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

Tommaso Proietti
Università di Roma "Tor Vergata"

November 28, 2022

Abstract

Serial dependence and predictability are two sides of the same coin. The literature has considered alternative measures of these two fundamental concepts. In this paper we aim at distilling the most predictable aspect of a univariate time series, i.e., the one for which predictability is optimized. Our target measure is the mutual information between past and future of a random process, a broad measure of predictability that takes into account all future forecast horizons, rather than focusing on the one-step-ahead prediction error mean square error. The first most predictable aspect is defined as the measurable transformation of the series for which the mutual information between past and future 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 the process. 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. The second most predictable aspect can also be defined, subject to suitable orthogonality restrictions. We also investigate the use of the most predictable aspect for testing the null of no predictability.

Keywords: Mutual information, Nonlinear dependence, Canonical Analysis.

1 Introduction

The issue of measuring serial dependence is at the heart of time series analysis, being intimately related to the predictability of a random process. The traditional Bravais-Pearson autocorrelation function (ACF) provides essential information on the dependence structure of a Gaussian process. Outside the Gaussian class, it is less informative. In particular, zero correlation implies only lack of linear association, or no linear predictability. Other classical measures, such as Spearman's rank correlation coefficient and Kendall's tau, measure monotonic association and fail to detect more general forms of nonlinear dependence.

Alternative broader measures of serial dependence have been developed; for a recent overview see Tjøstheim et al. (2018). There is also a substantive literature dealing with testing the null of (serial) independence and iid-ness, reviewed in Teräsvirta et al. (2010, sec. 7.7).

For a stationary time series X_t, 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 concepts of distance covariance and correlation, introduced by Székely et al. (2007) (see also Székely and Rizzo, 2009). Fokianos and Pitsillou (2017, 2018) proposed a test of serial independence based on the auto-distance covariance function and 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 possibly nonlinear forms of serial dependence. See Edelmann et al. (2019) for a comprehensive review of these developments.

Escanciano and Velasco (2006) proposed conditional mean dependence measures based on the covariance between X_t and the characteristic function of X_{t-k}, and a test of the martingale difference hypothesis based on the sample spectral distribution function. Shao and Zhang (2014) measured the degree of conditional mean independence of X_t from its past by the martingale difference correlation. Linton and Whang (2007) introduced a measure of directional predictability named the quantilogram. Otneim and Tjøstheim (2021) introduced a measure of conditional dependence, the locally Gaussian partial correlation. The Gini autocovariance function, measuring the degree of monotonicity in the relationship between X_t and X_{t-k}, was proposed by Carcea and Serfling (2015).

The generality of many of such measures is such that the most interpretable outcome is: *do we fail to reject (conditional mean) independence*? The answer for many applications, e.g., in economics and finance, 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. For asset returns time series it coincides with the volatility, which is a monotonic transformation of $|X_t|$, although we should allow for asymmetries in the volatility process due, for instance to leverage or asymmetric shocks. A second possibility is a robust transformation of the series, which reduces the effect that outlying observations exert on the dependence measure.

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 analysis, see 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 contributions see Reshef et al. (2011), who proposed the maximal information coefficient (MIC), and Kinney and Atwal (2014). As any general measure of dependence, it is difficult to estimate. In particular, we need to estimate the joint pdf of possibly high-dimensional vector random variables. It is available in closed form only for Gaussian r.v.s. Popular estimation methods rely on vine copulae, and nonparametric density estimates by the k-NN method, see Berrett and Samworth (2019) (2020). For alternative dependence measures based on alternative divergence functionals, see Skaug and Tjøstheim (1993), Granger et al. (2004) and Geenens and Lafaye de Micheaux (2020).

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). The asymptotic theory for their kernel based nonparametric estimators has been established by Hong and White (2005).

Our aim is to determine the transformation of a stationary stochastic process X_t, for which the MI between the past and the future of a time series is a maximum. Our approach is related to Gourieroux and Jasiak (2002), who proposed the nonlinear autocorrelogram, the nonlinear transformation which maximizes the autocorrelation at selected lags, and to Owen (1983), who develops the optimal transformation of an autoregressive processes by an adaptation of the alternating conditional expectation algorithm (Breiman and Friedman, 1985).

We contribute to the above referenced literature in several ways. First, we adopt 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 square integrable 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 or a monotonic transformation thereof. 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.

The paper is structured in the following way. In the next section we review the definition and the properties of the mutual information between two sets of random variables. Section 3 states the main assumptions about the univariate stochastic process under consideration, defines the mutual information between past and future, and deals with its evaluation in the special case of a Gaussian process. Section 2 defines the most predictable aspect of time series and presents alternative basis functions for eliciting it. Estimation and statistical inference is presented in section B. Section 6 illustrates our methodology. Finally, section 7 proposes a test for unpredictability. In section 8 we draw some conclusions.

2 Mutual Information: Definitions and Properties

Let X and Y denote a pair of possibly multivariate continuous random variables with probability density function (pdf) $f(X,Y)$ and marginal densities $f(X)$ and $f(Y)$, respectively. The mutual information (MI) 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, for any measurable function $g(U)$ of U with pdf $f(U)$, $\mathrm{E}_U(g(U)) = \int_{-\infty}^{\infty} g(u)f(u)du$; $I(X,Y)$ is interpreted as the Kullback-Leibler distance between the joint distribution and product of the marginal distribution.

The MI has the following main 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, see Granger and Lin (1994, Theorem 3).
- It 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 $H(Y|X)=-\mathrm{E}_{(Y,X)}\{\log f(Y|X)\}$.

Kinney and Atwal (2014) show that MI satisfies self-equitability: if $g(X)$ is a deterministic function of X and Y is conditionally independent of X given $g(X)$, then $I(g(X),Y)=I(X,Y)$.

The mutual information index is defined as $\mathscr{I}(X,Y)=1-\exp(-2I(X,Y))$. It provides a measure of association satisfying the properties of an ideal measure of dependence established by Rényi (1959) with the following properties: i. $0\le\mathscr{I}(X,Y)\le 1$, ii. $\mathscr{I}(X,Y)=0$ if X and Y are independent, iii. if $u(X)=v(Y)$ for u and v measurable functions then $\mathscr{I}(X,Y)=1$.

Finally, the conditional or partial mutual information (PMI) between X and Y, given the random variable Z, is defined as $I(X,Y|Z)=\mathrm{E}_{(X,Y,Z)}\left\{\log\frac{f(X,Y|Z)}{f(X|Z)f(Y|Z)}\right\}$.

3 Stationary random processes and their characteristics

Let $\{X_t, t=1,\dots\}$ be a strictly stationary zero mean process, with continuous density $f(X_t)$ and characterised by the autocovariance function $\gamma(k)=\mathrm{E}(X_tX_{t-k})<\infty, k=0,\pm1,\pm2,\dots$.

We denote by $\mathbf{\Gamma}_k=\{\gamma(|i-j|,i,j=1,\dots,k\}$ the autocovariance matrix of $X_{t-k+1:t}=(X_{t-k+1},X_{t-k+2},\dots,X_{t-1},X_t)$ and by $\rho(k)=\gamma(k)/\gamma(0), k\in\mathbb{Z}$ 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 $v_k=\mathrm{E}\{(X_t-\hat{X}_{t,k})^2\}$, given recursively as $v_k=v_{k-1}(1-\phi_{kk}^2)$, with $v_0=\gamma(0)$. The partial ACF (PACF) is $\phi_{kk}=\frac{\mathrm{Cov}(X_t-\hat{X}_{k-1,t},X_{t-k}-\hat{X}^*_{k-1,t-k})}{\sqrt{\mathrm{Var}(X_t-\hat{X}_{k-1,t})\mathrm{Var}(X_t-\hat{X}^*_{k,t-k})}}$, $k=1,2,\dots$, 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:

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

3.1 The mutual information between past and future

We now turn our consideration to the mutual information between the past and the future of a stochastic process. The following theorem establishes that it can be obtained as the sum of the pairwise partial mutual information of the random variables involved.

Theorem 1. *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:*

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

Remark 1. *The mutual information between* $X_{1:n}$ *and* $X_{n+1:n+m}$ *is equal to sum of the elements of the* $n \times m$ *Hankel matrix* $\mathbf{\Pi}_{mn}$ *with elements* $\Pi_{ij} = \pi(i+j-1), i = 1, \ldots, n, j = 1, \ldots, m$.

Remark 2. *The conditional mutual information* $\pi(k)$ *is the expected conditional log copula density of* X_t *and* X_{t+k}*, given the intermediate variables:*

$$\pi(k) = \int \cdots \int f(X_{t:t+k}) \ln c(F_{t+1:t+k-1}(X_t), F_{t+1:t+k-1}(X_{t+k})) dX_t \cdots dX_{t+k},$$

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

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.*

Denote by $X_p = X_{-\infty:n}$ the collection of random variables up to and including time n (generically, the "past" of the process) and by $X_f^{(h)} = X_{n+h:\infty}$, $h \in \mathbb{Z}^+$ 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, originally formulated for Gaussian processes (Ibragimov and Rozanov, 2012; Jewell and Bloomfield, 1983; Jewell et al., 1983; Pourahmadi, 2001),

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

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

An important concept related to mutual information is the information regularity of a stochastic process (Ibragimov and Rozanov, 2012). 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.

3.2 The Gaussian case

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

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

The partial autocorrelation sequence is computed by the Durbin-Levinson algorithm. The MI can be alternatively obtained via canonical correlation analysis; see Appendix C for details.

Example 1. 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 2. 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 ϵ_t. Then, $I(X_p, X_f) = -\frac{1}{2}\log(1-\phi^2)$ and $\mathscr{I}(X_p, X_f) = \phi^2$.

Example 3. 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 information regular.

The next example illustrates the difficulty of obtaining analytical formulae for a class of conditionally Gaussian processes.

