Detecting the Direction of Causal Time Series

Jonas Peters JONAS.PETERS@TUEBINGEN.MPG.DE
Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Arthur Gretton ARTHUR.GRETTON@TUEBINGEN.MPG.DE
Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max-Planck-Institute for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany

Abstract

We propose a method that detects the true direction of time series, by fitting an autoregressive moving average model to the data. Whenever the noise is independent of the previous samples for one ordering of the observations, but dependent for the opposite ordering, we infer the former direction to be the true one. We prove that our method works in the population case as long as the noise of the process is not normally distributed (for the latter case, the direction is not identifiable). A new and important implication of our result is that it confirms a fundamental conjecture in causal reasoning — if after regression the noise is independent of signal for one direction and dependent for the other, then the former represents the true causal direction — in the case of time series. We test our approach on two types of data: simulated data sets conforming to our modeling assumptions, and real world EEG time series. Our method makes a decision for a significant fraction of both data sets, and these decisions are mostly correct. For real world data, our approach outperforms alternative solutions to the problem of time direction recovery.

1. Introduction

The field of causal discovery has as its purpose the inference of causal directions from observed data, where explicit intervention is not permitted (due to considerations of practicality or ethics). Several recent advances in causal inference (Shimizu et al., 2006; Hoyer et al., 2009) have been based around the following design principle: assume continuous random variables X and Y are related according to $Y = f(X) + \epsilon$, where ϵ is additive i.i.d. noise independent of X. To fit the model, regress both X on Y and Y on X, and test for which direction the obtained noise is independent of the predictor variable: when the residual noise ϵ is independent of X, we determine X to be the cause and Y to be the effect. We will refer to this principle as *causal direction through noise dependence*.

While the above strategy makes intuitive sense, it is necessary to specify conditions on f and ϵ under which it yields an unambiguous answer. Hoyer et al. (2009) show that apart from certain highly contrived artificial examples, the forward and reverse models will *both* exhibit independence between predictors and noise *only* when f is linear and the noise is Gaussian. Thus, apart from these cases, the underlying causal direction can be found according to the procedure outlined above (Shimizu et al., 2006; Hoyer et al., 2009).

To further support this kind of causal inference rule we demonstrate in the current paper that the principle of causal direction through noise dependence can be applied to determine the direction of a time series. In our main theorem, we prove that when the time series is a causal autoregressive moving average (ARMA) process with a non-Gaussian noise distribution, then the noise is independent of all preceding values of the time series *only* when the correct ordering is used. On the other hand, when Gaussian noise is present, the random process has no underlying direction. We use these insights as the basis for a framework to detect time series direction, and provide a specific implementation of this framework based on a kernel independence test (Gretton et al., 2008). A preliminary version of the method (without proofs) can be found in (Peters et al., 2009).

While our main motivation is the issue of causality,

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

the asymmetry between past and future is considered a fundamental problem of physics (Reichenbach, 1999; Hawking, 1995). Thermodynamics suggests to search for asymmetries that can be phrased in terms of entropy criteria. This is because the entropy of a closed physical system can only increase or remain constant, but never decrease in time (the entropy remains constant but increases after appropriate coarse-graining of the physical state space (Balian, 1992)). In this paper, however, we will consider time series that do not describe the state of any closed physical system. Instead, they stem from the measurement of a particular quantity in a complex system (an EEG measurement of brain activity). Some of the time-series are stationary, which shows that no entropy increase can indicate the time direction. We will find the time direction by a method that assumes that the corresponding stochastic process has a simpler description in forward time direction than in backward time direction. Models in (Janzing, 2007) suggest that this kind of asymmetry is linked to the arrow of time in statistical physics, even though the connection to the above entropy criterion is not obvious. Hence, our results can also help to further understand subtle implications of the physical arrow of time that appear in real-world data.

