Stat Comput (2015) 25:755–766
DOI 10.1007/s11222-015-9558-5

Computing functions of random variables via reproducing kernel Hilbert space representations

Bernhard Schölkopf[1] · **Krikamol Muandet**[1] · **Kenji Fukumizu**[2] · **Stefan Harmeling**[3] · **Jonas Peters**[4]

Accepted: 8 March 2015 / Published online: 11 June 2015

Abstract We describe a method to perform functional operations on probability distributions of random variables. The method uses reproducing kernel Hilbert space representations of probability distributions, and it is applicable to all operations which can be applied to points drawn from the respective distributions. We refer to our approach as *kernel probabilistic programming*. We illustrate it on synthetic data and show how it can be used for nonparametric structural equation models, with an application to causal inference.

Keywords Kernel methods · Probabilistic programming · Causal inference

1 Introduction

Data types, derived structures, and associated operations play a crucial role for programming languages and the computations we carry out using them. Choosing a data type, such as Integer, Float, or String, determines the possible values, as well as the operations that an object permits. Operations typically return their results in the form of data types. Composite or derived data types may be constructed from simpler ones, along with specialized operations applicable to them.

The goal of the present paper is to propose a way to represent distributions over data types, and to generalize operations built originally for the data types to operations applicable to those distributions. Our approach is nonparametric and thus not concerned with what distribution models make sense on which *statistical* data type (e.g., binary, ordinal, categorical). It is also general, in the sense that in principle, it applies to all data types and functional operations. The price to pay for this generality is that

- our approach will, in most cases, provide approximate results only; however, we include a statistical analysis of the convergence properties of our approximations, and
- for each data type involved (as either input or output), we require a positive definite kernel capturing a notion of similarity between two objects of that type. Some of our results require, moreover, that the kernels be characteristic in the sense that they lead to injective mappings into associated Hilbert spaces.

In a nutshell, our approach represents distributions over objects as elements of a Hilbert space generated by a kernel and describes how those elements are updated by operations available for sample points. If the kernel is trivial in the sense that each distinct object is only similar to itself, the method reduces to a Monte Carlo approach where the operations are applied to sample points which are propagated to the next step. Computationally, we represent the Hilbert space elements as finite weighted sets of objects, and all opera-

✉ Bernhard Schölkopf
bs@tuebingen.mpg.de

Krikamol Muandet
krikamol@tuebingen.mpg.de

Kenji Fukumizu
fukumizu@ism.ac.jp

Stefan Harmeling
stefan.harmeling@uni-duesseldorf.de

Jonas Peters
peters@stat.math.ethz.ch

[1] Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany

[2] Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo

[3] Institut für Informatik, Heinrich-Heine-Universität, 40225 Düsseldorf, Germany

[4] ETH Zürich, Seminar für Statistik, 8092 Zürich, Switzerland

tions reduce to finite expansions in terms of kernel functions between objects.

The remainder of the present article is organized as follows. After describing the necessary preliminaries, we provide an exposition of our approach (Sect. 3). Section 4 analyzes an application to the problem of cause–effect inference using structural equation models. We conclude with a limited set of experimental results.

2 Kernel Maps

2.1 Positive definite kernels

The concept of representing probability distributions in a reproducing kernel Hilbert space (RKHS) has recently attracted attention in statistical inference and machine learning (Berlinet and Agnan 2004; Smola et al. 2007). One of the advantages of this approach is that it allows us to apply RKHS methods to probability distributions, often with strong theoretical guarantees (Sriperumbudur et al. 2008, 2010). It has been applied successfully in many domains such as graphical models (Song et al. 2010, 2011), two-sample testing (Gretton et al. 2012), domain adaptation (Huang et al. 2007; Gretton et al. 2009; Muandet et al. 2013), and supervised learning on probability distributions (Muandet et al. 2012; Szabó et al. 2014). We begin by briefly reviewing these methods, starting with some prerequisites.

We assume that our input data $\{x_1, \ldots, x_m\}$ live in a nonempty set $\mathscr{X}$ and are generated i.i.d. by a random experiment with Borel probability distribution p. By k, we denote a *positive definite kernel* on $\mathscr{X} \times \mathscr{X}$, i.e., a symmetric function

$$k : \mathscr{X} \times \mathscr{X} \to \mathbb{R} \tag{1}$$
$$(x, x') \mapsto k(x, x') \tag{2}$$

satisfying the following nonnegativity condition: for any $m \in \mathbb{N}$, and $a_1, \ldots, a_m \in \mathbb{R}$,

$$\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0. \tag{3}$$

If equality in (3) implies that $a_1 = \cdots = a_m = 0$, the kernel is called *strictly* positive definite.

2.2 Kernel maps for points

Kernel methods in machine learning, such as Support Vector Machines or Kernel PCA, are based on mapping the data into a reproducing kernel Hilbert space (RKHS) $\mathscr{H}$ (Boser et al. 1992; Schölkopf and Smola 2002; Shawe-Taylor and Cristianini 2004; Hofmann et al. 2008; Steinwart and Christmann 2008),

$$\Phi : \mathscr{X} \to \mathscr{H} \tag{4}$$
$$x \mapsto \Phi(x), \tag{5}$$

where the *feature map* (or *kernel map*) Φ satisfies

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle \tag{6}$$

for all $x, x' \in \mathscr{X}$. One can show that every k taking the form (6) is positive definite, and every positive definite k allows the construction of $\mathscr{H}$ and Φ satisfying (6). The canonical feature map, which is what by default we think of whenever we write Φ, is

$$\Phi : \mathscr{X} \to \mathbb{R}^{\mathscr{X}} \tag{7}$$
$$x \mapsto k(x, .), \tag{8}$$

with an inner product satisfying the *reproducing kernel property*

$$k(x, x') = \langle k(x, .), k(x', .) \rangle. \tag{9}$$

Mapping observations $x \in \mathscr{X}$ into a Hilbert space is rather convenient in some cases. If the original domain $\mathscr{X}$ has no linear structure to begin with (e.g., if the x are strings or graphs), then the Hilbert space representation provides us with the possibility to construct geometric algorithms using the inner product of $\mathscr{H}$. Moreover, even if $\mathscr{X}$ is a linear space in its own right, it can be helpful to use a nonlinear feature map in order to construct algorithms that are linear in $\mathscr{H}$ while corresponding to nonlinear methods in $\mathscr{X}$.

2.3 Kernel maps for sets and distributions

One can generalize the map Φ to accept as inputs not only single points, but also sets of points or distributions. It was pointed out that the *kernel map of a set of points* $\mathbf{X} := \{x_1, \ldots, x_m\}$,

$$\mu[\mathbf{X}] := \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i), \tag{10}$$

corresponds to a kernel density estimator in the input domain (Schölkopf and Smola 2002; Schölkopf et al. 2001), provided the kernel is nonnegative and integrates to 1. However, the map (10) can be applied for all positive definite kernels, including ones that take negative values or that are not normalized. Moreover, the fact that $\mu[\mathbf{X}]$ lives in an RKHS and the use of the associated inner product and norm will have a number of subsequent advantages. For these reasons, it would be misleading to think of (10) simply as a kernel density estimator.

Table 1 What information about a sample $\mathbf{X}$ does the kernel map $\mu[\mathbf{X}]$ [see (10)] contain?

