Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise*

Makoto Yamada† and **Masashi Sugiyama**†‡
†Department of Computer Science, Tokyo Institute of Technology
‡Japan Science and Technology Agency
{yamada@sg. sugi@}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 minimization of an estimator of the *squared-loss mutual information* between inputs and residuals. A notable advantage of LSIR over existing approaches 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 the state-of-the-art causal inference method.

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 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 the *Gaussian process* (GP) (Bishop 2006), and evaluate the independence between the inputs and the 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, Bach, and Jordan 2009), the lack of systematic model selection strategies is critical in causal inference. Setting the kernel width to the median distance between sample points seems to be 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 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.

*This work was supported by SCAT, AOARD, and the JST PRESTO program.

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 real-world datasets, we demonstrate the superiority of LSIR.

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).

Problem Formulation

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 $f(\cdot) : \mathbb{R} \to \mathbb{R}$ 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 $\widehat{f}(\cdot)$ 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 minimization regression, the regression parameter $\boldsymbol{\beta}$ may be learned 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 for avoiding overfitting.

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

$$\begin{aligned} I(X, \widehat{E}) &= \frac{1}{2} \iint \left(\frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} - 1 \right)^2 p(x)p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} \\ &= \frac{1}{2} \iint \frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} p(x, \widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} - \frac{1}{2}. \end{aligned}$$

$I(X, \widehat{E})$ is the *Pearson divergence* 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 independent. Note that ordinary *mutual information* corresponds to the *Kullback-Leibler divergence* 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 SMI since it allows us to obtain an analytic-form estimator as explained below.

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, Stögbauer, and Grassberger 2004). A key idea of LSMI is to directly estimate the *density ratio*:

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

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

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

$$w_{\boldsymbol{\alpha}}(x, \widehat{e}) = \sum_{l=1}^{b} \alpha_l \varphi_l(x, \widehat{e}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(x, \widehat{e}), \tag{2}$$

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 model $w_{\boldsymbol{\alpha}}(x, \widehat{e})$ is learned so that the following squared error $J_0(\boldsymbol{\alpha})$ is minimized:

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

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

$$J(\boldsymbol{\alpha}) = J_0(\boldsymbol{\alpha}) - C = \frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha}, \tag{3}$$

where

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

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

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

where a regularization term $\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$ 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:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}, \tag{4}$$

where $\boldsymbol{I}_b$ denotes the b-dimensional identity matrix. Consistency and the convergence rate of the above density ratio estimator has been theoretically studied in Kanamori, Suzuki, and Sugiyama (2009).

Given a density ratio estimator $\widehat{w} = w_{\widehat{\boldsymbol{\alpha}}}$, SMI can be simply approximated as

$$\widehat{I}(X, \widehat{E}) = \frac{1}{2n} \sum_{i=1}^{n} \widehat{w}(x_i, \widehat{e}_i) - \frac{1}{2} = \frac{1}{2} \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \frac{1}{2}. \tag{5}$$

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 $b = \min(50, n)$, and choose σ and λ by cross-validation (CV) with grid search as follows. First, the samples $\{z_i \mid z_i = (x_i, \widehat{e}_i)\}_{i=1}^n$ 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 $\{\mathcal{Z}_j\}_{j \neq k}$, and the approximation error for the hold-out samples $\mathcal{Z}_k$ is computed as

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

where, for $|\mathcal{Z}_k|$ being the number of samples in the subset $\mathcal{Z}_k$,

$$\widehat{\boldsymbol{H}}_{\mathcal{Z}_k} = \frac{1}{|\mathcal{Z}_k|^2} \sum_{x, \widehat{e} \in \mathcal{Z}_k} \boldsymbol{\varphi}(x, \widehat{e}) \boldsymbol{\varphi}(x, \widehat{e})^\top,$$

$$\widehat{\boldsymbol{h}}_{\mathcal{Z}_k} = \frac{1}{|\mathcal{Z}_k|} \sum_{(x, \widehat{e}) \in \mathcal{Z}_k} \boldsymbol{\varphi}(x, \widehat{e}).$$