We begin our presentation in Section 2, where we introduce our framework for determining time series direction, and specify the statistical tests used in our implementation (we note that other choices of tests would also yield valid algorithms). Section 3 provides a proof of Theorem 1, which is the main theoretical result of our document, and demonstrates that the direction of time is identifiable for an ARMA process with non-Gaussian noise. In Section 4 we provide results of experiments both with simulated and real (EEG) data. We compare with LiNGAM (Shimizu et al., 2006), which has been applied to test time series directionality (these earlier experiments were inconclusive). LiNGAM follows the principle of causal direction through noise dependence, but it infers the causal structure among n random variables from data generated by independent sampling from the same joint distribution. Given a single time series, we can artificially generate a statistical sample by cutting it into windows of equal length, but this does not take the full dependence structure into account: thus, while it can distinguish time direction in toy data, it provides ambiguous results on EEG data. By contrast, our approach performs well on both toy and EEG data.

2. Method

2.1. Model

A time series $(X_t)_{t\in\mathbb{Z}}$ is called stationary if the distribution of a random vector $(X_{t_1+h}, \ldots, X_{t_n+h})$ does not change for any value of h. It is called weakly stationary if the mean is constant: $\mathbf{E}X_t = \mu$ and the auto-covariance function only depends on the time gap: $\mathrm{cov}(X_t, X_{t+h}) = \gamma_h \, \forall t, h \in \mathbb{Z}$.

We call a time series $(X_t)_{t\in\mathbb{Z}}$ an *autoregressive moving average process of order* (p, q), if it is weakly stationary and there is an iid noise ϵ_t with mean zero, such that

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

For $q = 0$ the process reduces to an *autoregressive process* and for $p = 0$ to a *moving average process.* The short-hand notations are ARMA(p, q), AR(p) and MA(q).
Defining the backward shift operator B via $B^j X_t = X_{t-j}$ the ARMA equation simplifies to

$$\phi(B)X_t = \theta(B)\epsilon_t \qquad \forall t \in \mathbb{Z}\,, \tag{1}$$

with the polynomials $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$.

An ARMA process is called *causal* if the noise ϵ_t is independent of all preceding values of the time series $X_i\,, i < t$. There exist equivalent characterizations of causal time series:

Lemma 1 *For an ARMA(p, q) process satisfying $\phi(B)X_t = \theta(B)\epsilon_t$, where $\phi(z)$ and $\theta(z)$ have no common zeros, the following are equivalent:*[1]

(i) The process is causal.
(ii) There exists a sequence (ψ_i), such that $\sum_{i=0}^{\infty} |\psi_i| < \infty$ and

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}\,. \tag{2}$$

(iii) $\phi(z)$ does not have any zeros in the unit circle $|z| \leq 1$.

If this is the case, the coefficients ψ_i of (2) are determined by $\psi(z) = \sum_{i=0}^{\infty} \psi_i z^i = \frac{\theta(z)}{\phi(z)} \quad |z| \leq 1\,.$

$(i) \Leftrightarrow (ii)$ is easily checked, $(ii) \Leftrightarrow (iii)$ is proved by Brockwell & Davis (1991).

[1] Note that in (Brockwell & Davis, 1991) causal processes are actually defined as those satisfying condition (ii).

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 Section 3 we prove that causal ARMA processes are time-reversible if and only if the noise is Gaussian.

In their work (Weiss, 1975) and (Breidt & Davis, 1991) the authors call a strictly stationary process time-reversible if $(X_0, \ldots, X_h)$ and $(X_0, \ldots, X_{-h})$ equal in distribution for all h. Their theoretical result about the characterization of time-reversible processes is related to the Gaussian distribution, too. Diks et al. (1995) use this notion of time-reversibility to test whether a process is Gaussian. For our purposes, however, this notion is not appropriate because, a priori, both forward and backward process could be ARMA processes even though they do not coincide in distribution. Moreover, we do not want to restrict ourselves to strictly stationary processes.

From now on we assume the data to follow a causal ARMA process with non-Gaussian noise (model assumption).

Finally we remark that our definition of ARMA processes requires only *weak* stationarity at the cost of restricting the model to variables with finite variances. However, to include long-tailed distributions with infinite variance we will replace weak stationarity with strict stationarity. Brockwell & Davis (1991) show that then the expansion (2) is still valid. To ensure strict stationarity we consider Levy skew stable (or α-stable) distributed noise. As an interesting example of a non time-reversible process we will later use Cauchy distributed noise in the experiments.

2.2. Algorithm

The algorithm is built on the idea that non-Gaussian causal ARMA processes are not time-reversible. Thus we proceed as in Algorithm 1.

