right) \leq 2\sigma_\varepsilon^2 \qquad \text{for all } k = 1, \dots, p.$$

Hence $\text{Var}(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_l}{\sqrt{c_{l,1}}})$ and $\text{Var}(\frac{(\hat{\mathbf{Z}} - \hat{\mathbf{G}}\beta^0)_k}{\sqrt{c_{k,1}}})$ are of the same order for all $k, l = 1, \dots, p$.

5.2. *Choosing* λ. Large segments of the regularization path of the *causal Dantzig* are usually poor estimates of β^0. Hence it is crucial to use an appropriate value of the regularization parameter λ. From a theoretical perspective, one would choose λ as in Theorem 4. However, σ_ε and $\sigma_{\max}^e$ are usually unknown in real-world datasets. Hence, in practice we propose to choose λ by k-fold cross-validation. Concretely, in each environment $e \in \mathcal{E}$ the samples are split into k groups of approximately equal size. Denote $\hat{\beta}^{\lambda, -i}$ the *causal Dantzig* estimator that is calculated on all samples except the samples from group i. Let $\hat{\mathbf{Z}}^i$ and $\hat{\mathbf{G}}^i$
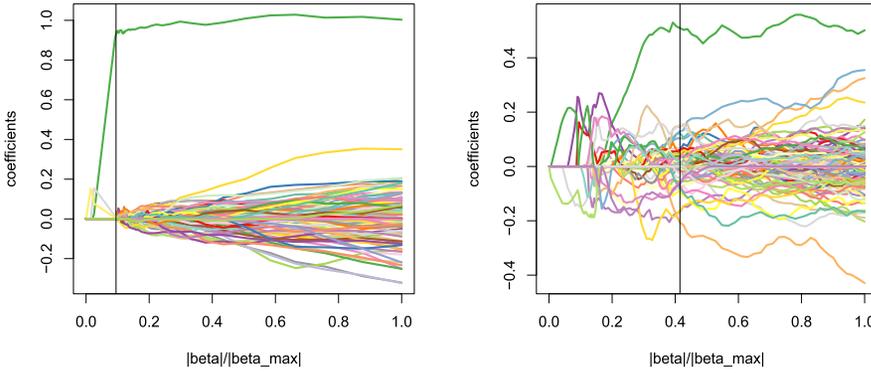
FIG. 4. *Two typical regularization paths for the causal Dantzig. The black vertical line specifies the solution chosen by* 10-*fold cross-validation. On the left-hand side $p = 100$, $n = 200$. On the right-hand side $p = 200$, $n = 60$. In both cases the standard deviation of the interventions is 2.5 and the variance of the errors is 1. One component of β^0 is equal to one (upper green line), all others are zero.*

be defined as in equation (5), using the samples from group i. Then we can choose $\hat{\lambda}^{cv}$ as a solution to

$$\hat{\lambda}^{cv} = \arg\min_{\lambda} \frac{1}{k} \sum_{i=1}^{k} \|\hat{\mathbf{Z}}^i - \hat{\mathbf{G}}^i \hat{\beta}^{\lambda, -i}\|_{\infty}.$$

We define the cross-validated *causal Dantzig* as $\hat{\beta}^{cv} := \hat{\beta}^{\hat{\lambda}^{cv}}$. Two exemplary regularization paths and the solution chosen by cross-validation are depicted in Figure 4.

5.3. *Preselection with hidden variables.* An alternative of running the *causal Dantzig* directly on a high-dimensional dataset is doing preselection. In the first stage, we recommend to run Lasso on observational data, if available. If observational data is not available, one could run Lasso on the pooled dataset. In the second stage, one would run the *causal Dantzig* with or without regularization on the active set of the first stage. Ideally, the first stage would screen out as many variables as possible, except for the parental set of the target variable Y. Quite often this will result in a set that contains a superset of the parental set implying a very useful dimensionality reduction. The following lemma provides some justification for this approach.

