Gaussianity Measures for Detecting the Direction of Causal Time Series

José Miguel Hernández-Lobato, Pablo Morales-Mombiela and Alberto Suárez

Computer Science Department, Universidad Autónoma de Madrid

IJCAI 2011, Barcelona, Spain, July 20th

Outline

Outline

Detecting the Direction of Causal Time Series

Motivation:

Interesting problem for the evaluation of novel methods for causal inference [Peters et al., 2009].

New insights about the asymmetries between past and future in physical systems [Janzing, 2010].

Simplified setting:

1. Linear models (AR and ARMA processes).
2. Causes $X_{t-1}, X_{t-2}, \ldots$ and effect X_t have the same marginals.

Reversibility in Linear Time Series

The time series $\{X_t\}_{t=-\infty}^{\infty}$ is reversible when $\{X_{t_1}, \ldots, X_{t_n}\}$ and $\{X_{-t_1}, \ldots, X_{-t_n}\}$ have the same joint distribution $\forall n, t_1 < \ldots < t_n$.

Let $\{\epsilon_t\}$ be i.d. white noise and consider the ARMA process

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

If the process is reversible then $\{\epsilon_t\}$ is Gaussian [Weiss, 1975].

Alternative definition of reversibility [Peters et al. (2009)]

Let $\epsilon_t \perp X_{t-k}\ \forall k > 0$. The ARMA process is reversible if

$X_t = \sum_{i=1}^{p} \phi_i X_{t+i} + \sum_{j=1}^{q} \theta_j \tilde{\epsilon}_{t+j} + \tilde{\epsilon}_t,$

where $\{\tilde{\epsilon}_t\}$ is an i.i.d. sequence with $\tilde{\epsilon}_t \perp X_{t+k}, \forall k > 0$.

An ARMA process with i.i.d. noise is reversible $\Leftrightarrow$ $\{\epsilon_t\}$ is Gaussian.

Overview of the Contribution

Main Result

Given a linear time series with non-Gaussian innovations, the residuals of a linear fit to the time-reversed series are more Gaussian than the residuals of the correctly ordered series.

Proved for AR(1) processes. We obtain empirical evidence that it also holds for ARMA processes.

New causal inference rules (based on Gaussianity measures) for detecting the correct ordering of linear time-series data.

These rules outperform state-of-the-art methods in experiments with simulated and real-world data.

Outline

Stationary Causal AR(1) Process

Forward model:

$$X_t = \phi X_{t-1} + \epsilon_t\,, \qquad |\phi| < 1\,, \qquad \phi \neq 0\,,$$

1. $\{\epsilon_t\}$ is i.d. whie noise.
2. $\epsilon_t \perp X_{t-k}\ \forall k > 0.$
3. $\phi = \mathrm{Cov}(X_t, X_{t-1})\,[\mathrm{Var}(X_t)]^{-1}.$

Backward model:

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

1. $\tilde{\phi} = \mathrm{Cov}(X_t, X_{t+1})\,[\mathrm{Var}(X_t)]^{-1} = \phi.$
2. What are the statistical properties of $\{\tilde{\epsilon}_t\}$?

Characteristics of $\{\tilde{\epsilon}_t\}$

When $\{\epsilon_t\}$ is Gaussian:

1. $\{\tilde{\epsilon}_t\}$ is i.d. white noise.
2. $\tilde{\epsilon}_t \perp X_{t+k} \; \forall k > 0$.
2. $\tilde{\epsilon}_t$ has also a Gaussian distribution.

When $\{\epsilon_t\}$ is not Gaussian:

1. $\{\tilde{\epsilon}_t\}$ is i.d. white noise.
2. $\tilde{\epsilon}_t$ is not independent of $X_{t+k} \; \forall k > 0$.
3. $\tilde{\epsilon}_t$ is more Gaussian than ϵ_t (New Result).

