Gaussianity Measures for Detecting the Direction of Causal Time Series*

José Miguel Hernández-Lobato, Pablo Morales-Mombiela, Alberto Suárez
Universidad Autónoma de Madrid, Madrid, Spain
{josemiguel.hernandez, alberto.suarez}@uam.es
pablo.morales@estudiante.uam.es

Abstract

We conjecture that the distribution of the time-reversed residuals of a causal linear process is closer to a Gaussian than the distribution of the noise used to generate the process in the forward direction. This property is demonstrated for causal AR(1) processes assuming that all the cumulants of the distribution of the noise are defined. Based on this observation, it is possible to design a decision rule for detecting the direction of time series that can be described as linear processes: The correct direction (forward in time) is the one in which the residuals from a linear fit to the time series are *less* Gaussian. A series of experiments with simulated and real-world data illustrate the superior results of the proposed rule when compared with other state-of-the-art methods based on independence tests.

1 Introduction

Identifying the true direction of a causal time series is an interesting problem for the evaluation of novel methods for causal inference [Peters *et al.*, 2009]. Furthermore, solving this problem can provide new insights about the asymmetries between past and future in physical systems [Janzing, 2010].

We conjecture that, when a linear process is appropriate for modeling a time series with non-Gaussian innovations, the residuals of a linear fit to the time-reversed series are more Gaussian than the residuals in the chronologically (forward in time) ordered series. This property can be derived for series that follow a causal AR process, assuming that the noise is non-Gaussian and that the cumulants of the distribution of the noise are defined. Based on this property it is possible to use Gaussianity measures to detect the true direction of a time series. The proposed method works by fitting a linear model to the time series in the original ordering and in the inverted ordering and then selecting the direction in which the corresponding residuals are less Gaussian.

*This research has been supported by the Spanish Dirección General de Investigación under project TIN2010-21575-C02-02. AS acknowledges partial financial support from Shizuoka University (Japan).

Different metrics can be used to measure the distance of the empirical distribution of the residuals to the Gaussian. A simple method is to use an estimate of a cumulant of order higher than two to make the decision. More sophisticated metrics, such as the Maximum Mean Discrepancy (MMD) have been recently proposed [Gretton *et al.*, 2007]. The performance of the novel causal inference rules is evaluated in experiments with data simulated from AR and ARMA processes and with real-world data formed by time series of EEG measurements of brain activity. In these experiments, the proposed approach outperforms state-of-the-art decision methods that work by quantifying the independence of the empirical residuals with respect to the previous time series values [Peters *et al.*, 2009].

2 Reversibility in Time Series

The time series $\{X_t\}_{t=-\infty}^{\infty}$ is said to be reversible when the vectors $\{X_{t_1}, X_{t_2}, \ldots X_{t_n}\}$ and $\{X_{-t_1}, X_{-t_2}, \ldots X_{-t_n}\}$ have the same joint distribution for all $n > 0$ and t_i, where $1 \leq i \leq n$ [Lawrance, 1991]. Identifying the true direction of a time series is only possible when the series is not reversible.

Consider the autoregressive-moving average linear process

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \,, \qquad (1)$$

where $\{\epsilon_t\}$ is i.d. white noise [Brockwell and Davis, 1991]. The previous definition of reversibility implies stationarity [Lawrance, 1991]. If the distribution of ϵ_t is Gaussian then $\{X_t\}$ is time-reversible. This can be derived using the fact that, when the noise is Gaussian, the joint distribution of $\{X_{t_1}, X_{t_2}, \ldots X_{t_n}\}$ for any $n > 0$ and t_i, where $1 \leq i \leq n$, is multivariate Gaussian. The symmetry of the covariance matrix guarantees that the process is strictly stationary and time-symmetric. Weiss showed that if an ARMA process[1] is reversible, then $\{\epsilon_t\}$ is normally distributed [Weiss, 1975].

An alternative definition of time reversibility not restricted to strictly stationary processes is proposed in [Peters *et al.*, 2009; 2010]. Consider a causal ARMA process, in which ϵ_t is independent of $X_{t-k} \;\; \forall k > 0, \forall t \in \mathbb{Z}$. The causal ARMA process is said to be reversible when there is an i.i.d. sequence $\{\tilde{\epsilon}_t\}$ such that $X_t = \sum_{i=1}^{p} \phi_i X_{t+i} + \sum_{j=1}^{q} \theta_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t$, with

[1] The process must satisfy $p > 0$ or, if $p = 0$, the cases $\theta_n = \theta_{q-n}$ and $\theta_n = -\theta_{q-n}$ are excluded.