Example 4. ARCH(1) process Let $X_t = \sigma_t\epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$, $\sigma^2_t = \omega + \alpha X^2_{t-1}$, $\omega > 0$, $0 \le \alpha < 1$. Then,

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

where the last line follows from the Gaussian approximation of $f(X_{t+1})$. Also, $\pi(k) = 0$ for all $k > 1$, and thus $I(X_{1:n}, X_{n+1:n+m}) = \pi(1)$.

The MIPF provides a measure of predictability across all possible future forecast horizons. For a Gaussian process, recalling that $v_n = \mathrm{Var}(X_{n+1}|X_{1:n})$ and $v_0 = \gamma(0)$, it holds that $v_n = v_0 \prod_{k=1}^n (1-\phi^2_{kk})$. An index of one-step-ahead mean square predictability is the following:

$$\mathscr{P} := 1 - \frac{v_n}{v_0} = \mathscr{I}(X_{1:n}, X_{n+1}),$$

see Jewell and Bloomfield (1983), and Pourahmadi and Miamee (1992). Similarly, the h-step ahead predictability, if $v_n^{(h)} = \mathrm{Var}(X_{n+h}|X_{1:n})$ is

$$\mathscr{P}(h) := 1 - \frac{v_n^{(h)}}{v_0} = \mathscr{I}(X_{1:n}, X_{n+h}).$$

Assumption 1. *X_t is absolutely regular, i.e., $\sum_{k=1}^\infty k\pi(k) < \infty$.*

4 Optimal transformations: the most predictable aspects of time series

Let

$$h_{1t} = h_1(X_t),\ h_{2t} = h_2(X_t),\ \dots,\ h_{rt} = h_r(X_t)$$

denote a set of Borel measurable functions, such that $\mathrm{E}(h_{jt}) = \mu_{hj}$, $\mathrm{Var}(h_{jt}) > 0$ and $|\mathrm{Cov}(h_{kt}, h_{jt})| < \sqrt{\mathrm{Var}(h_{kt})}\sqrt{\mathrm{Var}(h_{jt})}$, and let $\mathbf{h}_t$ denote the $r\times 1$ vector $\mathbf{h}_t = (h_{1t}, \dots, h_{rt})'$. The cross-covariance matrix of $\mathbf{h}_t$ at lag k is $\mathrm{Cov}(\mathbf{h}_t, \mathbf{h}_{t-k}) = \mathbf{\Gamma}_h(k), k \in \mathbb{Z}$.

Assumption 2. *The set of measurable transformations $h_j(X_t)$ is non-singular, i.e., $\mathbf{\Gamma}_h(0)$ is positive definite with distinct eigenvalues.*

Assumption 2 rules out affine transformations of X_t, e.g., $h_j(X_t) = a_j + b_j X_t$. If the transformation and nonlinear then Ass. 2 is implied by the absolute continuity of the marginal density $f(X_t)$.

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

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

Notice that Z_t^* has unit variance. The additional constraint $\mathrm{E}(Z_t^*) = 0$ can be enforced by imposing the linear constraint $\boldsymbol{\beta}'\mathrm{E}(\mathbf{h}_t) = 0$, but this is not essential.

For $g(\cdot)$ we consider two cases: the identity transformation, $Z_t = Z_t^*$ and the normalizing transformation $g(Z_t^*) = \Phi^{-1}\left(F(Z_t^*)\right)$, where F is the cumulative distribution function (CDF) of Z_t^*, estimated by the empirical CDF, and Φ is the standard normal CDF.

Definition 1. *The most predictable aspect of* X_t *is the transformation* $Z_t = g\left(\sum_{j=1}^r \beta_j h_{jt}\right)$, *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* $\mathcal{Z}_f = \{Z_{n+j}, j \geq 1\}$.

Definition 2. *The second best predictable aspect of* X_t *is the transformation* $W_t = g\left(\sum_{j=1}^r \theta_j h_{jt}\right)$, *such that* $\theta'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 0$, $\theta'\boldsymbol{\Gamma}_h(0)\theta = 1$, *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* $\mathcal{W}_f = \{W_{n+j}, j \geq 1\}$.

4.1 Basis functions

The vector $\mathbf{h}_t$ can be thought as a feature vector, and the choice of the functions $h_j(X_t)$ can be considered 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

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

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 ,

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

  There are $r = 2r^*$ basis functions. The transformation encompasses the identity transformation, $Z_t^* = X_t$, which occurs if $\beta_{2j-1} = 1$ and $\beta_{2j} = -1$, the absolute value transformation, $Z_t^* = |X_t|$, if j is odd and $\beta_{r^*} = \beta_{r^*+1} = 1$ and $\beta_j = 0$ for $j \neq (r^*, r^* + 1)$.

- *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 $\tau = \pi^{-1}\sqrt{3\mathrm{Var}(X_t)}$. The logistic transformation is bounded between -0.5 and 0.5.

- Natural cubic spline basis, consisting of $h_1(X_t) = 1$, $h_2(X_t) = X_t$, $h_j(X_t) = (X_t - q_j)^3_+$, $j = 3, \dots, r^*+2$. The coefficients are subject to the following natural boundary constraints: $\sum_{j=3}^{r^*+2} \beta_j = 0$, $\sum_{j=3}^{r^*+2} \beta_j q_{j-1} = 0$, see Hastie et al. (2009, Chapter 5)

Other bases are obviously possible. Hard and soft thresholding transformation and Huber's ψ function can be used as well. As another instance, if the range of X_t is defined as $R = q_{0.99} - q_{0.01}$, we could adopt the set of Fourier pairs $\left\{\cos\left(\frac{2\pi}{R} j X_t\right), \sin\left(\frac{2\pi}{R} j X_t\right)\right\}$, for $j = 1, 2, \dots, r/2$.

Remark 3. *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*

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

A polinomial basis, such as a cubic spline basis, possibly considering natural boundary constraints, could be considered after performing a normalizing transformation of X_t.

An alternative interesting direction is to adopt the set of check functions $\{\alpha_j - I(x_t < q_j), j = 1, \dots, r\}$, that define Linton and Whang (2007) quantilogram.

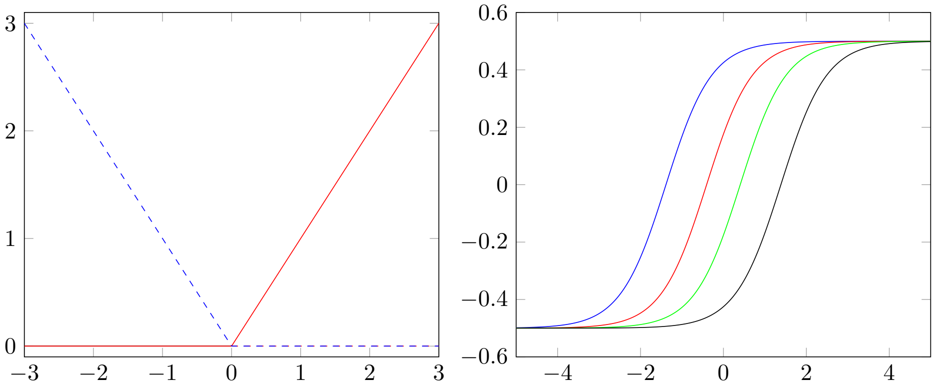

Figure 1: Left: Plot of $h(u) = \max\{0, u\}$ (solid red) and $h(u) = \max\{0, -u\}$ (dashed blue). Right: Plot of $h_j(u) = \{1 + \exp((q_j - u)/\tau)\}^{-1}$, for $\tau = 1$ and $q_j = \ln(\alpha_j/(1-\alpha_j))$ and $\alpha_j = j/5, j = 1, 2, 3, 4$.

The left plot of figure 1 displays the generic constituent pair of the hinge basis, $h(u) = \max\{0, u\}$ (solid red) and $h(u) = \max\{0, -u\}$ (dashed blue).The right plot displays the logistic functions obtained by setting $\tau = 1$ and choosing $q_j = \ln(\alpha_j/(1-\alpha_j))$, $\alpha_j = j/5, j = 1, 2, 3, 4$, i.e., the quintiles of the logistic distribution.

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., X. Chen and Shen (1998), and the references therein, and Gourieroux and Jasiak (2002) for applications to the estimation of nonlinear correlograms.

Appendix D shows that if $g(\cdot)$ is the identity transformation, i.e., $Z_t = \sum_{j=1}^{r} \beta_j h_{jt}$, the problem of estimating the most predictable aspect of the time series can be traced back to a nonlinear canonical correlation analysis of past and future of h_t.

5 Statistical Inference

Let $\{x_t, t = 1, \ldots, T\}$ denote the observed time series. The quantile corresponding to the probability α_j is estimated as the minimizer of the total check loss function

$$\hat{q}_j = \arg\min_{q \in \mathbb{R}} \sum_{t=1}^{T} \ell_{\alpha_j}(x_t - q), \tag{3}$$

where $\ell_{\alpha_j}(u) = u\{\alpha_j - I(u < 0)\}$. Then, with a slight abuse of notation, denote $\mathbf{h}_t = (h_1(x_t), \ldots, h_r(x_t))'$. The sample mean and the covariance matrix of the vector $\mathbf{h}_t$ are respectively $\bar{\mathbf{h}} = T^{-1} \sum_{t=1}^{T} \mathbf{h}_t$ and $\hat{\boldsymbol{\Gamma}}_h(0) = T^{-1} \sum_{t=1}^{T} \left(\mathbf{h}_t - \bar{\mathbf{h}}\right)\left(\mathbf{h}_t - \bar{\mathbf{h}}\right)'$.

