The arrow of time in time series

Alberto Suárez, José Miguel Hernández Lobato, Pablo Morales Mombiela

Machine learning group

EPS, Universidad Autónoma de Madrid (Spain)

Escuela Politécnica Superior

alberto.suarez@uam.es

Please, ask questions!

The arrow of time

The past and the future are distinguishable

- **In thermodynamics**

  In an isolated system the thermodynamic entropy is an increasing quantity (unless all processes in the system are reversible, in which case it is constant)

- **In cosmology:**

  The universe expands towards the future /contracts towards the past (initial singularity: The big bang)

- **In causality:**

  A cause precedes its effect.

How does the arrow of time arise?

- **The basic laws of physics are time-reversible**
  - Newton's equations in classical mechanics
  - Schrödinger's equation in quantum mechanics
- **Macroscopic / mesoscopic equations describe irreversible behavior**
  - Diffusion equation
  - Hydrodynamic equations

QUESTION: How can irreversibility arise (at the meso and macroscopic level) from time-reversible laws describing the evolution of the system at the microscopic level?

Possible answers

- **Boltzmann:** For the purpose of macro and mesoscopic descriptions, it is sufficient to replace the exact many-particle distribution function by a form that satisfies irreversible equations. The differences between the exact and the approximate form would be apparent only if the observation time is very long (in systems of realistic size, the Poincaré recurrences would take longer than the age of the universe) or if the initial state is highly correlated, which is a very improbable situation.

- **Prigogine:** The basic laws of physics need to be modified and should include sources of irreversibility at the fundamental level.

The arrow of time in time series

Problem:

- We are given N ordered values $X_1, X_2, \ldots, X_N$ from a time series.
- The direction of time is unknown.
- Which of the orderings

$$X_1, X_2, X_3, \ldots, X_{N-1}, X_N$$

$$X_N, X_{N-1}, X_{N-2}, \ldots, X_2, X_1$$

is the correct chronological order?

Temporal symmetry of stochastic processes

- A stochastic process $\{X_t;\ t \in Z\}$ is stationary if

  $P(X_{t_1}, X_{t_2}, X_{t_3}, \ldots, X_{t_r}) = P(X_{t_1+\tau_1}, X_{t_2+\tau}, X_{t_3+\tau}, \ldots, X_{t_r+\tau})$

- A stationary stochastic process $\{X_t;\ t \in Z\}$ is temporally symmetric if

  $P(X_{t_1}, X_{t_2}, X_{t_3}, \ldots, X_{t_r}) = P(X_{-t_1}, X_{-t_2}, X_{-t_3}, \ldots, X_{-t_r})$
- If the stationary process $\{X_t;\ t \in Z\}$ is temporally symmetric, then

  $P(X_1, X_2, \ldots, X_T) = P(X_T, X_{T-1}, \ldots, X_1)$

Gaussian processes

- Stationary Gaussian processes are temporally symmetric
  - The joint distributions of Gaussian processes are fully determined by the covariance matrix

    $Cov(X_{t_i}, X_{t_j})$
  - The covariance matrix is symmetric

    $Cov(X_{t_i}, X_{t_j}) = Cov(X_{t_j}, X_{t_i})$
  - By stationarity

    $Cov(X_{t_i}, X_{t_j}) = Cov(X_{t_j}, X_{t_i}) = Cov(X_{t_j+\tau}, X_{t_i+\tau})$

    $= Cov(X_{-t_i}, X_{-t_j})$ [choose $\tau = -t_i - t_j$]

Therefore: $P(X_{t_1}, X_{t_2}, X_{t_3}, \ldots, X_{t_r}) = P(X_{-t_1}, X_{-t_2}, X_{-t_3}, \ldots, X_{-t_r})$

Markov processes

- A stationary process $\{X_t;\ t \in Z\}$ is first order Markov if

  $$P(X_t | X_{t-1}, X_{t-2}, \ldots) = P(X_t | X_{t-1}) \quad \forall t \in Z$$

- If $\{X_t;\ t \in Z\}$ is a stationary Markov process

  $\{X_{-t};\ t \in Z\}$ is also a stationary Markov process

- Therefore, an equivalent definition is

  $$P(X_t | X_{t+1}, X_{t+2}, \ldots) = P(X_t | X_{t+1}) \quad \forall t \in Z$$

- Using Markov property + Chapman-Kolmogorov eqn.

$$P_{X_{t-\tau}, X_t}(x,y) = \int dz \, \frac{P_{X_{t-\tau'-\tau''}, X_{t-\tau''}}(x,z) P_{X_{t-\tau''}, X_t}(z,y)}{P_X(z)} \qquad \tau' + \tau'' = \tau$$

