Journal of Machine Learning Research 16 (2015) 2901-2907 Submitted 9/14; Revised 2/15; Published 12/15

The Randomized Causation Coefficient

David Lopez-Paz* DAVID@LOPEZPAZ.ORG
Max-Planck-Institute for Intelligent Systems,
Spemannstrasse 38, 72076 Tübingen, Germany

Krikamol Muandet KRIKAMOL@TUEBINGEN.MPG.DE
Max-Planck-Institute for Intelligent Systems,
Spemannstrasse 38, 72076 Tübingen, Germany

Benjamin Recht BRECHT@BERKELEY.EDU
Department of EECS, University of California Berkeley,
387 Soda Hall, Berkeley, CA 94720

Editor: Isabelle Guyon and Alexander Statnikov

Abstract

We are interested in learning causal relationships between pairs of random variables, purely from observational data. To effectively address this task, the state-of-the-art relies on strong assumptions on the mechanisms mapping causes to effects, such as invertibility or the existence of additive noise, which only hold in limited situations. On the contrary, this short paper proposes to *learn* how to perform causal inference directly from data, without the need of feature engineering. In particular, we pose causality as a kernel mean embedding classification problem, where inputs are samples from arbitrary probability distributions on pairs of random variables, and labels are types of causal relationships. We validate the performance of our method on synthetic and real-world data against the state-of-the-art. Moreover, we submitted our algorithm to the ChaLearn's "Fast Causation Coefficient Challenge" competition, with which we won the fastest code prize and ranked third in the overall leaderboard.

Keywords: causality, cause-effect inference, kernel mean embeddings, random features

1. Introduction

According to Reichenbach's common cause principle (Reichenbach, 1956), the dependence between two random variables X and Y implies that either X causes Y (denoted by $X \to Y$), or that Y causes X (denoted by $Y \to X$), or that X and Y have a common cause. In this note, we are interested in distinguishing between these three possibilities by using samples drawn from the joint probability distribution P on (X, Y).

Two of the most successful approaches to tackle this problem are the information geometric causal inference method (Daniusis et al., 2012; Janzing et al., 2014), and the additive noise model (Hoyer et al., 2009; Peters et al., 2014). First, the Information Geometric Causal Inference (IGCI) is designed to infer causal relationships between variables related by invertible, noiseless relationships. In particular, assume that there exists a pair of functions or *mapping mechanisms* f and g such that $Y = f(X)$ and $X = g(Y)$. The IGCI method

*. This project was conceived while DLP was visiting BR at University of California, Berkeley.

decides that $X \to Y$ if $\rho(P(X), |\log(f'(X))|) < \rho(P(Y), |\log(g'(Y))|)$, where ρ denotes Pearson's correlation coefficient. IGCI decides $Y \to X$ if the opposite inequality holds, and abstains otherwise. The assumption here is that the cause random variable is independently generated from the mapping mechanism; therefore it is unlikely to find correlations between the density of the former and the slope of the latter. Second, the additive noise model (ANM) assumes that the effect variable is equal to a nonlinear transformation of the cause variable plus some independent random noise, i.e., $Y = f(X) + N_Y$. If $X \perp\!\!\!\perp N_Y$, then there exists no model of the form $X = g(Y) + N_X$ for which $Y \perp\!\!\!\perp N_X$. As a result, one can find the causal direction by performing independence test between the input variable and residual variable in both directions. Specifically, the algorithm will conclude that $X \to Y$ if the pair of random variables (X, N_Y) are independent but the pair (Y, N_X) is not. The algorithm will conclude $Y \to X$ if the opposite claim is true, and abstain otherwise. The additive noise model has been extended to study post-nonlinear models of the form $Y = h(f(X) + N_Y)$, where h a monotone function (Zhang and Hyvärinen, 2009). The consistency of causal inference under the additive noise model was established by Kpotufe et al. (2013) under some technical assumptions.

As it becomes apparent from the previous exposition, there is a lack of a general method to infer causality without assuming strong knowledge about the underlying causal mechanism. Moreover, it is desirable to readily extend inference to other new model hypotheses without incurring in the development of a new, specific algorithm. Motivated by this issue, we raise the question:

Is it possible to automatically learn patterns revealing causal relationships between random variables from large amounts of labeled data?