$k(x, x') = \langle x, x' \rangle$	Mean of $\mathbf{X}$
$k(x, x') = (\langle x, x' \rangle + 1)^n$	Moments of $\mathbf{X}$ up to order $n \in \mathbb{N}$
$k(x, x')$ strictly p.d.	All of $\mathbf{X}$ (i.e., μ injective)

Table 2 What information about p does the kernel map $\mu[p]$ [see (11)] contain? For the notions of characteristic/universal kernels, see Steinwart (2002); Fukumizu et al. (2008, 2009); an example thereof is the Gaussian kernel (26)

$k(x, x') = \langle x, x' \rangle$	Expectation of p
$k(x, x') = \big(\langle x, x' \rangle + 1\big)^n$	Moments of p up to order $n \in \mathbb{N}$
$k(x, x')$ characteristic/universal	All of p (i.e., μ injective)

The *kernel map of a distribution* p can be defined as the expectation of the feature map (Berlinet and Agnan 2004; Smola et al. 2007; Gretton et al. 2012),

$$\mu[p] := \mathbb{E}_{x\sim p}\big[\Phi(x)\big], \tag{11}$$

where we overload the symbol μ and assume, here and below, that p is a Borel probability measure, and

$$\mathbb{E}_{x,x'\sim p}\big[k(x, x')\big] < \infty. \tag{12}$$

A sufficient condition for this to hold is the assumption that there exists an $M \in \mathbb{R}$ such that

$$\big\|k(x, .)\big\| \leq M < \infty, \tag{13}$$

or equivalently $k(x, x) \leq M^2$, on the support of p. Kernel maps for sets of points or distributions are sometimes referred to as *kernel mean maps* to distinguish them from the original kernel map. Note, however, that they include the kernel map of a point as a special case, so there is some justification in using the same term. If $p = p_X$ is the law of a random variable X, we sometimes write $\mu[X]$ instead of $\mu[p]$.

In all cases, it is important to understand what information we retain, and what we lose, when representing an object by its kernel map. We summarize the known results (Steinwart and Christmann 2008; Fukumizu et al. 2008; Smola et al. 2007; Gretton et al. 2012; Sriperumbudur et al. 2010) in Tables 1 and 2.

We conclude this section with a discussion of how to use kernel mean maps. To this end, first assume that Φ is injective, which is the case if k is strictly positive definite (see Table 1) or characteristic/universal (see Table 2). Particular cases include the *moment generating function* of a RV with distribution p,

$$M_p(.) = \mathbb{E}_{x\sim p}\left[e^{\langle x, \cdot \rangle}\right], \tag{14}$$

which equals (11) for $k(x, x') = e^{\langle x, x' \rangle}$ using (8).

We can use the map to test for equality of data sets,

$$\big\|\mu[\mathbf{X}] - \mu[\mathbf{X}']\big\| = 0 \Longleftrightarrow \mathbf{X} = \mathbf{X}', \tag{15}$$

or distributions,

$$\big\|\mu[p] - \mu[p']\big\| = 0 \Longleftrightarrow p = p'. \tag{16}$$

Two applications of this idea lead to tests for *homogeneity* and *independence*. In the latter case, we estimate $\big\|\mu[p_x p_y] - \mu[p_{xy}]\big\|$ (Bach and Jordan 2002; Gretton et al. 2005); in the former case, we estimate $\big\|\mu[p] - \mu[p']\big\|$ (Gretton et al. 2012).

Estimators for these applications can be constructed in terms of the empirical mean estimator (the kernel mean of the empirical distribution)

$$\mu[\hat{p}_m] = \frac{1}{m}\sum_{i=1}^{m} \Phi(x_i) = \mu[\mathbf{X}], \tag{17}$$

where $\mathbf{X} = \{x_1, \ldots, x_m\}$ is an i.i.d. sample from p [cf. (10)]. As an aside, note that using ideas from James–Stein estimation (James and Stein 1961), we can construct shrinkage estimators that improve upon the standard empirical estimator [see e.g., Muandet et al. (2014a, b)].

One can show that $\mu[\hat{p}_m]$ converges at rate $m^{-1/2}$ [cf. Smola et al. (2007), (Song 2008, Theorem 27), and Lopez-Paz et al. (2015)]:

Theorem 1 *Assume that* $\|f\|_\infty \leq 1$ *for all* $f \in \mathscr{H}$ *with* $\|f\|_{\mathscr{H}} \leq 1$. *Then with probability at least* $1 - \delta$,

$$\big\|\mu[\hat{p}_m] - \mu[p]\big\|_{\mathscr{H}} \leq 2\sqrt{\frac{\mathbb{E}_{z\sim p}[k(z, z)]}{m}} + \sqrt{\frac{2\log(1/\delta)}{m}}. \tag{18}$$

Independent of the requirement of injectivity, μ can be used to compute expectations of arbitrary functions f living in the RKHS, using the identity

$$\mathbb{E}_{x\sim p}[f(x)] = \big\langle \mu[p], f \big\rangle, \tag{19}$$

which follows from the fact that k represents point evaluation in the RKHS,

$$f(x) = \big\langle k(x, .), f \big\rangle. \tag{20}$$

A small RKHS, such as the one spanned by the linear kernel

$$k(x, x') = \langle x, x' \rangle, \tag{21}$$

may not contain the functions we are interested in. If, on the other hand, our RKHS is sufficiently rich [e.g., associated

with a universal kernel (Steinwart 2002)], we can use (19) to approximate, for instance, the probability of any interval (a, b) on a bounded domain, by approximating the indicator function $I_{(a,b)}$ as a kernel expansion $\sum_{i=1}^{n} a_i k(x_i, .)$, and substituting the latter into (19). See Kanagawa and Fukumizu (2014) for further discussion. Alternatively, if p has a density, we can estimate it using methods such as reproducing kernel moment matching and combinations with kernel density estimation (Song et al. 2008; Kanagawa and Fukumizu 2014).

This shows that the map is not a one-way road: we can map our objects of interest into the RKHS, perform linear algebra and geometry on them (Schölkopf and Smola 2002), and at the end answer questions of interest. In the next section, we shall take this a step further, and discuss how to implement rather general operations in the RKHS.

Before doing so, we mention two additional applications of kernel maps. We can map conditional distributions and perform Bayesian updates (Fukumizu et al. 2008, 2013; Zhang et al. 2011), and we can connect kernel maps to Fourier optics, leading to a physical realization as Fraunhofer diffraction (Harmeling et al. 2013).

3 Computing functions of independent random variables

3.1 Introduction and earlier work

A random variable (RV) is a measurable function mapping possible outcomes of an underlying random experiment to a set $\mathcal{Z}$ (often, $\mathcal{Z} \subset \mathbb{R}^d$, but our approach will be more general). The probability measure of the random experiment induces the distribution of the random variable. We will below not deal with the underlying probability space explicitly, and instead directly start from random variables X, Y with distributions p_X, p_Y and values in $\mathcal{X}, \mathcal{Y}$. Suppose we have access to (data from) p_X and p_Y, and we want to compute the distribution of the random variable $f(X, Y)$, where f is a measurable function defined on $\mathcal{X} \times \mathcal{Y}$.

For instance, if our operation is addition $f(X, Y) = X + Y$, and the distributions p_X and p_Y have densities, we can compute the density of the distribution of $f(X, Y)$ by convolving those densities. If the distributions of X and Y belong to some parametric class, such as a class of distributions with Gaussian density functions, and if the arithmetic expression is elementary, then closed-form solutions for certain favorable combinations exist. At the other end of the spectrum, we can resort to numerical integration or sampling to approximate $f(X, Y)$.

