Least-Squares Independence Regression for Non-Linear Causal Inference under Non-Gaussian Noise

Makoto Yamada
Yahoo Labs
makotoy@yahoo-inc.com

Masashi Sugiyama
Tokyo Institute of Technology, Japan.
sugi@cs.titech.ac.jp http://sugiyama-www.cs.titech.ac.jp/~sugi

Jun Sese
Tokyo Institute of Technology, Japan.
sesejun@cs.titech.ac.jp

Abstract

The discovery of non-linear causal relationship under additive non-Gaussian noise models has attracted considerable attention recently because of their high flexibility. In this paper, we propose a novel causal inference algorithm called *least-squares independence regression* (LSIR). LSIR learns the additive noise model through the minimization of an estimator of the *squared-loss mutual information* between inputs and residuals. A notable advantage of LSIR is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized by cross-validation, allowing us to avoid overfitting in a data-dependent fashion. Through experiments with real-world datasets, we show that LSIR compares favorably with a state-of-the-art causal inference method.

Keywords

Causal inference, Non-Linear, Non-Gaussian, Squared-loss mutual information, Least-Squares Independence Regression

1 Introduction

Learning *causality* from data is one of the important challenges in the artificial intelligence, statistics, and machine learning communities (Pearl, 2000). A traditional method

of learning causal relationship from observational data is based on the linear-dependence Gaussian-noise model (Geiger and Heckerman, 1994). However, the linear-Gaussian assumption is too restrictive and may not be fulfilled in practice. Recently, non-Gaussianity and non-linearity have been shown to be beneficial in causal inference, allowing one to break symmetry between observed variables (Shimizu *et al.*, 2006; Hoyer *et al.*, 2009). Since then, much attention has been paid to the discovery of non-linear causal relationship through non-Gaussian noise models (Mooij *et al.*, 2009).

In the framework of non-linear non-Gaussian causal inference, the relation between a cause X and an effect Y is assumed to be described by $Y = f(X) + E$, where f is a non-linear function and E is non-Gaussian additive noise which is independent of the cause X. Given two random variables X and X', the causal direction between X and X' is decided based on a hypothesis test of whether the causal model $X' = f(X) + E$ or the alternative model $X = f'(X') + E'$ fits the data well—here, the goodness of fit is measured by independence between inputs and residuals (i.e., estimated noise). Hoyer *et al.* (2009) proposed to learn the functions f and f' by *Gaussian process* (GP) regression (Bishop, 2006), and evaluate the independence between inputs and residuals by the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005).

However, since standard regression methods such as GP are designed to handle Gaussian noise, they may not be suited for discovering causality in the non-Gaussian additive noise formulation. To cope with this problem, a novel regression method called *HSIC regression* (HSICR) has been introduced recently (Mooij *et al.*, 2009). HSICR learns a function so that the dependence between inputs and residuals is directly minimized based on HSIC. Since HSICR does not impose any parametric assumption on the distribution of additive noise, it is suited for non-linear non-Gaussian causal inference. Indeed, HSICR was shown to outperform the GP-based method in experiments (Mooij *et al.*, 2009).

However, HSICR still has limitations for its practical use. The first weakness of HSICR is that the kernel width of HSIC needs to be determined manually. Since the choice of the kernel width heavily affects the sensitivity of the independence measure (Fukumizu *et al.*, 2009), lack of systematic model selection strategies is critical in causal inference. Setting the kernel width to the median distance between sample points is a popular heuristic in kernel methods (Schölkopf and Smola, 2002), but this does not always perform well in practice. Another limitation of HSICR is that the kernel width of the regression model is fixed to the same value as HSIC. This crucially limits the flexibility of function approximation in HSICR.

To overcome the above weaknesses, we propose an alternative regression method for causal inference called *least-squares independence regression* (LSIR). As HSICR, LSIR also learns a function so that the dependence between inputs and residuals is directly minimized. However, a difference is that, instead of HSIC, LSIR adopts an independence criterion called *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009), which is a consistent estimator of the *squared-loss mutual information* (SMI) with the optimal convergence rate. An advantage of LSIR over HSICR is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized through cross-validation (CV) with respect to the LSMI criterion.

Furthermore, we propose to determine the kernel width of the regression model based on CV with respect to SMI itself. Thus, the kernel width of the regression model is determined independent of that in the independence measure. This allows LSIR to have higher flexibility in non-linear causal inference than HSICR. Through experiments with benchmark and real-world biological datasets, we demonstrate the superiority of LSIR.

A preliminary version of this work appeared in Yamada and Sugiyama (2010); here we provide a more comprehensive derivation and discussion of LSIR, as well as a more detailed experimental section.

2 Dependence Minimizing Regression by LSIR

In this section, we formulate the problem of dependence minimizing regression and propose a novel regression method, *least-squares independence regression* (LSIR). Suppose random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ are connected by the following additive noise model (Hoyer *et al.*, 2009):

$$Y = f(X) + E,$$

where $f : \mathbb{R} \to \mathbb{R}$ is some non-linear function and $E \in \mathbb{R}$ is a zero-mean random variable independent of X. The goal of dependence minimizing regression is, from i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to obtain a function $\widehat{f}$ such that input X and estimated additive noise $\widehat{E} = Y - \widehat{f}(X)$ are independent.

Let us employ a linear model for dependence minimizing regression:

$$f_{\boldsymbol{\beta}}(x) = \sum_{l=1}^{m} \beta_l \psi_l(x) = \boldsymbol{\beta}^\top \boldsymbol{\psi}(x), \tag{1}$$

where m is the number of basis functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ are regression parameters, $^\top$ denotes the transpose, and $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_m(x))^\top$ are basis functions. We use the Gaussian basis function in our experiments:

$$\psi_l(x) = \exp\left(-\frac{(x - c_l)^2}{2\tau^2}\right),$$

where c_l is the Gaussian center chosen randomly from $\{x_i\}_{i=1}^n$ without overlap and τ is the kernel width. In dependence minimizing regression, we learn the regression parameter $\boldsymbol{\beta}$ as

$$\min_{\boldsymbol{\beta}} \left[I(X, \widehat{E}) + \frac{\gamma}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

where $I(X, \widehat{E})$ is some measure of independence between X and $\widehat{E}$, and $\gamma \geq 0$ is the regularization parameter to avoid overfitting.

In this paper, we use the *squared-loss mutual information* (SMI) (Suzuki *et al.*, 2009) as our independence measure:

$$\mathrm{SMI}(X, \widehat{E}) = \frac{1}{2} \iint \left(\frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} - 1 \right)^2 p(x)p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e}. \tag{2}$$

$\mathrm{SMI}(X, \widehat{E})$ is the *Pearson divergence* (Pearson, 1900) from $p(x, \widehat{e})$ to $p(x)p(\widehat{e})$, and it vanishes if and only if $p(x, \widehat{e})$ agrees with $p(x)p(\widehat{e})$, i.e., X and $\widehat{E}$ are statistically independent. Note that ordinary *mutual information* (MI) (Cover and Thomas, 2006),

$$\mathrm{MI}(X, \widehat{E}) = \iint p(x, \widehat{e}) \log \frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} \mathrm{d}x\mathrm{d}\widehat{e}, \tag{3}$$

corresponds to the *Kullback-Leibler divergence* (Kullback and Leibler, 1951) from $p(x, \widehat{e})$ and $p(x)p(\widehat{e})$, and it can also be used as an independence measure. Nevertheless, we adhere to using SMI since it allows us to obtain an analytic-form estimator, as explained below.

2.1 Estimation of Squared-Loss Mutual Information

SMI cannot be directly computed since it contains unknown densities $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$. Here, we briefly review an SMI estimator called *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009).

Since density estimation is known to be a hard problem (Vapnik, 1998), avoiding density estimation is critical for obtaining better SMI approximators (Kraskov *et al.*, 2004). A key idea of LSMI is to directly estimate the *density ratio*,

$$r(x, \widehat{e}) = \frac{p(x, \widehat{e})}{p(x)p(\widehat{e})},$$

without going through density estimation of $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$.