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

$$J^{(K\text{-CV})} = \frac{1}{K} \sum_{k=1}^{K} J_{\mathcal{Z}_k}^{(K\text{-CV})}. \tag{6}$$

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 model that minimizes $J^{(K\text{-CV})}$. Note that $J^{(K\text{-CV})}$ is an almost unbiased estimator of the objective function (3), 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 below:

Input: $\{(x_i, \widehat{e}_i)\}_{i=1}^n$, $\{\sigma_i\}_{i=1}^p$, and $\{\lambda_j\}_{j=1}^q$
Output: LSMI parameter $\widehat{\boldsymbol{\alpha}}$

Compute CV score for $\{\sigma_i\}_{i=1}^p$ and $\{\lambda_j\}_{j=1}^q$ by Eq.(6);
Choose $\widehat{\sigma}$ and $\widehat{\lambda}$ that minimize the CV score;
Compute $\widehat{\boldsymbol{\alpha}}$ by Eq.(4) with $\widehat{\sigma}$ and $\widehat{\lambda}$;

Least-Squares Independence Regression

Given the SMI estimator (5), our next task is to learn the parameter $\boldsymbol{\beta}$ in the regression model (1) as

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

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

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

$$\boldsymbol{\beta} \longleftarrow \boldsymbol{\beta} - \eta \Big(\frac{\partial \widehat{I}(X, \widehat{E})}{\partial \boldsymbol{\beta}} + \gamma \boldsymbol{\beta} \Big), \tag{7}$$

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 $\widehat{I}(X, \widehat{E})$ with respect to $\boldsymbol{\beta}$ can be approximately expressed as

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

where

$$\frac{\partial \widehat{h}_l}{\partial \boldsymbol{\beta}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \varphi_l(x_i, \widehat{e}_i)}{\partial \boldsymbol{\beta}},$$

$$\frac{\partial \widehat{H}_{l,l'}}{\partial \boldsymbol{\beta}} = \frac{1}{n^2} \sum_{i,j=1}^{n} \Bigg(\frac{\partial \varphi_l(x_i, \widehat{e}_j)}{\partial \boldsymbol{\beta}} \varphi_{l'}(x_j, \widehat{e}_i) + \varphi_l(x_i, \widehat{e}_j) \frac{\partial \varphi_l(x_j, \widehat{e}_i)}{\partial \boldsymbol{\beta}} \Bigg),$$

$$\frac{\partial \varphi_l(x, \widehat{e})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} \varphi_l(x, \widehat{e}) (\widehat{e} - \widehat{v}_l) \boldsymbol{\psi}(x).$$

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

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

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

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 $m = \min(50, n)$, and choose τ and γ by CV with grid search as follows. First, the samples $\{z_i \mid z_i = (x_i, \widehat{e}_i)\}_{i=1}^n$ are divided into T disjoint subsets $\{\mathcal{Z}_t\}_{t=1}^T$ of (approximately) the same size (we set $T = 2$ in experiments). Then, an estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{Z}_t}$ is obtained using $\{\mathcal{Z}_j\}_{j \neq t}$, and the independence criterion for the hold-out samples $\mathcal{Z}_t$ is computed as

$$\widehat{I}_{\mathcal{Z}_t}^{(T\text{-CV})} = \frac{1}{2}\widehat{\boldsymbol{h}}_{\mathcal{Z}_t}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t} - \frac{1}{2}.$$

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

$$\widehat{I}^{(T\text{-CV})} = \frac{1}{T}\sum_{t=1}^{T} \widehat{I}_{\mathcal{Z}_t}^{(T\text{-CV})}. \tag{8}$$

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

The LSIR algorithm is summarized below:

Input: $\{(x_i, y_i)\}_{i=1}^n$, $\{\tau_i\}_{i=1}^p$, and $\{\gamma_j\}_{j=1}^q$
Output: LSIR parameter $\widehat{\boldsymbol{\beta}}$

Compute CV score for all $\{\tau_i\}_{i=1}^p$ and $\{\gamma_j\}_{j=1}^q$ by Eq.(8);
Choose $\widehat{\tau}$ and $\widehat{\gamma}$ that minimize the CV score;
Compute $\widehat{\boldsymbol{\beta}}$ by gradient descent (7) with $\widehat{\tau}$ and $\widehat{\gamma}$;

Causal Direction Inference by LSIR

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

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 exists). Then we compute SMI estimates for the permuted data $\{(x_i, \widehat{e}_{\kappa(i)})\}_{i=1}^n$ by LSMI, changing the permutation function $\kappa(\cdot)$ randomly. This random permutation process is repeated many times (in experiments, the number of repetitions is set to 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.

In order to decide the causal direction, we 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). For a given significance level δ, if $p_{X \to Y} > \delta$ and $p_{X \leftarrow Y} \leq \delta$, the model $X \to Y$ is chosen; if $p_{X \leftarrow Y} > \delta$ and $p_{X \to Y} \leq \delta$, the model $X \leftarrow Y$ is selected. If $p_{X \to Y}, p_{X \leftarrow Y} \leq \delta$, then we conclude that there is no causal relation between X and Y. If $p_{X \to Y}, p_{X \leftarrow Y} > \delta$, perhaps our modeling assumption is not correct.

When we have prior knowledge that there exists a causal relation between X and Y but their the causal direction is unknown, we may simply compare the values of $p_{X \to Y}$ and $p_{X \leftarrow Y}$: if $p_{X \to Y} > p_{X \leftarrow Y}$, we conclude that X causes Y; otherwise we conclude that Y causes X. This allows us to avoid the computational expensive permutation process.

In our preliminary experiments, we empirically observed that SMI estimates obtained by LSIR tend to be affected by the way data samples were split in the CV procedure of LSIR. To mitigate this problem, we run LSIR and compute an SMI estimate 10 times, randomly changing the data split in the CV procedure of LSIR. Then the regression function which gave the median SMI estimate among 10 repetitions is selected and the permutation test is performed for that regression function.

Experiments

In this section, we first illustrate the behavior of LSIR using a toy example, and then we evaluate the performance of LSIR using real-world datasets.

Illustrative Examples

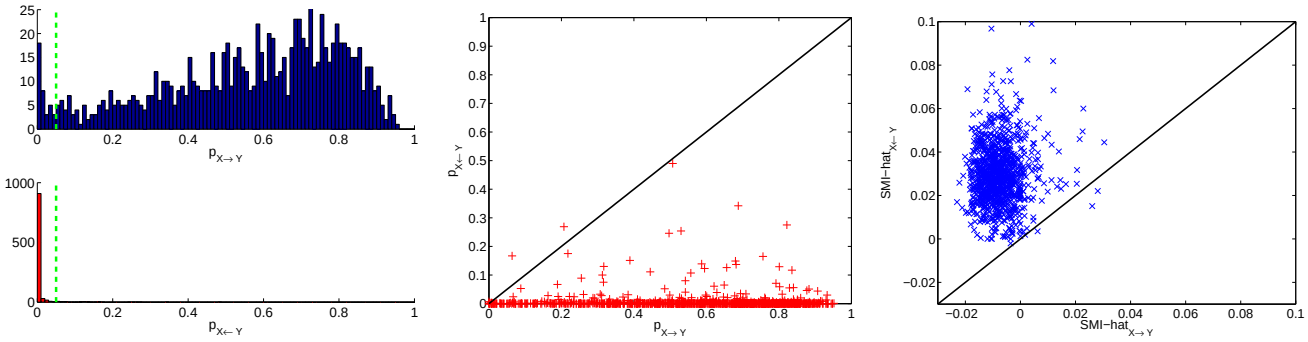
Let us consider the following additive noise model:

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

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 have mean zero). We drew 300 paired samples of X and Y following the above generative model (see Figure 1), where the ground truth is that X and E are independent. Thus, the null-hypothesis should be accepted (i.e., the p-values should be large).