Arithmetic operations on random variables are abundant in science and engineering. Measurements in real-world systems are subject to uncertainty, and thus subsequent arithmetic operations on these measurements are operations on random variables. An example due to Springer (1979) is signal amplification. Consider a set of n amplifiers connected together in a serial fashion. If the amplification of the i-th amplifier is denoted by X_i, then the total amplification, denoted by Y, is $Y = X_1 \cdot X_2 \cdots X_n$, i.e., a product of n random variables.

A well-established framework for arithmetic operation on independent random variables (iRVs) relies on *integral transform* methods (Springer 1979). The above example of addition already suggests that Fourier transforms may help, and indeed, people have used transforms such as the ones due to Fourier and Mellin to derive the distribution function of either the sum, difference, product, or quotient of iRVs (Epstein 1948; Springer and Thompson 1966; Prasad 1970; Springer 1979). Williamson (1989) proposes an approximation using Laguerre polynomials and a notion of *envelopes* bounding the cumulative distribution function. This framework also allows for the treatment of dependent random variables, but the bounds can become very loose after repeated operations. Milios (2009) approximates the probability distributions of the input random variables as mixture models (using uniform and Gaussian distributions), and apply the computations to all mixture components.

Jaroszewicz and Korzen (2012) consider a numerical approach to implement arithmetic operations on iRVs, representing the distributions using piecewise Chebyshev approximations. This lends itself well to the use of approximation methods that perform well as long as the functions are well-behaved. Finally, Monte Carlo approaches can be used as well and are popular in scientific applications [see e.g., Ferson (1996)].

The goal of the present paper is to develop a derived data type representing a distribution over another data type and to generalize the available computational operations to this data type, at least approximately. This would allow us to conveniently handle error propagation as in the example discussed earlier. It would also help us perform inference involving conditional distributions of such variables given observed data. The latter is the main topic of a subject area that has recently begun to attract attention, *probabilistic programming* (Gordon et al. 2014). A variety of probabilistic programming languages has been proposed (Wood et al. 2014; Paige and Wood 2014; Cassel 2014). To emphasize the central role that kernel maps play in our approach, we refer to it as *kernel probabilistic programming (KPP)*.

3.2 Computing functions of independent random variables using kernel maps

The key idea of KPP is to provide a consistent estimator of the kernel map of an expression involving operations on random variables. This is done by applying the expression to the

sample points and showing that the resulting kernel expansion has the desired property. Operations involving more than one RV will increase the size of the expansion, but we can resort to existing RKHS approximation methods to keep the complexity of the resulting expansion limited, which is advisable in particular if we want to use it as a basis for further operations. The benefits of KPP are threefold. First, we do not make parametric assumptions on the distributions associated with the random variables. Second, our approach applies not only to real-valued random variables, but also to multivariate random variables, structured data, functional data, and other domains, as long as positive definite kernels can be defined on the data. Finally, it does not require explicit density estimation as an intermediate step, which is difficult in high dimensions.

We begin by describing the basic idea. Let f be a function of two independent RVs X, Y taking values in the sets $\mathcal{X}, \mathcal{Y}$, and suppose we are given i.i.d. m-samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$. We are interested in the distribution of $f(X, Y)$, and seek to estimate its representation $\mu\big[f(X,Y)\big] := \mathbb{E}\Big[\Phi\big(f(X,Y)\big)\Big]$ in the RKHS as

$$\frac{1}{m^2}\sum_{i,j=1}^{m} \Phi\left(f(x_i, y_j)\right). \tag{22}$$

Although $x_1, \ldots, x_m \sim p_X$ and $y_1, \ldots, y_m \sim p_Y$ are i.i.d. observations, this does not imply that the $\big\{f(x_i, y_j)|i, j = 1, \ldots, m\big\}$ form an i.i.d. m^2-sample from $f(X, Y)$, since—loosely speaking—each x_i (and each y_j) leaves a footprint in m of the observations, leading to a (possibly weak) dependency. Therefore, Theorem 1 does not imply that (22) is consistent. We need to do some additional work:

Theorem 2 *Given two independent random variables X, Y with values in $\mathcal{X}, \mathcal{Y}$, mutually independent i.i.d. samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, a measurable function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$, and a positive definite kernel on $\mathcal{Z} \times \mathcal{Z}$ with RKHS map Φ, then*

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \Phi\big(f(x_i, y_j)\big) \tag{23}$$

is an unbiased and consistent estimator of $\mu\big[f(X,Y)\big]$.

Moreover, we have convergence in probability

$$\left\| \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \Phi\big(f(x_i, y_j)\big) - \mathbb{E}\Big[\Phi\big(f(X,Y)\big)\Big] \right\| = O_p\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}\right), \quad (m, n \to \infty). \tag{24}$$

As an aside, note that (23) is an RKHS-valued two-sample U-statistic.

Proof For any i, j, we have $\mathbb{E}\big[\Phi\big(f(x_i, y_j)\big)\big] = \mathbb{E}\big[\Phi\big(f(X, Y)\big)\big]$; hence, (23) is unbiased.

The convergence (24) can be obtained as a corollary to Theorem 3, and the proof is omitted here. □

3.3 Approximating expansions

To keep computational cost limited, we need to use approximations when performing multi-step operations. If for instance, the outcome of the first step takes the form (23), then we already have $m \times n$ terms, and subsequent steps would further increase the number of terms, thus quickly becoming computationally prohibitive.

We can do so by using the methods described in Chap. 18 of Schölkopf and Smola (2002). They fall in two categories. In reduced set *selection* methods, we provide a set of expansion points [e.g., all points $f(x_i, y_j)$ in (23)], and the approximation method sparsifies the vector of expansion coefficients. This can be for instance done by solving eigenvalue problems or linear programs. Reduced set *construction* methods, on the other hand, construct new expansion points. In the simplest case, they proceed by sequentially finding approximate pre-images of RKHS elements. They tend to be computationally more demanding and suffer from local minima; however, they can lead to sparser expansions.

Either way, we will end up with an approximation

$$\sum_{k=1}^{p} \gamma_k \Phi(z_k) \tag{25}$$

of (23), where usually $p \ll m \times n$. Here, the z_k are either a subset of the $f(x_i, y_j)$, or other points from $\mathcal{Z}$.

It is instructive to consider some special cases. For simplicity, assume that $\mathcal{Z} = \mathbb{R}^d$. If we use a Gaussian kernel

$$k(x, x') = \exp\big(-\|x - x'\|^2/(2\sigma^2)\big) \tag{26}$$

whose bandwidth σ is much smaller than the closest pair of sample points, then the points mapped into the RKHS will be almost orthogonal, and there is no way to sparsify a kernel expansion such as (23) without incurring a large RKHS error. In this case, we can identify the estimator with the sample itself, and KPP reduces to a Monte Carlo method. If, on the other hand, we use a linear kernel $k(z, z') = \langle z, z'\rangle$ on $\mathcal{Z} = \mathbb{R}^d$, then Φ is the identity map and the expansion (23) collapses to one real number, i.e., we would effectively represent $f(X, Y)$ by its mean for any further processing. By choosing kernels that lie 'in between' these two extremes, we retain a varying amount of information which we can thus tune to our wishes, see Table 1.