The vector $\boldsymbol{\beta}$ is estimated by maximizing the mutual information

$$Q_T(\boldsymbol{\beta}, l) = -\frac{1}{2} \sum_{k=1}^{2\lceil l \rceil + 1} k \log\left(1 - \tilde{\phi}^2_{z,kk}(\boldsymbol{\beta})\right), \tag{4}$$

which is also a function of a bandwidth parameter, l, 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 $z_t = \boldsymbol{\beta}'\mathbf{h}_t$ 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 $w_k = \kappa(k/l)$. Here $l \in \mathbb{R}^+$ denotes the bandwidth parameter of the trapezoidal kernel $\kappa(u)$ defined as

$$\kappa(u) = \begin{cases} 1, & |u| \le 1, \\ 2 - |u|, & 1 < |u| \le 2, \\ 0, & |u| > 2. \end{cases} \tag{5}$$

The kernel weights are thus equal to 1 for $k \le l$, decrease linearly to zero for $l < k \le< 2l+1$, and are identically zero for $k > 2l + 1$. By construction, $Q_T(\boldsymbol{\beta})$ is a finite sum, since the regularized partial autocorrelations are zero after lag $2\lceil l \rceil + 1$.

The bandwidth is an important parameter. As it is shown in Proietti and Giovannelli (2018), the optimal choice of the bandwidth depends on the rate of decay of the autocovariance function of Z_t, $\gamma_z(j)$. In practice, given $\boldsymbol{\beta}$, its value can be estimated from z_t. Several selection criteria can be used. Proietti and Giovannelli (2018) adopt a data-based selection criterion, adapted from McMurry and Politis (2010), which chooses $\hat{l}_\beta$ as the smallest value of l such that

$$|\hat{\phi}_{z,kk}(l + k)| < c\{\log_{10} n/n\}^{1/2},\ k = 1, \ldots, K_n,\ K_n = o(\log_{10} n). \tag{6}$$

For the sample sizes typically used in applied work, McMurry and Politis recommend $c = 2$ and $K_n = 5$. The rule amounts to conducting an approximate 95% simultaneous test of $\phi_{z,kk}(l+k) = 0$ $(k = 1, \ldots, K_n)$. See also Politis (2003).

Alternatively, the selection of the bandwidth can be made according to a modified Akaike Information Criterion (AIC). In particular, letting $L = \lceil 2l \rceil$, we select $\hat{l}_\beta$ as the minimiser of

$$AIC_C(l) = T \ln \tilde{v}_L + T \frac{1 + L/T}{1 - (L+2)/T} \tag{7}$$

where $\tilde{v}_L = \hat{\gamma}_z(0) \prod_{k=1}^{L-1}\{1 - \tilde{\phi}^2_{z,kk}(\boldsymbol{\beta})\}$ is the prediction error variance of the RDL predictor of Z_t based on L past observations and using bandwidth $\hat{l}$.

In practice, the maximization of (4) 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.

The selection of r For the selection of r an information criterion based on Li and Xie (1996) LIC can be used

$$Q_T(\hat{\boldsymbol{\beta}}, \hat{l}_\beta) - \frac{c \log\log(T)}{T}\left(\hat{L}(\hat{L}+1)/2 + r\right),$$

where $\hat{L} = \lceil 2\hat{l}_\beta \rceil$ and $c > 2$. The rationale is that we add a penalty for the number of elements in the basis, r.

6 Illustrations

6.1 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. $\mathrm{N}(0,1)$, 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 $Q_T(\hat{\boldsymbol{\beta}}, \hat{l}_\beta)$.

Figure 2 displays in the first panel a simulated series with $T = 500$ and in (ii) its sample ACF. The estimated transformation, plotted in panel (v), is essentially the logarithmic transformation. The z_t series is plotted in panel (iii) and its ACF (panel (iv)) displays larger autocorrelations, the largest being close to 0.5. The ability to estimate the true MI is considered in the last panel, which shows the MC sampling distribution for different sample sizes.

6.2 Nonlinear MA(2) process

The process $X_t = \epsilon_t \epsilon_{t-1} \epsilon_{t-2}, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1)$, is serially uncorrelated, but not independent, as X_t^2 is positively autocorrelated at lags 1 and 2. Figure 3 displays a series of length $T = 1000$ generated by this process, along with its ACF, which displays no statistically significant autocorrelations. Interestingly, the second most predictable aspect, which is the level of the series is unpredictable, and has mutual information close to zero. The optimized value $Q_T(\hat{\boldsymbol{\beta}}, \hat{l}_\beta)$ did not vary with the choice of l and we plot the tranformation

$$z_t = 1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}$$

versus the original series and versus time. The transformation removes the concentration of values around zero and unearths the serial correlation, and in particular the second order moving average feature.

The second best predictable aspect of the series (not shown) is a white noise process arising from a sigmoid transformation of the series.

6.3 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/),

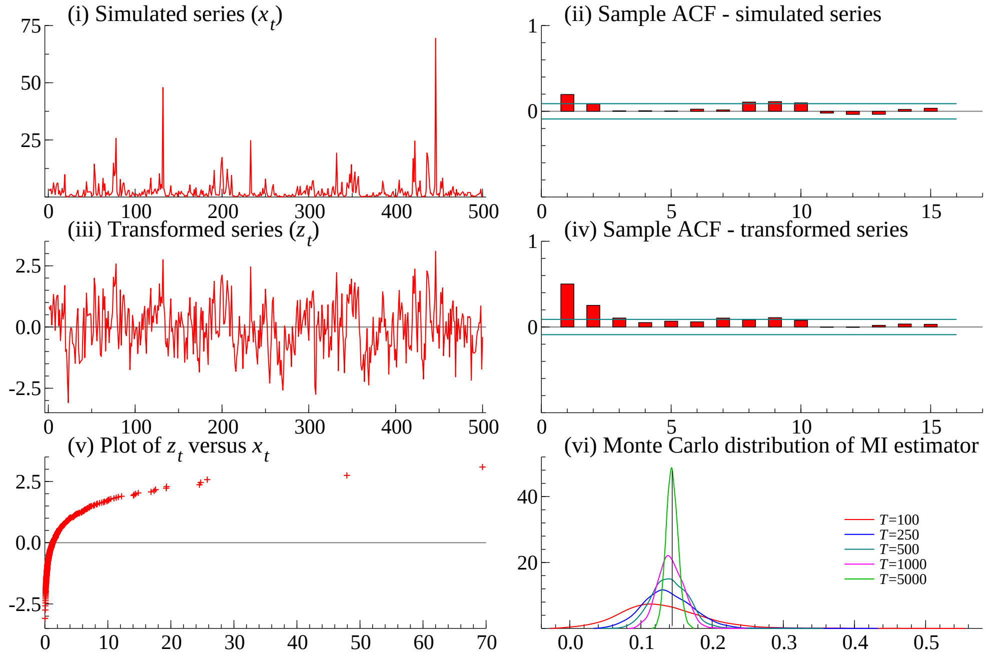

Figure 2: Log-normal AR(1). (i) Simulated series, x_t, of length $T = 500$ observations, generated by the lognormal process $X_t = e^{Y_t}, Y_t = 0.2 + 0.5Y_{t-1} + \epsilon_t, \epsilon_t \sim$ i.i.d. N(0, 1). (ii) Sample ACF of x_t. (iii) Transformed time series, z_t. (iv) Sample ACF of z_t. (v) Plot of z_t versus x_t. (vi) Kernel density estimates of the sampling distribution of the MI estimator $Q_T(\hat{\boldsymbol{\beta}}, \hat{l}_\beta)$, for $T = 100, 250, 500, 1000, 50000$.

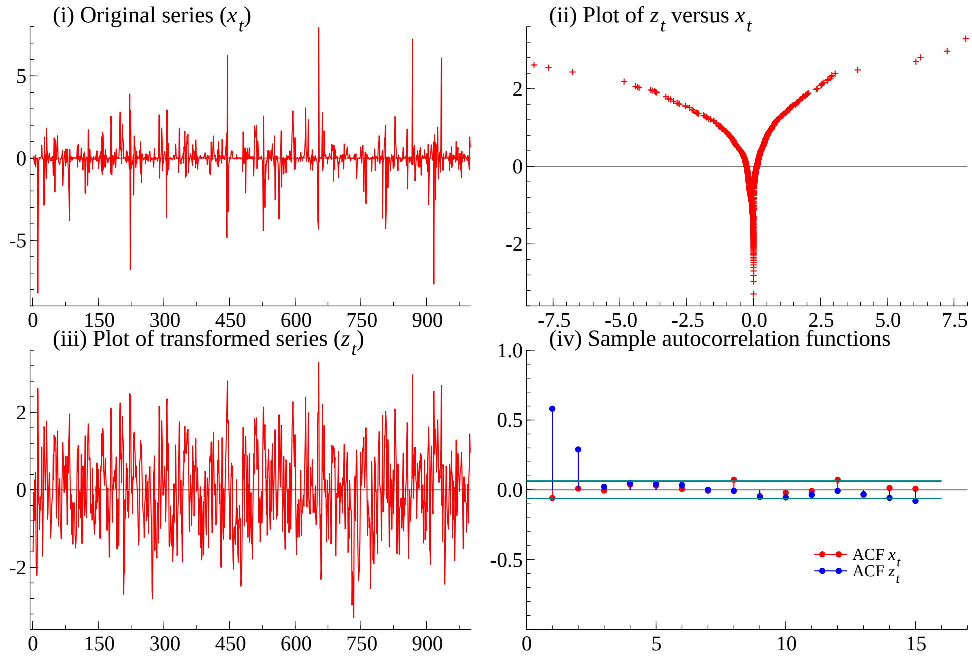

Figure 3: Nonlinear MA(2) process (i) Plot of the simulated series, x_t, $t = 1, \dots, 1000$. (ii) Plot of $z_t = 1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}$ versus x_t. (iii) Time series plot of z_t. (iv) Sample ACFs of x_t (red) and z_t (blue).

available for the period 1960.1-2022.1. The autocorrelation structure of the series is strongly affected by the downfall and subsequent recovery following the Covid-19 pandemic, as it is seen from panel (i) of figure 4. For the analysis of this series we adopted a logistic basis, although the main results are quite insensitive to this choice. We present the results for $r = 3$.

