Regression by dependence minimization and its application to causal inference in additive noise models

Joris Mooij JORIS.MOOIJ@TUEBINGEN.MPG.DE
Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Jonas Peters JONAS.PETERS@TUEBINGEN.MPG.DE
Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics, Dept. Schölkopf, Spemannstraße 38, 72076 Tübingen, Germany

Abstract

Motivated by causal inference problems, we propose a novel method for regression that minimizes the statistical dependence between regressors and residuals. The key advantage of this approach to regression is that it does not assume a particular distribution of the noise, i.e., it is non-parametric with respect to the noise distribution. We argue that the proposed regression method is well suited to the task of causal inference in additive noise models. A practical disadvantage is that the resulting optimization problem is generally non-convex and can be difficult to solve. Nevertheless, we report good results on one of the tasks of the NIPS 2008 Causality Challenge, where the goal is to distinguish causes from effects in pairs of statistically dependent variables. In addition, we propose an algorithm for efficiently inferring causal models from observational data for more than two variables. The required number of regressions and independence tests is quadratic in the number of variables, which is a significant improvement over the simple method that tests all possible DAGs.

1. Introduction

Most existing methods for learning causal models from observational data assume that continuous variables are multivariate Gaussian (Spirtes et al., 1993; Geiger & Heckerman, 1994; Bollen, 1989). This corresponds to structural equations where effects are linear functions of their causes up to an additive Gaussian noise term that is independent of the causes. Apart from the fact that those assumptions are often not met in practice, it has been pointed out recently that they can actually exacerbate the problem of causal inference. Indeed, for linear models, *non-Gaussianity* in the data can actually aid in distinguishing causal directions (Shimizu et al., 2006); similarly, *nonlinearity* of the functional relationships can aid in identifying the causal structure (Hoyer et al., 2009).

An important class of causal models are *additive noise models*. The structure of these models is determined by a directed acyclic graph (DAG), with a random variable corresponding to each node, and each random variable is assumed to be a (possibly nonlinear) function of its parents plus an additive noise term, with the important restriction that all noise terms are assumed to be jointly independent. For this class of causal models with additive noise, a causal inference method has been proposed recently (Hoyer et al., 2009) that estimates the causal model from a finite sample by exploiting nonlinearities and non-Gaussianities in the data. The basic idea of their method is the following: for a given candidate DAG, one solves a regression problem for each node, modelling it as a (possibly nonlinear) function of its parents. Then, a statistical independence test is performed to assess whether all residuals are jointly independent. If that is the case, the candidate DAG is accepted, otherwise it is rejected. The two basic ingredients of this method are regression methods and independence tests; Hoyer et al. (2009) use Gaussian Process Regression in combination with the Hilbert-Schmidt Independence Criterion (HSIC) independence test (Gretton et al., 2005).

Since the method of Hoyer et al. (2009) benefits from non-Gaussian noise, we argue that it is not entirely consistent to use a regression method that assumes Gaussian noise (as the standard Gaussian Process Regression does). One could use other regression meth-

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

ods that assume different noise distributions, but then the problem becomes how to choose the regression method if the noise distribution is unknown (which is usually the case).

The solution we propose here is a novel regression method that makes no assumptions about the noise distribution. The basic idea is to simply minimize the *dependence* between residuals and regressors, measured by some dependence measure. The underlying intuition is that regression tries to model the dependence of the output on the input, and it is successful when the residuals, i.e., the difference between the actual and the predicted output, are no longer dependent on the input. Here we have chosen the empirical HSIC estimator (Gretton et al., 2005) as a measure of dependence, because of its good performance for scalar and vector-valued continuous variables. In the context of causal inference, it also has been successfully applied in (Hoyer et al., 2009; Zhang & Hyvärinen, 2008). Apart from making the method nonparametric with respect to the noise distribution, it can be argued that our solution is more elegant from a theoretical point of view for the causal inference task discussed above, because it unifies the regression with the subsequent independence test. Indeed, the regression and independence test now both use the *same* "loss function": the statistical dependence between residuals and the regressors (parent variables), according to the same dependence measure.

Another contribution of the present work is an efficient algorithm for inferring DAGs from data. It improves upon the algorithm proposed by Hoyer et al. (2009) by reducing the computational complexity from super-exponential in the number of variables to quadratic.

This paper is organized as follows. In Section 2 we discuss our novel regression method. In Section 3 we dicuss the causal inference task and show how the regression method can be applied successfully for distinguishing cause and effect in pairs of statistically dependent variables. In Section 4, we propose an efficient method for inferring complete DAGs consisting of $d \geq 2$ variables, which has computational complexity $\mathcal{O}(d^2)$, and illustrate its performance on a toy example.

