Kernel Methods for Detecting the Direction of Time Series

Jonas Peters[1], Dominik Janzing[1], Arthur Gretton[1], and Bernhard Schölkopf[1]

MPI for biological Cybernetics, 72076 Tübingen
jonas.peters@tuebingen.mpg.de, dominik.janzing@tuebingen.mpg.de,
arthur@tuebingen.mpg.de and bernhard.schoelkopf@tuebingen.mpg.de

Summary. We propose two kernel based methods for detecting the time direction in empirical time series. First we apply a Support Vector Machine on the finite-dimensional distributions of the time series (*classification method*) by embedding these distributions into a Reproducing Kernel Hilbert Space. For the *ARMA method* we fit the observed data with an autoregressive moving average process and test whether the regression residuals are statistically independent of the past values. Whenever the dependence in one direction is significantly weaker than in the other we infer the former to be the true one.
Both approaches were able to detect the direction of the true generating model for simulated data sets. We also applied our tests to a large number of real world time series. The ARMA method made a decision for a significant fraction of them, in which it was mostly correct, while the classification method did not perform as well, but still exceeded chance level.

Key words: Classification, Support Vector Machines, Time Series

1 Introduction

Consider the following problem: We are given m ordered values $X_1, \ldots, X_m$ from a time series, but we do not know their direction in time. Our task is to find out whether $X_1, \ldots, X_m$ or $X_m, \ldots, X_1$ represents the true direction. The motivation to study this unusual problem is two-fold:

(1) The question is a simple instantiation of the larger issue of what characterizes the direction of time, which is an issue related to philosophy and physics [13], in particular to the second law of thermodynamics. One possible formulation of the latter states that the entropy of a closed physical system can only increase but never decrease (from a microphysical perspective the entropy is actually constant in time but only increases after appropriate *coarse-graining* the physical state space [1]). This may suggest the use of entropy criteria to identify the time direction in empirical data. However, most real-life time series (such as that given by data from stock markets, EEGs,

or meteorology) do not stem from closed physical systems, and there is no obvious way to use entropy for detecting the time direction. Moreover, we also want to detect the direction of *stationary* processes.

(2) Analyzing such asymmetries between past and future can provide new insights for *causal* inference. Since every cause precedes its effect (equivalently, the future cannot influence the past), we have, at least, *partial* knowledge of the ground truth [4].

In this work we propose the classification method for solving this problem: Consider a strictly stationary time series, that is a process for which the w-dimensional distribution of $(X_{t+h}, X_{t+1+h} \ldots, X_{t+w+h})$ does not depend on h for any choice of $w \in \mathbb{N}$. We assume that the difference between a forward and backward sample ordering manifests in a difference in the finite-dimensional distributions $(X_t, X_{t+1} \ldots, X_{t+w})$ and $(X_{t+w}, X_{t+w-1} \ldots, X_t)$. For many time series we thus represent both distributions in a Reproducing Kernel Hilbert Space and apply a Support Vector Machine within this Hilbert Space.

In [14] Shimizu et al. applied their causal discovery algorithm LINGAM to this problem. Their approach was able to propose a hypothetical time direction for 14 out of 22 time series (for the other cases their algorithm did not give a consistent result); however, only 5 out of these 14 directions turned out to be correct. Possible reasons for this poor performance will be discussed below. Nevertheless, we now describe the idea of LINGAM because our ARMA method builds upon the same idea. Let ϵ be the residuum after computing a least squares linear regression of Y on X for real-valued random variables X, Y. If X and ϵ are statistically independent (note that they are *uncorrelated* by construction) we say that the joint distribution $P(X, Y)$ admits a linear model from X to Y. Then the only case admitting a linear model in both directions is that $P(X, Y)$ is bivariate Gaussian (except for the trivial case of independence). The rationale behind LINGAM is to consider the direction as causal that can better be fit by a linear model. This idea also applies to causal inference with directed acyclic graphs (DAGs) having n variables $X_1, \ldots, X_n$ that linearly influence each other.