3.4 Computing functions of RKHS approximations

More generally, consider approximations of kernel means $\mu[X]$ and $\mu[Y]$

$$\hat{\mu}[X] := \sum_{i=1}^{m'} \alpha_i \Phi_x(x_i'), \qquad \hat{\mu}[Y] := \sum_{j=1}^{n'} \beta_j \Phi_y(y_j'). \tag{27}$$

In our case, we think of (27) as RKHS-norm approximations of the outcome of previous operations performed on random variables. Such approximations typically have coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{n'}$ and $\boldsymbol{\beta} \in \mathbb{R}^{m'}$ that are not uniform, that may not sum to one, and that may take negative values (Schölkopf and Smola 2002), e.g., for conditional mean maps (Song et al. 2009; Fukumizu et al. 2013).

We propose to approximate the kernel mean $\mu\big[f(X, Y)\big]$ by the estimator

$$\hat{\mu}\big[f(X, Y)\big] := \frac{1}{\sum_{i=1}^{m'} \alpha_i \sum_{j=1}^{n'} \beta_j} \times \sum_{i=1}^{m'} \sum_{j=1}^{n'} \alpha_i \beta_j \Phi_z\big(f(x_i', y_j')\big), \tag{28}$$

where the feature map Φ_z defined on $\mathscr{Z}$, the range of f, may be different from both Φ_x and Φ_y. The expansion has $m' \times n'$ terms, which we can subsequently approximate more compactly in the form (25), ready for the next operation. Note that (28) contains (23) as a special case.

One of the advantages of our approach is that (23) and (28) apply for general data types. In other words, $\mathscr{X}$, $\mathscr{Y}$, $\mathscr{Z}$ need not be vector spaces—they may be arbitrary nonempty sets, as long as positive definite kernels can be defined on them.

3.4.1 Convergence analysis in an idealized setting

We analyze the convergence of (28) under the assumption that the expansion points are actually samples $x_1, \ldots, x_m$ from X and $y_1, \ldots, y_n$ from Y, which is for instance the case if the expansions (27) are the result of reduced set selection methods (cf. Sect. 3.3). Moreover, we assume that the expansion coefficients $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_n$ are constants, i.e., independent of the samples.

The following proposition gives a sufficient condition for the approximations in (27) to converge. Note that below, the coefficients $\alpha_1, \ldots, \alpha_m$ depend on the sample size m, but for simplicity we refrain from writing them as $\alpha_{1,m}, \ldots, \alpha_{m,m}$; and likewise, for $\beta_1, \ldots, \beta_n$. We make this remark to ensure that readers are not puzzled by the below statement that $\sum_{i=1}^m \alpha_i^2 \to 0$ as $m \to \infty$.

Proposition 1 *Let $x_1, \ldots, x_m$ be an i.i.d. sample and $(\alpha_i)_{i=1}^m$ be constants with $\sum_{i=1}^m \alpha_i = 1$. Assume $\mathbb{E}\big[k(X, X)\big] > \mathbb{E}\big[k(X, \tilde{X})\big]$, where X and $\tilde{X}$ are independent copies of x_i. Then, the convergence*

$$\mathbb{E}\left\| \sum_{i=1}^m \alpha_i \Phi(x_i) - \mu[X] \right\|^2 \to 0 \qquad (m \to \infty)$$

holds true if and only if $\sum_{i=1}^m \alpha_i^2 \to 0$ as $m \to \infty$.

Proof From the expansion

$$\begin{aligned} &\mathbb{E}\left\| \sum_{i=1}^m \alpha_i \Phi(x_i) - \mu[X] \right\|^2 \\ &= \sum_{i,s=1}^m \alpha_i \alpha_s \mathbb{E}\big[k(x_i, x_s)\big] - 2 \sum_{i=1}^m \alpha_i \mathbb{E}\big[k(x_i, X)\big] \\ &\quad + \mathbb{E}\big[k(X, \tilde{X})\big] \\ &= \Big(1 - \sum_i \alpha_i\Big)^2 \mathbb{E}\big[k(X, \tilde{X})\big] + \Big(\sum_i \alpha_i^2\Big)\Big\{\mathbb{E}\big[k(X, X)\big] \\ &\quad - \mathbb{E}\big[k(X, \tilde{X}\big]\Big\}, \end{aligned}$$

the assertion is straightforward. □

The next result shows that if our approximations (27) converge in the sense of Proposition 1, then the estimator (28) (with expansion coefficients summing to 1) is consistent.

Theorem 3 *Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be mutually independent i.i.d. samples, and the constants $(\alpha_i)_{i=1}^m, (\beta_j)_{j=1}^n$ satisfy $\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 1$. Assume $\sum_{i=1}^m \alpha_i^2$ and $\sum_{j=1}^n \beta_j^2$ converge to zero as $n, m \to \infty$. Then*

$$\begin{aligned} &\left\| \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j \Phi\big(f(x_i, y_j)\big) - \mu\big[f(X, Y)\big] \right\| \\ &= O_p\left(\sqrt{\sum_i \alpha_i^2} + \sqrt{\sum_j \beta_j^2} \right) \end{aligned}$$

as $m, n \to \infty$.

Proof By expanding and taking expectations, one can see that

$$\mathbb{E}\left\| \sum_{i=1}^m \sum_{j=1}^n \alpha_i \beta_j \Phi\big(f(x_i, y_j)\big) - \mathbb{E}\Big[\Phi\big(f(X, Y)\big)\Big] \right\|^2$$

equals

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_i^2\beta_j^2\mathbb{E}\left[k\big(f(X,Y),f(X,Y)\big)\right]\\
&\quad+\sum_{s\neq i}\sum_{j}\alpha_i\alpha_s\beta_j^2\mathbb{E}\left[k\big(f(X,Y),f\left(\tilde{X},Y\right)\big)\right]\\
&\quad+\sum_{i}\sum_{t\neq j}\alpha_i^2\beta_j\beta_t\mathbb{E}\left[k\left(f(X,Y),f(X,\tilde{Y})\right)\right]\\
&\quad+\sum_{s\neq i}\sum_{t\neq j}\alpha_i\alpha_s\beta_j\beta_t\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},\tilde{Y})\right)\right]\\
&\quad-2\sum_{i}\sum_{j}\alpha_i\beta_j\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},\tilde{Y})\right)\right]\\
&\quad+\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},\tilde{Y})\right)\right]\\
&=\Big(\sum_i\alpha_i^2\Big)\Big(\sum_j\beta_j^2\Big)\mathbb{E}\left[k\left(f(X,Y),f(X,Y)\right)\right]\\
&\quad+\left\{\Big(1-\sum_i\alpha_i\sum_j\beta_j\Big)^2+\sum_i\alpha_i^2\sum_j\beta_j^2\right.\\
&\quad\left.-\sum_i\alpha_i^2(\sum_j\beta_j)^2-(\sum_i\alpha_i)^2\sum_j\beta_j^2\right\}\\
&\quad\times\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},\tilde{Y})\right)\right]\\
&\quad+\Big((\sum_i\alpha_i)^2-\sum_i\alpha_i^2\Big)\Big(\sum_j\beta_j^2\Big)\\
&\quad\times\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},Y)\right)\right]\\
&\quad+\Big(\sum_i\alpha_i^2\Big)\Big((\sum_j\beta_j)^2-\sum_j\beta_j^2\Big)\\
&\quad\times\mathbb{E}\left[k\left(f(X,Y),f(X,\tilde{Y})\right)\right]\\
&=\Big(\sum_i\alpha_i^2\Big)\Big(\sum_j\beta_j^2\Big)\mathbb{E}\left[k(f(X,Y),f(X,Y))\right]\\
&\quad+\left\{\sum_i\alpha_i^2\sum_j\beta_j^2-\sum_i\alpha_i^2-\sum_j\beta_j^2\right\}\\
&\quad\times\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},\tilde{Y})\right)\right]\\
&\quad+\Big(1-\sum_i\alpha_i^2\Big)\Big(\sum_j\beta_j^2\Big)\mathbb{E}\left[k\left(f(X,Y),f(\tilde{X},Y)\right)\right]\\
&\quad+\Big(\sum_i\alpha_i^2\Big)\Big(1-\sum_j\beta_j^2\Big)\mathbb{E}\left[k\left(f(X,Y),f(X,\tilde{Y})\right)\right],
\end{aligned}
$$

which implies that the norm in the assertion of the theorem converges to zero at $O_p\left(\sqrt{\sum_i\alpha_i^2}+\sqrt{\sum_j\beta_j^2}\right)$ under the assumptions on α_i and β_j. Here, $(\tilde{X},\tilde{Y})$ is an independent copy of (X,Y). This concludes the proof. □