Temporal symmetry of Markov processes

A stationary Markov process is temporally symmetric iff

$$P_{X_{t-1}, X_t}(x,y) = P_{X_{t-1}, X_t}(y,x)$$

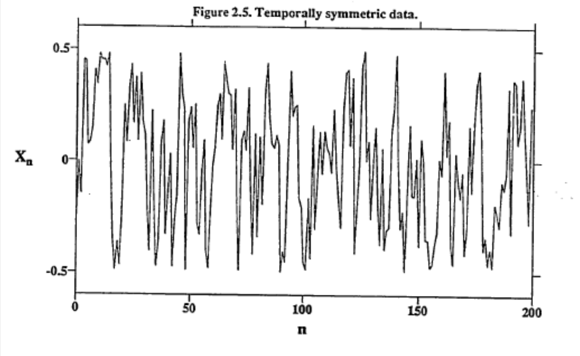
Examples:

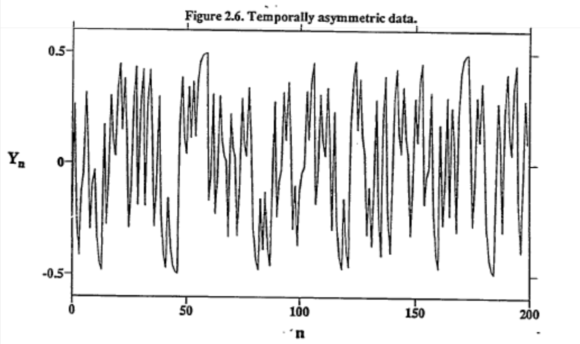
- Temporally symmetric and not Gaussian [$|a| \leq 1/3$]

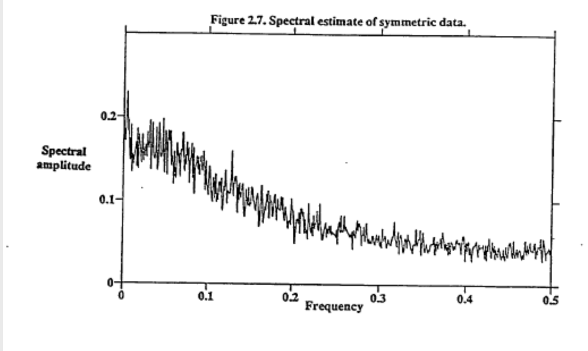
  $$P_{X_{t-1}, X_t}(x,y) = P_X(x) P_X(y) \{1 + 3a(2P_X(x) - 1)(2P_X(y) - 1)\}$$

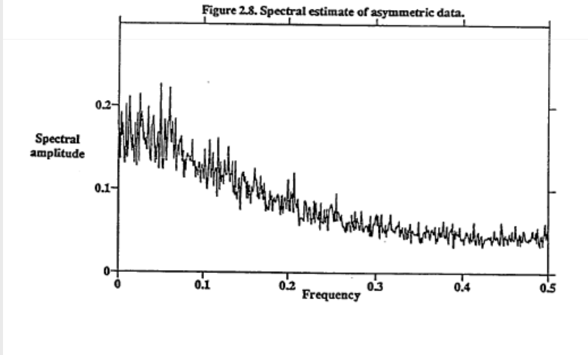
- Temporally asymmetric **[Non-Gaussian AR(1)]**