LEMMA 4. *Assume that the distribution $X_1, \ldots, X_p, Y$ is generated by a linear acyclic Gaussian structural equation model with directed acyclic graph D_{total} that consists of both the observed variables $X_1, \ldots, X_p, Y$ and (potentially) hidden confounders $H_1, \ldots, H_q$. Assume that the joint distribution of the variables*

$X_1, \ldots, X_p, Y, H_1, \ldots, H_q$ *is faithful (*Pearl *(2009)) to* D_{total}. *Let S denote the active set of regressing Y on* $X_1, \ldots, X_p$ *in the population case. Then*

$$\{k : X_k \text{ is a parent or a child of } Y \text{ in } D_{\text{total}}\} \subset S.$$

The proof can be found in the Supplementary Material (see Rothenhäusler, Bühlmann and Meinshausen (2018)). We test this two-step procedure on real world data in Section 6.4. However, note that for valid p-values (with the unregularized *causal Dantzig*) we would have a post-selection problem due to the screening step.

6. Numerical examples. Section 6.1 explores actual coverage and length of the asymptotic confidence intervals as defined in Section 3.3. In Section 6.2, we compare the *causal Dantzig* to instrumental variable regression for $p = 1$ under different types of interventions. In Section 6.3, we evaluate the performance of parameter selection by cross-validation as defined in Section 5.2. Finally, in Section 6.4 we discuss an application to real-world data that has been published in Meinshausen et al. (2016).

6.1. *Causal Dantzig in low dimensions*: *Confidence intervals.* In this section, we explore the actual coverage and average length of the asymptotic confidence intervals constructed according to Theorem 1.

We simulate data from two linear SEMs shown in Figure 5. Specifically, the data are generated according to the equations

$$(A): \begin{cases} X_2 & \leftarrow & \eta_2 \\ Y & \leftarrow & X_2 + \eta_y \\ X_1 & \leftarrow & Y - X_2 + \eta_1 \end{cases},$$

$$(23)$$

$$(B): \begin{cases} X_3 & \leftarrow & \eta_3 \\ X_2 & \leftarrow & X_3 + \eta_2 \\ Y & \leftarrow & -X_3 + X_2 + \eta_y \\ X_1 & \leftarrow & -X_2 + Y + \eta_1 \\ X_4 & \leftarrow & -Y + X_1 + \eta_4 \end{cases},$$
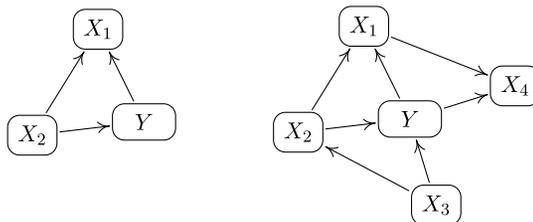

FIG. 5. *The graphs (A) and (B) used in the simulations. The noise distributions at all variables follow a factor model which allows for hidden confounding.*

TABLE 2

The first two rows contain actual coverage and average length of confidence intervals of causal
Dantzig *for the first variable in SEM* (A) *of equation* (23). *The last row contains the actual coverage
of ICP in these settings. The nominal coverage is* 0.95 *for causal Dantzig and at least* 0.95 *for ICP.
For small sample sizes, the variance is relatively large. As discussed in Section* 4, *regularization can
be helpful in these settings*

