Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference

SETH R. FLAXMAN, Machine Learning Department and Event and Pattern Detection Laboratory, H. J. Heinz III College, Carnegie Mellon University
DANIEL B. NEILL, Event and Pattern Detection Laboratory, H. J. Heinz III College, Carnegie Mellon University
ALEXANDER J. SMOLA, Machine Learning Department, Carnegie Mellon University, Marianas Labs

In applied fields, practitioners hoping to apply causal structure learning or causal orientation algorithms face an important question: which independence test is appropriate for my data? In the case of real-valued iid data, linear dependencies, and Gaussian error terms, partial correlation is sufficient. But once any of these assumptions is modified, the situation becomes more complex. Kernel-based tests of independence have gained popularity to deal with nonlinear dependencies in recent years, but testing for conditional independence remains a challenging problem. We highlight the important issue of non-iid observations: when data are observed in space, time, or on a network, "nearby" observations are likely to be similar. This fact biases estimates of dependence between variables. Inspired by the success of Gaussian process regression for handling non-iid observations in a wide variety of areas and by the usefulness of the Hilbert-Schmidt Independence Criterion (HSIC), a kernel-based independence test, we propose a simple framework to address all of these issues: first, use Gaussian process regression to control for certain variables and to obtain residuals. Second, use HSIC to test for independence. We illustrate this on two classic datasets, one spatial, the other temporal, that are usually treated as iid. We show how properly accounting for spatial and temporal variation can lead to more reasonable causal graphs. We also show how highly structured data, like images and text, can be used in a causal inference framework using a novel structured input/output Gaussian process formulation. We demonstrate this idea on a dataset of translated sentences, trying to predict the source language.

Categories and Subject Descriptors: A.1.1 [**Machine Learning**]: Artificial Intelligence

General Terms: Kernels, Independence Tests, Structured Data, Causal Inference

Additional Key Words and Phrases: Reproducing kernel Hilbert space, Gaussian process, causal structure learning, causal inference

ACM Reference Format:
Seth R. Flaxman, Daniel B. Neill, and Alexander J. Smola. 2015. Gaussian processes for independence tests with non-iid data in causal inference. ACM Trans. Intell. Syst. Technol. 7, 2, Article 22 (November 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2806892

This work was partially supported by the National Science Foundation, grants IIS-0916345, IIS-0911032, and IIS-0953330.
Authors' addresses: S. R. Flaxman and A. J. Smola, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213; emails: sflaxman@cs.cmu.edu, alex@smola.org; D. B. Neill, Carnegie Mellon University, 4800 Forbes Ave, Pittsburgh, PA 15213; email: neill@cs.cmu.edu.

DOI: http://dx.doi.org/10.1145/2806892

1. INTRODUCTION

It is common for observational data to violate the typical assumption of independent and identically distributed (iid) observations. For instance, data about user behavior usually has a temporal structure. Environmental measurements often have both temporal and spatial structure. This structure poses a particular problem when inferring dependence between random variables. As a motivating example, consider two independently generated autoregressive time series AR(1) on random variables X and Y according to the model

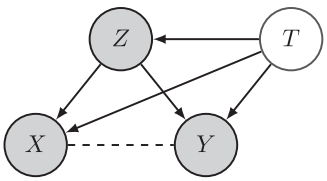
$$x_t = 0.9 \cdot x_{t-1} + \epsilon_{x,t} \text{ and } y_t = 0.8 \cdot y_{t-1} + \epsilon_{y,t} \text{ where } \epsilon_{x,t}, \epsilon_{y,t} \sim \mathcal{N}(0, 1).$$

That is, X and Y are *independent* time series, each of which is corrupted at each step by adding normally distributed iid random variables. Despite the fact that X and Y are independent, the Pearson correlation between X and Y may be large in magnitude due to the underlying autocorrelation structure of each time series, as shown in Figure 1. Whereas Fisher's z-transformation can be used to derive the distribution of the Pearson correlation statistic under linear independence, this assumes iid observations. But in the case of X and Y, our observations are neither independent nor identically distributed. The general guidance in the time series literature is to fit an appropriate autoregressive model to the data and to obtain residuals from this model [Box et al. 2008]. The intuition is that this *pre-whitening* should yield residuals that are iid, after which independence testing proceeds as usual. We formalize this notion in this article.

Many algorithmic approaches to causal inference rely on statistical tests of independence between variables. The most popular default methods are the Fisher z-score, Pearson correlation (and partial correlation) [Pearson 1983], and, more recently, the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al. 2008]. More generally, the entire framework of graphical models for causal inference [Pearl 2009] relies crucially on assumptions about d-separation in graphs, and testing these assumptions with observational data requires applying a valid conditional independence test.

As in the case just described, each of these tests is prone to spuriously reporting large correlations when used on non-iid data due to the underlying autocorrelation structure. Furthermore, causal inference tools such as the PC algorithm [Spirtes et al. 2001] rely on conditional independence tests, asking whether $X \perp\!\!\!\perp Y|Z$. It is not clear a priori what effect non-iid data will have in this case. If the true model is that $X \perp\!\!\!\perp Y|Z$, underlying autocorrelation affecting both X and Y might lead us to believe that $X \not\perp\!\!\!\perp Y|Z$.

Example 1.1. Consider the graphical model here: It illustrates the problem of confounding due to non-iid data. T represents time. Shaded nodes X, Y, and Z are observed, and T may be either observed or unobserved.



The true causal relationship is that $X \perp\!\!\!\perp Y|Z$. However, if T is unobserved, it acts as a latent confounding variable, meaning that a spurious edge may be inferred between X and Y (i.e., a conditional independence test rejects the hypothesis $X \perp\!\!\!\perp Y|Z$). Once T is observed and controlled for, a conditional independence test will correctly conclude that $X \perp\!\!\!\perp Y|Z$.

We are not the first to point out that *every* scientific observation was generated at some specific point in time [Cressie and Wikle 2011]. But, in most cases, this

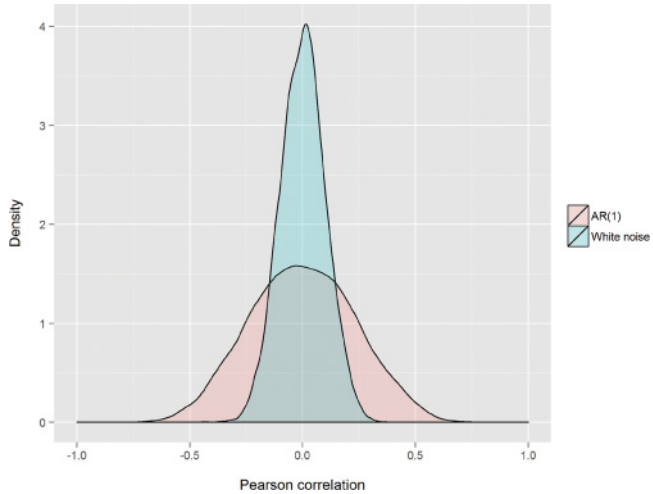


Fig. 1. Pairs of time series processes were generated 10,000 times, with $n = 100$ observations for each. Each time, the Pearson correlation between the two processes was calculated. When both pairs were white noise (i.e. iid $\sim \mathcal{N}(0, 1)$), 95% of the correlations were between -0.2 and 0.2. But when the two pairs were independently generated AR(1) processes, with $x_t = 0.9x_{t-1} + \epsilon_{x,t}$ and $y_t = 0.8y_{t-1} + \epsilon_{y,t}$, only 60% of the correlations were between -0.2 and 0.2. This is an example of the way that temporal autocorrelation can bias an independence test (in this case, linear independence tested with Pearson correlation) that assumes iid data: Many more correlations are significant than we would expect by chance. A simple correction is to first pre-whiten x_t and y_t by fitting an AR(1) model and obtaining residuals.

information is discarded for convenience. In Section 5, we consider a spatial and a temporal dataset that are usually analyzed as if the data were iid. We perform tests (Moran's I for spatial data and partial autocorrelation for time series data [Moran 1950]) that conclusively reject the hypothesis that the observations are iid, and we show how causal inference algorithms yield more reasonable results after controlling for the underlying spatial and temporal autocorrelation. Our framework also opens up the possibility of causal inference with structured data, and we develop a novel approach to Gaussian process (GP) regression and independence testing that we apply to textual data to determine which language is a translation of another for pairs of texts.

We propose a simple framework for using GP regression to reduce questions about conditional independence with non-iid data to questions about unconditional independence with iid data, which can be answered with HSIC. Mechanically, our approach is similar to that taken in recent papers on bivariate causal orientation [Peters et al. 2013], in which it is termed *Regression with Subsequent Independence Test* (RESIT), but the motivation is different. The most similar approach to ours is the conditional independence tests proposed by Moneta et al. [2011], which are specifically designed for time series data modeled by a Vector Autoregression (VAR) model and thus not directly applicable to, for example, spatial data. Insofar as our method combines kernel-based independence tests with the PC algorithm, it is similar to the Kernel PC algorithm proposed by Tillman et al. [2009], but our conditional independence tests are different. The strategy we propose is straightforward, generally applicable wherever GPs can be used, and it works for both pre-whitening non-iid data and for testing conditional independence.

2. CONTRIBUTIONS

2.1. Approach

The centerpiece of the approach is to use regression to remove dependence on space, time, or a set of conditioning variables. We assume that we have random variables

(X, Y, Z), observed at locations S (in time, space, or on a network). Conditional independence testing then proceeds in the following three steps:

(1) We first use separate GP regressions of $X|S$, $Y|S$ and $Z|S$ to obtain residuals

$$r_x = x - \hat{\mathbf{E}}[x|s] \text{ and } r_y = y - \hat{\mathbf{E}}[y|s] \text{ and } r_z = z - \hat{\mathbf{E}}[z|s], \tag{1}$$

thus pre-whitening each variable and eliminating its dependence on S.

(2) Next, we again use GP regression to obtain residuals

$$\epsilon_{xz} = r_x - \hat{\mathbf{E}}[r_x|r_z] \text{ and } \epsilon_{yz} = r_y - \hat{\mathbf{E}}[r_y|r_z] \tag{2}$$

from regressing both r_x and r_y on r_z separately.

(3) Finally, we use HSIC to test for independence:

$$\epsilon_{xz} \perp\!\!\!\perp \epsilon_{yz}. \tag{3}$$

At a mechanical level, in each step we use GP regression to obtain residuals for which we have controlled for variation—in the case of the dependence structure of the data, we are controlling for, say, temporal variation. In the case of conditional independence, we are controlling for the variation due to variable Z.