2. Learning to Learn Causal Inference

Unlike the methods described above, we propose a *data-driven approach* to build a flexible causal inference engine. To do so, we assume access to some set of pairs $\{(S_i, l_i)\}_{i=1}^n$, where the sample $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i}$ are drawn i.i.d. from the joint distribution P_i of the two random variables X_i and Y_i, which obey the causal relationship denoted by the label l_i. To simplify exposition, the labels $l_i = 1$ denotes $X \to Y$ and $l_i = -1$ stands for $Y \to X$. Using these data, we build a causal inference algorithm in two steps. First, an m-dimensional feature vector $\mathbf{m}_i$ is extracted from each sample S_i, to meaningfully represent the corresponding distribution P_i. Second, we use the set $\{(\mathbf{m}_i, l_i)\}_{i=1}^n$ to train a binary classifier, later used to predict the causal relationship between previously unseen pairs of random variables. This framework can be straightforwardly extended to also infer the "common cause" and "independence" cases, by introducing two extra labels.

Our setup is fundamentally different from the standard classification problem in the sense that the inputs to the learners are samples from probability distributions, rather than real-valued vectors of features (Muandet et al., 2012; Szabó et al., 2014). In particular, we place two assumptions. First, the existence of a *Mother distribution* $\mathcal{M}(\mathcal{P}, \{-1, +1\})$ from which all paired probability distributions $P_i \in \mathcal{P}$ on (X_i, Y_i) and causal labels $l_i \in \{-1, +1\}$ are sampled, where $\mathcal{P}$ denotes the set of all distributions on two real-valued random variables. Second, the causal relationships l_i can be inferred in most cases from observable properties of

the distributions P_i. While these assumptions may not hold in generality, our experimental evidence suggests their wide applicability in real-world data.

The rest of this paper is organized as follows. Section 3 elaborates on how to extract the $m-$dimensional feature vectors $\mathbf{m}_i$ from each causal sample S_i. Section 4 provides empirical evidence to validate our methods. Section 5 closes the exposition by commenting on future research directions.

3. Featurizing Distributions with Kernel Mean Embeddings

Let P be the probability distribution of some random variable Z taking values in $\mathbb{R}^d$. Then, the *kernel mean embedding* of P associated with the positive definite kernel function k is

$$\mu_k(P) := \int_{\mathbb{R}^d} k(z, \cdot) \mathrm{d}P(z) \in \mathcal{H}_k, \tag{1}$$

where $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) endowed with the kernel k (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007). A sufficient condition which guarantees the existence of μ_k is that the kernel k is bounded, i.e., $\sup_{z \in \mathcal{Z}} k(z, z) < \infty$. One of the most attractive property of μ_k is that it uniquely determines each distribution P when k is a characteristic kernel (Sriperumbudur et al., 2010). In another words, $\|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = 0$ iff $P = Q$. Examples of characteristic kernels include the popular squared-exponential

$$k(z, z') = \exp\left(-\gamma \|z - z'\|_2^2\right), \text{ for } \gamma > 0, \tag{2}$$

which will be used throughout this work.

However, in practice, we do not have access to the true distribution P, and consequently to the true embedding μ_k. Instead, we often have access to a sample $S = \{z_i\}_{i=1}^n$ drawn i.i.d. from P. Then, we can construct the empirical measure $P_S = \frac{1}{n} \sum_{i=1}^n \delta_{(z_i)}$, where $\delta_{(z)}$ is the Dirac mass at z, and estimate (1) by

$$\mu_k(P_S) := \frac{1}{n} \sum_{i=1}^n k(z_i, \cdot) \in \mathcal{H}_k. \tag{3}$$

Though it can be improved (Muandet et al., 2014), the estimator (3) is the most common due to its ease of implementation. We can essentially view (1) and (3) as the feature representations of the distribution P and its sample S, respectively.