	$n = 50$	100	500	1000
Coverage	0.93 ± 0.01	0.95 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
Average length	65.79 ± 2918.53	4.11 ± 602.53	0.27 ± 0.62	0.18 ± 0.01
Coverage ICP	0.92 ± 0.01	0.84 ± 0.01	0.42 ± 0.02	0.3 ± 0.03

where the noise distributions of (η_1, η_2, η_y) and $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_y)$, respectively,
depend on the environment. Specifically, for SEM (A), we assume a factor model
for the noise

$$(\eta_1, \eta_2, \eta_y)^t = AH + \sigma_j(\varepsilon_1, \varepsilon_2, \varepsilon_y)^t,$$

where $(\varepsilon_1, \varepsilon_2, \varepsilon_y)^t \sim \mathcal{N}(0, 1_3)$, and the entries in both the factor loading matrix
$A \in \mathbb{R}^{3 \times 5}$ and the factor values $H \in \mathbb{R}^5$ are chosen i.i.d. standard normal. The 5-
dimensional variable H act as hidden confounders between the observed variables.
The noise contribution σ_j is chosen as 1 in environment $e = 1$ and as $1 + \kappa$ in
environment $e = 2$. We call $\kappa = \sigma_1 - \sigma_2$ the intervention strength as it measures
the variance of the additional noise input in environment $e = 1$ over environment
$e = 2$. In our simulations, it is chosen as 8. For SEM (B), we generate the data
analogously with the dimension of the hidden variable H being five.

We draw $n \in \{50, 100, 500, 1000\}$ samples in total (across both environments)
and compute the confidence intervals for the causal coefficients β^0 of Y with the
unregularized *causal Dantzig*. For SEM (A), the true causal coefficients for Y are
given by $\beta^0 = (0, 1)$ and the actual coverage and average length of the constructed
intervals at confidence level 0.05 with the unregularized *causal Dantzig* is shown in
the two upper rows of Table 2 for variable X_1. The bottom row show the coverage
of the confidence intervals for invariant causal prediction (ICP). For large n, ICP
often (rightfully) rejects all models and outputs neither coefficient estimates nor
confidence intervals. These cases were ignored in the table. ICP is not consistent,
and hence has incorrect coverage for growing sample size, as clearly visible in the
table.

The *causal Dantzig* has approximately correct coverage for all sample sizes
in this example. For small sample sizes, the variance of the *causal Dantzig* is
large and consequently the average length of the confidence intervals of the *causal
Dantzig* is large, too. In such regimes, regularization is recommended, as discussed
in Section 4. For larger sample sizes, the confidence intervals are shrinking con-
siderably with the $\sqrt{n}$-rate. For SEM (A), this effect is depicted in Table 2. Table 3

TABLE 3
Actual coverage and average length of confidence intervals for first variable in SEM (B) of equation (23) *with causal Dantzig. The nominal coverage is* 0.95

	$n = 50$	100	500	1000
Coverage	0.95 ± 0.01	0.95 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
Average length	$11{,}354.75 \pm 2776.95$	$57.27 \pm 28{,}842.69$	0.69 ± 7.28	0.39 ± 3.73

shows these effects for SEM (B). Note that also in this case the actual coverage of the *causal Dantzig* is approximately correct.

6.2. *Causal Dantzig and the instrumental variable approach.* To compare the *causal Dantzig* to instrumental variables, consider a binary instrument $e \in \{0, 1\}$. To be more precise, we consider the model

$$H, \epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1), \qquad e \in \{0, 1\},$$

$$(24) \qquad X = H + 2e + \epsilon_1,$$

$$Y = 2X + H + 2\epsilon_2.$$

The corresponding DAG is depicted in Figure 6. In words, X is a direct cause of Y, there is a hidden confounder H that causes both X and Y, and e is an instrument for X, meaning that e is a root node and a direct cause of X, but not of H or Y. Note that the conditional mean differs between settings, that is, $\mathbb{E}[X|e = 1] \neq \mathbb{E}[X|e = 0]$. Hence the IV approach is consistent for the true causal effect from X to Y, as discussed in Section 3.6.

