# Another Look at Dependence: the Most Predictable Aspects of Time Series

Tommaso Proietti

University of Rome Tor Vergata

Séminaire ECODEP
IHP Paris, March 15, 2023

# Introduction

- In time series analysis, the traditional Bravais-Pearson autocorrelation function (ACF) provides essential information on the dependence structure of a Gaussian process. Outside this class, it is less informative. In particular, zero correlation only implies lack of **linear association**, or zero linear predictability.
- Other classical measures, such as Spearman's rank correlation coefficient and Kendall's tau, measure **monotonic association**.
- Alternative dependence measures have been developed, see Tjostheim (2018) for a recent overview. Interest lies in broader, less restrictive measures of the dependence. There is a substantial literature on testing the null of (serial) independence and iid-ness.
- Hong (1999) defined a measure of dependence based on the covariance between the characteristic functions of $X_t$ and $X_{t-k}$.
- Zhou (2012) extended to strictly stationary time series the dependence measure based on the concept of **distance covariance and correlation**, introduced by Szekely et al. (2007).

- Zhou defined the **auto-distance covariance function**, and Fokianos and Pitsillou (2018) proposed a test of serial independence based on the auto-distance correlation function. Fokianos and Pitsillou (2017) extended the theory to multivariate processes.
- Compared with the classical ACF, the auto-distance correlation function and its Fourier transform, the generalized spectral density by Hong, can capture general possibly **nonlinear** forms of serial **dependence**. See Edelmann et al. (2019) for a review.
- Escanciano and Velasco (2006) propose **conditional mean dependence** measures based on the covariance between $X_t$ and the characteristic function of $X_{t-k}$ and proposed a test of the martingale difference hypothesis based on the sample spectral distibution function constructed on the empirical covariances.
- Shao and Zhang (2014) proposed to measure the degree of **conditional mean independence** of $X_t$ from its past by the martingale difference correlation.
- Carcea and Serfling (2015) introduce the Gini autocovariance function, measuring the degree of monotonicity in the relationship between $X_t$ and $X_{t-k}$.

The generality of serial independence tests is such that the most interpretable outcome is: *do we fail to reject (conditional mean) independence*? The answer for economic and financial time series is typically 'no', raising thereby the issue as to why independence is rejected.

The main motivation for this paper is to provide an answer to both questions, by establishing what is the transformation of the series which is most predictable from its past, and estimating its predictability.

Our focus is on mutual information (MI), a measure of dependence defined as the Kullback-Leibler distance between the joint probability density function (pdf) and the product of the marginal pdfs.

It has a long tradition in time series, see Ibragimov and Rozanov (1978, 2012), Jewell and Bloomfield (1983), Jewell et al. (1983), and Pourahmadi (2001), in information and communication theory (Cover and Thomas, 2006), and data science. For recent contribution see Reshef et al. (2011), Kinney and Atwal (2014) and Berrett and Samworth (2019).

For the analysis of univariate time series measures of serial dependence based on the mutual information between $(X_t, X_{t-j})$ have been proposed by Granger and Lin (1994).

Our approach is related to Gourieroux and Jasiak (2002) nonlinear autocorrelogram, and to Owen (1983), who develops the optimal transformation of an autoregressive processes by an adaptation of the alternating conditional expectation algorithm by Breiman and Friedman (1985).

This is a summary of our contribution:

- We consider the mutual information between past and future (MIPF) as the target measure of predictability. This is a broad measure that takes into account all future forecast horizons, rather than focusing on the one-step-ahead forecast mean square error.
- The most predictable aspect is defined as the measurable transformation of $X_t$ for which the MIPF is a maximum. The proposed transformation arises from of the linear combination of a set of basis functions localized at the quantiles of the unconditional distribution of $X_t$.
- We consider several basis functions and consider their merits.
- The mutual information is estimated as a function of the sample partial autocorrelations, by a semiparametric method which estimates an infinite sum by a regularized finite sum.

# Mutual Information

Let $X$ and $Y$ denote a pair of multivariate continuous random variables, with probability density function (pdf) $f(X,Y)$ and marginal densities $f(X)$ and $f(Y)$, respectively.

The mutual information between $X$ and $Y$ is defined as

$$I(X,Y) = \mathrm{E}_{(X,Y)}\left\{\log \frac{f(X,Y)}{f(X)f(Y)}\right\},$$

where $\mathrm{E}_U(g(U)) = \int_{-\infty}^{\infty} g(u)f(u)du$, for any measurable function $g(U)$ of $U$ with pdf $f(U)$.

This is interpreted as the Kullback-Leibler distance between the joint distribution and product of the marginal distribution.

**Properties**

- Nonnegativity: $I(X, Y) \geq 0$.
- $I(X, Y) = 0$ if and only if $X$ and $Y$ are independent.
- Symmetry: $I(Y, X) = I(X, Y)$.
- $I(X, Y)$ is invariant to one-to-one transformations of $Y$ and $X$.
- MI is related to entropy via $I(X, Y) = H(Y) - H(Y|X)$, or, equivalently, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where, e.g., $H(Y) = -\mathrm{E}_Y\{\log f(Y)\}$ and $H(Y|X) = -\mathrm{E}_{(Y,X)}\{\log f(Y|X)\}$.

The **mutual information index** is defined as $\mathscr{I}(X, Y) = 1 - \exp(-2I(X, Y))$.

The **partial mutual information** between $X$ and $Y$, given $Z$, is defined as

$$I(X, Y|Z) = \mathrm{E}_{(X,Y,Z)}\left\{\log \frac{f(X, Y|Z)}{f(X|Z)f(Y|Z)}\right\}.$$

# Stationary random processes and their characteristics

- We assume that $\{X_t, t = 1, \dots\}$ is a strictly stationary and ergodic zero mean process, characterised by the autocovariance function $\gamma(k) = \mathsf{E}(X_t X_{t-k}), k = 0, \pm1, \pm2, \dots$
- $\mathbf{\Gamma}_k = \{\gamma(|i-j|, i, j = 1, \dots, k\}$ denotes the (auto)covariance matrix of $X_{t-k+1:t} = (X_{t-k+1}, X_{t-k+2}, \dots, X_{t-1}, X_t)$
- $\rho(k) = \gamma(k)/\gamma(0), k \in \mathbb{Z}$, the autocorrelation function (ACF) of $X_t$.
- The optimal linear predictor of $X_t$ based on $X_{t-k:t-1} = (X_{t-k}, \dots, X_{t-1})$,

$$\hat{X}_{kt} = \phi_{1k}X_{t-1} + \phi_{2k}X_{t-2} + \cdots + \phi_{kk}X_{t-k},$$

  has coefficients $\boldsymbol{\phi}_k = (\phi_{1k}, \dots, \phi_{kk})'$ equal to $\boldsymbol{\phi}_k = \mathbf{\Gamma}_k^{-1}\boldsymbol{\gamma}_k$, where $\boldsymbol{\gamma}_k = (\gamma(1), \gamma(2), \dots, \gamma(k))'$, and mean square prediction error $v_k = \mathsf{E}\{(X_t - \hat{X}_{t,k})^2\}$, given recursively as $v_k = v_{k-1}(1 - \phi_{kk}^2)$, with $v_0 = \gamma(0)$.
- The partial ACF (PACF) is $\phi_{kk} = \frac{\mathsf{Cov}(X_t - \hat{X}_{k-1,t}, X_{t-k} - \hat{X}^*_{k-1,t-k})}{\sqrt{\mathsf{Var}(X_t - \hat{X}_{k-1,t})\mathsf{Var}(X_t - \hat{X}^*_{k,t-k})}}$, $k = 1, 2, \dots,$

  where $\hat{X}^*_{k-1,t-k}$ is the linear predictor of $X_{t-k}$ based on $X_{t-k+1:t-1} = (X_{t-k+1}, X_{t-k+2}, \dots, X_{t-1})$.