  $X_t = \phi X_{t-1} + W_t \quad \{W_t;\ t \in Z\}$ Non-Gaussian white noise

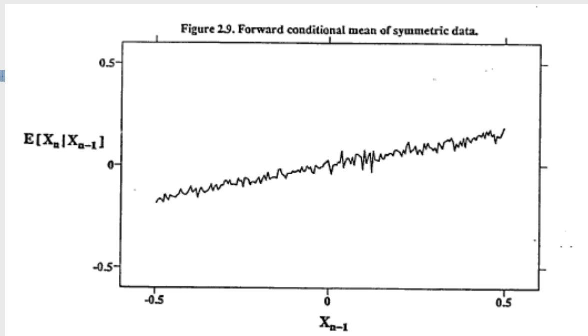


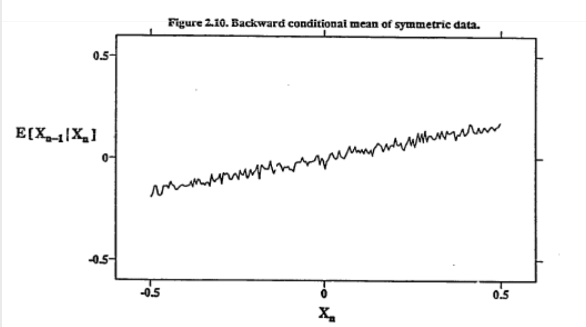


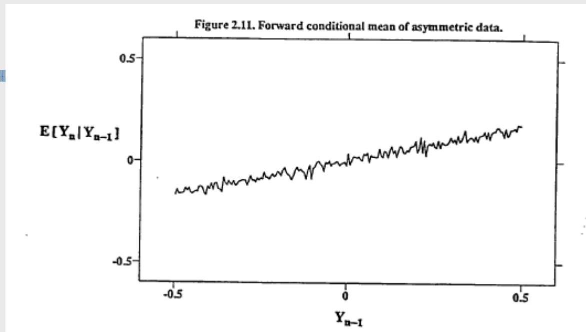


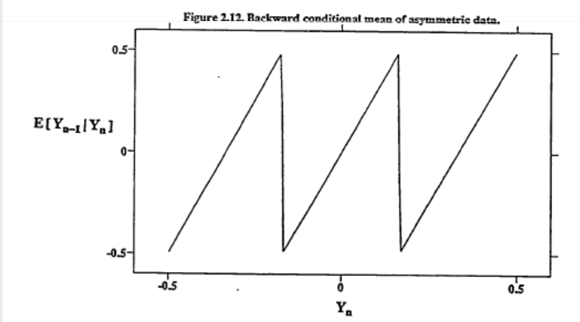


Source: P.S. Rao, "Non-Gaussian Markov time series" (1988)









Linear AR(1) processes

$$X_t = \phi X_{t-1} + W_t$$

$\{W_t;\ t \in Z\}$ white noise

- A linear AR(1) process is temporally symmetric iff $\{W_t;\ t \in Z\}$ is Gaussian.
- A Gaussian Markov process follows a linear AR(1) with Gaussian white noise.

Linear AR(1) processes

$$X_t = \phi X_{t-1} + W_t$$

$|\phi| < 1;\quad \{W_t;\ t \in Z\}$ white noise

$W_t \perp X_{t-k} \quad \forall k > 0$ [independence]

- A linear AR(1) process is temporally symmetric iff $\{W_t;\ t \in Z\}$ is Gaussian.
- A Gaussian Markov process satisfies the Stochastic Difference Equation of a linear AR(1) with Gaussian white noise.

Gaussian AR(1) processes

- Forward representation

  $X_t = \phi X_{t-1} + W_t \quad \{W_t;\ t \in Z\}$ Gaussian white noise

  $|\phi| < 1; \quad W_t \perp X_{t-k}\ \forall k > 0$ [independence]

  $E[X_t | X_{t-1}] = \phi X_{t-1}$

- Time-reversed representation

  $X_t = \phi X_{t+1} + \hat{W}_t \quad \{\hat{W}_t;\ t \in Z\}$ Gaussian white noise

  $|\phi| < 1; \quad \hat{W}_t \perp X_{t+k}\ \forall k > 0$ [independence]

  $E[X_t | X_{t+1}] = \phi X_{t+1}$

There is no arrow of time

Non-Gaussian linear AR(1) processes

- Forward representation

  $X_t = \phi X_{t-1} + W_t$ **$\{W_t;\ t \in Z\}$ Non-Gaussian white noise**

  $|\phi| < 1; \phi \neq 0 \quad W_t \perp X_{t-k}\ \forall k > 0$

  $E[X_t | X_{t-1}] = \phi X_{t-1}$

- Time-reversed residuals

  $\hat{W}_t = X_t - \phi X_{t+1}$ **$\{\hat{W}_t;\ t \in Z\}$ Non-Gaussian white noise**

  $E[\hat{W}_t X_{t+1}] = 0$ BUT $\hat{W}_t$ is not independent of X_{t+1}

  $E[X_{t-1} | X_t] = f(X_t);$ **f is non-linear**

There is an arrow of time

Properties of the time reversed linear AR(1)

$X_t = \phi X_{t-1} + W_t$ **$|\phi| < 1; \phi \neq 0;$ $\{W_t; t \in Z\}$ white noise**

$\hat{W}_t = X_t - \phi X_{t+1}$ **[time-reversed residuals]**

(i) $\{X_t\}$ and $\{\hat{W}_t\}$ are Gaussian iff $\{W_t\}$ is Gaussian.

(ii) $\hat{W}_t$ is independent of X_{t+k} (k>0) iff $\{W_t\}$ is Gaussian.

(iii) $\{\hat{W}_t\}$ is white noise iff $\{W_t\}$ Gaussian.