For each environment $e \in \{0, 1\}$, we generate n samples and estimate the direct causal effect via *causal Dantzig* and instrumental variables regression using the function ivreg in the R-package AER. Table 4 shows the mean square error for both methods. For few observations, the *causal Dantzig* is relatively unstable. For larger values of n, this is not the case and the mean square error shrinks at the $\sqrt{n}$-rate for both estimators. The instrumental variables (IV) approach outperforms the *causal Dantzig* in this example. This is due to the fact that IV is a fraction of conditional means, whereas the *causal Dantzig* is a fraction of conditional covariances. Estimating conditional means is statistically easier, but it comes at a certain price as we will see below.
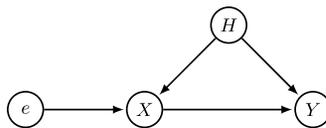

FIG. 6. *The DAG corresponding to the model of equation* (24).

TABLE 4
Mean square error for varying n. Instrument is not weak

	n = 20	**50**	**100**
causal Dantzig	0.46 ± 0.41	0.03 ± 0.01	0.01 ± 0
ivreg	0.07 ± 0.01	0.02 ± 0	0.01 ± 0

For the second model, we change the edge function between e and X. Notably,

$$H, \epsilon_1, \epsilon_2, \epsilon_3 \sim \mathcal{N}(0, 1), \qquad e \in \{0, 1\},$$
$$(25) \qquad X = H + 2e \cdot (0.25 + \epsilon_3) + \epsilon_1,$$
$$Y = 2X + H + 2\epsilon_2.$$

Both the conditional variance $\text{Var}(X|e = \bullet)$, $\bullet \in \{0, 1\}$ and the conditional mean $\mathbb{E}[X|e = \bullet]$, $\bullet \in \{0, 1\}$ change between the environments. However, the conditional mean changes only slightly, imposing difficulties for the IV approach. Again, for each environment $e \in \{0, 1\}$ we generate n samples and estimate the direct causal effect via *causal Dantzig* and ivreg. As seen in Table 5, for very few observations, both ivreg and *causal Dantzig* are comparatively far from the target quantity. For larger values of n, the *causal Dantzig* converges with the $\sqrt{n}$-rate. The instrumental variables approach is consistent but unstable for these small sample sizes as the instrument is weak. It exhibits large MSE as it does not use the changing variance for inference.

6.3. *Causal Dantzig in high dimensions.* We consider a structural equation model, where the observed variables $X_1, \ldots, X_p, Y$ form a chain and the distribution of the unobserved confounder η changes between the environments. The corresponding directed acyclic graph is depicted in Figure 7. To be more precise, the distribution of the observed variables e, X and Y is generated according to the

TABLE 5
Mean square error for varying n. The instrument is weak, but causal Dantzig *can leverage changes in variance*

	n = 20	**50**	**100**
causal Dantzig	24.05 ± 90.75	0.03 ± 0	0.01 ± 0
ivreg	$36{,}634.21 \pm 161{,}096.94$	$4244.29 \pm 15{,}557.82$	1862.7 ± 8171.14
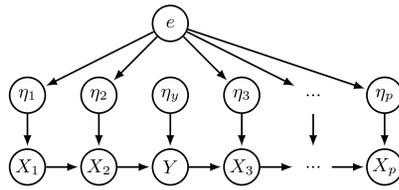
FIG. 7.   *The directed acyclic graph corresponding to SEM (C).*

following structural equation model:

$$\eta_k = \eta_k^0 + \delta_k^e,$$

$$(26) \quad (C): \begin{cases} X_1 & \leftarrow & \eta_1 \\ X_2 & \leftarrow & X_1 + \eta_2 \\ X_{p+1} = Y & \leftarrow & X_2 + \eta_y \\ X_3 & \leftarrow & Y + \eta_3 \\ X_4 & \leftarrow & X_3 + \eta_4 \\ \vdots & & \vdots \quad \vdots \\ X_p & \leftarrow & X_{p-1} + \eta_p \end{cases} \quad \text{with} \quad \begin{aligned} & \delta_k^e = \begin{cases} 0, & e = 0 \text{ or} \\ & k = p+1, \\ z_k, & e = 1, \end{cases} \\ & z_k \sim \mathcal{N}(0, \sigma^2) \quad \text{i.i.d.,} \\ & \eta_k^0 \sim \mathcal{N}(0, 1) \quad \text{i.i.d.,} \\ & e \in \{0, 1\}, \end{aligned}$$