2. Regression by dependence minimization

Suppose X is a random variable with values in $\mathbb{R}^m$ and Y, E are random variables with values in $\mathbb{R}$. We assume that $Y = f(X) + E$ for some function $f : \mathbb{R}^m \to \mathbb{R}$ and that $X \perp\!\!\!\perp E$ (i.e., X is independent of E). Now, given an i.i.d. sample $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,N}$ of the pair (X, Y), our goal is to approximate the function f. Given some estimate $\hat{f}$ of the function f and a sample $\mathcal{D}$, we define the *residuals* $\{\hat{e}^{(i)} := y^{(i)} - \hat{f}(x^{(i)})\}_{i=1,\ldots,N}$. An example of a situation in which this regression problem occurs is where Y is an *effect* of several *causes* $X_1, \ldots, X_m$ (each with values in $\mathbb{R}$) and of some additional unobserved causes, summarized in an *additive noise* variable E. We will not make assumptions about the probability distribution of E other than that it has mean 0 and is independent of X.

Regression is an old and important problem and has been extensively studied. Most regression methods optimize some "loss function" over a class of functions, that is they solve a minimization problem

$$\hat{f} = \operatorname*{argmin}_{f' \in \mathcal{F}} \mathcal{L}\big(f', \mathcal{D}\big),$$

where $\mathcal{F}$ is a set of functions and $\mathcal{L}$ is a loss function that measures how well a function $f' \in \mathcal{F}$ fits the data $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1,\ldots,N}$. A concrete example is ℓ_2-regularized linear least-squares regression, where the unknown function f is approximated with a linear combination of n basis functions $\phi_r : \mathbb{R}^m \to \mathbb{R}$ with weights α_r, i.e.,

$$f_\alpha(x) := \sum_{r=1}^{n} \alpha_r \phi_r(x), \tag{1}$$

by minimizing the following ℓ_2-regularized ℓ_2-loss function:

$$\hat{\alpha} := \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \left(\sum_{i=1}^{N} \big(y^{(i)} - f_\alpha(x^{(i)})\big)^2 + \frac{\lambda_N}{2} \|\alpha\|_2^2 \right) \tag{2}$$

and taking $\hat{f} = f_{\hat{\alpha}}$ as the estimate for the function f. The first term in (2) is the ℓ_2-loss of the residuals and the second (regularization) term is important to avoid overfitting. The regularization constant λ_N is often chosen by cross-validation. Other regularization terms can be used, but this particular one has the advantage of mathematical simplicity. We will henceforth refer to (2) simply as "least-squares regression". Asymptotically, as $N \to \infty$, the corresponding estimator $f_{\hat{\alpha}}$ will (under certain technical conditions, see e.g., (Györfi et al., 2002)) converge to the conditional mean $\mathbb{E}(Y \,|\, X = x) = f(x)$.

The ℓ_2-loss function in (2) is adapted to a Gaussian distribution of the noise E, because it corresponds with the log-likelihood in that case. If E actually has a non-Gaussian distribution, the estimation procedure (2) will still converge to the true f asymptotically, but

for any finite N, other loss functions may achieve better approximations to f. For example, the ℓ_1 loss $\sum_{i=1}^{N} \left|y^{(i)} - f_\alpha(x^{(i)})\right|$ may yield better results if the noise E has a Laplace distribution. The choice of the loss function can have a large influence on the result of the regression for any finite N.

If the distribution of the noise E is unknown, it is not clear which loss function will give optimal results. Here we propose a method for regression that does not assume a particular distribution of the noise E; instead, we minimize the *dependence* between the residuals $Y - f_\alpha(X)$ and the regressor X. A theoretically well-motivated dependence measure would be the mutual information between X and $Y - f_\alpha(X)$. In practice, however, estimating this mutual information can be difficult. In this work, we will minimize a mutual-information like quantity instead—the empirical estimator of the Hilbert Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). HSIC uses kernels for measuring dependence and has good uniform convergence guarantees. We continue our exposition with a short introduction to the HSIC.

2.1. Hilbert-Schmidt Independence Criterion

Let X be a random variable with values in some set $\mathcal{X}$. Consider a Hilbert space $\mathcal{H}_X$ of functions from $\mathcal{X}$ to $\mathbb{R}$. Then $\mathcal{H}_X$ is a reproducing kernel Hilbert space (see e.g., (Schölkopf & Smola, 2002)) if for each $x \in \mathcal{X}$, the Dirac evaluation operator $\delta_x : \mathcal{H}_X \to \mathbb{R} : f \mapsto f(x)$ is a bounded linear functional. To each point $x \in \mathcal{X}$ there corresponds an element $\phi_X(x) \in \mathcal{H}_X$ such that $\langle \phi_X(x), \phi_X(x') \rangle = k_X(x, x')$, where $k_X : \mathcal{X}^2 \to \mathbb{R}$ is a unique positive definite kernel. Similarly, for a random variable Y with values in $\mathcal{Y}$, let $\mathcal{H}_Y$ be a reproducing kernel Hilbert space with kernel $k_Y : \mathcal{Y}^2 \to \mathbb{R}$ and feature mapping $\phi_Y : \mathcal{Y} \to \mathcal{H}_Y$. We assume throughout the paper that the kernel functions are bounded and continuous.