For a Gaussian processes we have the enhanced interpretation and results:

- $\phi_{kk} = \frac{\text{Cov}(X_t, X_{t-k} | X_{t-1}, \ldots, X_{t-k+1})}{\sqrt{\text{Var}(X_t | X_{t-1}, \ldots, X_{t-k+1}) \text{Var}(X_{t-k} | X_{t-1}, \ldots, X_{t-k+1})}}$, $k = 1, 2, \ldots$.
- $I(X_t, X_{t+k}) = -\frac{1}{2} \log(1 - \rho^2(k))$, $\mathscr{I}(X_t, X_{t+k}) = \rho^2(k)$.
- $I(X_t, X_{t+k} | X_{t-1:t-k}) = -\frac{1}{2} \log(1 - \phi_{kk}^2)$, $\mathscr{I}(X_t, X_{t+k} | X_{t+1:t+k-1}) = \phi_{kk}^2$.

The coefficients $\phi_{kj}$ and thus the PACF ($j = k$) are computed by the Durbin–Levinson (DL) algorithm.

# The mutual information between past and future

## Theorem

*Let* $\pi(k) = I(X_t, X_{t+k}|X_{t+1}, X_{t+2}, \ldots, X_{t+k-1})$*, the partial mutual information of* $X_t$ *and* $X_{t+k}$*, given all the intermediate random variables. The mutual information between the* $n$ *past variables* $X_{1:n} = (X_1, X_2, \ldots, X_n)$ *and the* $m$ *future variables* $X_{n+1:n+m} = (X_{n+1}, X_{n+2}, \ldots, X_{n+m})$*, can be decomposed as follows:*

$$I(X_{1:n}, X_{n+1:n+m}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi(n+j-i). \tag{1}$$

Note that $\pi(k)$ is the expected conditional log copula density of $X_t$ and $X_{t+k}$, given the intermediate variables:

$$\pi(k) = \mathrm{E}_{(X_t,\ldots,X_{t+k})}\{\log c(F_{t+1:t+k-1}(X_t), F_{t+1:t+k-1}(X_{t+k}))\}.$$

where $f(X_t, X_{t+k}|X_{t+1:t+k-1}) = f(X_t|X_{t+1:t+k-1})f(X_{t+k}|X_{t+1:t+k-1})c(F_{t+1:t+k-1}(X_t), F_{t+1:t+k-1}(X_{t+k}))$ and $c(\cdot)$ is the copula density. It is a general measure of partial dependence for two random variables, which generalizes the notion of partial autocorrelation function.

Define

- $X_p = X_{-\infty:n}$ (the collection of random variables up to and including time $n$, the "past" of the process)
- $X_f^{(h)} = X_{n+h:\infty}$, $h \in \mathbb{Z}^+$ (the collection of future random variables, with a gap of $h$ time units).
- For $h = 1$, we write $X_f^{(1)} = X_f$.

By Theorem 1, we can provide the following generalization of the **mutual information between past and future** (MIPF), originally formulated for Gaussian processes (Ibragimov and Rozanov,2012):

$$I(X_p, X_f) = \sum_{k=1}^{\infty} k\pi(k).$$

This arises simply as the limit of $I(X_{-n:0}, X_{1:m})$ as $n, m \to \infty$.

A stationary random process is said to be **information regular** if $I(X_p, X_f^{(h)}) \to 0$ as $h \to \infty$, and **absolutely regular** if $I(X_p, X_f) < \infty$. Absolute regularity implies information regularity.

# Gaussian processes

For a Gaussian process the mutual information is a function of the (squared) partial autocorrelations, as it is shown by the following corollary.

## Corollary

*If* $\{X_t, t \in \mathbb{Z}\}$ *is a Gaussian process,*
$I(X_{1:n}, X_{n+1:n+m}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\log(1-\phi^2_{i+j-1,i+j-1}),$
$I(X_p, X_f) = -\frac{1}{2}\sum_{k=1}^{\infty} k\log(1-\phi^2_{kk}).$

## Example (Gaussian AR(1) process)

Let $X_t = \phi X_{t-1} + \epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,\sigma^2)$. Then, $I(X_p, X_f) = -\frac{1}{2}\log(1-\phi^2)$ and $\mathscr{I}(X_p, X_f) = \phi^2$.

## Example (Lognormal stochastic volatility process)

Let $X_t = \exp(Y_t/2)\epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$ and
$Y_{t+1} = \mu(1-\phi) + \phi Y_t + \eta_t, \eta_t \sim$ i.i.d. $\mathrm{N}(0,\sigma^2_\eta)$, independently of $\epsilon_t$. Then,
$I(X_p, X_f) = -\frac{1}{2}\log(1-\phi^2)$ and $\mathscr{I}(X_p, X_f) = \phi^2$.

## Example (Infinite mutual information)

Let $(1-B)^d X_t = \epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0, \sigma^2)$, $d < 1/2$. Then, $\phi_{kk} = d/(k-d)$. Hence $I(X_p, X_f) \to \infty$: long memory and noninvertible processes are not absolutely regular.

## Example (Gaussian ARCH(1) process)

Let $X_t = \sigma_t \epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$, $\sigma_t^2 = \omega + \alpha X_{t-1}^2$, $\omega > 0$, $0 \leq \alpha < 1$. Then,

$$\begin{aligned} \pi(1) &= I(X_t, X_{t+1}) \\ &= H(X_{t+1}) - H(X_{t+1}|X_t) \\ &= -\int f(X_{t+1}) \log f(X_{t+1}) dX_{t+1} - \frac{1}{2}\mathrm{E}_{X_t}(\log \sigma_{t+1}^2) - \frac{1}{2}\log(2\pi) \\ &\simeq \frac{1}{2}\log \omega - \frac{1}{2}\log(1-\alpha) - \frac{1}{2}\mathrm{E}(\log \sigma_t^2), \end{aligned}$$

where the last line follows from the Gaussian approximation of $f(X_{t+1})$.