Strategies like these are standard practice in statistical modeling. In econometrics, this approach is justified by the Frisch-Waugh-Lovell theorem [Frisch and Waugh 1933] (which was originally stated in a time series context), which proves that in the case of linear regression, partial correlations can be calculated by finding the correlations between residuals. In the spatial statistics and time series literature, pre-whitening by fitting models and obtaining residuals, removing trends, and taking first differences are all standard approaches [Box et al. 2008]. However, to our knowledge, a full formulation of this strategy, combining a nonparametric regression and independence test, has not been stated explicitly before. Moreover, beyond the case of linear models with Gaussian noise, the conditions under which it holds are not known. In Section 4.1, we state precise conditions under which our test is valid.

We believe that our method can serve as a default template when testing for conditional independence with non-iid data. It is equally useful as a simple method for testing for conditional independence even when observations are iid, in which case the pre-whitening step can be skipped. We highlight a few reasons for relying on GP regression for pre-whitening and conditioning rather than using parametric tests or relying solely on kernel-based tests:

(1) GPs provide a principled Bayesian approach. Yet, for regression, their convenient analytic form means that hyperparameters can be learned much more efficiently than in many other fully Bayesian models since we can integrate out additive noise. This provides considerable computational savings and increased numerical accuracy.
(2) A variety of packages already exist to fit GP regression [Rasmussen and Nickisch 2010; Kalaitzis et al. 2013; Vanhatalo et al. 2013; Karatzoglou et al. 2004], and these perform inference using either optimization methods, grid search, or sampling (MCMC) strategies.[1]
(3) In the case of time series and especially spatial data, the GP framework is a long-standing, proven method, typically referred to as "kriging" in geostatistics [Salkauskas 1982]. In applied fields where it has been used, practitioners are adept at designing appropriate covariance functions (Mercer kernels) adapted to their problem domains. For instance, the Matérn kernel is a popular choice. With spatiotemporal

[1]See also http://www.gaussianprocess.org/#code for more details.

data, much recent work has focused on designing classes of sophisticated nonseparable and nonstationary covariance functions for capturing complex dependencies [Gneiting et al. 2007]. These covariance functions could be directly imported into the kernel-based statistical tests, but their use requires model-checking and diagnostics. Recent work suggests that complicated time series dynamics can be automatically fit through combinations of covariances [Wilson and Adams 2013; Duvenaud et al. 2013].

(4) By design, GPs allow for easy graphical model-checking: Diagnostic plots can be inspected to check for autocorrelation and overfitting.
(5) Our new GP formulation for structured inputs and outputs, as introduced in Section 4.4, opens up the possibility of conditional independence and causal inference with structured data such as text, images, and anything else on which a Mercer kernel can be defined.
(6) In the case of real-valued data, our formulation allows for testing conditional independence without first discretizing the conditioning set. This is useful because discretization is fraught with information loss—we may lose the relevant time scale or we might even introduce dependence due to the quantization level inherent in binning.

2.2. Related Work

We are aware of only one general test [Zhang et al. 2008] for unconditional independence with non-iid data. It requires precisely specifying the dependence structure of the data as a graphical model and then decomposing this model into cliques, exploiting the connection between the exponential family of distributions and kernels over graphical models. The analysis is by no means simple—for instance, it has not been extended to a lattice structure; this is unfortunate because assuming that points are on a lattice is a basic starting point in the spatial statistics literature.

In the case of conditional independence, several tests have been proposed, including a test based on characteristic functions [Su and White 2007], the Normalized Conditional Cross-Covariance Operator (NOCCO) [Fukumizu et al. 2007], Kernel-based Conditional Independence (KCI) [Zhang et al. 2011], a scale-invariant measure [Reddi and Póczos 2013], a scalable method called Conditional Correlation Independence (CCI) [Ramsey 2014], and a permutation-based conditional independence test [Doran et al. 2014]. However, these tests will all be biased for non-iid data, just like the unconditional tests. Although CCI does not address the non-iid case, for conditional independence, it takes an approach with a similar flavor to our method and makes similar asymptotic claims. However, CCI is based on a finite basis expansion, so consistency only holds in the limit as the number of basis functions goes to infinity along with the number of samples, whereas we use a consistent nonparametric regression method, so consistency holds in the large-sample limit.

A few works focus specifically on the time series domain, but it is not clear if they can be generalized to spatial or continuous/partially observed time series data. Moneta et al. [2011] proposed a conditional independence test, appropriate for time series data that can be modeled as a VAR process, based on calculating divergence between density estimates using smoothing kernels. Besserve et al. [2013] proposed a powerful kernel cross-spectral density operator for characterizing independence between time series and Chwialkowski and Gretton [2014] explored the behavior of HSIC for random processes (e.g., time series data), showing a new consistent estimate of the p-value for non-iid data, but neither of these works address the conditional independence case.

Even with iid data, these tests have not found widespread application. Closed-form distributions under the null are not available, except in the cases of KCI and the test in Su and White [2007], so permutation testing is required. Valid permutation testing

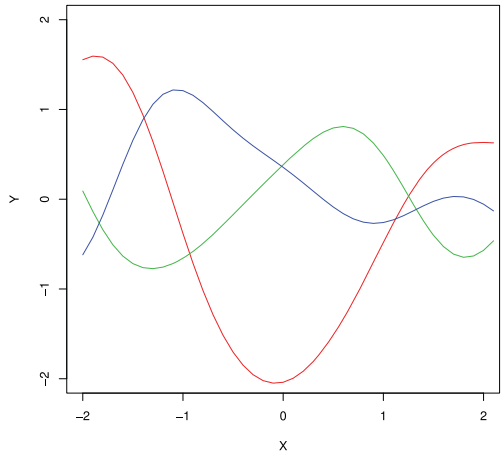


Fig. 2. Three draws from a GP prior with mean 0 and Gaussian RBF covariance function.

of $X \perp\!\!\!\perp Y|Z$ must preserve the marginal structure $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$. Assuming that Z is categorical, for each value of Z, one can consider permuting X. But when Z is real-valued, discretization is necessary first. Clustering is a common approach, as in Tillman et al. [2009]. By contrast, our regression-based approach naturally handles categorical, real-valued, and even structured (image or text) data.

3. BACKGROUND

3.1. Gaussian Processes

A GP is a stochastic process over an index set $\mathcal{X}$. It is entirely defined by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. These two functions are chosen such as to jointly define a normal distribution whenever we draw $f|X$ from a $\mathcal{GP}(\mu, k)$ on a finite set of locations $X := \{x_1, \ldots x_n\}$. More specifically, we have

$$f|X \sim \mathcal{N}(\mu(X), k(X, X)) \text{ where } \mu(X)_i = \mu(x_i) \text{ and } [K(X, X)]_{ij} = k(x_i, x_j). \tag{4}$$

By its very construction this means that $\mu(X)$ is an m dimensional vector, and $k(X, X) \in \mathbb{R}^{d \times d}$ is a positive semidefinite matrix. In other words, k generates symmetric matrices with nonnegative eigenvalues.

Note that this does *not* introduce a function over X. In fact, although there are some kernels leading to smooth processes [Wahba 1990], this is in general not the case. In particular, quite often the realization $f(x)$ is nonsmooth whereas its prior is smooth. A well-known example is the Brownian Bridge.

Note that there is a subtle difference between functions and function values in the construction of a GP. For any infinite-dimensional GP (i.e., where the rank of $k(X, X)$ is unbounded), it is only possible to evaluate the GP pointwise. The technical challenge is that distributions over infinite-dimensional objects are nontrivial to define. Evaluating a GP on a finite number of locations sidesteps the entire problem.

As an illustration, consider a GP with mean function $\mu = 0$ and Gaussian Radial Basis Function (RBF) kernel $k(x_i, x_j) = e^{-\|x_i - x_j\|^2}$. These parameters give a GP from which we can draw a realization. Since we want to know its value for a range of locations, we draw f for a grid of points. By construction, they are drawn from a multivariate Gaussian distribution with mean $\mu = 0$ and covariance K.

Three different draws are shown in Figure 2. In a Bayesian framework, these should be thought of as draws from the prior distribution over joint values $f(x)$ before seeing any data. How do we update our prior given observations $Z = (X, Y)$? We start

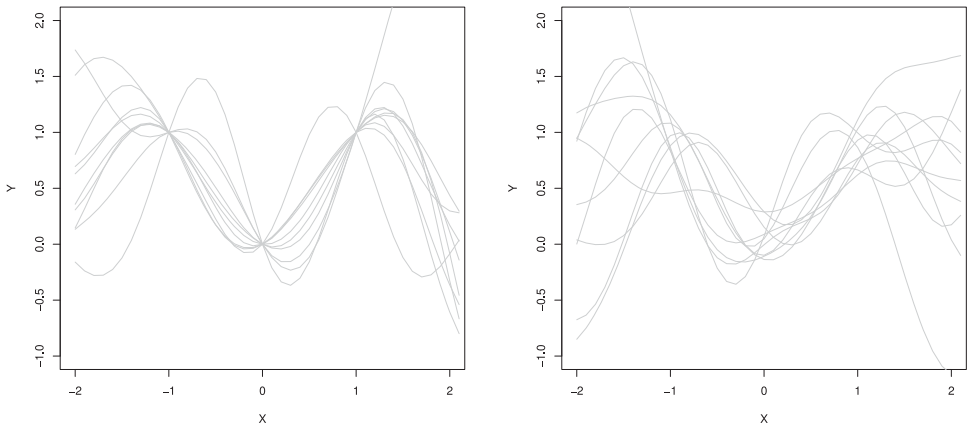


Fig. 3. Draws from a Gaussian Process posterior with Gaussian RBF kernel after observations at $\{(-1,1),(0,0),(1,1)\}$. Left: Noise-free observations. Right: Noisy observations with $\sigma^2 = 0.2$. Notice the difference in terms of uncertainty at the locations of measurement and the relative similarity otherwise.

by specifying the joint distribution over both observed outputs Y and unobserved outputs Y^*:

$$\begin{bmatrix} Y & Y^* \end{bmatrix} \sim \mathcal{N}(\mu(\vec{x}), K),$$

where we can calculate $K(x_i, x_j)$ for any pair of x's, observed or unobserved; that is:

$$K = \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}.$$

Since we've observed (X, Y), we can find the conditional distribution using the properties of multivariate Gaussian distributions (see, e.g., Rasmussen and Williams [2006]):

$$Y^*|Y \sim \mathcal{N}(K(X^*,X)K(X,X)^{-1}Y, K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*))$$

We give an illustration in Figure 3, where the observations $(-1,1),(0,0),(1,1)$ are shown in black circles and 10 posterior function draws f^* are plotted. Notice that there is no uncertainty at the observed points.