In analogy with a covariance matrix, we define a *cross-covariance operator*, which is a linear operator $\mathcal{C}_{XY} : \mathcal{H}_Y \to \mathcal{H}_X$ satisfying

$$\mathcal{C}_{XY} = \mathbb{E}_{X,Y}[(\phi_X - \mu_X) \otimes (\phi_Y - \mu_Y)],$$

where $\otimes$ is the tensor product and $\mu_X = \mathbb{E}_X(\phi_X) \in \mathcal{H}_X$ is the mean element corresponding to the probability distribution of X (and similarly for μ_Y). The square of the Hilbert-Schmidt norm of the cross-covariance operator (HSIC),

$$\mathrm{HSIC}(\mathcal{H}_X, \mathcal{H}_Y, \mathbb{P}_{XY}) := \|\mathcal{C}_{XY}\|^2_{\mathrm{HS}}, \tag{3}$$

is a measure of the statistical dependence of X and Y. Gretton et al. (2005) show that whenever the kernels k_X, k_Y are *universal* on respective compact domains $\mathcal{X}$ and $\mathcal{Y}$ in the sense of Steinwart (2002), then $\mathrm{HSIC}(\mathcal{H}_X, \mathcal{H}_Y, \mathbb{P}_{XY}) = 0$ if and only if X and Y are independent. A universal kernel such as the Gaussian RBF kernel or the Laplace kernel permits HSIC to detect any dependence between X and Y.

Gretton et al. (2005) define the following *empirical* HSIC estimator for an i.i.d. sample $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1,\dots,N}$:

$$\widehat{\mathrm{HSIC}}(\mathcal{H}_X, \mathcal{H}_Y, \mathcal{D}) := \frac{1}{N^2} \mathrm{tr}\,(KHLH), \tag{4}$$

where N is the number of data points, K is the $N \times N$ kernel matrix for X and L for Y, i.e.,

$$K_{ij} = k_X(x^{(i)}, x^{(j)}), \qquad L_{ij} = k_Y(y^{(i)}, y^{(j)}),$$

and H is the $N \times N$ matrix defined by

$$H := I - \frac{1}{N} 1 \cdot 1^T, \qquad \text{i.e.,} \qquad H_{ij} = \delta_{i,j} - \frac{1}{N}.$$

Gretton et al. (2005) show that for $N \to \infty$, $\widehat{\mathrm{HSIC}} \to 0$ if and only if X is independent of Y. Furthermore, the empirical HSIC estimator (4) has a bias of $\mathcal{O}(N^{-1})$, but this is negligible with respect to finite sample fluctuations. The empirical estimate (4) converges to the population value (3) at rate $\mathcal{O}(N^{-1/2})$. Finally, Gretton et al. (2005) propose a statistical test of independence based on the empirical HSIC estimator, which accepts the null hypothesis $H_0 : X \perp\!\!\!\perp Y$ if $p > \alpha$, but rejects it if $p < \alpha$, for some threshold α.

With some abuse of notation, we will henceforth simply write $\widehat{\mathrm{HSIC}}(X, Y)$ instead of $\widehat{\mathrm{HSIC}}(\mathcal{H}_X, \mathcal{H}_Y, \{(x^{(i)}, y^{(i)})\}_{i=1,\dots,N})$.

2.2. Regression by minimizing the HSIC

To return to our regression problem: we propose to replace the "log-likelihood" term in the loss-function by the empirical HSIC estimator. In particular, in (2) we replace the sum of the squares of the residuals by the empirical HSIC estimate of the dependence of the residuals with the regressor:

$$\hat{\alpha} := \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \left(\widehat{\mathrm{HSIC}}\big(X, Y - f_\alpha(X)\big) + \frac{\lambda_N}{2} \|\alpha\|_2^2 \right). \tag{5}$$

This yields an estimate for the function f, modulo some additive constant, assuming that the kernel k_Y is translation invariant. The missing constant can be estimated using the assumption that the mean of E is zero, by using the final estimate

$$\hat{f} = f_{\hat{\alpha}} + \sum_{i=1}^{N} \left(y^{(i)} - f_{\hat{\alpha}}(x^{(i)}) \right),$$

with $\hat{\alpha}$ as in (5). A noteworthy property of the particular loss function (5) is that it is *not* a sum over data points of some quantity.