# Optimal transformations: the most predictable aspects of time series

- Let
$$h_{1t} = h_1(X_t),\ h_{2t} = h_2(X_t),\ \ldots,\ h_{rt} = h_r(X_t)$$
denote a set of measurable functions of $X_t$, such that $\mathrm{E}(h_{jt}) = \mu_{hj}$, $\mathrm{Var}(h_{jt}) > 0$ and $|\mathrm{Cov}(h_{kt}, h_{jt})| < \sqrt{\mathrm{Var}(h_{kt})}\sqrt{\mathrm{Var}(h_{jt})}$.
- Let $\boldsymbol{h}_t$ denote the $r \times 1$ vector $\boldsymbol{h}_t = (h_{1t}, \ldots, h_{rt})'$. The cross-covariance matrix of $\boldsymbol{h}_t$ at lag $k$ is $\mathrm{Cov}(\boldsymbol{h}_t, \boldsymbol{h}_{t-k}) = \boldsymbol{\Gamma}_h(k), k \in \mathbb{Z}$.
- Hence, we assume that the set $\{h_j(X_t), j = 1, \ldots, r\}$ is non-singular, i.e., $\boldsymbol{\Gamma}_h(0)$ is positive definite.

Consider the process resulting from a monotonic transformation $Z_t = g(Z_t^*)$ of the contemporaneous aggregation of the elements of $\boldsymbol{h}_t$, with coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)'$, satisfying the a normalization constraint:

$$Z_t^* = \boldsymbol{\beta}' \boldsymbol{h}_t, \quad \boldsymbol{\beta}' \boldsymbol{\Gamma}_h(0) \boldsymbol{\beta} = 1. \tag{2}$$

For $g(\cdot)$ we consider two cases: the identity transformation, $Z_t = Z_t^*$, and the normalizing transformation $g(Z_t^*) = \Phi^{-1}(F_Z(Z_t^*))$, where $F_Z$ is the cumulative distribution function (cdf) of $Z_t^*$, and $\Phi$ is the standard normal cdf.

## Definition

The most predictable aspect of $X_t$ is the transformation $Z_t = g\left(\boldsymbol{\beta}'\boldsymbol{h}_t\right)$, with $\boldsymbol{\beta}$ satisfying the constraint $\boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1$, such that the mutual information between the past and future $I(\mathcal{Z}_p, \mathcal{Z}_f)$ is a maximum, where $\mathcal{Z}_p = \{Z_{n-j}, j \geq 0\}$ and $\mathcal{Z}_f = \{Z_{n+j}, j \geq 1\}$.

The second most predictable aspect of $X_t$ is the transformation $W_t = g\left(\boldsymbol{\zeta}'\boldsymbol{h}_t\right)$, such that $\boldsymbol{\zeta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 0$, $\boldsymbol{\zeta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\zeta} = 1$, and the mutual information between the past and future $I(\mathcal{W}_p, \mathcal{W}_f)$ is a maximum, where $\mathcal{W}_p = \{W_{n-j}, j \geq 0\}$ and $\mathcal{W}_f = \{W_{n+j}, j \geq 1\}$.

The most predictable aspects of the time series are difficult to evaluate, as they depend on the partial mutual information coefficients of $Z_t$, denoted $\pi_Z(k)$, that are difficult to estimate. A workable definition takes into consideration linear predictability.

## Definition

The most linearly-predictable aspect of $X_t$ is the transformation $Z_t = g\left(\boldsymbol{\beta}'\boldsymbol{h}_t\right)$, with $\boldsymbol{\beta}$ satisfying the constraint $\boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1$, which maximises the linear mutual information measure

$$I^*(\mathcal{Z}_p, \mathcal{Z}_f) = -\frac{1}{2}\sum_{k=1}^{\infty} \log(1 - \phi^2_{Z,kk}),$$

where $\phi_{Z,kk}$ denotes the PACF of $Z^*_t = \boldsymbol{\beta}'\boldsymbol{h}_t$, $\mathcal{Z}_p = \{Z^*_{n-j}, j \geq 0\}$ and $\mathcal{Z}_f = \{Z^*_{n+j}, j \geq 1\}$.

The second most linearly-predictable aspect is defined as in Definition 2 with reference to the target measure $I^*(\mathcal{W}_p, \mathcal{W}_f) = -\frac{1}{2}\sum_{k=1}^{\infty} \log(1 - \phi^2_{W,kk})$, where $\phi_{W,kk}$ is the PACF of $W^*_t = \boldsymbol{\zeta}'\boldsymbol{h}_t$.

# Basis functions

The vector $\boldsymbol{h}_t$ can be thought as a feature vector, and the choice of the functions $h_j(X_t)$ can be considered as context specific. However, we concentrate on sets of basis functions that can be used for the purpose of eliciting the most predictable aspect of a time series.

The basis functions are evaluated at location shifts of $X_t$, namely $X_t - q(\alpha_j)$, where

$$q(\alpha_j) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha_j\}, \quad j = 1, \dots, r,$$

is the quantile corresponding to the probability $\alpha_j \in (0, 1)$.

Some relevant choices are the following.

- *Hinge basis* functions with knots at the $r^*$ quantiles $q_j = q(\alpha_j)$, $\alpha_j = \frac{j}{r^*+1}$, such that,

  $$h_{2j-1}(X_t) = \max\{0, X_t - q_j\}, \;\; h_{2j}(X_t) = \max\{0, q_j - X_t\}, \quad j = 1, 2, \dots, r^*.$$

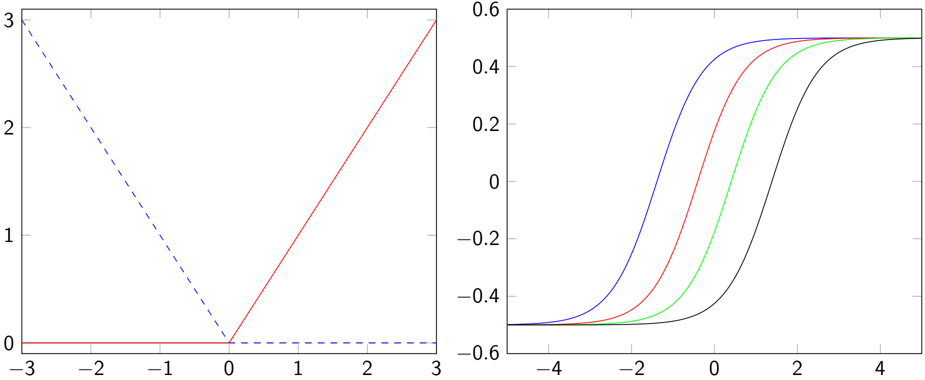
  There are $r = 2r^*$ basis functions.

- *Logistic basis.* Define

$$h_j(X_t) = \frac{1}{1 + \exp\left(-\frac{X_t - q_j}{\tau}\right)} - \frac{1}{2},$$

where $\tau > 0$ is a scale parameter, related to the variance of $X_t$ by $\tau = \pi^{-1}\sqrt{3\text{Var}(X_t)}$. The logistic transformation is bounded between -0.5 and 0.5.

- Natural cubic spline basis, consisting of $h_1(X_t) = 1$, $h_2(X_t) = X_t$, $h_j(X_t) = (X_t - q_j)_+^3$, $j = 3, \ldots, r^* + 2$. The coefficients are subject to the following natural boundary constraints: $\sum_{j=3}^{r^*+2} \beta_j = 0$, $\sum_{j=3}^{r^*+2} \beta_j q_{j-1} = 0$.