(iv) $E[X_t | X_{t+1}] = \phi X_{t+1}$ iff $\{W_t\}$ is Gaussian

Non-Gaussian linear AR(1)

$X_t = \phi X_{t-1} + W_t$ **$|\phi| < 1; \phi \neq 0;$ $\{W_t; t \in Z\}$ white noise**

$\hat{W}_t = X_t - \phi X_{t+1}$ **[time-reversed residuals]**

If $\{W_t\}$ is non-Gaussian

(I) The **optimal predictor** $E[X_t | X_{t+1}]$ in the **time-reversed direction** is **non-linear**.
The **optimal prediction error** in the **time-reversed direction** is **smaller than** the optimal error **in the forward direction**.

$$E[(X_t - E[X_t | X_{t+1}])^2] < E[(X_t - E[X_t | X_{t-1}])^2]$$

(II) **$\hat{W}_t$ and X_{t+k}** (k>0) become **dependent** quantities.

(III) **The time-reversed residual** $\{\hat{W}_t\}$ is **more Gaussian** than $\{W_t\}$.

Detecting the direction of time (I)

- **If $\{W_t\}$ is non-Gaussian** the **optimal predictor** E $[X_t|X_{t+1}]$ in the **time-reversed direction** is **non-linear** [Rao + Johnson 1992]

$$X_t = \begin{cases} \frac{1}{2}X_{t-1} - \frac{1}{4} & \text{w.p. } \frac{1}{2}; \\ \frac{1}{2}X_{t-1} + \frac{1}{4} & \text{w.p. } \frac{1}{2} \end{cases} \qquad X_t \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

$$P_{X_t X_{t-1}}(x,y) = \frac{1}{2}\delta\left(x - \frac{1}{2}y + \frac{1}{4}\right) + \frac{1}{2}\delta\left(x - \frac{1}{2}y - \frac{1}{4}\right)$$

$$E[X_t|X_{t-1}] = \frac{1}{2}X_{t-1};$$

$$E[X_t|X_{t+1}] = 2X_{t+1} \bmod 1 - \frac{1}{2}$$

Detecting the direction of time (I)

- The **optimal prediction error** in the **time-reversed direction** is **smaller than** the optimal error **in the forward direction**.

$$E[(X_t - E[X_t|X_{t+1}])^2] < E[(X_t - E[X_t|X_{t-1}])^2]$$

$$X_t = \begin{cases} \frac{1}{2}X_{t-1} - \frac{1}{4} & \text{w.p. } \frac{1}{2}; \\ \frac{1}{2}X_{t-1} + \frac{1}{4} & \text{w.p. } \frac{1}{2} \end{cases} \qquad X_t \in \left[-\frac{1}{2}, \frac{1}{2}\right]$$

$$X_t = 2X_{t+1} \bmod 1 - \frac{1}{2}; \qquad \text{[deterministic chaotic map]}$$

$$E[X_t|X_{t-1}] = \frac{1}{2}X_{t-1}; \qquad E[X_t|X_{t+1}] = 2X_{t+1} \bmod 1 - \frac{1}{2}$$

$$E\left[(X_t - E[X_t|X_{t-1}])^2\right] = \frac{1}{16}; \qquad E\left[(X_t - E[X_t|X_{t+1}])^2\right] = 0$$

Detecting the direction of time (II)

- **If {W_t} is non-Gaussian $\hat{W}_t$ and X_{t+k}** (k>0) become **dependent** quantities [Peters, Janzig, Gretton, Schölkopf, 2009]

Algorithm 1 Detecting true Time Direction

```
Input: a = (x_1, ..., x_n), b = (x_n, ..., 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, ..., x_n) correct time direction"
   end if
else if (res_a, a) dependent then
   if (res_b, b) independent then
      output = "(x_n, ..., x_1) correct time direction"
   else if (res_b, b) dependent then
      output = "I do not know (bad fit)"
   end if
end if
```

- Gaussianity test: Jarque-Bera
- Independence test: HSIC
  Hilbert-Schmidt Independence Criterion

$$\max_{\substack{f,g \\ \|f\|=\|g\|=1}} \mathrm{cov}\left[f(X_n), g(W_{n-1})\right]$$

Detecting the direction of time (III)

- **{$\hat{W}_t$} is more Gaussian than {W_t}**
- **Cumulant / product-cumulant generating function**

$$E[\exp\{aW_t\}] = \exp\left\{\sum_{n=1}^{\infty}\frac{a^n}{n!}K_n[W_t]\right\}; \qquad E[\exp\{aW_t + bX_{t-1}\}] = \exp\left\{\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{a^n}{n!}\frac{b^m}{m!}K_{nm}[W_t, X_{t-1}]\right\}$$

$$E[\exp\{a\hat{W}_t\}] = \exp\left\{\sum_{n=1}^{\infty}\frac{a^n}{n!}K_n[\hat{W}_t]\right\}; \qquad E[\exp\{a\hat{W}_t + bX_{t+1}\}] = \exp\left\{\sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{a^n}{n!}\frac{b^m}{m!}K_{nm}[\hat{W}_t, X_{t+1}]\right\}$$