$\{\tilde{\epsilon}_t\}$ being independent of $X_{t+k},\ \forall k > 0, \forall t \in \mathbb{Z}$. They then show, using the Darmois-Skitovich theorem, that a causal ARMA process with i.i.d. noise and $p > 0$ is time-reversible if and only if the noise is normally distributed. Finally, Peters et al. propose to identify the true direction of a time series by fitting an ARMA model and testing whether the residuals of the model are independent of the previous time series values in one direction and dependent in the opposite one.

Our approach to the problem of directionality detection is different. Assuming a linear time-series model, we show that the residuals of the time-reversed series are more Gaussian than the residuals of the original series. Therefore, once we have obtained the two series of empirical residuals, we can use measures of Gaussianity to identify the correct direction.

3 Main Result

To simplify the presentation, the main result is derived for a stationary causal AR(1) process

$$X_t = \phi X_{t-1} + \epsilon_t\,, \quad |\phi| < 1\,,$$

$\{\epsilon_t\}$ is an i.d. white noise process, which is not necessarily Gaussian. The noise is independent of the delayed values of the process: $\epsilon_t \perp X_{t-k},\ \forall k > 0,\ \forall t \in \mathbb{Z}$. The condition of stationarity is fulfilled if $|\phi| < 1$. The case $\phi = 0$ is trivially time-reversible ($X_t = \epsilon_t$) and will not be considered further.

If the process is reversible, it can be expressed as

$$X_t = \phi X_{t+1} + \tilde{\epsilon}_t\,,$$

where $\{\tilde{\epsilon}_t\}$ is a white noise sequence and $\tilde{\epsilon}_t \perp X_{t+k}$, $\forall k > 0,\ \forall t \in \mathbb{Z}$. The coefficient ϕ is the same as in the forward equation because it is the one-lag autocorrelation $\phi = cov(X_t X_{t-1})/var(X_t) = cov(X_t X_{t+1})/var(X_t)$, which is symmetric in time.

If the process is not reversible, we define the time-reversed residuals as

$$\tilde{\epsilon}_t \equiv X_t - \phi X_{t+1}\,.$$

In this case, $\{\tilde{\epsilon}_t\}$ is not a white noise sequence and $\tilde{\epsilon}_t$ is in fact dependent of X_{t+k}, for some $k > 0,\ \forall t \in \mathbb{Z}$.

Assuming that the cumulants of the white noise process exist, the cumulants of $\tilde{\epsilon}_t$ can be expressed in terms of the cumulants of ϵ_t:

$$\begin{aligned} \kappa_n(\tilde{\epsilon}_t) &= c_n(\phi)\kappa_n(\epsilon_t), \quad n > 0 \\ c_n(\phi) &= (-\phi)^n + (1-\phi^2)^n(1-\phi^n)^{-1} \end{aligned} \qquad (2)$$

The derivation of this expression can be found in Appendix A. Figure 1 shows plots of $c_n(\phi)$ as a function of ϕ for $n = 1, \ldots, 14$. The following result can be stated for causal AR(1) processes with $\phi \neq 0$ and $|\phi| < 1$:

Theorem 3.1. *The cumulants of $\tilde{\epsilon}_t$ of order higher than 2 are smaller in magnitude than the cumulants of ϵ_t, if these exist.*

Proof. Using (2), and the fact that $|c_n(\phi)| < 1, \forall n > 2$, we obtain $|\kappa_n(\tilde{\epsilon}_t)| < |\kappa_n(\epsilon_t)|, \forall n > 2$. □

For a Gaussian distribution, the cumulants of order higher than two are zero. Therefore, Theorem 3.1 implies that the cumulants of $\tilde{\epsilon}_t$ are closer to the cumulants of a Gaussian than the cumulants of ϵ_t. In this precise sense, the distribution of

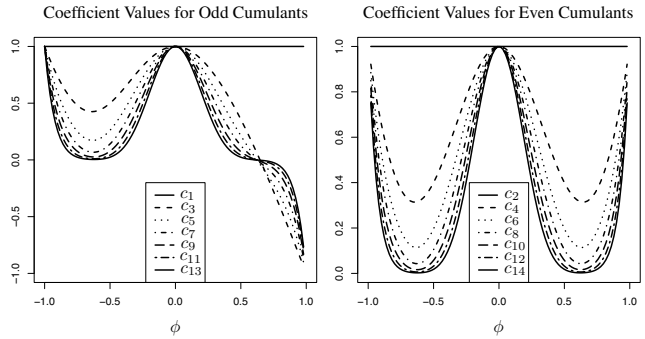

Figure 1: Coefficients $c_n(\phi)$ (odd in the left plot, even in the right plot) in a causal AR(1) process.