Left: Plot of $h(u) = \max\{0, u\}$ (solid red) and $h(u) = \max\{0, -u\}$ (dashed blue).
Right: Plot of $h_j(u) = \{1 + \exp((q_j - u)/\tau)\}^{-1}$, for $\tau = 1$ and $q_j = \ln(\alpha_j/(1-\alpha_j))$ and $\alpha_j = j/5, j = 1, 2, 3, 4$.

## Remark

*A variant of the above bases can be adopted when $X_t$ does not have a finite second moment, entailing a preliminary transformation of the original series. For instance, in the logistic case, let $X_t^* = L^{-1}(F(X_t))$, where $F$ is the CDF or $X_t$, which is estimated by the empirical CDF of $X_t$, and $L^{-1}(u) = \log(u/(1-u))$ is the standard (unit scale) logistic quantile function. Then, considering the quantiles of the standard logistic distribution $q_j(\alpha_j) = \log(\alpha_j/(1-\alpha_j))$, $\alpha_j = \frac{j}{r+1}$, we can set*

$$h_j(X_t) = \{1 + \exp(q_j - X_t^*)\}^{-1} - 0.5.$$

## Remark

*If $r$ is allowed to vary with $T$, the estimation of the optimal tranformation can be considered as a particular instance of the method of sieve extremum estimation, see, e.g., Chen and Shen (1998), and the references therein, and Gourieroux and Jasiak (2002) for applications to the estimation of nonlinear correlograms.*

# Statistical Inference

- Let $\{x_t, t = 1, \dots, T\}$ denote the observed time series.
- The quantile corresponding to the probability $\alpha_j$ is estimated by the empirical quantile of $x_t$.
- Denoting $\hat{\boldsymbol{h}}_t = (h_1(x_t), \dots, h_r(x_t))'$, the sample mean and covariance matrix of the vector $\boldsymbol{h}_t$ are respectively $\bar{\boldsymbol{h}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{h}_t$ and $\hat{\boldsymbol{\Gamma}}_h(0) = T^{-1}\sum_{t=1}^{T}\left(\boldsymbol{h}_t - \bar{\boldsymbol{h}}\right)\left(\boldsymbol{h}_t - \bar{\boldsymbol{h}}\right)'$.
- The vector $\boldsymbol{\beta}$ is estimated by maximizing the mutual information

$$\hat{Q}_T(\boldsymbol{\beta}) = -\frac{1}{2}\sum_{k=1}^{\lfloor 2\ell_T \rfloor} k \log\left(1 - \tilde{\phi}^2_{z,kk}(\boldsymbol{\beta})\right), \qquad (3)$$

  which is also a function of a bandwidth parameter, $\ell_T$, allowing for the truncation of the infinite sum.
- The coefficients $\tilde{\phi}_{z,kk}(\boldsymbol{\beta})$ are the regularized Durbin-Levinson estimators of the PACF of $z_t = \boldsymbol{\beta}'\boldsymbol{h}_t$ at lag $k$, under the constraint $\boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta} = 1$.

- For given $\boldsymbol{\beta}$, we construct $z_t$; letting $\hat{\phi}_{z,kk}(\boldsymbol{\beta})$ denote the sample PACF of $z_t$, then, the regularized PACF is $\tilde{\phi}_{z,kk}(\boldsymbol{\beta}) = w_k \hat{\phi}_{z,kk}(\boldsymbol{\beta})$, where the weight $w_k \in [0, 1]$ is obtained as $w_k = \kappa(k/\ell_T)$.
- Here, $\ell_T \in \mathbb{R}^+$ denotes the bandwidth parameter of the trapezoidal kernel $\kappa(u)$ defined as

$$\kappa(u) = \begin{cases} 1, & |u| \leq 1, \\ 2 - |u|, & 1 < |u| \leq 2, \\ 0, & |u| > 2. \end{cases} \tag{4}$$

- For the selection of the bandwidth we adopt a data-based selection criterion, which chooses $\hat{\ell}_T$ as the smallest value of $\ell_T$ such that

$$|\hat{\phi}_{z,kk}(\ell_T + k)| < c\{\log_{10} n/n\}^{1/2}, \; k = 1, \ldots, K_n, \; K_n = o(\log_{10} n). \tag{5}$$

We set $c = 2$ and $K_n = 5$. The rule amounts to conducting an approximate 95% simultaneous test of $\phi_{z,kk}(\ell_T + k) = 0$ $(k = 1, \ldots, K_n)$.

- In practice, the maximization of (3) is carried out by a numerical optimization routine handling nonlinear equality constraints, such as fmincon in Matlab.
- The initial value of $\hat{\boldsymbol{\beta}}$ is obtained from the eigenvector of $\boldsymbol{\Gamma}_h(0)$ (scaled by the square root of the corresponding eigenvalue) for which the mutual information of the corresponding $z_t$ variable is largest.
- For the selection of $r$ a criterion based MI (Li and Xie, 1996) can be used.

# Large sample properties

Assumptions

1. $X_t$ is strictly stationary with absolutely continuous marginal distribution function $F(x)$, with continous density $f(x)$, and $\upsilon_j$-quantiles $q_j = F^{-1}(\upsilon_j)$, $j = 1, \ldots, r$, such that $-\infty < a \leq q_1 < q_2 < \cdots < q_r \leq b < \infty$, and $0 < f(q_j) < \infty$.
2. $X_t$ is absolutely regular with strong mixing coefficient $\alpha_m$ of size -$\varphi_0$, with $\varphi_0 = 1 + \frac{1}{1+\delta}$, $\delta > 0$, and $\mathrm{E}|X_t|^{4+2\delta}$.
3. The set of basis functions is chosen so that their number if fixed and known, $\mathrm{E}|h_j(X_t)|^{4+2\delta} < \infty$, $\mathbf{\Gamma}_h(0)$ is non singular, and $h_j(X_t)$ is a Lipschitz continuous function of the quantile $q_j$.
4. The bandwidth parameter of the trapezoidal kernel is chosen so that $\ell_T = o(T^{1/4})$ and $\ell_T \geq r/2$.
5. Let $Q_0(\boldsymbol{\beta}) = \lim_{T\to\infty} \hat{Q}_T(\boldsymbol{\beta})$. The value $\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}\in\mathbb{B}} Q_0(\boldsymbol{\beta})$, where $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^r : \boldsymbol{\beta}'\mathbf{\Gamma}_h(0)\boldsymbol{\beta} = 1\}$, is unique (apart from a sign change), i.e., $\boldsymbol{\beta}_0$ is the unique fixed point of the nonlinear system $\boldsymbol{\beta} = \mathbf{\Gamma}_h^{-1}(0)g(\boldsymbol{\beta})/\left(\boldsymbol{\beta}'g(\boldsymbol{\beta})\right)$, where $g(\boldsymbol{\beta}) = \partial Q_0(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