Here, "(res_a, a) independent" means that the residuals $res_a = (\epsilon_1, \ldots, \epsilon_n)$ are independent of the *preceding* time series values, id est ϵ_{t+1} is independent of x_t.

To implement this algorithm we need tests for normality and independence, and a method to fit ARMA processes. We now describe these components, bearing in mind that the method will also work with other choices (though possibly with altered performance).

Algorithm 1 Detecting true Time Direction

Input: $a = (x_1, \ldots, x_n)$, $b = (x_n, \ldots, x_1)$
$model_a$=armafit(a)
$res_a = model_a.residuals$
$model_b$=armafit(b)
$res_b = model_b.residuals$
if res_a normally distributed **then**
  output ="I do not know (Gaussian process)"
  break
end if
if (res_a, a) independent **then**
  if (res_b, b) dependent **then**
    output ="$(x_1, \ldots, x_n)$ correct time direction"
  end if
else if (res_a, a) dependent **then**
  if (res_b, b) independent **then**
    output ="$(x_n, \ldots, x_1)$ correct time direction"
  else if (res_b, b) dependent **then**
    output ="I do not know (bad fit)"
  end if
end if

- Normality Test
  We chose the Jarque-Bera test (Jarque & Bera, 1987), which is a simple test for normality based on skewness and kurtosis of the distribution.

- Independence Test
  Choosing a good independence test is important for the performance of our method. Since there is no obvious way to discretize the continuous data, standard tests (like χ^2) are not very well-suited for this method. In our implementation we used a statistical test of independence based on the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005; Smola et al., 2007; Gretton et al., 2008). This criterion estimates the distance between the joint distribution of two random variables and the product of their distributions in a Reproducing Kernel Hilbert Space. Depending on the choice of the kernel it can be shown that the HSIC is zero if and only if the two random variables are independent. The distribution of the HSIC itself under the hypothesis of independence converges (scaled with the number of data points) to a weighted sum of Chi-Squares, which can be estimated by a Gamma distribution (Kankainen, 1995).
  In our method we test if the residuals res_a are independent of the time series values a and if the residuals res_b are independent of b. This yields

two different p-values: p_a and p_b. If $p_a = 0.003$ and $p_b = 0.43$, for example, we reject the independence in the first case and we do not reject it in the second case. Thus we infer direction b to be the true one (see Figure 1). But how shall we decide if $p_a = 0.021$ and $p_b = 0.033$? To this end we introduce two parameters to our method (which would be needed for any other independence test): the minimal difference in p-values δ, and a significance level α. We only decide if the difference in p-values is larger than δ and if exactly one p-value is below α. If we chose a large value for δ and a small value for α our method makes fewer decisions, but also fewer mistakes.
Technically we need iid data of the residuals ϵ and the time series values X_t to perform the independence test. Clearly, the time series values $(x_1, \ldots, x_n)$ are not iid. If the ARMA process is strictly stationary, the values are identically distributed. But it lies in the nature of time series models that $(x_1, \ldots, x_n)$ are not independent. This dependence within the values of x_t, however, does not influence the independence test much, since the test is looking for a dependence between the joint samples of (res and x_t). To reduce the effect further we also introduce a gap (3 in all experiments) between the values of the time series being considered.

- Fitting an ARMA process
  We use the program R (RProject, 2009) to fit an ARMA process to the data. R computes a Gaussian Likelihood via a state-space representation of the ARMA process and uses a Kalman filter to find the innovations. We use an information criterion (AICC) to select the order of the process. In the experiments we fitted ARMA processes of order up to (5,5).

  We are aware of the ambiguity between the need for a non-Gaussian distributed process for our method to work and the assumption of a normal distribution for the fit at the same time. It can be shown, however, (see Chapter 9.3 in (Brockwell & Davis, 1991)) that fitting a model to non-Gaussian ARMA processes using a Gaussian Likelihood and AICC still leads to good results. It is possible to avoid the normality assumption by performing a fit using the innovations algorithm (see Chapter 8.4 in (Brockwell & Davis, 1991)).

