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

Tommaso Proietti*
Department of Economics and Finance, University of Rome Tor Vergata

March 13, 2023

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 to distill 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 the 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, which maximizes the mutual information between past and future. The proposed transformation arises from 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 using the most predictable aspect for testing the null of no predictability.

Keywords: Mutual Information; Partial Dependence Measures; Testing unpredictability.

*The author gratefully acknowledges financial support by the Italian Ministry of Education, University and Research, Progetti di Ricerca di Interesse Nazionale, research project 2020-2023, project 2020N9YFFE.

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 a 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, but 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 $\{X_t, t \in \mathbb{Z}\}$, 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) proposed a test of serial independence based on the auto-distance covariance function. 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 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 mea-

sure of directional predictability named the quantilogram. Otneim and Tjøstheim (2021) have recently proposed the locally Gaussian partial correlation, a measure of conditional dependence.

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', thereby raising the issue as to why independence is rejected. The main motivation for this paper is to provide an answer to both questions, by establishing 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 analysis, see Jewell and Bloomfield (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, and Kinney and Atwal (2014).

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).

Against this background, 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).

Our contribution to the literature referenced above is the following. 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. Secondly, we provide a general result which decomposes the mutual information between past and future as the sum of all partial mutual information between the pair of random variables in the future and the past. The most predictable aspect is then 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 5. Section 6 illustrates our methodology. Finally, section 7 uses the most predictable aspect for testing (lack of) predictability. 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. i. Nonnegativity: $I(X,Y) \geq 0$. ii. $I(X,Y) = 0$ if and only if X and Y are independent. iii. Symmetry: $I(Y,X) = I(X,Y)$. iv. $I(X,Y)$ is invariant to one-to-one transformations of Y and X, see Granger and Lin (1994, Theorem 3). v. $I(X,Y)$ 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 $\mathscr{I}(X,Y) = 1 - \exp(-2I(X,Y))$. It provides a measure of association satisfying the properties of an ideal measure of dependence established by Rényi (1959), with the following properties: i. $0 \leq \mathscr{I}(X,Y) \leq 1$, ii. $\mathscr{I}(X,Y) = 0$ if X and Y are independent, iii. if $u(X) = v(Y)$ for u and v measurable functions, $\mathscr{I}(X,Y) = 1$.

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 $\{X_t, t = 1, \ldots\}$ be a strictly stationary zero mean process, with continuous density $f(X_t)$ and characterised by the autocovariance function $\gamma(k) = \mathrm{E}(X_t X_{t-k}) < \infty, k = 0, \pm 1, \pm 2, \ldots$.

We denote by $\mathbf{\Gamma}_k = \{\gamma(|i-j|), i, j = 1, \ldots, k\}$ the autocovariance matrix of $X_{t-k+1:t} = (X_{t-k+1}, X_{t-k+2}, \ldots, X_{t-1}, X_t)$ 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 $X_{t-k:t-1} = (X_{t-k}, \ldots, X_{t-1})$,

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

has coefficients $\boldsymbol{\phi}_k = (\phi_{1k}, \ldots, \phi_{kk})'$ equal to $\boldsymbol{\phi}_k = \mathbf{\Gamma}_k^{-1} \boldsymbol{\gamma}_k$, where $\boldsymbol{\gamma}_k = (\gamma(1), \gamma(2), \ldots, \gamma(k))'$, 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 $\phi_{kk} = \frac{\mathrm{Cov}(X_t - \hat{X}_{k-1,t}, X_{t-k} - \hat{X}^*_{k-1,t-k})}{\sqrt{\mathrm{Var}(X_t - \hat{X}_{k-1,t})\mathrm{Var}(X_t - \hat{X}^*_{k,t-k})}}$, $k = 1, 2, \ldots$, where $\hat{X}^*_{k-1,t-k}$ is the linear predictor of X_{t-k} based on $X_{t-k+1:t-1} = (X_{t-k+1}, X_{t-k+2}, \ldots, X_{t-1})$.

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

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

The partial autocorrelation sequence and the partial autogressive coefficients are computed by the Durbin-Levinson algorithm (see Appendix B).

3.1 The mutual information between past and future

We now turn our consideration to the mutual information between the past and the future (MIPF) 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:*

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

Proof. See Appendix A. □

Remark 1. *The partial mutual information $\pi(k)$ is the expected conditional log copula density of X_t and X_{t+k}, given the intermediate variables:*

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

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 MIPF, originally formulated

for Gaussian processes (Ibragimov and Rozanov, 2012; Jewell and Bloomfield, 1983),

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

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

An important related concept is that of 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, whose proof is direct since $\pi(k) = -\frac{1}{2}\log(1-\phi_{kk}^2)$.

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

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. $\mathrm{N}(0, \sigma_\eta^2)$, independently of ϵ_t. Then, $I(X_p, X_f) = -\frac{1}{2}\log(1-\phi^2)$ and $\mathscr{I}(X_p, X_f) = \phi^2$.

The MIPF provides a measure of predictability across all possible future forecast horizons.

4 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 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 $\mathbf{h}_t = (h_{1t}, \ldots, h_{rt})'$. The cross-covariance matrix of $\mathbf{h}_t$ at lag k is $\mathrm{Cov}(\mathbf{h}_t, \mathbf{h}_{t-k}) = \boldsymbol{\Gamma}_h(k), k \in \mathbb{Z}$.

For identifiability we will assume that the set of measurable transformations $h_j(X_t), j = 1, \ldots, r$, is non-singular, i.e., $\boldsymbol{\Gamma}_h(0)$ is positive definite, with distinct eigenvalues; this rules out affine transformations of X_t, e.g., $h_j(X_t) = a_j + b_j X_t$.

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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)'$, $Z_t^* = \boldsymbol{\beta}'\mathbf{h}_t$, satisfying the normalization constraint $\boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1$. Notice that Z_t^* has unit variance.

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 Φ is the standard normal cdf.

We are now ready to define the most and second most predictable aspects of the series.

Definition 1. *The most predictable aspect of X_t is the transformation $Z_t = g(\boldsymbol{\beta}'\mathbf{h}_t)$, 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 $W_t = g(\boldsymbol{\zeta}'\mathbf{h}_t)$, 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 2. *The most linearly-predictable aspect of* X_t *is the transformation* $Z_t = g(\boldsymbol{\beta}'\mathbf{h}_t)$*, with* $\boldsymbol{\beta}$ *satisfying the constraint* $\boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1$*, which maximises the linear mutual information measure* $I^*(\mathcal{Z}_p, \mathcal{Z}_f) = -\frac{1}{2}\sum_{k=1}^{\infty}\log(1-\phi_{Z,kk}^2)$*, where* $\phi_{Z,kk}$ *denotes the PACF of* $Z_t^* = \boldsymbol{\beta}'\mathbf{h}_t$*,* $\mathcal{Z}_p = \{Z_{n-j}^*, j \geq 0\}$ *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* $I^*(\mathcal{W}_p, \mathcal{W}_f) = -\frac{1}{2}\sum_{k=1}^{\infty}\log(1-\phi_{W,kk}^2)$*, where* $\phi_{W,kk}$ *is the PACF of* $W_t^* = \boldsymbol{\zeta}'\mathbf{h}_t$*.*

In the sequel, the most predictable aspect of X_t will refer to Definition 2. We refer to $I^*(\mathcal{Z}_p, \mathcal{Z}_f)$ as the *linear* MIPF.

