Extremal Dependence-Based Specification Testing of Time Series

Yannick Hoga
Faculty of Economics and Business Administration, University of Duisburg-Essen, Essen, Germany

ABSTRACT
We propose a specification test for conditional location–scale models based on extremal dependence properties of the standardized residuals. We do so comparing the left-over serial extremal dependence—as measured by the pre-asymptotic tail copula—with that arising under serial independence at different lags. Our main theoretical results show that the proposed Portmanteau-type test statistics have nuisance parameter-free asymptotic limits. The test statistics are easy to compute, as they only depend on the standardized residuals, and critical values are likewise easily obtained from the limiting distributions. This contrasts with some extant tests (based, e.g., on autocorrelations of squared residuals), where test statistics depend on the parameter estimator of the model and critical values may need to be bootstrapped. We show that our tests perform well in simulations. An empirical application to S&P 500 constituents illustrates that our tests can uncover violations of residual serial independence that are not picked up by standard autocorrelation-based specification tests, yet are relevant when the model is used for, for example, risk forecasting.

1. Motivation

The dynamics of many economic and financial time series can be successfully captured by location–scale models of the form $Y_t = \mu_t(\theta^0) + \sigma_t(\theta^0)\epsilon_t$, where $\theta^0$ is some unknown parameter vector, and $\mu_t(\theta^0)$ and $\sigma_t(\theta^0)$ are parametric location and scale curves, respectively. The benchmark models for incorporating mean changes are ARMA models, and GARCH-type processes have become the standard volatility models. The main aim of this article is to derive tests that check for serial tail dependence in the innovations $\{\epsilon_t\}_{t \in \mathbb{Z}}$. Rejecting the null of no tail dependence also immediately implies innovations are not independent, identically distributed (iid). Of course, failing to reject does not imply that innovations are iid, and by construction a test focusing on tail dependence will have no power against other deviations from iid innovations. On the positive side, focusing on serial extremal dependence properties implies that misspecifications in the tails can be detected more easily, which is important, for example, in a risk management context as further outlined below. In the following, we will often refer to our proposal as a test for iid innovations both for expository simplicity (we in fact assume iidness under the null) and because tail dependence is a particularly relevant deviation from iidness.

Next to risk management, the iid assumption is crucial in the context of location–scale models, because it plays a key role in drawing inferences on $\theta^0$, in testing the correct parametric specification of $\mu_t(\theta^0)$ and $\sigma_t(\theta^0)$, and for purposes of option pricing. In risk management, the iid assumption has implications for the computation of risk forecasts. For instance, for iid innovations, Value-at-Risk (VaR) forecasts can be computed as $\text{VaR}_t = \mu_t(\theta^0) + \sigma_t(\theta^0)F^{-1}_\epsilon(\theta)$, where $F^{-1}_\epsilon(\theta)$ denotes the $\theta$-quantile of the $\epsilon_t$. Such risk measure forecasts are key inputs in risk management procedures of financial institutions, due to regulatory requirements in the Basel framework of the Basel Committee on Banking Supervision (2019). The Basel framework penalizes sustained underprediction of risk by imposing higher capital requirements. On the other hand, if risk is overpredicted, too much capital is put aside as a buffer against large losses. In both cases of over- and underprediction, some portion of the capital can no longer earn premiums, leading to foregone profits. This underscores the importance of accurate risk forecasts from correctly specified models. Consequently, diagnostic tests should be routinely applied to the chosen forecasting model, because they provide valuable information on whether the above formula should be used for out-of-sample VaR forecasting.

Several such diagnostic tests are available that verify whether the model residuals are close to iid This is done by checking some implication of the iid property; for example, the absence of autocorrelation. However, standard tests of this—such as Box and Pierce (1970) and Ljung and Box (1978) tests—cannot be applied “as usual,” since the residuals are only estimated. Li and Mak (1994) were the first to propose a Portmanteau-type test for the autocorrelations of squared residuals that corrects for this fact in (conditionally Gaussian) location–scale models. Berkes, Horváth, and Kokoszka (2003) and Ling and Li (1997) extend the applicability of the Li and Mak (1994) test to more general GARCH and $N(0,1)$–FARIMA–GARCH models, respectively. Similarly, Fisher and Gallagher (2012) consider weighted...
Portmanteau statistics applied to ARCH residuals. Hidalgo and Zaffaroni (2007) develop spectral-based tests for ARCH(∞) series. The limiting distributions for all these proposed test statistics have only been derived for the quasi-maximum likelihood (QML) estimator. Moreover, practical application of these tests is complicated by the fact that the limiting distributions depend on nuisance parameters induced by parameter estimation. Valid tests thus require consistent estimators of the nuisance parameters (that have to be computed on a case-by-case basis) or involved bootstrap-based procedures to compute critical values. By construction, all these tests diagnose the complete serial dependence structure of the standardized residuals.

Here, instead of considering the complete dependence structure, we take a different tack by focusing on the left-over serial extremal dependence in the standardized residuals. We emphasize that as we test a very narrow implication of the iid null (namely the absence of serial extremal dependence) the properties of our tests may be invalidated by deviations from the null not captured by our test statistics (such as non-identically distributed innovations). In particular, there are alternatives for which our tests have no power at all (e.g., remaining serial dependence only in the non-tail region). However, when the deviation from the iid null is most marked in the serial extremal dependence, our tests can be expected to have higher power than alternative tests considering the complete dependence structure. This is because when diagnosing the complete dependence structure, the effect of any remaining residual extremal dependence may be “washed-out.” However, overlooked serial dependence in the extremes may be very harmful as it invalidates, for example, the model’s risk forecasts, such as VaR and expected shortfall (ES) forecasts. As pointed out above, misspecified risk forecasts are penalized under the Basel framework. Hence, a separate diagnostic for the extremes—such as the one proposed in this article—is desirable. This point is underscored by the empirical application in Section 6, where our extremal dependence-based (in-sample) diagnostic is highly indicative of inadequate out-of-sample risk forecasts.

In empirical work, Davis, Mikosch, and Cribben (2012) and Davis, Mikosch, and Zhao (2013) use the pre-asymptotic (PA) extremogram as a diagnostic. For example, to assess the GARCH fit for FTSE returns, Davis, Mikosch, and Cribben (2012) show that the estimated PA-extremogram of the standardized residuals exhibits no signs of serial extremal dependence. However, such a model check lacks theoretical justification so far, because the central limit theory for extremal dependence measures has not been developed for model residuals. It is the main theoretical contribution of this article to do so.

To date, the literature has mainly explored the residual-based estimation of tail quantities, such as the tail index, VaR or ES. For instance, Chan et al. (2007) and Hoga (2019) derive confidence intervals for VaR and ES forecasts based on Hill estimates of residuals of GARCH-type models. Hill (2015b) provides an elaborate theory of tail index estimation based on residuals under high-level conditions. However, to the best of our knowledge, theoretical results on residual-based serial extremal dependence estimation (as developed in this article) do not exist. Our theorems show that—similarly as in Chan et al. (2007), Hoga (2019), and Hill (2015b)—the limit theory is unaffected by the use of estimated residuals, because the convergence rate of the parameter estimator is faster than that of the tail quantity. For actual observables, our theory most closely resembles that of Schmidt and Stadtmüller (2006) and Hill (2009, 2011b), who consider the tail copula and tail event correlation, respectively. However, our theory has to take care of the filtering step and the tail dependence measures we consider are more flexible.

Specifically, we base our tests on the PA-tail copula of the absolute values of the standardized residuals, that is, on $\Lambda^{(d)}(x, y) = (n/k)P[|\varepsilon_t| > b(n/\{kx\}), |\varepsilon_{t-d}| > b(n/\{ky\})]$, where $b(x)$ denotes the $(1 - 1/x)$-quantile of the $|\varepsilon_t|$. $n$ is the sample size and $k = k(n) \to \infty$ is an intermediate sequence with $k/n \to 0$. We use the PA-tail copula because of its flexibility in measuring extremal serial dependence for different $x, y > 0$. Under serial independence (and continuity sufficiently far out in the tail), the PA-tail copula equals $(k/n)^x$. Thus, we take any deviation from this null hypothetical value as evidence against the null of iid innovations.

Testing for any left-over serial extremal dependence in the residuals directs the power of our specification tests to the tails, thus, rendering them more powerful when, say, a GARCH model captures well the volatility dynamics in the body of the distribution, but not so much in the extremes (see also the simulations for such an example). It is exactly for these cases that our tests are designed. Thus, the diagnostic tests developed here are not meant to replace standard tests based (e.g.) on the auto-correlations of the squared residuals, but rather to supplement them.

It turns out that—unlike the above mentioned traditional Portmanteau tests due to Li and Mak (1994) and others—our tests are easy to apply: We only require the standardized residuals to compute the test statistics. In particular, it does not matter for our tests which parameter estimator (QML estimator, Self-Weighted QML estimator, Least Squares, etc.) was used in fitting the model, as long as it is $n^\kappa$-consistent for some $\kappa \in (0, 1/2]$. This gives our tests wide practical appeal. Only requiring $n^\kappa$-convergent (instead of $\sqrt{n}$-convergent) estimators may also be convenient in some applications. For instance, empirical results suggest that $E[\varepsilon_t^2] < \infty$ and $E[\varepsilon_t^4] = \infty$ for models of certain asset returns (Zhu and Ling 2011; Zheng, Li, and Li 2018). Thus, estimators may not be $\sqrt{n}$-convergent (Hall and Yao 2003), which—to the best of our knowledge—is always required by alternative specification tests; see Zheng, Li, and Li (2018, sec. 7).

As for other Portmanteau-type tests, the maximal lag length to be included in our test statistics is a tuning parameter. In Appendix F, supplementary materials, we adapt the proposal of Escanciano and Lobato (2009) to choose the maximal lag length data-adaptively to our context, which, however, requires the choice of another tuning parameter. Simulations show that our adaptation works well in some cases, but not in others (see Appendix F3, supplementary materials).