$\tilde{\epsilon}_t$ is closer to a Gaussian than the distribution of ϵ_t. This Gaussianization effect can be observed even when the higher order cumulants diverge. In fact, it obtains even when none of the cumulants are defined, as in the Cauchy distribution. This effect is stronger for AR(1) processes in which $|\phi|$ is in the vicinity of the golden ratio conjugate: $\phi \approx \pm(\sqrt{5}-1)/2$ (see appendix A and the plots in Figure 1).

Theorem 3.1 allows us to determine the correct direction of the series $\{X_t\}_{t=1}^T$ by measuring distances to the Gaussian of the distributions of the residuals: Assuming that the AR(1) model is correct, one performs a fit to the series and calculates the empirical residuals in the original ordering, $\{e_t\}_{t=2}^T$, and in the inverted ordering of the series, $\{\tilde{e}_t\}_{t=1}^{T-1}$, where

$$e_t = X_t - \hat{\phi} X_{t-1}\,, \qquad \tilde{e}_{t-1} = X_{t-1} - \hat{\phi} X_t\,,$$

$t = 2, \ldots, T$ and $\hat{\phi}$ is the empirical correlation between X_t and X_{t-1}. Note the difference between $\{e_t\}$, the empirical residuals, and $\{\epsilon_t\}$, the i.d. white noise process that was used to generate the time-series values. The correct direction of the series is the one in which the corresponding empirical residuals are *less* Gaussian.

Theorem 3.1 can be readily generalized to causal AR(p) processes and is expected to hold also for causal ARMA and vector AR processes. An analysis of simulated and real-world data provides strong evidence for this.

4 Gaussianity Measures

We consider two measures of Gaussianity for detecting the direction of causal time series that follow a linear model. The first one is directly based on the inequality given by Theorem 3.1 for a particular order of the cumulants. We focus on the fourth order cumulant because κ_4 is the cumulant with lowest order which is non-zero for symmetric non-Gaussian noise. In leptokurtic distributions, the estimates of κ_4 have lower variance than the estimates of cumulants of higher order.

Given a sample $\{Y_i\}_{i=1}^N$, the unbiased estimator of κ_4 with lowest variance is the fourth k-statistic [Kendall *et al.*, 1994]:

$$\begin{aligned} k_4 = \frac{1}{N^{[4]}} \big[&(N^3+N^2)S_4 - 4(N^2+N)S_3 S_1 \\ &-3(N^2-N)S_2^2 + 12N S_2 S_1^2 - 6S_1^4\big]\,, \end{aligned} \qquad (3)$$

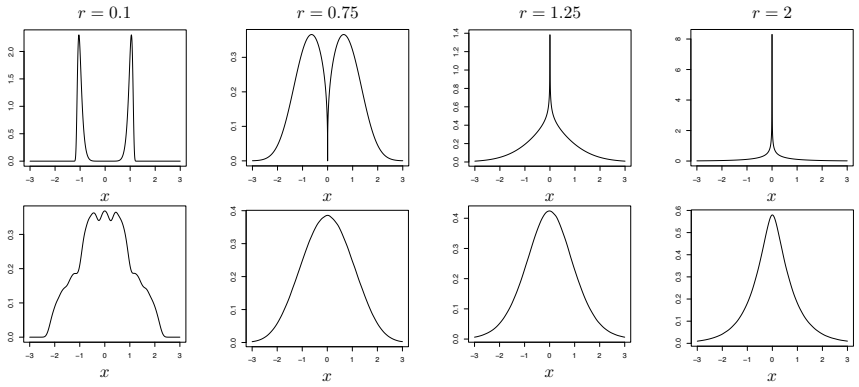

Figure 2: Density for the noise (top row) and the reversed residuals (bottom row) in an AR(1) model with $\phi = (\sqrt{5}-1)/2$.

where $N^{[4]} = N(N-1)(N-2)(N-3)$ and $S_1, \ldots, S_4$ are the first 4 power sums, that is, $S_r = \sum_{i=1}^{N} Y_i^r$. The inference rule based on κ_4 selects the direction for which the empirical residuals have lower values of (3).