There are three major problems when using conventional causal inference tools [11, 16] in determining time series direction. First, the standard framework refers to data matrices obtained by iid sampling from a joint distribution on n random variables. Second, for interesting classes of time series like MA and ARMA models (introduced in Section 2), the observed variables (X_t) are not causally sufficient since the (hidden) noise variables influence more than one of the observed variables. Third, the finitely many observations are typically preceded by instances of the same time series, which have not been observed.

Besides the classification method mentioned before we propose the following ARMA approach: for both time directions, we fit the data with an ARMA model and check whether the residuals occurring are indeed independent of the past observations. Whenever the residuals are significantly less dependent

for one direction than for the converse one, we infer the true time direction to be the former. To this end, we need a good dependence measure that is applicable to continuous data and finds dependencies beyond second order. Noting that the ARMA method might work for other dependence measures, too, we use the Hilbert Schmidt Independence Criterion (HSIC) [7]. This recently developed kernel method will be described together with the concept of Hilbert Space embeddings and ARMA processes in Section 2. Section 3 explains the method we employ for identifying the true time direction of time series data. In Section 4 we present results of our methods on both simulated and real data.

2 Statistical Methods

2.1 A Hilbert Space Embedding for Distributions

Recall that for a positive definite kernel k a Hilbert space $\mathcal{H}$ of functions $f: \mathcal{X} \to \mathbb{R}$ is called a *Reproducing Kernel Hilbert Space (RKHS)* if $k(x,.) \in \mathcal{H} \quad \forall x \in \mathcal{X}$ and $\langle f, k(x,.)\rangle = f(x) \quad \forall f \in \mathcal{H}$. Here, $k(x,.)$ denotes a function $\mathcal{X} \to \mathbb{R}$ with $y \mapsto k(x,y)$. We can represent received data in this RKHS using the feature map $\Phi(x) := k(x,.)$.

We can further represent probability distributions in the RKHS [15]. To this end we define the *mean elements* $\mu[\mathbf{P}](.) = \mathbf{E}_{X\sim\mathbf{P}} k(X,.)$, which are vectors obtained by averaging all $k(X,.)$ over the probability distribution $\mathbf{P}$. Gretton et al. [6] now introduced the Maximum Mean Discrepancy (MMD) between two probability measures $\mathbf{P}$ and $\mathbf{Q}$, which is defined in the following way: mapping the two measures into an RKHS via $\mathbf{P} \mapsto \mu[\mathbf{P}]$, the MMD is the RKHS distance $\|\mu[\mathbf{P}] - \mu[\mathbf{Q}]\|_{\mathcal{H}}$ between these two points. Assume the following conditions on the kernel are satisfied: $k(x,y) = \psi(x-y)$ for some positive definite function ψ, and ψ is bounded and continuous. Bochner's theorem states that ψ is the Fourier transform of a nonnegative measure Λ. Assume further that Λ has a density with respect to the Lebesgue measure, which is strictly positive almost everywhere. It can be shown that under these conditions on the kernel the embedding μ is injective [17] and therefore the MMD is zero if and only if $\mathbf{P} = \mathbf{Q}$. Note that the Gaussian kernel $k(x,y) = \exp\big(-\|x-y\|^2/(2\sigma^2)\big)$ on $\mathbb{R}^d$ satisfies all conditions mentioned above. In our experiments we chose $2\sigma^2$ to be the median of all distances $\|x-y\|^2$, following [6].

If only a finite sample $(X_1, \ldots, X_m)$ of a random variable is given we can estimate the mean element by $\mu[\hat{\mathbf{P}}_m^X] = \frac{1}{m}\sum_{i=1}^m k(X_i,.)$. If the kernel is strictly positive definite the two function values are the same if and only if the samples are of the same size and consist of exactly the same points. In this sense these Hilbert space representations inherit all relevant statistical information of the finite sample.