As is common for specification tests focusing on the iid property of the innovations, a rejection of our test points to a problem in the mean dynamics or variance dynamics or to non-iid-innovations, but does not establish which of the three. If the goal is to precisely identify the source of model failure, specification...
tests for the conditional mean only, and specification tests for the conditional mean and conditional variance jointly are required. Examples of the former include Francq, Roy, and Zakoïan (2005), Escanciano (2006), and Delgado and Velasco (2011), while Escanciano (2008, 2010) and Velasco and Wang (2015) focus on the latter. We mention that some of these tests share several desirable features with our test, such as a pivotal limiting distribution and reasonable generality regarding the choice of the parameter estimator. Note that our test is not a parametric conditional distribution test, that would verify, for example, conditional normality, such as in Rothe and Wied (2013) or Chen and Hong (2014).

We illustrate the good size and power of our tests in simulations. Specifically, we show that size is indeed unaffected by parameter estimation effects in sufficiently large samples, as predicted by theory. We also demonstrate that empirically relevant misspecifications can be detected more easily using our extremal dependence-based tests instead of (e.g.) autocorrelation-based tests.

The empirical application shows that for the S&P 500 components an autocorrelation-based test can often not detect any significant departures from a fitted GARCH-type model. However, in many of these cases, applying our tests suggests some significant departures from the fitted GARCH-type model to be introduced in Section 2.2. As pointed out in the Motivation, our goal is to verify whether their CAViaR models at different risk levels that “our findings suggest that the process governing the behavior of the tails is, of course, particularly worrisome when using GARCH-type models to forecast risk measures, such as VaR and ES, far out in the tail. Finally, Section 7 concludes. The Appendices in the supplementary materials contain all proofs, and additional simulations and theoretical results for our test statistics.

2. Preliminaries

2.1. The Tail Copula and its Estimators

Consider a sample \( \varepsilon_1, \ldots, \varepsilon_n \) of strictly stationary random variables. These will later on play the role of the innovations in the location–scale model to be introduced in Section 2.2. As pointed out in the Motivation, our goal is to verify whether their serial extremal dependence structure. Since serial extremal dependence in both tails of the \( \varepsilon_i \) provides evidence against the iid hypothesis, we introduce the serial extremal dependence measures in terms of the \( |\varepsilon_i| \) in the following.

Denote the distribution function (df) of the \( |\varepsilon_i| \)'s as \( F(\cdot) \), and let \( b(x) = F^-(1 - 1/x) \), where \( F^{-} \) denotes the left-continuous inverse. Then, following Schmidt and Stadtmüller (2006), the (upper) tail copula is defined as

\[
\Lambda_n^{(d)}(x, y) = \lim_{s \to \infty} sP\left\{ |\varepsilon_i| > b\left( \frac{s}{n} \right), |\varepsilon_{i-1,d}| > b\left( \frac{N}{n} \right), x, y > 0, \right\},
\]

where the limit is assumed to exist. Thus, the tail copula describes the serial extremal dependence structure of the \( |\varepsilon_i| \) at different lags \( d \). For \( x = y = 1 \), the tail copula simplifies to the tail dependence coefficient of Sibuya (1960). In this case, it also coincides with the most popular version of the extremogram due to Davis and Mikosch (2009).

In this article, we prefer to measure serial extremal dependence via the PA-tail copula. To introduce it, we let \( k = k(n) \) denote an intermediate sequence satisfying \( k \to \infty \) as \( n \to \infty \). The PA-tail copula then is

\[
\Lambda_n^{(d)}(x, y) = \frac{n}{k} P\left\{ |\varepsilon_i| > b\left( \frac{n}{k} \right), |\varepsilon_{i-1,d}| > b\left( \frac{n}{k} \right), x, y > 0, \right\}.
\]

Note that the PA-tail copula is closely related to the PA-extremogram of Davis, Mikosch, and Cribben (2012). For instance, for \( x = y = 1 \), the PA-tail copula is the perhaps most popular version of the PA-extremogram.

To highlight the main advantage of the PA-tail copula vis-à-vis the tail copula, recall that for iid \( \varepsilon_i \), following a serially dependent stochastic volatility (SV) model and for \( \left( |\varepsilon_i|, |\varepsilon_{i-1,d}| \right)' \) possessing a Gaussian copula, we have that \( \Lambda_n^{(d)}(x, y) \equiv 0 \) (Hefernan 2000; Hill 2011a). Thus, the tail copula cannot adequately discriminate between independent and dependent variables, which would result in compromised power of a specification test based on the tail copula. In contrast, for the PA-tail copula we have

\[
\Lambda_n^{(d)}(x, y) = \left( n/k \right) P\left\{ |\varepsilon_i| > b\left( \frac{n}{k} \right), |\varepsilon_{i-1,d}| > b\left( \frac{n}{k} \right) \right\} =: \Lambda_n^{(d), \text{ind}}(x, y) \quad \text{for iid } \varepsilon_i,
\]

yet \( \Lambda_n^{(d)}(x, y) \neq \Lambda_n^{(d), \text{ind}}(x, y) \) when the \( \varepsilon_i \) follow a SV model (Hill 2011a) or \( \left( |\varepsilon_i|, |\varepsilon_{i-1,d}| \right)' \) possesses a Gaussian copula (Hefernan 2000). Thus, the PA-tail copula captures finer differences in pre-asymptotic levels of extremal dependence than the tail copula, ensuring that our tests have power against more subtle misspecifications.

The PA-tail copula can be estimated non-parametrically by replacing population quantities with empirical counterparts:

\[
\tilde{\Lambda}_n^{(d)}(x, y) = \frac{1}{k} \sum_{i=d+1}^n I\left\{ |\varepsilon_i| > |\varepsilon_{i-1,d}|, |\varepsilon_{i-1,d}| > |\varepsilon_{i-1,d+1}| \right\},
\]

where \( |\varepsilon_i| \geq \cdots \geq |\varepsilon_n| \) denote the order statistics, \( I \) is the indicator function, and \( \lfloor \cdot \rfloor \) rounds to the nearest smaller integer. This estimator allows us to detect deviations from independence, where \( \Lambda_n^{(d)}(x, y) = \Lambda_n^{(d), \text{ind}}(x, y) \), via the test statistic

\[
\tilde{F}_n^{(d)}(x, y) = \frac{n}{xy} \sum_{d=1}^D \left[ \tilde{\Lambda}_n^{(d)}(x, y) - (k/n)xy \right]^2.
\]

Here, \( (k/n)xy \) nonparametrically “estimates” \( \Lambda_n^{(d), \text{ind}}(x, y) \). (Under Assumption 1 (ii) imposed below, we even have \( \Lambda_n^{(d), \text{ind}}(x, y) = (k/n)xy \). The feasible residual-based analog
of \( \hat{F}_n(x, y) \) is explored in Section 3.1. Since independence implies \( \Lambda_n^{(d)}(x, y) = \Lambda_n^{(d, \text{ind})}(x, y) \) for all \( x, y > 0 \), Section 3.2 also considers (residual-based) test statistics based on functionals of \( (x, y) \mapsto \Lambda_n^{(d)}(x, y) - (k/n)xy \).

### 2.2. Location–Scale Model

Denote by \( Y_1, \ldots, Y_n \) the observations of interest to which some parametric model will be fitted. For instance, the \( Y_1, \ldots, Y_n \) may denote log-returns on some speculative asset. We define the \( \sigma \)-field generated by \( Y_1, Y_{1-1}, \ldots \) and some exogenous (possibly multivariate) variables \( x_1, x_{1-1}, \ldots \) by \( \mathcal{F}_t = \sigma(Y_1, Y_{1-1}, \ldots, x_1, x_{1-1}, \ldots) \). We assume the \( Y_t \) to follow the conditional location–scale model

\[
Y_t = \mu_t(\theta^o) + \sigma_t(\theta^o) e_t, \tag{2}
\]

where \( \theta^o \in \Theta \) is the true parameter vector from the parameter space \( \Theta \subset \mathbb{R}^m \) \( (m \in \mathbb{N}) \), \( \mu_t(\theta^o) \), and \( \sigma_t(\theta^o) \) are the \( \mathcal{F}_t \)-measurable conditional mean and volatility, and the \( e_t \) are strictly stationary with mean zero and unit variance.

In the following, we require a smoothness condition on the tail of the \( |e_t| \).

**Assumption 1.** The df \( F(\cdot) \) of the \( |e_t| \) satisfies:

(i) The upper endpoint of \( F(\cdot) \) is infinite.

(ii) There exists some constant \( C_F > 0 \) such that \( F(\cdot) \) is differentiable with density \( f(\cdot) \) on \( \{C_F, \infty\} \).

(iii) \( \lim_{x \to -\infty} x f(x)/(1 - F(x)) = \alpha \in (0, \infty) \).

Assumption 1 is known as one of the von Mises’ conditions (see, e.g., de Haan and Ferreira 2006, Rem. 1.1.15) and ensures a sufficiently well-behaved tail of the \( |e_t| \). This is needed to justify the replacement of \( |e_t| \) in our test statistic with the standardized residuals. In their autocorrelation-based diagnostic test, Berkes, Horváth, and Kokoszka (2003) impose a similar regularity condition; see in particular their eq. (2.4). Assumption 1 is only needed to ensure the validity of Lemma 1 in Appendix B.1, supplementary materials. Thus, it may be replaced by any other assumption on the tail decay, as long as the conclusions of Lemma 1 remain valid. For instance, it may be replaced by imposing second-order regular variation on \( F(\cdot) \) (de Haan and Ferreira 2006).

Next, we require a sufficiently precise estimator of the unknown \( \theta^o \) in the sense of

**Assumption 2.** There exists an estimator \( \hat{\theta} \) satisfying \( n^k |\hat{\theta} - \theta^o| = O_P(1) \), as \( n \to \infty \), where \( k \in (0, 1/2) \) and \( |\cdot| \) denotes the Euclidean norm.

The standard case of \( \sqrt{n} \)-consistent estimators is recovered for \( k = 1/2 \) in Assumption 2. Such estimators are available under various regularity conditions for ARMA–GARCH and GARCH-type models (e.g., Francq and Zakoïan 2010). However, it may happen that \( \mathbb{E}[e_t^2] = 0 < \infty \) and \( \mathbb{E}[\epsilon_t^2] = \infty \), such as for exchange rates in Zheng, Li, and Li (2018, sec. 6) or for crude oil returns in Zhu and Ling (2011, sec. 5). Then, estimators may only be \( n^k \)-convergent with some \( k \in (0, 1/2) \); see, for example, Hall and Yao (2003) for the quasi-maximum likelihood estimator (QMLE) in GARCH models and Hill (2015a) for a tail-trimmed QMLE. In such a case, the additional generality afforded by Assumption 2 is beneficial.