## Theorem

*Under Assumptions 1-5,*

$$\hat{\beta} \to_p \beta. \quad (6)$$

*Also, denoting* $\hat{\mathbf{g}}_T(\beta) = \frac{\partial \hat{Q}_T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ *and* $\hat{\mathbf{G}}_T(\beta) = \frac{\partial^2 \hat{Q}_T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$*, and letting* $\boldsymbol{\Sigma}_0 = \lim_{T\to\infty} Var\left(\sqrt{T}\hat{\mathbf{g}}_T(\beta_0)\right)$ *and* $\mathbf{G}_0 = plim\left\{\hat{\mathbf{G}}_T(\beta_0)\right\}$,

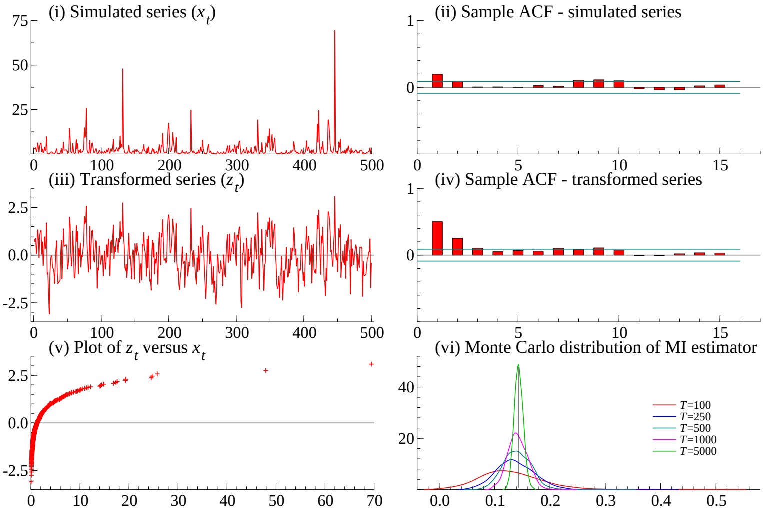
$$\sqrt{T}\left(\hat{\beta} - \beta_0\right) \to_d N(\mathbf{0}, \mathbf{E}_0 \boldsymbol{\Sigma}_0 \mathbf{E}_0'), \quad (7)$$

*where* $\mathbf{E}_0 = \mathbf{G}_0^{-1} - \frac{1}{\boldsymbol{\beta}_0' \boldsymbol{\Gamma}_h(0) \mathbf{G}_0^{-1} \boldsymbol{\Gamma}_h(0) \boldsymbol{\beta}_0} \mathbf{G}_0^{-1} \boldsymbol{\Gamma}_h(0) \beta_0 \beta_0' \boldsymbol{\Gamma}_h(0) \mathbf{G}_0^{-1}$.

# Lognormal AR(1)

- Consider the log-normal first order autoregressive process $X_t = e^{Y_t}, Y_t = 0.2 + 0.5Y_{t-1} + \epsilon_t, \epsilon_t \sim$ i.i.d. N$(0, 1)$, for which the mutual information is equal to 0.1438.
- The ability to estimate this value has been assessed via a Monte Carlo (MC) simulation experiment, according to which 1,000 simulated time series $x_t, t = 1, \dots, T$, with lengths $T = 100, 250, 500, 1000, 5000$ have been generated.
- The most predictable aspect have been estimated by adopting a hinge basis with $r^* = 3$ functions located at the quartiles of the marginal distribution of $x_t$, and the MI estimated by $\hat{Q}_T(\hat{\beta})$.

Log-normal AR(1). (i) Simulated series. (ii) Sample ACF of $x_t$. (iii) Transformed time series, $z_t$. (iv) Sample ACF of $z_t$. (v) Plot of $z_t$ versus $x_t$. (vi) Sampling distribution of the MI estimator $\hat{Q}_T(\hat{\beta})$, for $T = 100, 250, 500, 1000, 50000$.

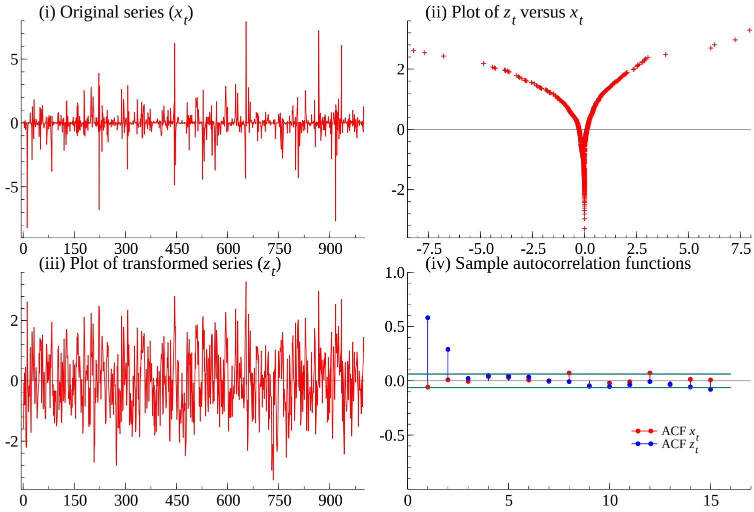


# Nonlinear MA(2) process

- The process $X_t = \epsilon_t\epsilon_{t-1}\epsilon_{t-2}, \epsilon_t \sim$ i.i.d. N(0, 1), is serially uncorrelated, but not independent: $X_t^2$ is positively autocorrelated at lags 1 and 2.
- A sample realization of size $T = 1,000$ is generated.
- The most predictable aspect using 2 hinge basis functions located at the median is

$$z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}\right)\right),$$

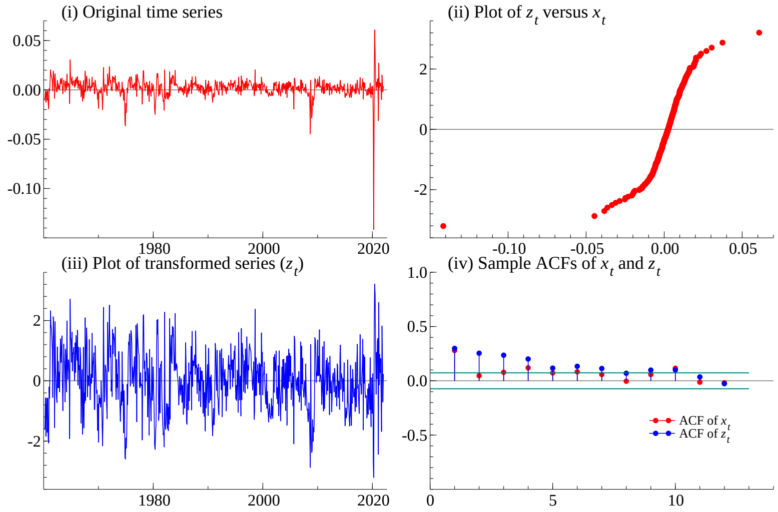
  where $\hat{F}_Z(z)$ is the empirical cdf of $z_t^* = 1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}$.
- The second best predictable aspect of the series (not shown) is a white noise process arising from a sigmoid-shaped monotonic transformation of the series.