For some kernels such as (2), the feature maps (1) and (3) do not have a closed form, or are infinite dimensional. This translates into the need of kernel matrices, which require at least $O(n^2)$ computation. In order to alleviate these burdens, we propose to compute a low-dimensional approximation of (3) using random Fourier features (Rahimi and Recht, 2007). In particular, if the kernel k is shift-invariant, we can exploit Bochner's theorem (Rudin, 1962) to construct a randomized approximation of (3), with form

$$\mu_{k,m}(P_S) = \frac{1}{n} \sum_{i=1}^n \left[\cos(w_1' z_i + b_1), \ldots, \cos(w_m' z_i + b_m)\right]' \in \mathbb{R}^m, \tag{4}$$

where the vectors $w_1, \ldots, w_m \in \mathbb{R}^d$ are sampled from the normalized Fourier transform of k, and $b_1, \ldots, b_m \sim \mathcal{U}(0, 2\pi)$. The squared-exponential kernel in (2) is shift-invariant, and

can be approximated in this fashion when setting $w_i \sim \mathcal{N}(0, 2\gamma I)$. These features can be computed in $O(mn)$ time and stored in $O(1)$ memory. Importantly, the low dimensional representation $\mu_{k,m}$ is amenable for the off-the-shelf use with any standard learning algorithm, and not only kernel-based methods.

Using the assumptions introduced in Section 1, the data $\{(\mathbf{m}_i, l_i)\}_{i=1}^n := \{(\mu_{k,m}(P_{S_i}), l_i)\}_{i=1}^n$ and a binary classifier, we can now pose causal inference as a supervised learning problem.

4. Numerical Simulations

We conduct an array of experiments to test the effectiveness of a simple implementation of the presented causal learning framework[1]. Given the use of random embeddings (4) in our classifier, we term our method the *Randomized Causation Coefficient* (RCC). Throughout our simulations, we featurize each sample $S = \{(x_i, y_i)\}_{i=1}^n$ as

$$\nu(S) = (\mu_{k,m}(P_{S_x}), \mu_{k,m}(P_{S_y}), \mu_{k,m}(P_S)), \tag{5}$$

where the three elements forming (5) stand for the low-dimensional representations (4) of the empirical kernel mean embeddings of $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, and $\{(x_i, y_i)\}_{i=1}^n$, respectively. This representation is motivated by the typical conjecture in causal inference about the existence of asymmetries between the marginal and conditional distributions of causally-related pairs of random variables (Schölkopf et al., 2012). Each of these three embeddings has random features sampled to approximate the sum of three Gaussian kernels (2) with hyper-parameters 0.1γ, γ, and 10γ, where γ is set using the median heuristic. In practice, we set $m = 1000$, and observe no significant improvements when using larger amounts of random features. To classify the embeddings (5) in each of the experiments, we use the random forest implementation from Python's `sklearn-0.16-git`. The number of trees forming the forest is chosen from the set $\{100, 250, 500, 1000, 5000\}$, via cross-validation.

4.1 Tübingen Data

The *Tübingen cause-effect pairs* is a collection of heterogeneous, hand-collected, real-world cause-effect samples[2]. Given the small size of this data set, we resort to the synthesis of some Mother distribution to sample our training data from. To this end, assume that sampling a synthetic cause-effect sample $\hat{S}_i := \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n$ equals the following generative process:

1. A *cause* vector $(\hat{x}_{ij})_{j=1}^n$ is sampled from a mixture of Gaussians with c components. The mixture weights are sampled from $\mathcal{U}(0, 1)$, and normalized to sum to one. The mixture means and standard deviations are sampled from $\mathcal{N}(0, \sigma_1)$, and $\mathcal{N}(0, \sigma_2)$, respectively, accepting only positive standard deviations. The cause vector is standardized.

2. A *noise* vector $(\hat{\epsilon}_{ij})_{j=1}^n$ is sampled from a centered Gaussian, with variance sampled from $\mathcal{U}(0, \sigma_3)$.

1. The source code of our experiments is available at `https://github.com/lopezpaz/causation_learning_theory`.
2. The Tübingen cause-effect pairs data set can be downloaded at `https://webdav.tuebingen.mpg.de/cause-effect/`.