Once the parameters have been estimated, volatility \( \hat{\sigma}_t = \hat{\sigma}_t(\hat{\theta}) \), the conditional mean \( \hat{\mu}_t = \hat{\mu}_t(\hat{\theta}) \) and the standardized residuals \( \hat{\epsilon}_t = \hat{\epsilon}_t(\hat{\theta}) = (Y_t - \hat{\mu}_t(\hat{\theta}))/\hat{\sigma}_t(\hat{\theta}) \) can be computed. Note that \( \mu_t(\theta) \) and \( \sigma_t(\theta) \) may depend on the infinite past of \( Y_t \) (and, possibly, \( x_t \)) and, hence, can only be approximated via \( \hat{\mu}_t(\hat{\theta}) \) and \( \hat{\sigma}_t(\hat{\theta}) \) using some artificial initial values; see, for example, the truncated recursion (A.2) in Appendix A, supplementary materials.

**Assumption 3.** There exists some neighborhood \( \Theta^o \) of \( \theta^o \), such that:

(i) \( \min_{\theta \in \Theta^o} \sigma_t(\theta) \geq \epsilon > 0 \) almost surely (a.s.).

(ii) \( \mu_t(\theta) \) and \( \sigma_t(\theta) \) are differentiable with respect to \( \theta \) in \( \Theta^o \) a.s., with derivatives \( \hat{\mu}_t(\theta) = \partial \mu_t(\theta)/\partial \theta \) and \( \hat{\sigma}_t(\theta) = \partial \sigma_t(\theta)/\partial \theta \).

(iii) \( \mathbb{E}[\sup_{\theta \in \Theta} |\hat{\mu}_t(\theta) - \mu_t(\theta)|] < \infty \) and \( \mathbb{E}[\sup_{\theta \in \Theta} |\hat{\sigma}_t(\theta) - \sigma_t(\theta)|] < \infty \) for some \( \nu > 1/k \) and for all \( t \in \mathbb{Z} \).

**Assumption 4.** Let \( \ell_n \to \infty \) with \( \ell_n < n \) and \( \ell_n = o(k/\sqrt{n}) \) denote the truncation. Then:

(i) \( \hat{\mu}_t(\theta) \) and \( \hat{\sigma}_t(\theta) \) are measurable with respect to \( \mathcal{F}_{t-1} \) for all \( \theta \in \Theta^o \) and all \( t \in \mathbb{Z} \).

(ii) \( \sum_{t=\ell_n}^{n} \mathbb{E}[\sup_{\theta \in \Theta} |\hat{\mu}_t(\theta) - \mu_t(\theta)|] = o(1) \) and \( \sum_{t=\ell_n}^{n} \mathbb{E}[\sup_{\theta \in \Theta} |\hat{\sigma}_t(\theta) - \sigma_t(\theta)|] = o(1) \).

Assumption 3 imposes standard regularity conditions on the location–scale model; see, for example, Hong and Lee (2003, Assumption A.3) or Escanciano (2010, Assumption A2).

Assumption 4 provides a bound on the effect of information truncation, and is even weaker than Assumption A5 in Escanciano (2006) or Assumption A4 in Escanciano (2010). Assumptions 2–4 jointly ensure that the residuals \( \hat{\epsilon}_t \) approximate the true innovations sufficiently well for \( t = \ell_n, \ldots, n \). Note that the first few residuals for \( t < \ell_n \) are often imprecise, due to initialization effects caused by using artificial initial values in the mean and variance recursions.

Of course, Assumptions 3 and 4 are high-level conditions that must be verified for each model on a case-by-case basis. For instance, Escanciano (2010, sec. 3) verifies even more stringent conditions than imposed in Assumptions 3 and 4 for ARMA–GARCH models under some regularity conditions. We additionally check Assumptions 3 and 4 for APARCH models in Appendix A, supplementary materials.

### 3. Main Results under the Null

#### 3.1. A Portmanteau-Type Test for Residual Extremal Dependence

For reasons pointed out in the Motivation, we are interested in verifying whether the observations \( Y_1, \ldots, Y_n \) are generated by some parametric model (2) with iid innovations. Formally:

\[ H_0 : \text{The innovations } \{e_t\}_{t \in \mathbb{Z}} \text{ in the parametric model (2) are iid.} \]
When $H_0$ holds, we write $e_t \overset{iid}{\sim} (0, 1)$ for short. Typically, $H_0$ is tested by fitting the specific model (2) to the $Y_1, \ldots, Y_n$. Under $H_0$, the residuals should then be approximately iid, which is verified by testing some implication of the iid property. This implication can be either very broad or rather specific. The test of Du and Escanciano (2015) provides an example of the former case, as it is based on the implication that
\[
P(e_t \leq x, e_{t-j} \leq y) - P(e_t \leq x)P(e_{t-j} \leq y) = 0
\]
\[
\forall j = 1, 2, \ldots, \forall (x,y) \in \mathbb{R}^2.
\]
However, when a specific alternative is entertained, one may also focus on a more narrow implication of $H_0$, such as the absence of autocorrelation in Ljung–Box-type tests (e.g., Carbon and D). poorly, suggesting some form of misspecification in the tails; see used for (out-of-sample) tail risk forecasting they often perform sample) specification tests don’t provide evidence against the motivation comes from the observation that classic (in-

As pointed out in the Motivation, we test $H_0$ with the specific alternative in mind that the left-over serial dependence in the residuals is particularly strong in the extremes. Once again, the motivation comes from the observation that classic (in-sample) specification tests often do not provide evidence against some fitted GARCH-type model, yet when the fitted models are used for (out-of-sample) tail risk forecasting they often perform poorly, suggesting some form of misspecification in the tails; see also the empirical application in Section 6 for evidence on this.

As a measure of the serial extremal dependence in the residuals, we use the PA-tail copula in (1), which equals $(k/n)xy$ under serial independence (and Assumption 1(ii)). Hence, we test the implication of $H_0$ that $\Lambda_n^{(d)}(x,y) = (k/n)xy$ for $d = 1, 2, \ldots$. Indeed, when the $e_t$ are serially dependent, then, for fixed $n \in \mathbb{N}$, there necessarily exist some $d \in \mathbb{N}$ and $x, y > 0$, such that $\Lambda_n^{(d)}(x,y) \neq (k/n)xy$. This gives our test power against deviations from the null, though not from all deviations (i.e., it is not an omnibus test) because the specific lag $d$ and the specific values of $x, y > 0$ may not be included in the test statistic.

For purposes of model checking, the estimator $\Lambda_n^{(d)}(x,y)$ from Section 2.1 is infeasible, as the $e_t$ are unobserved. Hence, we rely on the feasible counterpart
\[
\hat{\Lambda}_n^{(d)}(x,y) = \frac{1}{k} \sum_{t = d+1}^n \mathbb{I}\{x_{t-d} \geq \hat{y}_{t-d} > \hat{y}_{t-d+1}, \hat{y}_{t-d} > \hat{y}_{t-d+1}, \hat{y}_{t-d+1} > \hat{y}_{t-d+2}\}.
\]

With this estimator, we can detect deviations from $\Lambda_n^{(d)}(x,y) = (k/n)xy$ using the Portmanteau-type test statistic
\[
P_n^{(d)}(x,y) = \frac{n}{xy} \sum_{d=1}^D \left[ \hat{\Lambda}_n^{(d)}(x,y) - (k/n)xy \right]^2,
\]
where $D \in \mathbb{N}$ is some fixed user-specific integer. Once again, the fact that we compare $\Lambda_n^{(d)}(x,y)$ with the null hypothetical value of the PA tail copula $(k/n)xy$ (and not with the null hypothetical value of the tail copula, i.e., 0) gives our test power in situations, where there is independence in the limit (as measured by the tail copula), but not independence pre-asymptotically (as measured by the PA-tail copula).

As pointed out above, by testing $H_0$ via $P_n^{(D)}(x,y)$, we have a very specific alternative in mind, namely that the misspecification is largely in the serial extremal dependence. When this is true, a test based on $P_n^{(D)}(x,y)$ can be expected to have more power than alternative tests focusing on the complete dependence structure. However, by construction, $P_n^{(D)}(x,y)$ cannot detect all deviations from the complete independence null. Thus, our test should be viewed as a complement to extant "omnibus" specification tests, not a substitute.

In choosing $D$, there is the usual tradeoff. Using a small $D$ possibly leads to undetected misspecifications at higher lags, resulting in a loss of power. However, choosing $D$ too large, the estimates $\hat{\Lambda}_n^{(d)}(x,y)$ may be based on too few observations, thus, distorting size. We explore the choice of $D$ in detail in the simulations in Appendix E.2, supplementary materials. There, we show that $D = 5$ typically leads to a good balance between size and power. Additionally, in Appendix F of the supplementary materials, we adapt the automatic lag-length choice of Escanciano and Lobato (2009) to our test statistic $P_n^{(D)}(x,y)$ (and also to next section’s $\hat{P}_n^{(D)}$). While size of the test statistics with the data-driven choice of $D$ is good, power may suffer; see Appendix F.3, supplementary materials.

Regarding the sequence $k$, we impose

**Assumption 5.** For $n \to \infty$, the sequence $k$ satisfies $k \to \infty$, $k/\sqrt{n} \to \infty$, and $k = o(n^{2\kappa-\delta})$ for some $\delta > 0$ and $\kappa \in (0, 1/2)$ from Assumption 2.