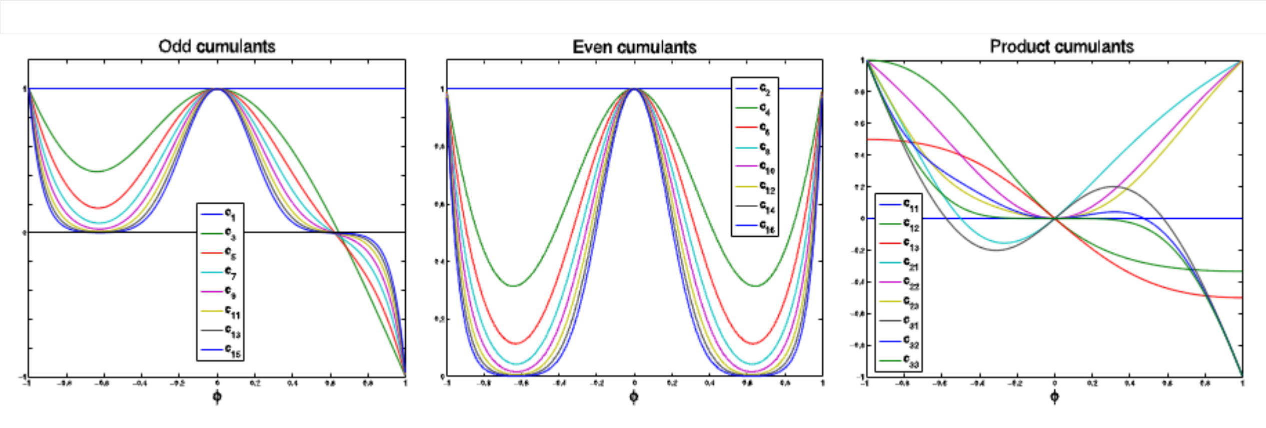
- **Cumulant / product-cumulant moments**

$$K_n[\hat{W}_t] = c_n(\phi)K_n[W_t]; \qquad c_n(\phi) = (-\phi)^n + \frac{(1-\phi^2)^n}{1-\phi^n}; \qquad n > 0$$

$$K_{nm}[\hat{W}_t, X_{t+1}] = c_{nm}(\phi)K_{n+m}[W_t]; \qquad c_{nm}(\phi) = (-\phi)^n + \frac{(1-\phi^2)^n}{1-\phi^{n+m}}\phi^m; \quad n+m > 0$$

Detecting the direction of time (III)

$$K_n\left[\hat{W}_t\right] = c_n(\phi) K_n\left[W_t\right]; \qquad K_{nm}\left[\hat{W}_t, X_{t+1}\right] = c_{nm}(\phi) K_{n+m}\left[W_t\right];$$



Detecting the direction of time (III)

Assuming cumulants / product-cumulants are defined

- A random variable is Gaussian iff $K_n = 0 \quad \forall n > 2$
  - $\hat{\mathbf{W}}_t$ is more Gaussian than $\mathbf{W}_t$

    $|c_n(\phi)| < 1; \quad \Rightarrow \left|K_n\left[\hat{W}_t\right]\right| < \left|K_n\left[W_t\right]\right|; \quad n > 2$

- Two variables are independent iff

  $K_{nm} = 0 \quad n > 0, m > 0, n+m > 2$

  - $(\hat{\mathbf{W}}_t, X_{t+1})$ are independent only iff $K_n\left[W_t\right] = 0 \ \forall n > 2$