3. The *mapping mechanism* $\hat{f}_i$ is a spline fitted using an uniform grid of d_f elements from $\min((\hat{x}_{ij})_{j=1}^n)$ to $\max((\hat{x}_{ij})_{j=1}^n)$ as inputs, and d_f normally distributed outputs.

4. An *effect* vector is built as $(\hat{y}_{ij} := \hat{f}_i(\hat{x}_{ij}) + \hat{\epsilon}_{ij})_{j=1}^n$, and standardized.

5. Return the cause-effect sample $\hat{S}_i := \{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n$.

To choose a $\theta = (c, \sigma_1, \sigma_2, \sigma_3, d_f)$ that best resembles the unlabeled test data, we minimize the distance between the embeddings of N synthetic pairs and the Tübingen samples

$$\arg\min_{\theta} \sum_i \min_{1 \le j \le N} \|\nu(S_i) - \nu(\hat{S}_j)\|_2^2,$$

over $c, d_f \in \{1, \ldots, 10\}$, and $\sigma_1, \sigma_2, \sigma_3 \in \{0, 0.5, 1, \ldots, 5\}$, where the $\hat{S}_j$ is sampled using the generative process described above, the S_i are the Tübingen cause-effect pairs, and ν is as in (5). This strategy can be thought of as transductive learning, since we have access to the test inputs (but not their underlying causal relation) at the training time.

We set $n = 1000$, and $N = 10,000$. Using the generative process described above, and the best found parameter vector $\theta = (3, 2, 2, 2, 5)$, we construct the synthetic training data

$$\{\{\nu(\{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n), +1)\}_{i=1}^N,$$
$$\{\nu(\{(\hat{y}_{ij}, \hat{x}_{ij})\}_{j=1}^n), -1)\}_{i=1}^N\},$$

where $\{(\hat{x}_{ij}, \hat{y}_{ij})\}_{j=1}^n = \hat{S}_i$, and train our classifier on it. Figure 1 plots the classification accuracy of RCC, IGCI (Daniusis et al., 2012), and ANM (Mooij et al., 2014) versus the fraction of decisions that the algorithms are forced to make out of the 82 scalar Tüebingen cause-effect pairs. To compare these results to other lower-performing methods, refer to Janzing et al. (2012). Overall, RCC surpasses the state-of-the-art in these data, with a classification accuracy of 81.61% when inferring the causal directions on all pairs. The confidence of RCC is computed using the random forest's output class probabilities.

4.2 ChaLearn's "Fast Causation Coefficient" Challenge

We tested RCC at the ChaLearn's *Fast Causation Coefficient* challenge (Guyon, 2014). We trained a Gradient Boosting Classifier (GBC), with hyper-parameters chosen via a 4-fold cross validation, on the featurizations (5) of the training data. In particular, we built two separate classifiers: a first one to distinguish between causal and non-causal pairs (i.e., $X - Y$ vs $\{X \to Y, X \leftarrow Y\}$), and a second one to distinguish between the two possible causal directions on the causal pairs (i.e., $X \to Y$ vs $X \leftarrow Y$). The final causation coefficient for a given sample S_i was computed as

$$\text{score}(S_i) = p_1(S_i) \cdot (2 \cdot p_2(S_i) - 1),$$

where $p_1(x)$ and $p_2(x)$ are the class probabilities output by the first and the second GBCs, respectively. We found it easier to distinguish between causal and non-causal pairs than to infer the correct direction on the causal pairs.

RCC ranked third in the ChaLearn's "Fast Causation Coefficient Challenge" competition, and was awarded the prize to the fastest running code (Guyon, 2014). At the time of the