In LSMI, the density ratio function $r(x, \widehat{e})$ is directly modeled by the following linear model:

$$r_{\boldsymbol{\alpha}}(x, \widehat{e}) = \sum_{l=1}^{b} \alpha_l \varphi_l(x, \widehat{e}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(x, \widehat{e}), \tag{4}$$

where b is the number of basis functions, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$ are parameters, and $\boldsymbol{\varphi}(x, \widehat{e}) = (\varphi_1(x, \widehat{e}), \ldots, \varphi_b(x, \widehat{e}))^\top$ are basis functions. We use the Gaussian basis function:

$$\varphi_l(x, \widehat{e}) = \exp\left(-\frac{(x - u_l)^2 + (\widehat{e} - \widehat{v}_l)^2}{2\sigma^2}\right),$$

where $(u_l, \widehat{v}_l)$ is the Gaussian center chosen randomly from $\{(x_i, \widehat{e}_i)\}_{i=1}^n$ without replacement, and σ is the kernel width.

The parameter $\boldsymbol{\alpha}$ in the density-ratio model $r_{\boldsymbol{\alpha}}(x, \widehat{e})$ is learned so that the following squared error $J_0(\boldsymbol{\alpha})$ is minimized:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2} \iint (r_{\boldsymbol{\alpha}}(x, \widehat{e}) - r(x, \widehat{e}))^2 p(x)p(\widehat{e}) \mathrm{d}x\mathrm{d}\widehat{e} \\ &= \frac{1}{2} \iint r_{\boldsymbol{\alpha}}^2(x, \widehat{e}) p(x)p(\widehat{e}) \mathrm{d}x\mathrm{d}\widehat{e} - \iint r_{\boldsymbol{\alpha}}(x, \widehat{e}) p(x, \widehat{e}) \mathrm{d}x\mathrm{d}\widehat{e} + C, \end{aligned}$$

where C is a constant independent of $\boldsymbol{\alpha}$ and therefore can be safely ignored. Let us denote the first two terms by $J(\boldsymbol{\alpha})$:

$$J(\boldsymbol{\alpha}) = J_0(\boldsymbol{\alpha}) - C = \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha}, \tag{5}$$

where

$$\boldsymbol{H} = \iint \boldsymbol{\varphi}(x,\widehat{e})\boldsymbol{\varphi}(x,\widehat{e})^\top p(x)p(\widehat{e})\mathrm{d}x\mathrm{d}\widehat{e},$$
$$\boldsymbol{h} = \iint \boldsymbol{\varphi}(x,\widehat{e})p(x,\widehat{e})\mathrm{d}x\mathrm{d}\widehat{e}.$$

Approximating the expectations in $\boldsymbol{H}$ and $\boldsymbol{h}$ by empirical averages, we obtain the following optimization problem:

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}}\Big[\frac{1}{2}\boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}}\boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha}\Big],$$

where a regularization term $\frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{\alpha}$ is included for avoiding overfitting, and

$$\widehat{\boldsymbol{H}} = \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{\varphi}(x_i,\widehat{e}_j)\boldsymbol{\varphi}(x_i,\widehat{e}_j)^\top,$$
$$\widehat{\boldsymbol{h}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(x_i,\widehat{e}_i).$$

Differentiating the above objective function with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1}\widehat{\boldsymbol{h}}, \tag{6}$$

where $\boldsymbol{I}_b$ denotes the b-dimensional identity matrix. It was shown that LSMI is consistent under mild assumptions and it achieves the optimal convergence rate (Kanamori *et al.*, 2012).

Given a density ratio estimator $\widehat{r} = r_{\widehat{\boldsymbol{\alpha}}}$, SMI defined by Eq.(2) can be simply approximated by samples via the *Legendre-Fenchel convex duality* of the divergence functional as follows (Rockafellar, 1970; Suzuki and Sugiyama, 2013):

$$\begin{aligned}\widehat{\mathrm{SMI}}(X,\widehat{E}) &= \frac{1}{n}\sum_{i=1}^{n}\widehat{r}(x_i,\widehat{e}_i) - \frac{1}{2n^2}\sum_{i,j=1}^{n}\widehat{r}(x_i,\widehat{e}_j)^2 - \frac{1}{2}\\ &= \widehat{\boldsymbol{h}}^\top\widehat{\boldsymbol{\alpha}} - \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{H}}\widehat{\boldsymbol{\alpha}} - \frac{1}{2}.\end{aligned} \tag{7}$$

2.2 Model Selection in LSMI

LSMI contains three tuning parameters: the number of basis functions b, the kernel width σ, and the regularization parameter λ. In our experiments, we fix $b = \min(200, n)$ (i.e., $\boldsymbol{\varphi}(x, e) \in \mathbb{R}^b$), and choose σ and λ by cross-validation (CV) with grid search as follows. First, the samples $\mathcal{Z} = \{(x_i, \widehat{e}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$ of (approximately) the same size (we set $K = 2$ in experiments). Then, an estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}$ is obtained using $\mathcal{Z}\backslash\mathcal{Z}_k$ (i.e., without $\mathcal{Z}_k$), and the approximation error for the hold-out samples $\mathcal{Z}_k$ is computed as

$$J_{\mathcal{Z}_k}^{(K\text{-CV})} = \frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}^\top \widehat{\boldsymbol{H}}_{\mathcal{Z}_k} \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k} - \widehat{\boldsymbol{h}}_{\mathcal{Z}_k}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}, \tag{8}$$

where, for $\mathcal{Z}_k = \{(x_i^{(k)}, \widehat{e}_i^{(k)}\}_{i=i}^{n_k}$,

$$\widehat{\boldsymbol{H}}_{\mathcal{Z}_k} = \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=1}^{n_k} \boldsymbol{\varphi}(x_i^{(k)}, \widehat{e}_j^{(k)})\boldsymbol{\varphi}(x_i^{(k)}, \widehat{e}_j^{(k)})^\top,$$

$$\widehat{\boldsymbol{h}}_{\mathcal{Z}_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{\varphi}(x_i^{(k)}, \widehat{e}_i^{(k)}).$$

This procedure is repeated for $k = 1, \ldots, K$, and its average $J^{(K\text{-CV})}$ is calculated as

$$J^{(K\text{-CV})} = \frac{1}{K}\sum_{k=1}^{K} J_{\mathcal{Z}_k}^{(K\text{-CV})}. \tag{9}$$

We compute $J^{(K\text{-CV})}$ for all model candidates (the kernel width σ and the regularization parameter λ in the current setup), and choose the density-ratio model that minimizes $J^{(K\text{-CV})}$. Note that $J^{(K\text{-CV})}$ is an almost unbiased estimator of the objective function (5), where the almost-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting (Schölkopf and Smola, 2002).

The LSMI algorithm is summarized in Figure 1.

2.3 Least-Squares Independence Regression

Given the SMI estimator (7), our next task is to learn the parameter $\boldsymbol{\beta}$ in the regression model (1) as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left[\widehat{\mathrm{SMI}}(X, \widehat{E}) + \frac{\gamma}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right].$$

We call this method *least-squares independence regression (LSIR)*.

For regression parameter learning, we simply employ a gradient descent method:

$$\boldsymbol{\beta} \longleftarrow \boldsymbol{\beta} - \eta\left(\frac{\partial\widehat{\mathrm{SMI}}(X, \widehat{E})}{\partial\boldsymbol{\beta}} + \gamma\boldsymbol{\beta}\right), \tag{10}$$

Input: Paired samples $\mathcal{Z} = \{(x_i, e_i)\}_{i=1}^n$,
Gaussian widths $\{\sigma_r\}_{r=1}^R$,
regularization parameters $\{\lambda_s\}_{s=1}^S$,
the number of basis functions b

Output: SMI estimator $\widehat{\mathrm{SMI}}(X, E)$

Split $\mathcal{Z}$ into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$
For each Gaussian width candidate σ_r
  For each regularization parameter candidate λ_s
    For each split $k = 1, \ldots, K$
      Compute $\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}$ by Eq.(6) with $\mathcal{Z} \backslash \mathcal{Z}_k$, σ_r and λ_s
      Compute hold-out error $J_{\mathcal{Z}_k}^{(K\text{-CV})}(r, s)$ by Eq.(8)
    End
    Compute average hold-out error $J^{(K\text{-CV})}(r, s)$ by Eq.(9)
  End
End
$(\widehat{r}, \widehat{s}) \leftarrow \mathrm{argmin}_{(r,s)} J^{(K\text{-CV})}(r, s)$
Compute $\widehat{\boldsymbol{\alpha}}$ by Eq.(6) with $\mathcal{Z}$, $\sigma_{\widehat{r}}$ and $\lambda_{\widehat{s}}$
Compute SMI estimator $\widehat{\mathrm{SMI}}(X, E)$ by Eq.(7)

Figure 1: Pseudo code of LSMI with CV.

where η is a step size which may be chosen in practice by some approximate line search method such as *Armijo's rule* (Patriksson, 1999).