In some cases, like modeling computer simulations, this noise-free behavior might be desirable, but for real data generated by nature we need to include an extra noise term. If we believe our noise is iid, we can use the following covariance function:

$$k(x_i, x_j) = e^{-\|x_i - x_j\|^2} + \sigma^2 \delta_{i,j}.$$

What does this extra variance σ^2 (called the "nugget" in geostatistics) do? It only appears when $i = j$, meaning that the diagonal of the covariance matrix has entries $1 + \sigma^2$ instead of 1. If we use the same K as before, we have:

$$Y^*|Y \sim \mathcal{N}(\bar{\mu}, \bar{K}) \tag{5}$$

where $\bar{\mu} = K(X^*,X)(K(X,X) + \sigma^2 I)^{-1}Y$

$$\bar{K} = K(X^*,X^*) - K(X^*,X)(K(X,X) + \sigma^2 I)^{-1}K(X,X^*).$$

Here, $\bar{K}$ is the well-known Schur complement of the joint covariance matrix over X and X'. Note that the noise term σ^2 is only used for observed data Y. If we use this prior, we can draw 10 posterior functions as before. In Figure 3 (right), we plotted these function draws. Notice that there is now some uncertainty, controlled by the parameter σ^2, at the observed points: Even if we were to observe $y|x$ at the same location repeatedly, we would have no assurance that the observations would be identical.

3.2. Hilbert-Schmidt Independence Criterion

Given observations from a joint distribution $P(X, Y)$, the HSIC [Gretton et al. 2005, 2008] is a statistical test for the null hypothesis of independence: $X \perp\!\!\!\perp Y$. It uses kernel embeddings of probability distributions to compare the joint distribution $P(X, Y)$ to the product of the marginal distributions $P(X)P(Y)$. After specifying kernels $k(x, x')$ with Hilbert space $\mathcal{H}_X$ and $\ell(y, y')$ with Hilbert space $\mathcal{H}_Y$, HSIC maximizes a kernelized covariance or, equivalently, the distance between the mean embedding of the joint distribution in Hilbert space and the product of the mean embeddings of the marginal distributions [Smola et al. 2007]:

$$\sup_{\|f\|,\|g\|\leq 1} \mathbf{E}_{x,y}[f(X)g(Y)] - \mathbf{E}_x[f(X)]\mathbf{E}_y[g(Y)].$$

For general f, g, this expression is 0 if and only if $P(X, Y) = P(X)P(Y)$, which is equivalent to $X \perp\!\!\!\perp Y$.

The test statistic that maximizes this expression if $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$ is given by

$$\text{HSIC} = \left\|\mathbf{E}_{x,y}\left[k(x,\cdot)\ell(y,\cdot)\right] - \mathbf{E}_x\left[k(x,\cdot)\mathbf{E}_y\ell(y,\cdot)\right]\right\|^2.$$

For so-called *characteristic* kernels [Sriperumbudur et al. 2010], such as the RBF kernel, this statistic is 0 if and only if $P(X, Y) = P(X)P(Y)$. An estimator can be derived:

$$\widehat{\text{HSIC}} = \frac{1}{n^2}\sum_{i,j} k(x_i, x_j)\ell(y_i, y_j) - \frac{2}{n^3}\sum_{i,j,q} k(x_i, x_j), \ell(y_i, y_q) + \frac{1}{n^4}\sum_{i,j,q,r} k(x_i, x_j)\ell(y_q, y_r).$$

This estimator can be written compactly in terms of Gram matrices K and L:

$$\widehat{\text{HSIC}} = \frac{1}{n^2}\text{tr}(KHLH),$$

where $K_{ij} = k(x_i, x_j)$, $L_{ij} = \ell(y_i, y_j)$, and $H = I - \frac{1}{n}11^T$ is a centering matrix. More details are in Gretton et al. [2012].

The distribution of HSIC under the null can be obtained by randomization testing: Given pairs (x_i, y_i), we shuffle the y's and recompute $\widehat{\text{HSIC}}$. Gretton et al. [2008] gives an asymptotic result based on the Gamma distribution; Zhang et al. [2011] gives a test based on the eigenvalues of the kernel matrices.

HSIC has been used to test for independence between structured data. In Gretton et al. [2008], a string kernel was used to test for independence between French and English sentences. Inspired by this approach, we develop a novel input/output GP regression method in Section 4.4, testing for independence with HSIC using residuals in feature space.

3.3. Causal Inference Methods

We focus on two classes of causal inference methods: constraint-based causal structure learning algorithms exemplified by the PC algorithm and bivariate causal orientation methods (i.e., the Additive Non-Gaussian (ANG) framework [Hoyer et al. 2008] and the Continuous Additive Noise Model (CANM) framework [Peters et al. 2013]).

The PC algorithm learns an equivalence class of Partially Directed Acyclic Graphs (PDAGs), which are consistent with the conditional independencies entailed by the data, as tested with statistical tests for conditional independence. After learning this "skeleton," the algorithm finds V-structures, also known as colliders, of the form $A \rightarrow B \leftarrow C$ that are consistent with the learned conditional independencies and orients edges accordingly. For example, a V structure $A \rightarrow B \leftarrow C$ would be implied by $A \perp\!\!\!\perp C$ and $A \not\perp\!\!\!\perp C|B$. Finally, the algorithm orients any other edges it can to be consistent

with the edges it has already oriented, so long as these orientations do not introduce any new V structures or cycles. Once a PDAG is learned, independence relations can be read off the graph using the rules of d-separation. For a detailed discussion of the PC algorithm, see Spirtes et al. [2001] and for causal DAGs and d-separation see, Pearl [2009].

The bivariate causal orientation methods compare two models, a forward model: $Y = f_1(X) + \epsilon_1$ and a backwards model: $X = f_2(Y) + \epsilon_2$. After fitting nonparametric regressions to obtain residuals $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$, an independence test such as HSIC is used to test whether $\hat{\epsilon}_1 \perp\!\!\!\perp X$ and $\hat{\epsilon}_2 \perp\!\!\!\perp Y$. If, for example, $\hat{\epsilon}_1 \perp\!\!\!\perp X$ but $\hat{\epsilon}_2 \not\perp\!\!\!\perp Y$, we reject the backward model and retain the forward model, $X \to Y$. For a detailed discussion see Peters et al. [2013].

4. THEORETICAL DEVELOPMENT

4.1. Testing Conditional Independence by Regression and Unconditional Independence

We start by assuming both faithfulness and the Markov condition, the same assumptions made for the PC algorithm:

Faithfulness. There exists a causal DAG G and a probability distribution over random variables X, Y, Z such that if $X \perp\!\!\!\perp Y|Z$, then X and Y are d-separated by Z in graph G
Markov. If X and Y are d-separated by Z in G, then $X \perp\!\!\!\perp Y|Z$.

Second, we assume that we have access to a conditional regression estimator to remove the dependence on Z from X and Y. More specifically, we assume that this can be done in an additive fashion:

Consistent Regressors. We assume that we have consistent nonparametric regressors $\hat{m}_x(Z)$ and $\hat{m}_y(Z)$ that converge to $\mathbf{E}[X|Z]$ and $\mathbf{E}[Y|Z]$, respectively, such as GP regression.
Additive Noise Model. If Z is the cause of X or Y, we assume an additive independent noise model. That is, if Z causes X (respectively Y), then $X = f(Z) + \epsilon$ where $Z \perp\!\!\!\perp \epsilon$. Notice that we are not assuming in this case that $Y \perp\!\!\!\perp \epsilon$ or that the noise is always additive. For example, if the true structure is $X \leftarrow Z \leftarrow Y$, then we assume $X = f(Z) + \epsilon$, but we do not assume $Y = g(Z) + \epsilon_2$ or $Z = g(Y) + \epsilon_2$.

Finally, we assume that we have a valid method for testing unconditional independence between random variables, such as HSIC. Given these assumptions, our method can be summarized in the following simple algorithm:

(1) Obtain residuals $\epsilon_{xz} = X - \hat{m}_x(Z)$ and $\epsilon_{yz} = Y - \hat{m}_y(Z)$
(2) Test whether $\epsilon_{xz} \perp\!\!\!\perp \epsilon_{yz}$.

We claim that $\epsilon_{xz} \perp\!\!\!\perp \epsilon_{yz} \iff X \perp\!\!\!\perp Y|Z$.

We remark upon the assumptions underlying our method. As explained here, Choi and Schervish [2007] demonstrate almost sure convergence for GP regression under mild conditions, whereas Van Der Vaart and Van Zanten [2011] provide convergence rates for GP regression. Additive noise models underlie many standard regression techniques such as linear regression, kernel ridge regression, GP regression, and generalized additive models. Furthermore, it is straightforward to test this assumption in our framework: If we assume that $X = f(Z) + \epsilon$, we can check this assumption by using GP regression to regress X on Z to estimate $\hat{\epsilon}$. Then we use HSIC to check whether $\hat{\epsilon} \perp\!\!\!\perp Z$.

As discussed in Section 4.3, an application of our method to synthetic data is illustrated in Figure 5 in the case where Z represents time.

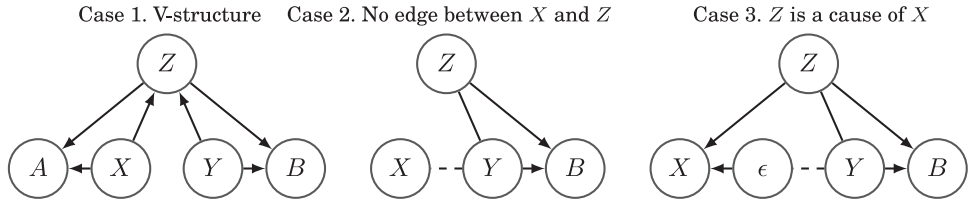


Fig. 4. Three cases of dependence between (X, Y, Z), corresponding to the cases in the proof of Theorem 4.1 that $X \perp\!\!\!\perp Y|Z$ if and only if $X - \mathbf{E}[X|Z] \perp\!\!\!\perp Y - \mathbf{E}[Y|Z]$. We define auxiliary variables $A := X - \mathbf{E}[X|Z]$ and $B := Y - \mathbf{E}[Y|Z]$, which are uniquely determined by their parents. **Case 1**: we have a V-structure $X \to Z \leftarrow Y$, so we see that $X \not\perp\!\!\!\perp Y|Z \Rightarrow X - \mathbf{E}[X|Z] \not\perp\!\!\!\perp Y - \mathbf{E}[Y|Z]$ because A and B are d-connected. **Case 2**: If there is no edge between X and Z, any path from X to B must go through Y. **Case 3**: Z and ϵ cause X, so the only possible path from ϵ to B is through Y.