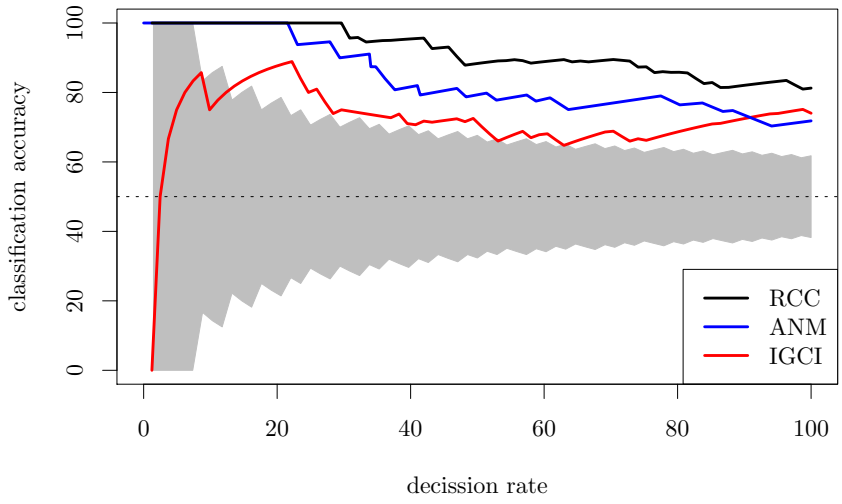

Figure 1: Accuracy of RCC, IGCI and ANM on the Tübingen cause-effect pairs, as a function of decision rate. The grey area depicts accuracies not statistically significant.

competition, we obtained a bidirectional AUC of 0.73 on the test pairs in two minutes of test-time (Guyon, 2014). On the other hand, the winning entry of the competition, which made use of hand-engineered features, took a test-time of 30 minutes, and achieved a bidirectional AUC of 0.82. Interestingly, the performance of IGCI on the $20,000$ training pairs is barely better than random guessing. The computational complexity of the additive noise model (usually implemented as two Gaussian Process regressions followed by two kernel-based independence tests) made it unfeasible to compare it on this data set.

5. Conclusions and Future Research

To conclude, we proposed to *learn how to perform causal inference* between pairs of random variables from observational data, by posing the task as a supervised learning problem. In particular, we introduced an effective and efficient featurization of probability distributions, based on kernel mean embeddings and random Fourier features. Our numerical simulations support the conjecture that patterns revealing causal relationships can be learnt from data.

In light of our encouraging results, we would like to mention four exciting research directions. First, the proposed ideas can be used to learn other *domain-general* statistics, such as measures of dependence (Lopez-Paz et al., 2013). Second, it is important to develop techniques to visualize and interpret the causal features learned by our classifiers. This direction is particularly essential for causal inference as it provides a data-dependent way of discovering new hypothesis on underlying causal mechanism. Third, RCC can be extended to operate not only on pairs, but also sets of random variables, and eventually reconstruct causal DAGs from multivariate data. Finally, one may adapt the distributional learning theory of Szabó et al. (2014) to analyze our randomized, classification setting. For preliminary results on the last two points, we refer the reader to (Lopez-Paz et al., 2015).

References

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, 2004.

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. *UAI*, 2012.

I. Guyon. Chalearn fast causation coefficient challenge, 2014. URL `https://www.codalab.org/competitions/1381`.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. R. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *NIPS*, 2009.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 2012.

D. Janzing, B. Steudel, N. Shajarisales, and B. Schölkopf. Justifying information-geometric causal inference. *arXiv prepring arXiv:1402.2499*, 2014.

S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf. Consistency of causal inference under the additive noise model. *ICML*, 2013.

D. Lopez-Paz, P. Hennig, and B. Schölkopf. The Randomized Dependence Coefficient. *NIPS*, 2013.

D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of causation. *ICML*, 2015.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*, 2014.

K. Muandet, K. Fukumizu, F. Dinuzzo, and B. Schölkopf. Learning from distributions via support measure machines. *NIPS*, 2012.

K. Muandet, K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Kernel mean estimation and Stein effect. *ICML*, 2014.

J. Peters, Joris M. M., D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 2014.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *NIPS*, 2007.

H. Reichenbach. *The direction of time*. Dover, 1956.

W. Rudin. *Fourier analysis on groups*. Wiley, 1962.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. M. Mooij. On causal and anticausal learning. In *ICML*, 2012.

A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *ALT*. Springer-Verlag, 2007.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 2010.

Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. *arXiv preprint arXiv:1402.1754*, 2014.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. *UAI*, 2009.