    $K_{nm}\left[\hat{W}_t, X_{t+1}\right] = c_{nm}(\phi) K_{n+m}\left[W_t\right];$

Gaussianization of the time-reversed residuals

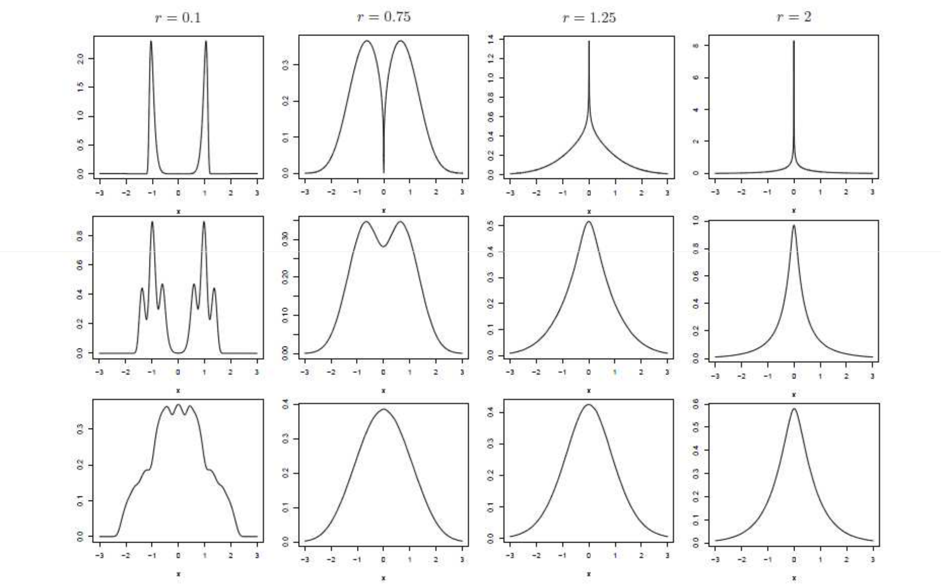
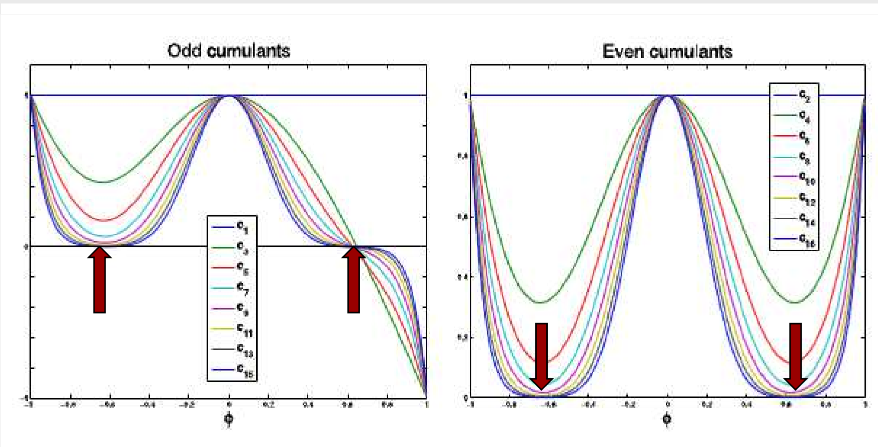


Figure 2: Density functions for the noise (top row) and the time-reversed residuals in an AR(1) model with $\phi = 0.2$ (middle row) and $\phi = (\sqrt{5}-1)/2$ (bottom row).

The golden ratio in AR(1)

- The Gaussianization of the time-reversed residuals in Non-Gaussian AR(1) is strongest when the autocorrelation in the time-series is the golden ratio

$$\phi = \pm \frac{\sqrt{5}-1}{2} = \pm 0.6180$$



Distance to a Gaussian

- **Magnitude of the fourth cumulant (excess of kurtosis)**
- **Kullback-Leibler divergence**
  Let $p(z)$, $q(z)$ be two pdf's on $\mathbb{R}$ $\quad KL(p \| q) = \int dz\, p(z) \log \frac{p(z)}{q(z)}$
- **Maximum Mean Discrepancy (MMD)**
  - Let p, q be two pdf's on $\mathbb{R}$
  - Let $\mathcal{F}$ be the unit ball in a universal reproducing kernel Hilbert space $\mathcal{H}$ two pdf's on $\mathbb{R}$

$$MMD(p, q, \mathcal{F}) = \sup_{f \in \mathcal{F}} \left(E_p[f(z)] - E_q[f(z)] \right) = \left\| \mu_p - \mu_q \right\|_{\mathcal{H}}$$

$\mu_p = E_p[k(z,\cdot)]$; $\mu_q = E_q[k(z,\cdot)]$; mappings of p, q onto $\mathcal{H}$

$\|\cdot\|_{\mathcal{H}}$: norm operator in $\mathcal{H}$

Tests for the arrow of time in linear processes

- **Fit AR(1)** to $\{X_1, X_2, \ldots, X_N\}$ **[original order (o)]**

$$X_n = \hat{\phi}\, X_{n-1} + W_n^o \quad \Rightarrow \quad W_n^o = X_n - \hat{\phi}\, X_{n-1}$$

- **Fit AR(1)** to $\{X_N, X_{N-1}, \ldots, X_1\}$ **[inverted order (i)]**

$$X_n = \hat{\phi}\, X_{n+1} + W_n^i \quad \Rightarrow \quad W_n^i = X_n - \hat{\phi}\, X_{n+1}$$