We assume that z_k and η_k are jointly independent. The regularization parameter λ is chosen by 10-fold cross-validation. Figure 8 shows the regularization path for two different values of p. Figure 9 shows the regularization path for varying intervention strength σ. Finally, in Figure 10 the number of samples collected from each environment $n := n_0 = n_1$ is varied. In a nutshell, cross-validation seems to select a reasonable regularization parameter in most cases, estimation performance deteriorates with increasing p, but improves with increasing n and drastically so with increasing intervention strength σ.

6.4. *Gene knockout experiments.*   We outline here an application which has appeared in Meinshausen et al. (2016). The authors consider gene expression in yeast (Saccharomyces cerevisiae) under deletion of single genes (Kemmeren et al. (2014)): 160 samples are wild-type (observational); and 1, 479 samples are measured under the deletion of a single gene (intervention). For each of those observations, genome-wide mRNA expression levels were measured. We denote these measurements by $X_1, \ldots, X_{p+1}$, where $p + 1 = 6170$. The goal is to predict whether mRNA expression level $Y = X_{p+1}$ changes significantly under a new and unobserved gene-deletion X_j, $j \neq p + 1$. Knocking out a gene is not always successful, and the measured activity of a gene is not constant (or zero) after knocking it out, that is, the intervention is "noisy." Overall, knockouts decrease the activity, which can be interpreted as a negative shift in the measured log-activity of a gene.

The data is split into training and validation data. To this end, the 1, 479 interventional samples are divided into five sets $B_1, \ldots, B_5$. For some $v \in \{1, \ldots, 5\}$,
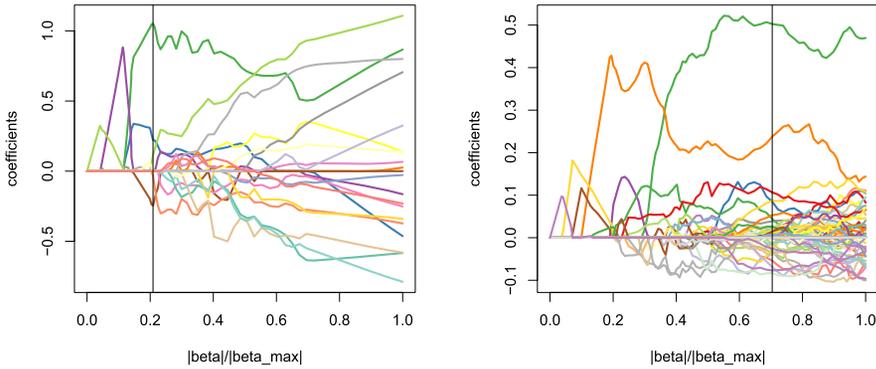
FIG. 8. *Two regularization paths for the* causal Dantzig *with n = 30 and σ = 2.5. The black vertical line specifies the solution chosen by* 10-*fold cross-validation. On the left p = 20, on the right p = 200. The true underlying coefficient is equal to one for one variable* (*upper green line*), *and equal to zero for all other variables. Though not flawless, cross-validation chooses a reasonable regularization parameter in both cases.*

the training data consists of the four sets $\{B_i\}_{i\in\{1,2,3,4,5\}\setminus\{v\}}$ and the 160 observational samples. The samples in B_v are held out for validation. The interventional effects on the validation set B_v were predicted using only training data. This procedure is carried out for all sets $B_v, v = 1, \ldots, 5$, that is, each gene perturbation is excluded from the training set once.