Note that in the simplest case, where $\alpha_i=1/m$ and $\beta_j=1/n$, we have $\sum_i\alpha_i^2=1/m$ and $\sum_j\beta_j^2=1/n$, which proves Theorem 2. It is also easy to see from the proof that we do not strictly need $\sum_i\alpha_i=\sum_j\beta_j=1$—for the estimator to be consistent, it suffices if the sums converge to 1. For a sufficient condition for this convergence, see Kanagawa and Fukumizu (2014).

3.4.2 More general expansion sets

To conclude our discussion of the estimator (28), we turn to the case where the expansions (27) are computed by reduced set construction, i.e., they are not necessarily expressed in terms of samples from X and Y. This is more difficult, and we do not provide a formal result, but just a qualitative discussion.

To this end, suppose the approximations (27) satisfy

$$\sum_{i=1}^{m'}\alpha_i=1 \text{ and for all } i,\ \alpha_i>0, \tag{29}$$

$$\sum_{j=1}^{n'}\beta_j=1 \text{ and for all } j,\ \beta_j>0, \tag{30}$$

and we approximate $\mu\big[f(X,Y)\big]$ by the quantity (28).

We assume that (27) are good approximations of the kernel means of two unknown random variables X and Y; we also assume that f and the kernel mean map along with its inverse are continuous. We have no samples from X and Y, but we can turn (27) into sample estimates based on *artificial* samples $\mathbf{X}$ and $\mathbf{Y}$, for which we can then appeal to our estimator from Theorem 2.

To this end, denote by $\mathbf{X}'=(x'_1,\ldots,x'_{m'})$ and $\mathbf{Y}'=(y'_1,\ldots,y'_{n'})$ the expansion points in (27). We construct a sample $\mathbf{X}=(x_1,x_2,\ldots)$ whose kernel mean is close to $\sum_{i=1}^{m'}\alpha_i\Phi_x(x'_i)$ as follows: for each i, the point x'_i appears in $\mathbf{X}$ with multiplicity $\lfloor m\cdot\alpha_i\rfloor$, i.e., the largest integer not exceeding $m\cdot\alpha_i$. This leads to a sample of size at most m. Note, moreover, that the multiplicity of x'_i, divided by m, differs from α_i by at most $1/m$, so effectively we have quantized the α_i coefficients to this accuracy.

Since m' is constant, this implies that for any $\varepsilon>0$, we can choose $m\in\mathbb{N}$ large enough to ensure that

$$\left\|\frac{1}{m}\sum_{1=1}^{m}\Phi_x(x_i)-\sum_{i=1}^{m'}\alpha_i\Phi_x(x'_i)\right\|^2<\varepsilon. \tag{31}$$

We may thus work with $\frac{1}{m}\sum_{1=1}^{m}\Phi_x(x_i)$, which for strictly positive definite kernels corresponds uniquely to the sample $\mathbf{X}=(x_1,\ldots,x_m)$. By the same argument, we obtain a sample $\mathbf{Y}=(y_1,\ldots,y_n)$ approximating the second expansion.

Substituting both samples into the estimator from Theorem 2 leads to

$$\hat{\mu}\big[f(X,Y)\big] = \frac{1}{\sum_{i=1}^{m'}\hat{\alpha}_i \sum_{j=1}^{n'}\hat{\beta}_j} \times \sum_{i=1}^{m'}\sum_{j=1}^{n'}\hat{\alpha}_i\hat{\beta}_j \Phi_z\big(f(x'_i, y'_j)\big), \tag{32}$$

where $\hat{\alpha}_i = \lfloor m \cdot \alpha_i \rfloor / m$, and $\hat{\beta}_i = \lfloor n \cdot \beta_i \rfloor / n$. By choosing sufficiently large m, n, this becomes an arbitrarily good approximation (in the RKHS norm) of the proposed estimator (28). Note, however, that we cannot claim based on this argument that this estimator is consistent, not the least since Theorem 2 in the stated form requires i.i.d. samples.

3.4.3 Larger sets of random variables

Without analysis, we include the estimator for the case of more than two variables: Let g be a measurable function of jointly independent RVs $U_j (j = 1, \ldots, p)$. Given i.i.d. observations $u_1^j, \ldots, u_m^j \sim U_j$, we have

$$\frac{1}{m^p}\sum_{m_1,\ldots,m_p=1}^{m} \Phi\left(g\big(u^1_{m_1}, \ldots, u^p_{m_p}\big)\right) \xrightarrow{m\to\infty} \mu\left[g\big(U_1, \ldots, U_p\big)\right] \tag{33}$$

in probability. Here, in order to keep notation simple, we have assumed that the sample sizes for each RV are identical.

As above, we note that (i) g need not be real-valued, it can take values in some set $\mathscr{Z}$ for which we have a (possibly characteristic) positive definite kernel; (ii) we can extend this to general kernel expansions like (28); and (iii) if we use Gaussian kernels with width tending to 0, we can think of the above as a sampling method.

4 Dependent RVs and structural equation models

For dependent RVs, the proposed estimators are not applicable. One way to handle dependent RVs is to appeal to the fact that any joint distribution of random variables can be written as a structural equation model with independent noises. This leads to an interesting application of our method to the field of causal inference.

We consider a model $X_i = f_i(\mathrm{PA}_i, U_i)$, for $i = 1, \ldots, p$, with jointly independent noise terms $U_1, \ldots, U_p$. Such models arise for instance in causal inference (Pearl 2009). Each random variable X_i is computed as a function f_i of its noise term U_i and its parents PA_i in an underlying directed acyclic graph (DAG). Every graphical model w.r.t. a DAG can be expressed as such a structural equation model with suitable functions and noise terms [e.g., Peters et al. (2014)].