Consistent with the notion of extremal dependence, the requirement that $k = o(n^{2\kappa-\delta})$ ensures that only a vanishing fraction of the residuals is used in estimation. In the standard case where $\kappa = 1/2$ this is close to $k/n \to 0$, which is the standard condition placed on $k$ in extreme value theory (de Haan and Ferreira 2006). On the other hand, sufficiently many extremes need to be used, as specified by $k/\sqrt{n} \to \infty$. In the related task of estimating the tail event correlation $r^{(d)}_n = \frac{n}{D} \mathbb{P} \{ |e_t| > b(n/k), |e_{t-d}| > b(n/k) \} - \left\{ P\{ |e_t| > b(n/k) \} P\{ |e_{t-d}| > b(n/k) \} \right\}$. Hill (2011b, sec. 5.2) even requires $k/n^{2/3} \to \infty$. Therefore, assuming $k/\sqrt{n} \to \infty$ may be regarded as a rather mild requirement. Yet, in light of $k = o(n^{2\kappa-\delta})$, it restricts $k$ to lie in the interval $(1/4, 1/2]$.

Before stating our first main theoretical result under the null, we require one final standard condition under $H_0$, that has also been used elsewhere (e.g., Du and Escanciano 2015, Assumption A.3).

**Assumption 6.** $e_t$ is independent of $\mathcal{F}_{t-1}$.

**Theorem 1.** Suppose Assumptions 1–6 are met. Then, under $H_0$, it holds for all $x, y > 0$ that, as $n \to \infty$,
\[
P_n^{(D)}(x,y) \overset{d}{\to} \chi^2_D,
\]
where $\chi^2_D$ denotes the $\chi^2$-distribution with $D$ degrees of freedom.
We defer the proof of Theorem 1 to Appendix B, supplementary materials. The proof reveals that the same $\chi^2$-limit as in Theorem 1 is obtained, if the $\hat{r}_i$'s are replaced with the true $\varepsilon_i$'s in $P^{(D)}_n(x, y)$. Thus, the effects of parameter estimation in $\hat{r}_i = \hat{r}(\theta)$ vanish asymptotically. In the context of residual-based estimation of univariate extremal quantities (such as the tail index or large quantiles), such vanishing estimation effects have been observed before (Hill 2015b; Hoga 2019). The explanation is that the convergence rate of the parameter estimator is faster than that of the tail quantity, leading to vanishing parameter estimation effects.

Hill (2009, sec. 5.2) develops a remarkably general limit theory for what he terms the joint/marginal tail probability discrepancy, which is closely related to the quantities $\hat{\Lambda}^{(d)}_n(x, y) - (k/n)xy$ appearing in $P^{(D)}_n(x, y)$. However, his theory only covers observables, not model residuals. Indeed, the bulk of the work in proving our theorems lies in showing that the limit theory is unchanged when replacing the $\varepsilon_i$'s with the $\hat{r}_i$'s.

For $x = y = 1$ the PA-tail copula is the most prominent version of Davis and Mikosch’s (2009) PA-extremogram. Thus, the choice $x = y = 1$ is the leading case for applications of $P^{(D)}_n(x, y)$. However, other choices of $x$ and $y$ may lead to different test results. Thus, the next section derives a test that combines the evidence across different values of $x$ and $y$. This may also lead to more powerful specification tests, as we will see in the simulations.

### 3.2. A Functional Portmanteau-Type Test

To obtain powerful functional versions of $P^{(D)}_n(x, y)$, we need to integrate it over a suitable area in the $(x, y)$-space. Observe that $\hat{\Lambda}^{(d)}_n(x, y)$ estimates $\Lambda^{(d)}_n(x, y)$, which converges to the tail copula $\Lambda^{(d)}(x, y)$ (if it exists) for $n \to \infty$. Theorem 1(ii) of Schmidt and Stadtmüller (2006) implies that the tail copula is homogeneous, that is, it satisfies $\Lambda^{(d)}(tx, ty) = t^{\alpha} \Lambda^{(d)}(x, y)$ for all $t > 0$. This suggests that $\hat{\Lambda}^{(d)}_n(tx, ty) \approx t^{\alpha} \hat{\Lambda}^{(d)}_n(x, y)$. Hence, it is sufficient to integrate over some sphere $S_c(\|\cdot\|) := \{(x, y) : \| (x, y) \| = c\}$, because all other estimates can be inferred from extrapolation using homogeneity, thus, providing no additional information against the null. Here, $c > 0$ is a constant and $\|\cdot\|$ some norm on $\mathbb{R}^2$. A typical choice in extreme value theory is the $\|\cdot\|_1$-norm. Since—as pointed out above—putting $x = y = 1$ is a natural choice (and hence to be included in the sphere) and $\{(1, 1)\} \subset S_c(\|\cdot\|_1)$, we fix $c = 2$ and, thus, integrate over the sphere $S_2(\|\cdot\|_1)$. This amounts to integrating over $(2 - 2z, 2z) \times (2 - 2\varepsilon, 2\varepsilon)$, such that $(2 - 2z, 2z) \subset S_2(\|\cdot\|_1)$. Using the $\|\cdot\|_1$-norm in defining the sphere also has the added advantage of giving a limiting distribution in terms of standard Brownian bridges in Theorem 2.

These considerations lead to the functional Portmanteau-type test statistic

$$F^{(D)}_n = \frac{1}{2} \sum_{d=1}^D \int_{\{1-|t|\}} \left[ \hat{\Lambda}^{(d)}_n(2 - 2z, 2\varepsilon) - \frac{k}{n} (2 - 2\varepsilon)2z \right] dz,$$

where $t \in (0, 1/2)$ is some small constant chosen by the practitioner. The parameter $t$ bounds the $z$-values in the integral away from 0 and 1, where our nonparametric estimator $\hat{\Lambda}^{(d)}_n(2 - 2\varepsilon, 2z)$ may be unreliable as very extreme quantile thresholds are considered, above which hardly any observations lie. Unreported simulations show that the test results are quite robust with respect to the choice of $t$. In our numerical experiments, we choose $t = 0.1$ and compute the integral in $F^{(D)}_n$ using standard numerical integration techniques, which allow to approximate integrals with any required accuracy.

**Theorem 2.** Suppose Assumptions 1–6 are met. Then, it holds under $H_0$ that, as $n \to \infty$,

$$\frac{F^{(D)}_n}{d} \to 2 \sum_{d=1}^D \int_{\{1-|t|\}} B^2(dx) dz,$$

where $\{B_t(\cdot)\}_{t=1,...,D}$ are mutually independent standard Brownian bridges.

**Remark 1.** The test statistics $P^{(D)}_n(x, y)$ and $F^{(D)}_n$ are of a Portmanteau-type. However, one may also consider (e.g.) max statistics $P^{(D)}_{n, s,\alpha}(x, y) = \max_{d=1,...,D} \left[ \hat{\Lambda}^{(d)}_n(x, y) - (k/n)xy \right]^{2s}$ and $F^{(D)}_{n, s,\alpha} = \max_{d=1,...,D} \int_{\{1-|t|\}} \left[ \hat{\Lambda}^{(d)}_n(2 - 2z, 2\varepsilon) - \frac{k}{n} (2 - 2\varepsilon)2z \right]^{2s} dz$ (Hill and Motegi 2020). Unreported simulations (available upon request) show that our Portmanteau-type test statistics have similar size as the max-type statistics, yet slightly larger power. One explanation for this may be that misspecified models generally do not produce serially dependent residuals at a single lag only (which would favor max-type statistics), but rather they tend to produce serial dependence in the residuals at several lags (favoring Portmanteau-type statistics).

**Remark 2.** The test statistics $P^{(D)}_n(x, y)$ and $F^{(D)}_n$ are based on absolute residuals. However, in some specific applications—such as forecasting VaR or ES—only one tail is of interest. In such cases, one could employ tests based on (say) $\hat{\varepsilon}_i < 0$ or $\hat{\varepsilon}_i > 0$ instead of $\hat{\varepsilon}_i$. Under easy adaptations of the test statistics and of our main assumptions, the limit theory would still go through with identical limits.

**Remark 3.** The idea behind $F^{(D)}_n$ is to combine evidence across different $(x, y)$-values in $P^{(D)}_n(x, y)$. For the moment, make explicit the dependence of $P^{(D)}_n(x, y)$ on $k$ by writing $P^{(D)}_{n, k}(x, y)$. Then, it is easy to see that $P^{(D)}_{n, s,\alpha}(x, y) = P^{(D)}_{n, s,\alpha}(x, y)$ for any $s > 0$. Thus, by integrating suitably over the $(x, y)$-space, we can also combine evidence across different values of $k$. This extension would be similar in spirit to the smoothed Hill estimator proposed by Resnick and Stărică (1997). We leave out the details for brevity.
4. Main Results under Local Alternatives

As mentioned in the Motivation, there are specification tests focusing on the mean dynamics, mean and variance dynamics, and the iid property of the innovations (such as our tests). For the former two types of tests, (local) power analyses are standard (Escanciano 2006, 2008, 2010; Delgado and Velasco 2011; Velasco and Wang 2015). However, when testing against deviations from iid innovations, the alternative is very broad and in general no natural alternative suggests itself. Therefore, only very specific alternatives have been considered; see Du and Escanciano (2015, Rem. 3) and Zheng, Li, and Li (2018, sec. 3). This may explain why the (local) power of iid specification tests is often not explored theoretically (Li and Mak 1994; Ling and Li 1997; Hidalgo and Zaffaroni 2007; Fisher and Gallagher 2012). Nonetheless, here we carry out a (local) power analysis to give a sense of the situations in which our tests can be expected to work well. Our specific local alternative is

$$H_{a,n} : \Lambda_{n}^{(d)}(x,y) - (k/n)xy = n^{-1/2}r_{d}(x,y)$$

for all $x, y > 0$, $d \in \mathbb{N}$.

To investigate our test statistics under $H_{a,n}$, we impose the following conditions.

**Assumption 7.** The $\{\varepsilon_{i}\}_{i \in \mathbb{Z}}$ are strictly stationary and $d^{*}$-dependent for some $d^{*} \in \mathbb{N}$.

**Assumption 8.** For all $x > 0$, it holds that $P[|\varepsilon_{i}| > b(n^{1/2}) \mid F_{t-1}] = \left[ 1 - \frac{\theta}{\sqrt{k_{c}n}} \right]^{k_{c}} + \frac{\theta}{\sqrt{k_{c}n}} R(x \mid F_{t-1})$, where $\theta$ is some fixed constant, $E|R(x \mid F_{t-1})|^{2} < \infty$ and $|R(x \mid F_{t-1}) - R(y \mid F_{t-1})| \leq C_{t-1}|x - y|$ with $C_{t-1} \leq 0$ uniformly bounded in $t$.