THEOREM 4.1. *Given structural assumptions of faithfulness and the Markov assumptions, and assuming that we have consistent regressors with an additive noise model, whenever Z is a cause of X or Y, it follows that*

$$X \perp\!\!\!\perp Y|Z \text{ if and only if } X - \mathbf{E}[X|Z] \perp\!\!\!\perp Y - \mathbf{E}[Y|Z].$$

PROOF. We consider three cases for the structure of the causal graph G corresponding to the joint distribution of X, Y, and Z here. For each, we prove both the forward and reverse directions of the theorem. The associated graphical models are given in Figure 4. Our three cases are exhaustive due to symmetry (i.e., given three variables, we might need to switch the variables called X and Y) and the fact that they cover all possible dependencies in a DAG between X and (Y, Z) and all possible dependencies between Z and Y.

Case 1. Assume that we have a graph G with V-structure as in Figure 4

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

This immediately implies $X \not\perp\!\!\!\perp Y|Z$, so we do not need to prove anything for the forward direction. We prove the reverse direction by contradiction. Thus, we assume

$$X - \mathbf{E}[X|Z] \perp\!\!\!\perp Y - \mathbf{E}[Y|Z] \text{ but } X \not\perp\!\!\!\perp Y|Z,$$

and specifically this is because we have the V-structure $X \to Z \leftarrow Y$. Adding a new set of variables $A := X - \mathbf{E}[X|Z]$ with parents X and Z and $B := Y - \mathbf{E}[Y|Z]$ with parents Y and Z to the DAG, as shown in Figure 4, does not change the model since these random variables are entirely determined by (X, Z) and (Y, Z) respectively. Now we see that the path $A \leftarrow Z \to B$ d-connects A and B. By the faithfulness assumption, it follows that $A \not\perp\!\!\!\perp B$, which is a contradiction.

Case 2. If there is no edge between Z and X or between Z and Y or both, the test reduces to that of testing unconditional independence between X and Y. Without loss of generality, let us assume there is no edge between X and Z. $\mathbf{E}[X|Z]$ is a constant, call it c, so $X - \mathbf{E}[X|Z] = X - c$. As before, add the auxiliary variable $B := Y - \mathbf{E}[Y|Z]$ with parents Y and Z to the DAG, as in the Figure 4, Case 2. Then we are testing whether $X - c \perp\!\!\!\perp B$. Since c is constant, this holds if and only if $X \perp\!\!\!\perp B$. Finally, $X \perp\!\!\!\perp B$ if and only if $X \perp\!\!\!\perp Y$: If X and Y are d-connected by a path p, then we can add the edge from Y to B to the path p to make X and B d-connected. If instead X and Y are d-separated, then so are X and B because any path from X to B must go through Y.

Case 3. Z is a cause of X or Y or both, so assume without loss of generality that Z is a cause of X. Then, by assumption, we can write $X = f(Z) + \epsilon$ with $Z \perp\!\!\!\perp \epsilon$ where $\epsilon = X - \mathbf{E}[X|Z]$ and check $\epsilon \perp\!\!\!\perp Y - \mathbf{E}[Y|Z]$. Once again, we add a variable

$B := Y - \mathbf{E}[Y|Z]$ with parents Y and Z to the DAG, as shown in Figure 4, Case 3. Now we prove $X \perp\!\!\!\perp Y|Z \iff \epsilon \perp\!\!\!\perp B$ by considering two subcases.

Subcase 1. If there is an edge in either direction between ϵ and Y in Figure 4, Case 3, then X and Y are d-connected by the path $X \leftarrow \epsilon - Y$ and ϵ and B are d-connected by the path $\epsilon - Y \rightarrow B$, so we conclude $X \not\perp\!\!\!\perp Y|Z$ and $\epsilon \not\perp\!\!\!\perp B$ by faithfulness. Thus, we have proved the forward and reverse directions for this subcase.

Subcase 2. If there is no edge between ϵ and Y in Figure 4, then ϵ and B are d-separated, since X is a collider in the path $\epsilon \rightarrow X \leftarrow Z \rightarrow B$, which is thus blocked. Moreover X and Y are d-separated given Z, since Z blocks the path $X \leftarrow Z - Y$. By the Markov assumption, this implies $X \perp\!\!\!\perp Y|Z$ and $\epsilon \perp\!\!\!\perp B$. This proves the forward and reverse directions for this subcase. □

4.2. Testing Conditional Independence with Gaussian Process Regression and HSIC

Based on the preceding general framework, we propose the use of GP regression, which is almost surely consistent assuming Gaussian errors [Choi and Schervish 2007] with good rates of convergence [Van Der Vaart and Van Zanten 2011], and the HSIC for testing for independence.[2] Neither of these choices is crucial—we picked GP regression because of its long history in spatial statistics and widespread use as a convenient nonparametric regression method. We picked HSIC because it is equal to 0 if and only if the distributions under consideration are independent, whenever the kernel is characteristic. But in cases where domain knowledge could be used to guide the choice of independence test or pre-whitening method, these will, of course, be preferable to generic choices.[3]

In the following, we assume without loss of generality that X (and Y) is embedded in a vector space. That is, we assume that regression on X is well-defined. Hence, given observations $(X, Y, Z) = \{x_i, y_i, z_i\}$, we use GP regression to fit the models $X = f(Z) + \epsilon_1$ and $Y = g(Z) + \epsilon_2$. This requires specifying (possibly different) covariance functions over Z. If $k(z, z')$ is a covariance function (Mercer kernel) over Z, then we could sample directly from the GP prior, where we follow general practice and set the mean to 0:

$$f \sim \mathcal{GP}(0, k).$$

Let K be the Gram matrix where $K_{ij} = k(z_i, z_j)$. Conditional on the observations (X, Z), our data follow a multivariate Gaussian distribution. For a new location

$$x^*|X, Z, z^* \sim \mathcal{N}(K_*(K + \sigma^2 I)^{-1}X, K_{**} - K_*(K + \sigma^2 I)^{-1}(K_*)^T),$$

where $K_* = [k(z^*, z_1), \ldots, k(z^*, z_n)]$, and the σ^2 term is added because we assume that our observations are noisy, as discussed earlier.

We are not actually interested in observations at new locations but in making point predictions at the existing locations. In other words, we use the GP as a smoother. Hence, we replace K_* by K and find a vector of mean predictions: $\hat{X} = K(K + \sigma^2 I)^{-1}X$. The residuals are:

$$\epsilon_{xz} = X - \hat{X} = X - K(K + \sigma^2 I)^{-1}X = (I + \sigma^{-2}K)^{-1}X. \tag{6}$$

Using a possibly different kernel, say l with kernel matrix L, we obtain residuals from smoothing Y via $\epsilon_{yz} = Y - \hat{Y} = (I + \sigma^{-2}L)^{-1}Y$ in exactly the same way.

[2]Since GP regression is a.s. consistent, it is possible that our estimated residuals will not converge to their true values on a set of measure 0; but, in the worst case, all that this could do is bias our estimate of HSIC on a set of measure 0, so the standard estimate of HSIC will still be consistent.

[3]Note that guarantees about consistency and convergence for GP regression apply in the large sample limit. It is entirely possible that for misspecified models and/or small samples, generic methods like GP regression will fail to remove all dependence.

Now, as just proved, $\epsilon_{xz} \perp\!\!\!\perp \epsilon_{yz}$ if and only if $X \perp\!\!\!\perp Y|Z$. To test the hypothesis $\epsilon_{xz} \perp\!\!\!\perp \epsilon_{yz,}$ we use HSIC, which requires specifying kernels on the residuals, say $\tilde{p}$ and $\tilde{q}$ on ϵ_{xz} and $\epsilon_{yz,}$ respectively. This leads to the kernel matrices P and Q. The associated HSIC test statistic is $\frac{1}{n^2}\text{tr}(PHQH)$ for the centering matrix $H = I - \frac{1}{n}11^T$.

4.3. Pre-whitening with Gaussian Processes

Given observations $(X, S) = \{x_i, s_i\}$ where s_i is a location in space or time, we consider the model:

$$f \sim \mathcal{GP}(0, K)$$

with observation model:

$$X = f(S) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus, we want an estimator:

$$\hat{f} = E[f|(x_1, s_1), \ldots, (x_n, s_n)]$$

so that we can obtain residuals $\epsilon_{xs} = X - \hat{f}(S)$. Notice that because we are using GP regression, all our observations $(x_1, s_1) \ldots, (x_n, s_n)$ are used to estimate f, and we explicitly account for the non-iid nature of our spatial or temporal observations by learning f. An intuitive way to think about this is as smoothing. All the observations play a role in our posterior prediction of f at a new or existing location s^* as seen in the algebra: Our posterior prediction at location s^* is given by $E[f(s^*)|s^*, X, S] = K_*(K + \sigma^2 I)^{-1}X$. Note further that whereas for small samples some residual dependence may remain, we have a consistent method; so, as our sample size increases, this dependence will go to 0.[4]

Because we consider S to be an environmental variable, we make the same assumptions as previously, of independent, additive noise. In other words, if S is a cause of X, we assume $X = f(S) + \epsilon_{xs}$ with $S \perp\!\!\!\perp \epsilon_{xs}$. Similarly, if S is a cause of Y, we assume $Y = f(S)+\epsilon_{ys}$ with $S \perp\!\!\!\perp \epsilon_{ys}$. Notice that we are not restricting ourselves to deterministic functions f. Any time series model, such as an autoregressive time series, with additive errors fits these requirements. Thus, we can use ϵ_{xs} and ϵ_{ys} in subsequent independence tests, continuing to assume independent, additive noise for causes and Markov and faithfulness without worrying about bias due to an underlying correlation structure.

