Auto-Hellinger Correlation for Serial Dependence: Rank-Based Tests

Burak Alparslan Eroğlu* Haluk Yener† Taner Yiğit‡ Ramazan Ekinci§

July 24, 2026

Abstract

In this paper, we propose the Auto-Hellinger Correlation Function (AHCF), a lag-specific copula-based measure of serial dependence. Unlike classical autocorrelation, the AHCF detects linear, nonlinear, nonmonotonic, and volatility-driven dependence while remaining compatible with Gaussian linear structures. We develop a rank-based estimator together with exactly distribution-free individual-lag and portmanteau tests based on an order-adaptive minimum-p procedure. Under the i.i.d. calibration null, the proposed tests attain finite-sample size no larger than the nominal level and are consistent against fixed alternatives. Simulations show that the AHCF provides more reliable short-sample size control and higher average power than auto-distance correlation for several nonlinear alternatives, while the kernel Hellinger benchmark achieves the highest average power. An application to professional survey forecast errors demonstrates the practical value of the proposed methodology by detecting robust pre-COVID dependence in US CPI forecast errors and correctly localizing dependence generated by overlapping ECB forecast horizons.

Keywords: Hellinger Correlation; Nonlinear Dependence; Copula; Bhattacharyya Affinity; Serial Dependence; Forecast Evaluation; Portmanteau Testing

JEL Codes: C14, C15, C22, C53

MSC Codes: 62H20, 62M10, 62G10

1 Introduction

Serial dependence is a central problem in time series econometrics. Classical approaches focus on linear dependence and measure it through the autocorrelation function and the

*Corresponding Author: İzmir Bakırçay University, Department of Economics, Gazi Mustafa Kemal District, Kaynaklar Road, Menemen/İzmir/Türkiye, email:burak.eroglu@bakircay.edu.tr, ORCID: 0000-0001-6814-747X

†Özyeğin University, International Trade and Business, Çekmeköy Campus Nişantepe District, Orman Street, Çekmeköy - İstanbul/Türkiye, email:haluk.yener@ozyegin.edu.tr

‡İhsan Doğramacı Bilkent University, Department of Economics, Çankaya, Ankara/Türkiye, email:tyigit@bilkent.edu.tr, ORCID: 0000-0003-0620-8905

§İzmir Bakırçay University, Department of Economics, Gazi Mustafa Kemal District, Kaynaklar Road, Menemen/İzmir/Türkiye, email:ramazan.ekinci@bakircay.edu.tr, ORCID: 0000-0001-7420-9841

associated portmanteau tests (Box and Pierce, 1970; Ljung and Box, 1978). These tools are well understood and widely used, but they capture only second-moment dependence and may miss nonlinear, nonmonotonic, or distributional forms of serial dependence. This limitation is particularly relevant for economic and financial time series, where dependence often persists even when linear autocorrelations are weak or negligible (Ashley et al., 1986; Giannerini et al., 2015; Granger et al., 2004; Rusticelli et al., 2008). In this paper, we introduce the Auto-Hellinger Correlation Function (AHCF), a lag-specific copula-based measure of serial dependence, together with exact, distribution-free tests of serial independence.

A growing literature has proposed alternative measures of serial dependence based on entropy, copulas, and related distributional concepts. Among these, the Hellinger-distance-based measure of Granger et al. (2004) captures both linear and nonlinear serial structure and provides both individual-lag and omnibus tests. Its implementation, however, relies on kernel density estimation, bandwidth selection, and computationally intensive permutation calibration. These features motivate a procedure with comparable generality but exact finite-sample validity and no bandwidth selection.

The AHCF extends the Hellinger correlation coefficient of Geenens and Lafaye de Micheaux (2022) to the time series setting. Unlike classical autocorrelation, which depends only on covariance, the AHCF depends on the full copula of (y_t, y_{t-q}), allowing it to detect nonlinear, nonmonotonic, and volatility-driven dependence while remaining compatible with Gaussian linear structures. Building on this lag-specific measure, we develop a rank-based testing framework for serial independence at individual lags and over fixed lag ranges. Because the tests are constructed entirely from within-sample ranks, they are exactly distribution-free under continuous marginal distributions. Our contribution is threefold.

First, we develop the population theory and estimation of the AHCF. Serial dependence is represented through the Bhattacharyya affinity between the lag-q copula and the independence copula, yielding a Hellinger correlation based on the Legendre expansion of the square-root copula density. We construct a Beta-transformed, rank-based nearest-neighbor estimator of the corresponding coefficient vector and establish its consistency under strict stationarity, geometric absolute regularity, and global smoothness conditions on the copula density. Unlike existing nearest-neighbor analyses, the proof avoids high-level moment assumptions through the combination of Beta-transformed coordinates and a deterministic boundary argument.

Second, we develop an inferential framework for this nonregular problem. Under serial independence, the Hellinger-correlation map has a boundary singularity, so inference is based on squared fixed-order AHCF statistics, $Z_{m,q} = T\widehat{\eta}_{m,m}(q)^2$, combined through an order-adaptive minimum-p procedure. Component p-values are calibrated using rank-based

Monte Carlo draws, and the minimum is calibrated using an independent set of combination draws. Under the i.i.d. calibration null, the resulting tests are exactly distribution-free over continuous marginal distributions and attain unconditional finite-sample size no larger than the nominal level. They are consistent against any fixed alternative represented by the chosen basis and require neither bandwidth selection nor an asymptotic distribution for the estimator.

Third, we extend the framework to portmanteau testing over fixed lag ranges. The same calibration applies to a broad class of lag aggregation rules, separating exact finite-sample validity from the choice of aggregation, which primarily affects power. We study a maximum-over-lags aggregation that is well suited to alternatives concentrated at a small number of unknown lags. Simulation results compare the proposed procedure with the literature implementations of auto-distance correlation and the kernel Hellinger benchmark of Granger et al. (2004). The AHCF tests retain exact finite-sample validity under the i.i.d. calibration null and require no bandwidth selection.

The empirical application considers professional survey forecast errors, where samples are short and dependence may operate through the full error distribution rather than only its conditional mean. For nonoverlapping US one-quarter-ahead forecasts, the AHCF detects robust pre-COVID lag-one dependence in CPI forecast errors that is stable across both median and mean consensus forecasts, whereas the Ljung–Box omnibus test does not reject. The GDP evidence is sensitive to the consensus definition and is therefore interpreted more cautiously. For overlapping one-year-ahead ECB forecasts, the AHCF correctly localizes the mechanically induced dependence to the expected short lags, with no evidence beyond the overlap after Holm adjustment. These applications illustrate the practical value of exact finite-sample calibration and lag localization. A rejection indicates dependence in the forecast-error distribution, but it does not imply that the point forecast itself could be improved.

The paper proceeds as follows. Section 2 develops the AHCF, introduces the Beta-transformed rank-based nearest-neighbor estimator, establishes consistency of the coefficient vector, and develops the order-adaptive individual-lag and portmanteau tests together with their finite-sample validity and consistency theory. Section 3 presents simulation evidence on size and power, Section 4 reports the empirical application, and Section 5 concludes.

Notation. The symbol $:=$ denotes definition. $\mathbb{1}\{\cdot\}$ is the indicator function. $\|\cdot\|$ is the Euclidean norm. $\mathbb{E}[\cdot]$ and $\Pr(\cdot)$ denote expectation and probability. $\mathbb{R}$ is the real line, and $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ are the floor and ceiling functions. $o_p(1)$ and $O_p(1)$ denote sequences that converge to zero and are bounded in probability, respectively. $\xrightarrow{p}$ denotes convergence in probability. $^\top$ denotes the transpose of a vector or matrix.

2 AHCF Diagnostics: Population Functional, Limit Theory, and Testing

This section develops the AHCF and its inferential framework. We first extend the Hellinger correlation coefficient of Geenens and Lafaye de Micheaux (2022) to lagged pairs (y_t, y_{t-q}), yielding the AHCF. We then develop a rank-based nearest-neighbor estimator and establish its consistency under primitive weak-dependence and copula-smoothness conditions. Finally, we construct order-adaptive individual-lag and portmanteau tests for serial independence with exact rank-based Monte Carlo calibration. Although conceptually related to the Hellinger-distance measure of Granger et al. (2004), the proposed framework differs in its estimator, finite-order normalization, and inferential approach.

2.1 Auto-Hellinger correlation function

Let $\{y_t\}_{t=1}^{\infty}$ be a real-valued strictly stationary process. For a fixed integer lag $q \geq 1$, we denote by F_{0q} the joint distribution of (y_t, y_{t-q}) and by F_0 and F_q the marginal distributions of y_t and y_{t-q}, respectively. Under strict stationarity these objects do not depend on t and $F_0 = F_q$ (we retain the separate notation F_0 and F_q for consistency with the copula-based representation developed below).

Assumption 1. (a) The marginal cumulative distribution functions F_0 and F_q are continuous.

(b) For each $q \in \mathbb{Z}^+$, the copula of (y_t, y_{t-q}) (the lag-q copula) has density c_q on $[0,1]^2$. Moreover, c_q is bounded away from 0 and ∞, and locally Lipschitz.

Assumption 1 collects the baseline regularity conditions that make the copula-based Hellinger dependence measure well defined. Part (a) ensures that the marginal probability integral transforms (PITs) are well defined and that their rank-based empirical counterparts are uniformly consistent. It also guarantees that ranks are almost surely free of ties, which underlies the exact distribution-freeness of the tests in Section 2.2. Part (b) imposes smoothness and nondegeneracy of the lag-q copula density. Boundedness away from zero and infinity prevents pathological local behavior of the copula density and keeps the dependence measure well behaved, while Lipschitz continuity controls the local approximation errors underlying the nonparametric estimation. The consistency theory in Section 2.1.3 adds the separate, primitive conditions of Assumption 2, which complement Assumption 1(b) by supplying the global weak-dependence and joint-density regularity required for estimation. In particular, geometric absolute regularity provides the dependence structure that drives the proofs. Because geometric absolute regularity implies

geometric strong mixing through $\alpha(j) \leq \beta(j)$, it makes available the covariance and Bernstein inequalities used in the estimation stage, so no separate strong-mixing assumption is required. At the independence boundary the AHCF transformation is nonregular, so the tests are calibrated exactly by rank-based Monte Carlo draws rather than by an asymptotic reference distribution.

Definition 1. We say that $\{y_t\}$ is serially independent at lag q if $F_{0q} = F_0 \times F_q$, equivalently if the copula satisfies $C_q(u, v) = uv$ (and $c_q(u, v) = 1$) on $[0, 1]^2$.

We now define the dependence measure and its estimators, and then derive tests from them.

2.1.1 Population AHCF

In this part, we briefly review Geenens and Lafaye de Micheaux's (2022) method and then adapt it to potentially serially dependent time series. We consider a pair of real-valued random variables (X_1, X_2) with continuous marginals and copula density c_{12}. We summarize their dependence by the Bhattacharyya affinity $\mathcal{B}$ between the copula and independence and by the copula-based Hellinger distance $\mathcal{H}^2$, defined as

$$\mathcal{B} := \int_0^1 \int_0^1 \sqrt{c_{12}(u_1, u_2)}\, du_1\, du_2, \qquad \mathcal{H}^2 := 1 - \mathcal{B}. \tag{1}$$

The affinity $\mathcal{B}$ equals one under independence and decreases as dependence strengthens, while $\mathcal{H}^2$ moves in the opposite direction. To obtain a correlation-like measure that equals $|\rho|$ under bivariate Gaussian dependence with correlation ρ, we follow Geenens and Lafaye de Micheaux (2022) and apply a monotone, continuous transformation to $\mathcal{B}$,

$$\eta = g(\mathcal{B}) := \left(\frac{2}{\mathcal{B}^2}\right) \sqrt{\mathcal{B}^4 + \sqrt{4 - 3\mathcal{B}^4} - 2}, \tag{2}$$

which maps independence to $\eta = 0$ and yields a dependence index that behaves like absolute correlation in the Gaussian case. Intuitively, η is near zero when (X_1, X_2) is close to independence and grows as the dependence strengthens. The transformation is undefined at $\mathcal{B} = 0$ (perfect dependence), but $g(\mathcal{B}) \to 1$ as $\mathcal{B} \downarrow 0$ by L'Hôpital's rule, so η extends continuously to one there.