Example 1. (continued) Let $\sigma^2 = 1-\phi^2$, so that $\text{Var}(X_t) = 1$, and choose two hinge functions located at the median, so that $h_{1t} = \max(0, X_t)$ and $h_{2t} = \max(0, -X_t)$. Then it holds that $\text{E}(h_{it}) = 1/\sqrt{2\pi}$, $\text{Var}(h_{it}) = (\pi-1)/(2\pi)$, $i = 1, 2$, and $\text{Cov}(h_{1t}, h_{2t}) = -1/(2\pi)$. The covariance matrix $\boldsymbol{\Gamma}_h(0) = \frac{1}{2}\left(\mathbf{I}_2 - \frac{1}{\pi}\mathbf{i}\mathbf{i}'\right)$, has eigenvalues $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{2} - \frac{1}{\pi}$, with corresponding eigenvectors $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1, -1)'$ and $\mathbf{v}_2 = \frac{1}{\sqrt{2}}\mathbf{i}$. Rescaling the two vectors by $\lambda_i^{-1/2}$, we obtain the two linear combinations $Z_{1t} = \max(0, X_t) - \max(0, -X_t) \equiv X_t$ and $Z_{2t} = \sqrt{\frac{\pi}{\pi-2}}\left(\max(0, X_t) + \max(0, -X_t)\right) \equiv \sqrt{\frac{\pi}{\pi-2}}|X_t|$, the first corresponding to the most predictable aspect and the second to the least. By the properties of the folded normal distribution, see Kan and Robotti (2017, page 933), it can be shown that Z_{2t} (equivalently $|X_t|$) is AR(1) with partial autocorrelation coefficient $\phi_{Z_2,11} = \phi c(\phi)$, where $|c(\phi)| < 1$, namely $c(\phi) = \frac{4\Phi_2(\phi) - 1 + \frac{2}{\pi}\left(\frac{\sqrt{1-\phi^2}}{\phi} - 1\right)}{1 - \frac{2}{\pi}}$, where $\Phi_2(\phi)$ denotes the bivariate standard normal cdf with correlation ϕ evaluated at (0,0). In this case, the most predictable aspect is X_t and the second (and least) predictable aspect is $|X_t|$. Notice that $Z_t^* = Z_t$, i.e., we have considered the identity transformation for $g(\cdot)$.

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 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 $X_t - q(\upsilon_j)$, where $q(\upsilon_j) = \inf\{x \in \mathbb{R} : F(x) \geq \upsilon_j\}$, $j = 1, \ldots, r$, is the quantile corresponding to the probability $\upsilon_j \in (0, 1)$. Some relevant choices are the following.

- *Hinge basis* functions with knots at the r^* quantiles $q_j = q(\upsilon_j)$, $\upsilon_j = \frac{j}{r^*+1}$, such that $h_{2j-1}(X_t) = \max\{0, X_t - q_j\}$, $h_{2j}(X_t) = \max\{0, q_j - X_t\}$, $j = 1, 2, \ldots, r^*$. 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.

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 $q_j = \ln(\upsilon_j/(1-\upsilon_j))$, $\upsilon_j = j/5, j = 1, 2, 3, 4$, i.e., the quintiles of the logistic distribution.

Remark 2. *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 $X_t^* = F_L^{-1}(F(X_t))$, where F is the distribution function of*

of X_t, estimated by the empirical distribution function of X_t, and $F_L^{-1}(u) = \log(u/(1-u))$ is the standard (unit scale) logistic quantile function. Then, considering the quantiles of the standard logistic distribution $q_j(\upsilon_j) = \log(\upsilon_j/(1-\upsilon_j))$, $\upsilon_j = \frac{j}{r+1}$, 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 $\{\upsilon_j - I(x_t < q_j), j = 1, \ldots, r\}$, that define Linton and Whang (2007) quantilogram.

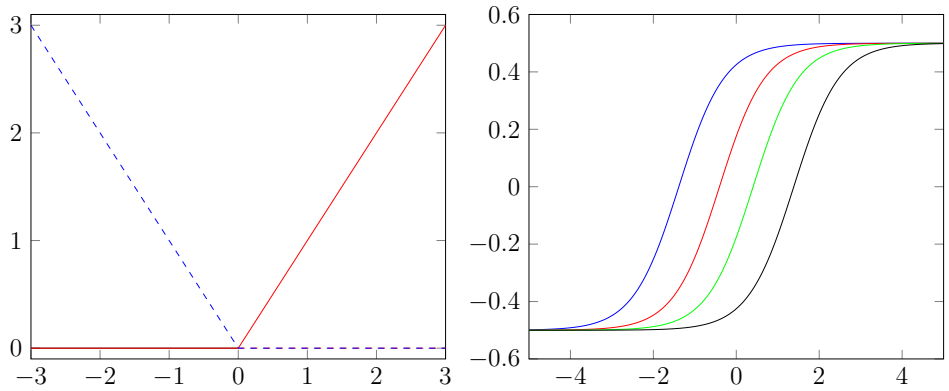

Figure 1: Left: Plot of $h(u) = \max\{0, u\}$ (solid red) and $h(u) = \max\{0, -u\}$ (dashed blue). Right: Plot of $h_j(u) = \{1+\exp((q_j - u)/\tau)\}^{-1}$, for $\tau = 1$ and $q_j = \ln(\upsilon_j/(1-\upsilon_j))$ and $\upsilon_j = j/5, j = 1, 2, 3, 4$.

In section S3 of the Supplementary Material, we shows that the problem of evaluating the most predictable aspect of the time series can be traced back to a nonlinear canonical correlation analysis of the past and the future of $(h_{1t}, \ldots, h_{rt})'$.

5 Statistical Inference

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

$$\hat{q}_j = \arg\min_{q \in \mathbb{R}} \sum_{t=1}^{T} \varsigma_{\upsilon_j}(x_t - q), \tag{2}$$

where $\varsigma_{v_j}(u) = u\{v_j - I(u < 0)\}$. Denoting $\hat{\mathbf{h}}_t = (h_1(x_t), \ldots, h_r(x_t))'$, the sample mean and covariance matrix of the vector $\mathbf{h}_t$ are respectively $\bar{\mathbf{h}} = T^{-1}\sum_{t=1}^{T}\mathbf{h}_t$ and $\hat{\mathbf{\Gamma}}_h(0) = T^{-1}\sum_{t=1}^{T}\left(\mathbf{h}_t - \bar{\mathbf{h}}\right)\left(\mathbf{h}_t - \bar{\mathbf{h}}\right)'$.

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

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

which is also a function of a bandwidth parameter, ℓ_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 $z_t = \boldsymbol{\beta}'\mathbf{h}_t$ at lag k, under the constraint $\boldsymbol{\beta}'\hat{\mathbf{\Gamma}}_h(0)\boldsymbol{\beta} = 1$. 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 $\kappa(u) = 1$, if $|u| \leq 1$, $\kappa(u) = 2 - |u|$, if $1 < |u| \leq 2$, and $\kappa(u) = 0$, if $|u| > 2$. The kernel weights are thus equal to 1 for $k \leq \ell_T$, decrease linearly to zero for $\ell_T < k \leq 2\ell_T$, and are identically zero for $k > 2\ell_T$. By construction, $\hat{Q}_T(\boldsymbol{\beta})$ is a finite sum, since the regularized partial autocorrelations are zero after lag $\lfloor 2\ell_T \rfloor$.

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 $\mathbf{\Gamma}_h(0)$ (scaled by the square root of the corresponding eigenvalue) for which the mutual information of the corresponding z_t variable is largest.

5.1 Large sample properties

Under regularity conditions concerning X_t, the nature of the basis functions $h_j(X_t), j = 1, \ldots, r$, and the design of the estimator, we can prove the consistency and the asymptotic normality of $\hat{\boldsymbol{\beta}}$.

The following assumptions will be made about X_t.