This pre-whitening process, which follows standard practice in the spatial statistics and time series literature, is illustrated in Figure 5 using the same setup described

[4]A proof of this fact uses the result in Van Der Vaart and Van Zanten [2011] that, for our model, there is some sequence $r_n \to 0$ for sample size n such that $\hat{f}$ converges to f with:

$$E_f \|\hat{f} - f\|_2^2 \leq r_n^2 \tag{7}$$

where the convergence rate of r_n depends on our choice of kernel. But, under a variety of conditions given in Van Der Vaart and Van Zanten [2011], we are guaranteed that it decreases to 0 in n. Since residuals are given by the vector $\hat{f} - f$, we would like a bound on the covariance off the diagonal (i.e., $\text{Cov}((\hat{f} - f)_i, (\hat{f} - f)_j)$), for all i and j. This is bounded above by the covariance on the diagonal:

$$\text{Cov}((\hat{f} - f)_i, (\hat{f} - f)_j) \tag{8}$$
$$\leq Var((\hat{f} - f)_i) \tag{9}$$
$$\leq E\big[(\hat{f} - f)_i^2\big] - E[(\hat{f} - f)_i]^2 \tag{10}$$
$$\leq E\big[(\hat{f} - f)_i^2\big] \tag{11}$$
$$\leq r_n^2 \tag{12}$$

The last step follows because the sum of the squared residuals are $\leq r_n^2$ by Equation (7), so any particular squared residual is also $\leq r_n^2$.

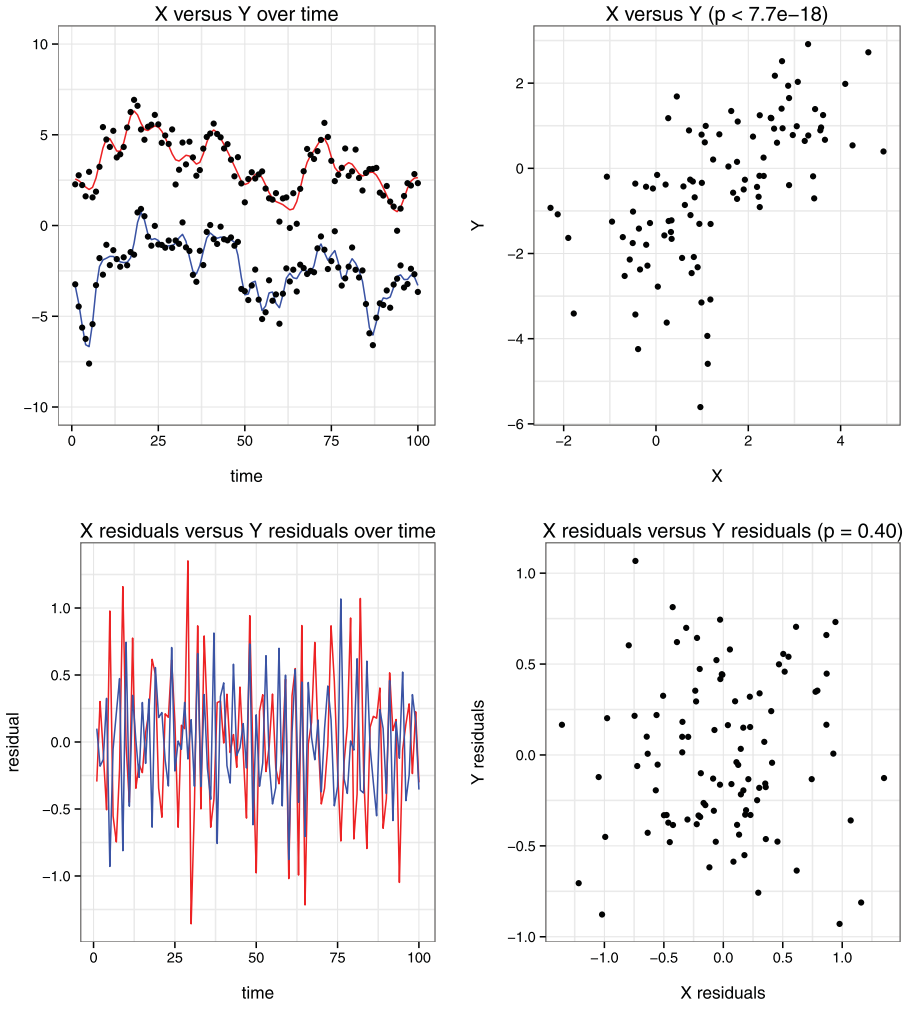


Fig. 5. X is a realization of an AR(1) process with $x_t = 0.9 \cdot x_{t-1} + \epsilon_{x,t}$ and Y is a realization of an AR(1) process with $y_t = 0.8 \cdot y_{t-1} + \epsilon_{y,t}$. X and Y are independent and $\epsilon_{x,t}, \epsilon_{y,t} \sim \mathcal{N}(0, 1)$. As shown in Figure 1, it is likely that there will be a spurious correlation between X and Y. We chose a specific realization, plotted as the black dots in the top left plot (for visual clarity, $X + 2$ and $Y - 2$ are shown), in which the correlation is 0.61 with highly significant p-value from HSIC $\leq 7.7 \times 10^{-18}$ (top right plot). We used GP regression with an RBF covariance function to obtain the fitted curves shown in red and blue in the top left. The residuals are shown in the bottom left and compared in the bottom right: The correlation between the residuals is 0.01 with insignificant p-value from HSIC $= 0.40$.

in the Introduction, where X and Y are independent AR(1) time series. We choose a particular realization with a large (but spurious) correlation of 0.61 between X and Y and a correspondingly highly significant value from HSIC ($p \leq 7.7 \times 10^{-18}$) for rejecting the null hypothesis of independence. We apply GP regression to X and Y separately, as shown in Figure 5, to estimate pre-whitened residuals ϵ_X and ϵ_Y. These residuals have a very low correlation of 0.01 and a correspondingly insignificant p-value from HSIC of 0.40[5].

[5]We note that the correct choice of kernel and method for obtaining residuals matters. We used a Gaussian RBF kernel and obtained residuals by smoothing. If we had been more concerned with trying to exactly mimic

In the case of the PC algorithm, an alternative approach would be to always include S in the conditioning set when testing for conditional independence. With enough data, this should be equivalent to the two-stage process we proposed. But because we believe that there is the potential for an important autocorrelation structure that we need to worry about, we think it is better to explicitly adjust for it in every variable first. This approach saves on computational time and modeling complexity: For moderately sized datasets, we can use a fully Bayesian analysis and carefully inspect the results of our pre-whitening step for each variable. By contrast, the PC algorithm could entail many conditional independence tests, so we need these to be automatic and relatively fast. Many conditional independence tests also rely on categorical conditioning sets, which are often obtained by first discretizing; this approach will be very difficult since observations are usually not repeated in space or time.

Finally, for the two-variable causal orientation task (e.g., as addressed by the RESIT framework), space or time would need to be included as part of the regression and again as part of the independence test, turning what was a simple univariate regression followed by unconditional independence test into two more complicated steps, a multivariate regression followed by conditional independence test.

4.4. Gaussian Processes for Structured Data

Since GPs depend on defining a kernel between observations, they can be used for highly structured data such as images and text. Given domains $\mathcal{X}, \mathcal{Y}$, we can define a joint GP over both domains; that is, using a kernel $k : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ such that random variables f, indexed by $(x, y) \in \mathcal{X} \times \mathcal{Y}$, are drawn from a multivariate Gaussian distribution with covariance matrix given by k, as evaluated on the index set and with mean function $\mu : (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ as evaluated on the index set.

A special but particularly interesting case arises whenever the kernel function k is given by a product over kernels on $\mathcal{X}$ and $\mathcal{Y}$ respectively; that is, whenever

$$k((x, y), (x', y')) = k_x(x, x')k_y(y, y').$$

Such a situation occurs, for example, in multivariate GP regression where $\mathcal{Y} = \{1, \ldots, d\}$ (i.e., where $\mathcal{Y}$ denotes the coordinate index of the regression problem and where k_y denotes the correlation between the coordinate-wise regressions). Likewise, when $\mathcal{Y}$ is the domain of images or documents, we therefore end up modeling the similarity between structured objects in $\mathcal{Y}$ using their covariates in $\mathcal{X}$.

We now exploit the duality discussed by Williams [1998] between feature space representations and GPs to introduce estimates of feature functions on $\mathcal{Y}$. That is, we adhere to the GP treatment for the covariate-dependent part of the kernel via $k_x(x, x')$ and use a feature space representation for the label-dependent part $l(y, y') = \langle \psi(y), \psi(y') \rangle$. The main motivation is that this will allow us to efficiently reason about feature space embeddings of distributions and of conditional probability distributions.

Before we do so, recall scalar GP regression as introduced in Section 3.1. There, one assumes that the random variable f, as indexed by $x \in \mathcal{X}$, follows a normal distribution with covariance function k and mean function μ. The idea is to extend the predictive

---

the behavior of a classical autoregressive fit, we would instead need to use the Ornstein-Uhlenbeck process, which is a GP with exponential kernel given by $k(t, t') = \frac{1}{1-\phi^2} \exp(\log(\phi)|t-t'|)$ where $\phi = 0.9$ for x and $\phi = 0.8$ for y, and we would also have performed one-step-ahead forecasting rather than smoothing in order to obtain the residuals. Ultimately, if the practitioner has domain knowledge supporting the use of a particular class of models, such as AR(1), we would absolutely recommend incorporating this knowledge, rather than relying on a generic choice like GP regression with a Gaussian RBF kernel. We advocate GP regression as a generally applicable method, especially in cases for which there is little domain expertise, and we further advocate carefully checking residuals for structure and refining one's modeling choices accordingly.

distribution $Y^*|Y, X, X^*$, as captured by Equation (5). We now extend this to vector-valued functions and, subsequently, to general index sets. In the standard treatment, we assume that:

$$f(X),\ f(X^*)|X, x^* \sim \mathcal{N}(0, K),$$

where $K_{ij} = k(x_i, x_j) + \delta_{ij}\sigma^2$. So conditioning, we find:

$$f(X^*)|Y, X, x^* = \mathcal{N}(K(x^*, X)(K + \sigma^2 I)^{-1}y,\ K(x_*, x_*) - K(x_*, x)(K + \sigma^2 I)^{-1}K(x, x_*)).$$

What if $f(X)$ isn't in $\mathcal{R}$, such as $Y \in \mathcal{R}^d$ or whenever Y is a string or an image? We begin with $\mathcal{Y} = \{1, \ldots, d\}$; thus, we could view the scalar case as $\mathcal{Y} = \{1\}$ and therefore with estimates in $\mathbb{R}^{\mathcal{Y}} = \mathbb{R}^1$. In general, the challenge is to deal with possible normalization problems of distributions over infinite-dimensional objects. The trick is to consider *evaluating* the GP on $\mathcal{Y}$ only on relevant points $y \in \mathcal{Y}$ rather than considering a possibly infinite dimensional set of evaluations.