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). The transformed series z_t is homoscedastic and displays stronger autocorrelations than the original time series. This is constructed as $z_t = \Phi^{-1}(1.19h_{1t} + 1.34h_{2t} + 1.19h_{3t})$, where $h_{jt} = 1/\{1+\exp(-(x_t^* - \hat{q}_j^*)\}$, $x_t^* = \log(F_T(x_t)/(1 - F_T(x_t))$. The estimated mutual information index is 0.22.

The second most predictable aspect of the time series is $w_t = \Phi^{-1}(-5.81h_{1t} - 2.14h_{2t} + 8.21h_{3t})$, which is a measure of volatility. This is characterized by a sizable persistence in the autocorrelation function, and the mutual information index is 0.14.

6.4 S&P500 index returns

Figure 5, panel (i), displays 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 mutual information index of z_t^* is equal to 0.63. The top graph of figure 6 displays the values of the objective function $Q_T(\boldsymbol{\beta}, 10)$ as a function of $\boldsymbol{\beta}$, evaluated at the points $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta} = 1$ (in grey), for $h_{1t} = \max\{0, x_t - q_{0.5}\}, h_{2t} = \max\{0, q_{0.5} - x_t\}$ and $l = 10$. The covariance matrix of the two functions is

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

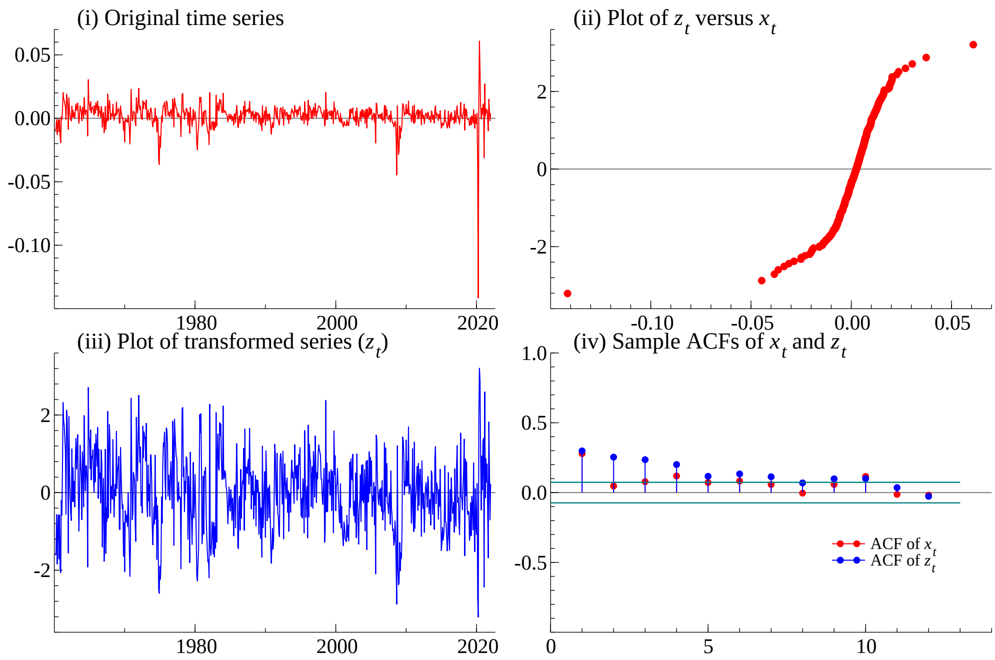

Figure 4: US Index of Industrial Production: relative changes with respect to the previous month. (i) Plot of the original time series (x_t). (ii) Plot of $z_t = \Phi^{-1}(1.19h_{1t} + 1.34h_{2t} + 1.19h_{3t})$, where $h_{jt} = 1/\{1 + \exp(-(x_t^* - \hat{q}_j^*)\}$, $x_t^* = \log(F_T(x_t)/(1 - F_T(x_t))$, versus the original x_t. (iii) Time series plot of the transformed series (z_t). (iv) Sample ACFs of x_t (red) and z_t (blue).

The sample cross-correlation between h_{1t} and h_{2t} is equal to -0.28. 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 mutual information has a local maximum in the vicinity of it, barely visible from figure 6.

The most predictable aspect of S&P 500 stock returns, X_t, is the volatility process

$$z_t = \Phi^{-1}(z_t^*), z_t^* = 1.133 \max\{0, x_t - q_{0.5}\} + 1.031 \max\{0, q_{0.5} - x_t\}.$$

Figure 6 plots z_t^* versus x_t (panel (ii)), and z_t versus x_t (panel (iii)). Both relations are slightly asymmetric. The sample ACF of z_t^*, plotted in panel (iv) of figure 5, is very persistent.

The second most predictable aspect w_t, orthogonal to the first is a robust level transformation.

$$w_t = \Phi^{-1}(w_t^*), w_t^* = 0.879 \max\{0, x_t - q_{0.5}\} - 0.746 \max\{0, q_{0.5} - x_t\}.$$

Its mutual information index is equal to 0.004. It is characterized by a significant autocorrelation at lag 1, equal to -0.108. This is less than the value of the first sample autocorrelation of the original time series (-0.102). Part of this covariability is absorbed by z_t.

7 Testing (un)predictability

The most predictable aspect z_t can be used for testing the null of no predictability of the series. The idea is to apply a serial correlation test, such as the Box and Pierce (1970) and the Ljung and Box (1978), or an independence test (see Teräsvirta et al., 2010, sec. 7.7) to the series z_t.

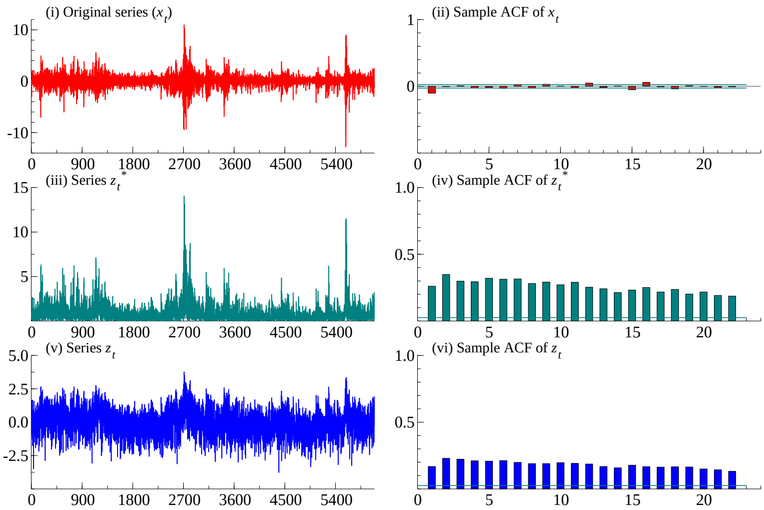

Figure 5: S&P 500 daily returns. (i) Plot of the original time series (x_t). (ii) Sample ACF plot of x_t. (iii) Time series plot of $z_t^* = 1.29 \max\{0, x_t - q_{0.5}\} + 0.85 \max\{0, q_{0.5} - x_t\}$. (iv) Sample ACF plot of z_t^*. (v) Time series plot of $z_t = \Phi^{-1}(z_t^*)$. (vi) Sample ACF plot of 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 κ (0.2, 0.3,0.4).

When appropriately standardized, the test statistic is asymptotically N(0,1). It has been shown by W. W. Chen and Deo (2004) that the (1996) tests suffer from size distortions in finite samples, which are resolved in W. W. Chen and Deo (2004) by taking a power transformation of the test statistic, aiming at reducing the skewness of the distribution. Their test statistic will be denoted $H_T^\delta(\kappa)$, where δ is a power parameter depending on the moments of the kernel.

The empirical size refers to the test conducted at the 5% nominal size for the following processes: i. $X_t \sim$ i.i.d. N$(0, 1)$; ii. $X_t = \exp(\epsilon_t), \epsilon_t \sim$ i.i.d. N$(0, 1)$; iii. $X_t \sim$ i.i.d.t_3 (Student's-t with 3 degrees of freedom); iv. $X_t \sim$ i.i.d. α-stable with characteristic exponent 1, skewness parameter 0, location 0 and scale 1; v. $X_t \sim$ i.i.d. α-stable with characteristic exponent 1.5, skewness parameter 0.8, location parameter 0 and scale parameter 1.

The results, reported in Appendix E (tables 1-5), show that the test tends to be slightly oversized in small sample; the size distortion is larger as r increases (overfitting generates more false discoveries), but tend to disappear as T increases. We also observe that the Chen and Deo modified test behaves better, and in particular with the choice of the bandwidth $\kappa = 0.2$.