The deviations from Gaussianity may also appear at higher order cumulants. In fact, it is possible to build non-Gaussian random variables with a finite number of zero cumulants, including κ_4. Therefore, properly designed metrics on the space of probability distributions should be a better way to quantify the deviations of the residuals from Gaussianity and to provide a more effective method for detecting the direction of the time series. We focus on metrics based on a Hilbert space embedding for probability distributions [Gretton *et al.*, 2007; Sriperumbudur *et al.*, 2010]. Let p and q denote two density functions defined on $\mathbb{R}$ and let $\mathcal{F}$ be the unit ball in a universal reproducing kernel Hilbert space $\mathcal{H}$ defined on $\mathbb{R}$ with kernel $k(\cdot,\cdot)$. The Maximum Mean Discrepancy (MMD) between p and q within the set of functions $\mathcal{F}$ is

$$\sup_{f \in \mathcal{F}} \left(\mathbf{E}_{y\sim p}[f(y)] - \mathbf{E}_{z\sim q}[f(z)]\right) = \|\mu_p - \mu_q\|_{\mathcal{H}} \,,$$

where $\mu_p = \mathbf{E}_{y\sim p}[k(y,\cdot)]$ and $\mu_q = \mathbf{E}_{z\sim q}[k(z,\cdot)]$ are the corresponding mappings of p and q onto $\mathcal{H}$ and $\|\cdot\|_{\mathcal{H}}$ denotes the norm operator in $\mathcal{H}$ [Smola *et al.*, 2007]. MMD vanishes when $p = q$ and is positive otherwise [Gretton *et al.*, 2007].

Given a sample $\{Y_i\}_{i=1}^N$ with zero-mean and unit standard deviation, we can compute the MMD statistic between the standard Gaussian distribution and the empirical distribution of the sample. For this, we define

$$p(y) = \frac{1}{n}\sum_{i=1}^{n} \delta(y - Y_i)\,, \quad q(z) = (2\pi)^{-\frac{1}{2}} \exp\left(-0.5z^2\right)\,,$$

where $\delta(y)$ represents a point probability mass at $y = 0$, and the target function q is the $\mathcal{N}(0,1)$ density. Using a Gaussian kernel $k(y,z) = \exp[-\frac{1}{2\sigma^2}(y-z)^2]$ of width σ, we obtain the square of the MMD statistic:

$$\begin{aligned}\|\mu_p - \mu_q\|_{\mathcal{H}}^2 &= \langle \mu_p, \mu_p\rangle_{\mathcal{H}} + \langle \mu_q, \mu_q\rangle_{\mathcal{H}} - 2\langle \mu_p, \mu_q\rangle_{\mathcal{H}} \\ &= \frac{1}{N^2}\sum_{i,j}^{N} k(Y_i, Y_j) + \frac{\sigma}{\sqrt{2+\sigma^2}} - \frac{2}{N}\sum_{i=1}^{N} k'(Y_i)\,, \quad (4)\end{aligned}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ denotes the dot product operator in $\mathcal{H}$ and $k'(y) = \frac{\sigma}{\sqrt{1+\sigma^2}} \exp[-\frac{1}{2(1+\sigma^2)}y^2]$. The MMD decision rule selects the direction of the time series in which the value of (4) for the empirical residuals is lower.

5 Related Work

PC-LINGAM [Hoyer *et al.*, 2008] combines conditional independence methods with ICA based approaches for causal inference. This method discriminates between models in the same d-separation-equivalence class by selecting the model that maximizes the ICA target function

$$\left(\mathbf{E}_{y\sim p}\left\{|y|\right\} - \mathbf{E}_{z\sim\mathcal{N}(0,1)}\left\{|z|\right\}\right)^2 \,,$$

where p is the empirical distribution of the standardized model residuals (zero mean and unit standard deviation). This target function is a specific case of the MMD distance in which the set $\mathcal{F}$ contains only the absolute value function $|\cdot|$.

A related method is described in [Daniušis *et al.*, 2010]. This technique employs Gaussianity measures for inferring causal directions between two random variables. The variable whose marginal distribution is closer to a Gaussian is selected as the cause and the other one as the effect. The distance to the Gaussian is computed using an empirical estimate of the entropy. This method cannot be directly applied to time series because, if the analyzed process is stationary, both cause (X_t) and effect (X_{t+1}) have the same distribution.