2.2.1. Implementation details

Note that $\widehat{\mathrm{HSIC}}\big(X, Y - f_\alpha(X)\big)$ can easily be computed in $\mathcal{O}(N^2)$, since only the kernel matrices K and L are needed. For regression, K is fixed through the whole process, so it can be precomputed and stored for speedup if needed. The resulting optimization problem is in general non-convex (assuming the kernels are nonlinear). We used `libLBFGS` (Okazaki & Nocedal, 2008), a C implementation of the L-BFGS method (Liu & Nocedal, 1989), to minimize the regularized HSIC loss function in (5). We chose Gaussian RBF kernels for both X and Y, i.e., $k_X(x, x') = \exp(-\|x - x'\|^2 \sigma_X^{-2})$, and similarly for k_Y. Further, we used Gaussian RBF functions with centers $\{x^{(i)}\}_{i=1,\ldots,N}$ and the same width σ_X for the basis functions in the expansion of the function f in (1). For convenience, the kernel width σ_X is fixed as the median distance between points in the sample (Schölkopf & Smola, 2002). The kernel width σ_Y is chosen in the same way, but based on an initial rough estimate of the residuals. We chose the regularization constant λ_N by 2-fold cross-validation, using the average empirical HSIC over the folds as optimization criterion for λ_N. Note that we take a small number of folds, because the empirical HSIC estimator is not an additive function of the data points; indeed, if one would take the number of folds equal to the number of data points ("leave-one-out cross-validation"), one would evaluate the empirical HSIC of a single data point, which obviously makes no sense. For the independence tests, we use the permutation test for estimating the p-value of the HSIC as described in (Gretton et al., 2005).

3. Application: causal inference

The regression method discussed in the previous section was motivated by an application to causal inference. In this Section, we discuss the causal inference problem and how the proposed regression method naturally applies to it.

A *causal model* (Pearl, 2000) is defined as follows. Given a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes $\mathcal{V} = \{1, \ldots, d\}$ and directed edges $\mathcal{E}$, we denote the parents of node $i \in \mathcal{V}$ as $\mathrm{pa}(i)$. Each node $i \in \mathcal{V}$ has a corresponding (observed) random variable X_i and an (unobserved) random variable E_i. We will assume that these random variables have values in $\mathbb{R}$. For each node $i \in \mathcal{V}$, the corresponding random variable X_i ("effect") is a function $X_i = f_i(X_{\mathrm{pa}(i)}, E_i)$ of the random variables $X_{\mathrm{pa}(i)}$ ("causes") associated with the parents $\mathrm{pa}(i)$ of i and an independent noise source E_i. The *causal inference* problem we consider here is to estimate the causal model, given only a finite sample of observational data $\mathcal{D} = \{(X_1^{(i)}, \ldots, X_d^{(i)})\}_{i=1,\ldots,N}$. A causal model with *additive noise* is a special case, where the functions are of the form $f_i(X_{\mathrm{pa}(i)}, E_i) = g_i(X_{\mathrm{pa}(i)}) + E_i$. An example of a causal model with additive noise is shown in Figure 2(a), with the corresponding DAG in panel (b).

Hoyer et al. (2009) showed for the case of additive noise that the causal structure is generically identifiable in the two-variable case (with one of the few exceptions being the case where the function is linear and the distribution of the cause and of the noise is Gaussian). They proposed a method for inferring the causal structure from a finite sample of observations, which basically works as follows. Given a candidate DAG $\mathcal{G}$, for each node $i \in \mathcal{V}$ one performs regression of X_i as a function of its parents $X_{\mathrm{pa}(i)}$ to obtain an estimate of the (hypothetical) function $\hat{g}_i$. The corresponding residuals $\hat{\epsilon}_i$ should be independent of the parents $X_{\mathrm{pa}(i)}$. However, this independence condition is not enough: in fact, all residuals $\{\hat{\epsilon}_i\}_{i=1,\ldots,d}$ should be jointly independent, in order for the candidate DAG $\mathcal{G}$ to be accepted as a possible model for the data. An independence test is employed to test if at least one of the residuals $\hat{\epsilon}_i$ is not independent of one of the others (in practice, if the corresponding p-value exceeds some threshold); if this is the case, the candidate DAG is rejected, otherwise it is accepted.

Hoyer et al. (2009) show that their method works in the two-variable case (which is theoretically justified by their identifiability theorem) and give some empirical evidence that it also works for more than two variables, where they simply enumerate all DAGs and test for each DAG whether it fits the data. However, since the number of DAGs grows super-exponential in the number of variables, this method is only feasible for a few variables.