For computational convenience of derivation, we adopt the argument of Williams [1998], which states that $f(y) = \langle v, \psi(y)\rangle$ is a linear function in the space of the features $\psi(y)$, where $v \sim \mathcal{N}(0, \mathbf{1})$ and therefore $f \sim \mathcal{GP}(0, l)$. It is understood that the kernel satisfies $l(y, y') = \langle \psi(y), \psi(y')\rangle$. This is entirely consistent whenever ψ is finite-dimensional. For the purpose of evaluation on a finite number of terms, we can always assume that ψ denotes the Cholesky factors of the covariance matrix L.

We now assume that we are given features $\psi(y_1), \ldots \psi(y_n)$, which are drawn from a GP with kernel k and mean 0. That is, we assume that this holds for any one-dimensional projection of $\psi(y)$ onto a unit-vector. Using Equation (5), we have that

$$\begin{aligned} &\psi(Y^*)|Y \sim \mathcal{N}(\bar{\mu}, \bar{K}), \qquad (13)\\ &\text{where } \bar{\mu} = K(X^*, X)(K(X, X) + \sigma^2 I)^{-1}\psi(Y)\\ &\qquad \bar{K} = K(X^*, X^*) - K(X^*, X)(K(X, X) + \sigma^2 I)^{-1}K(X, X^*). \end{aligned}$$

In it, we used the shorthand $\psi(Y) = (\psi(y_1), \ldots \psi(y_n))$ and, analogously, $\psi(Y^*) = (\psi(y_1^*), \ldots, \psi(y_{n^*}^*))$, whenever Y^* is a set. For instance, whenever $\mathcal{Y} = \{1, \ldots d\}$ and $\psi(y) = e_y$, this simply decomposes into d decoupled GPs. More generally, we can evaluate by taking inner products with test functions $\psi(y)$. At this point, the evaluation reduces to kernel computations $l(y, y')$.

As before, we employ the GP not for prediction but for smoothing only; that is, we are mostly interested in the residuals $\hat{\psi}_i - \psi(y_i)$ at locations x_i rather than the predictions $\hat{\psi}_i$ themselves. Since we will ultimately use HSIC, we do not need to explicitly compute the residuals; rather, we need to compute the Gram matrix R of the residuals with

$$\begin{aligned} R_{ij} &= \mathbf{E}_{\hat{\psi}}\left[\langle \hat{\psi}_i - \psi(y_i), \hat{\psi}_j - \psi(y_j)\rangle\right]. \qquad (14)\\ &= \operatorname*{Cov}_{\hat{\psi}}\left[\langle \hat{\psi}_i, \hat{\psi}_j\rangle\right] + \left\langle \mathbf{E}_{\hat{\psi}}\left[\hat{\psi}_i\right] + \psi(y_i), \mathbf{E}_{\hat{\psi}}\left[\hat{\psi}_j\right] - \psi(y_j)\right\rangle. \end{aligned}$$

The second line follows from the fact that $\mathrm{Cov}(A, B) = \mathbf{E}[AB] - \mathbf{E}[A][B]$. To evaluate this expression, we use the fact that the covariance is given in Equation (13). Its contribution to the entire matrix R is

$$K - K(K + \sigma^2 I)^{-1}K = K(I + \sigma^{-2}K)^{-1}, \qquad (15)$$

where we used the Woodbury matrix identity.[6] Next, we use the fact that

$$\mathbf{E}_{\hat{\psi}}\left[(\hat{\psi}_1, \ldots \hat{\psi}_n)\right] - \psi(Y) = K(K + \sigma^2 I)^{-1}\psi(Y) - \psi(Y) = -(I + \sigma^{-2}K)^{-1}\psi(Y), \qquad (16)$$

[6] $(A + B)^{-1} = A^{-1} - A^{-1}(B^{-1} + A^{-1})^{-1}A^{-1}$.

again using the Woodbury matrix identity. Taking inner products and plugging this back into Equation (14), we obtain

$$R = K(I + \sigma^{-2}K)^{-1} + (I + \sigma^{-2}K)^{-1}L(I + \sigma^{-2}K)^{-1}. \quad (17)$$

Note that R decomposes into two parts: The first is the contribution of the residuals due to smoothing in K. This converges to $\sigma^2 I$ for small σ^2 (i.e., whenever we assume that there is little additive noise associated with $y|x$, the contribution to the residuals matrix is very small off-diagonal and equal to σ^2 on the diagonal). Second, we have an appropriately smoothed term between K and L. Again, this vanishes for small additive noise but it also vanishes whenever K and L are coherent.

Once we have access to R, we can use HSIC to test independence: $\widehat{HSIC}(R, X) = \frac{1}{n^2}\text{tr}RHKH$. If we choose characteristic kernels for K (possibly a different K relative to the one used in the GP regression) and L, then we do not need to consider doing either a further embedding of the residuals or solving the pre-image problem and applying a different embedding because we only care about testing the independence between the residuals and X. Although we do not have access to the residuals, calculating HSIC only requires access to the Gram matrices corresponding to the feature space representation of the residuals and X. This is exactly what we have in the form of R and K respectively.

5. EXPERIMENTS

Source code for reproducing our experiments is provided in our supplementary materials.[7]

5.1. Spatial Data

The Boston Housing dataset, originally investigated in Harrison Jr and Rubinfeld [1978], has been widely used in statistics and machine learning. In the original paper, data were collected in 1970 and used in an analysis of the willingness of Boston area residents to pay for better air quality based on an economic model and regression analysis. As discussed in Pace and Gilley [1997], it is usually analyzed without taking into consideration the fact that the data are spatially observed.

There is significant spatial clustering in every single variable in the dataset, as revealed by Moran's I test (using a similarity matrix calculated as the reciprocals of the spatial distances between observations, p-values for each variable were significant, thus rejecting the null hypothesis of no spatial clustering) and confirmed by HSIC, which was used to test for independence between the locations in space (using an RBF kernel) and each variable separately. In addition to adding spatial coordinates to each observation, Pace and Gilley [1997] also corrected a few errors in the original dataset.

The variables in the dataset are crime rate (crim), proportion of residential land zoned for lots over 25,000 sq. ft (zn), proportion of nonretail business acres per town (indus), indicator variable for whether tract bounds the Charles River (chas), nitric oxides concentration (nox), average number of rooms per dwelling (rm), proportion of owner-occupied units built prior to 1940 (age), weighted average of distances to five Boston employment centers (dis), index of accessibility to radial highways (rad), full-value property-tax rate (tax), pupil-teacher ratio by town (ptratio), polynomial transformation of proportion of blacks by town (b), percentage of lower status people in the population (lstat), and median value of owner-occupied homes (medv).

[7]http://www.bitbucket.org/flaxter/tist-supporting-materials.

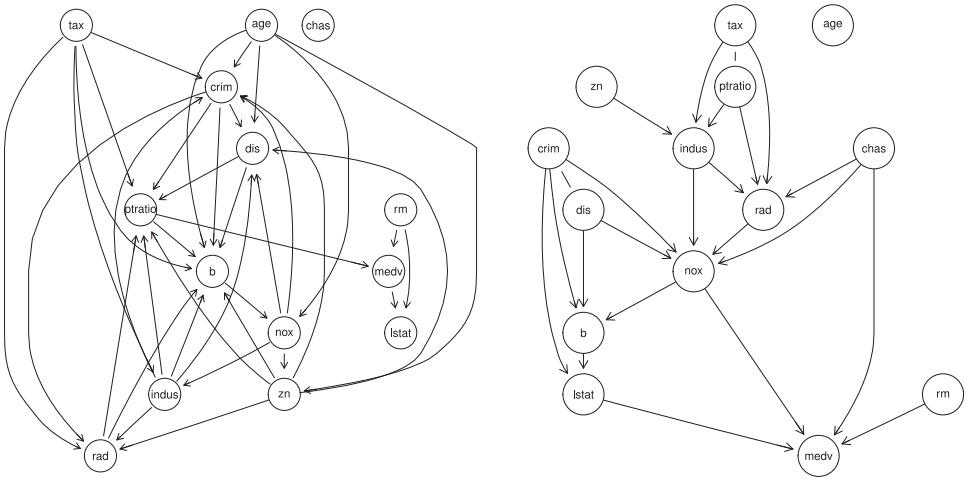


Fig. 6. Boston Housing data. Left: The PC algorithm was run on the data without pre-whitening (the data exhibit spatial autocorrelation), using GP/HSIC for conditional independence tests. The outcome variable of interest, median house value (medv), is caused by the number of rooms (rm) and the parent teacher ratio (ptratio), and it is a cause of the percentage of lower status people in the population (lstat). Other edge orientations seem dubious: Nitric oxide concentration (nox), a measure of pollution that causes industrial business activitity (indus), residential land zoned for large lots (zn), distance to employment centers (dis), and crime (crim). The substantive question in the original paper [Harrison Jr and Rubinfeld 1978] was about the effect of pollution (nox) on house value (medv), but, in the graph shown, there is no direct causal effect of pollution on house value. Right: After pre-whitening the data to remove spatial autocorrelation, the PC algorithm was run on it. The resulting causal graph has many fewer edges than the graph on the left. The outcome variable of interest, median house value (medv), is caused by percentage of lower status people in the population (lstat), number of rooms (rm), whether the tract bounds the Charles River (chas), and nitric oxide concentration (nox), a measure of pollution and the predictor variable of interest in the original paper [Harrison Jr and Rubinfeld 1978]. The graph shows that nitric oxide concentration (nox) is caused by industrial business activity (indus) (the opposite was found in the graph on the left), which is reasonable, but also by crime (crim), which seems unlikely.

The usual task with this dataset is to predict the median value of owner-occupied homes.

In the original analysis [Harrison Jr and Rubinfeld 1978], the authors carefully state their prior theoretical beliefs about the statistical (but not necessarily causal) relationship between each of the predictors in the dataset and the dependent variable. They included two "structural" variables that they expect to be related to home value, number of rooms and proportion of owner units built prior to 1940, eight neighborhood variables, two accessibility (in terms of transportation) variables, and two air pollution variables. Zhang et al. [2011] demonstrated KCI with the PC algorithm on these data. For the variable of interest, median value of house (medv), they found that number of rooms (rm), percentage of lower status people in the population (lstat), proportion of owner-occupied units built prior to 1940 (age), and crime rate (crim) are all parents of house value, with directed edges implying that these variables all cause house value.