The same construction applies to the lag-q pair (y_t, y_{t-q}). Under strict stationarity the copula of this pair depends only on the lag q, not on t, so we index its density by the lag alone and write c_q, the time series counterpart of the generic c_{12}. We relabel the coordinates (u_1, u_2) as (u_0, u_q), where 0 marks the current component y_t and q the lagged component y_{t-q}. We define

$$\mathcal{B}(q) := \int_0^1 \int_0^1 \sqrt{c_q(u_0, u_q)}\, du_0\, du_q, \quad \text{and} \quad \eta(q) := g\big(\mathcal{B}(q)\big). \tag{3}$$

Definition 2. The function $q \mapsto \eta(q)$ is the Auto-Hellinger Correlation Function (AHCF) of $\{y_t\}$. Under Assumption 1(b), $\eta(q) = 0$ if and only if y_t and y_{t-q} are serially independent at lag q.

Because $\eta(q)$ depends on the entire lag-q copula, it detects nonlinear and nonmonotonic dependence that classical autocorrelation, a function of covariance alone, cannot.

2.1.2 Empirical AHCF

We now describe how we estimate the affinity $\mathcal{B}(q)$, its Legendre coefficients, and the AHCF $\eta(q)$ from data, following Geenens and Lafaye de Micheaux (2022). The Beta transformation is part of the estimator itself, not a numerical refinement, and the consistency theory below applies to the transformed, weighted statistic that we actually compute.

2.1.2.1 Pseudo-observations.

For a fixed lag $q < T$, we set $n := T - q$ and form the n lag-q pairs $(y_{q+1}, y_1), \ldots, (y_T, y_{T-q})$. We map each component to its empirical marginal rank: $\widehat{U}_{t,0} := \widehat{F}_0(y_t)$ and $\widehat{U}_{t,q} := \widehat{F}_q(y_{t-q})$ for $t = q+1, \ldots, T$, where

$$\widehat{F}_0(y) := \frac{1}{n+1} \sum_{s=q+1}^{T} \mathbb{1}\{y_s \le y\}, \qquad \widehat{F}_q(y) := \frac{1}{n+1} \sum_{s=q+1}^{T} \mathbb{1}\{y_{s-q} \le y\},$$

are the (rescaled) empirical CDFs computed from the corresponding components of the n pairs. Under Assumption 1(a) ties occur with probability zero, and the implementation breaks any numerical ties by midranks. By construction, within each coordinate the pseudo-observations $\{\widehat{U}_{t,j}\}_t$ take exactly the grid values $\{1/(n+1), \ldots, n/(n+1)\}$. This step makes the procedure *margin-free*: we work on the unit square and focus on dependence, not on the scale of y_t. We stack $\widehat{\mathbf{U}}_t^{(q)} := (\widehat{U}_{t,0}, \widehat{U}_{t,q})$, and write $\mathbf{U}_t^{(q)} := (F_0(y_t), F_q(y_{t-q}))$ for the (infeasible) population pseudo-observations.

2.1.2.2 Beta transformation.

Pseudo-observations are nearly uniform on $[0,1]^2$, but nearest-neighbor distances computed directly on the unit square are distorted near the boundary. A point close to an edge has fewer neighbors on one side, which inflates its nearest-neighbor distance. Following Geenens and Lafaye de Micheaux (2022), we therefore transform each coordinate by the quantile function of a symmetric Beta law. We fix $a \ge 3$ (the implementation uses $a = 6$, as in Geenens and Lafaye de Micheaux, 2022), and write Ξ_a and ξ_a for the CDF and density

of the Beta(a, a) distribution on $[0, 1]$, so that

$$\xi_a(z) = \frac{z^{a-1}(1-z)^{a-1}}{B(a,a)}, \qquad z \in [0,1],$$

with $B(\cdot,\cdot)$ the Beta function. We define the transformed points

$$\widehat{Z}_{t,j} := \Xi_a^{-1}\big(\widehat{U}_{t,j}\big), \qquad j \in \{0, q\}, \qquad \widehat{\mathbf{Z}}_t := \big(\widehat{Z}_{t,0}, \widehat{Z}_{t,q}\big) \in (0,1)^2. \tag{4}$$

Because ξ_a vanishes at the endpoints, the map Ξ_a^{-1} pulls mass away from the edges of the square. Points with extreme ranks are compressed toward the interior, exactly matching a Beta(a, a) marginal design. If $\widehat{U}_{t,j}$ were exactly uniform, $\widehat{Z}_{t,j}$ would be exactly Beta(a, a)-distributed.

2.1.2.3 Weighted nearest-neighbor distances.

We compute nearest-neighbor distances in the transformed coordinates and weight them by the local Beta intensity:

$$\widehat{D}_t^{(q)} := \min_{s \neq t} \big\|\widehat{\mathbf{Z}}_t - \widehat{\mathbf{Z}}_s\big\|, \qquad \widehat{W}_t^{(q)} := \sqrt{\xi_a\big(\widehat{Z}_{t,0}\big)\,\xi_a\big(\widehat{Z}_{t,q}\big)}, \qquad \widehat{R}_t^{(q)} := \widehat{D}_t^{(q)}\,\widehat{W}_t^{(q)}. \tag{5}$$

The weight $\widehat{W}_t^{(q)}$ undoes, at first order, the density distortion introduced by the transformation. Near the center, where the Beta design places many points, the raw distances $\widehat{D}_t^{(q)}$ are small and the weight is large. Near the edges the converse holds. The product $\widehat{R}_t^{(q)}$ is therefore comparable across the square, and its average estimates the affinity. The (unnormalized) Bhattacharyya affinity estimator is

$$\widehat{\mathcal{B}}(q) := \frac{2\sqrt{n-1}}{n} \sum_{t=q+1}^{T} \widehat{R}_t^{(q)}. \tag{6}$$

We now describe the population mechanism behind Equation (6), which we make rigorous in the Online Supplementary Material. We write $Z_{t,j} := \Xi_a^{-1}(U_{t,j}^{(q)})$ for the transformed population pseudo-observations. A change of variables shows that $\mathbf{Z}_t := (Z_{t,0}, Z_{t,q})$ has density

$$f_{q,a}(z) \;=\; c_q\big(\Xi_a(z_0), \Xi_a(z_q)\big)\,\xi_a(z_0)\,\xi_a(z_q), \qquad z = (z_0, z_q) \in (0,1)^2. \tag{7}$$

For a sample of size n from a planar density f, the nearest-neighbor distance at a point z concentrates at the scale $\{nf(z)\}^{-1/2}$, so that $2\sqrt{n-1}\,\widehat{D}_t^{(q)} \approx f_{q,a}(\mathbf{Z}_t)^{-1/2}$. Multiplying by the weight and using Equation (7),

$$2\sqrt{n-1}\,\widehat{R}_t^{(q)} \;\approx\; \frac{W(\mathbf{Z}_t)}{\sqrt{f_{q,a}(\mathbf{Z}_t)}} \;=\; \frac{1}{\sqrt{c_q\big(\mathbf{U}_t^{(q)}\big)}}, \qquad W(z) := \sqrt{\xi_a(z_0)\xi_a(z_q)}\,:$$

The Beta factors cancel exactly, so that the weighted reciprocal square-root transformed density reduces to the reciprocal square-root copula density. Averaging over t and using

$\mathbf{U}_t^{(q)} \sim c_q$ gives $\mathbb{E}\big[c_q(\mathbf{U}_t)^{-1/2}\big] = \int_{[0,1]^2} \sqrt{c_q(u)}\, du = \mathcal{B}(q)$. Under serial independence, $c_q \equiv 1$, every rescaled weighted radius fluctuates around one, and $\widehat{\mathcal{B}}(q) \approx 1$. Under dependence, the population affinity satisfies $\mathcal{B}(q) < 1$, so the estimator fluctuates around a value below one. This weight cancellation is not merely a heuristic device; it is the key structural identity underlying the consistency theory in Section 2.1.3. It allows the required exponential tail bounds for the weighted radii to be derived without imposing high-level moment conditions on the estimator itself, with constants that are uniform over the unit square. Equivalently, the Beta transformation modifies the local sampling density in a completely explicit way, and dividing by the squared Beta weight exactly removes this induced distortion, leaving only the copula density, which is uniformly bounded by assumption.

2.1.2.4 Legendre coefficients and normalization.

The Bhattacharyya affinity captures only the average of the square-root copula density. To characterize departures from independence more fully and to construct omnibus tests, we expand the square-root copula density in an orthonormal Legendre basis. As we will see, the constant coefficient of this expansion equals the affinity $\widehat{\mathcal{B}}(q)$, while the remaining coefficients describe increasingly complex departures from independence.

To this end, we write $\{b_k(u)\}_{k\geq 0}$ for the shifted, orthonormal Legendre polynomials on $[0,1]$ ($b_0 \equiv 1$, $\int_0^1 b_k b_{k'} = \mathbb{1}\{k = k'\}$, $\|b_k\|_\infty = \sqrt{2k+1}$), and set $b_{k,l}(u_0, u_q) := b_k(u_0) b_l(u_q)$. For integers $K, L \geq 0$, we define the empirical coefficient array

$$\widehat{\beta}_{k,l}(q) := \frac{2\sqrt{n-1}}{n} \sum_{t=q+1}^{T} \widehat{R}_t^{(q)}\, b_k\big(\widehat{U}_{t,0}\big)\, b_l\big(\widehat{U}_{t,q}\big), \qquad 0 \leq k \leq K,\ 0 \leq l \leq L. \tag{8}$$

Since $b_0 \equiv 1$, the constant coefficient $\widehat{\beta}_{0,0}(q) = \widehat{\mathcal{B}}(q)$. The coefficients with $(k,l) \neq (0,0)$ estimate the projections $\beta_{k,l}(q) = \int b_{k,l}\sqrt{c_q}$ of the square root copula density on the non-constant basis directions. Under independence all of them are zero. The normalized affinity estimator and the normalized empirical AHCF are

$$\widehat{\mathcal{B}}_{K,L}(q) := \frac{\widehat{\beta}_{0,0}(q)}{\sqrt{\sum_{k=0}^{K}\sum_{l=0}^{L} \widehat{\beta}_{k,l}(q)^2}}, \qquad \widehat{\eta}_{K,L}(q) := g\big(\widehat{\mathcal{B}}_{K,L}(q)\big), \tag{9}$$

with $g(\cdot)$ as in Equation (2). By the Cauchy–Schwarz inequality in $\mathbb{R}^{(K+1)(L+1)}$, $\widehat{\mathcal{B}}_{K,L}(q) \in (0,1]$ always, so $\widehat{\eta}_{K,L}(q)$ is always well defined. This is the finite-sample stabilization role of the normalization emphasized by Geenens and Lafaye de Micheaux (2022). The normalized estimator targets the finite-order quantity $\mathcal{B}_{K,L}(q) := \mathcal{B}(q)/\sqrt{S_{K,L}(q)}$ rather than the exact affinity, where $S_{K,L}(q) := \sum_{k\leq K, l\leq L} \beta_{k,l}(q)^2$ collects the squared coefficients up to order (K,L). Because this sum contains the constant term $\beta_{0,0}(q)^2 = \mathcal{B}(q)^2$, it is bounded below by $\mathcal{B}(q)^2 > 0$, and $\mathcal{B}_{K,L}(q)$ is well defined. Under independence the nonconstant

coefficients vanish and $\mathcal{B}(q) = 1$, so $S_{K,L}(q) = 1$ and $\mathcal{B}_{K,L}(q) = 1$. The distinction from the exact affinity vanishes as (K, L) grow. For descriptive use of the AHCF, (K, L) may be selected by the leave-one-out cross-validation selector of Geenens and Lafaye de Micheaux (2022). The tests developed in Section 2.2 instead handle the basis order internally, by an order-adaptive construction over a prespecified grid, so that no data-dependent selector enters the validity theory.

We summarize the construction as follows: (i) convert (y_t, y_{t-q}) into ranks so that only dependence remains; (ii) move the rank points into Beta-designed coordinates so that the boundary of the square carries vanishing weight; (iii) measure local concentration through nearest-neighbor distances, compensated by the Beta intensity so that the boundary distortion cancels; (iv) project the compensated distances onto the Legendre basis; and (v) transform the normalized affinity into the correlation-like index $\widehat{\eta}_{K,L}(q)$, using its fixed-order squared values in the tests of Section 2.2.