The empirical powers are evaluated using the same experimental design, with reference to the following processes (where $\epsilon_t \sim$ i.i.d. N$(0, 1)$):

1. Non-Linear MA(2) process, $X_t = \epsilon_t\epsilon_{t-1}\epsilon_{t-2}$ (Fokianos and Pitsillou, 2017).

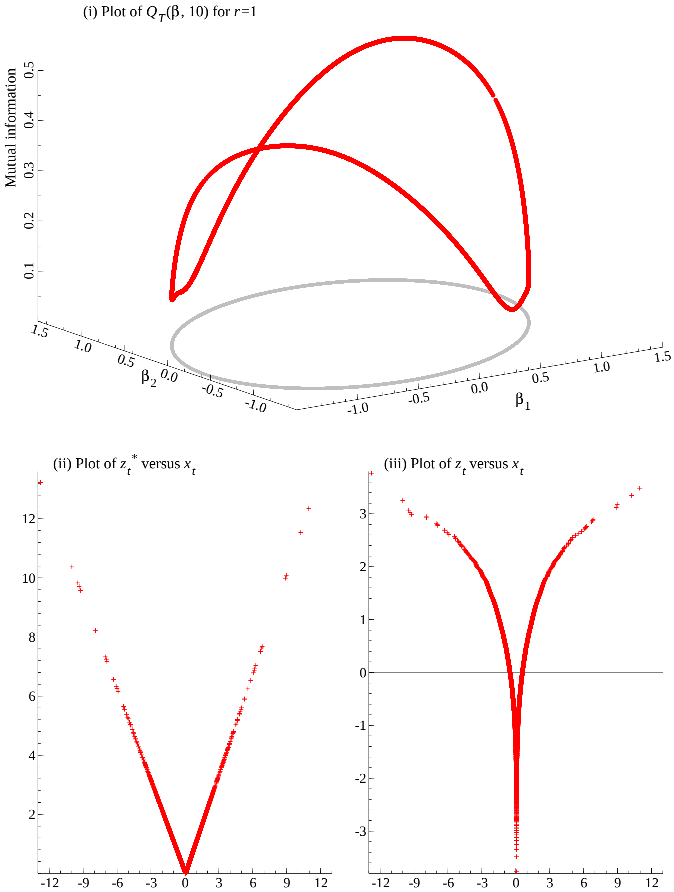

Figure 6: S&P 500 daily returns. (i) Plot of the mutual information as a function of $\boldsymbol{\beta}$, $Q_T(\boldsymbol{\beta}, l)$, evaluated at the points $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}_h}(0)\boldsymbol{\beta} = 1$ (in grey), for $h_{1t} = \max\{0, x_t - q_{0.5}\}, h_{2t} = \max\{0, q_{0.5} - x_t\}$ and $l = 10$. (ii) Plot of $z_t^* = 1.133 \max\{0, x_t - q_{0.5}\} + 1.031 \max\{0, q_{0.5} - x_t\}$ versus x_t. (iii) Plot of $z_t = \Phi^{-1}(z_t^*)$ versus x_t.

2. ARCH(1,1) process: $X_t = \sqrt{h_t}\epsilon_t, h_t = 0.5 + 0.8X_{t-1}^2 + 0.1X_{t-2}^2$, (Fokianos and Pitsillou, 2017).

3. GARCH(1,1) process: $X_t = \sqrt{h_t}\epsilon_t, h_t = 0.01 + 0.94X_{t-1}^2 + 0.05h_{t-1}$, (Granger et al., 2004, Model 11)

4. Threshold autoregressive process (Fokianos and Pitsillou, 2017),

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

5. Bilinear model $X_t = 0.6\epsilon_{t-1}X_{t-2} + \epsilon_t$ (Granger and Lin, 1994; Granger et al., 2004, Model 9).

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

For cases 1, 2 and 5 the power of the test is very high (tables 6 and 7), already with $r = 1$ and $T = 100$ ($T = 250$ in the bilinear case, see table 10). The lower power found in the GARCH(1,1) case, see table 8, can be explained as follows: our estimate of the MI for z_t relies on a $2\lceil \hat{l}_\beta \rceil$ autoregressive approximation; the data generating process is such that, e.g., the squares are ARMA(1,1) with MA parameter close to the noninvertibility region (-0.94). Hence, the goodness of the AR approximation can be poor. Finally, for the threshold autoregressive process (table 9), a large sample size, such as $T = 1000$, is needed to achieve a reasonable power. The most predictable transformation with $r = 3$ in the majority of cases (those leading to reject the null) takes into account the fact that the series is subject to large positive values when it switches from negative to positive and that the second regime (positive x_t) is more frequent.

8 Conclusions

The methods presented in this papers do not provide an easy way of estimating the mutual information between past and future of a time series for any stationary process. Our goal is less ambitious and consisted of estimating non-(or semi-)parametrically the nonlinear transformation that is most predictable from its past. If the second most predictable aspect is unpredictable, then we could think that we are close to the more ambitious target.

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 to predict aspects of the original time series.

A Proof of Theorem 1

The following result is known as the chain rule for mutual information, see Cover and Thomas (2006).

Theorem 2 (MI decomposition)**.** *Let $Y, X = (X_1, X_2, \dots, X_r)$ be continuous random variables with joint density $f(X, Y)$. The mutual information between Y and X is decomposed into the sum of the partial mutual information*

$$I(Y, X) = I(Y, X_r) + \sum_{i=1}^{r-1} I(Y, X_i | X_{i+1}, \dots, X_r).$$

In view of further developments, we provide an alternative proof.

Proof. The proof follows from the easily established factorization

$$f(Y, X) = f(Y, X_r) \prod_{i=1}^{r-1} \frac{f(Y, X_i | X_{i+1}, \dots, X_r)}{f(Y | X_{i+1}, \dots, X_r)},$$

so that,

$$\frac{f(Y, X)}{f(X) f(Y)} = \frac{f(Y, X_r)}{f(Y) f(X_r)} \prod_{i=2}^{r} \frac{f(Y, X_i | X_{i+1}, \dots, X_r)}{f(Y | X_{i+1}, \dots, X_r) f(X_i | X_{i+1}, \dots, X_r)}.$$

Then,

$$\begin{aligned} I(Y, X) &= \int \int f(Y, X) \log \tfrac{f(Y,X)}{f(X)f(Y)} dY dX \\ &= \textstyle\sum_{i=1}^{r} \int \cdots \int f(Y, X_i, X_{i+1}, \dots, X_r) \log \tfrac{f(Y, X_i | X_{i+1}, \dots, X_r)}{f(Y | X_{i+1}, \dots, X_r) f(X_i | X_{i+1}, \dots, X_r)} dY dX_i dX_{i+1} \cdots dX_r, \end{aligned}$$

gives the above decomposition.

□

Corollary 2. *Given the continuous random variable Z, the conditional mutual information $I(Y, X|Z)$ has the following decomposition*

$$I(Y, X|Z) = I(Y, X_r|Z) + \sum_{i=1}^{r-1} I(Y, X_i | X_{i+1}, \dots, X_r, Z).$$

We are now ready to prove Theorem 1. For $m = 1$, apply Theorem 2 with $Y = X_{n+1}$ and $r = n$, to show that $I(X_{1:n}, X_{n+1}) = \sum_{k=1}^{n} \pi(k)$. For $m > 1$, the following recursion holds:

$$\begin{aligned} \tfrac{f(X_{1:n}, X_{n+1:n+m})}{f(X_{1:n}) f(X_{n+1:n+m})} &= \tfrac{f(X_{1:n}, X_{n+1:n+m-1})}{f(X_{1:n}) f(X_{n+1:n+m-1})} \tfrac{f(X_{n+m} | X_{1:n+m-1})}{f(X_{n+m} | X_{n+1:n+m-1})}, \\ &= \tfrac{f(X_{1:n}, X_{n+1:n+m-1})}{f(X_{1:n}) f(X_{n+1:n+m-1})} \tfrac{f(X_{n+m}, X_{1:n} | X_{n+1:n+m-1})}{f(X_{n+m} | X_{n+1:n+m-1}) f(X_{1:n} | X_{n+1:n+m-1})}, \end{aligned}$$

so that, taking logarithms and the expectation with respect to the joint density of $(X_{1:n}, X_{n+1:n+m})$, Theorem 2, applied with $Y = X_{n+m}$, $X = X_{1:n}$ and $Z = X_{n+1:n+m-1}$, yields

$$\begin{aligned} I(X_{1:n}, X_{n+1:n+m}) &= I(X_{1:n}, X_{n+1:n+m-1}) + \textstyle\sum_{i=1}^{n} I(X_{n+m}, X_i | X_{i+1}, \dots, X_{n+m-1}), \\ &= I(X_{1:n}, X_{n+1:n+m-1}) + \textstyle\sum_{i=1}^{n} \pi(n + m - i) \\ &= I(X_{1:n}, X_{n+1:n+m-2}) + \textstyle\sum_{i=1}^{n} (\pi(n + m - i) \pi(n + m - i - 1) \\ &= \textstyle\sum_{j=1}^{m} \sum_{i=1}^{n} \pi(n + j - i). \end{aligned}$$

B Durbin Levinson algorithm

Let $v_0 = \gamma(0)$, $\phi_{11} = \gamma(1)/\gamma(0)$, $v_1 = (1-\phi_{11}^2)v_0$; then, for $k = 2, \dots, n-1$, the Durbin-Levinson algorithm is the following set of recursions (Levinson, 1946; Durbin, 1960):

$$\begin{array}{rcl} \phi_{kk} & = & \frac{\gamma(k)-\sum_{j=1}^{k-1}\phi_{k-1,j}\gamma(k-j)}{v_{k-1}}, \\ \phi_{kj} & = & \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j}, \quad (j = 1, \dots, k-1), \\ v_k & = & (1-\phi_{kk}^2)v_{k-1}. \end{array} \tag{8}$$

The DL algorithm performs the factorization of the autocovariance matrix of the random variables $\{X_t, t = 1, \dots, n\}$,

$$\mathbf{\Gamma}_n^{-1} = \mathbf{\Phi}_n' \mathbf{D}_n \mathbf{\Phi}_n,$$

where $\mathbf{D}_n = \text{diag}\left(v_0^{-1}, v_1^{-1}, \dots, v_{n-1}^{-1}\right)$,

$$\mathbf{\Phi}_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\phi_{11} & 1 & 0 & \cdots & 0 \\ -\phi_{22} & -\phi_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ -\phi_{n-1,n-1} & -\phi_{n-1,n-2} & -\phi_{n-1,n-3} & \cdots & 1 \end{bmatrix}.$$

C Canonical correlation analysis

In the Gaussian case the mutual information can be alternatively obtained via canonical correlation analysis of the variables $X_{1:n} = (X_1, \dots, X_n)'$ and $X_{n+1:n+m} = (X_{n+1}, \dots, X_{n+m})'$. Consider the variance-covariance matrix of the two sets of variables:

$$\mathbf{\Gamma}_{n+m} = \begin{bmatrix} \mathbf{\Gamma}_n & \mathbf{C}_{nm} \\ \mathbf{C}_{mn} & \mathbf{\Gamma}_m \end{bmatrix},$$

with $\mathbf{C}_{nm} = \text{Cov}(X_{1:n}, X_{n+1:n+m})$ and $\mathbf{C}_{mn} = \mathbf{C}_{nm}'$.