Assumption 1. *X_t is strictly stationary with absolutely continuous marginal distribution function $F(x)$, with continous density $f(x)$, and v_j-quantiles $q_j = F^{-1}(v_j)$, $j = 1, \dots, r$, such that $-\infty < a \leq q_1 < q_2 < \cdots < q_r \leq b < \infty$, and $0 < f(q_j) < \infty$.*

Assumption 2. *X_t is absolutely regular with strong mixing coefficient α_m of size -φ_0, with $\varphi_0 = 1 + \frac{1}{1+\delta}$, $\delta > 0$, and $E|X_t|^{4+2\delta}$.*

Recall that $\{X_t, t \in \mathbb{Z}\}$ is α-mixing (Dedecker et al., 2007; Davidson, 2021, Ch. 15) if $\lim_{m\to\infty} \alpha_m = 0$, where α_m is the mixing coefficient, defined as

$$\alpha_m = \sup_{B \in \mathcal{F}_{-\infty}^t, C \in \mathcal{F}_{t+m}^\infty} |P(B \cap C) - P(B)P(C)|,$$

and $\mathcal{F}_r^s, r < s$ is the σ-field generated by $\{X_r, X_{r+1}, \dots, X_s\}$. We also say that $\{X_t, t \in \mathbb{Z}\}$ is α-mixing of size $-\varphi_0$, $\varphi_0 > 0$, if $\alpha_m = O(m^{-\varphi}), \varphi > \varphi_0$.

If the transformation Z_t is bounded (see Remark 2), Assumption 2 can be relaxed.

Remark 3. *If $|Z_t| < C, C > 0$, then it suffices to assume that X_t is absolutely regular with α-mixing coefficients satisfying the summability condition $\sum_{m=1}^\infty m\alpha_m < \infty$.*

As for the selection of the features of the process, we will assume what follows.

Assumption 3. *The set of basis functions is chosen so that their number if fixed and known, $E|h_j(X_t)|^{4+2\delta} < \infty$, $\mathbf{\Gamma}_h(0)$ is non singular, and $h_j(X_t)$ is a Lipschitz continuous function of the quantile q_j.*

For the hinge and logistic bases the Lipschitz condition is satisfied, and $\mathrm{E}|h_j(X_t)|^{4+2\delta} < \infty$ is implied by Assumption 2. We stress that we assume in our setup that r is fixed. 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.

The next assumption deals with the design of the estimator of $\boldsymbol{\beta}$.

Assumption 4. *The bandwidth parameter of the trapezoidal kernel is chosen so that* $\ell_T = o(T^{1/4})$ *and* $\ell_T \geq r/2$.

Our last assumption deals with the existence of a solution to the problem of determining the most predictable aspect of X_t (in linear sense).

Assumption 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 $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^r : \boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1\}$, *is unique (apart from a sign change), i.e.,* $\boldsymbol{\beta}_0$ *is the unique fixed point of the nonlinear system* $\boldsymbol{\beta} = \boldsymbol{\Gamma}_h^{-1}(0)g(\boldsymbol{\beta})/\left(\boldsymbol{\beta}'g(\boldsymbol{\beta})\right)$, *where* $g(\boldsymbol{\beta}) = \partial Q_0(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Note that $\boldsymbol{\beta}_0$ is identified up to a sign change, i.e., $-\boldsymbol{\beta}_0$ is also a solution. The nonlinear system for $\boldsymbol{\beta}$ follows from the first order conditions for a maximum of the Lagrangian $\mathcal{L}(\boldsymbol{\beta}, \varpi) = Q_0(\boldsymbol{\beta}) - \frac{\varpi}{2}(\boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} - 1)$, where ϖ is the Lagrange multiplier.

The following theorem shows that under the stated assumptions $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ and its asymptotic sampling distribution is normal.

Theorem 2. *Under Assumptions 1-5,*

$$\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}. \tag{4}$$

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

and $\mathbf{G}_0 = plim\left\{\hat{\mathbf{G}}_T(\boldsymbol{\beta}_0)\right\}$,

$$\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \to_d N(\mathbf{0}, \mathbf{E}_0\boldsymbol{\Sigma}_0\mathbf{E}_0'), \tag{5}$$

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

Proof. See Appendix C. □

5.2 Selection of the bandwidth parameter and of the number of basis functions

The bandwidth of the trapezoidal kernel 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. Proietti and Giovannelli (2018) adopt a data-based selection criterion, adapted from McMurry and Politis (2010), which chooses $\hat{\ell}_T$ as the smallest value of ℓ_T such that $|\hat{\phi}_{z,kk}(\ell_T+k)| < c\{\log_{10} T/T\}^{1/2},\ k = 1, \ldots, K_n,\ K_n = o(\log_{10} T)$. 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 $\phi_{z,kk}(\ell_T + k) = 0$ $(k = 1, \ldots, K_n)$.

We have assumed r fixed. However, the number of basis functions should be selected. For this purpose, an information criterion based on Li and Xie (1996) can be used. This is evaluated as $\hat{Q}_T(\hat{\boldsymbol{\beta}}) - \frac{c\log\log(T)}{T}\left(\hat{L}_T(\hat{L}_T+1)/2 + r\right)$, where $c > 2$. 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 $x_t, t = 1, \ldots, T$, are generated with $T = 100, 250, 500, 1000, 5000$. 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 $\hat{Q}_T(\hat{\boldsymbol{\beta}})$.

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 higher autocorrelations with respect to x_t, the largest being close to the true value, equal 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 shows 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 $\hat{Q}_T(\hat{\boldsymbol{\beta}})$ did not vary with the choice of ℓ_T for values of ℓ_T between 1.5 and 4. In the top right panel we plot the transformation $z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}\right)\right)$, where $\hat{F}_Z(z)$ is the empirical

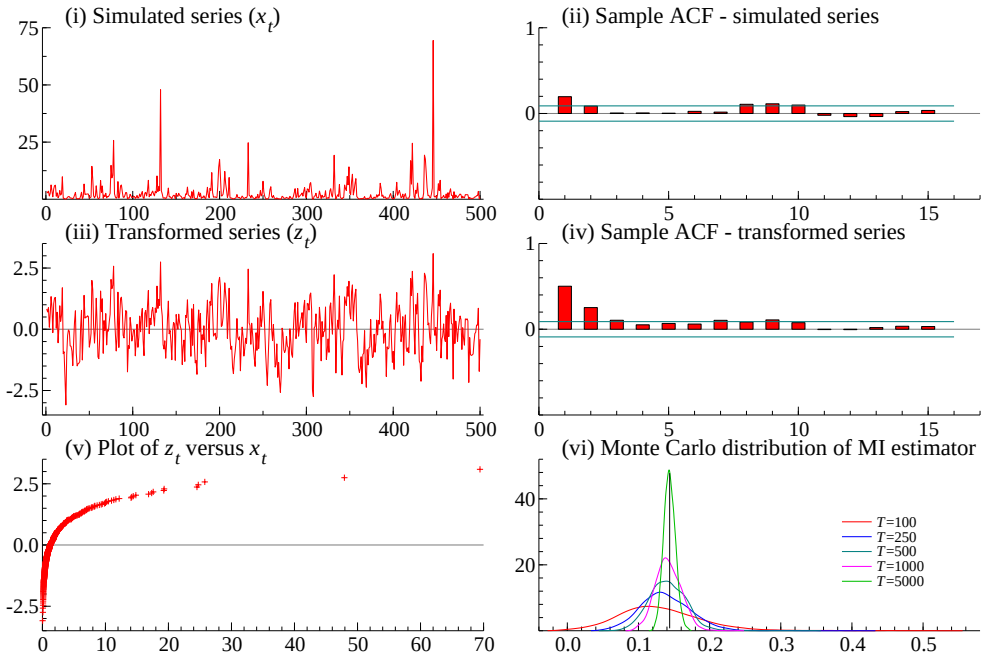

Figure 2: Log-normal AR(1). (i) Simulated series, x_t, with length $T = 500$, generated by the lognormal process $X_t = e^{Y_t}, Y_t = 0.2 + 0.5Y_{t-1} + \epsilon_t, \epsilon_t \sim$ i.i.d. $\mathrm{N}(0, 1)$. (ii) Sample ACF of x_t. (iii) Transformed time series, z_t. (iv) Sample ACF of z_t. (v) Plot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

cdf of $z_t^* = 1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}$, versus the original series and versus time. The transformation removes the concentration of values around zero and unveils 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/), 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 with $r = 3$ components, located at the quartiles of the distribution of $x_t^* =$