We used the corrected dataset given in Pace and Gilley [1997]. We ran the PC algorithm as implemented in the R package pcalg [Kalisch et al. 2012] using our new GP/HSIC approach for conditional independence with $\alpha = 0.001$. The results are shown in Figure 6. Throughout, we use the Gamma approximation to calculate p-values from HSIC. In this case, the outcome variable of interest, median house value, is caused by the number of rooms and pupil-teacher ratio, and it is a cause of the percentage of lower status people in the population. There is no direct causal effect of pollution on house value.

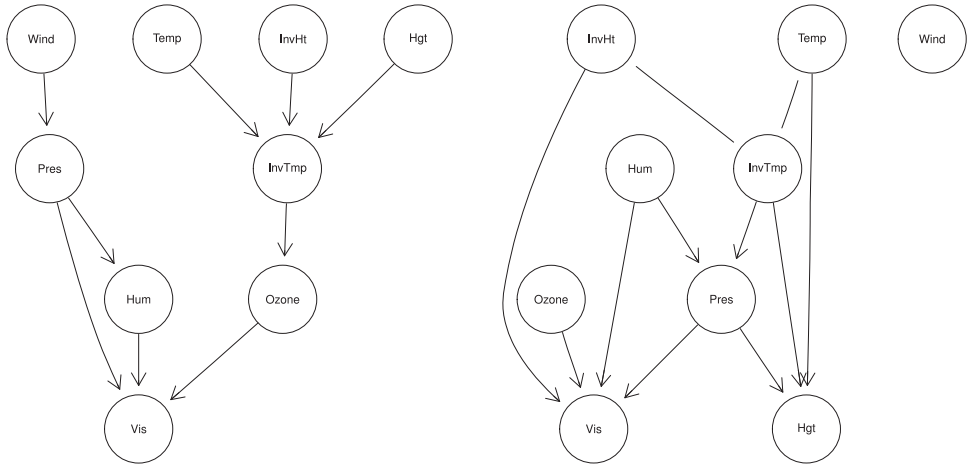


Fig. 7. Left: We used the PC algorithm with a dataset of environmental observations related to ozone in Upland, California, without first pre-whitening the time series data. The inferred causal CPDAG says that the temperature at the temperature inversion in the atmosphere (InvTmp) directly causes ozone, which causes visibility (Vis). Right: We pre-whitened the data using GP regression. Then we used the PC algorithm with the same dataset as previously. Ozone is still a cause of visibility (Vis), but no variables in the dataset were found to be a cause of ozone.

Next, we pre-whitened each variable using the spatial coordinates with a GP regression in which the hyperparameters of the squared exponential (RBF) covariance function are learned by maximizing the marginali likelihood using gradient descent (the default in GPStuff [Vanhatalo et al. 2013]). Using this new dataset, we ran the PC algorithm again with $\alpha = 0.001$, as shown in Figure 6 (right). The resulting causal graph has many fewer edges. The percentage of lower status people in the population and number of rooms cause house value, as in Zhang et al. [2011]. In addition, an indicator variable for whether the house is near the Charles River (which was not considered in Zhang et al. [2011]) also causes house value. Unlike the original graph in Figure 6 (left), we find that nitric oxide concentration, an indicator of air pollution, is a direct cause of house value, which addresses the original hypothesis explored by the authors in Harrison Jr and Rubinfeld [1978]. Furthermore, nitric oxide concentration is now found to be caused by industrial business activity, rather than the converse when using unwhitened data. But we see that nitric oxide concentration is also apparently caused by crime, which seems unlikely.

5.2. Time Series Data

We consider the ozone dataset used in Breiman and Friedman [1985]. These daily data clearly exhibit temporal autocorrelation, with 330 observations made over the course of 358 days. In Figure 7 (left), we show the results of the PC algorithm run on the data as-is, with conditional independence tests using GP regression for conditioning and HSIC for independence testing. We set $\alpha = .05$ and used the standard version of the PC algorithm implemented in Kalisch et al. [2012]. We used the Gamma approximation to calculate p-values for HSIC.

The ozone variable is directly caused by the temperature at the temperature inversion in the atmosphere (InvTmp) and is a cause of visibility. Figure 7 (right) contains the results of the PC algorithm run on the data after each variable has first been pre-whitened. To pre-whiten, we used GP regression with an exponential covariance function for time (which is analogous to an autoregressive fit), learning the

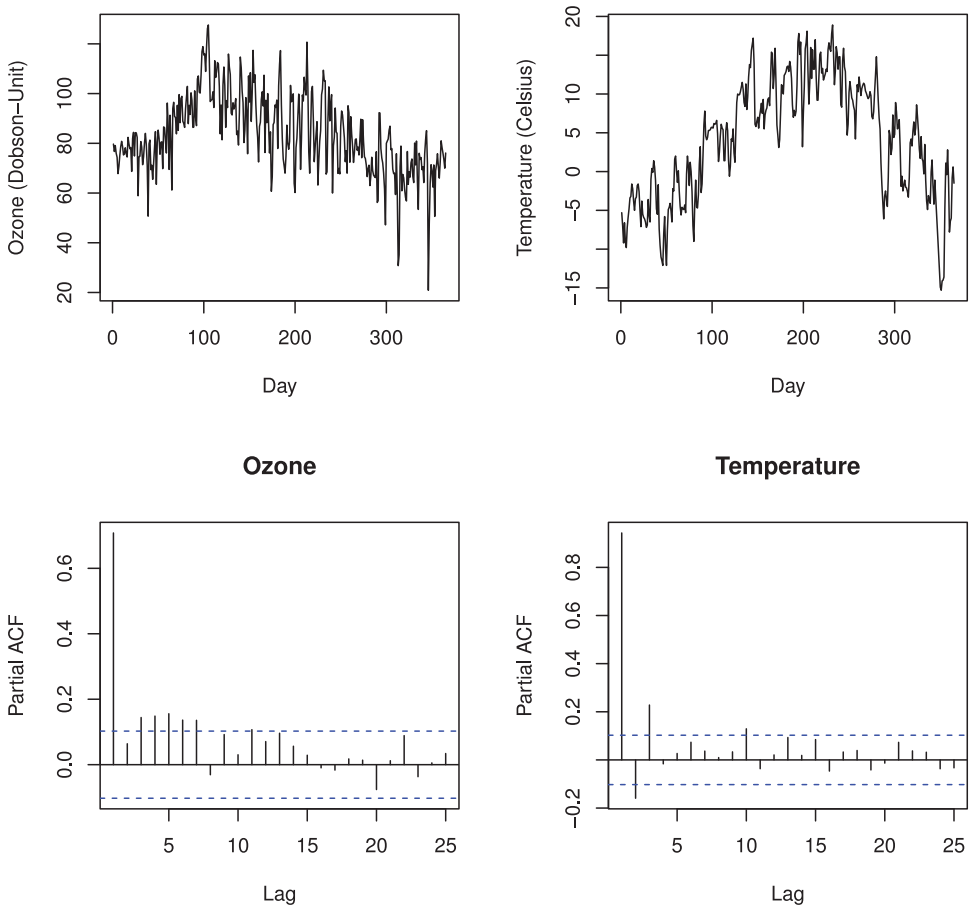


Fig. 8. Ozone and temperature data from Switzerland. A partial autocorrelation plot reveals significant temporal autocorrelation in both the ozone and temperature data. Before pre-whitening, bivariate causal orientation suggests incorrectly that ozone causes temperature. After pre-whitening, bivariate causal orientation correctly concludes that temperature causes ozone.

hyperparameters from the data by maximizing the marginal likelihood with gradient descent. Now we see that the ozone variable has no parents and is still a cause of visibility. Wind is no longer connected to any nodes, and two edges that were directed are no longer directed.

Next, we turn to the causal orientation (RESIT) framework for edge orientation and consider one of the pairs of data[8] that our replication of Peters et al. [2013] showed was misoriented using the same method considered in that paper, GP regression followed by HSIC, comparing a forward and backward model. Pair 51 consists of daily ozone and temperature data from Switzerland, where the ground truth is that temperature causes ozone. As shown in Figure 8, there is an underlying time trend, and a partial autocorrelation plot reveals temporal autocorrelation. Considering the data as-is, the p-value of the forward model ("ozone causes temperature") is 0.002 and the p-value of the backwards model ("temperature causes ozone") is 4×10^{-7}. Thus, the causal orientation method fails, incorrectly predicting the forward model because it fits better.

[8]http://webdav.tuebingen.mpg.de/cause-effect.

After pre-whitening, the p-values change. The forward model is still 0.002, but the backwards model is 0.34. The backwards model thus fits better, and the edge is correctly oriented.

5.3. Textual Data

We consider a novel causal orientation problem: Given pairs of translated sentences in two languages X and Y, determine whether X "causes" Y (meaning that the sentence in language Y was translated from the sentence in language X) or vice versa. We use the OpenOffice documentation corpus [Tiedemann 2009] that consists of sentence-aligned documentation in English, French, Spanish, Swedish, German, and Japanese. We use our GP formulation for structured data to calculate residuals, and then we test whether these residuals are independent of the predictor, as in the RESIT framework. We use a spectrum kernel (also called a string kernel, the default in Karatzoglou et al. [2004]) that matches substrings of length $m = 3$.

The corpus is relatively large, with 30,000–40,000 observations, so we use a bootstrap approach: For a pair of languages X, Y, we take a small sample ($n = 400$) and calculate a Gram matrix for the residuals R for the forward model X causes Y for half the sample ($n = 200$). Then we use HSIC to test whether R is independent of X on the other half of the sample ($n = 200$). We do the same for the reverse direction. We repeat this process 500 times with different subsets of the data and report the fraction of times that we predict the forward direction based on comparing the p-values of the forward and reverse directions. (Larger p-values indicate better fits, so we accept the direction with the larger p-value.)

The results are shown in Figure 9. The OpenOffice documentation was originally written in German for its predecessor, StarOffice. When it was purchased by Sun Microsystems, the documentation was translated into English. Subsequently, translations were made from English to other languages, and new additions to the documentation were made in English.[9] Thus, we consider English to be the "cause" of every other language except German. The algorithm correctly orients forward edges from English to every other language except German. The algorithm also orients forward edges from German to every language, which makes sense since German is a cause (though not direct) of every other language.