The partial derivative of $\widehat{\mathrm{SMI}}(X, \widehat{E})$ with respect to $\boldsymbol{\beta}$ can be approximately expressed as

$$\frac{\partial \widehat{\mathrm{SMI}}(X, \widehat{E})}{\partial \boldsymbol{\beta}} \approx \sum_{l=1}^{b} \widehat{\alpha}_l \frac{\partial \widehat{h}_l}{\partial \boldsymbol{\beta}} - \frac{1}{2} \sum_{l,l'=1}^{b} \widehat{\alpha}_l \widehat{\alpha}_{l'} \frac{\partial \widehat{H}_{l,l'}}{\partial \boldsymbol{\beta}},$$

where

$$\begin{aligned}
\frac{\partial \widehat{h}_l}{\partial \boldsymbol{\beta}} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \varphi_l(x_i, \widehat{e}_i)}{\partial \boldsymbol{\beta}}, \\
\frac{\partial \widehat{H}_{l,l'}}{\partial \boldsymbol{\beta}} &= \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\frac{\partial \varphi_l(x_i, \widehat{e}_j)}{\partial \boldsymbol{\beta}} \varphi_{l'}(x_j, \widehat{e}_i) + \varphi_l(x_i, \widehat{e}_j) \frac{\partial \varphi_{l'}(x_j, \widehat{e}_i)}{\partial \boldsymbol{\beta}} \right), \\
\frac{\partial \varphi_l(x, \widehat{e})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2} \varphi_l(x, \widehat{e})(\widehat{e} - \widehat{v}_l) \boldsymbol{\psi}(x).
\end{aligned}$$

In the above derivation, we ignored the dependence of $\widehat{\alpha}_l$ on $\boldsymbol{\beta}$. It is possible to exactly compute the derivative in principle, but we use this approximated expression since it is computationally efficient and the approximation performs well in experiments.

We assumed that the mean of the noise E is zero. Taking into account this, we modify the final regressor as

$$\widehat{f}(x) = f_{\widehat{\boldsymbol{\beta}}}(x) + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{\widehat{\boldsymbol{\beta}}}(x_i)\right).$$

2.4 Model Selection in LSIR

LSIR contains three tuning parameters—the number of basis functions m, the kernel width τ, and the regularization parameter γ. In our experiments, we fix $m = \min(200, n)$, and choose τ and γ by CV with grid search as follows. First, the samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{n}$ are divided into T disjoint subsets $\{\mathcal{S}_t\}_{t=1}^{T}$ of (approximately) the same size (we set $T = 2$ in experiments), where $\mathcal{S}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$ and n_t is the number of samples in the subset $\mathcal{S}_t$. Then, an estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_t}$ is obtained using $\mathcal{S}\backslash\mathcal{S}_t$ (i.e., without $\mathcal{S}_t$), and the noise for the hold-out samples $\mathcal{S}_t$ is computed as

$$\widehat{e}_{t,i} = y_{t,i} - \widehat{f}_{\mathcal{S}_t}(x_{t,i}),\ i = 1, \ldots, n_t,$$

where $\widehat{f}_{\mathcal{S}_t}(x)$ is the estimated regressor by LSIR.

Let $\mathcal{Z}_t = \{(x_{t,i}, \widehat{e}_{t,i})\}_{i=1}^{n_t}$ be the hold-out samples of inputs and residuals. Then the independence score for the hold-out samples $\mathcal{Z}_t$ is given as

$$I_{\mathcal{Z}_t}^{(T\text{-CV})} = \widehat{\boldsymbol{h}}_{\mathcal{Z}_t}^{\top}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t} - \frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t}^{\top}\widehat{\boldsymbol{H}}_{\mathcal{Z}_t}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t} - \frac{1}{2}, \tag{11}$$

where $\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t}$ is the estimated model parameter by LSMI. Note that, the kernel width σ and the regularization parameter λ for LSMI are chosen by CV using the hold-out samples $\mathcal{Z}_t$.

This procedure is repeated for $t = 1, \ldots, T$, and its average $I^{(T\text{-CV})}$ is computed as

$$I^{(T\text{-CV})} = \frac{1}{T}\sum_{t=1}^{T}\widehat{I}_{\mathcal{Z}_t}^{(T\text{-CV})}. \tag{12}$$

We compute $I^{(T\text{-CV})}$ for all model candidates (the kernel width τ and the regularization parameter γ in the current setup), and choose the LSIR model that minimizes $I^{(T\text{-CV})}$.

The LSIR algorithm is summarized in Figure 2. A MATLAB® implementation of LSIR is available from

'`http://sugiyama-www.cs.titech.ac.jp/~yamada/lsir.html`'.

2.5 Causal Direction Inference by LSIR

In the previous section, we gave a dependence minimizing regression method, LSIR, that is equipped with CV for model selection. In this section, following Hoyer *et al.* (2009), we explain how LSIR can be used for causal direction inference.

Input: Paired samples $\{(x_i, y_i)\}_{i=1}^n$,
Gaussian width τ,
regularization parameter γ,
the number of basis functions m
Output: LSIR parameter $\widehat{\boldsymbol{\beta}}$

Initialize $\boldsymbol{\beta}$ by kernel regression with τ and γ (Schölkopf and Smola, 2002)
Computing a residual $\widehat{e}_i$ with current $\boldsymbol{\beta}$
While convergence
  Estimate $\widehat{\mathrm{SMI}}(x, e)$ by LSMI with $\{(x, \widehat{e}_i)\}_{i=1}^n$
  Update $\boldsymbol{\beta}$ by Eq.(10) with τ and γ
  Compute a residual $\widehat{e}_i$ with current $\boldsymbol{\beta}$
  If $\boldsymbol{\beta}$ has converged
    Return the current $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$
  End
End

Figure 2: Pseudo code of LSIR.

Input: Paired samples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$,
Gaussian widths $\{\tau_p\}_{p=1}^P$,
regularization parameters $\{\gamma_q\}_{q=1}^Q$,
the number of basis functions m
Output: LSIR parameter $\widehat{\boldsymbol{\beta}}$

Split $\mathcal{S}$ into T disjoint subsets $\{\mathcal{S}_t\}_{t=1}^T$, $\mathcal{S}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$
For each Gaussian width candidate τ_p
  For each regularization parameter candidate γ_q
    For each split $t = 1, \ldots, T$
      Compute $\widehat{\boldsymbol{\beta}}_{\mathcal{S}_t}$ by LSIR with $\mathcal{S}\backslash\mathcal{S}_k$, τ_p and γ_q
      Compute a residual $\widehat{e}_{t,i}$ and make a set $\mathcal{Z}_t = \{(x_{t,i}, \widehat{e}_{t,i})\}_{i=1}^{n_t}$
      Compute hold-out independence criterion $I_{\mathcal{Z}_k}^{(T\text{-CV})}(r, s)$ by Eq.(11)
    End
    Compute average hold-out independence criterion $I^{(T\text{-CV})}(p, q)$ by Eq.(12)
  End
End
$(\widehat{p}, \widehat{q}) \leftarrow \mathrm{argmin}_{(p,q)} I^{(T\text{-CV})}(p, q)$
Compute $\widehat{\boldsymbol{\beta}}$ by LSIR with $\mathcal{S}$, $\tau_{\widehat{p}}$, and $\gamma_{\widehat{q}}$

Figure 3: Pseudo code of LSIR with CV.

Our final goal is, given i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to determine whether X causes Y or vice versa. To this end, we test whether the causal model $Y = f_Y(X) + E_Y$ or the alternative model $X = f_X(Y) + E_X$ fits the data well, where the goodness of fit is measured by independence between inputs and residuals (i.e., estimated noise). Independence of inputs and residuals may be decided in practice by the *permutation test* (Efron and Tibshirani, 1993).

More specifically, we first run LSIR for $\{(x_i, y_i)\}_{i=1}^n$ as usual, and obtain a regression function $\widehat{f}$. This procedure also provides an SMI estimate for $\{(x_i, \widehat{e}_i) \mid \widehat{e}_i = y_i - \widehat{f}(x_i)\}_{i=1}^n$. Next, we randomly permute the pairs of input and residual $\{(x_i, \widehat{e}_i)\}_{i=1}^n$ as $\{(x_i, \widehat{e}_{\kappa(i)})\}_{i=1}^n$, where $\kappa(\cdot)$ is a randomly generated permutation function. Note that the permuted pairs of samples are independent of each other since the random permutation breaks the dependency between X and $\widehat{E}$ (if it exists). Then we compute SMI estimates for the permuted data $\{(x_i, \widehat{e}_{\kappa(i)})\}_{i=1}^n$ by LSMI. This random permutation process is repeated many times (in experiments, the number of repetitions is set at 1000), and the distribution of SMI estimates under the null-hypothesis (i.e., independence) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of the SMI estimate computed from the original input-residual data over the distribution of SMI estimates for randomly permuted data.