3. Time Identifiability

This section justifies the key procedure of our method. As the main theoretical result of this work we prove

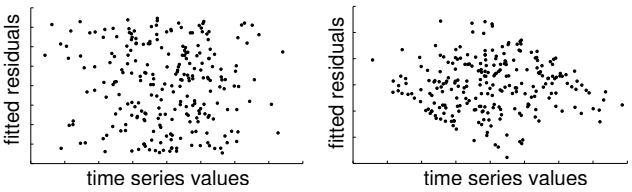

Figure 1. Simulated AR(1) process with uniformly distributed noise: The fitted residuals of the forward model (left) and of the backward model (right) are plotted against the preceding time series values. The fit in the wrong direction leads to a strong dependence between residuals and time series (p-value of 0.0001), the residuals of the forward model are regarded as independent (p-value of 0.6744).

that if a time series follows a causal ARMA model with non-Gaussian noise, it is not time-reversible.

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

Proof. For the *only if* part of this proof we need a characterization of the normal distribution that is based on Mamai and Skitovich (Darmois, 1953; Skitovic, 1962). Together with a technical detail of the proof this characterization is provided in the appendix. First we prove the easy direction of the theorem:

$\Leftarrow$: A Gaussian distribution is characterized by its mean and its covariance matrix, meaning weak and strict stationarity coincide. For an ARMA process with Gaussian noise we have $(X_{t+p}, \ldots, X_t) \stackrel{d}{=} (X_{-t-p}, \ldots, X_{-t})$ because the covariance matrix is symmetric. If for the forward direction an iid sequence (ϵ_t) exists such that (X_t) is a causal ARMA process, it is clear that there is a (different) iid sequence $(\tilde{\epsilon}_t)$ with the same properties with respect to the reversed direction.

$\Rightarrow$: By assumption, we have

$$X_t = \sum_{i=1}^{\tilde{p}} \tilde{\phi}_i X_{t+i} + \sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t \qquad \forall t \in \mathbb{Z}\,.$$

Thus using (2) we can write

$$\begin{aligned}\sum_{j=1}^{\tilde{q}} \tilde{\theta}_j \tilde{\epsilon}_{t-\tilde{p}+j} + \tilde{\epsilon}_{t-\tilde{p}} &= X_{t-\tilde{p}} - \sum_{j=1}^{\tilde{p}} \tilde{\phi}_j X_{t-\tilde{p}+j} \\ &= \sum_{i=0}^{\infty} \left(\psi_{i-\tilde{p}} - \sum_{j=1}^{\tilde{p}} \tilde{\phi}_j \psi_{i+j-\tilde{p}} \right) \epsilon_{t-i}\,,\end{aligned}$$

where $\psi_i = 0$ for all $i < 0$. Additionally we have

$$X_{t-\tilde{p}+\tilde{q}+1} = \sum_{i=0}^{\infty} \psi_i \, \epsilon_{t-\tilde{p}+\tilde{q}+1-i} = \sum_{i=\tilde{q}-\tilde{p}+1}^{\infty} \psi_{\tilde{p}-\tilde{q}-1+i} \, \epsilon_{t-i} \, .$$

Both sums are converging absolutely with probability one (see (Brockwell & Davis, 1991)) and by assumption, the left hand sides are independent of each other. Now we can apply Lemma 2, which is a generalization of the Darmois-Skitovich theorem and stated in the Appendix. Lemma 3 shows that the boundedness conditions from Lemma 2 are satisfied. We can therefore conclude that the noise ϵ_t is Gaussian distributed (note that some ϵ_t occur on both sides because of the non-vanishing AR part). But then X_t is Gaussian distributed, too: Define $X_t^{(n)} := \sum_{i=1}^{n} \psi_i \epsilon_{t-i}$. Since $(X_t^{(n)})_n$ is converging in $\mathcal{L}^2$ and in distribution, X_t is Gaussian distributed. □

Note that we excluded all MA processes. This is necessary because an MA(1) process with coefficient $\theta_1 = 1$, for example, can be reversed for all noise distributions. In the following section the experiments with simulated data underline the necessity of non-Gaussian noise.

4. Experiments

We applied our method to simulated and real data. Recall that the method does not apply to processes with Gaussian noise, for which no "true" direction exists. Normally distributed noise 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. (Mandelbrot, 1967), also (Pearlmutter & Parra, 1997)). In the finance sector especially, heavy-tailed distributions can explain the occurrence of extreme events better than the normal distribution.