Preselection with the LASSO was used on the pooled data to screen for a superset of the parental set of variable X_{p+1}. For some justification of this approach, see Section 5.3. Then the *causal Dantzig* without regularization was used, with
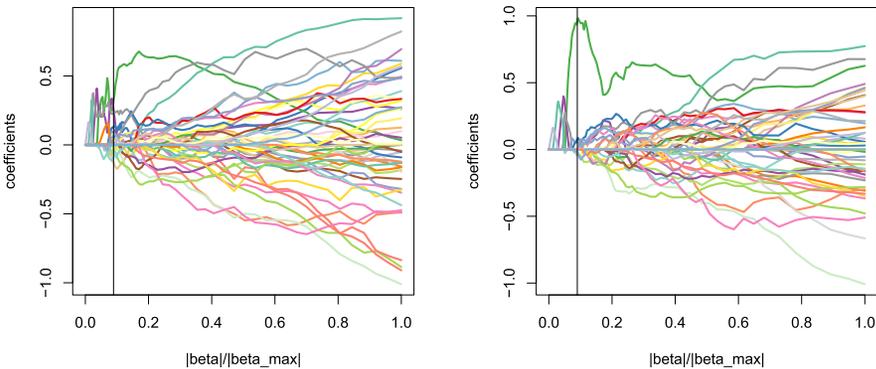

FIG. 9. *Two regularization paths for the* causal Dantzig *with p = 30 and n = 30. The black vertical line specifies the solution chosen by* 10-*fold cross-validation. On the left, the intervention strength is σ = 2.5, on the right it is σ = 3.5. The true underlying coefficient is equal to one for one variable* (*upper green line*), *and equal to zero for all other variables. Clearly, strong interventions improve estimation performance.*
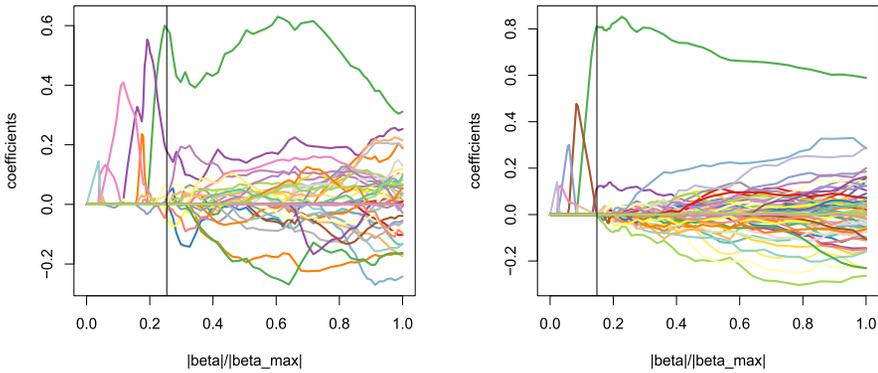
FIG. 10.   *Two regularization paths for the* causal Dantzig *with $p = 100$ and $\sigma = 3.5$. The black vertical line specifies the solution chosen by* 10*-fold cross-validation. On the left*, *the sample size is $n = 30$, on the right it is $n = 60$. The true underlying coefficient is equal to one for one variable* (*upper green line*), *and equal to zero for all other variables. Estimation performance is clearly better on the right*.

setting $e = 1$ for observational data and $e = 2$ for interventional training data. Using *causal Dantzig* without screening step is computationally prohibitive due to the large number of variables and as the procedure is repeated for each possible target variable $X_1, \ldots, X_{p+1}$. The s most often selected intervention predictions were compared to so-called "strong intervention effects" (SIEs) as defined in Meinshausen et al. (2016). SIEs are computed on the held-out data B_v and are a measure for the total causal effect. The results are depicted in Figure 11. As an example, for *causal Dantzig* the four most often selected intervention predictions correspond to SIEs.