2.1.3 Consistency theory for the Beta-transformed nearest-neighbor estimator

This section establishes the large-sample behavior used by the tests: consistency of the complete coefficient vector for a fixed finite pair (K, L). We fix a lag $q \geq 1$, set $n := T - q$, and retain the notation of Section 2.1.2: the population pseudo-observations $\mathbf{U}_t^{(q)} = (F_0(y_t), F_q(y_{t-q}))$, the lag-$q$ copula density c_q on $[0,1]^2$, the orthonormal Legendre basis $\{b_k\}_{k\geq 0}$ with $b_{k,l}(u_0, u_q) = b_k(u_0)b_l(u_q)$, the affinity $\mathcal{B}(q) = \int_{[0,1]^2} \sqrt{c_q}$, its Legendre coefficients $\beta_{k,l}(q) = \int_{[0,1]^2} b_{k,l}\sqrt{c_q}$, and the normalizing constant $S_{K,L}(q) = \sum_{k\leq K, l\leq L} \beta_{k,l}(q)^2$.

We order the index set $\mathcal{I}_{K,L} := \{(0,0), (0,1), \ldots, (0,L), (1,0), \ldots, (K,L)\}$ lexicographically and let $p := (K+1)(L+1)$. Since $b_0 \equiv 1$, $\beta_{0,0}(q) = \mathcal{B}(q)$. We collect the coefficients in

$$\theta(q) := \big(\beta_{0,0}(q), \beta_{0,1}(q), \ldots, \beta_{K,L}(q)\big)^\top \in \mathbb{R}^p,$$

with feasible sample analogue $\widehat{\theta}(q)$ from Equation (8), and let e_{00} select the constant-basis coordinate. We write $\theta_-(q) \in \mathbb{R}^{p-1}$ for the subvector of *nonconstant* coefficients, that is, all coordinates except $\beta_{0,0}(q)$, and define the sum of squared nonconstant coefficients

$$\Lambda_{K,L}(q) := \|\theta_-(q)\|^2 = S_{K,L}(q) - \beta_{0,0}(q)^2, \quad \widehat{\Lambda}_{K,L}(q) := \|\widehat{\theta}_-(q)\|^2 = \widehat{S}_{K,L}(q) - \widehat{\beta}_{0,0}(q)^2, \tag{10}$$

where $\widehat{S}_{K,L}(q) := \sum_{k\leq K, l\leq L} \widehat{\beta}_{k,l}(q)^2$. Because $S_{K,L}(q)$ includes the constant term $\beta_{0,0}(q)^2 = \mathcal{B}(q)^2$, subtracting it leaves the sum of squared nonconstant coefficients. Equivalently, $S_{K,L}(q) = \mathcal{B}(q)^2 + \Lambda_{K,L}(q)$. The quantity $\Lambda_{K,L}(q)$ provides a convenient identification device for the tests of Section 2.2. It is zero under independence and strictly positive precisely when some included nonconstant coefficient $\beta_{k,l}(q)$ is nonzero, equivalently when the corresponding finite-order AHCF is positive.

The consistency theory requires, in addition to the baseline conditions of Assumption 1, the following primitive geometric regularity conditions.

Assumption 2. For the fixed lag q under consideration:

(i) *(Geometric absolute regularity.)* The process $\{y_t\}$ is strictly stationary and β-mixing (absolutely regular). With $\mathcal{F}_s^t := \sigma(y_s, \dots, y_t)$ and time gap $j \geq 1$, the strong-mixing (α) and absolute-regularity (β) coefficients are

$$\alpha(j) := \sup_{A \in \mathcal{F}_{-\infty}^0,\, B \in \mathcal{F}_j^\infty} \big|\Pr(A \cap B) - \Pr(A)\Pr(B)\big|,$$
$$\beta(j) := \mathbb{E} \sup_{B \in \mathcal{F}_j^\infty} \big|\Pr(B \mid \mathcal{F}_{-\infty}^0) - \Pr(B)\big|,$$

and satisfy $\alpha(j) \leq \beta(j)$. We assume $\beta(j) \leq C_\beta \rho^j$ for some $C_\beta < \infty$ and $\rho \in (0,1)$.

(ii) *(Copula density bounds and smoothness.)* The lag-q copula density satisfies $0 < \underline{c} \leq c_q(u) \leq \bar{c} < \infty$ for all $u \in [0,1]^2$ and is globally Lipschitz on $[0,1]^2$. Moreover, $\sup_{h \geq 1} \|c_h\|_\infty \leq \bar{c}_2 < \infty$, and the joint densities of any collection of up to six probability integral transforms $F_{j_1}(y_{t_1}), \dots, F_{j_r}(y_{t_r})$ at pairwise distinct time indices (after removing lag-induced coordinate coincidences) are uniformly bounded by a constant $\bar{c}_* < \infty$.

Intuitively, Assumption 2 guarantees that the transformed observations remain sufficiently well spread over the unit square. The bounded copula densities prevent extremely concentrated or vanishing local probability regions, while the bounded low-dimensional joint PIT densities prevent several nearby observations from clustering together in pathological ways. These conditions ensure that nearest-neighbor distances behave regularly. Importantly, we do not assume moment bounds or stabilization properties for the weighted nearest-neighbor estimator itself; instead, these are derived in the Online Supplementary Material from the structural identity relating the transformed density, the Beta weights, and the copula density.

Proposition 1. *Let Assumptions 1(a) and 2 hold, and fix $K, L \geq 0$ and $a \geq 3$. Then, as $T \to \infty$, $\widehat{\beta}_{k,l}(q) \xrightarrow{p} \beta_{k,l}(q)$ for all $0 \leq k \leq K,\ 0 \leq l \leq L$, and consequently*

$$\widehat{S}_{K,L}(q) \xrightarrow{p} S_{K,L}(q), \quad \widehat{\Lambda}_{K,L}(q) \xrightarrow{p} \Lambda_{K,L}(q),$$
$$\widehat{\mathcal{B}}_{K,L}(q) \xrightarrow{p} \mathcal{B}_{K,L}(q), \quad \widehat{\eta}_{K,L}(q) \xrightarrow{p} \eta_{K,L}(q) := g\big(\mathcal{B}_{K,L}(q)\big).$$

We give the complete proof in Section A of the Online Supplementary Material and summarize only its main ideas here. The argument proceeds in five steps. (1) We first identify the density of the Beta-transformed observations using a change of variables. This provides the local probability approximation needed for nearest-neighbor methods. (2) We

then study an oracle version of the estimator based on the population PITs. The geometric β-mixing condition allows observations that are sufficiently far apart to behave almost independently, while the Beta weights exactly cancel the density distortion introduced by the transformation. As a result, the weighted nearest-neighbor radius estimates the reciprocal square-root copula density. (3) Averaging these local estimates yields the Legendre coefficient estimator. Its variability vanishes because temporal dependence is controlled by the mixing condition and the geometric fact that, in the plane, one point can be the nearest neighbor of at most six other points. (4) We next show that replacing the unobservable population PITs by empirical ranks changes the estimator only by an asymptotically negligible amount. (5) Finally, we show that observations near the boundary of the unit square have a negligible contribution, using a deterministic geometric argument together with the behavior of the Beta weights near the boundary.

In the next lemma, we collect several population characterizations of serial independence that have been introduced separately throughout the preceding discussion. The lemma establishes their formal equivalence for later reference and identifies the alternatives that the fixed-order AHCF statistics can detect. Its proof is given in the Online Supplementary Material.

Lemma 2. *Let Assumption 1(a)–(b) hold and write $c_{\min,q} > 0$ for a lower bound of c_q on $[0,1]^2$. Then:*

(a) $\sqrt{c_{\min,q}} \le \mathcal{B}(q) \le 1$, *with* $\mathcal{B}(q) = 1$ *if and only if* $c_q \equiv 1$, *that is, if and only if* y_t *and* y_{t-q} *are independent; and* $\eta(q) = 0$ *if and only if* $\mathcal{B}(q) = 1$.

(b) For any fixed $K, L \ge 0$: $\mathcal{B}_{K,L}(q) \in (0,1]$, *and the following are equivalent:* $\eta_{K,L}(q) > 0$; $\mathcal{B}_{K,L}(q) < 1$; $\Lambda_{K,L}(q) > 0$; $\beta_{k,l}(q) \neq 0$ *for some* $(k,l) \in \mathcal{I}_{K,L} \setminus \{(0,0)\}$.

(c) Under serial independence at lag q, $\theta(q) = e_{00}$ *and* $\Lambda_{K,L}(q) = 0$ *for every* K, L.

The next remark relates the proposed AHCF statistic to a more familiar quadratic-form statistic. It shows that, near the null hypothesis, the two are asymptotically equivalent, providing an interpretation of the local behavior of the proposed test.

Remark 1. *Fix* $(K,L) = (m,m)$ *and write* $\widehat{\theta} = \widehat{\theta}(q)$. *Since* $\widehat{\beta}_{0,0}(q) = \widehat{B}(q)$, *a simple algebraic identity gives*

$$1 - \widehat{B}_{m,m}(q) = \frac{\widehat{\Lambda}_{m,m}(q)}{\|\widehat{\theta}\| \left(\|\widehat{\theta}\| + \widehat{\beta}_{0,0}(q) \right)},$$

where $\widehat{\Lambda}_{m,m}(q) = \|\widehat{\theta}\|^2 - \widehat{\beta}_{0,0}(q)^2$. *Under any data-generating process for which* $\widehat{\theta}(q) \xrightarrow{p} \theta_\infty$, *the denominator converges to* 2 *under the null hypothesis. Hence,*

$$1 - \widehat{B}_{m,m}(q) = \frac{1}{2} \widehat{\Lambda}_{m,m}(q) \{1 + o_p(1)\}.$$

Using the expansion

$$g(b)^2 = 8(1-b) + O\big((1-b)^2\big), \qquad b \to 1,$$

gives

$$T_{\widehat{\eta}_{m,m}}(q)^2 = 4\,\widehat{\Lambda}_{m,m}(q)\{1 + o_p(1)\}.$$

Thus, close to independence, the squared AHCF statistic is asymptotically proportional to the sum of squared nonconstant Legendre coefficients. Nonetheless, the equivalence holds only in a neighborhood of the null hypothesis and should not be interpreted as saying that the two statistics behave identically under all forms of dependence. Motivated by this local equivalence, the test statistics introduced in Section 2.2 are scaled by the sample size T, matching the conventional normalization of Neyman smooth tests. This normalization facilitates comparison with the classical quadratic-form statistic while having no effect on the Monte Carlo p-values or on any rejection decision, since the same positive factor multiplies every statistic.

Although the squared AHCF statistic is asymptotically proportional to the corresponding quadratic-form statistic near the null hypothesis, this relationship does not determine its null distribution. Consequently, the proposed tests rely on the rank-based Monte Carlo calibration developed in the next section.

Remark 2. *One natural question is whether the estimated coefficient vector satisfies a root-n central limit theorem. We do not pursue such a result because it is not required for the proposed tests. Establishing a central limit theorem would require a joint analysis of nearest-neighbor boundary bias, the sample-dependent nearest-neighbor geometry, and the replacement of population PITs by empirical ranks. Such an analysis is beyond the scope of the present paper. Instead, the tests in Section 2.2 are calibrated by rank-based Monte Carlo, whose validity and consistency do not depend on a central limit theorem for the estimated coefficient vector.*

2.2 Rank-based tests for serial independence

This subsection develops the serial-independence tests. The construction follows three simple principles established in the previous section. First, the component statistic is the squared fixed-order AHCF, $T_{\widehat{\eta}_{m,m}}(q)^2$. As explained in Remark 1, this statistic has the same local behavior under independence as the corresponding quadratic-form statistic while retaining the interpretation of the proposed procedure in terms of the AHCF. Second, rather than relying on an asymptotic null distribution, we calibrate each statistic by rank-based Monte Carlo. Because the statistics depend only on the ranks, this calibration is exactly valid in finite samples under the i.i.d. null hypothesis with continuous marginal distribu-

tions. Finally, instead of selecting a single basis order from the data, we retain all candidate orders in a prespecified grid and combine their Monte Carlo p-values using a minimum-p rule. The final calibration automatically accounts for this multiple comparison step, so the validity of the test is preserved [1].

Choosing the order of a smooth test is a classical problem. Rather than selecting a single order from the data (Ledwina, 1994), we retain all candidate orders and combine their Monte Carlo p-values using the minimum-p principle of Westfall and Young (1993). Since the statistics are functions of ranks, the resulting Monte Carlo p-values are exactly valid in finite samples under the i.i.d. null hypothesis with continuous marginal distributions (Dwass, 1957; Hemerik and Goeman, 2018). Converting the statistics to p-values before combining them is an application of the prepivoting idea of Beran (1988). This strategy differs from related adaptive procedures in econometrics, which typically maximize over smoothing parameters or basis orders and rely on bootstrap calibration (Fan, 1996; Horowitz and Spokoiny, 2001).

The proposed tests address pairwise serial independence at lag q. Exact finite-sample critical values are obtained by Monte Carlo calibration under the stronger i.i.d. null, under which the distribution of the rank-based statistic is exactly known.