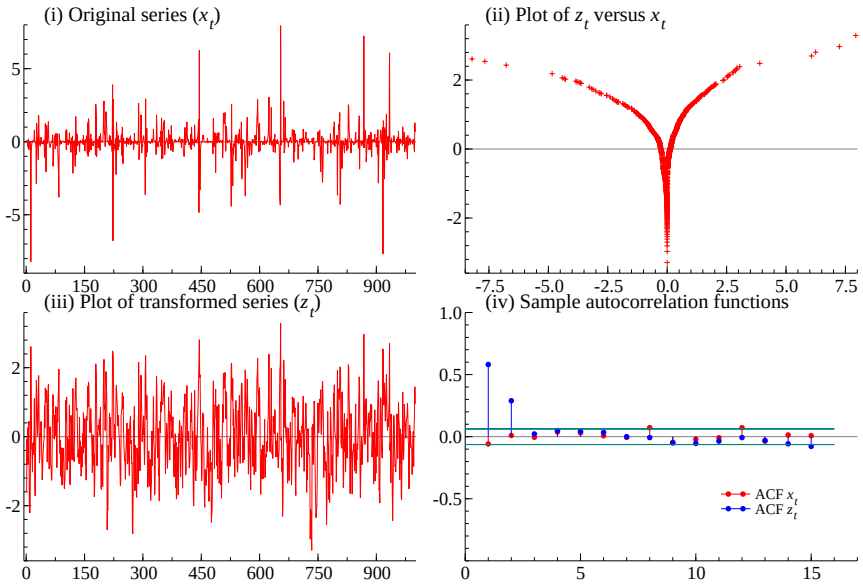

Figure 3: Nonlinear MA(2) process (i) Plot of the simulated series, x_t, $t = 1, \ldots, 1000$. (ii) Plot of $z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19 \cdot \max\{0, x_t - \hat{q}_{0.5}\} + 1.37 \cdot \max\{0, \hat{q}_{0.5} - x_t\}\right)\right)$ versus x_t. (iii) Time series plot of z_t. (iv) Sample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

$\log(\hat{F}_T(x_t)/(1 - \hat{F}_T(x_t))$. The latter is a bounded monotonic transformation, $\hat{F}_T(x)$ denoting the empirical distribution function of x_t.

The estimated 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 $z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19h_{1t} + 1.34h_{2t} + 1.19h_{3t}\right)\right)$, 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.

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 $T = 6088$

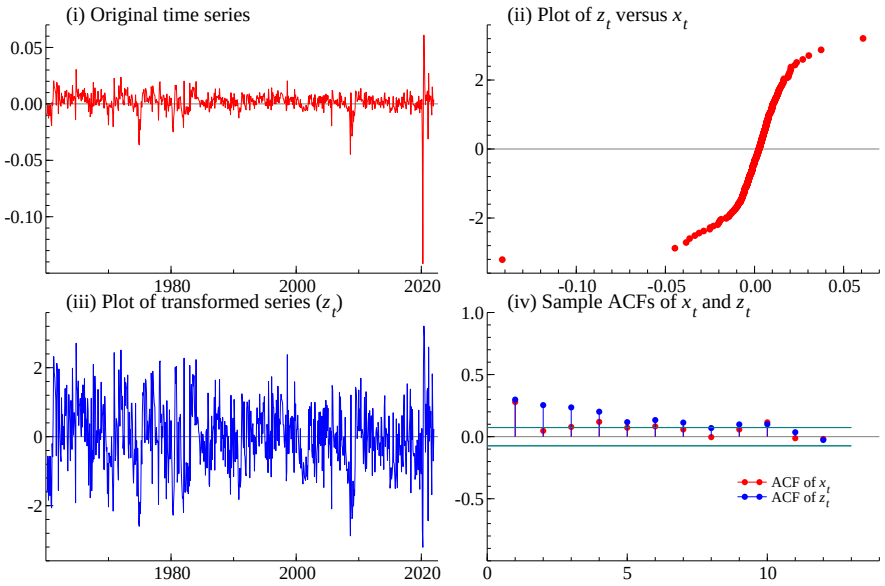

Figure 4: US Index of Industrial Production: relative changes with respect to the previous month. (i) Plot of the original time series (x_t). (ii) Plot of $z_t = \Phi^{-1}\left(\hat{F}_Z\left(1.19h_{1t} + 1.34h_{2t} + 1.19h_{3t}\right)\right)$, where $h_{jt} = 1/\{1 + \exp(-(x_t^* - \hat{q}_j^*)\}$, $x_t^* = \log(\hat{F}_T(x_t)/(1 - \hat{F}_T(x_t))$, 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).

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 $\hat{Q}_T(\boldsymbol{\beta})$ 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 $\ell_T = 10$. The covariance matrix of the two functions is $\hat{\boldsymbol{\Gamma}}_h(0) = \begin{pmatrix} 0.528 & -0.167 \\ -0.167 & 0.670 \end{pmatrix}$. 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 $z_t = \Phi^{-1}\left(\hat{F}_Z(z_t^*)\right), z_t^* = 1.133\max\{0, x_t - q_{0.5}\} + 1.031\max\{0, q_{0.5} - x_t\}$. Figure 6 plots z_t^*

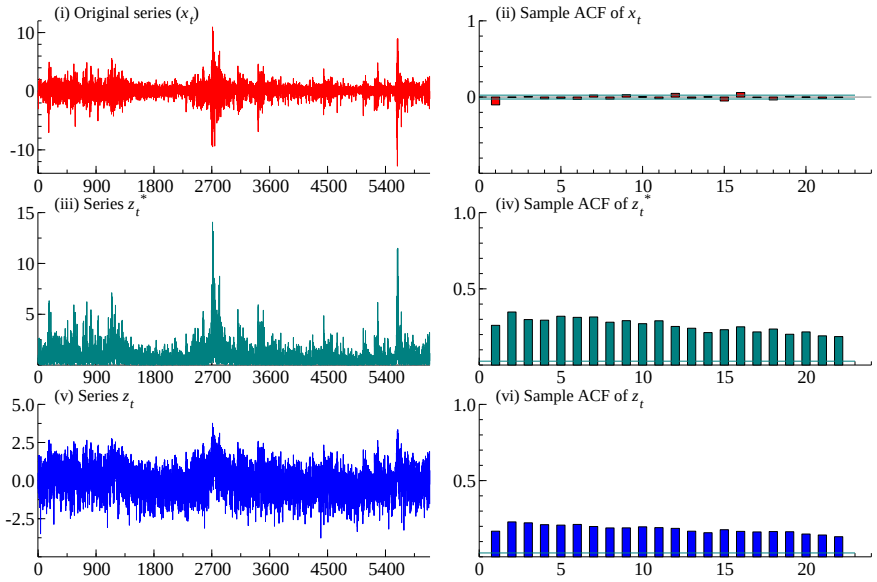

Figure 5: S&P 500 daily returns. (i) Plot of the original time series (x_t). (ii) Sample ACF plot of x_t. (iii) Time series plot of $z_t^* = 1.29 \max\{0, x_t - q_{0.5}\} + 0.85 \max\{0, q_{0.5} - x_t\}$. (iv) Sample ACF plot of z_t^*. (v) Time series plot of $z_t = \Phi^{-1}\left(\hat{F}_Z(z_t^*)\right)$. (vi) Sample ACF plot of 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 $w_t = \Phi^{-1}\left(\hat{F}_W w_t^*\right), w_t^* = 0.879 \max\{0, x_t - q_{0.5}\} - 0.746 \max\{0, q_{0.5} - x_t\}$. 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).