2.2 Hilbert Schmidt Independence Criterion

As mentioned earlier, the ARMA method requires an independence criterion that is applicable to continuous data. The Hilbert-Schmidt Independence Criterion (HSIC) is a kernel based statistic that, for sufficiently rich Reproducing Kernel Hilbert Spaces (RKHSs), is zero only at independence. The name results from the fact that it is the Hilbert-Schmidt norm of a cross-covariance operator [7]. Following [15], we will introduce HSIC in a slightly different way, however, using the Hilbert Space Embedding from Section 2.2.
For the formal setup let X and Y be two random variables taking values on $(\mathcal{X}, \Gamma)$ and $(\mathcal{Y}, \Delta)$, respectively; here, $\mathcal{X}$ and $\mathcal{Y}$ are two separable metric spaces, and Γ and Δ are Borel σ-algebras. We define positive definite kernels $k(.,.)$ and $l(.,.)$ on the spaces $\mathcal{X}$ and $\mathcal{Y}$ and denote the corresponding RKHS as $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, respectively. The product space $(\mathcal{X} \times \mathcal{Y}, \Gamma \otimes \Delta)$ is again a measurable space and we can define the product kernel $k(.,.) \cdot l(.,.)$ on it. X and Y are independent if and only if $\mathbf{P}^{(X,Y)} = \mathbf{P}^X \otimes \mathbf{P}^Y$. This means a dependence between X and Y is equivalent to a difference between the distributions $\mathbf{P}^{(X,Y)}$ and $\mathbf{P}^X \otimes \mathbf{P}^Y$.
The HSIC can be defined as the MMD between $\mathbf{P}^{(X,Y)}$ and $\mathbf{P}^X \otimes \mathbf{P}^Y$. It can further be shown [7] that for a finite amount of data, a biased empirical estimate of HSIC is given by a V-statistic,

$$\widehat{\mathrm{HSIC}} = m^{-2}\mathrm{tr}HKHL,$$

where $H = 1 - \frac{1}{m} \cdot (1, \ldots, 1)^t(1, \ldots, 1)$, K and L are the Gram matrices of the kernels k and l respectively, and m is the number of data points. Under the assumption of independence (where the true value of HSIC is zero), $m \cdot \widehat{\mathrm{HSIC}}$ converges in distribution to a weighted sum of Chi-Squares, which can be approximated by a Gamma distribution [9]. Therefore we can construct a test under the null hypothesis of independence.

2.3 Auto-Regressive Moving Average Models

Recall that a time series $(X_t)_{t\in\mathbb{Z}}$ is a collection of random variables and is called strictly stationary if $(X_{t_1}, \ldots, X_{t_n})$ and $(X_{t_1+h}, \ldots, X_{t_n+h})$ are equal in distribution for all $t_j, h \in \mathbb{Z}$. It is called weakly (or second-order) stationary if X_t has finite variance and

$$\mathbf{E}X_t = \mu \qquad \text{and} \qquad \mathrm{cov}(X_t, X_{t+h}) = \gamma_h \quad \forall t, h \in \mathbb{Z},$$

i.e., both mean and covariance do not depend on the time t, and the latter depends only on the time gap h.

Definition 1. *A time series $(X_t)_{t\in\mathbb{Z}}$ is called an autoregressive moving average process of order (p, q), written ARMA(p, q), if it is weakly stationary and if*

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \qquad \forall t \in \mathbb{Z}\,,$$

where ϵ_t are iid and have mean zero. The process is called an MA process if $p = 0$ and an AR process if $q = 0$.

An ARMA process is called causal if every noise term ϵ_t is independent of all X_i for $i < t$.

We call an ARMA process *time-reversible*, if there is an iid noise sequence $\tilde{\epsilon}_t$, such that $X_t = \sum_{i=1}^{\tilde{p}} \tilde{\phi}_i X_{t+i} + \sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t$ where $\tilde{\epsilon}_t$ is independent of all X_i with $i > t$.

In the theoretical work [18] and [2] the authors call a strictly stationary process time-reversible if $(X_0, \ldots, X_h)$ and $(X_0, \ldots, X_{-h})$ are equal in distribution for all h. However, this notion is not appropriate for our purpose because, a priori, it could be that forward and backward processes are both ARMA processes even though they do not coincide in distribution.