We compare our approach to LiNGAM (Shimizu et al., 2006), which has been proposed for the task of detecting the direction of real world time series. Using iid samples of a random vector $(X_1, \ldots, X_m)$, LiNGAM produces a causal graph with possible links between the random variables. Given a time series the data is separated into time windows of a length that equals the number p of included variables (e.g. $X_1, \ldots, X_5$). The different windows are then treated as samples of these variables ($x_1, x_6, x_{11}, \ldots$ are all assigned to X_1). If the resulting graph of the included variables is time consistent in the sense that all causal links go from lower to higher labelled variables (or vice versa), this direction is proposed to be the true one. Note that even if the data comes from an autoregressive process the problem of possibly confounding can cause difficulties for the algorithm. Another drawback of this approach is that there is no canonical way to choose the number p of included variables. In their experiments the authors used 3 or 5 variables (Shimizu et al., 2006). For 14 out of 22 time series LiNGAM proposed a direction (in the other cases the result was inconsistent); 5 out of these 14 decisions were correct. While we obtained somewhat more promising results for LiNGAM on our artificial data, our experience with LiNGAM on the real data set was not convincing.

Simulated Data. In order to support the theoretical result given above, we simulated many instances of an ARMA(2,2) process, each time with the same fixed parameters. Additionally we used varying noise distributions. For r ranging between 0.1 and 2 we sampled $\epsilon_t \sim \mathrm{sgn}(Z) \cdot |Z|^r$, where Z is normally distributed, and scaled it to variance 1. Only $r = 1$ results in a normal distribution. We expect our method to work for r sufficiently different from 1. In this experiment we did not perform the normality test (line 6-9 in the algorithm) and we set $\delta = 0.05$ and $\alpha = 0.96$; changing these values does not have a big influence on the results. The left panel of Figure 2 shows results. Note that the indistinguishability for the Gaussian case is not a problem of our method, but due to the theoretical result given by Theorem 1.
For the right panel of Figure 2 we used a χ^2 independence test instead of the HSIC independence test. The discretization for the χ^2 was chosen such that there was at least one data point in every bin. It performed worse than the HSIC. The same phenomenon occurs when comparing the two independence tests on real data.

We further simulated AR(p) processes of different orders ($p = 1, \ldots, 5$) with randomly chosen coefficients. Given five distributions for the noise (Gaussian, Laplace, Cauchy, Student-t and uniform) we used our method to predict the true time direction. Again, the noise was simulated such that the variance was one, except for the Cauchy distribution. For each order/noise combination we received the proportion of correctly and wrongly classified time series (both out of 100). See Figure 3 for details. The method works well on all distributions except for the Gaussian. Note that it works best for the Cauchy distribution, which is an example of an α stable distribution with infi-

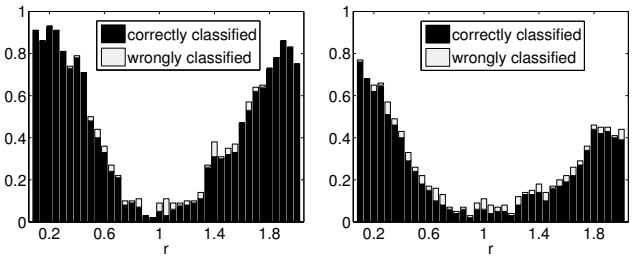

Figure 2. For each value of r (i.e. for each kind of noise) we simulated 100 instances of an ARMA(2,2) process. The graphs show the proportion of correctly and wrongly classified time series depending on r. Because Gaussian ARMA processes are symmetric in time, the method has to fail for $r \approx 1$. The left panel shows our method using the HSIC independence test, for the right one we used a χ^2 independence test instead, which results in a worse performance.

nite variance. For higher order processes estimating the parameters of the true model becomes more difficult. Thus we sometimes find dependencies between the noise and the values of the time series even in the true direction. This already shows possible difficulties in dealing with real data.

We also included a comparison with LiNGAM (see Figure 4). Although the authors of (Shimizu et al., 2006) did not mention it, the method clearly works for simulated AR processes. Since there is no way of choosing the number of included variables automatically, we had to set it manually to 4 and 7. For some orders the performance is comparable to our method, for others it is worse.