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)

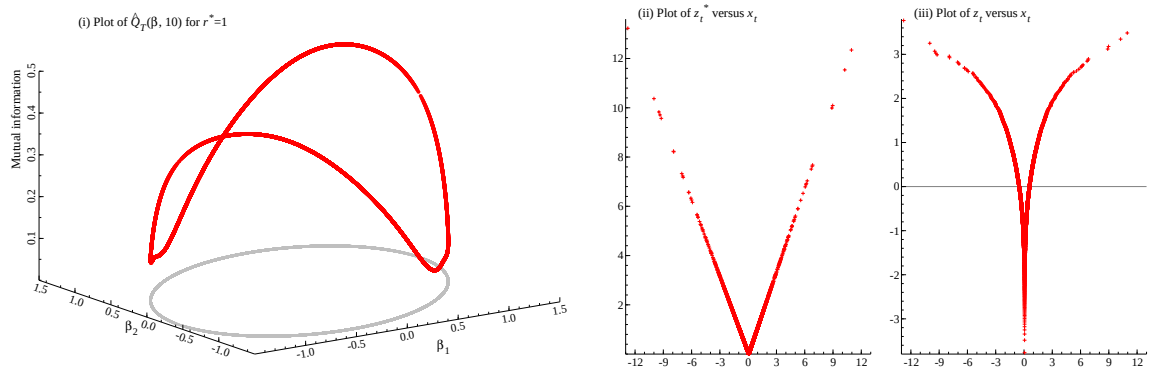

Figure 6: S&P 500 daily returns. (i) Plot of the mutual information as a function of $\boldsymbol{\beta}$, $\hat{Q}_T(\boldsymbol{\beta})$, evaluated at the points $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta} = 1$ (in grey), for $h_{1t} = \max\{0, x_t - q_{0.5}\}, h_{2t} = \max\{0, q_{0.5} - x_t\}$ and $\ell_T = 10$. (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.

to the series z_t. Here we propose considering Hong's (1996) test statistic:

$$\mathcal{H}_T(\kappa) = T\sum_{j=1}^{T-1}\mathcal{K}^2(j/B_T)\hat{\rho}_z^2(j), \quad B_T = 3T^{\kappa}, \tag{6}$$

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.

When appropriately standardized, the test statistic is asymptotically N(0,1). It has been shown by W. W. Chen and Deo (2004) that (6) suffers 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.

Section S3 of the Supplementary material 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 $r = 2r^*$, $r^* = 1, 2, 3, 5$, hinge-basis functions; As for the choice of κ, three values are considered: 0.2, 0.3, and 0.4.

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 the Supplementary Material (tables S1-S5), 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. $\mathrm{N}(0,1)$): 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 $X_t = 0.6\epsilon_{t-1}X_{t-2} + \epsilon_t$.

8 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.

Supplementary material

The supplement (a) relates the problem of determining the most predictable aspect of the time series to canonical correlation analysis of the cross-covariance matrix of the feature process $\mathbf{h}_t = (h_{1t}(X_t), \ldots, h_{rt}(X_t))$, subject to nonlinear constraints in the canonical vectors, and (b) and provides the full simulation results discussed in section 7.

A Proof of Theorem 1

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

Theorem A1 (MI decomposition)**.** *Let $Y, X = (X_1, X_2, \ldots, X_r)$ 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*

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

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

Proof. It follows from the easily established factorization $f(Y, X) = f(Y, X_r) \prod_{i=1}^{r-1} \frac{f(Y, X_i | X_{i+1}, \ldots, X_r)}{f(Y | X_{i+1}, \ldots, X_r)}$, that

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

Then, the above decomposition is obtained from

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

□

Corollary A1. *Given the continuous random variable Z, the conditional mutual information $I(Y, X|Z)$ has the following decomposition*

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

We are now ready to prove Theorem 1. For $m = 1$, apply Theorem A1 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 A1, applied with $Y = X_{n+m}$, $X = X_{1:n}$ and $Z = X_{n+1:n+m-1}$, yields

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

B Durbin–Levinson algorithm

The Durbin–Levinson algorithm (Durbin, 1960; Levinson, 1946) recursively computes the autoregressive coefficients of the optimal linear predictor based on $i = 1, 2, \ldots, k$, past observations and the variance of the corresponding prediction error, from the first k autocovariances.

Let $v_0 = \gamma(0)$, $\phi_{11} = \gamma(1)/\gamma(0)$, $v_1 = (1 - \phi_{11}^2)v_0$; then, for $i = 2, \ldots, k$, the Durbin-

Levinson (DL) algorithm is the following set of recursions:

$$\begin{aligned} \phi_{ii} &= \frac{\gamma(i)-\sum_{j=1}^{i-1}\phi_{i-1,j}\gamma(i-j)}{v_{i-1}}, \\ \phi_{ij} &= \phi_{i-1,j}-\phi_{ii}\phi_{i-1,i-j}, \quad (j=1,\ldots,i-1), \\ v_i &= (1-\phi_{ii}^2)v_{i-1}. \end{aligned} \tag{7}$$

The DL algorithm performs the factorization of the inverse of the autocovariance matrix of the random variables $\{X_{t-j}, j=0,\ldots,k-1\}$. Denoting $\mathbf{\Gamma}_k = \{\gamma(|i-j|), i,j=1,\ldots,k\}$, $\mathbf{\Gamma}_k^{-1} = \mathbf{\Phi}_k'\mathbf{D}_k\mathbf{\Phi}_k$, where $\mathbf{D}_k = \text{diag}\left(v_0^{-1}, v_1^{-1}, \ldots, v_{k-1}^{-1}\right)$, and

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

The mapping which transforms $\boldsymbol{\gamma} = (\gamma(0),\ldots,\gamma(k))$ into $(\phi_{11},\ldots,\phi_{kk},v_k)$ is one-to-one and continuously differentiable. The derivatives of the lag k partial autocorrelation with respect to $\boldsymbol{\gamma}$ is obtained by running the following recursions in parallel with the above DL recursions. Letting $\partial v_0/\partial\boldsymbol{\gamma} = (1,0,\ldots,0)'$, $\partial\phi_{11}/\partial\boldsymbol{\gamma} = \{\gamma(0)\}^{-1}(-\phi_{11},1,0,\ldots,0)'$, and defining $\mathbf{u}_i = (-\phi_{i-1,i-1},-\phi_{i-1,i-2},\ldots,-\phi_{i-1,1},1)'$, for $i=2,\ldots,k$,

$$\begin{aligned} \frac{\partial\phi_{ii}}{\partial\boldsymbol{\gamma}} &= \frac{1}{v_{i-1}}\left\{\mathbf{u}_i - \sum_{j=1}^{i-1}\gamma(i-j)\frac{\partial\phi_{i-1,j}}{\partial\boldsymbol{\gamma}} - \phi_{ii}\frac{\partial v_{i-1}}{\partial\boldsymbol{\gamma}}\right\}, \\ \frac{\partial\phi_{ij}}{\partial\boldsymbol{\gamma}} &= \frac{\partial\phi_{i-1,j}}{\partial\boldsymbol{\gamma}} - \phi_{i-1,i-j}\frac{\partial\phi_{ii}}{\partial\boldsymbol{\gamma}} - \phi_{ii}\frac{\partial\phi_{i-1,i-j}}{\partial\boldsymbol{\gamma}}, \quad (j=1,\ldots,i-1), \\ \frac{\partial v_i}{\partial\boldsymbol{\gamma}} &= (1-\phi_{ii}^2)\frac{\partial v_{i-1}}{\partial\boldsymbol{\gamma}} - 2\phi_{ii}v_{i-1}\frac{\partial\phi_{ii}}{\partial\boldsymbol{\gamma}}. \end{aligned} \tag{8}$$

C Proof of Theorem 2

Let $\mathbf{q} = (q_1,\ldots,q_r)$ denote the $r\times 1$ vector[1] containing the quantiles of the unconditional distribution of X_t, with $\mathbf{q}_0$ denoting the true quantiles and $\hat{\mathbf{q}}$ the estimated ones, and let

[1] Alternatively, r is replaced by $r^* = r/2$ if the hinge basis is considered.