In [Zhang and Hyvärinen, 2009], a class of acyclic causal models are investigated. In these models, the causal relations among the observed variables are nonlinear but the effect of the disturbances is linear. These authors show that, when the disturbances are additive, the correct causal model is the one with the lowest entropy for the disturbances. This agrees with the results of our investigation: For linear models, (2) implies that the differential entropy of the forward residuals is lower than that of the reversed residuals. However, in the problems investigated, entropy-based measures of Gaussianity, such as those used in [Daniušis *et al.*, 2010], achieve lower accuracies than the proposed methods (results not shown).

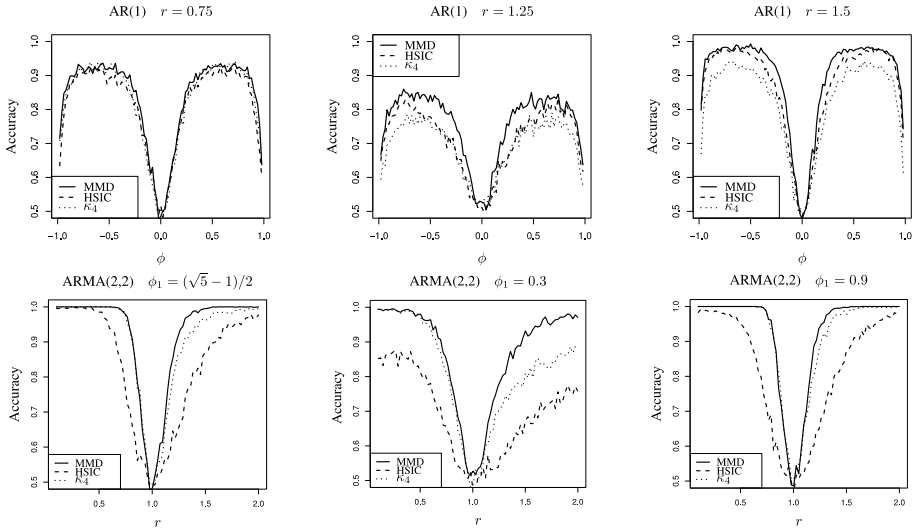

Figure 3: Results obtained by each method in the experiments with AR(1) and ARMA(2,2) time series.

6 Experiments

The performance of the decision rules based on MMD and κ_4 is investigated in experiments with simulated and real-world data. The statistical test described in [Peters *et al.*, 2009] is used as a benchmark for comparison. This method is based on determining whether the empirical residuals are independent of the previous time series values in one direction and dependent in the opposite direction. The degree of dependence is measured using the p-value given by a statistical test based on the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton *et al.*, 2008]. The smaller this p-value is, the larger the level of dependence in the data. The direction predicted by the HSIC test is the one for which the dependencies between the model residuals and the previous time series values up to a fixed lag γ are weaker. For example, when $\gamma = 2$ the HSIC method checks for (i) dependencies between e_t and X_{t-1}, X_{t-2} and (ii) dependencies between $\tilde{e}_t$ and X_{t+1}, X_{t+2}. The input to each decision rule (MMD, κ_4 and HSIC) are the empirical residuals and the time series values (the latter ones only for HSIC). These input series are standardized so that they have zero mean and unit standard deviation.

6.1 Experiments with Simulated Data

The protocol of these experiments is similar to the one used in [Peters *et al.*, 2009]. First, we simulate time series that follow AR(1) processes, using different values of ϕ and with $\epsilon_t \sim |Z|^r \text{sgn}(Z)$, where $Z \sim N(0, \sigma^2)$ and σ is such that ϵ_t has unit standard deviation. The parameter $0.1 \leq r \leq 2$ determines the shape of the noise distribution. For $r = 1$ ϵ_t follows a Gaussian distribution. In the region $r > 1$, the noise is leptokurtic. For values $r < 1$, the distribution of ϵ_t is platykurtic and bimodal. The density of ϵ_t is plotted in the top row of Figure 2 for $r = 0.1, 0.75, 1.25$ and 2. The bottom row in this figure displays the density functions of the time-reversed residuals in the AR(1) model when $\phi = (\sqrt{5}-1)/2$. The reversed residual densities are closer to a Gaussian than the original ones.