If we recursively substitute the parent equations, we can express each X_i as a function of only the independent noise terms $U_1, \ldots, U_p$,

$$X_i = g_i(U_1, \ldots, U_p). \tag{34}$$

Since we know how to compute functions of independent RVs, we can try to test such a model (assuming knowledge of all involved quantities) by estimating the distance between RKHS images,

$$\Delta = \big\|\mu[X_i] - \mu\big[g_i(U_1, \ldots, U_p)\big]\big\|^2 \tag{35}$$

using the estimator described in (33) (we discuss the bivariate case in Theorem 4). It may be unrealistic to assume that we have access to all quantities. However, there is a special case where this is conceivable, which we will presently discuss. This is the case of additive noise models (Peters et al. 2014)

$$Y = f(X) + U, \qquad \text{with } X \perp\!\!\!\perp U. \tag{36}$$

Such models are of interest for cause–effect inference since it is known (Peters et al. 2014) that in the generic case, a model (36) can only be fit in one direction, i.e., if (36) holds true, then we cannot simultaneously have

$$X = g(Y) + V, \qquad \text{with } Y \perp\!\!\!\perp V. \tag{37}$$

To measure how well (36) fits the data, we define an estimator

$$\Delta_{emp} := \left\| \frac{1}{m}\sum_{i=1}^{m}\Phi(y_i) - \frac{1}{m^2}\sum_{i,j=1}^{m}\Phi\big(f(x_i) + u_j\big)\right\|^2. \tag{38}$$

Analogously, we define the estimator in the backward direction

$$\Delta^{bw}_{emp} := \left\| \frac{1}{m}\sum_{i=1}^{m}\Phi(x_i) - \frac{1}{m^2}\sum_{i,j=1}^{m}\Phi\big(g(y_i) + v_j\big)\right\|^2. \tag{39}$$

Here, we assume that we are given the conditional mean functions $f : x \mapsto \mathbb{E}[Y \mid X = x]$ and $g : y \mapsto \mathbb{E}[X \mid Y = y]$.

In practice, we would apply the following procedure: we are given a sample $(x_1, y_1), \ldots, (x_m, y_m)$. We estimate the function f as well as the residual noise terms $u_1, \ldots, u_m$ by regression, and likewise for the backward function g (Peters

et al. 2014). Strictly speaking, we need to use separate subsamples to estimate function and noise terms, respectively, see Kpotufe et al. (2014).

Below, we show that Δ_{emp} converges to 0 for additive noise models (36). For the incorrect model (37), however, Δ^{bw}_{emp} will in the generic case not converge to zero. We can thus use the comparison of both values for deciding causal direction.

Theorem 4 *Suppose $x_1, \ldots, x_m$ and $u_1, \ldots, u_m$ are mutually independent i.i.d. samples, and $y_i = f(x_i) + u_i$. Assume further that the direction of the additive noise model is identifiable* (Peters et al. 2014) *and the kernel for x is characteristic. We then have*

$$\Delta_{emp} \to 0 \quad \textit{and} \tag{40}$$

$$\Delta^{bw}_{emp} \not\to 0 \tag{41}$$

in probability as $m \to \infty$.

Proof Equation (40) follows from Theorem 2 since $\left\| \frac{1}{m} \sum_{i=1}^{m} \Phi(y_i) - \mu[Y] \right\| \to 0$ and $\left\| \frac{1}{m^2} \sum_{i,j=1}^{m^2} \Phi(f(x_i) + u_j) - \mu[Y] \right\| \to 0$ (all convergences in this proof are in probability).

To prove (41), we assume that $\Delta^{bw}_{emp} \to 0$ which implies

$$\left\| \frac{1}{m^2} \sum_{i,j=1}^{m} \Phi\big(g(y_i) + v_j\big) - \mu[X] \right\| \to 0. \tag{42}$$

The key idea is to introduce a random variable $\tilde{V}$ that has the same distribution as V but is independent of Y and to consider the following decomposition of the sum appearing in (42):

$$\begin{aligned}
\frac{1}{m^2} \sum_{i,j=1}^{m} \Phi\big(g(y_i) + v_j\big) &= \frac{1}{m^2} \sum_{i=1}^{m} \Phi\big(g(y_i) + v_i\big) \\
&\quad + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{k=1}^{m-1} \Phi\big(g(y_i) + v_{i+k}\big) \\
&= \frac{1}{m} \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i) \\
&\quad + \frac{1}{m} \sum_{k=1}^{m-1} \frac{1}{m} \sum_{i=1}^{m} \Phi\big(g(y_i) + v_{i+k}\big) \\
&=: A_m + B_m,
\end{aligned}$$

where the index for v is interpreted modulo m, for instance, $v_{m+3} := v_3$. Since $v_{i+k} = x_{i+k} - g(y_{i+k})$ is independent of y_i, it further follows from Theorem 2 that $\| A_m - \frac{1}{m} \mu[X] \| \to 0$ and $\left\| B_m - \frac{m-1}{m} \mu\big[g(Y) + \tilde{V}\big] \right\| \to 0$. Therefore,

$$\left\| A_m + B_m - \frac{1}{m} \mu[X] - \frac{m-1}{m} \mu\big[g(Y) + \tilde{V}\big] \right\| \to 0.$$

Together with (42), this implies

$$\left\| \mu[X] - \frac{1}{m} \mu[X] - \frac{m-1}{m} \mu\big[g(Y) + \tilde{V}\big] \right\| \to 0$$

and therefore

$$\mu\big[g(Y) + \tilde{V}\big] = \mu[X].$$

Since the kernel is characteristic, this implies

$$g(Y) + \tilde{V} = X \quad \text{(in distribution)},$$

with $Y \perp\!\!\!\perp \tilde{V}$, which contradicts the identifiability of the additive noise model. □

As an aside, note that Theorem 4 would not hold if in (39) we were to estimate $\mu\big[g(Y) + V\big]$ by $\frac{1}{m} \sum_{i=1}^{m} \Phi\big(g(y_i) + v_i\big)$ instead of $\frac{1}{m^2} \sum_{i,j=1}^{m} \Phi\big(g(y_i) + v_j\big)$.

5 Experiments

5.1 Synthetic data

We consider basic arithmetic expressions that involve multiplication $X \times Y$, division X/Y, and exponentiation X^Y on two independent scalar RVs X and Y. Letting $p_X = \mathcal{N}(3, 0.5)$ and $p_Y = \mathcal{N}(4, 0.5)$, we draw i.i.d. samples $\mathbf{X} = \{x_1, \ldots, x_m\}$ and $\mathbf{Y} = \{y_1, \ldots, y_m\}$ from p_X and p_Y.

In the experiment, we are interested in the convergence (in RKHS norm) of our estimators to $\mu\big[f(X, Y)\big]$. Since we do not have access to the latter, we use an independent sample to construct a proxy $\hat{\mu}\big[f(X, Y)\big] = (1/\ell^2) \sum_{i,j=1}^{\ell} \Phi_z\big(f(x_i, y_j)\big)$. We found that $\ell = 100$ led to a sufficiently good approximation.

Next, we compare three estimators, referred to as μ_1, μ_2, and μ_3 below, for sample sizes m ranging from 10 to 50:

1. The sample-based estimator (23).
2. The estimator (28) based on approximations of the kernel means, taking the form $\hat{\mu}[X] := \sum_{i=1}^{m'} \alpha_i \Phi_x(x_i)$ and $\hat{\mu}[Y] := \sum_{j=1}^{m'} \beta_j \Phi_y(y_j)$ of $\mu[X]$ and $\mu[Y]$, respectively. We used the simplest possible reduced set selection method: we randomly subsampled subsets of size $m' \approx 0.4 \cdot m$ from $\mathbf{X}$ and $\mathbf{Y}$, and optimized the coefficients $\{\alpha_1, \ldots, \alpha_{m'}\}$ and $\{\beta_1, \ldots, \beta_{m'}\}$ to best approximate the original kernel means (based on $\mathbf{X}$ and $\mathbf{Y}$) in the RKHS norm (Schölkopf and Smola 2002, Sect. 18.3).