Screening for causal effects is a very challenging problem in this setting, mainly due to the high-dimensionality of the dataset and the presence of hidden confounders. The ground truth is not perfectly known but good proxies (strong intervention effects) can be computed on hold-out interventional data. The strongest discoveries of `InvariantCausalPrediction` (ICP) and `causalDantzig` correspond very well to the benchmark. Assuming hidden confounding and shift interventions (`causalDantzig`) leads to a different ranking of genes compared to assuming the absence of confounding and allowing for arbitrary interventions (ICP). Interestingly, while both methods miss some important variables, making "wrong" assumptions such as linearity or absence of latent confounding do not seem to lead to false positives for the first few variables in the ranking. This form of validation and the comparison to other methods are further discussed in Meinshausen et al. (2016).

7. Discussion.   Causal discovery is challenging, particularly in the presence of hidden confounders and feedback loops. However, hidden confounders can rarely
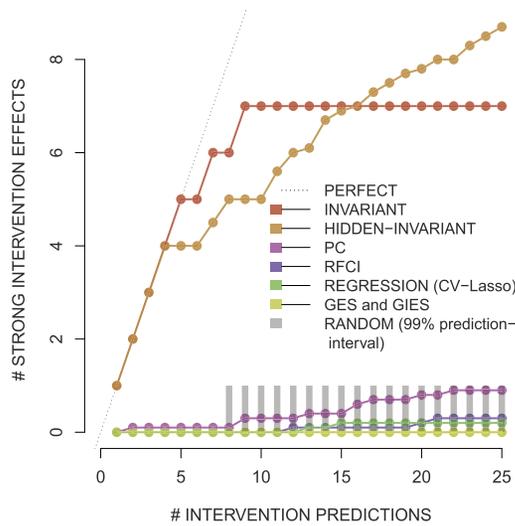
FIG. 11. *The results of the gene knockout experiments on the Kemmeren dataset (Kemmeren et al.* (2014)). *This figure has been published in Meinshausen et al. (2016). The method* HIDDEN-IN-VARIANT *is an unpublished early version of* causalDantzig.

be excluded and feedback loops are to be expected in many real-world applications (e.g., in biological systems). We introduced the notion of *inner-product invariance* and showed that inference in linear structural equation models under *inner-product invariance* is possible, both for low- and high-dimensional data.

The proposed methods have interesting parallels to widely-used statistical methods. For example, the functional form of the *causal Dantzig* estimator is similar to linear regression. The regularized *causal Dantzig* is similar to the Dantzig selector. For two environments ($|\mathcal{E}| = 2$), the *causal Dantzig* estimator can be compared with instrumental variable regression and is consistent in certain settings in which instrumental variable regression fails. Hence, we believe that the *causal Dantzig* will push the boundaries in the analysis of certain types of datasets, in particular in the analysis of datasets where potentially unknown interventions (or "perturbations") change both the mean and the variance of the observed error distribution. Empirical results show state-of-the-art performance of our proposed estimator on a real-world dataset.

We investigated the identifiability of direct causal effects under the proposed model class. Furthermore, we showed that the regularized *causal Dantzig* can be consistent in the high-dimensional case even if not all covariates have been intervened on. The estimator can be obtained by solving a linear program and as such is feasible for large-scale causal inference. We derived asymptotic confidence intervals for the unregularized *causal Dantzig*, as well as guarantees for statistical accuracy for the regularized *causal Dantzig*.

The notion of *inner-product invariance* pushes the boundaries for the types of datasets we can leverage for causal discovery. We expect it to be useful for practitioners, in particular as a simple and fast tool for screening for potential direct causal effects. From a theoretical perspective, the regularized and unregularized *causal Dantzig* provide new perspectives on invariant causal prediction, on the instrumental variable approach and on classical theory for high-dimensional estimation.

SUPPLEMENTARY MATERIAL

Supplement to "Causal Dantzig: Fast inference in linear structural equation models with hidden variables under additive interventions". (DOI: 10.1214/18-AOS1732SUPP; .pdf). The Supplementary Material contains detailed technical proofs.