Although not every causal mechanism can be described by an additive noise model, we assume that it is unlikely that the causal structure $Y \to X$ induces an additive noise model from X to Y, except for simple distributions like bivariate Gaussians. Janzing and Steudel (2010) support this assumption by an algorithmic information theory approach. In order to decide the causal direction based on the assumption, we first compute the p-values $p_{X\to Y}$ and $p_{X\leftarrow Y}$ for both directions $X \to Y$ (i.e., X causes Y) and $X \leftarrow Y$ (i.e., Y causes X). Then, for a given significance level δ_1 and δ_2 ($\delta_2 \geq \delta_1$), we determine the causal direction as follows:

- If $p_{X\to Y} > \delta_2$ and $p_{X\leftarrow Y} \leq \delta_1$, the causal model $X \to Y$ is chosen.
- If $p_{X\leftarrow Y} > \delta_2$ and $p_{X\to Y} \leq \delta_1$, the causal model $X \leftarrow Y$ is selected.
- If $p_{X\to Y}, p_{X\leftarrow Y} \leq \delta_1$, the causal relation is not an additive noise model.
- If $p_{X\to Y}, p_{X\leftarrow Y} > \delta_1$, the joint distribution seems to be close to one of the few exceptions that admit additive noise models in both directions.

In our preliminary experiments, we empirically observed that SMI estimates obtained by LSIR tend to be affected by the basis function choice in LSIR. To mitigate this problem, we run LSIR and compute an SMI estimate 5 times by randomly changing basis functions. Then the regression function that gives the smallest SMI estimate among 5 repetitions is selected and the permutation test is performed for that regression function.

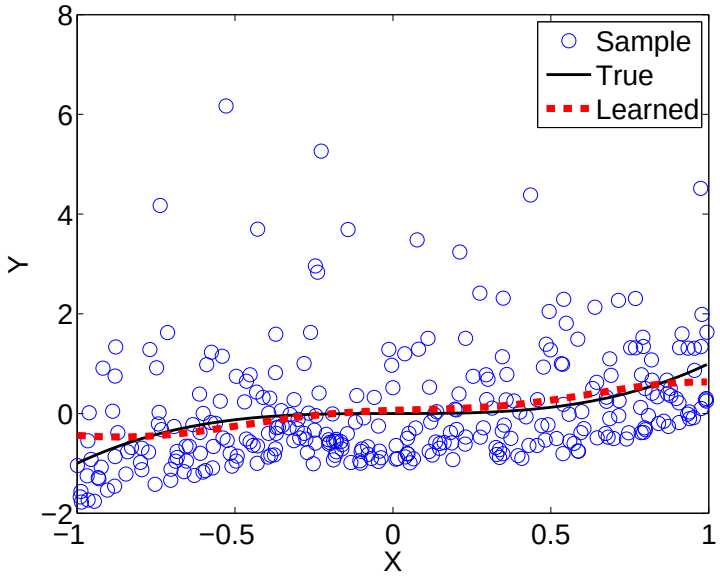


Figure 4: Illustrative example. The solid line denotes the true function, the circles denote samples, and the dashed line denotes the regressor obtained by LSIR.

2.6 Illustrative Examples

Let us consider the following additive noise model:

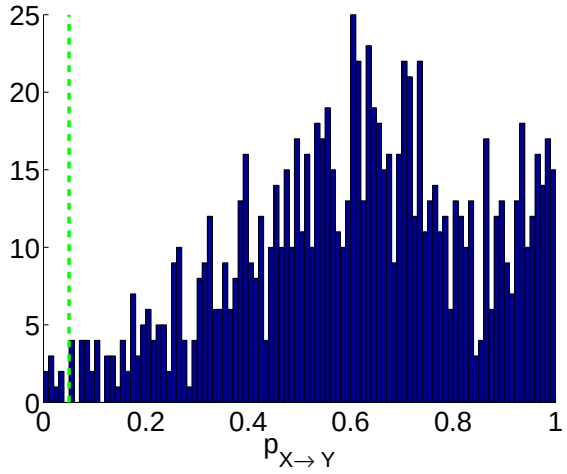
$$Y = X^3 + E,$$

where X is subject to the uniform distribution on $(-1, 1)$ and E is subject to the exponential distribution with rate parameter 1 (and its mean is adjusted to be zero). We drew 300 paired samples of X and Y following the above generative model (see Figure 4), where the ground truth is that X and E are independent of each other. Thus, the null-hypothesis should be accepted (i.e., the p-values should be large).

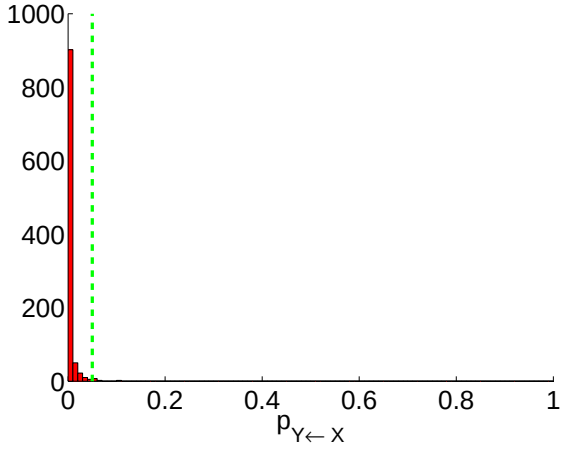
Figure 4 depicts the regressor obtained by LSIR, giving a good approximation to the true function. We repeated the experiment 1000 times with the random seed changed. For the significance level 5%, LSIR successfully accepted the null-hypothesis 992 times out of 1000 runs.

As Mooij *et al.* (2009) pointed out, beyond the fact that the p-values frequently exceed the pre-specified significance level, it is important to have a wide margin beyond the significance level in order to cope with, e.g., multiple variable cases. Figure 5(a) depicts the histogram of $p_{X \to Y}$ obtained by LSIR over 1000 runs. The plot shows that LSIR tends to produce much larger p-values than the significance level; the mean and standard deviation of the p-values over 1000 runs are 0.6114 and 0.2327, respectively.

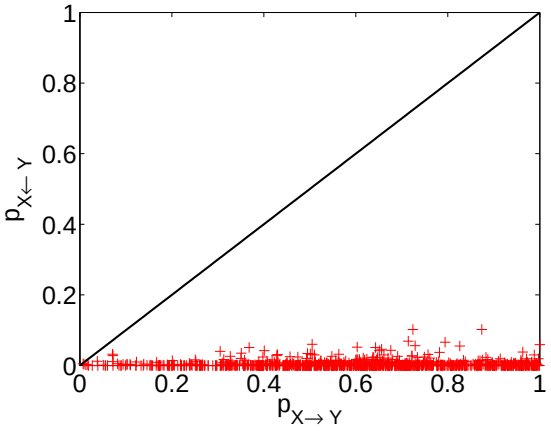
Next, we consider the backward case where the roles of X and Y are swapped. In this case, the ground truth is that the input and the residual are dependent (see Figure 4). Therefore, the null-hypothesis should be rejected (i.e., the p-values should be small). Figure 5(b) shows the histogram of $p_{X \leftarrow Y}$ obtained by LSIR over 1000 runs. LSIR rejected



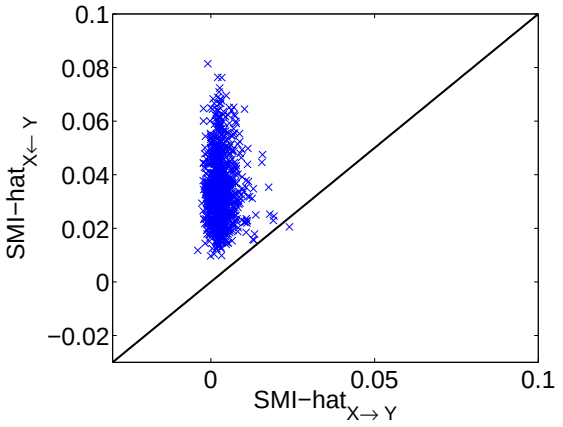
small(a) Histogram of $p_{X\to Y}$ obtained by LSIR over 1000 runs. The ground truth is to accept the null-hypothesis (thus the p-values should be large).



small(b) Histograms of $p_{X\leftarrow Y}$ obtained by LSIR over 1000 runs. The ground truth is to reject the null-hypothesis (thus the p-values should be small).



small(c) Comparison of p-values for both directions ($p_{X\to Y}$ vs. $p_{X\leftarrow Y}$). Points below the diagonal line indicate successful trials.



small(d) Comparison of values of independence measures for both directions ($\widehat{\mathrm{SMI}}_{X\to Y}$ vs. $\widehat{\mathrm{SMI}}_{X\leftarrow Y}$). Points above the diagonal line indicate successful trials.