3 Learning the True Time Direction

3.1 The Classification Method

In Section 2.1 we saw how to represent a sample distribution in an RKHS. Given a training and a test set we can perform a linear SVM on these representations in the RKHS. This linear classifier only depends on pairwise dot products, which we are able to compute:

$$\Big\langle \frac{1}{m}\sum_{i=1}^{m} k(x_i, .), \frac{1}{n}\sum_{j=1}^{n} k(\tilde{x}_j, .) \Big\rangle = \frac{1}{m\,n}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i, \tilde{x}_j)\,. \tag{1}$$

Note that this method allows us to perform an SVM on distributions. We note that this is only one possible kernel on the space of probability measures; see [8] for an overview.

We now apply this idea to the finite-dimensional distributions of a time series. Each time series yields two points in the RKHS (correct and reversed direction), on which we perform the SVM. The classification method can be summarized as follows:

- Choose a fixed window length w and take for each time series many finite-dimensional samples $\mathbf{X}_{t_1} = (X_{t_1}, \ldots, X_{t_1+w})$, $\mathbf{X}_{t_2} = (X_{t_2}, \ldots, X_{t_2+w})$, $\mathbf{X}_{t_m} = (X_{t_m}, \ldots, X_{t_m+w})$. The t_i can be chosen such that $t_{i+1} - (t_i + w) =$ const, for example. The larger the gap between two samples of the time series, the less dependent these samples will be (ideally, we would like to have iid data, which is, of course, impossible for structured time series). Represent the distribution of $(X_t, \ldots, X_{t+w})$ in the RKHS using the point $\frac{1}{m}\sum_{i=1}^{m} k(\mathbf{X}_{t_i}, .)$.
- Perform a (linear) soft margin SVM on these points using (1) for computing the dot product.

3.2 The ARMA Method

We state without proof the following theorem [12]:

Theorem 1. *Assume that (X_t) is a causal ARMA process with iid noise and non-vanishing AR part. Then the process is time-reversible if and only if the process is Gaussian distributed.*

If a time series is an ARMA process with non-Gaussian noise, the reversed time series is not an ARMA process anymore. Theorem 1 justifies the following procedure to predict the true time direction of a given time series:

- Assume the data come from a causal ARMA process with non-vanishing AR part and independent, non-Gaussian noise.
- Fit an ARMA process in both directions to the data (see e.g. [3]) and consider the residuals ϵ_t, $\tilde{\epsilon}_t$ respectively.
- Using HSIC and a significance level α test if ϵ_t depends on $X_{t-1}, X_{t-2}, \ldots$ or if $\tilde{\epsilon}_t$ depends on $X_{t+1}, X_{t+2}, \ldots$ and call the p-values of both tests p_1 and p_2, respectively. According to Theorem 1 only one dependence should be found.
  If the hypothesis of independence is indeed rejected for only one direction (i.e. exactly one p-value, say p_1, is smaller than α) and additionally, $p_2 - p_1 > \delta$, then propose the direction of p_2 to be the correct one.
- If the noise seems to be Gaussian (e.g. perform the Jarque-Bera test [5]) do not decide.
- If both directions lead to dependent noise, conclude that the model assumption is not satisfied and do not decide.

For the method described above, two parameters need to be chosen: the significance level α and the minimal difference in p-values δ. When α is smaller and δ is larger, the method makes fewer decisions, but these should be more accurate. Note that for the independence test we need iid data, which we cannot assume here. The time series values may have the same distribution (if the time series is strictly stationary), but two adjacent values are certainly not independent. We reduce this problem, however, by not considering every point in time, but leaving gaps of a few time steps in between.

Note that the ARMA method only works for ARMA processes with non-Gaussian noise. Gaussianity is often used in applications because of its nice computational properties, but there remains controversy as to how often this is consistent with the data. In many cases using noise with heavier tails than the Gaussian would be more appropriate (e.g. [10]).

4 Experiments

Simulated Data.

We show that the methods work for simulated ARMA processes provided that the noise is not Gaussian distributed. We simulated data from an ARMA(2,2)

time series with fixed parameters ($\phi_1 = 0.9, \phi_2 = -0.3, \theta_1 = -0.29$ and $\theta_2 = 0.5$) and using varying kinds of noise. The coefficients are chosen such that they result in a causal process (see [3] for details). For different values of r we sampled

$$\epsilon_t \sim \mathrm{sgn}(Z) \cdot |Z|^r,$$

where $Z \sim \mathcal{N}(0,1)$, and normalized in order to obtain the same variance for all r. Only $r = 1$ corresponds to a normal distribution. Theorem 1 states that the reversed process is again an ARMA(2,2) process only for $r = 1$, which results in the same finite-dimensional distributions as the correct direction. Thus we expect both methods to fail in the Gaussian case. However, we are dealing with a finite amount of data and if r is close to 1, the noise cannot be distinguished from a Gaussian distribution and we will still be able to fit a backward model reasonably well.