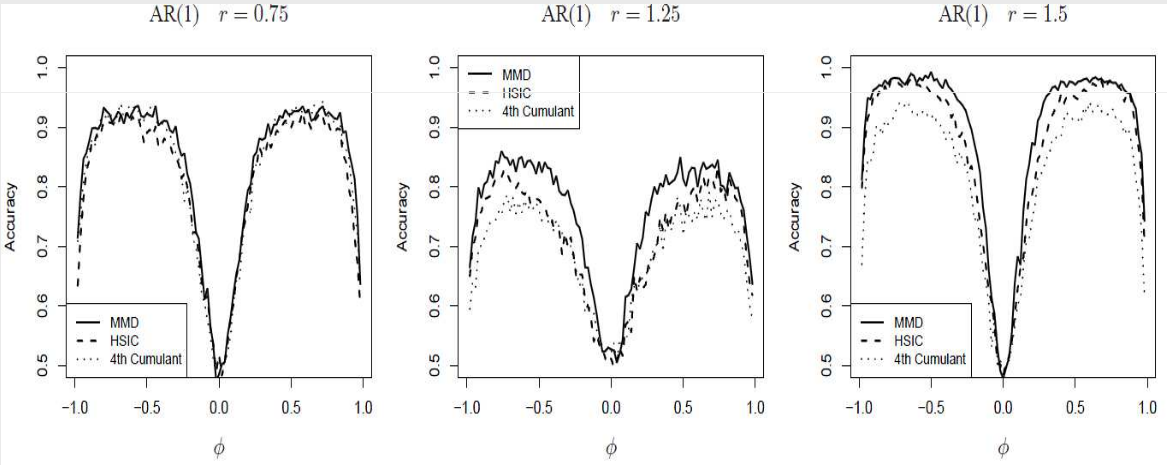
Which is the correct chronological order?

- **Measure of independence** (**HSIC**)
  The order in which $\{W_n^o, X_{n-1}\}$ / $\{W_n^i, X_{n+1}\}$ are independent
- **Measures of Gaussianity:**
  The order (x = o / i) in which $\{W_n^x\}$ has the largest
  - **4th Cumulant**
  - **MMD to a Gaussian**

Experiments on simulated AR(1) data (I)

$X_t = \phi X_{t-1} + W_t$ $\quad W_t = (Z_t)^r \quad Z_t \sim N(0,1)$

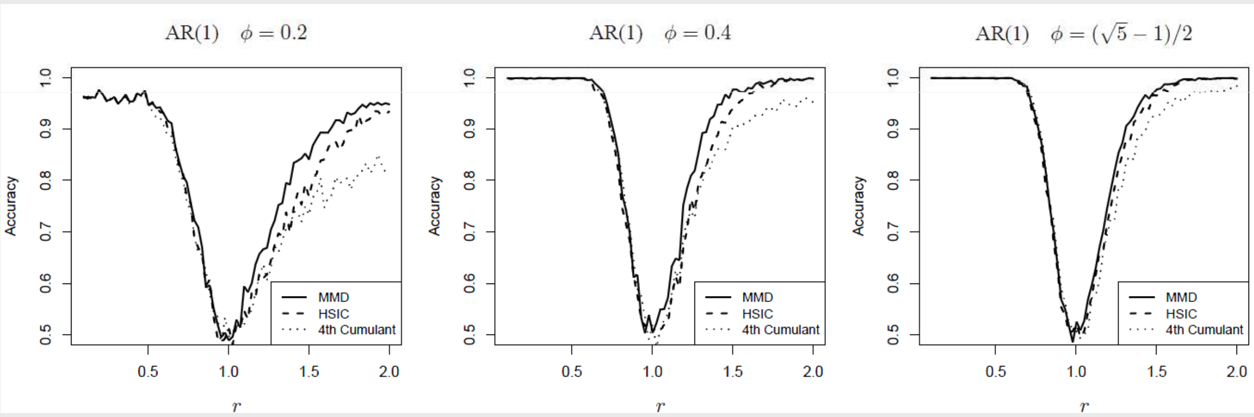
Keep r fixed, sweep ϕ



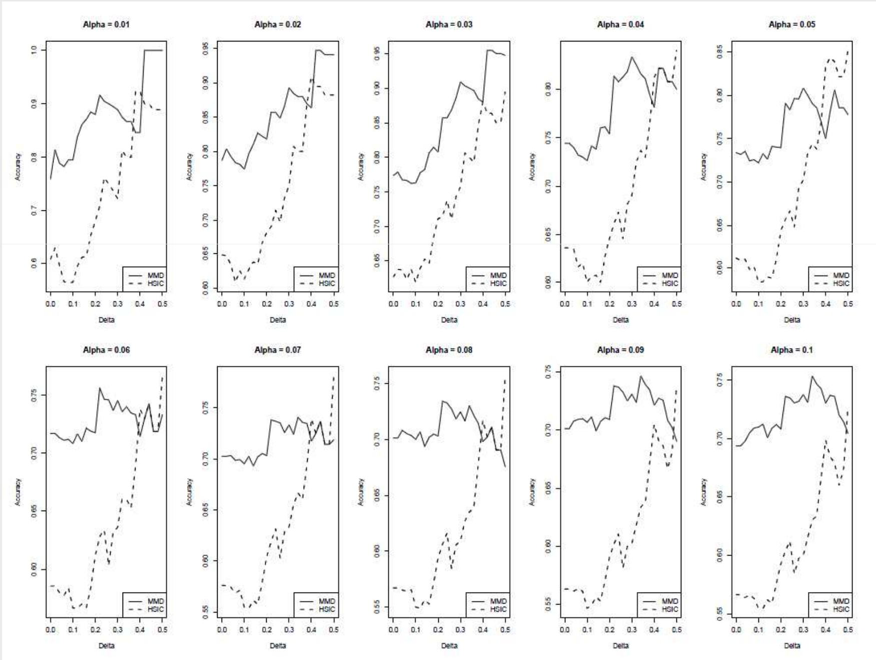
Experiments on simulated AR(1) data (II)