Figure 5: LSIR performance statistics in illustrative example.

the null-hypothesis 989 times out of 1000 runs; the mean and standard deviation of the p-values over 1000 runs are 0.0035 and 0.0094, respectively.

Figure 5(c) depicts the p-values for both directions in a trial-wise manner. The graph shows that LSIR perfectly estimates the correct causal direction (i.e., $p_{X\to Y} > p_{X\leftarrow Y}$), and the *margin* between $p_{X\to Y}$ and $p_{X\leftarrow Y}$ seems to be clear (i.e., most of the points are clearly below the diagonal line). This illustrates the usefulness of LSIR in causal direction inference.

Finally, we investigate the values of independence measure $\widehat{\mathrm{SMI}}$, which are plotted in Figure 5(d) again in a trial-wise manner. The graph implies that the values of $\widehat{\mathrm{SMI}}$ may be simply used for determining the causal direction, instead of the p-values. Indeed, the correct causal direction (i.e., $\widehat{\mathrm{SMI}}_{X\to Y} < \widehat{\mathrm{SMI}}_{X\leftarrow Y}$) can be found 999 times out of 1000 trials by this simplified method. This would be a practically useful heuristic since we can avoid performing the computationally intensive permutation test.

3 Existing Method: HSIC Regression

In this section, we review the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005) and *HSIC regression* (HSICR) (Mooij *et al.*, 2009).

3.1 Hilbert-Schmidt Independence Criterion (HSIC)

The *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005) is a state-of-the-art measure of statistical independence based on characteristic functions (see also Feuerverger, 1993; Kankainen, 1995). Here, we review the definition of HSIC and explain its basic properties.

Let $\mathcal{F}$ be a *reproducing kernel Hilbert space* (RKHS) with reproducing kernel $K(x,x')$ (Aronszajn, 1950), and $\mathcal{G}$ be another RKHS with reproducing kernel $L(e,e')$. Let C be a *cross-covariance operator* from $\mathcal{G}$ to $\mathcal{F}$, i.e., for all $f\in\mathcal{F}$ and $g\in\mathcal{G}$,

$$\langle f, Cg\rangle_{\mathcal{F}} = \iint \left(\left[f(x) - \int f(x)p(x)\mathrm{d}x \right] \left[g(e) - \int g(e)p(e)\mathrm{d}e \right] \right) p(x,e)\mathrm{d}x\mathrm{d}e,$$

where $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ denotes the inner product in $\mathcal{F}$. Thus, C can be expressed as

$$C = \iint \left(\left[K(\cdot,x) - \int K(\cdot,x)p(x)\mathrm{d}x \right] \otimes \left[L(\cdot,e) - \int L(\cdot,e)p(e)\mathrm{d}e \right] \right) p(x,e)\mathrm{d}x\mathrm{d}e,$$

where '$\otimes$' denotes the *tensor product*, and we used the reproducing properties:

$$f(x) = \langle f, K(\cdot,x)\rangle_{\mathcal{F}} \quad \text{and} \quad g(e) = \langle g, L(\cdot,e)\rangle_{\mathcal{G}}.$$

The cross-covariance operator is a generalization of the *cross-covariance matrix* between random vectors. When $\mathcal{F}$ and $\mathcal{G}$ are *universal RKHSs* (Steinwart, 2001) defined on compact domains $\mathcal{X}$ and $\mathcal{E}$, respectively, the largest singular value of C is zero if and only if x and e are independent. Gaussian RKHSs are examples of the universal RKHS.

HSIC is defined as the squared *Hilbert-Schmidt norm* (the sum of the squared singular

values) of the cross-covariance operator C:

$$\begin{aligned}\mathrm{HSIC} :=& \iiiint K(x,x')L(e,e')p(x,e)p(x',e')\mathrm{d}x\mathrm{d}e\mathrm{d}x\mathrm{d}e' \\ &+ \left[\iint K(x,x')p(x)p(x')\mathrm{d}x\mathrm{d}x'\right]\left[\iint L(e,e')p(e)p(e')\mathrm{d}e\mathrm{d}e'\right] \\ &- 2\iint\left[\int K(x,x')p(x')\mathrm{d}x'\right]\left[\int L(e,e')p(e')\mathrm{d}e'\right]p(x,e)\mathrm{d}x\mathrm{d}e.\end{aligned}$$

The above expression allows one to immediately obtain an empirical estimator—with i.i.d. samples $\mathcal{Z} = \{(x_k, e_k)\}_{k=1}^n$ following $p(x,e)$, a consistent estimator of HSIC is given as

$$\begin{aligned}\widehat{\mathrm{HSIC}}(X,E) :=& \frac{1}{n^2}\sum_{i,i'=1}^n K(x_i,x_{i'})L(e_i,e_{i'}) + \frac{1}{n^4}\sum_{i,i',j,j'=1}^n K(x_i,x_{i'})L(e_j,e_{j'}) \\ &- \frac{2}{n^3}\sum_{i,j,k=1}^n K(x_i,x_k)L(e_j,e_k) \\ =& \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K\Gamma L\Gamma}),\end{aligned} \tag{13}$$

where

$$K_{i,i'} = K(x_i,x_{i'}), \quad L_{i,i'} = L(e_i,e_{i'}), \quad \text{and} \quad \boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$

$\boldsymbol{I}_n$ denotes the n-dimensional identity matrix, and $\mathbf{1}_n$ denotes the n-dimensional vector with all ones.

$\widehat{\mathrm{HSIC}}$ depends on the choice of the universal RKHSs $\mathcal{F}$ and $\mathcal{G}$. In the original HSIC paper (Gretton *et al.*, 2005), the Gaussian RKHS with width set at the median distance between sample points was used, which is a popular heuristic in the kernel method community (Schölkopf and Smola, 2002). However, to the best of our knowledge, there is no strong theoretical justification for this heuristic. On the other hand, the LSMI method is equipped with cross-validation, and thus all the tuning parameters such as the Gaussian width and the regularization parameter can be optimized in an objective and systematic way. This is an advantage of LSMI over HSIC.

3.2 HSIC Regression

In *HSIC regression* (HSICR) (Mooij *et al.*, 2009), the following linear model is employed:

$$f_{\boldsymbol{\theta}}(x) = \sum_{l=1}^n \theta_l\phi_l(x) = \boldsymbol{\theta}^\top\boldsymbol{\phi}(x), \tag{14}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ are regression parameters and $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_n(x))^\top$ are basis functions. Mooij *et al.* (2009) proposed to use the Gaussian basis function:

$$\phi_l(x) = \exp\left(-\frac{(x - x_l)^2}{2\rho^2}\right),$$

where the kernel width ρ is set at the median distance between sample points:

$$\rho = 2^{-1/2}\mathrm{median}(\{\|x_i - x_j\|\}_{i,j=1}^n).$$

Given the HSIC estimator (13), the parameter $\boldsymbol{\theta}$ in the regression model (14) is obtained by

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\widehat{\mathrm{HSIC}}(X, Y - f_{\boldsymbol{\theta}}(X)) + \frac{\xi}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{15}$$

where $\xi \geq 0$ is the regularization parameter to avoid overfitting. This optimization problem can be efficiently solved by using the *L-BFGS quasi-Newton method* (Liu and Nocedal, 1989) or gradient descent. Then, the final regressor is given as

$$\widehat{f}(x) = f_{\widehat{\boldsymbol{\theta}}}(x) + \frac{1}{n}\sum_{i=1}^n \left(y_i - f_{\widehat{\boldsymbol{\theta}}}(x_i)\right).$$

In the HSIC estimator, the Gaussian kernels,

$$K(x, x') = \exp\left(-\frac{(x - x')^2}{2\sigma_{\mathrm{x}}^2}\right) \quad \text{and} \quad L(e, e') = \exp\left(-\frac{(e - e')^2}{2\sigma_{\mathrm{e}}^2}\right),$$

are used and their kernel widths are fixed at the median distance between sample points during the optimization (15):

$$\sigma_{\mathrm{x}} = 2^{-1/2}\mathrm{median}(\{\|x_i - x_j\|\}_{i,j=1}^n),$$
$$\sigma_{\mathrm{e}} = 2^{-1/2}\mathrm{median}(\{\|\widehat{e}_i - \widehat{e}_j\|\}_{i,j=1}^n),$$

where $\{\widehat{e}_i\}_{i=1}^n$ are initial rough estimates of the residuals. This implies that, if the initial choices of σ_{x} and σ_{e} are poor, the overall performance of HSICR will be degraded. On the other hand, the LSIR method is equipped with cross-validation, and thus all the tuning parameters can be optimized in an objective and systematic way. This is a significant advantage of LSIR over HSICR.