Since the $\{\varepsilon_{i}\}$ are 0-dependent under the null, it seems natural to only impose $d^{*}$-dependence under the alternative in Assumption 7 instead of some mixing condition. This is because under mixing, serial dependence may even be present at arbitrarily large lags, such as would be “further away” from the null. Note that $d^{*}$-dependence implies $r_{d}(x, y) \equiv 0$ for $d > d^{*}$. Assumption 8 replaces Assumption 6 and is similar in spirit to the local contiguity requirement in Rothe and Wied (2013, Ass. 6) in the context of testing conditional distribution models. Under independence of $\varepsilon_{i}$ from $F_{t-1}$, we have $R(x \mid F_{t-1}) = x$. Under the alternative, the magnitude $r_{d}(x, y)$ of the local alternative can be related to certain moments of the $R(\cdot \mid F_{t-1})$-function from Assumption 8. Specifically, $r_{d}(x, y) = \theta \{\bar{r}_{d}(x, y) - xy\}$ for $\bar{r}_{d}(x, y) := (n/k)E[1_{|\varepsilon_{i} - d| > b(\sqrt{n})}]R(x \mid F_{t-1})$; see Lemma 8 in Appendix D, supplementary materials.

With our Assumptions 7–8 we have not aimed for full generality. However, they allow the proof techniques from Theorems 1–2 to be easily generalized to the present setup. Strictly speaking, the $\varepsilon_{i}$ in Assumptions 7–8 form an array of random variables under $H_{a,n}$. Yet, we suppress the dependence of the $\varepsilon_{i}$ on $n$ for notational ease.

**Theorem 3.** Suppose Assumptions 1–5 and 7–8 are met. Then, under $H_{a,n}$:

(i) For all $x, y > 0$, it holds that, as $n \to \infty$,

$$P_{n}^{(d)}(x, y) \to \chi_{D}^{2} \left( \sum_{d=1}^{D} \left[ r_{d}(x, y) \right]^{2}/xy \right),$$

where $\chi_{D}^{2}(\cdot)$ denotes the $\chi^{2}$-distribution with $D$ degrees of freedom and noncentrality parameter $c > 0$.

(ii) For any $i \in (0, 1/2)$, it holds that, as $n \to \infty$,

$$P_{n}^{(d)} \to 4 \sum_{d=1}^{D} \int_{|z| \leq 1} \left[ B_{d}(z) - c_{d}(2 - 2z, 2z) \right]^{2} dz,$$

where $\{B_{d}(\cdot)\}_{d=1, \ldots, D}$ are mutually independent standard Brownian bridges.

It may seem unrealistic to assume an $n^{1/2}$-convergent estimator in Assumption 2 also under the alternative. However, for example, Escanciano (2009) shows that GARCH parameters can still be estimated $\sqrt{n}$-consistently when the $\varepsilon_{i}$ are certain martingale differences.

**Theorem 3** shows that our tests have nontrivial asymptotic local power (ALP) in $n^{-1/2}$-neighborhoods of the null hypothetical value $r_{d}(x, y) \equiv 0$. For $P_{n}^{(d)}(x, y)$ this is implied by $P(\chi_{D}^{2}(\cdot) > \chi_{D,1-\alpha}^{2}) \geq \alpha$, where $\chi_{D,1-\alpha}^{2}$ denotes the $(1 - \alpha)$-quantile of the $\chi_{D}^{2}$-distribution. For $P_{n}^{(d)}$, this follows from Anderson’s lemma (Anderson 1955, Cor. 2), because $B_{d}(\cdot)$ is a zero-mean Gaussian process, such that the ALP is bounded from below by

$$P \left[ 4 \sum_{d=1}^{D} \int_{|z| \leq 1} \left[ B_{d}(z) - c_{d}(2 - 2z, 2z) \right]^{2} dz \geq c_{d}^{(D)} \right] \geq P \left[ 4 \sum_{d=1}^{D} \int_{|z| \leq 1} B_{d}^{2}(z) dz \geq c_{d}^{(D)} \right] = \alpha;$$

see Rothe and Wied (2013, Proof of Thm. 3). We mention that consistency against fixed alternatives may also be shown, similarly as in Appendix F of the supplementary materials for the tests based on the data-driven choice of $D$. We leave details for brevity. Yet, from Theorem 3 it follows intuitively that as the magnitude of the local alternative diverges, the ALP converges to one.
Remark 4. As mentioned above, specification tests for the iid assumption often do not explore local power. When they do, as in Zheng, Li, and Li (2018) and our Theorem 3, the alternatives are tailored to the testing problem at hand, such that a direct comparison of the ALP is not possible. We also mention that, similar to Zheng, Li, and Li (2018), our tests focus on iid innovations and, hence, have power not only against deviations from idness, but also against incorrect specification of the mean and volatility model (because a misspecified mean / volatility filter also leads to serially dependent residuals). Here, we have only investigated the former deviations, but an exploration of the latter deviations is also possible; for example, along the lines of Zheng, Li, and Li (2018, sect. 3).

5. Simulations

5.1. Preliminaries

Here, we investigate size and power of our specification tests based on $P_n^{(D)}$ (with the standard choice $x = y = 1$) and $F_n^{(D)}$. We also compare the results with a test using the classical Ljung–Box statistic, $LB_n^{(D)}$, which is based on the autocorrelations of the squared residuals. We choose the Ljung–Box test as one of our competitors here for two reasons. First, due to its popularity in both applied work (see, e.g., Oh and Patton 2018, sec. 4.3) and theoretical work (Francq and Zakoïan 2010; Carbon and Francq 2011). And, second, because the comparison with another Portmanteau-type test is particularly transparent. However, it is well-known that the Ljung–Box test suffers from low power in certain situations. Hence, we also use the specification test of Hong and Lee (2003). The test statistic, $HL_n$, is based on Hong’s (1999) generalized spectral density approach.

We use several samples sizes $n \in \{500, 1000, 2000\}$ and we fix the significance level at $\alpha = 5\%$. Furthermore, we pick $\ell = 0.1$. Other choices of $\ell$ (e.g., $\ell = 0.05$ or even $\ell = 0.01$) do not materially change the results. All simulations use R version 4.1.1 (R Core Team 2021) and results are based on 10,000 replications.

The choice of $D$ is common to all tests (except for the Hong and Lee (2003) test) and we pick $D = 5$ in the following. This choice is supported by simulations in Appendix E.2, supplementary materials, where we investigate the influence of $D$ on the tests. Note that even the procedures that choose $D$ endogenously rely on a simulation-based recommendation for tuning parameters. For example, Escanciano and Lobato (2009) and Hill and Motegi (2020) rely on the choice of $q$ in their eqs. (4) and (13), respectively. They then resort to simulation evidence to justify their respective choices ($q = 2.4$ for Escanciano and Lobato (2009) and $q = 3$ for Hill and Motegi (2020)). So we also follow that tradition by inferring a simulation-based recommendation for “our” tuning parameter $D$.

Here, instead of using various $D$, we explore the effect of different $k$'s on our test statistics $P_n^{(D)}$ and $F_n^{(D)}$. In applications of extreme value methods the choice of $k$ is often a tricky issue, because the results may be very sensitive to the specific value of $k$ (de Haan and Ferreira 2006). Therefore, it is of interest to investigate the robustness of our tests with respect to $k$. A very popular choice in applications of extreme value theory is to follow DuMouchel’s (1983) rule by setting $k = \lfloor 0.1 \cdot n \rfloor$ (McNeil and Frey 2000; Quinots, Fan, and Phillips 2001; Chavez-Demoulin, Embrechts, and Sardy 2014). However, at least asymptotically, this is not a valid choice (here and elsewhere) since it fails the Assumption 5 requirement that $k$ be a vanishing fraction of $n$. Nonetheless, its widespread use suggests some merit in finite samples. Thus, we consider $k = \lfloor \rho n^{0.99} \rfloor$ for $\rho \in [0.05, 0.15]$, where the choice $\rho = 0.11$ roughly corresponds to DuMouchel’s (1983) rule. Since we consider $k = \lfloor \rho n^{0.99} \rfloor$ with $\rho \in [0.05, 0.15]$, the $k$'s are between $[23, 70]$ / $[46, 139]$ / $[92, 278]$ for $n = 500 / n = 1000 / n = 2000$. These values may be regarded as being sufficiently small relative to the sample size for extreme value methods to be applicable. Note that $k = \lfloor \rho n^{0.99} \rfloor$ satisfies both Assumption 5 requirements, because the estimators we consider throughout the simulations are $\sqrt{n}$-convergent, such that $\kappa = 1/2$ in Assumption 5.

For simplicity, we only consider model (2) with zero conditional mean, $\mu_t(\theta^0) \equiv 0$. We do so, because unmodeled dynamics in the tails of the conditional distribution are more likely to be caused by misspecified volatility dynamics in practice rather than misspecified mean dynamics. Since we simulate from ARARCH(1,1) models under the null, our high-level Assumptions 3 and 4 are satisfied; see Appendix A, supplementary materials.

We compare our tests with a classical Ljung–Box test and the Hong and Lee (2003) test when volatility dynamics are misspecified (Section 5.2) and for non-iid $\epsilon_t$ (Appendix E.1., supplementary materials). As pointed out above, Theorem 3 only investigates power in the latter case, but a misspecified volatility filter in the former case also leads to serial dependence in the standardized residuals. Throughout, we estimate our models in R using the rugarch package (Ghalanos 2020).