Definition 3. $H_0^{\text{pair}}(q)$: y_t and y_{t-q} are independent, equivalently $c_q \equiv 1$.

Definition 4. H_0^{iid}: $y_1, \ldots, y_T$ are independent and identically distributed with some continuous marginal distribution function.

The testing procedure involves two null hypotheses, each serving a different purpose. The hypothesis of scientific interest is $H_0^{\text{pair}}(q)$, which states that (y_t, y_{t-q}) is independent at lag q. For the portmanteau tests, the null hypothesis states that pairwise independence holds at every lag in the prescribed range. The stronger hypothesis H_0^{iid} is only used to calibrate the Monte Carlo procedure. Under H_0^{iid}, every ordering of the ranks is equally likely. Consequently, the null distribution of any rank-based statistic is exactly determined, making the Monte Carlo calibration exactly valid in finite samples. Since i.i.d. implies pairwise independence at every lag, but not conversely, H_0^{iid} provides a convenient calibration framework, whereas $H_0^{\text{pair}}(q)$ specifies the serial dependence that each component test is designed to detect.

The validity results show that the Monte Carlo test has exact finite-sample size under H_0^{iid}. The consistency results show that, under any fixed stationary alternative satisfying Assumptions 1 and 2, the rejection probability converges to one whenever the chosen basis captures the pairwise dependence.

[1] For example, if the candidate orders are $m = 1, \ldots, 4$, the procedure computes $T_{\widehat{\eta}_{1,1}}(q)^2, \ldots, T_{\widehat{\eta}_{4,4}}(q)^2$, obtains a Monte Carlo p-value for each statistic, and bases the final decision on the Monte Carlo-calibrated minimum of these four p-values.

Consequently, the proposed tests are sensitive to a broad class of linear and nonlinear pairwise alternatives, including dependence in conditional variance. Like any procedure based on lagged pairs, however, they are not designed to detect dependence that arises only through higher-order interactions while every pair (y_t, y_{t-q}) remains independent.

2.2.1 Component statistics and order adaptation at a fixed lag

We consider a finite collection of square basis orders $\mathcal{M} = \{m_1, \ldots, m_J\} \subset \mathbb{N}$, specified before observing the data (our implementation uses $\mathcal{M} = \{1, \ldots, 10\}$). For each lag q and basis order $m \in \mathcal{M}$, we compute the *component statistic*

$$Z_{m,q} := T\, \widehat{\eta}_{m,m}(q)^2,$$

where $\widehat{\eta}_{m,m}(q) = g\{\widehat{\mathcal{B}}_{m,m}(q)\}$ is constructed from the Beta-transformed rank-based coefficients in Equation (8). We *emphasize* that the component statistic $Z_{m,q}$ *is not the same object* as the transformed points $\hat{Z}_t$ in Equation (4) in Subsection 2.1.2.2. As mentioned in Remark 1, the scaling with sample size T is done to place $Z_{m,q}$ on the Neyman smooth scale and to give the rate-T divergence used in the consistency proofs. Large values of $Z_{m,q}$ indicate stronger evidence against pairwise independence at lag q captured by basis functions up to order m.

From the coefficient array computed at the largest basis order $\max \mathcal{M}$, all lower-order normalized affinities and AHCF statistics are obtained by cumulative sums of squared coefficients. Consequently, evaluating multiple basis orders requires little additional computation. Furthermore, because the proposed procedure is distribution-free, the reference samples may be generated from any continuous distribution. Throughout the paper we use the standard normal distribution $\mathcal{N}(0,1)$, although any continuous reference distribution would give the same null distribution.

Specifically, we first generate B_0 independent reference series of length T from the reference distribution F_{ref}. Applying exactly the same procedure used for the observed data to each reference series yields the statistics

$$Z_{m,q}^{(1,0)}, \ldots, Z_{m,q}^{(B_0,0)}.$$

These reference statistics are then used to compute the Monte Carlo p-value for each basis order,

$$\widehat{p}_{m,q} = \frac{1 + \#\{b \le B_0 :\, Z_{m,q}^{(b,0)} \ge Z_{m,q}\}}{B_0 + 1}. \tag{11}$$

Since the basis order that best captures the lag-q dependence is unknown, we compute the component test for every order in $\mathcal{M}$ and retain the smallest component p-value,

$$R_q = \min_{m \in \mathcal{M}} \widehat{p}_{m,q}. \tag{12}$$

However, R_q is not itself a valid p-value.[2] Moreover, since every component p-value is computed from the same observed data and the same reference series, the component p-values are generally dependent.

To calibrate this minimum-p selection rule, we generate a further B_1 independent series from F_{ref}. Each of these *calibration series* is treated exactly like the observed data: we compute its component statistics, obtain its component p-values using the same B_0 reference series, and record the resulting minimum component p-value, denoted by $R_q^{(b,1)}$. Repeating this procedure for all B_1 calibration series produces the null distribution of R_q. The order-adaptive Monte Carlo p-value is then

$$\widehat{p}_q^{\,\text{ord}} = \frac{1 + \#\{b \le B_1 : R_q^{(b,1)} \le R_q\}}{B_1 + 1}, \tag{13}$$

and the level-α order-adaptive test rejects whenever $\widehat{p}_q^{\,\text{ord}} \le \alpha$.

In what follows (see Table 1) we provide a summary of the Monte Carlo algorithm employed in the paper to ease understanding.

Table 1: Monte Carlo implementation of the order-adaptive AHCF test at lag q.

Step	Operation	Output
1	Choose the basis-order grid $\mathcal{M} = \{1, \dots, 10\}$. Compute the AHCF statistic $Z_{m,q} = T\widehat{\eta}_{m,m}(q)^2$ for every basis order.	$Z_{m,q}$
2	Generate B_0 i.i.d. reference series under the null hypothesis. Apply exactly the same estimation procedure to each series.	$Z_{m,q}^{(b,0)}$
3	Compare the observed statistic with the reference statistics to obtain a Monte Carlo component p-value for every basis order.	$\widehat{p}_{m,q}$
4	Adapt over basis orders by selecting the smallest component p-value.	$R_q = \min_{m \in \mathcal{M}} \widehat{p}_{m,q}$
5	Generate another B_1 independent calibration series. For each one, repeat Steps 1–4 using the same B_0 reference samples.	$R_q^{(b,1)}$
6	Compare the observed minimum p-value with the calibration minima to obtain the final adaptive Monte Carlo p-value. Reject if it is below α.	$\widehat{p}_q^{\,\text{ord}}$

With a second table (See Table 2) we also clarify further what B_0 and B_1 in the algorithm stands for.

The proposed testing procedure has two fundamental properties. First, it provides exact finite-sample size control under the i.i.d. null hypothesis, regardless of the choice of reference distribution or the Monte Carlo budgets. Second, it is consistent against detectable

[2] If only a single basis order were considered, the corresponding component p-value would be approximately Uniform$(0, 1)$ under the null hypothesis. In the proposed procedure, however, several basis orders are examined and the smallest component p-value is retained. Since the selection rule favors unusually small values, the distribution of the minimum is no longer Uniform$(0, 1)$, even when every component test is individually valid. This is why the minimum component p-value must be calibrated separately.

Table 2: Roles of the Monte Carlo samples.

Quantity	Purpose	Used to compute
Observed sample	Actual data	$Z_{m,q}$
B_0 reference series	Null reference distribution	$\widehat{p}_{m,q}$
B_1 calibration series	Distribution of the minimum p-value	$\widehat{p}_q^{\,\mathrm{ord}}$

alternatives, so its power converges to one as the sample size increases. The next two propositions formalize these two properties. We prove them in Section A of the Online Supplementary Material. The first proposition is finite-sample and exact: it holds for every sample size and every choice of B_0, B_1, $\mathcal{M}$, and a, with no assumptions beyond the i.i.d. null. The second proposition establishes asymptotic consistency under a detectable alternative.

Proposition 3. *(a) Every statistic in the procedure*

$$\widehat{\beta}_{k,l}(q),\ Z_{m,q},\ \widehat{p}_{m,q},\ R_q,\ \text{and}\ \widehat{p}_q^{\,\mathrm{ord}}$$

is a measurable function of the rank vectors of the observed and simulated samples. Consequently, under H_0^{iid}, the joint distribution of these statistics depends only on the ranks and is therefore identical for every continuous marginal distribution. Hence the particular choice of the reference distribution F_{ref} does not affect the null distribution of the test.

(b) Under H_0^{iid}, for any T, B_0, B_1, $\mathcal{M}$, $a \geq 3$, and $\alpha \in (0,1)$,

$$\Pr\left(\widehat{p}_q^{\,\mathrm{ord}} \leq \alpha\right) \leq \alpha.$$

The first proposition establishes two properties. Part (a) shows that the entire testing procedure depends only on the ranks of the observed and simulated samples, so the particular continuous reference distribution used in the Monte Carlo simulation does not affect the test. Part (b) then uses this rank invariance to establish exact finite-sample size control under the i.i.d. null hypothesis.

The second proposition addresses power. It shows that, under a detectable alternative, the order-adaptive procedure is asymptotically consistent. The result covers both increasing Monte Carlo budgets and the practically relevant case in which the simulation budgets remain fixed.

Proposition 4. *Let the data satisfy Assumptions 1(a) and 2 at lag q, and suppose $\eta_{m^*,m^*}(q) > 0$, equivalently $\Lambda_{m^*,m^*}(q) > 0$, for some $m^* \in \mathcal{M}$. Suppose first that $B_0 = B_0(T) \to \infty$[3] and that $B_1 = B_1(T)$ (fixed or growing) satisfy*

$$\frac{J\,B_1}{B_0+1} \longrightarrow 0, \qquad \frac{1}{B_1+1} \leq \alpha \ \ \textit{for all large } T,$$

[3]Throughout the asymptotic analysis, the Monte Carlo budgets are allowed to depend on the sample size. Thus $B_0(T)$ and $B_1(T)$ may either increase with T or remain fixed at the values used in practice.

and assume that the largest reference value of $\widehat{\eta}_{m,m}(q)^2$ *over all draws and orders,* $\Delta_{0,T}^{\eta} := \max_{b \le B_0} \max_{m \in \mathcal{M}} \widehat{\eta}_{m,m}^{(b,0)}(q)^2$*, converges to zero in probability. Then*

$$\Pr\left(\widehat{p}_q^{\,\mathrm{ord}} \le \alpha\right) \longrightarrow 1.$$

If instead the budgets (B_0, B_1) *are held fixed with* $1/(B_1+1) \le \alpha$*, then* $\Delta_{0,T}^{\eta} \xrightarrow{p} 0$ *holds automatically and*

$$\liminf_{T\to\infty} \Pr\left(\widehat{p}_q^{\,\mathrm{ord}} \le \alpha\right) \;\ge\; 1 - \frac{J\,B_1}{(B_0+1)\,\lfloor \alpha(B_1+1)\rfloor},$$

which for $(J, B_0, B_1, \alpha) = (10, 50{,}000, 2{,}000, 0.05)$ *is approximately* 0.996 *as a worst-case guarantee. The bound follows from Markov's inequality and is consequently conservative.*

The intuition behind Proposition 4 is simple. Under the alternative, the statistic $Z_{m^*,q}$ computed from the observed sample grows proportionally to T. In contrast, the statistics $Z_{m,q}^{b,0}$ computed from the B_0 simulated reference series remain asymptotically negligible[4]. Consequently, the observed statistic eventually exceeds all reference statistics, so that

$$\widehat{p}_{m^*,q} = \frac{1}{B_0+1} \quad \text{and} \quad R_q \le \frac{1}{B_0+1}$$

with probability approaching one. Exchangeability then implies that the expected number of calibration series satisfying $R_q^{(b,1)} \le R_q$ is at most $JB_1/(B_0+1)$, which converges to zero by assumption. Hence

$$\widehat{p}_q^{\,\mathrm{ord}} \xrightarrow{p} \frac{1}{B_1+1} \le \alpha.$$

The conditions $1/(B_1+1) \le \alpha$ and $JB_1/(B_0+1) \to 0$ therefore ensure consistency. When $B_0 \to \infty$, the smallest attainable component p-value converges to zero, yielding full consistency. If the Monte Carlo budgets are held fixed instead, the smallest attainable component p-value remains positive, giving the explicit lower bound in Proposition 4. No asymptotic distributional approximation is required at any stage.

In short, under the alternative, the observed sample produces a test statistic that eventually exceeds every statistic computed from the B_0 null reference samples. This forces the observed component p-value to attain its minimum possible value. Since almost no calibration sample produces an equally extreme minimum p-value, the adaptive Monte Carlo p-value also converges to its minimum possible value, leading to rejection with probability approaching one.