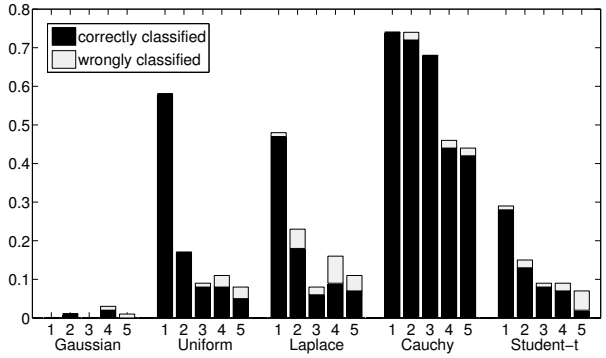

Figure 3. The histogram shows the proportion of correctly and wrongly classified time series (out of 100). The parameters are chosen as before: minimal difference in p-values $\delta = 5\%$, significance level $\alpha = 4\%$. A change in the parameters leads respectively to slightly more decisions losing accuracy or less decisions gaining accuracy. In most cases (except the Gaussian) the correct classification rate significantly exceeds 50%.

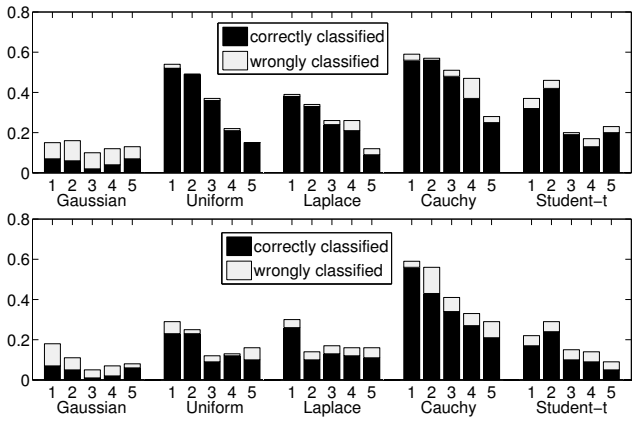

Figure 4. Same experiment as in Figure 3, but using LiNGAM. The performance is comparable to our method for $p = 4$, but worse for $p = 7$.

Real Data. In order to show that our method is applicable to real data, we used a publicly available EEG data set (EEGdata, 2008) consisting of 118 channels of a single subject. The sampling rate was 1000Hz and we considered the first 5 seconds of each channel, cut into 10 pieces. In total this gave 1180 time series of length 500. The results of our method for different values of α and δ are shown in Figure 5. As α shrinks and δ grows, the algorithm makes fewer mistakes, but also classifies fewer time series. That said, classification accuracy consistently exceeds 50%. By comparison, the performance of LiNGAM is provided in Table 1, and is worse than our approach: indeed, it does not always exceed chance level.

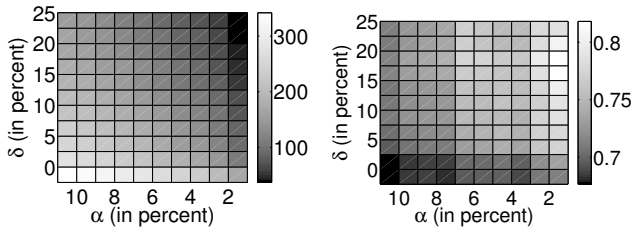

Figure 5. EEG data: The left panel shows the number of classified time series (out of 1180), and the right panel the proportion of correctly classified time series, depending on the parameters α and δ. The results are consistently better than chance, reaching a correct classification rate of up to 82%.

Note that our method is not restricted to EEG data. In (Peters et al., 2009) we did an experiment on a mixed collection of 200 time series from different areas. Our method also exceeded chance level on these data.

Table 1. Same experiment as in Figure 5, this time using LiNGAM for detecting the true time direction. The columns show the number of included variables p, the number of classified time series and the proportion of those that were correctly classified. The performance is worse than our method, even chance level is not always exceeded.