3.1. Experimental results

In this Subsection, we illustrate the advantage of the HSIC regression method over two standard regression methods for a simple toy example. Then, we apply the method to a dataset from the NIPS 2008 Causality Competition.

We will compare three different regression methods: regularized HSIC regression as discussed in Subsection 2.2.1, regularized linear least-squares regression as in (2) and Gaussian Process regression (Rasmussen

& Williams, 2006), using the implementation in (Rasmussen & Williams, 2007), with a Gaussian covariance function. We use the same kernels for the HSIC independence test as for the HSIC regression (with kernel widths set to the median distance between points in the sample), and also for the basis functions used in the least-squares regression.

3.1.1. Toy example

We start with a simple toy example, which illustrates the benefits of minimizing the HSIC dependence measure for the purposes of causal model selection in comparison with more traditional loss functions.

We consider the simple model $Y = X^2 + E$, with $X \sim U(-1,1)$ uniformly distributed. First we consider Gaussian noise $E \sim \mathcal{N}(0,1^2)$. We apply three different regression methods on a sample of $N = 300$ data points of the distribution; the results are shown in the top row of Figure 1. For the naked eye, it is difficult to decide which regression method gives the best result. Note that each regression method has its own definition of "best": the least-squares regression tries to minimize the ℓ_2-norm of the residuals, whereas the HSIC regression tries to minimize the HSIC measure of dependence between residuals and X. Applying the HSIC independence test to the residuals and X yields the following p-values (from left to right): 0.96, 0.66 and 0.87, each of which is clearly larger than $\alpha = 0.01$. Therefore, according to the HSIC measure, the residuals are indeed independent of X for each regression method, as we would expect them to be.

Now let us look at non-Gaussian noise; we alter the distribution of E to be $E \sim \text{Exp}(1) - 1$, i.e., an exponential distributed with mean 1 shifted so that it has mean 0. The bottom row of Figure 1 shows the results of the three regression methods. Again, to the naked eye, it is difficult to decide which regression method gives the best fit. However, the HSIC independence test now yields the following p-values: 0.55, 0.0010 and 0.0028. This means that independence of the residuals and X is only accepted for the HSIC regression, and rejected for the least-squares regression and the Gaussian Process regression methods. Note that the true noise E is actually independent of X because this is how the data were generated; this is verified by the HSIC independence test for E and X, which gives a p-value of 0.24. The failure of the least-squares and Gaussian Process regression methods in this case is not surprising when one considers that they incorrectly assume that the noise is Gaussian.

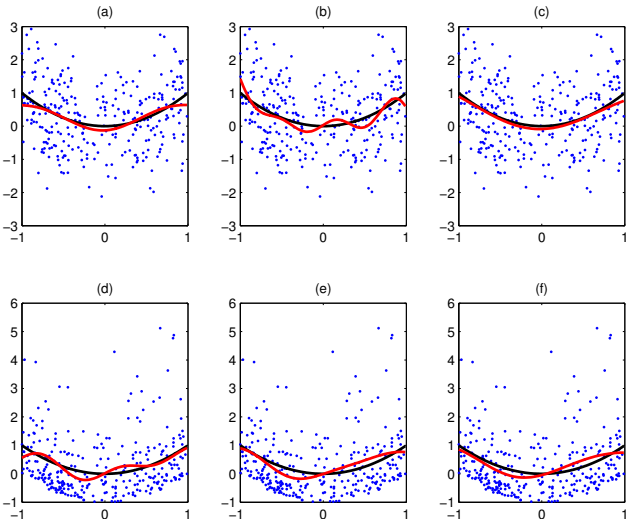


Figure 1. Regression results for the toy model $X \sim U(-1,1), Y = X^2 + E$. Top row: $E \sim \mathcal{N}(0,1^2)$; bottom row: $E \sim \text{Exp}(1) - 1$. From left to right: regularized HSIC regression, regularized least-squares regression, Gaussian Process regression. Black line is $Y = X^2$, blue points are samples and red line is regression result.

3.1.2. Real-world datasets

We also tested our method on the datasets of the `Cause-effect pairs` task for the NIPS 2008 Causality Competition (Mooij et al., 2008). Each dataset consists of a sample of two statistically dependent random variables, say X and Y, where one variable is known to causally influence the other (e.g., altitude and average temperature of weather stations). The task is to infer from the sample which variable is the cause and which one the effect. We use the approach proposed by (Hoyer et al., 2009), i.e., we test whether the causal model $Y = f_Y(X) + E_Y$, $E_Y \perp\!\!\!\perp X$ ("$X \to Y$") fits the data best, or the alternative model $X = f_X(Y) + E_X$, $E_X \perp\!\!\!\perp Y$ ("$Y \to X$"). The functions f_X, f_Y are estimated by regression and the independence of the residuals is tested using the HSIC independence test, which yields a small p-value if the data does not support the null hypothesis of independence, in which case the model is rejected.