Figure 1 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 963 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. The upper graph of Figure 2(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



(a) (b) (c)

Figure 2: Illustrative example. (a: upper) 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). (a: lower) 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). (b) Comparison of p-values for both directions ($p_{X\to Y}$ vs. $p_{X\leftarrow Y}$). (c) Comparison of values of independence measures for both directions ($\widehat{\mathrm{SMI}}_{X\to Y}$ vs. $\widehat{\mathrm{SMI}}_{X\leftarrow Y}$).

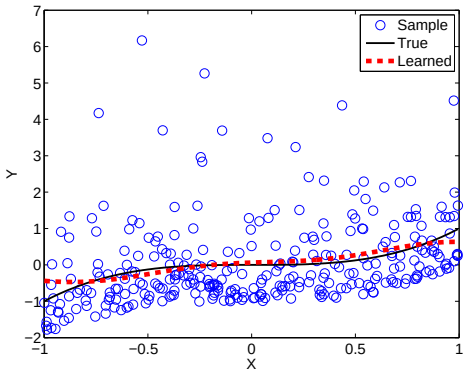


Figure 1: Illustrative example. The solid line denotes the true function, the circles denote samples, and the dashed line denotes the regressor obtained by LSIR.

significance level; the mean and standard deviation of the p-values over 1000 runs are 0.5638 and 0.2404, respectively.

Next, we consider the backward case where the roles of X and Y were swapped. In this case, the ground truth is that the input and the residual are dependent (see Figure 1). Therefore, the null-hypothesis should be rejected (i.e., the p-values should be small). The lower graph of Figure 2(a) shows the histogram of $p_{X\leftarrow Y}$ obtained by LSIR over 1000 runs. LSIR rejected the null-hypothesis 966 times out of 1000 runs; the mean and standard deviation of the p-values over 1000 runs are 0.0067 and 0.0309, respectively.

Figure 2(b) depicts the p-values for both directions in a trial-wise manner. The graph shows that LSIR results in the correct causal direction (i.e., $p_{X\to Y} > p_{X\leftarrow Y}$) 996 times out of 1000 trials, 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 2(c) 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 997 times out of 1000 trials by this simple method. This would be a practically useful heuristic since we can avoid performing the computationally intensive permutation test.

Real-world datasets

Next, we evaluate the performance of LSIR on the datasets of the '*Cause-Effect Pairs*' task in the *NIPS 2008 Causality Competition* (Mooij, Janzing, and Schölkopf 2008). The task contains 8 datasets, each has two statistically dependent random variables possessing inherent causal relationship. The goal is to identify the causal direction from the 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 p-values and the independence scores for each dataset and each direction are summarized in Table 1. LSIR with kernel width τ in the regression model optimized by CV is denoted by 'LSIR(CV)'. We also tested 'LSIR(med)', where the kernel width τ was set to the median distance between samples. This is a popular heuristic in kernel methods, and is also used in HSICR. The values of HSICR, which were also computed by the permutation test, were taken from Mooij et al. (2009), but the p-values were rounded off to three decimal places to be consistent with the results of LSIR. When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?').

Table 1 shows that LSIR(CV) successfully found the correct causal direction for 7 out of 8 cases, while LSIR(med) performed correctly only for 5 out of 8 cases. This illustrates the usefulness of CV in causal direction inference. HSICR gave the correct decision only for 5 out of 8 cases, implying that LSIR(CV) compares favorably with HSICR. For dataset 2, the p-values obtained by LSIR are large for both directions. We conjecture that our modeling assumption was not

Table 1: Results on datasets of the '*Cause-Effect Pairs*' task in the *NIPS 2008 Causality Competition* (Mooij, Janzing, and Schölkopf 2008). When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?').