Proof Based on Cumulants

The cumulants of $\tilde{\epsilon}_t$ are written in terms of the cumulants of ϵ_t:

$$\kappa_n(\tilde{\epsilon}_t) = c_n(\phi)\, \kappa_n(\epsilon_t),$$

where

$$c_n(\phi) = (-\phi)^n + (1 - \phi^2)^n (1 - \phi^n)^{-1},$$

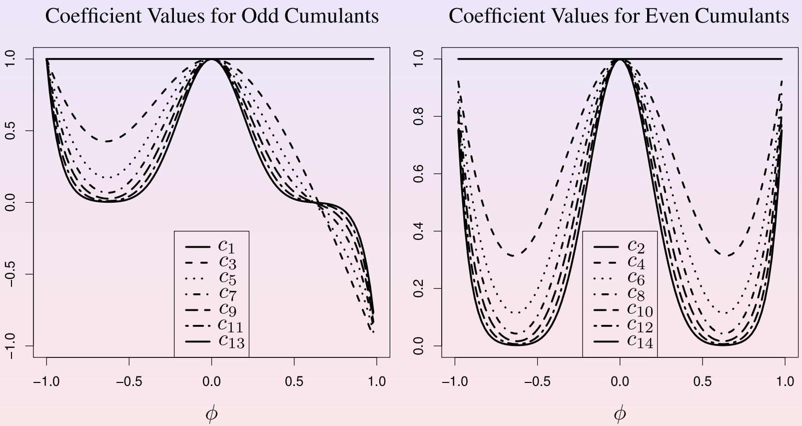
which satisfies $|c_n(\phi)| < 1$, $\forall n > 2$.

The Gaussian is the only distribution whose cumulants of order larger than two are zero.

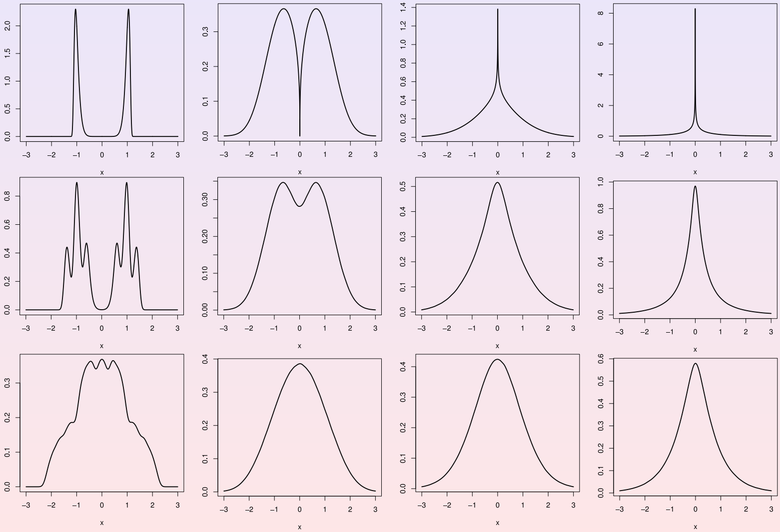
What values of ϕ produce the strongest Gaussianisation?

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

Plots of $c_n(\phi)$ for $n = 1, \ldots, 14$



Gaussianization Effect



Outline

Unbiased Estimator of κ_4

Given a sample $\{Y_i\}_{i=1}^N$, the unbiased estimator of κ_4 with lowest variance is the 4-th k-statistic [Kendall et al., 1994]:

$$k_4 = \frac{1}{N^{[4]}}\left[(N^3+N^2)S_4 - 4(N^2+N)S_3S_1 - 3(N^2-N)S_2^2 + 12NS_2S_1^2 - 6S_1^4\right],$$

where $N^{[4]} = N(N-1)(N-2)(N-3)$ and $S_j = \sum_{i=1}^N Y_i^j$.

Causal Inference Rule κ_4

Select the direction for which the empirical residuals have the largest value of $|k_4|$.

Maximum Mean Discrepancy (MMD)

The MMD between p and q within the unit ball $\mathcal{F}$ in $\mathcal{H}$ 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 mappings of p and q onto $\mathcal{H}$.

MMD is 0 when $p = q$, positive otherwise [Gretton et al., 2007].

Causal Inference Rule MMD

Select the direction for which the empirical distribution of the residuals has largest MMD distance to the Gaussian.

Outline

Benchmark Method (HSIC)

Employs a test of independence based on the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2008].

The test computes the MMD between the empirical joint distribution and the product of the empirical marginal distributions.

Recall: when ϵ_t is not Gaussian $\tilde{\epsilon}_t$ is not independent of X_{t+k}, $k > 0$.

Causal Inference Rule HSIC

Select the direction in which the empirical residuals are less dependent of the previous time series values [Peters et al., 2009].

Experiments with Simulated Data

AR(1) processes. 1000 series of length 100.

$\epsilon_t \sim |Z|^r \mathrm{sgn}(Z)$, where $Z \sim N(0, \sigma^2)$. σ such that $\mathrm{std}(\epsilon_t) = 1$.