[4]Since $Z_{m,q}^{(b,0)} = T\,\widehat{\eta}_{m,m}^{(b,0)}(q)^2$, the condition $\Delta_{0,T}^{\eta} \xrightarrow{p} 0$ is equivalent to

$$\max_{b \le B_0} \max_{m \in \mathcal{M}} \frac{Z_{m,q}^{(b,0)}}{T} \xrightarrow{p} 0.$$

Thus, after scaling by T, even the largest statistic computed from the simulated reference samples converges to zero.

2.2.2 Portmanteau tests over a fixed lag range

The preceding procedure tests serial independence at a given lag q. In many applications, however, the relevant lag is not known in advance. We therefore extend the test to a fixed lag range $\mathcal{Q} = \{1, \ldots, q_{\max}\}$. The component statistics are first aggregated across lags for each basis order, after which the same minimum-p adaptation over basis orders is applied. Aggregating before order adaptation produces a single adaptive test for the entire lag range, rather than a separate adaptive calibration at each lag.

Definition 5. A Borel measurable function

$$A : [0, \infty)^{q_{\max}} \to [0, \infty)$$

is called a *valid aggregation* if

1. it is coordinatewise nondecreasing; and
2. for every $q \in \mathcal{Q}$,
$$A(0, \ldots, 0, x, 0, \ldots, 0) \longrightarrow \infty \qquad \text{as } x \to \infty,$$
where x occupies the qth coordinate.

Examples of valid aggregations include $A_{\mathrm{BP}}(x) = \sum_{q \in \mathcal{Q}} x_q$, the Box-Pierce-type aggregation, $A_{\mathrm{LB}}(x) = \sum_{q \in \mathcal{Q}} \frac{T+2}{T-q} x_q$, the Ljung-Box-type aggregation, and $A_{\mathrm{MAX}}(x) = \max_{q \in \mathcal{Q}} x_q$. The simulation study and empirical application use $A = A_{\mathrm{MAX}}$, which is particularly well suited to alternatives concentrated at a small number of unknown lags.

For each $m \in \mathcal{M}$, we define the aggregated component statistic and run the identical construction:

$$Q_m := A\big(Z_{m,1}, \ldots, Z_{m,q_{\max}}\big), \qquad \widehat{p}_m := \frac{1 + \#\{b : Q_m^{(b,0)} \geq Q_m\}}{B_0 + 1}, \tag{14}$$

$$R := \min_{m \in \mathcal{M}} \widehat{p}_m, \qquad \widehat{p}^{\,\mathrm{port}} := \frac{1 + \#\{b : R^{(b,1)} \leq R\}}{B_1 + 1}, \tag{15}$$

where $p^{\,\mathrm{port}}$ is the final p-value of the portmanteau procedure.

Proposition 5. *(a) Let A be any Borel measurable function fixed before seeing the data. Under H_0^{iid}, for every T, B_0, B_1, $\mathcal{M}$, and $\alpha \in (0, 1)$,*

$$\Pr\left(\widehat{p}^{\,\mathrm{port}} \leq \alpha\right) \;\leq\; \alpha,$$

and, under H_0^{iid}, the procedure is distribution-free over continuous marginal distributions.

(b) Let A be a valid aggregation (Definition 5), fixed in advance, that also satisfies the comparability bounds

$$\gamma_1 \max_q x_q \;\leq\; A(x) \;\leq\; \gamma_2 \sum_q x_q, \qquad 0 < \gamma_1 \leq \gamma_2 < \infty,$$

with constants independent of T. These hold for A_{BP}, A_{LB}, and A_{MAX}, as verified in the Online Supplementary Material. Suppose the data satisfy Assumptions 1(a) and 2 at every $q \in \mathcal{Q}$, and that $\eta_{m^*,m^*}(q^*) > 0$, equivalently $\Lambda_{m^*,m^*}(q^*) > 0$, for some $m^* \in \mathcal{M}$ and $q^* \in \mathcal{Q}$. Then, under the conditions on B_0 and B_1 of Proposition 4 applied to the lagwise reference statistics over $\mathcal{Q}$,

$$\Pr\left(\widehat{p}^{\,\mathrm{port}} \le \alpha\right) \;\to\; 1,$$

and the worst-case guarantee of Proposition 4 carries over.

Part (b) extends the consistency result from the individual-lag test to the aggregated statistic. The comparability bounds ensure that a signal at any lag is preserved by the aggregation, while the reference statistics remain of smaller order under the null. Consequently, the same argument as in Proposition 4 establishes consistency of the portmanteau test.

The basis grid $\mathcal{M}$, lag range $\mathcal{Q}$, and aggregation rule A determine the sensitivity of the test to different forms of serial dependence. For example, sum-based aggregations are more effective when dependence is spread across many lags, whereas the maximum is better suited to dependence concentrated at a small number of lags. Regardless of these choices, finite-sample validity under H_0^{iid} is preserved, provided they are fixed before observing the data.

Remark 3. *Because the proposed statistics depend only on the sample ranks, the Monte Carlo reference distributions depend only on the sample size T. Consequently, for a given T, the same reference and calibration draws can be generated once and reused across different data sets, substantially reducing computational cost. The implementation details used in the simulations and empirical application are reported in Section 3.*

2.2.3 Competing methods

The comparison uses two benchmark procedures implemented as proposed in the original papers. The first is the auto-distance correlation function (ADCF) test of Fokianos and Pitsillou (2017). At a known lag q, the statistic is $(T-q)\widehat{R}^2(q)$, where $\widehat{R}^2(q)$ is the biased sample squared auto-distance correlation. For an unknown lag, we use their Bartlett-kernel portmanteau statistic

$$T_{\mathrm{ADCF}} = \sum_{q=1}^{T-1} (T-q)k^2(q/p)\widehat{R}^2(q), \qquad p = \left\lceil 3T^{0.1} \right\rceil,$$

with Gaussian wild-bootstrap calibration and 499 bootstrap draws. This preserves the statistic, kernel, bandwidth rule, and calibration proposed for serial-independence testing in the ADCF literature.

The second benchmark is the kernel Hellinger dependence statistic S_ρ of Granger et al. (2004). We use $S_\rho(q)$ for the known-lag experiments and the cumulative statistic $\sum_{q=1}^{5} S_\rho(q)$ for the unknown-lag experiments. Inference is based on 99 whole-series random permutations, with bandwidths selected once from the observed series and held fixed throughout the permutation procedure. Complete implementation details, including the bandwidth selection and numerical integration scheme, are provided in Section B of the Online Supplementary Material.

The comparison is restricted to omnibus tests designed to detect lag-specific pairwise serial dependence. Accordingly, ADCF (Fokianos and Pitsillou, 2017) and the kernel Hellinger statistic S_ρ (Granger et al., 2004) serve as the primary benchmarks. The Ljung-Box test is reported only as a linear benchmark, Kendall's tau is omitted because it has no standard portmanteau extension, and the BDS test (Brock et al., 1996) is excluded because it targets dependence in embedded vectors rather than lag-specific dependence.

3 Simulation Studies

3.1 Design and calibration

The simulation study addresses two distinct questions. The size experiments examine whether the proposed procedure controls type-I error under serial independence across a wide range of continuous marginal distributions. The power experiments instead fix the innovation distribution and vary the mechanism through which serial dependence is introduced. This separation reflects the two objectives of the methodology: distribution-freeness under the null and sensitivity to different forms of serial dependence under the alternative.

The AHCF procedures use shifted Legendre polynomial order $a = 6$, $\mathcal{M} = \{1, \ldots, 10\}$, $B_0 = 50{,}000$, and $B_1 = 2{,}000$. For a known lag, adaptation is performed over $m \in \mathcal{M}$ using the minimum-p rule. For the unknown-lag procedure, the lag-specific statistics are first aggregated by $\max_{1 \le q \le 5} Z_{m,q}$, followed by the same minimum-p adaptation across m.

Unless otherwise stated, all experiments are based on 5,000 Monte Carlo replications. ADCF uses 499 Gaussian wild-bootstrap draws, whereas the more computationally demanding S_ρ procedure uses 99 whole-series permutations. Size is evaluated at $T \in \{126, 1008\}$ under twelve continuous i.i.d. marginal distributions, considering both known-lag tests ($q = 1, 2, 3$) and the corresponding unknown-lag procedures. The two sample sizes represent short- and long-sample settings.

3.1.1 Null marginal distributions

Under the size null, $y_t = X_t$, where $\{X_t\}$ is an i.i.d. sequence. The twelve continuous marginal distributions in Table 3 span a broad range of distributional shapes while remaining standardized to have population mean zero and variance one. Standardization removes scale differences across designs, although it does not affect the ranks on which AHCF is based. The size experiment therefore examines the claimed marginal distribution-freeness of AHCF and the finite-sample calibration of the competing procedures under diverse continuous marginals.

Table 3: Continuous marginal distributions used in the size exercise

Law	Standardized draw X	Feature isolated
Gaussian	Z	Reference calibration law
Chi-square	$(C-3)/\sqrt{6}$, $C \sim \chi^2_3$	Moderate right skewness
Laplace	$L/\sqrt{2}$, $L \sim \text{Laplace}(0,1)$	Symmetric exponential tails
Lognormal	$\{\exp(Z) - \exp(1/2)\}/\sqrt{(e-1)e}$	Strong right skewness and tail
Logistic	$G/(\pi/\sqrt{3})$, $G \sim \text{Logistic}(0,1)$	Symmetric non-Gaussian tails
Contaminated normal	$\{(1-B)Z+B(15Z')\}/\sqrt{5.48}$, $B \sim \text{Bernoulli}(0.02)$	Rare, extreme observations
Product-normal, $r=0$	Z_1Z_2, $\text{Corr}(Z_1,Z_2)=0$	Symmetric leptokurtosis
Product-normal, $r=0.6$	$\{Z_1Z_2-0.6\}/\sqrt{1.36}$, $\text{Corr}(Z_1,Z_2)=0.6$	Asymmetric product mixture
Student t (three laws)	$\sqrt{(\nu-2)/\nu}\,W_\nu$, $W_\nu \sim t_\nu$, $\nu \in \{3,5,7\}$	Decreasing tail thickness
Symmetric variance-gamma	$\sqrt{V}Z$, $V \sim \Gamma(1.5, 2/3)$	Continuous variance mixture

Notes: All latent variables in a row are mutually independent unless a correlation is stated. Here Z, Z', Z_1, Z_2 are standard normal. The gamma distribution is parameterized by shape and scale. The Student row represents three distinct marginal distributions, giving twelve distributions in total.

The twelve distributions include symmetric and asymmetric laws, heavy- and light-tailed families, contaminated distributions, and variance-mixture models commonly used in statistics, economics, and finance (Cont, 2001; Huber, 1964; Madan et al., 1998; Mandelbrot, 1963; Nadarajah and Pogány, 2016). Their main features are summarized in the final column of Table 3.

3.1.2 Alternative DGPs Size Adjusted Power Exercise

Power is evaluated at $T \in \{126, 1008\}$, with dependence introduced at lag $q = 1$,[5] over $\rho \in \{0.05, 0.10, \ldots, 1.00\}$. The DGPs represent the principal forms of nonlinear serial dependence encountered in econometrics, with each design isolating a different dependence mechanism (Ashley et al., 1986; Patterson and Ashley, 2000).

Let $e_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $\phi = 0.95\rho$, $\mu_{|Z|} = \sqrt{2/\pi}$, and $m_0 = 1/\sqrt{2}$. The six recursive designs

[5]The analogous designs with dependence at lags $q = 2$ and $q = 3$ produce qualitatively similar rankings and are available upon request.

are

$$
\begin{aligned}
\text{AR:} \quad & y_t = \phi y_{t-1} + \sqrt{1-\phi^2}\, e_t, \\
\text{Bilinear:} \quad & y_t = e_t + \rho y_{t-1} e_{t-1}, \\
\text{Logistic:} \quad & y_t = \sqrt{1-\phi^2}\, e_t + \phi\, 3.3541 \left[4\Phi(y_{t-1})\{1-\Phi(y_{t-1})\} - \frac{2}{3} \right], \\
\text{SignSwitch-AR:} \quad & y_t = \phi s_t y_{t-1} + \sqrt{1-\phi^2}\, e_t, \qquad \Pr(s_t = 1) = \Pr(s_t = -1) = \frac{1}{2}, \\
\text{AbsV:} \quad & y_t = 0.8\phi\big(|y_{t-1}| - \mu_{|Z|}\big) + \sqrt{1-\phi^2}\, e_t, \\
\text{CenterVol:} \quad & y_t = \left[\max\left\{ 10^{-8},\, 1 + 2\phi \left\{ \exp(-y_{t-1}^2/2) - m_0 \right\} \right\} \right]^{1/2} e_t.
\end{aligned}
$$

Each recursive process is generated after a burn-in of 500 observations and standardized over the retained sample. We set $\phi = 0.95\rho$ to keep autoregressive-type recursions away from the unit-root boundary while preserving a common dependence scale across DGPs. The constants in the Logistic, AbsV, and CenterVol designs fix the scale of the nonlinear component and were chosen a priori.