(i) Original series ($x_t$)
5
0
-5
0 150 300 450 600 750 900
(ii) Plot of $z_t$ versus $x_t$
2
0
-2
-7.5 -5.0 -2.5 0.0 2.5 5.0 7.5
(iii) Plot of transformed series ($z_t$)
2
0
-2
0 150 300 450 600 750 900
(iv) Sample autocorrelation functions
1.0
0.5
0.0
-0.5
0 5 10 15
ACF $x_t$
ACF $z_t$

# US Index of Industrial Production

- The series considered for this illustration is the monthly growth of industrial production in the U.S. (Source: Board of Governors of the Federal Reserve System, https://fred.stlouisfed.org/), available for the period 1960.1-2022.1.
- For the analysis of this series we adopted a logistic basis with $r = 3$.
- The most predictable aspect of the series turns out to be a robust transformation of the series, cutting down the extreme values, see panel (ii).
- This is constructed as $z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19h_{1t} + 1.34h_{2t} + 1.19h_{3t}\right)\right)$, where $h_{jt} = 1/\{1 + \exp(-(x_t^* - \hat{q}_j^*)\}$. The estimated mutual information index is 0.22.
- The second most predictable aspect of the time series (not shown) is a measure of the volatility of the series, $w_t = \Phi^{-1}\left(\hat{F}_W\left(-5.81h_{1t} - 2.14h_{2t} + 8.21h_{3t}\right)\right)$. This is characterized by a sizable persistence in the autocorrelation function, and its mutual information index is estimated to be equal to 0.14.


(i) Original time series
0.05
0.00
-0.05
-0.10
1980
2000
2020
(ii) Plot of $z_t$ versus $x_t$
2
0
-2
-0.10
-0.05
0.00
0.05
(iii) Plot of transformed series ($z_t$)
2
0
-2
1980
2000
2020
(iv) Sample ACFs of $x_t$ and $z_t$
1.0
0.5
0.0
-0.5
0
5
10
ACF of $x_t$
ACF of $z_t$

# S&P500 index returns

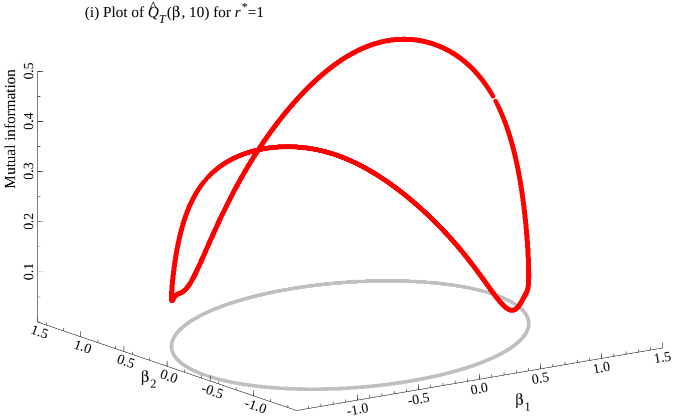
- We consider the time series of daily returns of the Standard & Poor 500 (SP500) stock market index from January 3, 1998, to March 11, 2022, for a total of $T = 6088$ observations.
- We considered a hinge basis function and the value maximising the MI selection criterion is $r^* = 1$. The MI index of $z_t^*$ is equal to 0.63.
- There are two basis functions, $h_{1t} = \max\{0, x_t - q_{0.5}\}, h_{2t} = \max\{0, q_{0.5} - x_t\}$ and $\ell_T = 10$, with covariance matrix

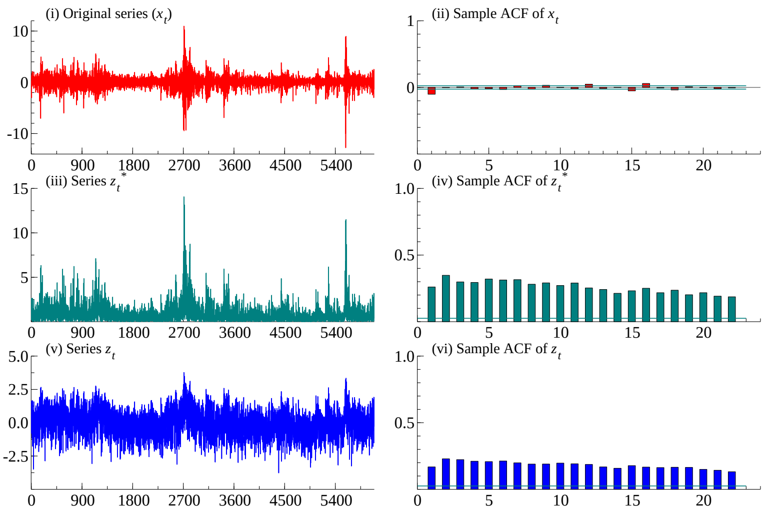
$$\hat{\mathbf{\Gamma}}_h(0) = \begin{pmatrix} 0.528 & -0.167 \\ -0.167 & 0.670 \end{pmatrix}.$$

The first eigenvector, scaled by the square root of the first eigenvalue (0.780), is $(0.624, -0.944)'$; the mutual information has a local maximum in the vicinity of it. The second eigenvector, scaled by the square root of the corresponding eigenvalue (0.418), is $(1.291, 0.850)'$

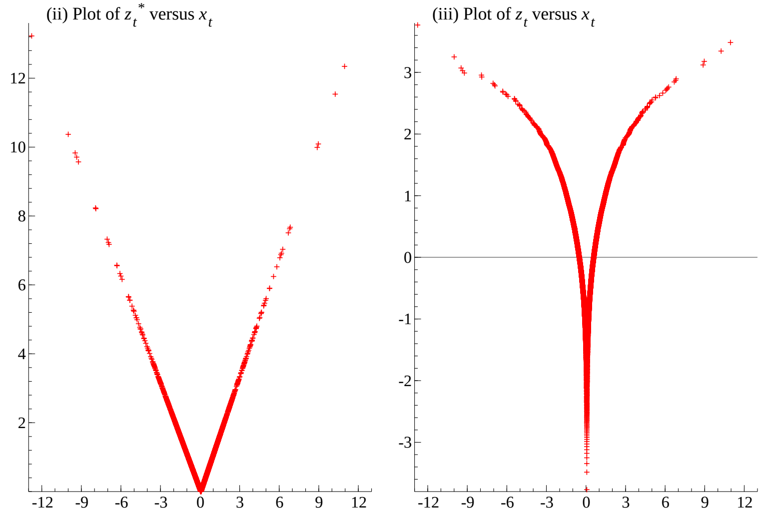
- The most predictable aspect of S&P 500 stock returns, $X_t$, is the volatility process
  $z_t = \Phi^{-1}\left(\hat{F}_Z(z_t^*)\right), z_t^* = 1.133\max\{0, x_t - q_{0.5}\} + 1.031\max\{0, q_{0.5} - x_t\}.$
- The second most predictable aspect $w_t$, orthogonal to the first is a robust level transformation
  $w_t = \Phi^{-1}\left(\hat{F}_W w_t^*\right), w_t^* = 0.879\max\{0, x_t - q_{0.5}\} - 0.746\max\{0, q_{0.5} - x_t\}.$
  It is characterized by a significant autocorrelation at lag 1, equal to -0.108, which is very close to the value of the first sample autocorrelation of the original time series (-0.102).