In general, both MMD and HSIC use Gaussian kernels whose width σ is $\sigma^2 = m/2$, where m is the median distance between sample points. However, MMD does not perform well using this rule when the modes of the distribution for $r < 1$ are far apart. In this specific case, the median distance between sample points measures mainly the distance between the modes, rather than the characteristic scale of the data. To avoid this problem, we only consider positive values of the residuals when computing m in MMD for $r < 1$. The results of the HSIC test are similar using either the original or the modified rule. In practice, the performances of MMD and of the HSIC test do not strongly depend on the particular choice of σ as long as reasonable values are considered.

Using different values of ϕ and r, we sampled 1000 series of length 100 and evaluated the accuracy of each decision method. The accuracy is defined as the fraction of series for which the method predicts the correct direction. The top row in Figure 3 displays the results for $r = 0.75, 1.25, 1.5$ when ϕ is varied between -0.98 and 0.98. For $r < 1$, all the decision rules perform similarly. However, when $r > 1$ the method based on κ_4 obtains worse results than MMD and HSIC. The reason for this is the larger sampling variance of (3) when the residual distribution is heavy-tailed. When $r > 1$ and ϕ is small, MMD consistently outperforms HSIC. All the methods perform similarly as random guessing for $|\phi|$ close to 0 or 1. The reason is that, for these parameter values, the process is close to being time-reversible. The highest accuracies are obtained for $\phi \approx \pm(\sqrt{5}-1)/2$. This is the value of ϕ for which the Gaussianization effect is the largest.

We also simulate ARMA(2,2) models $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t$, where the distribution of the noise has the same functional form as before. The parameters are $\phi_1 = 0.9$, $\phi_2 = -0.3$, $\theta_1 = -0.29$ and $\theta_2 = 0.5$ and r is varied between 0.1 and 2, as in [Peters *et al.*, 2009]. For each value of r, 1000 series of length 200 are generated. The analysis is made also for other choices of ϕ_1: $\phi_1 = 0.3$ and

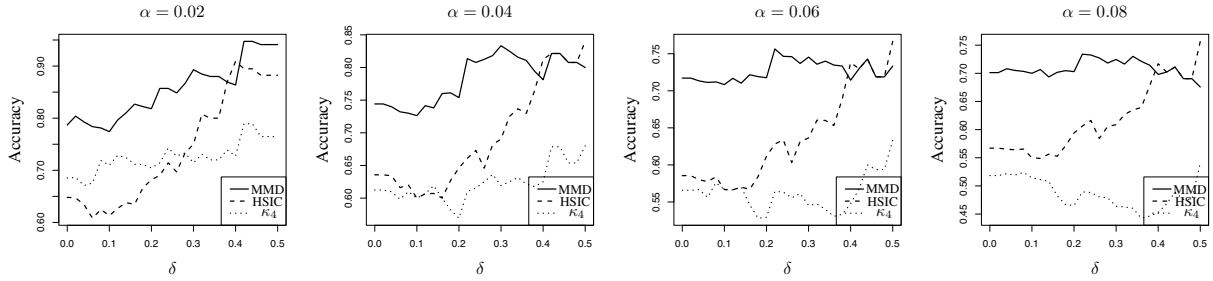

Figure 4: Results obtained by each method in the experiments with EEG data.

$\phi_1 = (\sqrt{5} - 1)/2$. The lag used in the HSIC independence test is $\gamma = 1$, except in the experiments with $\phi_1 = 0.3$, where $\gamma = 2$ performs better. The bottom row in Figure 3 displays the results obtained by MMD, HSIC and the method based on κ_4. MMD performs best, followed by the κ_4 method. The performance of HSIC for $r > 1$ and $\phi_1 = 0.3$ is rather poor. All the methods perform as random guessing for $r = 1$. In this case, the noise distribution is Gaussian and the time series are time-reversible. Similar results are obtained for ARMA models of different orders and different parameter values.