4 Experiments

In this section, we evaluate the performance of LSIR using benchmark datasets and real-world gene expression data.

4.1 Benchmark Datasets

Here, we evaluate the performance of LSIR on the '*Cause-Effect Pairs*' task[1]. The task contains 80 datasets[2], each has two statistically dependent random variables possessing inherent causal relationship. The goal is to identify the causal direction from observational data. Since these datasets consist of real-world samples, our modeling assumption may be only approximately satisfied. Thus, identifying causal directions in these datasets would be highly challenging. The 'pair0001' to 'pair0006' datasets are illustrated in Figure 6.

Table 1 shows the results for the benchmark data with different threshold values δ_1 and δ_2. As can be observed, LSIR compares favorably with HSICR. For example, when $\delta_1 = 0.05$ and $\delta_2 = 0.10$, LSIR found the correct causal direction for 20 out of 80 cases and the incorrect causal direction for 6 out of 80 cases, while HSICR found the correct causal direction for 14 out of 80 cases and the incorrect causal direction for 15 out of 80 cases. Also, the correct identification rate (the number of correct causal directions detected/the number of all causal directions detected) of LSIR and HSICR are 0.77 and 0.48, respectively. We note that the cases with $p_{X\to Y}, p_{Y\to X} < \delta_1$ and $p_{X\to Y}, p_{Y\to X} \geq \delta_1$ happened frequently both for LSIR and HSICR. Thus, although many cases were not identifiable, LSIR still compares favorably with HSICR.

Moreover, we compare LSIR with HSICR on the binary causal direction detection setting[3] (see Mooij *et al.* (2009)). In this experiment, we compare the p-values and choose the direction with a larger p-value as the causal direction. The p-values for each dataset and each direction are summarized in Figures 7(a) and 7(b), where the horizontal axis denotes the dataset index. When the correct causal direction can be correctly identified, we indicate the data by '$*$'. The results show that LSIR can successfully find the correct causal direction for 49 out of 80 cases, while HSICR gave the correct decision only for 31 out of 80 cases.

Figure 7(c) shows that merely comparing the values of $\widehat{\mathrm{SMI}}$ is actually sufficient to decide the correct causal direction in LSIR; using this heuristic, LSIR successfully identified the correct causal direction 54 out of 80 cases. Thus, this heuristic allows us to identify the causal direction in a computationally efficient way. On the other hand, as shown in Figure 7(d), HSICR gave the correct decision only for 36 out of 80 cases with this heuristic.

4.2 Gene Function Regulations

Finally, we apply our proposed LSIR method to real-world biological datasets, which contain known causal relationships about gene function regulations from transcription factors to gene expressions.

Causal prediction is biologically and medically important because it gives us a clue for disease-causing genes or drug-target genes. Transcription factors regulate expression

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

[2] There are 86 datasets in total, but since 'pair0052'–'pair0055' and 'pair0071' are a multivariate and 'pair0070' is categorical, we decided to exclude them from our experiments.

[3] `http://www.causality.inf.ethz.ch/cause-effect.php`

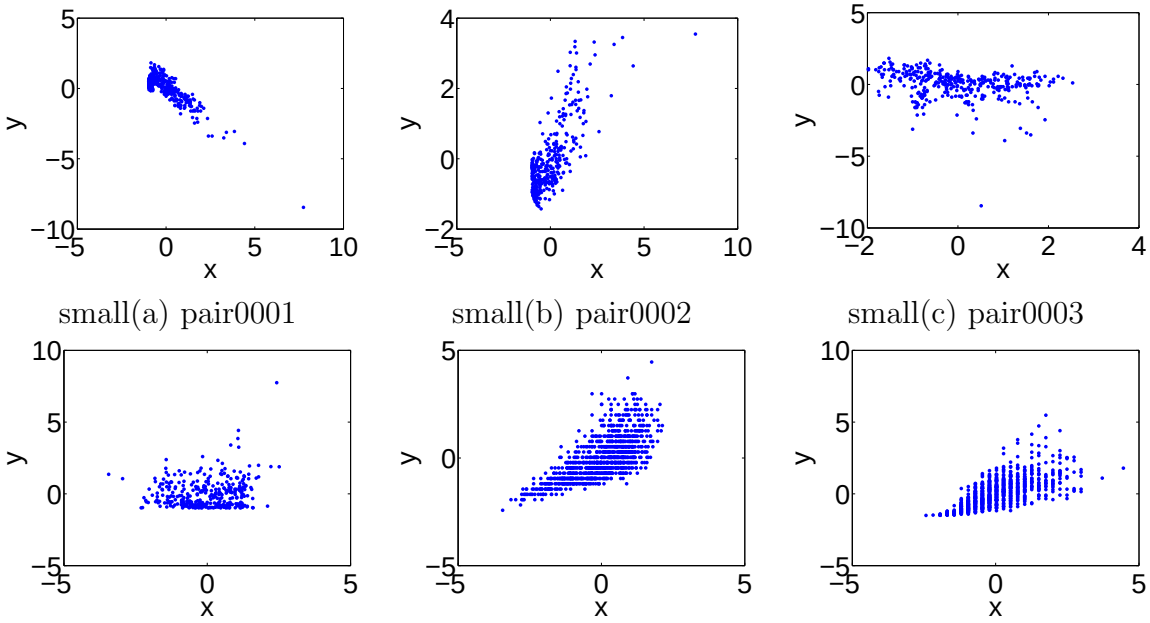


Figure 6: Datasets of the '*Cause-Effect Pairs*' task.

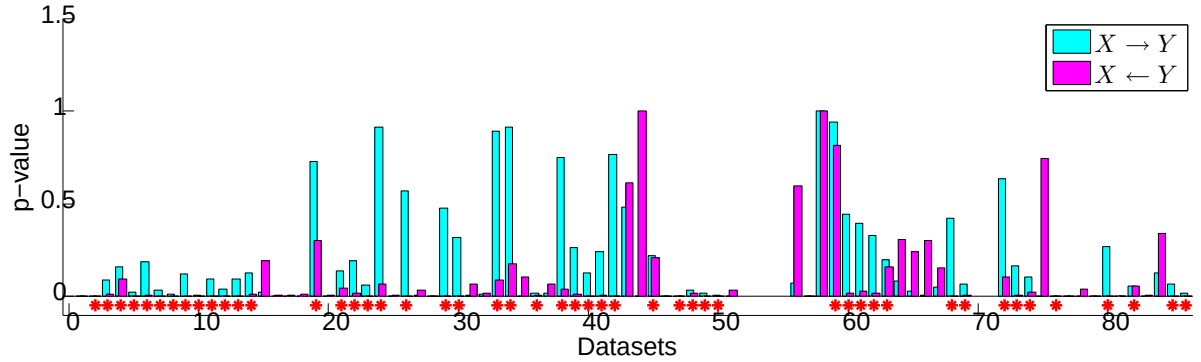
Table 1: Results for the '*Cause-Effect Pairs*' task. Each cell in the tables denotes 'the number of correct causal directions detected/the number of incorrect causal directions detected (the number of correct causal directions detected/the number of all causal directions detected)'.