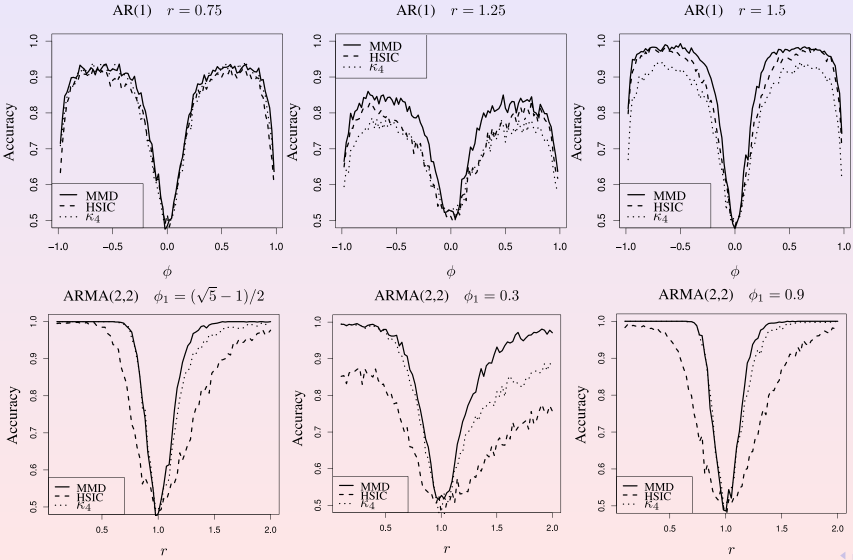
$r = 1$ The noise is Gaussian.
$r > 1$ The noise is leptokurtic.
$r < 1$ The noise is platykurtic and bimodal.

ARMA(2,2) processes. 1000 series of length 200.

$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \epsilon_t,$

Parameter values: $\phi_1 = 0.9$, $\phi_2 = -0.3$, $\theta_1 = -0.29$ and $\theta_2 = 0.5$ used by Peters et al., 2009.

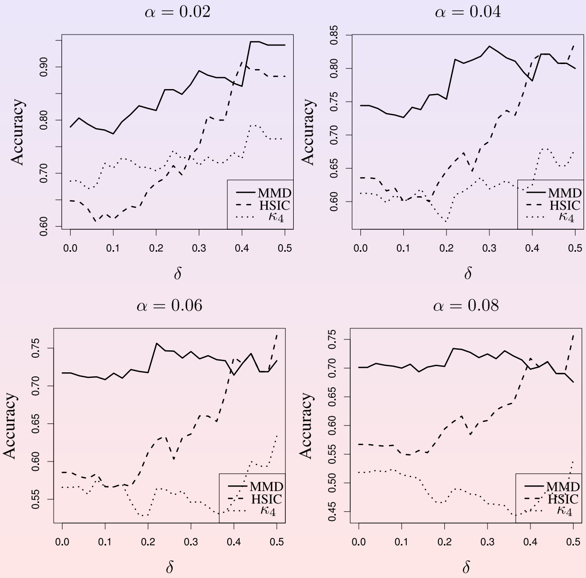
Results on Simulated Data



Experiments on Real World Data

- ⋆ 1180 series of length 500 with EEG measurements.
- ⋆ ARMA processes of order up to (5,5) calibrated to each direction.
- ⋆ Best model selected using AIC.
- ⋆ Experimental protocol of Peters et al., 2009. Given the two *p*-values generated by the HSIC method, only make a decision if
  - ⋆ One of them is larger than α and the other one is not.
  - ⋆ The distance between them is larger than δ.
  - ⋆ Usually α is small and δ is large.

Results on Real World Data



Outline

Conclusions

- For an AR(1) process, the reversed residuals are more Gaussian than the original residuals (proved using Cumulants).
- Using measures of Gaussianity, we can detect the correct direction of an AR(1) time series.
- Also seems to work very well for ARMA processes.
- On simulated and real-world data, we outperform existing state-of-the-art methods based on testing for independencies.

References

- J. Peters, D. Janzing, A. Gretton, and B. Schölkopf. Detecting the direction of causal time series. In ICML 2009, pages 801-808.
- D. Janzing. On the entropy production of time series with unidirectional linearity. Journal of Statistical Physics, 138(4):767-779, 2010.
- G. Weiss. Time-reversibility of linear stochastic processes. Journal of Applied Probability, 12(4):831-836, 1975.
- 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.
- 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 513520. MIT Press, 2007.
- 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 585592. MIT Press, 2008.

Thank you for your attention!