S&P 500 daily returns. (i) Plot of the mutual information as a function of $\beta$, $\hat{Q}_T(\beta, \ell_T)$, evaluated at the points $\beta$ such that $\beta' \hat{\Gamma}_h(0)\beta = 1$ (in grey), for $h_{1t} = \max\{0, x_t - q_{0.5}\}$, $h_{2t} = \max\{0, q_{0.5} - x_t\}$ and $\ell_T = 10$.




(i) Original series ($x_t$)
10
0
-10
0 900 1800 2700 3600 4500 5400
(ii) Sample ACF of $x_t$
1
0
0 5 10 15 20
(iii) Series $z_t^*$
15
10
5
0 900 1800 2700 3600 4500 5400
(iv) Sample ACF of $z_t^*$
1.0
0.5
0 5 10 15 20
(v) Series $z_t$
5.0
2.5
0.0
-2.5
0 900 1800 2700 3600 4500 5400
(vi) Sample ACF of $z_t$
1.0
0.5
0 5 10 15 20

S&P 500 daily returns. Plot of $z_t = \Phi^{-1}(z_t^*)$ and
$z_t^* = 1.133 \max\{0, x_t - q_{0.5}\} + 1.031 \max\{0, q_{0.5} - x_t\}$, versus $x_t$.



# Testing (un)predictability

The most predictable aspect $z_t$ can be used for testing the null of no predictability of the series, $H_0 : \mathrm{E}(Z_t|\mathcal{F}_{t-1}) = \mathrm{E}(Z_t)$. The idea is to apply a serial correlation test or an independence test to the series $z_t$.

This section reports the results of a Monte Carlo simulation experiment according to which

- we generate $M = 1000$ time series of length $T = 100, 250, 500, 1000, 5000$;
- for each series we determine the most predictable aspect, $z_t$, by using a set of $r = 1, 2, 3, 5$, hinge-basis functions;
- we test for (no) serial correlation using Hong's (1996) test statistic:

$$\mathcal{H}_T(\kappa) = T \sum_{j=1}^{T-1} \mathcal{K}^2(j/B_T)\hat{\rho}_z^2(j), \quad B_T = 3T^{\kappa},$$

  where $\mathcal{K}(j) = 0.5[1 + \cos(\pi u)]$, for $|u| \leq 1$, $\mathcal{K}(u) = 0$, for $|u| > 1$ is the Tukey-Hanning kernel, and $B_T$ is the bandwidth parameter.
- We consider 3 values of $\kappa$ (0.2, 0.3,0.4).

When appropriately standardized, the test statistic is asymptotically N(0,1).

Chen and Deo (2004) proposed a modification of Hong's test involving its power transformation, aiming at reducing the skewness of the distribution in finite samples. Their test statistic will be denoted $H_T^{\delta}(\kappa)$, where $\delta$ is a power parameter depending on the kernel moments.

We evaluate the empirical size of the test conducted at the 5% level for the following i.i.d. processes:

1. $X_t \sim$ i.i.d. $\mathrm{N}(0,1)$;
2. $X_t = \exp(\epsilon_t), \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$;
3. $X_t \sim$ i.i.d.$t_3$ (Student's-$t$ with 3 degrees of freedom);
4. $X_t \sim$ i.i.d. $\alpha$-stable with characteristic exponent 1, skewness parameter 0, location 0 and scale 1;
5. $X_t \sim$ i.i.d. $\alpha$-stable with characteristic exponent 1.5, skewness parameter 0.8, location parameter 0 and scale parameter 1.

r=1

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.077 | 0.064 | 0.057 | 0.069 | 0.067 |
| $\mathcal{H}_T(0.3)$ | 0.082 | 0.067 | 0.067 | 0.065 | 0.062 |
| $\mathcal{H}_T(0.4)$ | 0.111 | 0.092 | 0.070 | 0.061 | 0.064 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.058 | 0.047 | 0.042 | 0.040 | 0.049 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.081 | 0.061 | 0.060 | 0.053 | 0.055 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.135 | 0.083 | 0.063 | 0.056 | 0.067 |

r=2

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.091 | 0.099 | 0.086 | 0.073 | 0.091 |
| $\mathcal{H}_T(0.3)$ | 0.094 | 0.085 | 0.082 | 0.056 | 0.065 |
| $\mathcal{H}_T(0.4)$ | 0.127 | 0.096 | 0.087 | 0.069 | 0.059 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.059 | 0.060 | 0.050 | 0.040 | 0.055 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.079 | 0.062 | 0.064 | 0.058 | 0.055 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.127 | 0.088 | 0.073 | 0.056 | 0.052 |

Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^{\delta}(\kappa)$ tests of no predictability when the true model is $X_t \sim$ i.i.d. $N(0,1), t = 1, \dots, T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.069 | 0.075 | 0.094 | 0.075 | 0.071 |
| $\mathcal{H}_T(0.3)$ | 0.081 | 0.067 | 0.082 | 0.061 | 0.066 |
| $\mathcal{H}_T(0.4)$ | 0.101 | 0.078 | 0.083 | 0.066 | 0.056 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.055 | 0.061 | 0.074 | 0.045 | 0.055 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.077 | 0.056 | 0.074 | 0.049 | 0.050 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.107 | 0.074 | 0.084 | 0.060 | 0.053 |

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.107 | 0.091 | 0.101 | 0.094 | 0.102 |
| $\mathcal{H}_T(0.3)$ | 0.091 | 0.085 | 0.086 | 0.080 | 0.066 |
| $\mathcal{H}_T(0.4)$ | 0.120 | 0.093 | 0.082 | 0.073 | 0.063 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.064 | 0.060 | 0.068 | 0.055 | 0.060 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.088 | 0.062 | 0.068 | 0.062 | 0.050 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.121 | 0.080 | 0.075 | 0.058 | 0.053 |

Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is $X_t = \exp(\epsilon_t), \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), t = 1,\dots,T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.089 | 0.079 | 0.074 | 0.064 | 0.064 |
| $\mathcal{H}_T(0.3)$ | 0.088 | 0.065 | 0.053 | 0.056 | 0.047 |
| $\mathcal{H}_T(0.4)$ | 0.113 | 0.088 | 0.053 | 0.076 | 0.054 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.075 | 0.065 | 0.052 | 0.050 | 0.047 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.079 | 0.055 | 0.052 | 0.060 | 0.046 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.113 | 0.086 | 0.057 | 0.077 | 0.044 |

r=2

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.119 | 0.097 | 0.085 | 0.087 | 0.086 |
| $\mathcal{H}_T(0.3)$ | 0.118 | 0.082 | 0.079 | 0.061 | 0.067 |
| $\mathcal{H}_T(0.4)$ | 0.131 | 0.083 | 0.063 | 0.053 | 0.055 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.085 | 0.065 | 0.052 | 0.051 | 0.064 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.103 | 0.067 | 0.053 | 0.049 | 0.060 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.123 | 0.074 | 0.063 | 0.056 | 0.043 |

Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^{\delta}(\kappa)$ tests of no predictability when the true model is $X_t \sim$ i.i.d. $\alpha$-stable with ch. exponent 1.5, skewness parameter 0.8, location parameter 0 and scale parameter 1.