The recursive DGPs isolate different forms of serial dependence. AR represents purely linear dependence, Bilinear introduces multiplicative feedback, Logistic produces a smooth bounded nonlinear response, AbsV and SignSwitch-AR generate dependence with weak or vanishing linear correlation, and CenterVol isolates conditional heteroskedasticity. Together they cover the principal nonlinear alternatives considered in econometrics.

Additionally, we utilize three representative geometric DGPs from Geenens and Lafaye de Micheaux (2022). The lag pairs are generated around a circle, spiral, or four-circle manifold with Gaussian perturbations before standardization. These designs deliberately violate a single-valued regression representation and therefore assess sensitivity to nonlinear dependence beyond linear or monotone association. For all recursive and geometric DGPs, $\rho = 0$ corresponds to serial independence, while increasing ρ strengthens dependence.

Figure 1 illustrates one realization from each DGP at $T = 126$ and $\rho = 0.85$. The time series alone often conceals the dependence structure, whereas the lag scatter reveals the underlying geometry. Several nonlinear designs exhibit pronounced dependence despite negligible lag-one Pearson correlation. We note that the figure is descriptive rather than inferential and uses the same DGP definitions as the power experiment.

All power comparisons are size-adjusted using independent null simulations. For AHCF and ADCF, the cutoff is the largest empirical p-value threshold whose null rejection frequency does not exceed 0.05. The six nongeometric DGPs share an i.i.d. Gaussian null. Each geometric DGP uses its own $\rho = 0$ marginal distribution. The 99-permutation S_ρ p-values lie on a grid with spacing 0.01, for which the nonrandomized rule can be excessively

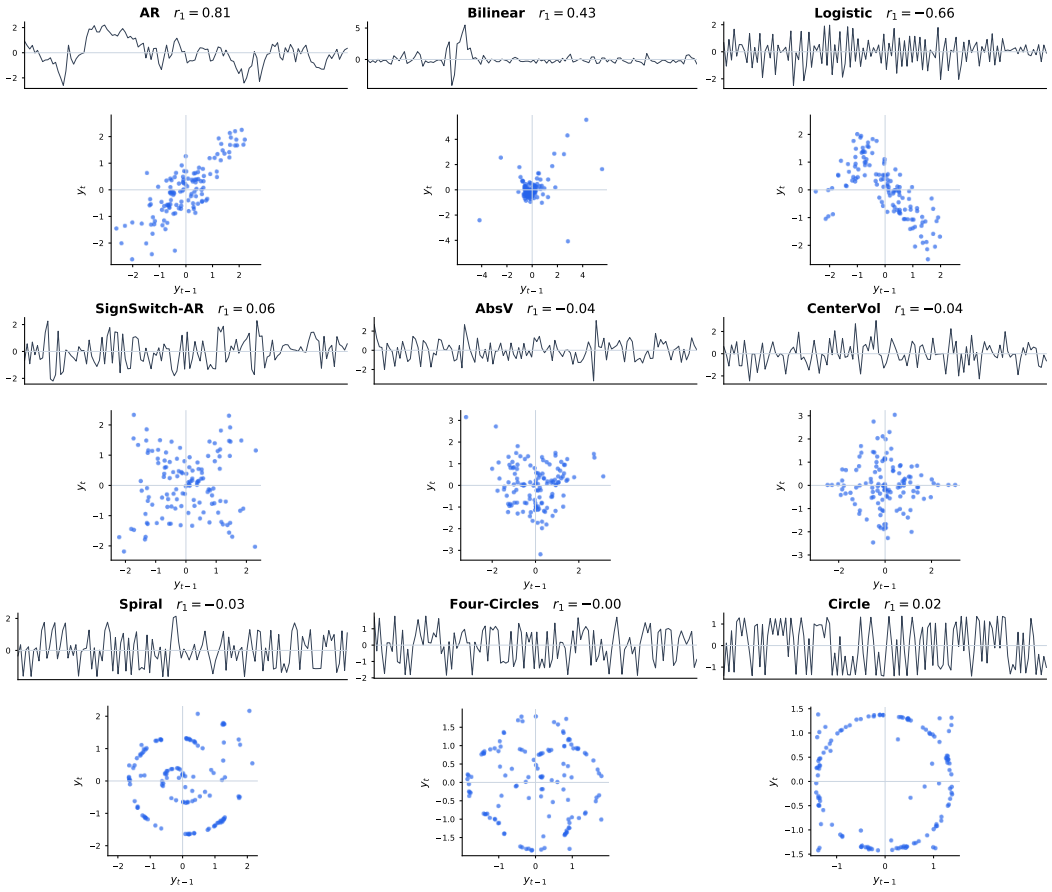


Figure 1: Representative DGP realizations and lag-one scatter plots, $T = 126$, $\rho = 0.85$, and $q = 1$

Notes: Each panel uses a fixed DGP-specific seed. The upper inset is the standardized time-series realization. The lower plot contains (y_{t-1}, y_t). Each panel heading reports the sample lag-one Pearson correlation r_1.

conservative. We therefore use boundary randomization for S_ρ. For a matched empirical null distribution $\widehat{F}_{\text{null}}$, let c_α satisfy

$$\widehat{F}_{\text{null}}(c_\alpha^-) \leq 0.05 \leq \widehat{F}_{\text{null}}(c_\alpha), \qquad \gamma = \frac{0.05 - \widehat{F}_{\text{null}}(c_\alpha^-)}{\widehat{F}_{\text{null}}(c_\alpha) - \widehat{F}_{\text{null}}(c_\alpha^-)}.$$

The randomized boundary rule modifies only the simulation evaluation threshold; it leaves the S_ρ statistic, bandwidth selection, and permutation algorithm unchanged. The rule rejects when $p < c_\alpha$, rejects with probability γ when $p = c_\alpha$, and does not reject otherwise, giving an empirical null rejection probability of exactly 0.05. As a robustness check, applying the same rule to AHCF and ADCF changes their average powers by at most 0.33 percentage points and does not affect any substantive ranking.

3.2 Size evidence

We report nominal rejection frequencies from the size experiment; the independent size adjustments used in the power comparison are not applied. Under the Gaussian benchmark (reported in full in the Online Supplementary Material), AHCF remains close to the nominal 5% level at both sample sizes. ADCF is mildly conservative at $T = 126$, whereas the cumulative S_ρ rejects 10.7% of the time. At $T = 1008$, all three methods are substantially closer to the nominal level, although the AHCF $q = 2$ cell rejects 6.1% of the time and cumulative S_ρ rejects 6.2%.

Table 4 summarizes the complete twelve-law exercise. AHCF has the most uniform rejection frequencies in the short sample: the known-lag mean is 0.0497 and the largest absolute distortion is 0.0118, while the unknown-lag mean is 0.0465 with maximum distortion 0.0102. ADCF is materially conservative at $T = 126$, especially for contaminated-normal, product-normal, and t_3 marginal distributions. Its known- and unknown-lag means are 0.0382 and 0.0353. The known-lag S_ρ tests are reasonably close to nominal on average, but the short-sample cumulative test is liberal: the mean rejection frequency is 0.0684 and the maximum distortion is 0.0570. At $T = 1008$, all procedures improve, and the AHCF unknown-lag rejection frequencies lie between 0.0456 and 0.0522 over all twelve marginal distributions.

Table 4: Distributional summary of empirical size

		Known lag		Unknown lag	
T	Method	Mean	Maximum |error|	Mean	Maximum |error|
126	AHCF	0.0497	0.0118	0.0465	0.0102
	ADCF	0.0382	0.0278	0.0353	0.0278
	S_ρ	0.0565	0.0220	0.0684	0.0570
1008	AHCF	0.0514	0.0110	0.0490	0.0044
	ADCF	0.0480	0.0150	0.0472	0.0168
	S_ρ	0.0527	0.0220	0.0555	0.0180

Notes: Known-lag summaries contain 36 equally weighted cells (12 marginal distributions and $q = 1, 2, 3$). Unknown-lag summaries contain 12 cells. Maximum |error| is the largest absolute deviation from 0.05. At a 5% rejection probability, the cell-level Monte Carlo standard error is approximately 0.0031 for AHCF and ADCF and 0.0069 for S_ρ. Distribution-level results appear in the Online Supplementary Material.

Across the twelve marginal distributions, AHCF varies more across lags than across marginal laws, consistent with its rank-based distribution-freeness. ADCF remains conservative under several non-Gaussian marginals at $T = 126$, while the cumulative S_ρ test exhibits the largest short-sample distortion. Detailed distribution-by-distribution results are reported in the Online Supplementary Material.

For computational reproducibility, we keep the reference and calibration draws fixed throughout all Monte Carlo replications, as in the power and empirical exercises. Consequently, the reported rejection frequencies may differ slightly from the nominal 5% level. The theoretical validity result instead averages over all possible reference and calibration draws and therefore guarantees exact size. This explains the small systematic deviations observed in the size experiment. Further discussion is provided in the Online Supplementary Material.

3.3 Size-adjusted power comparison

Table 5 reports average size-adjusted power over all nine DGPs and twenty positive values of ρ. AHCF has higher average power than ADCF at both sample sizes and in both panels. The kernel S_ρ benchmark has the highest overall average: relative to AHCF, its advantage is 5.1 percentage points at the known lag and 3.1 points in the unknown-lag panel when $T = 126$, and 5.2 and 3.4 points, respectively, when $T = 1008$.

Table 5: Average size-adjusted power over the simulation grid

	Known lag $q = 1$			Unknown lag		
T	AHCF	ADCF	S_ρ	AHCF	ADCF	S_ρ
126	0.560	0.402	0.611	0.477	0.393	0.508
1008	0.816	0.752	0.868	0.781	0.734	0.815

Notes: Entries average size-adjusted rejection frequencies over nine DGPs and $\rho = 0.05, 0.10, \ldots, 1.00$. AHCF and ADCF use 5,000 replications per cell, and S_ρ uses 1,000. The AHCF procedure uses $\mathcal{M} = \{1, \ldots, 10\}$. Its unknown-lag version aggregates by the maximum over $q = 1, \ldots, 5$. ADCF uses the fixed-lag and Bartlett procedures of Fokianos and Pitsillou (2017). The coarse S_ρ p-values use the boundary-randomized adjustment defined in the text. A common-randomization sensitivity check leaves the conclusions unchanged.

The average rankings mask important differences across DGPs. At $T = 126$, AHCF performs best for the CenterVol and Four-Circles designs, with average known-lag powers of 0.322 and 0.523, compared with 0.263 and 0.468 for S_ρ. The same ordering holds in the unknown-lag panel (0.218 versus 0.190 and 0.429 versus 0.337). By contrast, S_ρ is more powerful for the Bilinear, SignSwitch-AR, AbsV, Circle, and Spiral designs, while ADCF performs best for the linear AR benchmark and remains competitive for the Logistic and AbsV alternatives. At $T = 1008$, S_ρ achieves the highest average known-lag power across all designs. Even so, AHCF remains competitive for CenterVol in the unknown-lag setting and continues to outperform ADCF for the geometric and centered-volatility alternatives.

Figures 2 and 3 show that these patterns are not driven by a few isolated cases. AHCF's largest advantages over ADCF arise for the symmetry-based and geometric alternatives, where low-order distance correlation may require much stronger dependence to detect de-

partures from the null. As the sample size increases, S_ρ adapts well to most of these alternatives, although its cumulative test is less well calibrated in short samples. By contrast, AHCF combines exact distribution-free calibration under the i.i.d. null for continuous marginals with no density estimation or bandwidth selection, natural lag localization, and competitive or superior power in several challenging short-sample settings. Overall, the results support AHCF as a theoretically simpler and computationally lighter complement to the two benchmark procedures rather than as a uniformly most powerful test.