Let ϱ_i denote the $i-th$ canonical correlation between the future $X_{n+1:n+m}$ and the past $X_{1:n}$ random variables, $i = 1, \dots, q = \min(n, m)$, $\varrho_1 \geq \varrho_2 \geq \dots, \varrho_q \geq 0$.

The mutual information between $X_{1:n}$ and $X_{n+1:n+m}$ is equal to

$$I(X_{1:n}, X_{n+1:n+m}) = -\frac{1}{2}\sum_{i=1}^{q} \log\left(1-\varrho_i^2\right), \quad q = \min(n, m). \tag{9}$$

The canonical correlations are the singular values of the matrix $\mathbf{A}_{nm} = \mathbf{\Gamma}_n^{-1/2}\mathbf{C}_{nm}\mathbf{\Gamma}_m^{-1/2'}$. Equivalently, the squared canonical correlations, ϱ_i^2, are the eigenvalues of

$$\mathbf{A}_{nm}'\mathbf{A}_{nm} = \mathbf{\Gamma}_m^{-1/2}\mathbf{C}_{mn}\mathbf{\Gamma}_n^{-1}\mathbf{C}_{nm}\mathbf{\Gamma}_m^{-1/2'}.$$

In terms of the DL factorization, $\mathbf{A}_{nm} = \mathbf{D}_n^{1/2}\mathbf{\Phi}_n\mathbf{C}_{nm}\mathbf{\Phi}_m'\mathbf{D}_m^{1/2}$.

D A Nonlinear Canonical Correlation Problem

Conditional on the choice of r and the basis function, the problem of determining the most predictable aspect of X_t can be traced back to a nonlinear constrained correlation analysis of the multivariate process $\mathbf{h}_t$. We discuss this referring first to the finite past and finite future.

Stacking n past and m future vectors $\mathbf{h}_t$ into $\mathbf{h}_p = (\mathbf{h}'_1, \ldots, \mathbf{h}'_n)'$ and $\mathbf{h}_f = (\mathbf{h}'_{n+1}, \ldots, \mathbf{h}'_{n+m})'$, respectively, we let $\boldsymbol{\mu}_{hp} = \mathrm{E}(\mathbf{h}_p)$ and $\mathrm{E}\{(\mathbf{h}_p - \boldsymbol{\mu}_{hp})(\mathbf{h}_p - \boldsymbol{\mu}_{hp})'\} = \boldsymbol{\Gamma}_{pp}^{(h)}$, $\boldsymbol{\mu}_{hf} = \mathrm{E}(\mathbf{h}_f)$ and $\mathrm{E}\{(\mathbf{h}_f - \boldsymbol{\mu}_{hf})(\mathbf{h}_f - \boldsymbol{\mu}_{hf})'\} = \boldsymbol{\Gamma}_{ff}^{(h)}$, and $\mathrm{E}\{(\mathbf{h}_p - \boldsymbol{\mu}_{hp})(\mathbf{h}_f - \boldsymbol{\mu}_{hf})'\} = \boldsymbol{\Gamma}_{pf}^{(h)}$. Similarly, stacking the past and future values of Z_t in the vectors $\mathbf{Z}_p = (Z_1, \ldots, Z_n)'$ and $\mathbf{Z}_f = (Z_{n+1}, \ldots, Z_{n+m})'$, so that $\mathbf{Z}_p = (\mathbf{I}_n \otimes \boldsymbol{\beta}')\mathbf{h}_p$ and $\mathbf{Z}_f = (\mathbf{I}_m \otimes \boldsymbol{\beta}')\mathbf{h}_f$, we let $\boldsymbol{\mu}_{zp} = \mathrm{E}(\mathbf{Z}_p)$ and $\mathrm{E}\{(\mathbf{Z}_p - \boldsymbol{\mu}_{zp})(\mathbf{Z}_p - \boldsymbol{\mu}_{zp})'\} = \boldsymbol{\Gamma}_{pp}^{(z)}$, $\boldsymbol{\mu}_{zf} = \mathrm{E}(\mathbf{Z}_f)$ and $\mathrm{E}\{(\mathbf{Z}_f - \boldsymbol{\mu}_{zf})(\mathbf{Z}_f - \boldsymbol{\mu}_{zf})'\} = \boldsymbol{\Gamma}_{ff}^{(z)}$, and $\mathrm{E}\{(\mathbf{Z}_p - \boldsymbol{\mu}_{zp})(\mathbf{Z}_f - \boldsymbol{\mu}_{zf})'\} = \boldsymbol{\Gamma}_{pf}^{(z)}$.

Consider the problem of choosing the $nr \times 1$ vector $\mathbf{v}_p = (\boldsymbol{\alpha}_p \otimes \boldsymbol{\beta})$ and the $mr \times 1$ vector $\mathbf{v}_f = (\boldsymbol{\alpha}_f \otimes \boldsymbol{\beta})$, where $\otimes$ is the Kronecker product, so that the mutual information between the n past and m future values of Z_t,

$$Q_T(\boldsymbol{\beta}) = -\frac{1}{2} \sum_{i=1}^{n \wedge m} \log(1 - \varrho_i^2), \tag{10}$$

is maximised, where

$$\begin{aligned} \varrho_i &= \mathbf{v}'_{ip} \boldsymbol{\Gamma}_{pf}^{(h)} \mathbf{v}_{if} \\ &= \boldsymbol{\alpha}'_{ip} \boldsymbol{\Gamma}_{pf}^{(z)} \boldsymbol{\alpha}_{if}, \end{aligned} \tag{11}$$

are the canonical correlations between $\mathbf{Z}_p$ and $\mathbf{Z}_f$, ordered so that $\varrho_1 > \varrho_2 > \cdots > \varrho_{n \wedge m}$. is a maximum, under the constraints:

$$\boldsymbol{\alpha}'_p \boldsymbol{\Gamma}_{pp}^{(z)} \boldsymbol{\alpha}_p = 1, \quad \boldsymbol{\alpha}'_f \boldsymbol{\Gamma}_{ff}^{(z)} \boldsymbol{\alpha}_f = 1, \quad \boldsymbol{\beta}' \boldsymbol{\Gamma}_h(0) \boldsymbol{\beta} = 1. \tag{12}$$

Notice that $\mathbf{v}'_p \boldsymbol{\Gamma}_{pp}^{(h)} \mathbf{v}_p = \boldsymbol{\alpha}'_p \boldsymbol{\Gamma}_{pp}^{(z)} \boldsymbol{\alpha}_p$ and $\mathbf{v}'_f \boldsymbol{\Gamma}_{ff}^{(h)} \mathbf{v}_f = \boldsymbol{\alpha}'_f \boldsymbol{\Gamma}_{ff}^{(z)} \boldsymbol{\alpha}_f$. The solution of the optimization problem delivers the weights for the linear combination of the basis functions. Hence, the most predictable aspect of X_t can be estimated by the canonical analysis of the past and the future of the feature vector $\mathbf{h}_t$ subject to a nonlinear constraint on the canonical vectors.

E Size and Power of Tests of no Predictability

The following tables are obtained from a simulation experiment with $M = 1000$ replications, dealing with the frequency of rejection of the null of no serial correlation in the most predictable aspect of the simulated time series, z_t, estimated from x_t using r hinge-basis functions located at the $j/(r+1), j = 1, \ldots, r$, quantiles of the empirical distribution of x_t.

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

r=1

	$T=100$	$T=250$	$T=500$	$T=1000$	$T=5000$
$\mathcal{H}_T(0.2)$	0.077	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

r=3

	$T=100$	$T=250$	$T=500$	$T=1000$	$T=5000$
$\mathcal{H}_T(0.2)$	0.124	0.110	0.100	0.106	0.110
$\mathcal{H}_T(0.3)$	0.113	0.096	0.093	0.086	0.080
$\mathcal{H}_T(0.4)$	0.156	0.110	0.097	0.081	0.067
$\mathcal{H}_T^\delta(0.2)$	0.079	0.073	0.062	0.062	0.059
$\mathcal{H}_T^\delta(0.4)$	0.095	0.078	0.060	0.054	0.064
$\mathcal{H}_T^\delta(0.6)$	0.146	0.102	0.080	0.066	0.067

r=5

	$T=100$	$T=250$	$T=500$	$T=1000$	$T=5000$
$\mathcal{H}_T(0.2)$	0.163	0.139	0.136	0.140	0.146
$\mathcal{H}_T(0.3)$	0.149	0.114	0.102	0.094	0.093
$\mathcal{H}_T(0.4)$	0.177	0.122	0.103	0.082	0.080
$\mathcal{H}_T^\delta(0.2)$	0.108	0.078	0.081	0.077	0.084
$\mathcal{H}_T^\delta(0.4)$	0.120	0.089	0.070	0.067	0.068
$\mathcal{H}_T^\delta(0.6)$	0.159	0.113	0.090	0.067	0.066