The empirical power is evaluated using the same experimental design, with reference to the following processes:

1. Non-Linear MA(2) process, $X_t = \epsilon_t \epsilon_{t-1} \epsilon_{t-2}$.
2. ARCH(1,1) process: $X_t = \sqrt{h_t}\epsilon_t, h_t = 0.5 + 0.8X_{t-1}^2 + 0.1X_{t-2}^2$.
3. GARCH(1,1) process: $X_t = \sqrt{h_t}\epsilon_t, h_t = 0.01 + 0.94X_{t-1}^2 + 0.05h_{t-1}$.
4. Threshold autoregressive process,

$$X_t = (-1.5X_{t-1} + \epsilon_t)I(X_{t-1} < 0) + (0.5X_{t-1} + \epsilon_t)I(X_{t-1} \geq 0).$$

5. Bilinear model $X_t = 0.6\epsilon_{t-1}X_{t-2} + \epsilon_t$.

Here, $\epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$.

r=1

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.963 | 0.999 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.946 | 0.999 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.928 | 0.997 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.955 | 0.998 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.936 | 0.997 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.916 | 0.997 | 1.000 | 1.000 | 1.000 |

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.812 | 0.984 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.746 | 0.970 | 0.999 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.701 | 0.947 | 0.999 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.777 | 0.977 | 1.000 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.710 | 0.962 | 0.999 | 1.000 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.684 | 0.941 | 0.997 | 1.000 | 1.000 |

Rejection frequency (empirical power) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is the non-linear MA(2) process $X_t = \epsilon_t\epsilon_{t-1}\epsilon_{t-2}, \ \ \epsilon_t \sim$ i.i.d. N(0, 1).

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.767 | 0.984 | 0.997 | 0.999 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.667 | 0.963 | 0.996 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.615 | 0.931 | 0.994 | 0.998 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.702 | 0.976 | 0.997 | 0.999 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.633 | 0.955 | 0.995 | 0.999 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.585 | 0.925 | 0.994 | 0.998 | 1.000 |

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.742 | 0.969 | 0.996 | 0.996 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.635 | 0.940 | 0.995 | 0.998 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.590 | 0.907 | 0.986 | 0.994 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.684 | 0.952 | 0.995 | 0.996 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.584 | 0.925 | 0.992 | 0.997 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.560 | 0.894 | 0.984 | 0.994 | 1.000 |

Rejection frequency (empirical power) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is the Gaussian ARCH(2) process $X_t = \sqrt{h_t}\epsilon_t, h_t = 0.5 + 0.8X_{t-1}^2 + 0.1X_{t-2}^2$.

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.181 | 0.356 | 0.638 | 0.885 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.164 | 0.353 | 0.651 | 0.895 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.153 | 0.333 | 0.646 | 0.890 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.125 | 0.284 | 0.591 | 0.868 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.133 | 0.323 | 0.611 | 0.884 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.147 | 0.303 | 0.624 | 0.883 | 1.000 |

r=2

| | $T=100$ | $T=250$ | $T=500$ | $T=1000$ | $T=5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.282 | 0.444 | 0.675 | 0.896 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.220 | 0.392 | 0.664 | 0.886 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.191 | 0.340 | 0.642 | 0.875 | 1.000 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.214 | 0.349 | 0.613 | 0.870 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.172 | 0.342 | 0.625 | 0.874 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.170 | 0.317 | 0.619 | 0.866 | 1.000 |

Rejection frequency (empirical power) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is the Gaussian GARCH(1,1) process $X_t = \sqrt{h_t}\epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), h_t = 0.1 + 0.05X_{t-1}^2 + 0.94h_{t-1}$.

r=1

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.099 | 0.219 | 0.392 | 0.736 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.081 | 0.151 | 0.266 | 0.525 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.107 | 0.138 | 0.181 | 0.347 | 0.991 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.071 | 0.146 | 0.309 | 0.634 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.078 | 0.113 | 0.208 | 0.446 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.116 | 0.128 | 0.149 | 0.299 | 0.988 |

r=2

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.148 | 0.270 | 0.474 | 0.788 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.122 | 0.189 | 0.331 | 0.592 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.133 | 0.161 | 0.221 | 0.387 | 0.994 |
| $\mathcal{H}_T^\delta(0.2)$ | 0.102 | 0.190 | 0.378 | 0.714 | 1.000 |
| $\mathcal{H}_T^\delta(0.4)$ | 0.099 | 0.154 | 0.253 | 0.506 | 1.000 |
| $\mathcal{H}_T^\delta(0.6)$ | 0.123 | 0.147 | 0.189 | 0.334 | 0.991 |

Rejection frequency (empirical power) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is the TAR(1) process $X_t = (-1.5X_{t-1} + \epsilon_t)I(X_{t-1} < 0) + (0.5X_{t-1} + \epsilon_t)I(X_{t-1} \geq 0)$.

r=1

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.374 | 0.760 | 0.970 | 0.998 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.317 | 0.655 | 0.931 | 0.998 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.284 | 0.527 | 0.865 | 0.997 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.310 | 0.700 | 0.962 | 0.998 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.269 | 0.599 | 0.915 | 0.998 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.253 | 0.490 | 0.852 | 0.996 | 1.000 |

r=2

| | $T = 100$ | $T = 250$ | $T = 500$ | $T = 1000$ | $T = 5000$ |
|---|---|---|---|---|---|
| $\mathcal{H}_T(0.2)$ | 0.393 | 0.734 | 0.956 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.3)$ | 0.328 | 0.635 | 0.916 | 1.000 | 1.000 |
| $\mathcal{H}_T(0.4)$ | 0.297 | 0.512 | 0.865 | 0.992 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.2)$ | 0.317 | 0.681 | 0.945 | 1.000 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.4)$ | 0.281 | 0.601 | 0.897 | 1.000 | 1.000 |
| $\mathcal{H}_T^{\delta}(0.6)$ | 0.278 | 0.469 | 0.840 | 0.989 | 1.000 |

Rejection frequency (empirical power) of Hong's $\mathcal{H}_T(\kappa)$ and $\mathcal{H}_T^{\delta}(\kappa)$ tests of no predictability when the true model is the Bilinear process $X_t = 0.6\epsilon_{t-1}X_{t-2} + \epsilon_t$.

# Conclusions

- This paper has defined and estimated the most predictable aspect of a time series in a linear sense, which is defined as the measurable transformation of the series which maximizes the linear mutual information between the past and the future.
- The most predictable feature can be used for testing the null of unpredictability.
- The next issue, left unexplored here, is how we can use the most predictable aspect, $Z_t$, to predict aspects of the original time series, $X_t$. This entails the local inversion of the nonlinear transformation relating the former to the latter, so as to map the predictions of $Z_t$ into those for $X_t$. A similar idea has been explored by McNeil(2021) in a different framework.