(a) LSIR

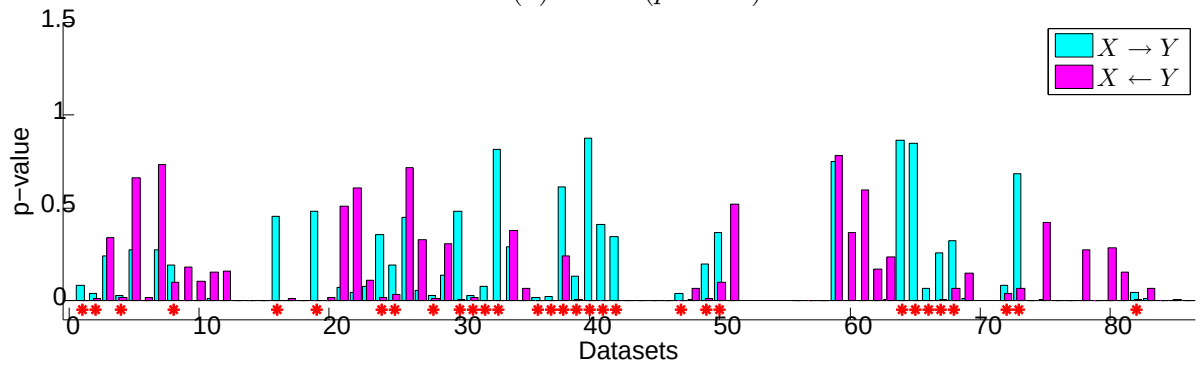
δ_1 \ δ_2	0.01	0.05	0.10	0.15	0.20
0.01	23/9 (0.72)	17/5 (0.77)	12/4 (0.75)	9/3 (0.75)	7/3 (0.70)
0.05	—	26/8 (0.77)	20/6 (0.77)	15/5 (0.75)	12/4 (0.75)
0.10	—	—	23/9 (0.72)	18/8 (0.69)	14/6 (0.70)
0.15	—	—	—	19/9 (0.68)	15/7 (0.68)
0.20	—	—	—	—	16/7 (0.70)

(b) HSICR

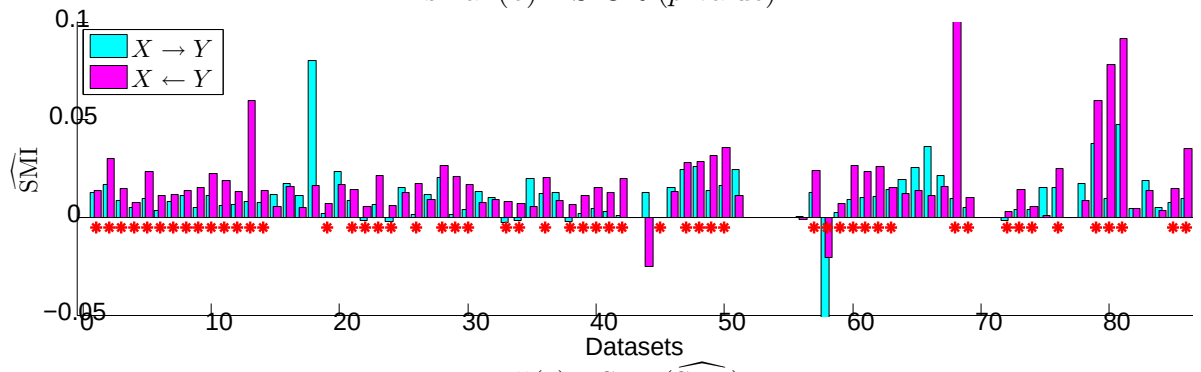
δ_1 \ δ_2	0.01	0.05	0.10	0.15	0.20
0.01	18/17 (0.51)	14/14 (0.50)	11/12 (0.48)	10/11 (0.48)	10/7 (0.59)
0.05	—	18/18 (0.50)	14/15 (0.48)	13/13 (0.50)	11/8 (0.58)
0.10	—	—	16/18 (0.47)	15/15 (0.50)	13/10 (0.57)
0.15	—	—	—	17/16 (0.52)	14/11 (0.56)
0.20	—	—	—	—	14/11 (0.56)



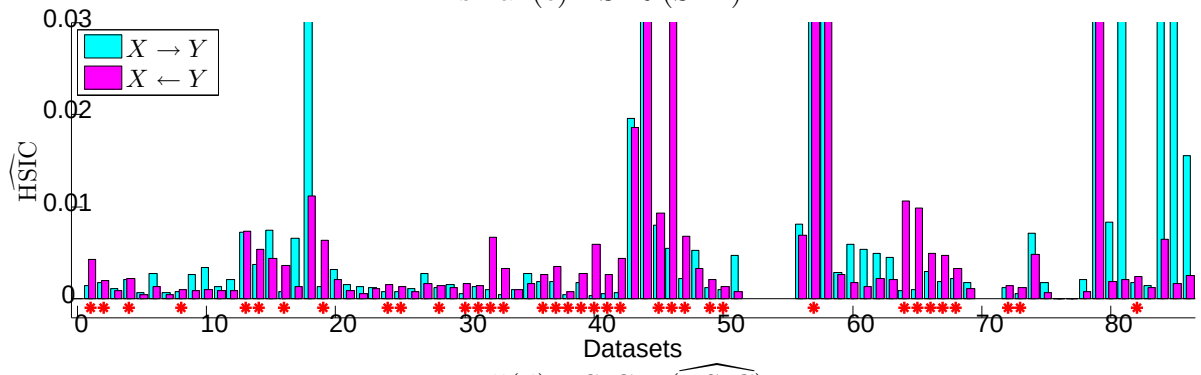
small(a) LSIR (p-value)



small(b) HSICR (p-value)



small(c) LSIR ($\widehat{\mathrm{SMI}}$)



small(d) HSICR ($\widehat{\mathrm{HSIC}}$)

Figure 7: Results for the '*Cause-Effect Pairs*' task. The horizontal axis denotes the dataset index. When the true causal direction can be correctly identified, we indicate the data by '$*$'.

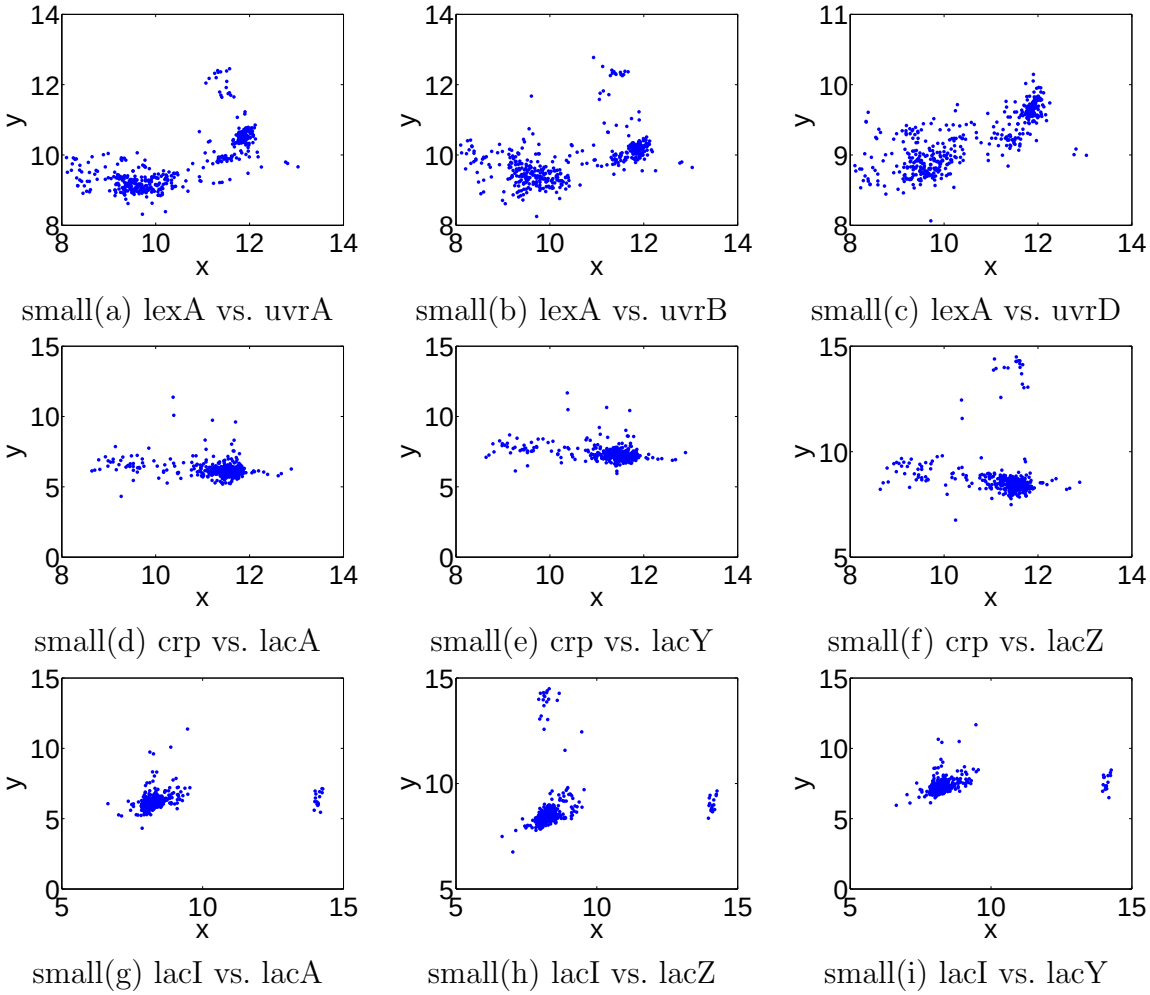


Figure 8: Datasets of the E. coli task (Faith *et al.*, 2007).

levels of their relating genes. In other words, when the expression level of transcription factor genes is high, genes regulated by the transcription factor become highly expressed or suppressed.