$\boldsymbol{\theta} = (\mathbf{q}', \boldsymbol{\beta}')' \in \Theta$, $\Theta = \mathbb{R}^r \times \mathbb{B}$, $\mathbb{B} = \{\boldsymbol{\beta} \in \mathbb{R}^r : \boldsymbol{\beta}'\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta} = 1\}$. Rewrite $\mathbf{h}_t(\mathbf{q})$ to denote the $r \times 1$ vector containing the values of the basis functions evaluated at $\mathbf{q}$, and $\hat{\mathbf{h}}_t(\mathbf{q})$ for sample counterpart.

Recall that the transformed time series $\{z_t, t = 1, \ldots, T\}$ depends on $\mathbf{q}$ via the set of basis functions $\hat{\mathbf{h}}_t(\mathbf{q})$ and linearly on $\boldsymbol{\beta}$. To stress this dependence we will write $z_t(\boldsymbol{\theta}) = \boldsymbol{\beta}'\hat{\mathbf{h}}_t(\mathbf{q})$, the corresponding stochastic process as $Z_t(\boldsymbol{\theta}) = \boldsymbol{\beta}'\mathbf{h}_t(\mathbf{q})$. Let $\gamma_z(k;\boldsymbol{\theta}) = \mathrm{Cov}(Z_t(\boldsymbol{\theta}), Z_{t-k}(\boldsymbol{\theta}))$ and $\hat{\gamma}_z(k;\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=k+1}^{T}(z_t(\boldsymbol{\theta}) - \bar{z}(\boldsymbol{\theta}))(z_{t-k}(\boldsymbol{\theta}) - \bar{z}(\boldsymbol{\theta}))$, $\bar{z}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} z_t(\boldsymbol{\theta})$.

An intermediate large sample property concerns the convergence of $\hat{\gamma}_z(k;\boldsymbol{\theta})$, equivalently written $\hat{\gamma}_z(k;\mathbf{q},\boldsymbol{\beta})$, to $\gamma_z(k;\boldsymbol{\theta})$, also written $\gamma_z(k;\mathbf{q},\boldsymbol{\beta})$, for all $\boldsymbol{\theta} \in \Theta$. Recall that $\mathbf{q}$ does not depend on $\boldsymbol{\beta}$ and is estimated by the sample quantiles, $\hat{\mathbf{q}}$, of the marginal distribution of X_t, so that $\hat{\gamma}(k;\hat{\mathbf{q}},\boldsymbol{\beta}) - \hat{\gamma}(k;\mathbf{q},\boldsymbol{\beta}) = o_p(1)$, under Assumption 1.

Lemma C1. *The sample autocovariance function of $z_t(\boldsymbol{\theta}) = \boldsymbol{\beta}'\hat{\mathbf{h}}_t(\mathbf{q})$ converges uniformly in probability to $\gamma_z(k;\boldsymbol{\theta}) = Cov(Z_t(\boldsymbol{\theta}), Z_{t-k}(\boldsymbol{\theta}))$, i.e., $\sup_{\theta\in\Theta}|\hat{\gamma}_z(k;\boldsymbol{\theta}) - \gamma_z(k;\boldsymbol{\theta})| \to_p 0$.*

Proof. Assumptions 1-3 imply that $\hat{\gamma}_z(k;\boldsymbol{\theta})$ is convergent in mean square to $\gamma_z(k;\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$. Consider the centred random process $Y^*_{kt} = [Z_t(\boldsymbol{\theta}) - \mathrm{E}\{Z_t(\boldsymbol{\theta})\}][Z_{t-k}(\boldsymbol{\theta}) - \mathrm{E}\{Z_t(\boldsymbol{\theta})\}]$, and its sample analogue, $y^*_{kt} = [z_t(\boldsymbol{\theta}) - \bar{z}(\boldsymbol{\theta})][z_{t-k}(\boldsymbol{\theta}) - \bar{z}(\boldsymbol{\theta})]$. The process Y^*_{kt} is strong mixing with coefficient of size $-\varphi_0$, and defining $\hat{\gamma}^*_z(k;\boldsymbol{\theta}) = T^{-1}\sum_{t=1}^{T} Y^*_{kt}$, it holds that

$$\begin{aligned}
\mathrm{E}\left\{[\hat{\gamma}^*_z(k;\boldsymbol{\theta}) - \gamma_z(k;\boldsymbol{\theta})]^2\right\} &= \tfrac{1}{T^2}\mathrm{E}\left\{[\textstyle\sum_t (Y^*_{kt} - \mathrm{E}(Y^*_{kt})]^2\right\} \\
&\leq \tfrac{2}{T} C \textstyle\sum_{m=1}^{\infty}\{\alpha(m)\}^{\frac{\delta}{2+\delta}} \\
&= O(T^{-1}),
\end{aligned}$$

where the second line follows from Davidov's covariance inequality (Davydov, 1968), with $C = 12\{\mathrm{E}(|Y^*_{kt}|^{2+\delta})\}^2$, whose finiteness is implied by Assumption 2. Then, simple manipulations show that $\hat{\gamma}^*(k;\boldsymbol{\theta}) - \hat{\gamma}(k;\boldsymbol{\theta}) = O_p\left(T^{-2}[\sum_t (Y_{kt} - \mathrm{E}(Y_{kt}))]^2\right) \to_p 0$.

Finally, $\hat{\gamma}_z(k;\boldsymbol{\theta})$ is a Lipschitz continuous function of $\boldsymbol{\theta} \in \Theta$: let $\boldsymbol{\theta} = (\mathbf{q}', \boldsymbol{\beta}')', \boldsymbol{\theta}^* =$

$(\mathbf{q}^{*'}, \boldsymbol{\beta}^{*'})' \in \Theta$, then

$$
\begin{aligned}
|\hat{\gamma}_z(k;\mathbf{q},\boldsymbol{\beta}) - \hat{\gamma}_z(k;\mathbf{q}^*,\boldsymbol{\beta}^*)| &\leq |\hat{\gamma}_z(k;\mathbf{q},\boldsymbol{\beta}) - \hat{\gamma}_z(k;\mathbf{q}^*,\boldsymbol{\beta})| + |\hat{\gamma}_z(k;\mathbf{q}^*,\boldsymbol{\beta}) - \hat{\gamma}_z(k;\mathbf{q}^*,\boldsymbol{\beta}^*)| \\
&= |\boldsymbol{\beta}'\{\hat{\boldsymbol{\Gamma}}_h(k,\mathbf{q}) - \hat{\boldsymbol{\Gamma}}_h(k,\mathbf{q}^*)\}\boldsymbol{\beta}| \\
&\quad + |\boldsymbol{\beta}^{*'}\hat{\boldsymbol{\Gamma}}_h^{'}(k,\mathbf{q}^*)(\boldsymbol{\beta}-\boldsymbol{\beta}^*) + (\boldsymbol{\beta}-\boldsymbol{\beta}^*)'\hat{\boldsymbol{\Gamma}}_h(k,\mathbf{q}^*)(\boldsymbol{\beta}-\boldsymbol{\beta}^*)| \\
&\leq C_{1k,T}\|\mathbf{q}-\mathbf{q}^*\|^2 + C_{2k,T}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2 \\
&\leq C_{3k,T}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2,
\end{aligned}
$$

with $C_{ik,T} = O_p(1), i = 1,2,3$. The first inequality arises since, if we refer to the hinge basis, the j-th component of the vector $|\hat{h}_{jt}(\mathbf{q}) - \hat{h}_{jt}(\mathbf{q}^*)| = |\max(0, x_t - q_j) - \max(0, x_t - q_j^*)|$, equals 0 if $x_t \leq (q_j \wedge q_j^*)$, and it is not greater than $|q_j - q_j^*|$ otherwise. Similar considerations hold for the logistic basis.