$X_t = \phi X_{t-1} + W_t$ $\quad W_t = (Z_t)^r \quad Z_t \sim N(0,1)$

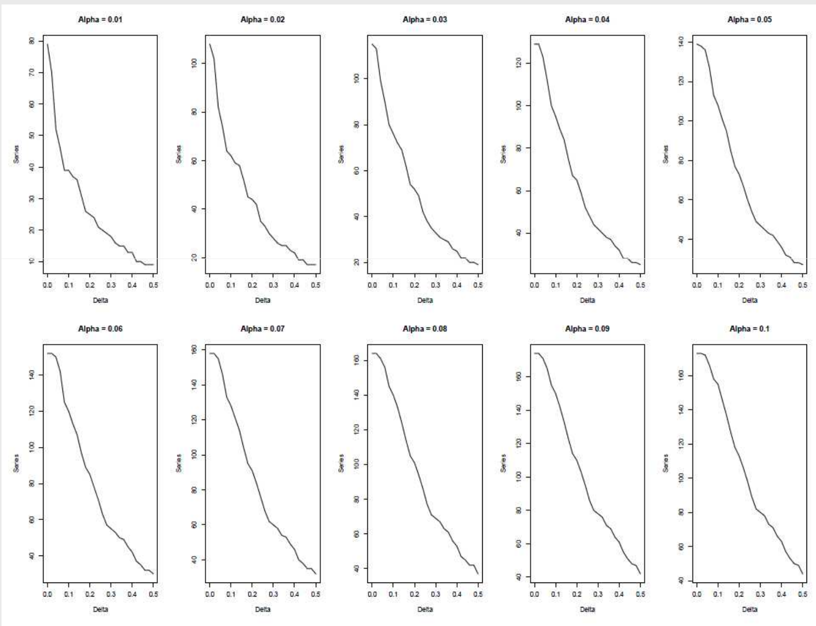
Keep ϕ fixed, sweep r



Experiments on EEG data



Experiments on EEG data



Applications

- **Detection of the time direction** in empirical time series.
- **Diagnostics** based on changes in the behavior under reversal of the direction of time **in medical time series**.
- **Compression of time-reversed signals** should be more effective.
- **Detection of causality**.
- **Detection of leading indicators**.
- **Irreversibility in physical systems**.
- **Irreversibility in machine learning**: why do we learn, rather than unlearn, as more examples become available.

Thank you!

Please, ask questions!

Bibliography

References

[1] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Sch¨olkopf, 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.

[2] 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.*

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

[4] Peter J. Brockwell and Richard A. Davis. *Time series: Theory and methods. Springer, second* edition edition, 1991.

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

[6] Q. Cheng. On time-reversibility of linear processes. *Biometrika, 86(2):483–486, 1999.*

[7] Howell Tong and Zhiqiang Zhang. On time-reversibility of multivariate linear processes. *Statistica Sinica, 15:495–504, 2005.*

[8] Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Sch¨olkopf. 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.*

[9] P. S. Rao, D. H. Johnson, and D. D. Becker. Generation and analysis of non-Gaussian Markov time series. *IEEE Transactions on Singal Processing, 40(4):845–856, apr 1992.*

Bibliography

[10] 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.*

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

[12] A. Smola, A. Gretton, L. Song, and B. Sch¨olkopf. 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.

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

[14] S. Shimizu, P. O. Hoyer, A. Hyv¨arinen, and A. Kerminen. A linear non-Gaussian acyclic modelfor causal discovery. *The Journal of Machine Learning Research, 7:2003–2030, 2006.*

[15] P.O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Sch¨olkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, pages* 689–696. MIT Press, 2009.

[16] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Proceedings of the Twenty-Fifth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-09), pages 647–655, Corvallis, Oregon, 2009. AUAI* Press.