Using at most 1000 data points from each dataset, we obtain the results shown in Table 1. Note that the HSIC regression yields the highest p-values (as one would expect) and that it correctly infers the causal direction in 6 out of 8 cases (using $\alpha = 0.01$). In one case it infers the wrong direction (which seems to be due to overfitting), in another case it rejects both directions (which may be due to a non-additive noise distribution, or a strong confounder). Qualitatively, the other regression methods yield similar decisions,

Algorithm 1 Find a DAG consistent with the data

input data matrix X of size $N \times d$, critical value α
  $S \leftarrow \{1, \ldots, d\}$
  for $j = d$ **downto** 1 **do**
    for all $i \in S$ **do**
      $\hat{\epsilon}_i \leftarrow \texttt{Residuals}(X_{S\setminus\{i\}}, X_i)$
      $p_i \leftarrow \texttt{TestIndependence}(X_{S\setminus\{i\}}, \hat{\epsilon}_i)$
    end for
    $i^* \leftarrow \operatorname{argmax} p_i$
    if $p_{i^*} < \alpha$ **then**
      return no consistent DAGs
    end if
    $\sigma_j \leftarrow i^*$
    $S \leftarrow S \setminus \{i^*\}$
  end for
  for $j = 1$ **to** d **do**
    $i \leftarrow \sigma_j$
    $\mathrm{pa}(i) \leftarrow \{\sigma_1, \ldots, \sigma_{j-1}\}$
    for $k = 1$ **to** $j - 1$ **do**
      $\hat{\epsilon}_i \leftarrow \texttt{Residuals}(X_{\mathrm{pa}(i)\setminus\{\sigma_k\}}, X_i)$
      if $\texttt{TestIndependence}(X_{\mathrm{pa}(i)}, \hat{\epsilon}_i) \geq \alpha$ **then**
        $\mathrm{pa}(i) \leftarrow \mathrm{pa}(i) \setminus \{\sigma_k\}$
      end if
    end for
  end for
output parent sets $(\mathrm{pa}(i))_{i \in \mathcal{V}}$

but their p-values are much lower. For this binary decision case, this is not so important, but if one has more than two variables or has a third possible decision ("neither variable causes the other"), the absolute p-values are important.

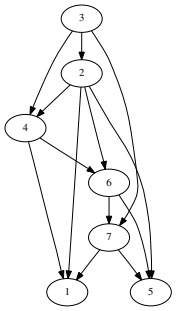
4. Efficient causal inference algorithm

In this Section, we propose a more efficient algorithm to find a causal model fitting the data than the algorithm that simply tests all possible DAGs. See Algorithm 1. It invokes two subroutines: $\texttt{Residuals}(X, Y)$, which fits Y as a function of X and returns the residuals (if X is empty, it should just return Y itself as the residuals); and $\texttt{TestIndependence}(X, Y)$, which tests independence of X and Y, returning the p-value corresponding to the null hypothesis of independence.

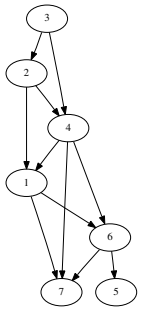
In the first sweep, a possible causal ordering $\sigma \in S_d$ of the variables is inferred (where S_d denotes the symmetric group consisting of all permutations of $\{1, \ldots, d\}$). In the second sweep, unnecessary arrows are removed. The result is a minimal DAG consistent with the data. The time complexity of the algorithm is $\mathcal{O}(d^2)$ if we count regression and independence tests as atomic operations. This should be compared with the super-exponential number of DAGs with d variables which have to be tested for the enumeration algorithm proposed in (Hoyer et al., 2009).

$X_1 = \sin(X_4^2) + X_2^2 + \cos(X_7) + E_1$	$E_1 \sim U(-0.1, 0.1)$
$X_2 = X_3^2 + E_2$	$E_2 \sim U(-0.5, 0.5)$
$X_3 = E_3$	$E_3 \sim U(-1.0, 1.0)$
$X_4 = \sin(X_2) + \sin(2X_3) + E_4$	$E_4 \sim U(-0.5, 0.5)$
$X_5 = \tanh(X_6 + X_7 + X_2) + E_5$	$E_5 \sim U(-0.2, 0.2)$
$X_6 = \sin(X_2) + \cos(2X_4) + E_6$	$E_6 \sim U(-0.5, 0.5)$
$X_7 = \cos(X_6 + X_3) + E_7$	$E_7 \sim U(-0.3, 0.3)$

(a) Probability distribution of ground truth



(b) Ground truth DAG



(c) Reconstructed DAG

Figure 2. Toy example consisting of 7 variables: (a) ground truth causal model; (b) DAG corresponding to the ground truth causal model; (c) reconstructed DAG based on a sample of $N = 300$ datapoints.