In summary, $\hat{\gamma}_z(k;\boldsymbol{\theta})$ converges pointwise to $\gamma_z(k;\boldsymbol{\theta})$ and satisfies the above Lipschitz condition. Hence, by Theorems 2.9 and 2.11 in Davidson (2021), it converges uniformly to $\gamma_z(k;\boldsymbol{\theta})$. □

The above Lipschitz property holds also for the sample partial autocorrelation of $Z_t(\boldsymbol{\theta})$, $\hat{\phi}_{z,kk}(\boldsymbol{\theta})$, which are a continuous and differentiable function of the sample autocovariances. The differential, which can be obtained recursively by differentiating the Durbin-Levinson recursions, is linear in $\hat{\gamma}(i;\boldsymbol{\theta}) - \hat{\gamma}(i;\boldsymbol{\theta}^*), i = 0,1,\ldots,k$, with bounded coefficients, see Appendix B.

The second large sample property that will be needed in the sequel concerns the asymptotic normality of $\hat{\gamma}_z(k;\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$.

Lemma C2. *Let* $\hat{\boldsymbol{\gamma}}_L(\boldsymbol{\theta}) = (\hat{\gamma}_z(0,\boldsymbol{\theta}), \hat{\gamma}_z(1,\boldsymbol{\theta}), \ldots, \hat{\gamma}_z(L,\boldsymbol{\theta}))'$, *and define* $\boldsymbol{\gamma}_L(\boldsymbol{\theta})$ *the corresponding population vector. Then,* $\sqrt{T}\{\hat{\boldsymbol{\gamma}}_L(\boldsymbol{\theta}) - \boldsymbol{\gamma}_L(\boldsymbol{\theta})\} \to_d N(\mathbf{0}, \mathbf{W}_L)$, *where* $\mathbf{W} = \{w_{kl}; k,l = 1,2,\ldots,L\}$, *with generic element given by the Bartlett's formula* $w_{kl} = \sum_{j=-\infty}^{\infty}\{\gamma_z(j)\gamma_z(j-k+l) + \gamma_z(j+k)\gamma_z(j+l) + \kappa(k,-j,l-j)\}$, *and* $\kappa(i,j,k)$ *is*

the fourth cumulant of $(Z_t(\boldsymbol{\theta}), Z_{t+i}(\boldsymbol{\theta}), Z_{t+j}(\boldsymbol{\theta}), Z_{t+k}(\boldsymbol{\theta}))$.

Proof. The result follows from Theorem 18.5.3 in Ibragimov and Linnik (1971): Assumptions 2 and 3 and Theorem 15.1 in Davidson (2021) imply that $Z_t(\boldsymbol{\theta})$ and $Y_t(\boldsymbol{\theta}) = Z_t(\boldsymbol{\theta})Z_{t-k}(\boldsymbol{\theta})$ are α-mixing of size $-\varphi_0$, $\mathrm{E}(|Y_t(\boldsymbol{\theta})|^{4+2\delta}) < \infty$, and $\sum_m \{\alpha_Y(m)\}^{\delta/(2+\delta)} < \infty$, where $\alpha_Y(m)$ is the mixing coefficient of $Y_t(\boldsymbol{\theta})$. The evaluation of the long run variance of $Y_t(\boldsymbol{\theta})$ leads to the above Bartlett's formula, see e.g. Anderson (1971, ch. 8) and Keenan (1997). Finally, $\mathrm{E}(|Y_t(\boldsymbol{\theta})|^{4+2\delta}) < \infty$ implies that $\sum_j |\kappa(k, -j, l-j)| < \infty$. □

The sample partial autocorrelations of $Z_t(\boldsymbol{\theta})$ are continuous and differentiable functions of the sample autocovariances. The derivatives can be obtained recursively by differentiating the Durbin-Levinson recursions, as in Appendix B. Hence, writing $\hat{\boldsymbol{\phi}}_L = \left(\hat{\phi}_{11}(\boldsymbol{\beta}), \ldots, \hat{\phi}_{LL}(\boldsymbol{\beta})\right)$ and $\boldsymbol{\phi}_L(\boldsymbol{\theta}) = (\phi_{11}(\boldsymbol{\beta}), \ldots, \phi_{LL}(\boldsymbol{\beta}))$, and applying the delta method, it holds that $\sqrt{T}\left\{\hat{\boldsymbol{\phi}}_L(\boldsymbol{\theta}) - \boldsymbol{\phi}_L(\boldsymbol{\theta})\right\} \to_d \mathrm{N}(\underline{0}, \mathbf{J}_\phi \mathbf{W}_L \mathbf{J}'_\phi)$, where $\mathbf{J}_\phi = \partial \boldsymbol{\phi}_L(\boldsymbol{\theta})/\partial \boldsymbol{\gamma}_L$.

Proof of $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$ (consistency). Under Assumptions 1 and 2 the sample quantiles converge in probability to the population quantiles, $\hat{\mathbf{q}} \to_p \mathbf{q}$. Notice that $\hat{Q}_T(\boldsymbol{\beta})$ in (3) can be written as $\hat{Q}_T(\hat{\mathbf{q}}, \boldsymbol{\beta})$, and $|\hat{Q}_T(\hat{\mathbf{q}}, \boldsymbol{\beta}) - \hat{Q}_T(\mathbf{q}, \boldsymbol{\beta})| = o_p(1)$.

The consistency of $\hat{\boldsymbol{\beta}}$ follows from Theorem 2.1 in Newey and McFadden (1994), since all the assumptions are satisfied: i. Θ is a compact set; ii. $\hat{Q}_T(\boldsymbol{\theta}) = \hat{Q}_T(\mathbf{q}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \Theta$ and is a measurable function of $\{X_t, t = 1, \ldots, T\}$. iii. $\hat{Q}_T(\boldsymbol{\theta})$ converges uniformly in probability to $Q_0(\boldsymbol{\theta})$, i.e., $\sup_{\theta \in \Theta} \left|\hat{Q}_T(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\right| \to_p 0$. This last property follows from Lemma 2.9 in Newey and McFadden (1994), as $Q_0(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$, and, for all $\boldsymbol{\theta} \in \Theta$, $\hat{Q}_T(\boldsymbol{\theta}) \to_p Q_0(\boldsymbol{\theta})$ (pointwise convergence); finally, for all $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta$, $|\hat{Q}_T(\boldsymbol{\theta}) - \hat{Q}_T(\boldsymbol{\theta}^*)| \leq B_T \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$ with $B_T = O_p(1)$ (Lipschitz condition).

Pointwise convergence is proved as follows. By the triangle inequality,

$$\begin{aligned}\left|\hat{Q}_T(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})\right| &\leq \left|\hat{Q}_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta})\right| + |Q_T(\boldsymbol{\theta}) - Q_0(\boldsymbol{\theta})| \\ &\leq C_1 \textstyle\sum_{k=1}^{\lfloor \ell_T \rfloor} k|\hat{\phi}_{z,kk} - \phi_{z,kk}| + C_2 \sum_{\lfloor \ell_T \rfloor + 1}^{\infty} k\phi_{z,kk}^2 \\ &\leq C_3 \tfrac{\lfloor \ell_T \rfloor (\lfloor \ell_T \rfloor + 1)}{2T^{1/2}} + C_2 \textstyle\sum_{\lfloor \ell_T \rfloor + 1}^{\infty} k\phi_{z,kk}^2\end{aligned}$$

where $C_i, i = 1, 2, 3$, are positive constants. The first term on the right hand results from the mean value theorem expansion $-\frac{1}{2}\log(1 - \hat{\phi}_{z,kk}^2) = -\frac{1}{2}\log(1 - \phi_{z,kk}^2) + \frac{\phi_{z,kk}^2}{1-\bar{\phi}_{kk}^2}(\hat{\phi}_{z,kk} - \phi_{z,kk}) + 0.5\frac{1+\bar{\phi}_{kk}^2}{(1-\bar{\phi}_{z,kk}^2)^2}(\hat{\phi}_{z,kk} - \phi_{z,kk})^2$, where $\bar{\phi}_{z,kk}$ is an intermediate point between ϕ_{kk} and $\hat{\phi}_{z,kk}$. Its probability limit is zero by Assumption 4. The second addend uses the first order Maclaurin expansion $-\frac{1}{2}\log(1 - \phi_{z,kk}^2) = \phi_{kk}^2 + O(\phi_{z,kk}^4)$. By Assumption 2 this term is $O(\sum_{k=\lfloor \ell_T \rfloor + 1}^{\infty} k\alpha_k^2)$ and thus converges to zero as $\ell_T \to \infty$.