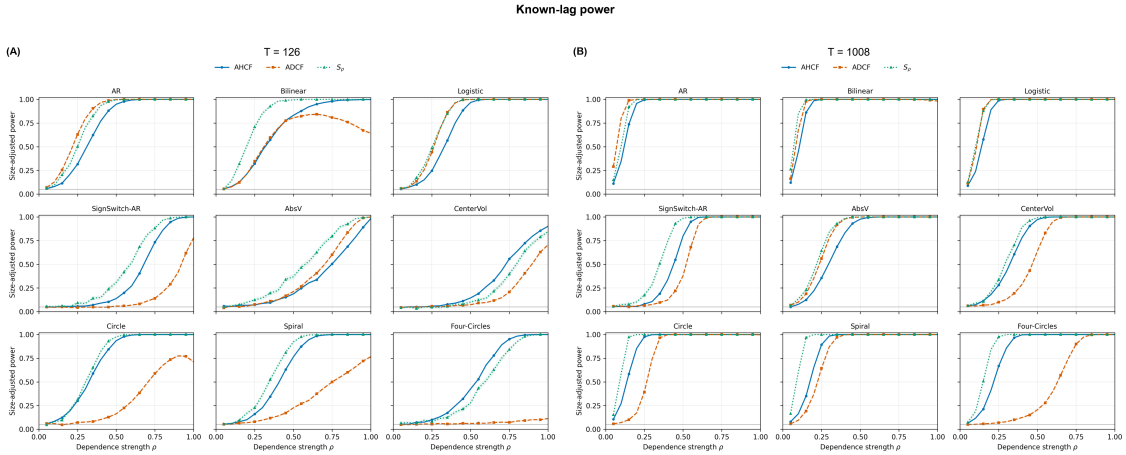


Figure 2: Size-adjusted power at the known lag $q = 1$, $T = 126$ (left) and $T = 1008$ (right)

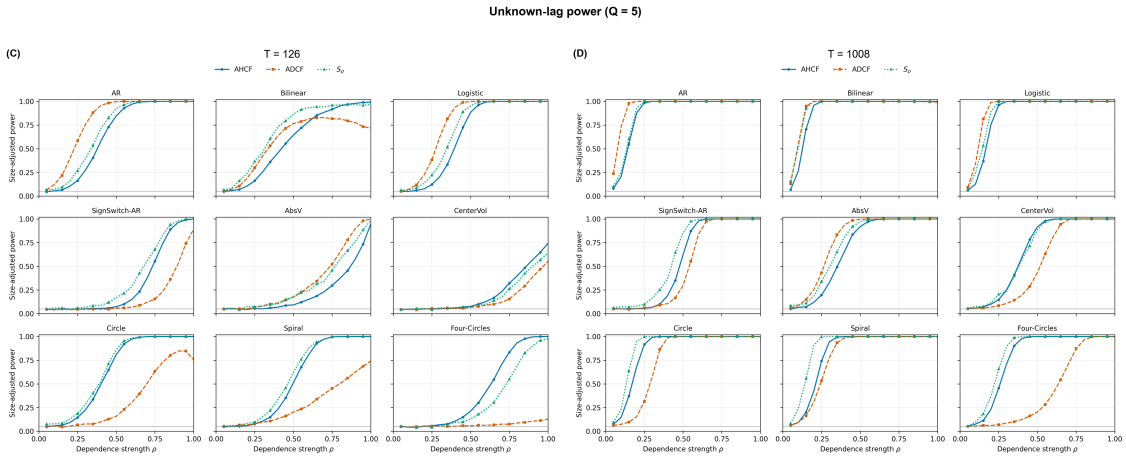


Figure 3: Size-adjusted power with unknown lag, $Q = 5$, $T = 126$ (left) and $T = 1008$ (right)

4 Empirical Application

4.1 Forecast data and error construction

We study professional forecast errors for US consumer-price inflation, US real GDP growth, and euro-area Harmonised Index of Consumer Prices (HICP) inflation. The application is deliberately in the short-sample range, with effective sample sizes between 81 and 176 observations.

For the United States, forecasts are obtained from the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters (SPF)[6]. We analyze next-quarter CPI (`CPI3`) and real GDP growth (`DRGDP3`) forecasts, using the median consensus as the primary specification and the mean as a robustness check. CPI realizations are constructed from the seasonally adjusted CPI-U all-items index obtained from the Bureau of Labor Statistics, while GDP realizations are taken from the Bureau of Economic Analysis advance estimates archived in the Philadelphia Fed Real-Time Data Set for Macroeconomists[7]. We retain the uninterrupted 1996Q1-2025Q2 block thereby avoiding the treatment of observations separated by missing target periods as consecutive quarters.

For the euro area, we use the published mean ECB SPF point forecast for one-year-ahead rolling HICP inflation and the current-vintage annual rate of change in the all-items HICP[8]. A reconstructed ECB median is excluded because 54 of its 105 forecast errors are tied, whereas the exact AHCF calibration assumes continuous marginal distributions.

For each series, forecast errors are defined as $e_t = y_t - \widehat{y}_{t|t-h}$, so positive values indicate underprediction. Observations are ordered by target period, and no demeaning, scaling, differencing, interpolation, or filtering is applied. In particular, differencing would alter the economic null hypothesis and mechanically introduce moving-average dependence.

The full samples span 1981Q4-2025Q3 for CPI ($n = 176$), 1996Q1-2025Q2 for GDP ($n = 118$), and 1999Q4-2025Q4 for ECB ($n = 105$). Pre-COVID robustness samples end in 2019Q4 and contain 153, 96, and 81 observations, respectively. Rather than deleting the pandemic period and reconnecting the remaining observations, we truncate each sample at 2019Q4.

The primary forecast-error series are shown in the Online Supplementary Material. CPI errors exhibit several episodes of elevated volatility, GDP errors are dominated by the extreme observations around 2020, and ECB errors evolve more smoothly because successive

[6]The SPF data and documentation are available from https://www.philadelphiafed.org/surveys-and-data/cpi-spf

[7]The real-time output files and documentation are available from https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/routput.

[8]The forecast and realization series are https://data.ecb.europa.eu/data/datasets/SPF and https://data.ecb.europa.eu/data/datasets/HICP.

one-year-ahead forecast horizons overlap. Since AHCF depends only on the ranks, isolated large forecast errors cannot dominate the test statistic.

4.2 Testing protocol and interpretation

We implement the fixed-lag and omnibus AHCF tests proposed in Section 3 using basis orders $m = 1, \ldots, 10$ and $q = 1, \ldots, 8$. For the fixed-lag test, we evaluate the AHCF statistic for each of the ten basis orders at the lag of interest. This produces ten p-values, one for each basis order. Rather than selecting a basis order in advance, we use the smallest of these p-values as the adaptive statistic and obtain its significance by Monte Carlo calibration. For the omnibus test, we first evaluate the AHCF statistic at all eight lags for each basis order and retain the largest statistic within each basis order. The resulting ten statistics are converted into p-values and combined exactly as in the fixed-lag test, with Monte Carlo calibration providing the final omnibus p-value. In Table 6 below summarizes the aforementioned procedure.

Table 6: Summary of the AHCF testing procedures.

Procedure	Computational steps
Fixed-lag test	10 AHCF statistics $\rightarrow$ 10 p-values $\rightarrow$ smallest p-value $\rightarrow$ Monte Carlo calibration $\rightarrow$ final adaptive p-value.
Omnibus test	80 AHCF statistics (10 basis orders $\times$ 8 lags) $\rightarrow$ 10 within-order maxima $\rightarrow$ 10 p-values $\rightarrow$ smallest p-value $\rightarrow$ Monte Carlo calibration $\rightarrow$ final omnibus p-value.

Calibration is based on $B_0 = 50{,}000$ reference draws and $B_1 = 2{,}000$ calibration draws with tuning parameter $a = 6$. Hypothesis testing uses the adaptive combination procedure described in Section 3 over all basis orders. Leave-one-out cross-validation is used only to select a basis order for estimating the strength of AHCF dependence.

The linear benchmark uses the classical single-lag statistic $T\widehat{r}(q)^2$ with its chi-square reference, and the Ljung-Box test ($Q = 8$) for omnibus inference. ADCF is implemented using its known-lag and Bartlett portmanteau statistics with 499 Gaussian wild-bootstrap draws. Its automatic bandwidth selects lags 1-4 or 1-5 in our samples. The kernel Hellinger procedure uses deterministic integration, likelihood-cross-validated bandwidths estimated once from the observed data and held fixed throughout the 99 whole-series permutations, and the cumulative S_ρ through lag eight. Because the original article does not specify how bandwidths should be handled under permutation, we follow the defaults of the accompanying R implementation used in our replication files; the implications are discussed in the

Online Supplementary Material. No procedure is retuned or empirically size-adjusted.

The portmanteau p-value is the primary omnibus result. Known-lag tests identify the lags at which dependence occurs and are adjusted by Holm's step-down method separately within each series, sample period, and method across $q = 1, \ldots, 8$ (Holm, 1979). Portmanteau p-values are not included in those lagwise families. The order-by-lag component p-values shown below are descriptive and unadjusted. Formal inference is based on the jointly calibrated combined p-values.

Two points are worth noting when interpreting the results. First, under squared-error loss, forecast rationality requires only that forecast errors have zero conditional mean, whereas the AHCF tests examine the stronger property of serial independence of the entire error distribution. Consequently, rejecting the null indicates serial dependence in the forecast errors but does not by itself imply that the conditional-mean forecast is inefficient or can be improved. This interpretation is consistent with the broader conditional predictive ability framework of Giacomini and White (2006).

Second, the ECB forecasts have a one-year horizon but are issued quarterly. Consequently, consecutive forecast errors overlap in realized outcomes, so dependence at lags 1-3 can arise mechanically even under forecast efficiency (Hansen and Hodrick, 1980). We therefore report the ECB omnibus results for completeness, but focus our economic interpretation on the Holm-adjusted known-lag results for $q = 4, \ldots, 8$, where horizon overlap is absent.

4.3 Empirical evidence

Table 7 reports the method-specific omnibus p-values. Panel A contains the primary consensus series and Panel B reports the alternative US mean consensus. No empirical size adjustment is applied. For ADCF, the smallest attainable p-value is $1/499$, shown as < 0.002, while the 99-permutation S_ρ has a minimum attainable p-value 0.01.

4.3.1 US CPI forecast errors

The CPI forecasts provide the clearest evidence that AHCF contributes information beyond conventional linear diagnostics. In the full sample, all four omnibus procedures reject the null of serial independence, indicating that dependence is readily detectable regardless of the method. The more informative comparison is therefore the pre-COVID sample. Before COVID, AHCF continues to reject the omnibus null for the primary median consensus (p=0.049), whereas the Ljung-Box test does not (p=0.118). The dependence is localized at lag one, where the AHCF known-lag test remains significant after Holm adjustment (p=0.048). In contrast, the corresponding Pearson test loses significance after adjustment.

Table 7: Portmanteau tests of forecast-error serial independence

Forecast error	Sample	n	AHCF	Ljung–Box	ADCF	S_ρ
Panel A: primary consensus series						
US CPI, median	Full	176	**0.003**	**0.005**	< 0.002	**0.010**
	Pre-COVID	153	**0.049**	0.118	**0.040**	**0.010**
US real GDP, median	Full	118	0.674	0.312	0.252	**0.010**
	Pre-COVID	96	0.186	0.236	**0.014**	**0.010**
ECB HICP, mean	Full	105	< 0.001	< 0.001	< 0.002	**0.010**
	Pre-COVID	81	< 0.001	< 0.001	< 0.002	**0.010**
Panel B: US mean-consensus robustness						
US CPI, mean	Full	176	**0.002**	**0.010**	< 0.002	**0.010**
	Pre-COVID	153	**0.009**	0.191	0.056	**0.010**
US real GDP, mean	Full	118	0.635	0.335	0.312	**0.010**
	Pre-COVID	96	**0.016**	0.280	**0.028**	**0.010**

Notes: Bold entries reject at 5%. AHCF combines orders $m = 1, \ldots, 10$ after maximum aggregation over $q = 1, \ldots, 8$. Ljung–Box uses $Q = 8$. ADCF uses the Bartlett bandwidth as specified in its source article and 499 Gaussian wild-bootstrap draws. S_ρ uses 99 whole-series permutations. ECB omnibus tests include the mechanically overlapping lags 1–3 and are descriptive only. Pre-COVID samples end at 2019Q4.

Thus, AHCF identifies distributional dependence at lag one that is not captured by the linear benchmark.

This conclusion is robust to the alternative consensus definition. For the mean consensus, the AHCF omnibus test rejects more strongly (p=0.009), while neither Ljung-Box nor ADCF rejects at the 5% level. The lag-one AHCF result also remains significant after Holm adjustment.