5.2. Misspecified Volatility

We generate time series from an ARARCH model with an exogenous covariate $x_t$ in the volatility equation. Specifically, we simulate from the ARARCH–X(1,1) model

$$Y_t = \sigma_t(\theta^0) \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} (0, 1),$$

$$\sigma_t(\theta^0) = \omega^0 + \alpha^0_{+1}(Y_{t-1})_{+} + \alpha^0_{-1}(Y_{t-1})_{-} + \beta_1^0 \sigma_{t-1}(\theta^0) + \pi_1^0 x_{t-1},$$

(3)

where $\theta^0 = (\omega^0, \alpha^0_{+1}, \alpha^0_{-1}, \beta_1^0, \pi_1^0)'$ and the $\epsilon_t$ follow a (standardized) Student’s $t$-distribution with 4.1 degrees of freedom. This ensures $E[\epsilon_t^4] < \infty$, which is required for $\sqrt{n}$-consistent QML estimation in Assumption 2 (Francq and Thieu 2019). The exogenous $x_t = \exp(z_t)$ are stationary with $z_t = 0.9 \cdot z_{t-1} + \epsilon_t \overset{iid}{\sim} N(0, 1)$. For the parameters, we take $\theta^0 := (\omega^0, \alpha^0_{+1}, \alpha^0_{-1}, \beta_1^0, \pi_1^0)' = (0.046, 0.027, 0.092, 0.843, 0)$ (to compute size) and $\theta^0 := (\omega^0, \alpha^0_{+1}, \alpha^0_{-1}, \beta_1^0, \pi_1^0)' = (0.046, 0.027, 0.092, 0.843, 0.089)$ (to compute power). Irrespective of the choice of $\theta^0$, we estimate an ARARCH(1,1) model (i.e., we estimate (3) imposing $\pi_1^0 = 0$) via Gaussian QML. Thus, when $\theta^0 = \theta^0_p$ we fit a correctly specified model.

The data-generating process is taken from Francq and Thieu (2019, sec. 3.1), who obtained the parameters in (3) from fitting an ARARCH–X model to Boeing returns.
Rejection frequencies at the 5%-level under the null and alternative for $P_n^{(D)}$-test (black), $F_n^{(D)}$-test (red), $LB_n^{(D)}$-test (green), $HL_n$-test (blue) for $D = 5$ and $k = \lfloor \rho n^{0.9} \rfloor$ with $\rho \in [0.05, 0.15]$. Nominal level of 5% indicated by the dotted horizontal line. Dashed vertical line indicates value of $k = \lfloor 0.11 \cdot n^{0.99} \rfloor$. Results under the null (alternative) are for model (3) with $\theta^\circ = \theta^\circ = \theta^p$. For purposes of comparison, we also include the results of a classical Ljung–Box test (with test statistic $LB_n^{(D)}$) based on the first $D$ autocorrelations of the squared residuals. We employ the corrected Ljung–Box test statistic of Carbon and Francq (2011), which ensures that $LB_n^{(D)}$ is asymptotically $\chi^2_D$-distributed. Note that this correction is specific to APARCH models estimated via Gaussian QML. We also use the Hong and Lee (2003) test with test statistic $HL_n$ based on $p = 5$ in their notation. Under some regularity conditions, $HL_n$ is asymptotically standard normal (Hong and Lee 2003, Theorem 1).

Figure 1 displays the rejection frequencies of all four tests at the 5%-level. Both the $P_n^{(D)}$-test (black) and the $F_n^{(D)}$-test (red) have almost identical size, which—as the sample size increases—converges rapidly to the nominal level for all $k$. Larger differences in power between the two tests only emerge for large $n$. We also see that size and power are reasonably stable across different choices of $k$, suggesting that test results are quite robust to the particular value of $k$. This is encouraging given that extreme

(yielding size), yet when $\theta^\circ = \theta^p$ the fitted model is misspecified (yielding power). In the latter case, the exogenous variable serves to introduce some unmodeled dynamics in the variance equation, with infrequent outbursts of $z_t$ inducing some dependence in the tails of the residuals.\footnote{The second-to-top panel of Figure 6 in Appendix E.2 of the supplementary materials shows a representative trajectory of $z_t$.} Hence, our simulation setup is well-suited to $P_n^{(D)}$ and $F_n^{(D)}$ and, at the same time, relevant in practice as exogenous volatility shocks are often empirically plausible; see Han and Kristensen (2014) for references.

For $v = 10$, we generate $\{Y_t\}_{t=-v+1,\ldots,n}$ from (3). Since volatility estimates $\hat{\sigma}_t(\hat{\theta})$ may be imprecise for the first few $t$ due to initialization effects in the variance equation (see also Hall and Yao 2003), we discard the first $v$ standardized residuals and only consider $\{\hat{e}_t = Y_t/\hat{\sigma}_t(\hat{\theta})\}_{t=1,\ldots,n}$ in our test statistics $P_n^{(D)}$ and $F_n^{(D)}$. For purposes of comparison, we also include the results of a classical Ljung–Box test (with test statistic $LB_n^{(D)}$) based on the first $D$ autocorrelations of the squared residuals. We employ the corrected Ljung–Box test statistic of Carbon and Francq (2011), which ensures that $LB_n^{(D)}$ is asymptotically $\chi^2_D$-distributed. Note that this correction is specific to APARCH models estimated via Gaussian QML. We also use the Hong and Lee (2003) test with test statistic $HL_n$ based on $p = 5$ in their notation. Under some regularity conditions, $HL_n$ is asymptotically standard normal (Hong and Lee 2003, Theorem 1).

Figure 1 displays the rejection frequencies of all four tests at the 5%-level. Both the $P_n^{(D)}$-test (black) and the $F_n^{(D)}$-test (red) have almost identical size, which—as the sample size increases—converges rapidly to the nominal level for all $k$. Larger differences in power between the two tests only emerge for large $n$. We also see that size and power are reasonably stable across different choices of $k$, suggesting that test results are quite robust to the particular value of $k$. This is encouraging given that extreme

}\end{figure}
value methods can be very sensitive to the choice of the cutoff. Our approximation \( k = \lfloor 0.11 \cdot n^{0.99} \rfloor \) to DuMouchel’s (1983) rule—indicated by the dashed vertical lines in Figure 1—yields good results in terms of both size and power. Note that for \( n = 500 \), where at first sight our heuristic does not lead to optimal power, size and power decrease with increasing \( k \), such that the size-corrected power of our proposal for the choice of \( k \) still appears very satisfactory.

Now, we compare our two tests with the LB\(_n\)\(_{(D)}\)-test (green), which is of course independent of \( k \). We find that—despite rejecting more often under the null—the LB\(_n\)\(_{(D)}\)-test rejects less often under the alternative. Hence, the size-corrected power of our tests is even larger than the differences in the right-hand side panels of Figure 1 suggest.

Finally, the HL\(_n\)-test (blue), which is again independent of \( k \), is oversized, particularly compared with our tests. Taking this into account, our tests seem to have higher power for all \( n \). However, as \( n \) increases the size differences shrink, such that power can be directly compared. Even though our tests reject less often under the null for \( n = 2000 \), they are much more likely to detect a misspecified model than the HL\(_n\)-test. Again, this translates into even higher size-corrected power of our tests than indicated by the right-hand panels of Figure 1.

To summarize the simulation results of this section and those of Appendix E.1, supplementary materials (where we consider dependent innovations as a source of misspecification), we find that size and power of our tests may depend somewhat on the choice of \( k \). However, setting \( k = \lfloor 0.11 \cdot n^{0.99} \rfloor \), which approximates DuMouchel’s (1983) rule, gives good results—irrespective of the type of misspecification. Although our tests work well for \( k = \lfloor 0.11 \cdot n^{0.99} \rfloor \), in practice we recommend to apply the tests for several values of \( k \) as a “robustness check” (see Figure 4 in the empirical application for an example), much like results for standard Ljung–Box tests are also routinely reported for several lags. Our tests also depend on the number of included lags. Even though we find that \( D = 5 \) leads to good results in Appendix E.2, supplementary materials, we again recommend—if possible—to report results for different \( D \). We generally find that the \( F_{D,n}^{(D)} \)-test is more powerful than the \( P_{D,n}^{(D)} \)-test. Yet both have markedly higher power than the LB\(_n\)\(_{(D)}\)-test and (to a lesser extent) the HL\(_n\)-test for the types of mis specifications considered here. As we have argued, these mis specifications are empirically plausible, such that our tests should have substantial merit in practice. We investigate this next.

### 6. Diagnosing S&P 500 Constituents

Consider the 500 components of the S&P 500 as of 31/5/2021. We apply the \( F_{D,n}^{(D)} \) and LB\(_n\)\(_{(D)}\)-test to the log-returns of each constituent. We do so for the recommended values \( D = 5 \), \( k = \lfloor 0.11 \cdot n^{0.99} \rfloor \) and \( \iota = 0.1 \). For brevity, we do not report results for \( P_{D,n}^{(D)} \), which showed slightly inferior performance to \( F_{D,n}^{(D)} \) in the simulations. We consider the 495 components of the S&P 500 for which we have complete data for our sample period from January 1, 2010 to December 31, 2019. We split the 10-year sample into an “in-sample” period of 8 years (used for specification testing) and an “out-of-sample” period of 2 years (used for backtesting VaR forecasts). The goal of our analysis is 2-fold. First, we want to illustrate that (similarly as in the simulations) cases may arise where \( F_{D,n}^{(D)} \) warrants a rejection, but LB\(_n\)\(_{(D)}\) does not. Our second goal is to show that in these cases, a rejection by \( F_{D,n}^{(D)} \) is not spurious, but—on the contrary—indicative of “out-of-sample” forecast failure of the model.

Specifically, we fit an APARCH(1,1) model without covariates as in (3) (i.e., with \( \pi_0^2 = 0 \)) based on the “in-sample” period, and use the fitted model for one-step-ahead VaR forecasting in the “out-of-sample” part. The VaR at level \( \theta \) is the loss at time \( t \) that is only exceeded with probability \( \theta \) given the current state of the market (embodied by some information set \( F_{t-1} \)). Formally, VaR\(_t\) is the \( F_{t-1} \)-measurable random variable satisfying

\[
P(Y_t \leq \text{VaR}_t | F_{t-1}) = \theta.
\]

The multiplicative structure in (3) implies that \( \text{VaR}_t = \sigma_t(\theta^*) F_{t-1}^{-1}(\theta) \). Thus, we can easily compute the “out-of-sample” VaR forecasts from the volatility forecasts of the fitted model (say, \( \sigma_t \)) and an estimate of \( F_{t}^{-1}(\theta) \) from the standardized residuals of the “in-sample” period. Then, we backtest the VaR forecasts using the dynamic quantile (DQ) test of Engle and Manganelli (2004) with 4 lags. The DQ test does not take into account estimation effects (Escanciano and Olmo 2010; Du and Escanciano 2017) and, hence, is a test of the whole forecasting procedure and not only of the forecasting model. We refer to Hoga and Demetrescu (in press, sec. 5) for some in-depth discussion. Particularly, Hoga and Demetrescu (in press) argue that in risk management practice it may be undesirable to factor out estimation effects. Therefore, the fact that the DQ-test does not take into account estimation effects may be seen as an advantage.