Table 2: Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and Chen and Deo $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is $X_t = \exp(\epsilon_t), \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), t = 1,\ldots,T$, 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

r=3

	$T=100$	$T=250$	$T=500$	$T=1000$	$T=5000$
$\mathcal{H}_T(0.2)$	0.134	0.112	0.127	0.112	0.132
$\mathcal{H}_T(0.3)$	0.120	0.091	0.092	0.091	0.094
$\mathcal{H}_T(0.4)$	0.141	0.094	0.102	0.084	0.062
$\mathcal{H}_T^\delta(0.2)$	0.089	0.083	0.083	0.072	0.089
$\mathcal{H}_T^\delta(0.4)$	0.099	0.078	0.067	0.069	0.062
$\mathcal{H}_T^\delta(0.6)$	0.126	0.087	0.092	0.082	0.056

r=5

	$T=100$	$T=250$	$T=500$	$T=1000$	$T=5000$
$\mathcal{H}_T(0.2)$	0.179	0.162	0.151	0.138	0.131
$\mathcal{H}_T(0.3)$	0.139	0.123	0.117	0.111	0.096
$\mathcal{H}_T(0.4)$	0.153	0.126	0.098	0.088	0.068
$\mathcal{H}_T^\delta(0.2)$	0.106	0.103	0.093	0.078	0.080
$\mathcal{H}_T^\delta(0.4)$	0.108	0.094	0.087	0.086	0.054
$\mathcal{H}_T^\delta(0.6)$	0.131	0.116	0.084	0.072	0.056

Table 3: Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and Chen and Deo $\mathcal{H}_T^\delta(\kappa)$ tests of no predictability when the true model is $X_t \sim$ i.i.d. $t_3, t = 1, \ldots, T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.071	0.086	0.083	0.071	0.061
$\mathcal{H}_T(0.3)$	0.085	0.091	0.060	0.058	0.054
$\mathcal{H}_T(0.4)$	0.124	0.094	0.069	0.067	0.055
$\mathcal{H}_T^\delta(0.2)$	0.050	0.070	0.058	0.042	0.046
$\mathcal{H}_T^\delta(0.4)$	0.072	0.078	0.057	0.054	0.047
$\mathcal{H}_T^\delta(0.6)$	0.118	0.094	0.061	0.062	0.051

r=2

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.090	0.090	0.102	0.084	0.078
$\mathcal{H}_T(0.3)$	0.105	0.079	0.077	0.071	0.058
$\mathcal{H}_T(0.4)$	0.121	0.096	0.086	0.075	0.059
$\mathcal{H}_T^\delta(0.2)$	0.064	0.061	0.076	0.052	0.048
$\mathcal{H}_T^\delta(0.4)$	0.090	0.063	0.070	0.057	0.050
$\mathcal{H}_T^\delta(0.6)$	0.112	0.091	0.076	0.070	0.053

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.119	0.120	0.126	0.104	0.092
$\mathcal{H}_T(0.3)$	0.119	0.100	0.099	0.073	0.064
$\mathcal{H}_T(0.4)$	0.133	0.109	0.102	0.070	0.068
$\mathcal{H}_T^\delta(0.2)$	0.079	0.068	0.082	0.054	0.048
$\mathcal{H}_T^\delta(0.4)$	0.098	0.072	0.084	0.052	0.040
$\mathcal{H}_T^\delta(0.6)$	0.132	0.086	0.081	0.064	0.057

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.168	0.173	0.152	0.118	0.112
$\mathcal{H}_T(0.3)$	0.159	0.133	0.118	0.080	0.079
$\mathcal{H}_T(0.4)$	0.153	0.126	0.104	0.077	0.061
$\mathcal{H}_T^\delta(0.2)$	0.115	0.111	0.098	0.064	0.066
$\mathcal{H}_T^\delta(0.4)$	0.119	0.105	0.090	0.066	0.053
$\mathcal{H}_T^\delta(0.6)$	0.138	0.114	0.085	0.066	0.049

Table 4: Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and Chen and Deo $\mathcal{H}_T^{\delta}(\kappa)$ tests of no predictability when the true model is $X_t \sim$ i.i.d. α-stable with characteristic exponent 1, skewness parameter 0, location 0 and scale 1, $t = 1, \ldots, T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.075	0.060	0.066	0.057	0.082
$\mathcal{H}_T(0.3)$	0.085	0.054	0.051	0.052	0.062
$\mathcal{H}_T(0.4)$	0.102	0.063	0.065	0.050	0.067
$\mathcal{H}_T^{\delta}(0.2)$	0.058	0.046	0.041	0.034	0.054
$\mathcal{H}_T^{\delta}(0.4)$	0.083	0.046	0.064	0.050	0.056
$\mathcal{H}_T^{\delta}(0.6)$	0.111	0.056	0.066	0.044	0.062

r=2

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.102	0.080	0.077	0.072	0.094
$\mathcal{H}_T(0.3)$	0.110	0.071	0.066	0.063	0.065
$\mathcal{H}_T(0.4)$	0.128	0.089	0.065	0.058	0.059
$\mathcal{H}_T^{\delta}(0.2)$	0.060	0.051	0.045	0.040	0.062
$\mathcal{H}_T^{\delta}(0.4)$	0.100	0.058	0.060	0.064	0.043
$\mathcal{H}_T^{\delta}(0.6)$	0.123	0.080	0.063	0.049	0.052

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.125	0.113	0.092	0.093	0.102
$\mathcal{H}_T(0.3)$	0.117	0.091	0.075	0.068	0.070
$\mathcal{H}_T(0.4)$	0.155	0.098	0.070	0.053	0.065
$\mathcal{H}_T^{\delta}(0.2)$	0.081	0.062	0.048	0.059	0.065
$\mathcal{H}_T^{\delta}(0.4)$	0.107	0.076	0.049	0.059	0.059
$\mathcal{H}_T^{\delta}(0.6)$	0.139	0.100	0.058	0.057	0.056

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.178	0.133	0.139	0.124	0.128
$\mathcal{H}_T(0.3)$	0.156	0.129	0.087	0.077	0.095
$\mathcal{H}_T(0.4)$	0.176	0.126	0.083	0.078	0.076
$\mathcal{H}_T^{\delta}(0.2)$	0.119	0.086	0.085	0.068	0.078
$\mathcal{H}_T^{\delta}(0.4)$	0.120	0.102	0.068	0.069	0.064
$\mathcal{H}_T^{\delta}(0.6)$	0.152	0.123	0.078	0.062	0.057

Table 5: Rejection frequency (empirical size) of Hong's $\mathcal{H}_T(\kappa)$ and Chen and Deo $\mathcal{H}_T^{\delta}(\kappa)$ tests of no predictability when the true model is $X_t \sim$ i.i.d. α-stable with characteristic exponent 1.5, skewness parameter 0.8, location parameter 0 and scale parameter 1, $t = 1, \ldots, T$, 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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.117	0.099	0.105	0.095	0.099
$\mathcal{H}_T(0.3)$	0.123	0.081	0.079	0.074	0.081
$\mathcal{H}_T(0.4)$	0.128	0.091	0.073	0.081	0.069
$\mathcal{H}_T^{\delta}(0.2)$	0.074	0.070	0.070	0.054	0.063
$\mathcal{H}_T^{\delta}(0.4)$	0.092	0.069	0.064	0.062	0.062
$\mathcal{H}_T^{\delta}(0.6)$	0.117	0.082	0.069	0.075	0.063

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.165	0.145	0.152	0.127	0.130
$\mathcal{H}_T(0.3)$	0.146	0.119	0.106	0.074	0.082
$\mathcal{H}_T(0.4)$	0.170	0.113	0.090	0.075	0.059
$\mathcal{H}_T^{\delta}(0.2)$	0.111	0.090	0.096	0.065	0.079
$\mathcal{H}_T^{\delta}(0.4)$	0.113	0.088	0.077	0.053	0.063
$\mathcal{H}_T^{\delta}(0.6)$	0.137	0.092	0.079	0.061	0.053

Table 6: 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 $X_t = \epsilon_t \epsilon_{t-1} \epsilon_{t-2}, \quad \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), t = 1, \ldots, T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.963	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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.892	0.996	1.000	1.000	1.000
$\mathcal{H}_T(0.3)$	0.802	0.994	1.000	1.000	1.000
$\mathcal{H}_T(0.4)$	0.726	0.983	1.000	1.000	1.000
$\mathcal{H}_T^{\delta}(0.2)$	0.842	0.993	1.000	1.000	1.000
$\mathcal{H}_T^{\delta}(0.4)$	0.767	0.992	1.000	1.000	1.000
$\mathcal{H}_T^{\delta}(0.6)$	0.703	0.980	1.000	1.000	1.000

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.874	0.990	1.000	1.000	1.000
$\mathcal{H}_T(0.3)$	0.795	0.983	1.000	1.000	1.000
$\mathcal{H}_T(0.4)$	0.708	0.970	0.999	1.000	1.000
$\mathcal{H}_T^{\delta}(0.2)$	0.817	0.985	0.999	1.000	1.000
$\mathcal{H}_T^{\delta}(0.4)$	0.751	0.979	0.999	1.000	1.000
$\mathcal{H}_T^{\delta}(0.6)$	0.683	0.962	0.999	1.000	1.000

Table 7: 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 $X_t = \sqrt{h_t}\epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), h_t = 0.5 + 0.8X_{t-1}^2 + 0.1X_{t-2}^2, t = 1,\ldots,T,$ for $T = 100, 250, 500, 1000, 5000,$ and $\kappa = 0.2, 0.4, 0.6$.