The order-by-lag decomposition in Figure 4 indicates that the evidence is concentrated in the lowest basis orders. Before COVID, significant raw component statistics are confined to orders $m = 1, \ldots, 4$ for the median consensus and $m = 1, \ldots, 5$ for the mean consensus, with the strongest contribution occurring at $m = 1$. These component statistics explain the omnibus rejection but are descriptive only, since formal inference is based on the jointly calibrated combined p-values.

The lagged rank pseudo-observations, reported in the Online Supplementary Material, display only a modest diagonal pattern rather than a pronounced nonlinear structure. Accordingly, we interpret the result as evidence of distributional dependence that is not fully summarized by the linear omnibus test, rather than evidence that the conditional-mean forecast is inefficient.

4.3.2 US real GDP forecast errors

The GDP results do not support a primary AHCF rejection. For the median consensus, the omnibus AHCF p-value is 0.674 in the full sample and 0.186 before COVID. In the

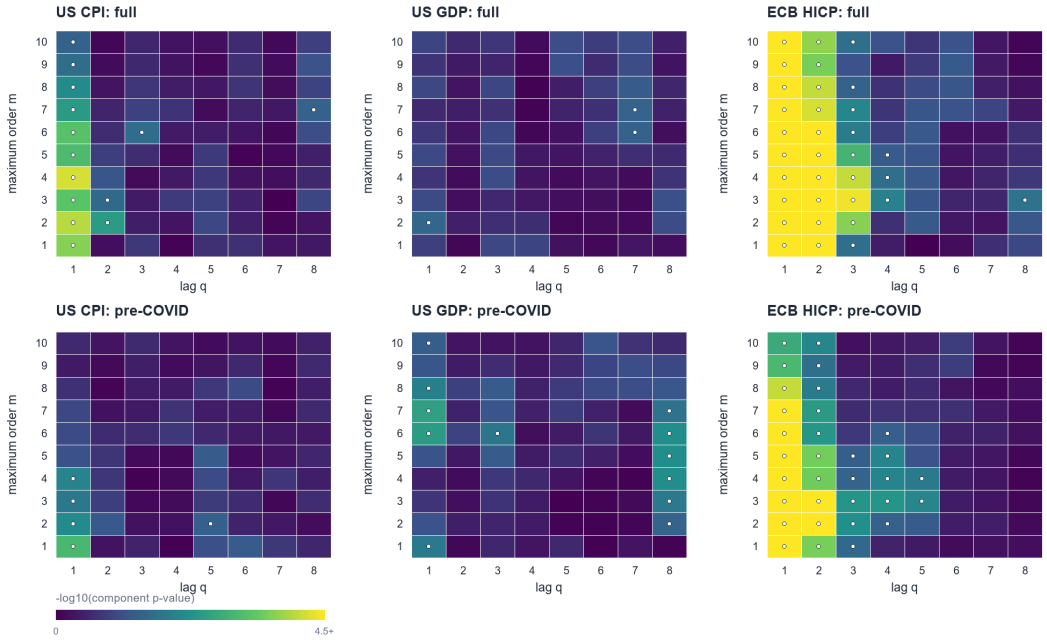


Figure 4: Order and lag composition of AHCF evidence

Notes: Colors report $-\log_{10}(p)$ for $Z_{m,q} = T\widehat{\eta}_{m,m}(q)^2$, $m = 1, \ldots, 10$, $q = 1, \ldots, 8$. White dots mark raw component p-values no greater than 0.05. Each panel contains eighty component cells, so roughly four raw rejections at the 5% level would arise by chance even under the null. These nested component values are descriptive and unadjusted. Formal inference uses the jointly calibrated combined AHCF p-values. ECB lags 1–3 are affected by horizon overlap.

pre-COVID median sample, raw known-lag p-values of 0.020 at lag one and 0.042 at lag eight do not survive Holm adjustment. ADCF rejects its pre-COVID omnibus test at 0.014, while S_ρ reports 0.010.

The alternative mean consensus yields a different pre-COVID result. AHCF rejects the omnibus null (p=0.016), with Holm-adjusted rejections at lags one, three, and eight. Ljung-Box again does not reject, whereas ADCF also rejects (p=0.028). Unlike the CPI application, the supporting AHCF evidence is concentrated in relatively high nested orders, as shown in the Online Supplementary Material. Because this finding is not reproduced for the primary median consensus, we regard it as consensus-sensitive rather than a principal empirical result. The absence of a full-sample rejection further suggests that the COVID period changes the ordering of forecast-error ranks rather than merely their scale.

4.3.3 ECB HICP forecast errors

All four ECB omnibus procedures reject in both samples because they include lags one through three, where consecutive one-year-ahead forecast errors necessarily share realization periods. The known-lag AHCF results confirm this interpretation. In the full sample, AHCF rejects at lags one, two, and three after Holm adjustment, and the heatmap shows broad

evidence across nearly all nested orders. There is no reliable AHCF evidence beyond the overlapping horizon. The full-sample lag-four p-value is 0.053 before adjustment and 0.267 after adjustment. Before COVID, lags four and five have raw p-values 0.031 and 0.047, but their adjusted values are 0.183 and 0.188. That is, although lags four and five are marginally significant before adjustment, no lag beyond the overlap survives Holm adjustment. The ECB application therefore illustrates AHCF's ability to localize dependence implied by the forecast design without suggesting additional serial dependence beyond the mechanical overlap.

Table 8: AHCF omnibus and Holm-adjusted lag localization

Forecast error	Full sample		Pre-COVID	
	Omnibus	Holm-significant lags	Omnibus	Holm-significant lags
US CPI, median	**0.003**	$q = 1$ (0.016)	**0.049**	$q = 1$ (0.048)
US CPI, mean	**0.002**	$q = 1$ (0.008)	**0.009**	$q = 1$ (0.008)
US GDP, median	0.674	None	0.186	None
US GDP, mean	0.635	None	**0.016**	$q = 1$ (0.024), $q = 3$ (0.042), $q = 8$ (0.020)
ECB HICP, mean	< 0.001	$q = 1, 2, 3$	< 0.001	$q = 1, 2$

Notes: Parentheses contain Holm-adjusted known-lag p-values. For ECB HICP, all reported significant lags are affected by the overlapping forecast horizon. No lag $q \geq 4$ survives adjustment.

4.3.4 Comparison across procedures

The empirical comparison is necessarily qualified rather than absolute. AHCF provides the clearest additional evidence in the pre-COVID CPI forecasts, where both its omnibus rejection and lag-one localization are robust across consensus definitions despite the absence of corresponding linear evidence. By contrast, the nonlinear procedures do not always agree. ADCF rejects for the pre-COVID GDP median whereas AHCF does not, while AHCF provides stronger evidence for the pre-COVID CPI mean. These differences are consistent with the complementary power properties documented in Section 3.

The kernel Hellinger procedure behaves differently. It attains its minimum possible p-value (0.010) in every omnibus application, including cases where the other procedures provide little evidence of serial dependence. This pattern is consistent with the coarse permutation resolution and short-sample size distortion documented in Section 3. Because those simulations were conducted for $Q = 5$, whereas the empirical analysis uses $Q = 8$, we avoid drawing strong conclusions from isolated S_ρ rejections.

Overall, the empirical contribution is not one of uniform dominance. Rather, AHCF combines exact finite-sample calibration, formal lag localization, and an interpretable order decomposition. These features are illustrated most clearly by the robust CPI findings and by the correct localization of the mechanically induced dependence in the ECB forecasts.

5 Conclusion and Discussion

This paper develops the Auto-Hellinger Correlation Function as a lag-specific measure of serial dependence together with a rank-based framework for testing serial independence at individual lags and over fixed lag ranges. Because the AHCF depends on the full copula of (y_t, y_{t-q}), it detects nonlinear, nonmonotonic, and volatility-driven dependence while remaining compatible with Gaussian linear structures. We establish consistency of the Beta-transformed, weighted nearest-neighbor estimator of the full Legendre coefficient vector under geometric absolute regularity and finite-dimensional copula density bounds. Two key ingredients simplify the analysis: Beta-transformed coordinates yield uniform interior moment bounds, while a deterministic geometric argument controls boundary effects without probabilistic assumptions.

On the inference side, serial independence is a nonregular boundary point of the Hellinger-correlation map, so inference is based on the squared coefficient statistic $Z_{m,q} = T\widehat{\eta}_{m,m}(q)^2$ together with an order-adaptive minimum-p procedure calibrated by nested rank-based Monte Carlo draws. Under the i.i.d. calibration null, the resulting tests are exactly distribution-free over continuous marginal distributions and attain unconditional finite-sample size no greater than the nominal level. They are consistent against fixed alternatives with nonzero Legendre coefficients, without requiring an asymptotic distribution for the estimator. In simulations, AHCF remains close to the nominal level across a wide range of continuous distributions and achieves higher average size-adjusted power than ADCF for several nonlinear alternatives, while ADCF performs better for some smooth linear alternatives. The kernel S_ρ benchmark often achieves the highest average power but at the cost of noticeable short-sample size distortions.

The empirical application illustrates the practical value of exact finite-sample calibration and lag localization. For US CPI forecasts, AHCF detects robust pre-COVID lag-one dependence that is missed by the Ljung-Box omnibus test. This finding indicates distributional dependence in the forecast errors rather than conditional-mean inefficiency. For US GDP forecasts, the evidence is sensitive to the consensus definition and is therefore treated as secondary. For the overlapping one-year-ahead ECB forecasts, AHCF correctly localizes the detected dependence to the mechanically overlapping horizons, with no evidence beyond the overlap after Holm adjustment. Together, these applications illustrate both the strengths and the appropriate interpretation of the proposed methodology.

The asymptotic analysis is intentionally limited to fixed-order consistency. Extending the theory to root-n inference for the feasible estimator would require a joint treatment of nearest-neighbor boundary bias, sample-dependent local geometry, and empirical-rank replacement, none of which is needed for the rank-calibrated tests developed here. Natural

extensions include growing basis and lag dimensions, multivariate and conditional versions of AHCF, structural breaks, long memory, and formal criteria for selecting lag aggregations. These extensions would broaden the scope of the methodology while preserving its central role as a nonparametric, lag-specific, and exactly calibrated complement to classical autocorrelation and other nonlinear serial-dependence diagnostics.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

The contribution of each author is summarized as follows: **Ramazan Ekinci:** Conceptualization, Software, Formal analysis, Investigation, Writing - Original Draft; **Burak Alparslan Eroğlu:** Conceptualization, Software, Formal analysis, Methodology, Investigation, Writing - Original Draft; **Haluk Yener:** Conceptualization, Methodology, Writing - Review & Editing; **Taner Yiğit:** Conceptualization, Methodology, Writing - Review & Editing. All authors reviewed and approved the final manuscript.

References

Ashley, R. A., Patterson, D. M., and Hinich, M. J. (1986). A diagnostic test for nonlinear serial dependence in time series fitting errors. *Journal of Time Series Analysis*, 7(3):165–178.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697.

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.

Brock, W. A., Dechert, W. D., Scheinkman, J. A., and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3):197–235.

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, 91(434):674–688.

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

Geenens, G. and Lafaye de Micheaux, P. (2022). The Hellinger correlation. *Journal of the American Statistical Association*, 117(538):639–653.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Giannerini, S., Maasoumi, E., and Dagum, E. B. (2015). Entropy testing for nonlinear serial dependence in time series. *Biometrika*, 102(3):661–675.

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

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748.

Hansen, L. P. and Hodrick, R. J. (1980). Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. *Journal of Political Economy*, 88(5):829–853.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *TEST*, 27(4):811–825.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica*, 69(3):599–631.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, 89(427):1000–1005.

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

Madan, D. B., Carr, P. P., and Chang, E. C. (1998). The variance gamma process and option pricing. *European Finance Review*, 2(1):79–105.

Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4):394–419.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume*, pages 273–292. Institute of Mathematical Statistics.

Nadarajah, S. and Pogány, T. K. (2016). On the distribution of the product of correlated normal random variables. *Comptes Rendus Mathématique*, 354(2):201–204.

Patterson, D. M. and Ashley, R. A. (2000). *A nonlinear time series workshop: A toolkit for detecting and identifying nonlinear serial dependence*, volume 2 of *Dynamic Modeling and Econometrics in Economics and Finance*. Springer, New York.

Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer, Berlin, Heidelberg.

Rusticelli, E., Ashley, R. A., Dagum, E. B., and Patterson, D. M. (2008). A new bispectral test for nonlinear serial dependence. *Econometric Reviews*, 28(1-3):279–293.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116.