As we run two specification tests, there are four potential outcomes because the LB\(_n\)\(_{(D)}\) and the \( F_{D,n}^{(D)} \)-test can either accept (indicated by a “0” in Table 2) or reject (indicated by a “1” in Table 2). The column “Total” in Table 2 indicates the total number of times each of the four cases occurs for the 495 S&P 500 stocks. In 181 out of 495 cases, none of the tests reject. For these 181 stocks, the “out-of-sample” VaR forecasts are rejected 60 times by the DQ-test. One reason for this rather large number of rejections may be the long out-of-sample period of 2 years. Typically, models are re-estimated at least quarterly, yet we only fitted the model once based on the “in-sample” data. Nonetheless, the rather long out-of-sample period of 2 years (yielding roughly 500 daily observations) is needed for the DQ-test to have reasonable power. Table 2 further shows that for 122 stocks, \( F_{D,n}^{(D)} \) but not LB\(_n\)\(_{(D)}\) leads to a rejection. The fact that for these 122 stocks the DQ test rejects in 69 cases suggests some unmodeled dynamics in these stocks, that were picked up by our \( F_{D,n}^{(D)} \)-test, yet not by the Ljung–Box test. When both tests agree in rejecting a model (which happens in 79 cases), the model produces inadequate VaR forecasts 38 times. For the remaining

**Table 2. Results of specification and DQ tests carried out at 5%-level.**

|  | \( F_{D,n}^{(D)} \) | Total | DQ-test |
|---|---|---|---|
| 0 | 0 | 181 | 60 |
| 0 | 1 | 122 | 69 |
| 1 | 0 | 113 | 31 |
| 1 | 1 | 79 | 38 |
113 stocks, only $LB_n^{(D)}$ rejects the null, yet the DQ-test rejects only 31 times.

Overall, we find that in most of the 198 cases when the DQ-test rejects, at least one of the specification test signals problems in advance (which occurs 138 times). Of these 138 cases, 107 cases are identified by our $F_n^{(D)}$-test, yet the $LB_n^{(D)}$-test rejects in only 69 of these cases. This suggests that our extreme value-based test provides complementary information to more classical specification tests.

Table 3 complements these results by providing an analog of Table 2, where the $LB_n^{(D)}$-test is replaced by the $HL_n$-test. Again, of the 495 VaR forecasts 198 (i.e., 40%) are rejected by the DQ-test. A rejection of a specification test that has any merit in identifying poor risk forecasting models in advance must therefore imply a larger rejection probability than 40% of the subsequent DQ-test. Yet, as we can see from Table 3, of the 177 models rejected exclusively by the $HL_n$-test only 57 (i.e., 32.2%) are rejected subsequently by the DQ-test. In contrast, for our $F_n^{(D)}$-test that percentage is much higher with $16/31 \approx 51.6\%$.

When both tests agree in rejecting the model, the percentage of DQ-test rejections is highest with $92/168 \approx 54.8\%$.

We mention that the larger number of rejections by the $HL_n$-test vis-à-vis the $F_n^{(D)}$-test may have two reasons. First, it may reflect that the $HL_n$-test is oversized, for which Figure 1 displays some simulation evidence. Second, it may be that while the tail dynamics are adequately captured by the model (resulting in low frequencies of DQ-test rejections), some non-tail dynamics are not captured adequately (resulting in high frequencies of $HL_n$-test rejections). In any case, a larger number of rejections by itself is not necessarily helpful in identifying unsuitable risk forecasting models in advance.

Next, we illustrate one of the cases where $LB_n^{(D)}$ does not reject, yet $F_n^{(D)}$ does. We do so for Cisco log-returns, shown in the top panel of Figure 2. Volatility estimates are superimposed in red. During our in-sample period from 2010 to 2017 there are only mild signs of volatility clustering. A cursory inspection of the residuals in the lower panel suggests that the model successfully captures this. This is also confirmed by the plots of the squared residual autocorrelations in the top part of Figure 3, which shows that autocorrelations up to lag 10 are

| $HL_n$ | $F_n^{(D)}$ | Total | DQ-test |
|--------|-------------|-------|---------|
| 0      | 0           | 119   | 33      |
| 0      | 1           | 31    | 16      |
| 1      | 0           | 177   | 57      |
| 1      | 1           | 168   | 92      |

**Table 3.** Results of specification and DQ tests carried out at 5%-level.

![Figure 2.](image1.png)

Top: Cisco returns ($Y_t$) in black and estimated volatility ($\hat{\sigma}_t$) in red. Bottom: Standardized residuals ($\hat{\varepsilon}_t$).

![Figure 3.](image2.png)

Top: Estimates $\hat{\rho}_1^{(D)}$ of the autocorrelation at lag $d$ for the squared standardized residuals; bottom: Estimates $\hat{\lambda}_1^{(D)}(1,1)$ for the standardized residuals $\hat{\varepsilon}_t$. Dashed red lines indicate pointwise 95%-confidence intervals.
insignificant. However, focusing on the extremes, we find that the \( \hat{\Lambda}_n^{(D)} (1, 1) \) -estimates in the bottom panel show strong (positive) serial extremal dependence at all lags. This may be due to occasional exogenous shocks to the share price. For instance, the largest drop on November 11, 2010 followed a profit warning of Cisco. This and other outliers apparent from the plot of the \( Y_t \) in Figure 2 seem to have been unpredictable from past information in \( F_{t-1} \), thus, violating the modeling paradigm in (2). Similarly as for model (3) in the simulations, such exogenous shocks may have caused the serial extremal dependence in the standardized residuals.

The rejection of the APARCH(1,1) model by the \( F_n^{(D)} \) -test implies that there is some misspecification either in the conditional mean and/or the conditional variance and/or the innovation dynamics. To identify the precise source of failure, one may employ tests for a correctly specified conditional mean / variance model, such as that of Escanciano (2008). If these do reject, then more complex conditional mean / variance specifications should be entertained. If, however, these do not reject, the reason for model failure must lie in unmodeled dynamics in the innovations, suggesting that an autoregressive conditional density model of Hansen (1994) may be more suitable (see Appendix E.1 of the supplementary materials for an example).

The results for the S&P 500 components in this section are for \( k = \lfloor 0.11 \cdot n^{0.89} \rfloor \). For a single time series one may want to assess the robustness of the test outcome with respect to \( k \). To do so, it is common in extreme value theory to plot the results as a function of \( k \). We illustrate this for the Cisco returns in Figure 4. There, the test statistics \( F_n^{(D)} \) and \( P_n^{(D)} \) are plotted as functions of \( k \). As in the simulations, the test results are quite robust to \( k \), with any reasonable choice leading to a rejection.

![Figure 4.](image)

**Figure 4.** \( F_n^{(D)} \) (top) and \( P_n^{(D)} \) (top) for Cisco returns as a function of \( k \). Horizontal red lines (dashed) indicate respective 5%-critical values.

### 7. Conclusion

We propose two specification tests based on the PA-tail copula of time series residuals. By relying on the PA-tail copula, our tests direct the power to any remaining serial extremal dependence and, thus, complement the evidence by more classical “omnibus”-type tests. The test statistics are easy to compute and their limiting distributions are free of nuisance parameters. Thus, our tests are simple to implement and enjoy broad applicability. This contrasts with more classical diagnostic tests, which often require involved bootstrap procedures for valid inference and/or depend on the specific estimator used to fit the model, hence, limiting practical applications. Simulations demonstrate the good size of our proposals. They also show that when a misspecified model captures well the serial dependence in the body of the distribution but not in the tail, our tests have higher power than competitor tests. The misspecified models considered in the simulations are realistic in that they display exogenous shocks in the volatility equation or unmodeled higher-order dynamics in the innovations, both of which being empirically relevant sources of misspecification (Hansen 1994; Han and Kristensen 2014). We also exemplify the better detection properties of our tests on S&P 500 constituents, where a rejection of our test is a more reliable indicator of poor out-of-sample risk predictions. Therefore, our tests can help to identify unsuitable risk forecasting models in advance, which is economically desirable under the Basel framework. Consequently, by focusing on the extremes, our tests are useful complements to standard specification tests, where valuable information on the serial extremal dependence may be “washed-out.”

### Supplementary Materials

The appendix contains the verification of Assumptions 3 and 4 for APARCH models (Appendix A), the proofs of Theorems 1–2 (Appendices B and C), the proof of Theorem 3 (Appendix D), additional simulations (Appendix E), and theoretical results along with simulations for a test with an automatic choice of \( D \) (Appendix F). Furthermore, the supplementary material contains the \( R \) code to reproduce the simulation study and the empirical application.

### Acknowledgments

The author is indebted to the Editor Christian Hansen, the Associate Editor, and three anonymous referees for their detailed comments, that significantly improved the quality of the article. The author is also grateful to seminar participants at CREST and Erasmus University Rotterdam for valuable suggestions, in particular Christian Francq, Jeroen Rombouts, Jean-Michel Zakoian and Chen Zhou. Finally, the author would like to thank Christoph Hanck and Till Massing for their careful reading of an earlier version of this manuscript.
Disclosure Statement

The author reports there are no competing interests to declare.

Funding

This work was supported by the German Research Foundation (DFG) under grant HO 6305/1-1.