6. CONCLUSION

We proposed a simple, unified framework for coherently addressing the problem of algorithmic causal inference with non-iid observations (e.g., when data points are distributed in space and time), and we demonstrated its use on two real datasets. When using the PC algorithm or any other method based on independence tests, non-iid data presents a problem, and we showed how a pre-whitening step, using GP regression, can address this problem. We further showed how this same idea, of obtaining residuals from a GP regression, can be used to turn an unconditional independence test like HSIC into a conditional independence test.

We also showed that highly structured data, like text, can be considered in a causal framework, again using GP regression. In this case, we presented a novel formulation of a GP for structured inputs and outputs. The key derivation was that of the Gram matrix of the residuals because, once this is calculated, we can use HSIC to test independence.

HSIC is but one of the many measures of statistical independence that have been proposed. It might be fruitful to consider other measures instead, such as mutual

[9]Uwe Fischer, personal communication, 9 July 2014.

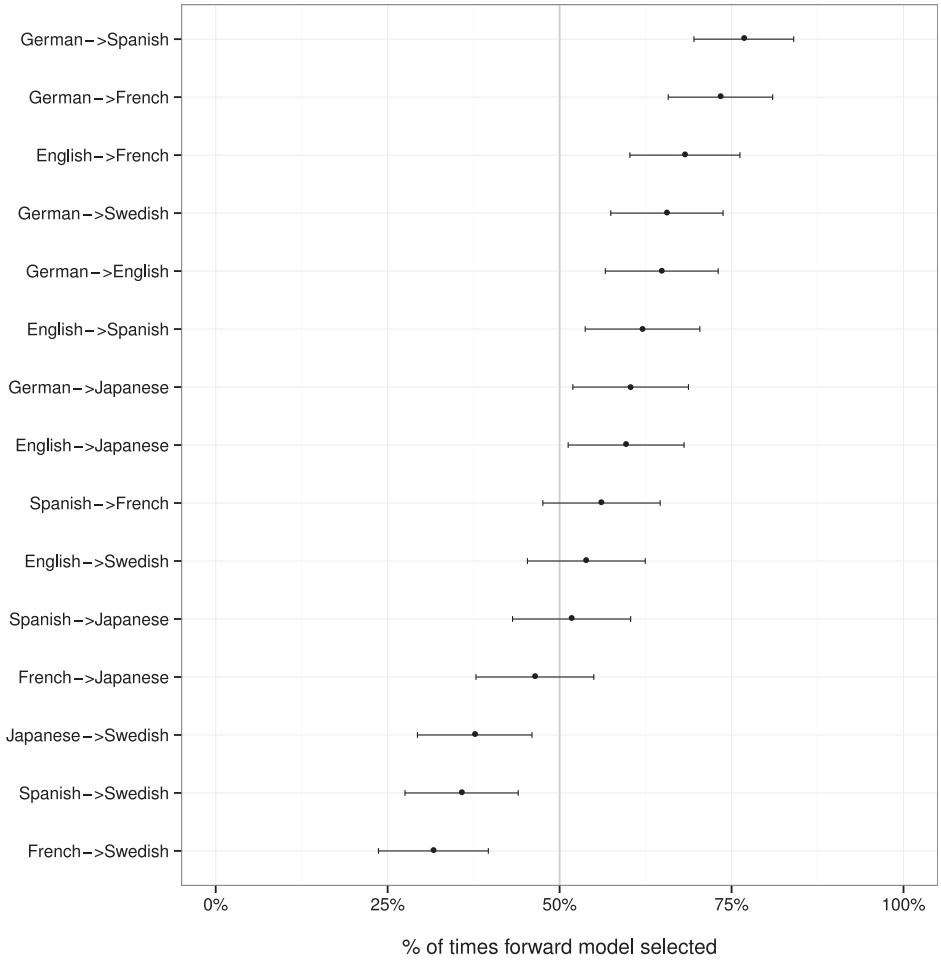


Fig. 9. Causal orientation with text: Given pairs of sentences in two languages, the task is to determine which language is a translation of the other. We used a bootstrapping approach to repeatedly apply a causal orientation algorithm based on Gaussian processes for structured data and HSIC for independence testing to determine which language "causes" the other language. Shown are the fraction of times that the algorithm selected the forward causal direction, along with 95% confidence intervals. The top line, for example, means that in comparing German and Spanish, the algorithm concluded that German caused Spanish 77% of the time. The sentences come from OpenOffice documentation, portions of which were originally written in German and translated into English. After this one-time translation, which occurred when Sun Microsystems bought what was then StarOffice, new documentation was written in English, and English became the source language for translations into Spanish, Swedish, French, Japanese, and back into German. The algorithm thus correctly orients edges such that German and English are the cause of every other language. The algorithm definitively concludes that German causes English.

information or distance correlation [Székely et al. 2009] or to determine whether the consistency results in Kpotufe et al. [2014] hold for our method. We do believe that GP regression is the most flexible and general tool for the purposes of pre-whitening non-iid data due to its long-standing use in the spatial statistics and time series literature. In future work, we intend to look more deeply at the connections between GP regression and kernel-based measures of independence.

REFERENCES

Michel Besserve, Nikos K. Logothetis, and Bernhard Schölkopf. 2013. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 2535–2543.

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. 2008. *Time Series Analysis*. John Wiley & Sons, Inc. DOI:http://dx.doi.org/10.1002/9781118619193.ch1

Leo Breiman and Jerome H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 391 (1985), 580–598.

Taeryon Choi and Mark J Schervish. 2007. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis* 98, 10 (2007), 1969–1987.

K. Chwialkowski and A. Gretton. 2014. A kernel independence test for random processes. In *ICML*. http://arxiv.org/abs/1402.4501.

N. Cressie and C. K. Wikle. 2011. *Statistics for Spatio-Temporal Data*. Vol. 465. Wiley.

G. Doran, K. Muandet, K. Zhang, and B. Scholkopf. 2014. A permutation-based kernel conditional independence test. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*. 132–41.

David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. 2013. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922* (2013).

Ragnar Frisch and Frederick V. Waugh. 1933. Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society* (1933), 387–401.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2007. Kernel measures of conditional dependence. In *NIPS*, Vol. 20. 489–496.

Tilmann Gneiting, M. Genton, and Peter Guttorp. 2007. Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems* (2007), 151–175.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13 (2012), 723–773.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*. Springer, 63–77.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schoelkopf, and A. Smola. 2008. A kernel statistical test of independence. (2008). http://books.nips.cc/papers/files/nips20/NIPS2007_0730.pdf.

David Harrison Jr. and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5, 1 (1978), 81–102.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. *Advances in Neural Information Processing Systems (NIPS)*, Vol. 21. 689–696.

Alfredo Kalaitzis, Antti Honkela, Pei Gao, and Neil D. Lawrence. 2013. *gptk: Gaussian Processes Tool-Kit*. http://CRAN.R-project.org/package=gptk R package version 1.07.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. 2012. Causal inference using graphical models with the r package pcalg. *Journal of Statistical Software* 47, 11 (2012), 1–26. http://www.jstatsoft.org/v47/i11/

Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. 2004. Kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software* 11, 9 (2004), 1–20.

Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Scholkopf. 2014. Consistency of causal inference under the additive noise model. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*. Beijing, China.

Alessio Moneta, Nadine Chlaß, Doris Entner, and Patrik O. Hoyer. 2011. Causal search in structural vector autoregressive models. *Journal of Machine Learning Research-Proceedings Track* 12 (2011), 95–114.

Patrick A. P. Moran. 1950. Notes on continuous stochastic phenomena. *Biometrika* (1950), 17–23.

R. Kelley Pace and Otis W. Gilley. 1997. Using the spatial configuration of the data to improve estimation. *The Journal of Real Estate Finance and Economics* 14, 3 (1997), 333–340.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York.

K. Pearson. 1983. *Handbook of Applied Mathematics*. Van Nostrand Reinhold Company, New York.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. 2013. Causal discovery with continuous additive noise models. *ArXiv preprint ArXiv:1309.6779v4* (2013).

J. D. Ramsey. 2014. A scalable conditional independence test for nonlinear, non-gaussian data. *ArXiv e-prints* (Jan. 2014).

Carl Edward Rasmussen and Hannes Nickisch. 2010. Gaussian processes for machine learning (GPML) toolbox. *The Journal of Machine Learning Research* 11 (2010), 3011–3015.

Carl Edward Rasmussen and Christopher K. I. Williams. 2006. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, Massachusetts.

Sashank Reddi and Barnabás Póczos. 2013. Scale invariant conditional dependence measures. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*. Atlanta, GA, USA, 1355–1363.

K. Salkauskas. 1982. Some relationships between surface splines and kriging. In *Multivariate Approximation Theory II*, W. Schempp and K. Zeller (Eds.). Birkhauser, Basel, 313–325.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. 2007. A hilbert space embedding for distributions. In *Proceedings of the International Conference on Algorithmic Learning Theory*, Vol. 4754. Springer, 13–31.

P. Spirtes, C. Glymour, and R. Scheines. 2001. *Causation, Prediction, and Search*. Vol. 81. MIT Press.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. 2010. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* 99 (2010), 1517–1561.

Liangjun Su and Halbert White. 2007. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics* 141, 2 (2007), 807–834.

Gábor J. Székely, Maria L. Rizzo, et al. 2009. Brownian distance covariance. *The Annals of Applied Statistics* 3, 4 (2009), 1236–1265.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov (Eds.). Vol. V. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 237–248.

Robert E. Tillman, Arthur Gretton, and Peter Spirtes. 2009. Nonlinear directed acyclic structure learning with weakly additive noise models. In *NIPS*. 1847–1855.

Aad Van Der Vaart and Harry Van Zanten. 2011. Information rates of nonparametric Gaussian process methods. *The Journal of Machine Learning Research* 12 (2011), 2095–2119.

Jarno Vanhatalo, Jaakko Riihimäki, Jouni Hartikainen, Pasi Jylänki, Ville Tolvanen, and Aki Vehtari. 2013. GPstuff: Bayesian modeling with Gaussian processes. *The Journal of Machine Learning Research* 14, 1 (2013), 1175–1179.

G. Wahba. 1990. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59. SIAM, Philadelphia.

C. K. I. Williams. 1998. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning and Inference in Graphical Models*, M. I. Jordan (Ed.). Kluwer Academic, 599–621.

A. G. Wilson and R. P. Adams. 2013. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning*.

K. Zhang, J. Peters, D. Janzing, B., and B. Schölkopf. 2011. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*. 804–813.

Xinhua Zhang, Le Song, Arthur Gretton, and Alex J. Smola. 2008. Kernel measures of independence for non-iid data. In *NIPS*, Vol. 22.

Received July 2014; revised April 2015; accepted July 2015