The classification method performed well on the simulated data (see Figure 1). Notice, however, that using the same coefficients in all simulations makes the problem for the SVM considerably easier. When we used different parameters for each simulated time series the classification method performed much worse (at chance level), while the ARMA method could still detect the correct direction in most cases.

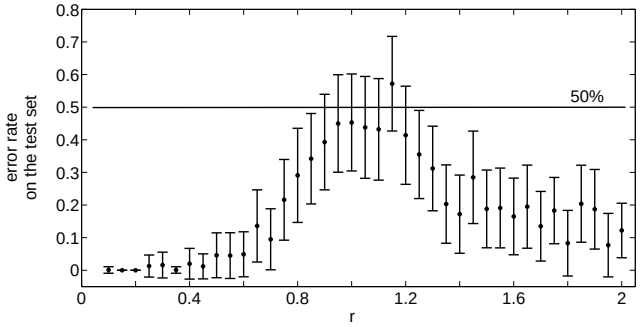

Fig. 1. Classification method on the ARMA processes. For each value of r (i.e. for each kind of noise) we simulated 100 instances of an ARMA(2,2) process with fixed coefficients and divided them into 85 time series for training and 15 for testing; this was done 100 times for each r. The graph shows the average classification error on the test set and the corresponding standard deviation.

For the ARMA method we fit an ARMA model to the data without making use of the fact that we already know the order of the process; instead we used the Akaike Information Criterion which penalizes a high order of the model. If we detected a dependence between residuals and past values of the time series, we rejected this direction, otherwise we accepted it. Obviously, for the true direction we expect that independence will only be rejected in very few

cases (depending on the significance level). For the independence test we used a significance level of $\alpha = 0.01$. See Figure 2 for details.

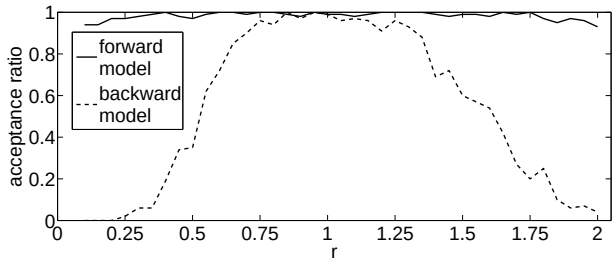

Fig. 2. For each value of r (expressing the non-Gaussianity of the noise) we simulated 100 instances of an ARMA(2,2) process with 500 time points and show the acceptance ratio for the forward model (solid line) and for the backward model (dashed line). When the noise is significantly different from Gaussian noise ($r = 1$), the correct direction can be identified.

Real World Data.

In order to see if the methods are applicable to real data as well, we collected data consisting of 200 time series with varying length (from 100 up to 10,000 samples) from very different areas: finance, physics, transportation, crime, production of goods, demography, economy, EEG data and agriculture. Roughly two thirds of our data sets belonged to the groups economy and finance.

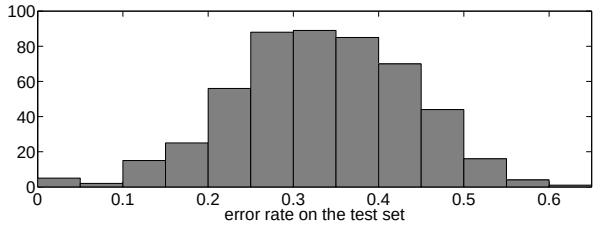

Fig. 3. Classification method on the time series collection. 500 times we chose randomly a test set of size 20, trained the method on the remaining 180 time series and looked at the performance on the test set. For the SVM regularization parameter we chose $C = 10$, which resulted in a training error of $29.8\% \pm 1.8\%$ and a test error of $35.7 \pm 10.5\%$. We reached the same performance, however, for values of C which were several orders of magnitude lower or higher.