INCLUDED VARIABLES (p)	CLASSIFIED TIME SERIES	COR. CLASSIFIED TIME SERIES
2	1100	0.4636
3	451	0.6563
4	102	0.5294
5	133	0.5489
6	109	0.4954
7	242	0.5372
8	200	0.5150
9	130	0.5538
10	114	0.4561
11	116	0.5603

5. Discussion and Future Work

In this work we have shown that causal inference based methods are able to detect the true time direction in time series. We have introduced a method which assumes the data to follow an ARMA process with independent non-Gaussian noise. We proved its ability to work in the population case and showed that it works well on simulated data. For real world data, our algorithm proposes a direction for a significant fraction of the time series, most of which are correct.

We argued that our method is closely related to causal inference methods like LiNGAM, which in principle can also be used for detecting the time direction. We showed that LiNGAM works on simulated data, but performs worse than our method.

In the experiments our method did not propose a time direction for many time series, because the fit led to dependent noise in both directions. This suggests that the model class of causal ARMA processes may still be too small. Generalizing our method and the theoretical results to other time series models like GARCH, which allows heteroscedastic noise, or to nonlinear models, may further improve our method. It is also worth investigating if we can change the fit of an ARMA process. Instead of minimizing a loss function related to the likelihood or the sum of squares, we should use the instance of an ARMA model that leads to the "most independent" noise. These are important topics for future work.

A. Characterizing the Normal Distribution

The Darmois-Skitovich theorem ((Darmois, 1953),(Skitovic, 1962)) states that if two non-trivial linear combinations of independent random variables are themselves independent then all summands are normally distributed. This fact can be used to prove the AR(1) case, for example. It turns out, however, that the Darmois-Skitovic can be generalized to an infinite sum, which we need for our purposes. This was first done by Mamai (Mamai, 1963):

Lemma 2 *Let (X_t) be a sequence of independent random variables and assume that both $\sum_{i=1}^{\infty} a_i X_i$ and $\sum_{i=1}^{\infty} b_i X_i$ converge almost surely. Further suppose that the sequences $\{\frac{a_i}{b_i} : b_i \neq 0\}$ and $\{\frac{b_i}{a_i} : a_i \neq 0\}$ are bounded. If*

$$\sum_{i=1}^{\infty} a_i X_i \qquad \text{and} \qquad \sum_{i=1}^{\infty} b_i X_i$$

are independent, then each X_i for which $a_i b_i \neq 0$ is normally distributed.

B. Boundedness Condition

Lemma 3 *For all possible causal backward models ARMA($\tilde{p}, \tilde{q}$) both*

$$\left| \frac{\psi_{\tilde{p}-\tilde{q}-1+i}}{\sum_{j=0}^{\tilde{p}} c_j \psi_{i+j-\tilde{p}}} \right| \qquad \text{and} \qquad \left| \frac{\sum_{j=0}^{\tilde{p}} c_j \psi_{i+j-\tilde{p}}}{\psi_{\tilde{p}-\tilde{q}-1+i}} \right| \tag{3}$$

are bounded in i (see (2) for the coefficients ψ_i).

Here, $c_1 := -\tilde{\phi}_1, \ldots, c_{\tilde{p}} := -\tilde{\phi}_{\tilde{p}} \in \mathbb{R}$ and $c_0 := 1$.

Proof.

We have the following expression for ψ_i (see Chapter 3.3 in (Brockwell & Davis, 1991)):

$$\psi_i = \sum_{s=1}^{S} \sum_{t=1}^{T_s - 1} \alpha_{s,t}\, i^t\, \xi_s^{-i},$$

where $\alpha_{s,t}$ are some coefficients, ξ_s are the distinct (possibly complex) roots of $\phi(z)$ and T_s their multiplicity. Wlog assume that $\alpha_{s,T_s-1} \neq 0 \ \forall s$. We can write the left fraction of (3) as

$$\frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, (\tilde{p}-\tilde{q}-1+i)^t\, \xi_s^{-\tilde{p}+\tilde{q}+1-i}}{\sum_{j=0}^{\tilde{p}} c_j \ \sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, (i+j-\tilde{p})^t\, \xi_s^{-i-j+\tilde{p}}}$$

$$= \frac{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, \xi_s^{-\tilde{p}+\tilde{q}+1} (\tilde{p}-\tilde{q}-1+i)^t\, \xi_s^{-i}}{\sum_{s=1}^{S} \sum_{t=1}^{T_s-1} \alpha_{s,t}\, \xi_s^{\tilde{p}} \sum_{j=0}^{\tilde{p}} c_j\, \xi_s^{-j} (i+j-\tilde{p})^t\, \xi_s^{-i}}. \tag{4}$$