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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.741	0.961	0.992	0.999	1.000
$\mathcal{H}_T(0.3)$	0.662	0.931	0.984	1.000	1.000
$\mathcal{H}_T(0.4)$	0.612	0.896	0.976	0.997	1.000
$\mathcal{H}_T^{\delta}(0.2)$	0.682	0.946	0.988	0.999	1.000
$\mathcal{H}_T^{\delta}(0.4)$	0.613	0.918	0.981	0.999	1.000
$\mathcal{H}_T^{\delta}(0.6)$	0.584	0.887	0.975	0.996	1.000

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.749	0.966	0.997	0.999	1.000
$\mathcal{H}_T(0.3)$	0.637	0.941	0.994	1.000	1.000
$\mathcal{H}_T(0.4)$	0.587	0.892	0.983	0.998	1.000
$\mathcal{H}_T^{\delta}(0.2)$	0.677	0.956	0.993	0.999	1.000
$\mathcal{H}_T^{\delta}(0.4)$	0.591	0.922	0.987	0.999	1.000
$\mathcal{H}_T^{\delta}(0.6)$	0.547	0.881	0.979	0.998	1.000

Table 8: 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. $N(0,1), h_t = 0.1 + 0.05\alpha X_{t-1}^2 + 0.94h_{t-1}, t = 1, \ldots, T,$ for $T = 100, 250, 500, 1000, 5000,$ and $\kappa = 0.2, 0.4, 0.6$.

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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.411	0.570	0.740	0.923	1.000
$\mathcal{H}_T(0.3)$	0.316	0.472	0.687	0.904	1.000
$\mathcal{H}_T(0.4)$	0.270	0.398	0.649	0.880	1.000
$\mathcal{H}_T^\delta(0.2)$	0.317	0.468	0.677	0.896	1.000
$\mathcal{H}_T^\delta(0.4)$	0.258	0.433	0.645	0.888	1.000
$\mathcal{H}_T^\delta(0.6)$	0.253	0.363	0.620	0.875	1.000

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.733	0.793	0.865	0.953	1.000
$\mathcal{H}_T(0.3)$	0.547	0.618	0.757	0.920	1.000
$\mathcal{H}_T(0.4)$	0.430	0.519	0.678	0.894	1.000
$\mathcal{H}_T^\delta(0.2)$	0.619	0.691	0.797	0.932	1.000
$\mathcal{H}_T^\delta(0.4)$	0.457	0.559	0.716	0.907	1.000
$\mathcal{H}_T^\delta(0.6)$	0.377	0.473	0.646	0.885	1.000

Table 9: 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 $X_t = (-1.5X_{t-1} + \epsilon_t)I(X_{t-1} < 0) + (0.5X_{t-1} + \epsilon_t)I(X_{t-1} \geq 0)$, $t = 1, \ldots, T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.099	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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.170	0.292	0.506	0.830	1.000
$\mathcal{H}_T(0.3)$	0.134	0.195	0.353	0.650	1.000
$\mathcal{H}_T(0.4)$	0.142	0.154	0.243	0.432	0.997
$\mathcal{H}_T^\delta(0.2)$	0.106	0.204	0.402	0.760	1.000
$\mathcal{H}_T^\delta(0.4)$	0.111	0.145	0.273	0.564	1.000
$\mathcal{H}_T^\delta(0.6)$	0.137	0.141	0.195	0.382	0.997

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.203	0.322	0.539	0.847	1.000
$\mathcal{H}_T(0.3)$	0.177	0.227	0.360	0.654	1.000
$\mathcal{H}_T(0.4)$	0.188	0.184	0.241	0.435	0.998
$\mathcal{H}_T^\delta(0.2)$	0.129	0.233	0.437	0.769	1.000
$\mathcal{H}_T^\delta(0.4)$	0.143	0.168	0.282	0.571	1.000
$\mathcal{H}_T^\delta(0.6)$	0.161	0.164	0.210	0.383	0.998

Table 10: 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 $X_t = 0.6\epsilon_{t-1}X_{t-2} + \epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0,1), t = 1,\ldots,T$, for $T = 100, 250, 500, 1000, 5000$, and $\kappa = 0.2, 0.4, 0.6$.

r=1

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.374	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

r=3

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.382	0.762	0.963	0.997	1.000
$\mathcal{H}_T(0.3)$	0.318	0.653	0.923	0.997	1.000
$\mathcal{H}_T(0.4)$	0.283	0.516	0.862	0.995	1.000
$\mathcal{H}_T^\delta(0.2)$	0.317	0.701	0.951	0.997	1.000
$\mathcal{H}_T^\delta(0.4)$	0.265	0.592	0.909	0.997	1.000
$\mathcal{H}_T^\delta(0.6)$	0.252	0.485	0.842	0.994	1.000

r=5

	$T = 100$	$T = 250$	$T = 500$	$T = 1000$	$T = 5000$
$\mathcal{H}_T(0.2)$	0.422	0.749	0.956	0.997	1.000
$\mathcal{H}_T(0.3)$	0.339	0.653	0.928	0.997	1.000
$\mathcal{H}_T(0.4)$	0.284	0.524	0.873	0.996	1.000
$\mathcal{H}_T^\delta(0.2)$	0.332	0.688	0.948	0.996	1.000
$\mathcal{H}_T^\delta(0.4)$	0.281	0.595	0.907	0.996	1.000
$\mathcal{H}_T^\delta(0.6)$	0.266	0.494	0.851	0.993	1.000

References

Berrett, T. B., and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika*, *106*(3), 547–566.

Box, G. E., and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, *65*(332), 1509–1526.

Breiman, L., and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, *80*(391), 580–598.

Carcea, M., and Serfling, R. (2015). A gini autocovariance function for time series modelling. *Journal of Time Series Analysis*, *36*(6), 817–838.

Chen, W. W., and Deo, R. S. (2004). Power transformations to induce normality and their applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(1), 117–130.

Chen, X., and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 289–314.

Cover, T. M., and Thomas, J. A. (2006). *Elements of information theory, 2nd ed.* John Wiley & Sons.

Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique*, 233–244.

Edelmann, D., Fokianos, K., and Pitsillou, M. (2019). An updated literature review of distance correlation and its applications to time series. *International Statistical Review*, *87*(2), 237–262.

Escanciano, J. C., and Velasco, C. (2006). Generalized spectral tests for the martingale difference hypothesis. *Journal of Econometrics*, *134*(1), 151–185.

Fokianos, K., and Pitsillou, M. (2017). Consistent testing for pairwise dependence in time series. *Technometrics*, *59*(2), 262–270.

Fokianos, K., and Pitsillou, M. (2018). Testing independence for multivariate time series via the auto-distance correlation matrix. *Biometrika*, *105*(2), 337–352.

Geenens, G., and Lafaye de Micheaux, P. (2020). The hellinger correlation. *Journal of the American Statistical Association*, 1–15.

Gourieroux, C., and Jasiak, J. (2002). Nonlinear autocorrelograms: an application to inter-trade durations. *Journal of Time Series Analysis*, *23*(2), 127–154.

Granger, C., and Lin, J.-L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *Journal of time series analysis*, *15*(4), 371–384.

Granger, C., Maasoumi, E., and Racine, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, *25*(5), 649–669.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in Statistics. New York.

Hong, Y. (1999). Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association*, *94*(448), 1201–1220.

Hong, Y., and White, H. (2005). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica*, *73*(3), 837–901.

Ibragimov, I. A., and Rozanov, Y. A. (2012). *Gaussian random processes* (Vol. 9). Springer Science & Business Media.

Jewell, N. P., and Bloomfield, P. (1983). Canonical correlations of past and future for time series: definitions and theory. *The Annals of Statistics*, *11*(3), 837–847.

Jewell, N. P., Bloomfield, P., and Bartmann, F. C. (1983). Canonical correlations of past and future for time series: bounds and computation. *The Annals of Statistics*, *11*(3), 848–855.

Kinney, J. B., and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, *111*(9), 3354–3359.

Levinson, N. (1946). The wiener (root mean square) error criterion in filter design and prediction. *Studies in Applied Mathematics*, *25*(1-4), 261–278.

Li, L., and Xie, Z. (1996). Model selection and order determination for time series by information between the past and the future. *Journal of time series analysis*, *17*(1), 65–84.

Linton, O., and Whang, Y.-J. (2007). The quantilogram: With an application to evaluating directional predictability. *Journal of Econometrics*, *141*(1), 250–282.

Ljung, G. M., and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

McMurry, T. L., and Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, *31*(6), 471–482.

Otneim, H., and Tjøstheim, D. (2021). The locally gaussian partial correlation. *Journal of Business & Economic Statistics*, 1–13.

Owen, A. (1983). Optimal transformations for autoregressive time series models. *Technical Report 20, Project ORION, Stanford University, Dept. of Statistics*.

Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of Nonparametric Statistics*, *15*(4-5), 517–533.

Pourahmadi, M. (2001). *Foundations of time series analysis and prediction theory* (Vol. 379). John Wiley & Sons.

Pourahmadi, M., and Miamee, A. (1992). Computation of canonical correlation between past and future of a time series. *Journal of Time Series Analysis*, *13*(4), 345–351.

Proietti, T., and Giovannelli, A. (2018). A durbin–levinson regularized estimator of high-dimensional autocovariance matrices. *Biometrika*, *105*(4), 783–795.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, *10*(3-4), 441–451.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518–1524.

Shao, X., and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, *109*(507), 1302–1318.

Skaug, H. J., and Tjøstheim, D. (1993). Nonparametric tests of serial independence. In T. Subba Rao (Ed.), *Developments in time series analysis* (pp. 207–29). Chapman and Hall London.

Székely, G. J., and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, *3*(4), 1236–1265.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, *35*(6), 2769–2794.

Teräsvirta, T., Tjostheim, D., and Granger, C. (2010). *Modelling nonlinear economic time series*. Oxford University Press.

Tjøstheim, D., Otneim, H., and Støve, B. (2018). Statistical dependence: Beyond Pearson's ρ. *arXiv preprint arXiv:1809.10455*.

Zhou, Z. (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. *Journal of Time Series Analysis*, *33*(3), 438–457.