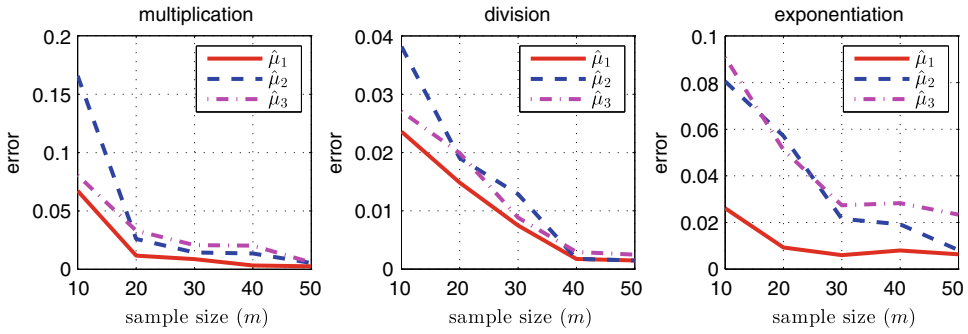
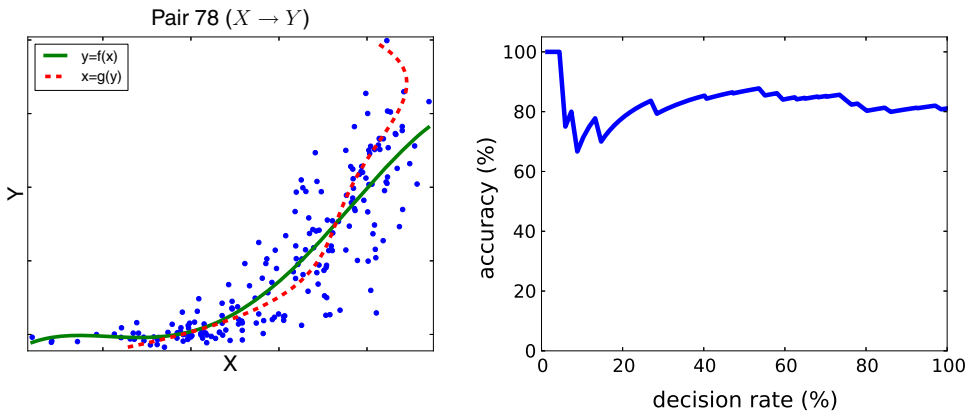


Fig. 1 Error of the proposed estimators for three arithmetic operations—multiplication $X \times Y$, division X/Y, and exponentiation X^Y—as a function of sample size m. The error reported is an average of 30 repetitions of the simulations. The expensive estimator $\hat{\mu}_1$ [see (23)] performs best. The approximation $\hat{\mu}_2$ [see (28)] works well as sample sizes increase.



(a) Pair 78 and regressors f, g **(b)** the accuracy curve

Fig. 2 **a** Scatter plot of the data of causal pair 78 in the `CauseEffectPairs` benchmarks, along with the forward and backward function fits, $y = f(x)$ and $x = g(y)$. **b** Accuracy of cause–effect decisions on all the 81 pairs in the `CauseEffectPairs` benchmarks

3. Analogously to the case of one variable (17), we may also look at the estimator $\hat{\mu}_3[\mathbf{X}, \mathbf{Y}] := (1/m) \sum_{i=1}^{m} \Phi_z\big(f(x_i, y_i)\big)$, which sums only over m mutually independent terms, i.e., a small fraction of all terms of (23).

For $i = 1, 2, 3$, we evaluate the estimates $\hat{\mu}_i\big[f(X, Y)\big]$ using the error measure

$$L = \big\|\hat{\mu}_i\big[f(X, Y)\big] - \hat{\mu}\big[f(X, Y)\big]\big\|^2. \tag{43}$$

We use (6) to evaluate L in terms of kernels. In all cases, we employ a Gaussian RBF kernel (26) whose bandwidth parameter is chosen using the median heuristic, setting σ to the median of the pairwise distances of distinct data points (Gretton et al. 2005).

Figure 1 depicts the error (43) as a function of sample size m. For all operations, the error decreases as sample size increases. Note that Φ_z is different across the three operations, resulting in different scales of the average error in Fig. 1.

5.2 Causal discovery via functions of kernel means

We also apply our KPP approach to bivariate causal inference problem (cf. Sect. 4). That is, given a pair of real-valued random variables X and Y with joint distribution p_{XY}, we are interested in identifying whether X causes Y (denote as $X \to Y$) or Y causes X (denote as $Y \to X$) based on observational data. We assume an additive noise model $E = f(C) + U$ with $C \perp\!\!\!\perp U$ where C, E, and U denote cause, effect, and residual (or "unexplained") variable, respectively. Below we present a preliminary result on the `CauseEffectPairs` benchmark data set (Peters et al. 2014).

For each causal pair $(\mathbf{X}, \mathbf{Y}) = \big\{(x_1, y_1), \ldots, (x_m, y_m)\big\}$, we estimate functions $y \approx f(x)$ and $x \approx g(y)$ as least-square fits using degree 4 polynomials. We illustrate one example in Fig. 2a. Next, we compute the residuals in both directions as $u_i = y_i - f(x_i)$ and $v_j = x_j - g(y_j)$.[1] Finally, we compute scores $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$ by

$$\Delta_{X\to Y} := \left\| \frac{1}{m}\sum_{i=1}^{m} \Phi(y_i) - \frac{1}{m^2}\sum_{i,j=1}^{m} \Phi\big(f(x_i) + u_j\big) \right\|^2,$$

$$\Delta_{Y\to X} := \left\| \frac{1}{m}\sum_{i=1}^{m} \Phi(x_i) - \frac{1}{m^2}\sum_{i,j=1}^{m} \Phi\big(g(y_i) + v_j\big) \right\|^2.$$

Following Theorem 4, we can use the comparison between $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$ to infer the causal direction. Specifi-

[1] For simplicity, this was done using the same data; but cf. our discussion following (38).

cally, we decide that $X \to Y$ if $\Delta_{X\to Y} < \Delta_{Y\to X}$, and that $Y \to X$ otherwise. In this experiment, we also use a Gaussian RBF kernel whose bandwidth parameter is chosen using the median heuristic. To speed up the computation of $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$, we adopted a finite approximation of the feature map using 100 random Fourier features [see Rahimi and Recht (2007) for details]. We allow the method to abstain whenever the two values are closer than $\delta > 0$. By increasing δ, we can compute the method's accuracy as a function of a decision rate (i.e., the fraction of decisions that our method is forced to make) ranging from 100 % to 0 %.

Figure 2b depicts the accuracy versus the decision rate for the 81 pairs in the `CauseEffectPairs` benchmark collection. The method achieves an accuracy of 80 %, which is significantly better than random guessing, when forced to infer the causal direction of all 81 pairs.

6 Conclusions

We have developed a kernel-based approach to compute functional operations on random variables taking values in arbitrary domains. We have proposed estimators for RKHS representations of those quantities, evaluated the approach on synthetic data, and showed how it can be used for cause–effect inference. While the results are encouraging, the material presented in this article only describes the main ideas, and much remains to be done. We believe that there is significant potential for a unified perspective on probabilistic programming based on the described methods, and hope that some of the open problems will be addressed in future work.

Acknowledgments Thanks to Dominik Janzing, Le Song, and Ilya Tolstikhin for dicussions and comments.