(a) LSIR(CV)

Data-set	p-values		$\widehat{\mathrm{SMI}}$ ($\times 10^3$)		Direction	
	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated	Truth
1	0.920	$< 10^{-3}$	-0.2404	6.1334	$\to$	$\to$
2	0.972	0.899	-0.3618	-0.1061	$\to$	$\to$
3	0.314	$< 10^{-3}$	-0.0994	4.4031	$\to$	$\to$
4	0.023	0.591	0.0017	-0.1624	$\leftarrow$	$\leftarrow$
5	$< 10^{-3}$	0.020	3.7799	-0.0406	$\leftarrow$	$\leftarrow$
6	0.946	0.040	-0.1628	-0.0989	$\to$	$\to$
7	0.001	0.208	3.4429	-0.0508	$\leftarrow$	$\leftarrow$
8	$< 10^{-3}$	$< 10^{-3}$	0.3468	0.4064	?	$\to$

(b) LSIR(med)

Data-set	p-values		$\widehat{\mathrm{SMI}}$ ($\times 10^3$)		Direction	
	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated	Truth
1	0.977	$< 10^{-3}$	-0.0823	6.5753	$\to$	$\to$
2	0.103	0.573	-0.0757	-0.0983	$\leftarrow$	$\to$
3	0.374	$< 10^{-3}$	-0.1031	4.2570	$\to$	$\to$
4	0.087	0.962	-0.0608	-0.3944	$\leftarrow$	$\leftarrow$
5	0.063	0.987	0.3794	-0.2078	$\leftarrow$	$\leftarrow$
6	0.953	0.974	-0.1946	-0.2830	$\leftarrow$	$\to$
7	0.168	0.972	-0.0637	-0.2481	$\leftarrow$	$\leftarrow$
8	$< 10^{-3}$	$< 10^{-3}$	0.0093	0.1267	?	$\to$

(c) HSICR

Data-set	p-values		$\widehat{\mathrm{HSIC}}$		Direction	
	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated	Truth
1	0.290	$< 10^{-3}$	0.0012	0.0060	$\to$	$\to$
2	0.037	0.014	0.0020	0.0021	$\to$	$\to$
3	0.045	0.003	0.0019	0.0026	$\to$	$\to$
4	0.376	0.012	0.0011	0.0023	$\to$	$\leftarrow$
5	$< 10^{-3}$	0.160	0.0028	0.0005	$\leftarrow$	$\leftarrow$
6	$< 10^{-3}$	$< 10^{-3}$	0.0032	0.0026	?	$\to$
7	$< 10^{-3}$	0.272	0.0021	0.0005	$\leftarrow$	$\leftarrow$
8	$< 10^{-3}$	$< 10^{-3}$	0.0015	0.0017	?	$\to$

really fulfilled for this dataset.

The values of independence measures described in Table 1 show that merely comparing the values of $\widehat{\mathrm{SMI}}$ is again sufficient for deciding the correct causal direction in LSIR(CV). Actually, this heuristic also allows us to correctly identify the causal direction in Dataset 8. On the other hand, this convenient heuristic does not seem to be useful in HSICR.

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 applied the LSIR method to the discovery of non-linear causal relationship in non-Gaussian additive noise models, and experimentally showed that LSIR is promising in real-world causal direction inference.

References

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

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

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)*, 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)*, 63–78.

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

Kanamori, T.; Suzuki, T.; and Sugiyama, M. 2009. Condition number analysis of kernel-based density ratio estimation. Technical report, arXiv. http://www.citebase.org/abstract?id=oai:arXiv.org:0912.2800

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

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)*, 745–752.

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

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

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

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

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.

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*. New York, NY: Wiley.