In this experiment, we select 9 well-known gene regulation relationships of *E. coli* (Faith *et al.*, 2007), where each data contains expression levels of the genes over 445 different environments (i.e., 445 samples, see Figure 8).

The experimental results are summarized in Table 2. In this experiment, we denote the estimated direction by '$\Rightarrow$' if $p_{X\to Y} > 0.05$ and $p_{Y\to X} < 10^{-3}$. If $p_{X\to Y} > p_{Y\to X}$, we denote the estimated direction by '$\to$'. As can be observed, LSIR can successfully detect 3 out of 9 cases, while HSICR can only detect 1 out of 9 cases. Moreover, in binary decision setting (i.e., comparison between p values), LSIR and HSICR successfully found the correct causal directions for 7 out of 9 cases and 6 out of 9 cases, respectively. In addition, all the correct causal directions can be efficiently chosen by LSIR and HSICR if the heuristic of comparing the values of $\widehat{\mathrm{SMI}}$ is used. Thus, the proposed method and

Table 2: Results for the 'E. coli' task. If $p_{X\to Y} > 0.05$ and $p_{Y\to X} < 10^{-3}$, we denote the estimated direction by $\Rightarrow$. If $p_{X\to Y} > p_{Y\to X}$, we denote the estimated direction by $\to$. When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?'). Estimated directions in the brackets are determined based on comparing the values of $\widehat{\mathrm{SMI}}$ or $\widehat{\mathrm{HSIC}}$.

(a) LSIR

Dataset		p-values		$\widehat{\mathrm{SMI}}$		Direction		
X	Y	$X\to Y$	$X\leftarrow Y$	$X\to Y$	$X\leftarrow Y$	Estimated		Truth
lexA	uvrA	$< 10^{-3}$	$< 10^{-3}$	0.0177	0.0255	?	$(\to)$	$\to$
lexA	uvrB	$< 10^{-3}$	$< 10^{-3}$	0.0172	0.0356	?	$(\to)$	$\to$
lexA	uvrD	0.043	$< 10^{-3}$	0.0075	0.0227	$\to$	$(\to)$	$\to$
crp	lacA	0.143	$< 10^{-3}$	-0.0004	0.0399	$\Rightarrow$	$(\to)$	$\to$
crp	lacY	0.003	$< 10^{-3}$	0.0118	0.0247	$\to$	$(\to)$	$\to$
crp	lacZ	0.001	$< 10^{-3}$	0.0122	0.0307	$\to$	$(\to)$	$\to$
lacI	lacA	0.787	$< 10^{-3}$	-0.0076	0.0184	$\Rightarrow$	$(\to)$	$\to$
lacI	lacZ	0.002	$< 10^{-3}$	0.0096	0.0141	$\to$	$(\to)$	$\to$
lacI	lacY	0.746	$< 10^{-3}$	-0.0082	0.0217	$\Rightarrow$	$(\to)$	$\to$

(b) HSICR

Dataset		p-values		$\widehat{\mathrm{HSIC}}$		Direction		
X	Y	$X\to Y$	$X\leftarrow Y$	$X\to Y$	$X\leftarrow Y$	Estimated		Truth
lexA	uvrA	0.005	$< 10^{-3}$	0.0013	0.0037	$\to$	$(\to)$	$\to$
lexA	uvrB	$< 10^{-3}$	$< 10^{-3}$	0.0026	0.0037	?	$(\to)$	$\to$
lexA	uvrD	$< 10^{-3}$	$< 10^{-3}$	0.0020	0.0041	?	$(\to)$	$\to$
crp	lacA	0.017	$< 10^{-3}$	0.0013	0.0036	$\to$	$(\to)$	$\to$
crp	lacY	0.002	$< 10^{-3}$	0.0018	0.0051	$\to$	$(\to)$	$\to$
crp	lacZ	0.008	$< 10^{-3}$	0.0013	0.0054	$\to$	$(\to)$	$\to$
lacI	lacA	0.031	$< 10^{-3}$	0.0012	0.0043	$\to$	$(\to)$	$\to$
lacI	lacZ	$< 10^{-3}$	$< 10^{-3}$	0.0019	0.0020	?	$(\to)$	$\to$
lacI	lacY	0.052	$< 10^{-3}$	0.0011	0.0027	$\Rightarrow$	$(\to)$	$\to$

HSICR are comparably useful for this task.

5 Conclusions

In this paper, we proposed a new method of dependence minimization regression called *least-squares independence regression* (LSIR). LSIR adopts the *squared-loss mutual information* as an independence measure, and it is estimated by the method of *least-squares mutual information* (LSMI). Since LSMI provides an analytic-form solution, we can explicitly compute the gradient of the LSMI estimator with respect to regression parameters.

A notable advantage of the proposed LSIR method over the state-of-the-art method of

dependence minimization regression (Mooij *et al.*, 2009) is that LSIR is equipped with a natural cross-validation procedure, allowing us to objectively optimize tuning parameters such as the kernel width and the regularization parameter in a data-dependent fashion.

We experimentally showed that LSIR is promising in real-world causal direction inference. We note that the use of LSMI instead of HSIC does not necessarily provide performance improvement of causal direction inference; indeed, experimental performances of LSMI and HSIC were on par if fixed Gaussian kernel widths are used. This implies that the performance improvement of the proposed method was brought by data-dependent optimization of kernel widths via cross-validation.

In this paper, we solely focused on the additive noise model, where noise is independent of inputs. When this modeling assumption is violated, LSIR (as well as HSICR) may not perform well. In such a case, employing a more elaborate model such as the post-nonlinear causal model would be useful (Zhang and Hyvärinen, 2009). We will extend LSIR to be applicable to such a general model in the future work.

Acknowledgments

The authors thank Dr. Joris Mooij for providing us the HSICR code. We also thank the editor and anonymous reviewers for their constructive feedback, which helped us to improve the manuscript. MY was supported by the JST PRESTO program, MS was supported by AOARD and KAKENHI 25700022, and JS was partially supported by KAKENHI 24680032, 24651227, and 25128704 and the support of Young Investigator Award of Human Frontier Science Program.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. the American Mathematical Society*, **68**, 337–404.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, **5**(1), e8.

Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review*, **61**(3), 419–433.

Fukumizu, K., Bach, F. R., and Jordan, M. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, **37**(4), 1871–1905.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *10th Annual Conference on Uncertainty in Artificial Intelligence (UAI1994)*, pages 235–243.

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

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Botton, editors, *Advances in Neural Information Processing Systems 21 (NIPS2008)*, pages 689–696, Cambridge, MA. MIT Press.

Janzing, D. and Steudel, B. (2010). Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Systems & Information Dynamics*, **17**(02), 189–212.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, **86**(3), 335–367.

Kankainen, A. (1995). *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*. Ph.D. thesis, University of Jyväskylä, Jyväskylä, Finland.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, **69**(066138).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

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

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *26th Annual International Conference on Machine Learning (ICML2009)*, pages 745–752, Montreal, Canada.

Patriksson, M. (1999). *Nonlinear Programming and Variational Inequality Problems*. Kluwer Academic, Dredrecht.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157–175.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ, USA.

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

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

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.

Suzuki, T. and Sugiyama, M. (2013). Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, **3**(25), 725–758.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. (2009). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, **10**(S52).

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, NY.

Yamada, M. and Sugiyama, M. (2010). Dependence minimizing regression with model selection for non-linear causal inference under non-gaussian noise. In *Proceedings of the twenty-fourth AAAI conference on artificial intelligence (AAAI2010)*, pages 643–648.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 647–655, Arlington, Virginia, United States. AUAI Press.