References

Bach, F.R., Jordan, M.I.: Kernel independent component analysis. J. Mach. Learn. Res. **3**, 1–48 (2002)

Berlinet, A., Agnan, T.C.: Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer Academic Publishers, Boston (2004)

Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT '92, pp. 144–152. ACM, New York (1992)

Cassel, J.: Probabilistic programming with stochastic memoization: implementing non-parametric bayesian inference. Math. J. **16** (2014). doi:10.3888/tmj.16-1

Epstein, B.: Some applications of the Mellin transform in statistics. Ann. Math. Stat. **19**(3), 370–379 (1948)

Ferson, S.: What Monte Carlo methods cannot do. Hum. Ecol. Risk Assess.: Int. J. **2**(4), 990–1007 (1996). doi:10.1080/10807039609383659

Fukumizu, K., Bach, F., Jordan, M.I.: Kernel dimension reduction in regression. Ann. Stat. **37**(4), 1871–1905 (2009)

Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B.: Kernel measures of conditional dependence. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol. 20, pp. 489–496. Curran, Red Hook (2008)

Fukumizu, K., Song, L., Gretton, A.: Kernel Bayes' rule: Bayesian inference with positive definite kernels. J. Mach. Learn. Res. **14**, 3753–3783 (2013)

Gordon, A.D., Henzinger, T.A., Nori, A.V., Rajamani, S.K.: Probabilistic programming. In: International Conference on Software Engineering (ICSE, FOSE track) (2014)

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. J. Mach. Learn. Res. **13**, 723–773 (2012)

Gretton, A., Bousquet, O., Smola, A.J., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) Algorithmic Learning Theory: 16th International Conference, pp. 63–78. Springer, Berlin (2005)

Gretton, A., Smola, A.J., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B.: Covariate shift by kernel mean matching. In: Candela, J.Q., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (eds.) Dataset Shift in Machine Learning, pp. 131–160. MIT Press, Cambridge (2009)

Harmeling, S., Hirsch, M., Schlkopf, B.: On a link between kernel mean maps and Fraunhofer diffraction, with an application to super-resolution beyond the diffraction limit. In: CVPR, pp. 1083–1090. IEEE (2013)

Hofmann, T., Schölkopf, B., Smola, A.J.: Kernel methods in machine learning. Ann. Stat. 36(3), 1171–1220 (2008)

Huang, J., Smola, A., Gretton, A., Borgwardt, K., Schölkopf, B.: Correcting sample selection bias by unlabeled data. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems (NIPS), pp. 601–608. MIT Press, Cambridge (2007)

James, W., Stein, C.: Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob. 361–379 (1961)

Jaroszewicz, S., Korzen, M.: Arithmetic operations on independent random variables: a numerical approach. SIAM J. Sci. Comput. **34**(3), A1241–A1265 (2012)

Kanagawa, M., Fukumizu, K.: Recovering distributions from Gaussian RKHS embeddings. In: JMLR W&CP 33 (Proc. AISTATS 2014), pp. 457–465. (2014)

Kpotufe, S., Sgouritsa, E., Janzing, D., Schölkopf, B.: Consistency of causal inference under the additive noise model. In: ICML (2014)

Lopez-Paz, D., Muandet, K., Schölkopf, B., Tolstikhin, I.: Towards a learning theory of cause-effect inference. In: Proceedings of the 32nd International Conference on Machine Learning, JMLR: W&CP, Lille, France, in press (2015)

Milios, D.: Probability Distributions as Program Variables. Master's thesis, School of Informatics, University of Edinburgh (2009)

Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. arXiv:1412.3773 (2014). Accessed 18 Jan 2015

Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML, pp. 10–18 (2013)

Muandet, K., Fukumizu, K., Dinuzzo, F., Schölkopf, B.: Learning from distributions via support measure machines. In: Bartlett, P., Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 10–18. Curran Associates, Inc., Red Hook (2012)

Muandet, K., Fukumizu, K., Sriperumbudur, B., Gretton, A., Schölkopf, B.: Kernel mean estimation and Stein effect. In: Xing, E.P., Jebara, T. (eds.). Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP, vol. 32. (2014a)

Muandet, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean estimation via spectral filtering. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 27. Curran Associates Inc. (2014b)

Paige, B., Wood, F.: A compilation target for probabilistic programming languages. In: Journal of Machine Learning Research; ICML 2014, pp. 1935–1943 (2014)

Pearl, J.: Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, New York (2009)

Peters, J., Mooij, J., Janzing, D., Schölkopf, B.: Causal discovery with continuous additive noise models. J. Mach. Learn. Res. **15**, 2009–2053 (2014)

Prasad, R.D.: Probability distributions of algebraic functions of independent random variables. SIAM J. Appl. Math. **18**(3), 614–626 (1970)

Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems, vol 20, pp. 1177–1184. Curran Associates, Inc. (2007)

Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)

Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge, MA, USA (2002)

Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)

Smola, A., Gretton, A., Song, L., Schölkopf, B.: A Hilbert space embedding for distributions. In: Proc. Algorithmic Learning Theory, pp. 13–31. Springer-Verlag (2007)

Song, L.: Learning via Hilbert space embedding of distributions. Ph.D. thesis, The School of Information Technologies, The University of Sydney (2008)

Song, L., Boots, B., Siddiqi, S.M., Gordon, G., Smola, A.J.: Hilbert space embeddings of hidden Markov models. In: Proceedings of the 27th International Conference on Machine Learning (ICML) (2010)

Song, L., Gretton, A., Bickson, D., Low, Y., Guestrin, C.: Kernel belief propagation. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS) (2011)

Song, L., Huang, J., Smola, A., Fukumizu, K.: Hilbert space embeddings of conditional distributions with applications to dynamical systems. In: ICML (2009)

Song, L., Zhang, X., Smola, A.J., Gretton, A., Schölkopf, B.: Tailoring density estimation via reproducing kernel moment matching. In: Cohen, W.W., McCallum, A., Roweis, S. (eds.). In: Proceedings of the 25th International Conference on Machine Learning, pp. 992–999. ACM Press, New York (2008)

Springer, M.: The Algebra of Random Variables, Wiley Series in Probability and Mathematical Statistics. Wiley, Hoboken (1979)

Springer, M.D., Thompson, W.E.: The distribution of products of independent random variables. SIAM J. Appl. Math. **14**(3), 511–526 (1966)

Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Lanckriet, G., Schölkopf, B.: Injective Hilbert space embeddings of probability measures. In: The 21st Annual Conference on Learning Theory (COLT) (2008)

Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.: Hilbert space embeddings and metrics on probability measures. J. Mach. Learn. Res. **99**, 1517–1561 (2010)

Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. J. Mach. Learn. Res. **2**, 67–93 (2002)

Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)

Szabó, Z., Gretton, A., Póczos, B., Sriperumbudur, B.: Two-stage sampled learning theory on distributions. arXiv:1402.1754 (2014). Accessed 18 Jan 2015

Williamson, R.C.: Probabilistic arithmetic. Ph.D. thesis, Department of Electrical Engineering, University of Queensland, St. Lucia, Queensland, Australia (1989)

Wood, F., van de Meent, J.W., Mansinghka, V.: A new approach to probabilistic programming inference. In: Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS), JMLR: W&CP, vol. 33. Reykjavik, Iceland (2014)

Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based Conditional Independence Test and Application in Causal Discovery. In: Cozman, F., Pfeffer, A. (eds.). In: 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011), pp. 804–813. AUAI Press, Corvallis, OR, USA (2011)