We show that Algorithm 1 is asymptotically consistent under the following assumptions:

(1) whenever $\{X_1, \ldots, X_l\}$ contains all the parents of Y and none of its descendants, the residuals $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ are independent of every set that contains no descendants of Y.

(2) whenever $\{X_1, \ldots, X_l\}$ contains a child of Y, independence of $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ and $\{X_1, \ldots, X_l\}$ is rejected.

(3) whenever there is a parent X of Y with $X \notin \{X_1, \ldots, X_l\}$ then $\texttt{Residuals}(\{X_1, \ldots, X_l\}, Y)$ is not independent of X.

Assumption (1) is satisfied if the joint distribution is generated by an *additive* noise model, because the noise of a variable is only relevant for the variable itself and its descendants. We conjecture that assumption (2) is satisfied in the generic case. This is suggested by (Hoyer et al., 2009, Theorem 1) regarding the two-variable case: generic additive noise models $X \to Y$ generate distributions that do not admit additive noise models $Y \to X$. Assumption (3) follows from causal faithfulness (Spirtes et al., 1993) because independence of the residual would imply $X \perp\!\!\!\perp Y \,|\, X_1, \ldots, X_l$, but conditional independence can only hold true for non-adjacent X, Y.

Table 1. Results on datasets of the `Cause-effect pairs` task for the NIPS 2008 Causality Competition (Mooij et al., 2008), using various regression methods. We report the empirical HSIC estimates for regression residuals and regressor values and their corresponding p-values. Two models are considered: $X_1 \to X_2$ (i.e., $X_2 = f_2(X_1) + E_2$ with $E_2 \perp\!\!\!\perp X_1$) where X_1 causes X_2, and the backwards model $X_2 \to X_1$ (i.e., $X_1 = f_1(X_2) + E_1$ with $E_1 \perp\!\!\!\perp X_2$) where X_2 causes X_1. If $p_{1\to 2} > p_{2\to 1}$, the model $X_1 \to X_2$ gives the best fit to the data, which we interpret as "X_1 causes X_2" (and *vice versa*).

Regularized HSIC-regression

Dataset	$p_{1\to 2}$	$p_{1\leftarrow 2}$	$\widehat{\text{HSIC}}_{1\to 2}$	$\widehat{\text{HSIC}}_{1\leftarrow 2}$	Decision	Ground truth
1	0.289823	$< 10^{-6}$	0.0012	0.0060	$\to$	$\to$
2	0.037262	0.014491	0.0020	0.0021	$\to$	$\to$
3	0.044745	0.002767	0.0019	0.0026	$\to$	$\to$
4	0.375563	0.011721	0.0011	0.0023	$\to$	$\leftarrow$
5	$< 10^{-6}$	0.159925	0.0028	0.0005	$\leftarrow$	$\leftarrow$
6	$< 10^{-6}$	$< 10^{-6}$	0.0032	0.0026	?	$\to$
7	$< 10^{-6}$	0.271836	0.0021	0.0005	$\leftarrow$	$\leftarrow$
8	0.000002	$< 10^{-6}$	0.0015	0.0017	$\to$	$\to$

Regularized linear least-squares regression

Dataset	$p_{1\to 2}$	$p_{1\leftarrow 2}$	$\widehat{\text{HSIC}}_{1\to 2}$	$\widehat{\text{HSIC}}_{1\leftarrow 2}$	Decision	Ground truth
1	0.034679	$< 10^{-6}$	0.0020	0.0067	$\to$	$\to$
2	$< 10^{-6}$	$< 10^{-6}$	0.0074	0.0075	?	$\to$
3	0.008914	0.000589	0.0023	0.0029	$\to$	$\to$
4	0.000011	0.002146	0.0040	0.0028	$\leftarrow$	$\leftarrow$
5	$< 10^{-6}$	0.024624	0.0047	0.0007	$\leftarrow$	$\leftarrow$
6	$< 10^{-6}$	$< 10^{-6}$	0.0059	0.0053	?	$\to$
7	$< 10^{-6}$	0.019524	0.0050	0.0008	$\leftarrow$	$\leftarrow$
8	$< 10^{-6}$	$< 10^{-6}$	0.0096	0.0029	?	$\to$