REFERENCES

ANDERSON, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist*. **1** 135–141. MR0331612

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist*. **25** 505–541. MR1439312

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc*. **91** 444–455.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025

BOWDEN, R. J. and TURKINGTON, D. A. (1990). *Instrumental Variables*. Cambridge Univ. Press, Cambridge. Reprint of the 1984 original. MR1113481

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist*. **35** 2313–2351. MR2382644

CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res*. **3** 507–554. MR1991085

DIDELEZ, V., MENG, S. and SHEEHAN, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statist. Sci*. **25** 22–40. MR2741813

HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res*. **13** 2409–2464. MR2973606

HAUSER, A. and BÜHLMANN, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **77** 291–318. MR3299409

HOYER, P. O., JANZING, D., MOOIJ, J. M., PETERS, J. and SCHÖLKOPF, B. (2009). Nonlinear causal discovery with additive noise models. *Adv. Neural Inf. Process. Syst*. **21** 689–696.

KEMMEREN, P., SAMEITH, K., VAN DE PASCH, L. A., BENSCHOP, J. J., LENSTRA, T. L., MAR-
    GARITIS, T., O'DUIBHIR, E., APWEILER, E., VAN WAGENINGEN, S. et al. (2014). Large-scale
    genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors.
    Cell **157** 740–752.

LEWBEL, A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous
    regressor models. *J. Bus. Econom. Statist.* **30** 67–80. MR2899185

MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional inter-
    vention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

MEINSHAUSEN, N., HAUSER, A., MOOIJ, J. M., PETERS, J., VERSTEEG, P. and BÜHLMANN, P.
    (2016). Methods for causal inference from gene perturbation experiments and validation. *Proc.
    Natl. Acad. Sci. USA* **113** 7361–7368.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Essay on
    principles. Section 9. [Translated and edited by D. Dabrowska and T. Speed] *Statist. Sci.* **5** (1990)
    465–480.

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press,
    Cambridge. MR2548166

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant
    prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–
    1012. With comments and a rejoinder. MR3557186

R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation
    for Statistical Computing, Vienna, Austria.

RICHARDSON, T. and ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): A unifica-
    tion of the counterfactual and graphical approaches to causality. Working Paper 128, Center for
    the Statistics and the Social Sciences, Univ. Washington Series.

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal
    inference in epidemiology. *Epidemiology* **11** 550–560.

ROTHENHÄUSLER, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2019). Supplement to "Causal
    Dantzig: Fast inference in linear structural equation models with hidden variables under additive
    interventions." DOI:10.1214/18-AOS1732SUPP.

ROTHENHÄUSLER, D., HEINZE, C., PETERS, J. and MEINSHAUSEN, N. (2015). Backshift: Learn-
    ing causal cyclic graphs from unknown shift interventions. *Adv. Neural Inf. Process. Syst.* **29**
    1513–1521.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized
    studies. *J. Educ. Psychol.* **66** 688.

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian
    acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030. MR2274431

TIAN, J. and PEARL, J. (2001). Causal discovery from changes. In *Proceedings of the* 17*th Confer-
    ence on Uncertainty in Artificial Intelligence* (*UAI*) 512–522.

VANDERWEELE, T. J. and ROBINS, J. M. (2010). Signed directed acyclic graphs for causal infer-
    ence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 111–127. MR2751246

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for
    the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

VERMA, T. and PEARL, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of
    the* 6*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 255–270.

WALD, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.*
    11 285–300. MR0002739

WANG, L. and TCHETGEN TCHETGEN, E. (2018). Bounded, efficient and multiply robust esti-
    mation of average treatment effects using instrumental variables. *J. R. Stat. Soc. Ser. B. Stat.
    Methodol.* **80** 531–550. MR3798877

WRIGHT, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. The Macmillan Company, New York.

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

SEMINAR FÜR STATISTIK
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: rothenhaeusler@stat.math.ethz.ch
        buehlmann@stat.math.ethz.ch
        meinshausen@stat.math.ethz.ch
URL: http://stat.ethz.ch