Finally, Lipschitz continuity of $\hat{Q}_T(\boldsymbol{\theta})$ follows from that of the sample autocovariance function. If we define the vector $\boldsymbol{\psi}_t(\boldsymbol{\theta})$ with L_T+1 elements $Y_{kt}^*, k = 0, 1, \ldots, L_T$, where Y_{kt^*} was defined above, then it can be seen that $\hat{Q}_T(\boldsymbol{\beta})$ is a continuous function of $T^{-1}\sum_t \boldsymbol{\psi}_t$, via the sample partial autocorrelations $\hat{\phi}_{z,kk}(\boldsymbol{\theta})$, which are a continuous and differentiable function of $\hat{\boldsymbol{\gamma}}_L(\boldsymbol{\theta})$, with bounded first derivative. Hence, there exist a constant $C > 0$, such that $|\hat{Q}_T(\boldsymbol{\theta}) - \hat{Q}_T(\boldsymbol{\theta}^*)| \leq B_T\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$ where $B_T = C\sum_{k=1}^{L_T} k\|\hat{\boldsymbol{\Gamma}}_h(k, \boldsymbol{\theta})\|$, and $B_T = O_p(1)$, as implied by Assumption 2 and the maximum norm inequality $\|\mathbf{A}\| \leq r\max\{|a_{ij}|\}$, where r is the row and column dimension of $\mathbf{A}$. Hence, $\hat{Q}_T(\boldsymbol{\theta})$ satisfies a Lipschitz condition and all the regularity conditions of Lemma 2.9 in Newey and McFadden (1994) are satisfied.

Asymptotic normality of $\hat{\boldsymbol{\beta}}$ Consider the Lagrangian $\mathcal{L}(\boldsymbol{\beta}, \varpi) = \hat{Q}_T(\boldsymbol{\beta}) - \frac{1}{2}\varpi\boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta}$. Let us denote $\hat{\mathbf{g}}_T(\boldsymbol{\beta}) = \frac{\partial \hat{Q}_T(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$. The first order conditions for the problem are $\hat{\mathbf{g}}_T(\boldsymbol{\beta}) - \varpi\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta} = 0,\ \boldsymbol{\beta}'\hat{\boldsymbol{\Gamma}}_h(0)\boldsymbol{\beta} - 1 = 0$; premultiplying the first equation by $\boldsymbol{\beta}'$ gives $\varpi = \boldsymbol{\beta}'\hat{\mathbf{g}}_T(\boldsymbol{\theta})$. Hence, the constrained solution $\hat{\boldsymbol{\beta}}$ satisfies the nonlinear system $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Gamma}}_h^{-1}(0)\hat{\mathbf{g}}_T(\hat{\boldsymbol{\beta}})/\hat{\boldsymbol{\beta}}'\hat{\mathbf{g}}_T(\boldsymbol{\theta})$.

By a first order Taylor's expansion of the first order conditions around $\boldsymbol{\beta}_0$, as in Davidson (2000, Sec. 12.3), after scaling by $\sqrt{T}$, we get

$$\begin{pmatrix} \hat{\mathbf{G}}_T(\boldsymbol{\beta}^*) & \hat{\boldsymbol{\Gamma}}_h(0)\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Gamma}}_h(0) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ \sqrt{T}\hat{\varpi} \end{pmatrix} = \begin{pmatrix} \sqrt{T}\hat{\mathbf{g}}_T(\boldsymbol{\beta}_0) \\ \mathbf{0} \end{pmatrix}.$$

Let now $\mathbf{G}_0 = \text{plim}\hat{\mathbf{G}}_T(\boldsymbol{\beta}^*)$, and recalling $\hat{\boldsymbol{\Gamma}}_h(0)\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\Gamma}_h(0)\boldsymbol{\beta}_0$, $\sqrt{T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \to_p \mathbf{E}_0\sqrt{T}\hat{\mathbf{g}}_T(\boldsymbol{\beta}_0)$, where $\mathbf{E}_0 = \mathbf{G}_0^{-1} - \frac{1}{\boldsymbol{\beta}_0'\boldsymbol{\Gamma}_h(0)\mathbf{G}_0^{-1}\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta}_0}\mathbf{G}_0^{-1}\boldsymbol{\Gamma}_h(0)\boldsymbol{\beta}_0\boldsymbol{\beta}_0'\boldsymbol{\Gamma}_h(0)\mathbf{G}_0^{-1}$ is the top-left block of the inverse matrix of the probability limit of the matrix on the left hand side.

The term on the right hand side is a function of the sample autocovariances. By the mean value theorem, $\hat{\mathbf{g}}_T(\boldsymbol{\beta}_0) = g_0(\boldsymbol{\beta}_0) + \mathbf{M}_T^*(\hat{\boldsymbol{\gamma}}_{L_T} - \boldsymbol{\gamma}_{L_T})$, where $\mathbf{M}_T^* = \left.\frac{\hat{\mathbf{g}}_T(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\gamma}_{L_T}}\right|_{\gamma_{L_T}=\gamma_{L_T}^*}$, $\boldsymbol{\gamma}_{L_T}^*$ is a point intermediate between $\hat{\boldsymbol{\gamma}}_{L_T}$, and $\boldsymbol{\gamma}_{L_T}$, and $\mathbf{g}_0(\boldsymbol{\beta}_0)$ converges to zero.

By the properties of the sample autocovariances (Lemma 2), denoting $\boldsymbol{\Sigma}_T = \mathbf{M}_T\mathbf{W}_{L_T}\mathbf{M}_T' \to_p \boldsymbol{\Sigma}_0$, we have that $\sqrt{T}\hat{\mathbf{g}}_T(\boldsymbol{\beta}_0) \to_d N(\mathbf{0}, \boldsymbol{\Sigma}_0)$, provided that $L_T \geq r$ (which follows from Assumption 4).

Hence, $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d \text{N}(\mathbf{0}, \mathbf{E}_0\boldsymbol{\Sigma}_0\mathbf{E}_0')$.

References

Anderson, T. W. (1971). *The statistical analysis of time series*. John Wiley & Sons.

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.

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.

Davidson, J. (2000). *Econometric theory.* John Wiley & Sons.

Davidson, J. (2021). *Stochastic limit theory: An introduction for econometricians.* Oxford University Press, Oxford.

Davydov, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, *13*(4), 691–696.

Dedecker, J., Doukhan, P., Lang, G., Rafael, L. R. J., Louhichi, S., and Prieur, C. (2007). *Weak dependence.* Springer.

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.

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.

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 Linnik, Y. V. (1971). *Independent and stationary sequences of random variables*.

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.

Kan, R., and Robotti, C. (2017). On moments of folded and truncated multivariate normal distributions. *Journal of Computational and Graphical Statistics*, *26*(4), 930–934.

Keenan, D. M. (1997). A central limit theorem for $m(n)$ autocovariances. *Journal of Time Series Analysis*, *18*(1), 61–78.

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.

McNeil, A. J. (2021). Modelling volatile time series with v-transforms and copulas. *Risks*, *9*(1), 14.

Newey, W. K., and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

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*.

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

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.

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.