6.2 Experiments with Real-world Data

The performance of the proposed rules is also evaluated in experiments with time series of EEG measurements of brain activity. Following [Peters *et al.*, 2009], we use 1180 time series of length 500 from the publicly available dataset [EEG data, 2010]. Different ARMA models of order up to (5,5) are calibrated to each time series in each possible direction. The best model is selected using the AIC method. For each time series, a Jarque-Bera test of Gaussianity is applied on the empirical residuals of the correct direction of the series [Peters *et al.*, 2009]. We discard the series for which this test does not reject the Gaussian hypothesis: When the noise is Gaussian the series is time-reversible and it is not possible to predict the correct ordering. Finally, we follow [Peters *et al.*, 2009] and select the final set of series on which we make a decision. For this, we use two p-values. The first one, p, is given by an HSIC test that checks for dependencies between e_t and X_{t-1}. The second one, $\tilde{p}$, is given by an HSIC test for dependencies between $\tilde{e}_t$ and X_{t+1}. We select those time series that satisfy $\min(p, \tilde{p}) < \alpha$ and $\max(p, \tilde{p}) > \alpha$ and additionally, $\max(p, \tilde{p}) - \min(p, \tilde{p}) > \delta$, where α and δ are two parameters whose value are fixed beforehand. We make a decision only when the difference between p-values is at least larger than δ and exactly one p-value is lower than α. Typically, α is small and δ large. The first filter (the one based on α) guarantees that the causal ARMA process provides a good description for some ordering of the series. The second filter (the one based on δ) selects series for which the HSIC method is more confident about its decision. For the HSIC test, the lag used in the experiments is $\gamma = 1$, which gives the best results. Both MMD and HSIC use Gaussian kernels whose width σ is $\sigma^2 = m/2$, where m is the median distance between sample points.

Figure 4 shows the accuracy of each method for $\alpha = 0.02$, 0.04, 0.06 and 0.08 and values of δ in the interval $[0, 0.5]$. The number of series selected varies from 164 for $\alpha = 0.8$ and $\delta = 0$ to 17 series for $\alpha = 0.02$ and $\delta = 0.5$. MMD has the best overall performance. The method based on κ_4 obtains rather poor results. The reason is that the noise in the EEG data usually presents heavy tails. These heavy tails mean that the variance of the estimator of κ_4 given by (3) due to the finite size of the sample is rather large. The accuracy of HSIC increases with δ and in some cases HSIC outperforms MMD (for large δ). Note that, by increasing δ, we are selecting the time series in which the HSIC method is more confident about its prediction. Hence, larger values of δ favor the HSIC method. The MMD approach could be modified to generate p-values under the null hypothesis that the distribution of the residuals is Gaussian [Gretton *et al.*, 2007]. These p-values could be used to select the time series for which MMD is more confident about its prediction. However, this possibility has not been addressed in the present investigation.

7 Concluding Remarks

Using the properties of cumulants we have shown that for a causal AR(1) process, the residuals $\{\tilde{\epsilon}_t\}$ of a time-reversed linear stochastic process are closer to a Gaussian than the noise $\{\epsilon_t\}$ used to generate the process in the chronologically ordered direction. The proof is based on showing that all the cumulants of $\tilde{\epsilon}_t$ of order higher than 2 are smaller in absolute value than the cumulants of ϵ_t. Using this property, we have designed a rule to identify the correct direction of a time series that admits a linear model and has non-Gaussian noise. This rule is based on measuring the distance of the residual distribution to the Gaussian. The correct direction is identified as the one for which the model residuals are less Gaussian. Experiments on several AR and ARMA processes with different distributions for the innovations and different parameter values confirm the excellent performance of the proposed rule. Similar results are obtained using real-world time series of EEG data. The performance is good even when the cumulants of the noise are not defined. This is probably related to the fact that the cumulants of a finite sample are finite.

A Derivations

The product cumulants of the random variables ϵ_t and X_{t-1} are defined as

$$\log \mathbf{E}[\exp\{a\epsilon_t + bX_{t-1}\}] \equiv \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{a^n}{n!} \frac{b^m}{m!} \kappa_{nm}(\epsilon_t, X_{t-1}) \,,$$

with $\kappa_{00}(\epsilon_t, X_{t-1}) = 0$, and a and b arbitrary constants. Since ϵ_t and X_{t-1} are independent, only the single-variable cumulants in the expression above are different from zero:

$$\kappa_{nm}(\epsilon_t, X_{t-1}) = \delta_{m0}\kappa_n(\epsilon_t) + \delta_{n0}\kappa_m(X_t)\,,\ n+m>0,$$

where δ is the Kronecker delta and $\kappa_m(X_{t-1}) = \kappa_m(X_t)$ by stationarity. The characteristic function of $\tilde{\epsilon_t}$ and X_{t+1} can be written in terms of the characteristic functions of ϵ_t and of X_t:

$$\begin{aligned}\mathbf{E}[\exp\{a\tilde{\epsilon}_t + bX_{t+1}\}] = \mathbf{E}[\exp\{(b-a\phi)\epsilon_t\}]\cdot \\ \mathbf{E}[\exp\left\{(b\phi + a(1-\phi^2))X_t\right\}]\,,\end{aligned}$$

where a and b are purely imaginary constants. The product cumulants of $\tilde{\epsilon}_t$ and X_{t+1} can also be expressed in terms of the cumulants of ϵ_t and X_t :

$$\begin{aligned}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{a^n}{n!}\frac{b^m}{m!}\kappa_{nm}(\tilde{\epsilon}_t, X_{t+1}) = \\ \sum_{j=0}^{\infty}\frac{1}{j!}\left\{(b-a\phi)^j\kappa_j(\epsilon_t) + (b\phi+a(1-\phi^2))^j\kappa_j(X_t)\right\}\,,\end{aligned}$$

for all $m>0,\ n>0,\ m+n>0$. Assuming stationarity, we can use the relation

$$\kappa_n(X_t) = (1-\phi^n)^{-1}\kappa_n(\epsilon_t)\,,$$

and express the product cumulants in terms of the cumulants of ϵ_t alone

$$\begin{aligned}\kappa_{nm}(\tilde{\epsilon}_t, X_{t+1}) = c_{nm}(\phi)\kappa_{n+m}(\epsilon_t)\,, \qquad (5)\\ c_{nm}(\phi) = (-\phi)^n + (1-\phi^2)^n(1-\phi^{n+m})^{-1}\phi^m\,.\end{aligned}$$

The cumulants of $\tilde{\epsilon}_t$ and X_t are particular cases of (5) with $m=0$ and $n=0$, respectively. Therefore, the cumulants of $\tilde{\epsilon}_t$ can be expressed as a function of the cumulants of ϵ_t:

$$\begin{aligned}\kappa_n(\tilde{\epsilon}_t) = c_n(\phi)\kappa_n(\epsilon_t)\,,\\ c_n(\phi) = (-\phi)^n + (1-\phi^2)^n(1-\phi^n)^{-1}\end{aligned}$$

and $n>0$. For causal AR(1) processes with $\phi\neq 0$ and $|\phi|<1$ we obtain $c_1(\phi)=c_2(\phi)=1$, $|c_n(\phi)|<1$, $\forall n>2$, $\lim_{n\to\infty}c_n(\phi)=0$ and

$$\phi_{min} = \lim_{n\to\infty}\operatorname*{argmin}_{\phi}\{|c_n(\phi)|\} = \pm(\sqrt{5}-1)/2\,.$$

References

[Brockwell and Davis, 1991] P. J. Brockwell and R. A. Davis. *Time series: Theory and methods*. Springer, second edition edition, 1991.

[Daniušis *et al.*, 2010] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence*, 2010.

[EEG data, 2010] EEG data. BCI competition III, dataset IVa, subject 3. http://bbci.de/competition/iii/, 2010.

[Gretton *et al.*, 2007] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007.

[Gretton *et al.*, 2008] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

[Hoyer *et al.*, 2008] P. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 282–289, 2008.

[Janzing, 2010] Dominik Janzing. On the entropy production of time series withunidirectional linearity. *Journal of Statistical Physics*, 138(4):767–779, 2010.

[Kendall *et al.*, 1994] M. G. Kendall, A. Stuart, J. K. Ord, and A. O'Hagan. *Kendall's Advanced Theory of Statistics: Volume I - Distribution Theory*. Arnold, sixth edition edition, 1994.

[Lawrance, 1991] A. J. Lawrance. Directionality and reversibility in time series. *International Statistical Review*, 59(1):67–79, 1991.

[Peters *et al.*, 2009] J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 801–808. ACM, 2009.

[Peters *et al.*, 2010] J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Kernel methods for detecting the direction of time series. In *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 57–66. Springer Berlin Heidelberg, 2010.

[Smola *et al.*, 2007] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *ALT '07: Proceedings of the 18th international conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, 2007.

[Sriperumbudur *et al.*, 2010] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Search*, 11:1517–1562, 2010.

[Weiss, 1975] G. Weiss. Time-reversibility of linear stochastic processes. *Journal of Applied Probability*, 12(4):831–836, 1975.

[Zhang and Hyvärinen, 2009] K. Zhang and A. Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine Learning and Knowledge Discovery in Databases*, volume 5782 of *Lecture Notes in Computer Science*, pages 570–585. Springer Berlin / Heidelberg, 2009.