Since the performance of the ARMA method strongly depends on the chosen parameters, we give the results for different values. The classification consistently exceeds 50%; and for more conservative parameter choices, a larger proportion of time series are correctly classified. See Figure 4 for details.

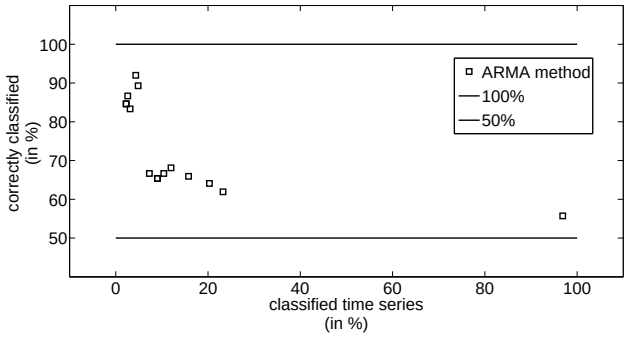

Fig. 4. ARMA method on the time series collection. We cut the longer time series into smaller pieces of length 400 and obtained 576 time series instead of 200. We show the results for different values of the parameters: the minimal difference in p-values δ ranges between 0% and 20%, and the significance level α between 10% and 0.1%. The point with the best classification performance corresponds to the highest value of δ and the lowest value of α.

5 Conclusion and Discussion

We have proposed two methods to detect the time direction of time series. One method is based on a Support Vector Machine, applied to the finite-dimensional distributions of the time series. The other method tests the validity of an ARMA model for both directions by testing whether the residuals from the regression were statistically independent of the past observations.

Experiments with simulated data sets have shown that we were able to identify the true direction in most cases unless the ARMA processes were Gaussian distributed (and thus time-reversible). For a collection of real world time series we found that in many cases the data did not admit an ARMA model in either direction, or the distributions were close to Gaussian. For a considerable fraction, however, the residuals were significantly less dependent for one direction than for the other. For these cases, the ARMA method mostly recovered the true direction. The classification method performed on average worse than the ARMA method, but still exceeded chance level.

Classification accuracies were not on par with the classification problems commonly considered in machine learning, but we believe that this is owed to the difficulty of the task; indeed we consider our results rather encouraging.

It is interesting to investigate whether there are more subtle asymmetries between past and future in time series that cannot be classified by our approach (i.e., if there is a simple generative model in the forward but not the backward direction in a more general sense). Results of this type would shed further light on the statistical asymmetries between cause and effect.

References

1. R. Balian. *From microphysics to macrophysics*. Springer, 1992.
2. F. J. Breidt and R. A. Davis. Time-reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis*, 13(5):379–390, 1991.
3. P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 2 edition, 1991.
4. M. Eichler and V. Didelez. Causal reasoning in graphical time series models. *In: Proceedings of the 23nd Annual Conference on Uncertainty in Artifical Intelligence*, pages 109–116, 2007.
5. C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172, 1987.
6. A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520, Cambridge, MA, 2007. MIT Press.
7. A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference*, pages 63–78. Springer-Verlag, 2005.
8. M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. Proceedings of AISTATS, 2005.
9. A. Kankainen. Consistent testing of total independence based on the empirical characteristic function. *PhD Thesis, University of Jyväskylä*, 1995.
10. B. Mandelbrot. On the distribution of stock price differences. *Operations Research*, 15(6):1057–1062, 1967.
11. J. Pearl. *Causality*. Cambridge University Press, 2002.
12. J. Peters. Asymmetries of Time Series under Inverting their Direction. Diploma Thesis, 2008.
13. H. Reichenbach. *The direction of time*. University of California Press, Berkeley, 1956.
14. S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.
15. A. J. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, pages 13–31. Springer-Verlag, 2007.
16. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, 1993. (2nd ed. MIT Press 2000).
17. B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective hilbert space embeddings of probability measures. In *COLT*, 2008.
18. G. Weiss. Time-reversibility of linear stochastic processes. *J. Appl. Prob.*, 12:831–836, 1975.