ORCID

Yannick Hoga  https://orcid.org/0000-0002-6332-5561

References

Anderson, T. W. (1955), "The Integral of a Symmetric Unimodal Function over a Symmetric Convex Set and some Probability Inequalities," Proceedings of the American Mathematical Society, 6, 170–176. [1280]
Basel Committee on Banking Supervision (2019), Basel Framework, Basel: Bank for International Settlements. Available at http://www.bis.org/basel_framework/index.html#export=pdf. [1274]
Berkes, I., Horváth, L., and Kokoszka, P. (2003), "Asymptotics for GARCH Squared Residual Correlations," Econometric Theory, 19, 515–540. [1274,1277]
Box, G. E. P., and Pierce, D. A. (1970), "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," Journal of the American Statistical Association, 65, 1509–1526. [1274]
Caamaño, M., and Francq, C. (2011), "Portmanteau Goodness-of-Fit Test for Asymmetric Power GARCH Models," Austrian Journal of Statistics, 40, 55–64. [1278,1281,1282]
Chan, N. H., Deng, S. J., Peng, L., and Xia, Z. (2007), "Interval Estimation of Value-at-Risk based on GARCH Models with Heavy-Tailed Innovations," Journal of Econometrics, 137, 556–576. [1275]
Chavez-Demoulin, V., Embrechts, P., and Sardy, S. (2014), "Extreme-Quantile Tracking for Financial Time Series," Journal of Econometrics, 181, 44–52. [1281]
Chen, B., and Hong, Y. (2014), "A Unified Approach to Validating Univariate and Multivariate Conditional Distribution Models in Time Series," Journal of Econometrics, 178, 22–44. [1276]
Davis, R. A., and Mikosch, T. (2009), "The Extremogram: A Correlogram for Extreme Events," Bernoulli, 15, 977–1009. [1276,1279]
Davis, R. A., Mikosch, T., and Cribben, I. (2012), "Towards Estimating Extremal Serial Dependence via the Bootstrapped Extremogram," Journal of Econometrics, 170, 142–152. [1275,1276]
Davis, R. A., Mikosch, T., and Zhao, Y. (2013), "Measures of Serial Extremal Dependence and their Estimation," Stochastic Processes and their Applications, 123, 2575–2602. [1275]
de Haan, L., and Ferreira, A. (2006), Extreme Value Theory, New York: Springer. [1277,1278,1281]
Delgado, M. A., and Velasco, C. (2011), "An Asymptotically Pivotral Transform of the Residuals Sample Autocorrelations with Application to Model Checking," Journal of the American Statistical Association, 106, 946–958. [1276,1280]
Du, Z., and Escanciano, J. C. (2015), "A Nonparametric Distribution-Free Test for Serial Independence of Errors," Econometric Reviews, 34, 1011–1034. [1278,1280]
— (2017), "Backtesting Expected Shortfall: Accounting for Tail Risk," Management Science, 63, 940–958. [1283]
DuMouchel, W. H. (1983), "Estimating the Stable Index α in Order to Measure Tail Thickness: A Critique," The Annals of Statistics, 11, 1019–1031. [1281,1283]
Engle, R. F., and Manganelli, S. (2004), "CVAir: Conditional Autoregressive Value at Risk by Regression Quantiles," Journal of Business & Economic Statistics, 22, 367–381. [1276,1283]
Escanciano, J. C. (2006), "Goodness-of-Fit Tests for Linear and Nonlinear Time Series Models," Journal of the American Statistical Association, 101, 531–541. [1276,1277,1280]
Frank, C., Roy, R., and Zakoian, J. M. (2005), "Diagnostic Checking in ARMA Models with Uncorrelated Errors," Journal of the American Statistical Association, 100, 532–544. [1276]
Francq, C., and Thiieu, L. Q. (2019), "QML Inference for Volatility Models with Covariates," Econometric Theory, 35, 37–72. [1281]
Francq, C., and Zakoian, J. M. (2010), GARCH Models: Structure, Statistical Inference and Financial Applications, Chichester: Wiley. [1277,1281]
Ghalanos, A. (2020), rugarch: Univariate GARCH Models. R package version 1.4-2. [1281]
Hall, P., and Yao, Q. (2003), "Inference in ARCH and GARCH Models with Heavy-Tailed Errors," Econometrica, 71, 285–317. [1275,1277,1282]
Han, H., and Kristensen, D. (2014), "Asymptotic Theory for the QMLE of GARCH-X Models with Stationary and Nonstationary Covariates," Journal of Business & Economic Statistics, 32, 416–429. [1282,1285]
Hansen, B. E. (1994), "Autoregressive Conditional Density Estimation," International Economic Review, 35, 705–730. [1285]
Heffernan, J. E. (2000), "A Directory of Coefficients of Tail Dependence," Extremes, 3, 279–290. [1276]
Hidalgo, J., and Zaffaroni, P. (2007), "A Goodness-of-Fit Test for ARCH(∞) Models," Journal of Econometrics, 141, 973–1013. [1275,1280]
Hill, J. B. (2009), "On Functional Central Limit Theorems for Dependent, Heterogeneous Arrays with Applications to Tail Index and Tail Dependence Estimation," Journal of Statistical Planning and Inference, 139, 2091–2110. [1275,1279]
— (2011a), "Extremal Memory of Stochastic Volatility with an Application to Tail Shape Inference," Journal of Statistical Planning and Inference, 141, 663–676. [1276]
— (2011b), "Tail and Nontail Memory with Applications to Extreme Value and Robust Statistics," Econometric Theory, 27, 844–884. [1275,1278]
— (2015a), "Robust Estimation and Inference for Heavy Tailed GARCH," Bernoulli, 21:1629–1669. [1277]
— (2015b), "Tail Index Estimation for a Filtered Dependent Time Series," Statistical Sinica, 25, 609–629. [1275,1279]
Hill, J. B., and Motegi, K. (2020), "A Max-Correlation White Noise Test for Weakly Dependent Time Series," Econometric Theory, 36, 907–960. [1279,1281]
Hoga, Y. (2019), "Confidence Intervals for Conditional Tail Risk Measures in ARMA–GARCH Models," Journal of Business & Economic Statistics, 37, 613–624. [1275,1279]
Hoga, Y., and Demetrescu, M. (in press), "Monitoring Value-at-Risk and Risk with Estimation Risk," Journal of Business & Economic Statistics.
Hogan, Y. (1999), "Hypothesis Testing in Time Series via the Empirical Characteristic Function: A Generalized Spectral Density Approach," Journal of the American Statistical Association, 94, 1201–1220. [1281]
Hong, Y., and Lee, T. H. (2003), "Diagnostic Checking for the Adequacy of Nonlinear Time Series Models," Econometric Theory, 19, 1065–1121. [1277,1281,1282]
Li, W. K., and Mak, T. K. (1994), "On the Squared Residual Autocorrelations in Non-linear Time Series with Conditional heteroskedasticity," Journal of Time Series Analysis, 15, 627–636. [1274,1275,1280]

(2008), "Joint and Marginal Specification Tests for Conditional Mean and Variance Models," Journal of Econometrics, 143, 74–87. [1276,1280,1285]
(2009), "Quasi-Maximum Likelihood Estimation of Semi-Strong GARCH Models," Econometric Theory, 25, 561–570. [1280]
(2010), "Asymptotic Distribution-Free Diagnostic Tests for Heteroskedastic Time Series Models," Econometric Theory, 26, 744–773. [1276,1277,1280]
Escanciano, J. C., and Lobato, I. N. (2009), "An Automatic Portmanteau Test for Serial Correlation," Journal of Econometrics, 151, 140–149. [1275,1278,1281]
Escanciano, J. C., and Olmo, J. (2010), "Backtesting Parametric Value-at-Risk with Estimation Risk," Journal of Business & Economic Statistics, 28, 36–51. [1283]
Fisher, T. I., and Gallagher, C. M. (2012), "New Weighted Portmanteau Statistics for Time Series Goodness of Fit Testing," Journal of the American Statistical Association, 107, 777–787. [1274,1280]
Ling, S., and Li, W. K. (1997), "On Fractionally Integrated Autoregressive
Moving-Average Time Series Models with Conditional Heteroscedas-
ticity," *Journal of the American Statistical Association*, 92, 1184–1194. [1274,1280]

Ljung, G. M., and Box, G. E. P. (1978), "On a Measure of Lack of Fit in Time
Series Models," *Biometrika*, 65, 297–303. [1274]

McNeil, A. J., and Frey, R. (2000), "Estimation of Tail-Related Risk Measures
for Heteroscedastic Financial Time Series: An Extreme Value Approach," *Journal of Empirical Finance*, 7, 271–300. [1281]

Oh, D. H., and Patton, A. J. (2018), "Time-Varying Systemic Risk: Evidence
from a Dynamic Copula Model of CDS Spreads," *Journal of Business &
Economic Statistics*, 36, 181–195. [1281]

Quintos, C., Fan, Z., and Phillips, P. C. B. (2001), "Structural Change Tests
in Tail Behaviour and the Asian Crisis," *Review of Economic Studies*, 68,
633–663. [1281]

R Core Team. 2021. *R: A Language and Environment for Statistical Comput-
ing*, Vienna, Austria: R Foundation for Statistical Computing. [1281]

Resnick, S., and Stáríča, C. (1997), "Smoothing the Hill Estimator," *Advances in Applied Probability*, 29, 271–293. [1279]

Rothe, C., and Wied, D. (2013), "Misspecification Testing in a Class of
Conditional Distributional Models," *Journal of the American Statistical
Association*, 108, 314–324. [1276,1280]

Schmidt, R., and Stadtmüller, U. (2006), "Nonparametric Estimation
of Tail Dependence," *Scandinavian Journal of Statistics*, 33, 307–335.
[1275,1276,1279]

Sibuya, M. (1960), "Bivariate Extreme Statistics," *Annals of the Institute of
Statistical Mathematics*, 11, 195–210. [1276]

Velasco, C., and Wang, X. (2015), "A Joint Portmanteau Test for Conditional
Mean and Variance Time-Series Models," *Journal of Time Series Analysis*,
36, 39–60. [1276,1280]

Zheng, Y., Li, W. K., and Li, G. (2018), "A Robust Goodness-of-Fit Test
for Generalized Autoregressive Conditional Heteroscedastic Models," *Biometrika*, 105, 73–89. [1275,1277,1280,1281]

Zhu, K., and Ling, S. (2011), "Global Self-Weighted and Local
Quasi-Maximum Exponential Likelihood Estimators for ARMA–
GARCH/IGARCH Models," *Annals of Statistics*, 39, 2131–2163.
[1275,1277]