To investigate the limit behaviour we again consider only leading terms in i. More specifically, all summands are going to zero since $|\xi_s^{-1}| < 1$. The root ξ_{s_0} with the smallest modulus converges towards zero with the slowest rate and thus the corresponding summand determines the overall convergence. We divide both numerator and denominator of (4) by $i^{T_{s_0}-1}\,\xi_{s_0}^{-i}$ to see that the fraction converges towards

$$\left| \frac{\alpha_{s_0,T_{s_0}-1}\,\xi_{s_0}^{-\tilde{p}+\tilde{q}+1}}{\alpha_{s_0,T_{s_0}-1}\,\xi_{s_0}^{\tilde{p}}\sum_{j=0}^{\tilde{p}} c_j\,\xi_{s_0}^{-j}} \right|$$

for $i \to \infty$. This surely implies boundedness.
Note that the coefficient

$$\alpha_{s_0,T_{s_0}-1}\,\xi_{s_0}^{\tilde{p}} \sum_{j=0}^{\tilde{p}} c_j\,\xi_{s_0}^{-j}$$

does not vanish because this implies $\sum_{j=0}^{\tilde{p}} c_j\,\xi_{s_0}^{-j} = \tilde{\phi}(\xi_{s_0}^{-1}) = 0$. That means $\xi_{s_0}^{-1}$ is a root of $\tilde{\phi}(z)$, which is contrary to the restriction of a causal backward model ($|\xi_{s_0}| > 1$, cf Lemma 1). □

References

Balian, R. (1992). *From microphysics to macrophysics*. Springer.

Breidt, F. J., & Davis, R. A. (1991). Time-reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis*, *13*(5), 379–390.

Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer, 2 edn.

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Rev. Inst.Internationale Statist.*, *21*, 2–8.

Diks, C., van Houwelingen, J., Takens, F., & DeGoede, J. (1995). Reversibility as a criterion for discriminating time series. *Physics Letters A*, *15*, 221–228.

EEGdata (2008). This data set (experiment 4a, subject 3) can be downloaded after registration. Website, 3.6.2008, 3:51pm. `http://ida.first.fraunhofer.de/projects/bci/competition_iii/#datasets`.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*, 63–78. Springer-Verlag.

Gretton, A., Fukumizu, K., Teo, C.-H., Song, L., Schölkopf, B., & Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, 585–592. Cambridge, MA: MIT Press.

Hawking, S. (1995). *A brief history of time*. Bantam.

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. *Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (NIPS)*.

Janzing, D. (2007). On causally asymmetric versions of Occam's Razor and their relation to thermodynamics. `http://arxiv.org/abs/0708.3411`.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, *55*(2), 163–172.

Kankainen, A. (1995). Consistent testing of total independence based on the empirical characteristic function. *PhD Thesis, University of Jyväskylä*.

Mamai, L. V. (1963). On the theory of characteristic functions. *Select. Transl. Math. Stat. Probab.*, *4*, 153–170.

Mandelbrot, B. (1967). On the distribution of stock price differences. *Operations Research*, *15*(6), 1057–1062.

Pearlmutter, B. A., & Parra, L. C. (1997). Maximum likelihood blind source separation: A context-sensitive generalization of ica. In *Advances in Neural Information Processing Systems 9*, 613–619. MIT Press.

Peters, J., Janzing, D., Gretton, A., & Schölkopf, B. (2009). Kernel methods for detecting the direction of time series. *In: Proccedings of the 32nd Annual Conference of the German Classification Society (GfKl 2008)*, 1–10.

Reichenbach, H. (1999). *The direction of time*. Dover.

RProject (2009). The r project for statistical computing. Website, 15.1.2009, 1:07pm. `http://www.r-project.org/`.

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

Skitovic, V. P. (1962). Linear combinations of independent random variables and the normal distribution law. *Select. Transl. Math. Stat. Probab.*, *2*, 211–228.

Smola, A. J., Gretton, A., Song, L., & Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference (ALT)*, 13–31. Springer-Verlag.

Weiss, G. (1975). Time-reversibility of linear stochastic processes. *J. Appl. Prob.*, *12*, 831–836.