Gaussian Process regression

Dataset	$p_{1\to 2}$	$p_{1\leftarrow 2}$	$\widehat{\text{HSIC}}_{1\to 2}$	$\widehat{\text{HSIC}}_{1\leftarrow 2}$	Decision	Ground truth
1	0.016375	$< 10^{-6}$	0.0022	0.0077	$\to$	$\to$
2	$< 10^{-6}$	$< 10^{-6}$	0.0078	0.0074	?	$\to$
3	0.007889	0.000702	0.0023	0.0029	$\to$	$\to$
4	0.000055	0.010831	0.0036	0.0023	$\leftarrow$	$\leftarrow$
5	$< 10^{-6}$	0.014970	0.0048	0.0008	$\leftarrow$	$\leftarrow$
6	$< 10^{-6}$	$< 10^{-6}$	0.0057	0.0052	?	$\to$
7	$< 10^{-6}$	0.012321	0.0053	0.0008	$\leftarrow$	$\leftarrow$
8	$< 10^{-6}$	$< 10^{-6}$	0.0097	0.0032	?	$\to$

To obtain a causal ordering, we search for a variable X_i for which the regression on the remaining $d-1$ variables (i.e., on $X_{S\setminus\{i\}}$) yields a residual that is independent of $X_{S\setminus\{i\}}$. Every childless node will be accepted by assumption (1), which shows that our search cannot fail. Conversely, X_i is childless by assumption (2), and is thus the last variable with respect to an appropriate ordering of nodes. Since X_i is therefore causally irrelevant for the remaining variables we can repeat the same procedure with $d-1$ variables and so on, until we have identified the first node. Induction over d shows that we have indeed found an allowed causal ordering. The corresponding complete DAG $\mathcal{G}'$ differs from the true graph $\mathcal{G}$ only by unnecessary links.

To remove irrelevant parents, we use the following iterative method. For every X_i, let $\text{pa}(i)$ be the set of parents with respect to the current preliminary graph.

For every $Y \in \mathrm{pa}(i)$, compute the regression of X_i on $\mathrm{pa}(i) \setminus \{Y\}$ and check whether the residual is still independent of $\mathrm{pa}(i)$. If Y is a true parent, independence will be rejected by assumption (3). Otherwise it will be accepted by assumption (1). Hence we keep exactly the links that are also present in $\mathcal{G}$.

To complete the consistency proof, the conjecture (assumption (2)) has to be proven. We consider this beyond the scope of the current work. In the next Subsection we give some empirical evidence that supports the conjecture.

4.1. Experimental results

We consider a toy example consisting of seven variables, specified in Figure 2(a). We applied Algorithm 1 to a sample of N variables, using regularized HSIC regression as the regression method in combination with the HSIC independence test. The number of regressions needed is reduced from 448 for the naïve algorithm reported in (Hoyer et al., 2009) to 48 for Algorithm 1; moreover, we only need to perform 48 independence tests, instead of at least one for each of the approximately 1 billion DAGs of 7 variables. Using 300 data points, a few errors are made, as shown in Figure 2(c), but the resulting DAG is already close to the ground truth, and becomes closer for larger N.

5. Conclusions

We introduced a novel regression method that minimizes the dependence of the residuals and the regressors. We successfully applied the method, using the HSIC independence measure, to causal inference tasks. We expect that the regression method may prove to be more generally useful, in particular whenever the noise distribution is unknown.

Acknowledgments

We thank Arthur Gretton for fruitful discussions.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Geiger, D., & Heckerman, D. (1994). Learning Gaussian networks. *Proc. of the 10th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 235–243).

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. *Algorithmic Learning Theory: 16th International Conference (ALT 2005)* (pp. 63–78).

Györfi, L., Kohler, M., Krzyżak, A., & Walk, H. (2002). *A distribution - free theory of nonparametric regression*. New York: Springer Verlag.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21 (NIPS*2008)*, 689–696.

Liu, D. C., & Nocedal, J. (1989). On the limited memory method for large scale optimization. *Mathematical Programming B*, *45*, 503–528.

Mooij, J., Janzing, D., & Schölkopf, B. (2008). Distinguishing between cause and effect. `http://www.kyb.tuebingen.mpg.de/bs/people/jorism/causality-data/`.

Okazaki, N., & Nocedal, J. (2008). libLBFGS: C library of limited-memory BFGS (L-BFGS). `http://www.chokkan.org/software/liblbfgs/`.

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

Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rasmussen, C. E., & Williams, C. (2007). GPML code. `http://www.gaussianprocess.org/gpml/code`.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels*. MIT Press.

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer-Verlag. (2nd ed. MIT Press 2000).

Steinwart, I. (2002). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.

Zhang, K., & Hyvärinen, A. (2008). Distinguishing causes from effects using nonlinear acyclic causal models. `http://videolectures.net/coa08_zhang_hyvarinen_dcfeu/`. Talk at the NIPS 2008 Workshop on Causality: objectives and assessment.