Research Paper

The efficiency of the Anderson–Darling test with a limited sample size: an application to backtesting counterparty credit risk internal models

Matteo Formenti,1,2 Luca Spadafora,3 Marcello Terraneo1 and Fabio Ramponi1

1UniCredit, Piazza Gae Aulenti 4, 20154 Milan, Italy; emails: marcello.terraneo@unicredit.eu, fabio.ramponi@unicredit.eu
2Università Carlo Cattaneo – LIUC, Corso G. Matteotti 22, 21053 Castellanza VA, Italy; email: mformenti@liuc.it
3UBS Business Solutions AG, 5 Broadgate, London EC2M 2QS, UK; email: luca.spadafora@gmail.com

(Received May 24, 2018; revised August 8, 2018; accepted August 9, 2018)

ABSTRACT

This work presents a theoretical and empirical evaluation of the Anderson–Darling test when the sample size is limited. The test can be used to backtest risk factor dynamics in the context of counterparty credit risk modeling. We show the limits of the test when backtesting the distributions of an interest rate model over long time horizons, and we propose a modified version of it that can more efficiently detect the underestimation of a model’s volatility. Finally, we provide an empirical application.

Keywords: Anderson–Darling (AD) test; backtesting; counterparty credit risk (CCR).

Corresponding author: M. Formenti
1 INTRODUCTION

Backtesting is defined as “the quantitative comparison of a model’s forecasts against realized values” (Basel Committee on Banking Supervision 1996b, 2010). In counterparty credit risk (CCR), model forecasts consider the estimates of interest rates, credit spreads and equity or commodity values, which are the underlying risk factors driving the mark-to-market of over-the-counter (OTC) derivatives up to the longest maturity of the contracts. As stated by the Basel Committee on Banking Supervision (2010), banks choose their own best and most appropriate methods to aggregate and then validate the overall quality of model forecasts. This can be done through a synthetic value, such as the outcome of a statistical test, with the goal of detecting any weaknesses in a model’s forecasts. Such forecasts are computed up to the longest maturity and directly affect the exposure toward a counterparty. Backtesting is one of the instruments via which risk managers assess the forecasts of counterparty exposures and the bank’s risk-weighted asset value. A failure in backtesting requires a change to the model, such as a different model parameterization, or even a change in the modeling assumptions used to forecast counterparty exposures.

Further, the Capital Requirements Regulation (CRR) asks banks to use at least three years of historical data for model estimation purposes (CRR, Article 292-2); to have procedures in place to identify and control the risks for counterparties where exposure extends beyond the one-year horizon (CRR, Article 289-6); and to verify their modeling choice through a backtesting program (CRR, Article 293-1(b)). Further, banks having an approved CCR internal model need to compute backtesting

(i) on a risk factor level, with the aim of validating the properties of the stochastic process used to simulate interest rates, credit spreads, foreign exchange and equities;

(ii) on a trade level, such as via plain vanilla or exotic options, with the aim of validating single deal exposure; and

(iii) on a counterparty level, in order to validate the soundness of the estimated exposures.

The CRR is a Capital Requirements Directive consisting of Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC, and Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012.
Hence, regulatory constraints and good management practice lead risk managers to be interested in backtesting all of the forecast distribution shapes of a given model. Uniformity tests are the instruments that can detect the underestimation of counterparty exposure as well as the likely underestimation of regulatory capital measures, risk-weighted assets and managerial measures, such as expected positive exposure (EPE) or potential future exposure (PFE). In turn, the constraints on the calibration window and the availability of historical market data (eg, USD/EUR, started in 1999) lead to test models with a very limited sample size.\(^2\)

This work attempts to improve the ability of statistical tests to accept/reject models when they are applied for risk management purposes, particularly in the case of limited sample sizes (eg, five to ten observations). Specifically, the concern in risk management is catching the under-/overestimation of volatility in an unknown model distribution to prevent the underestimation of a bank’s regulatory capital or credit limit toward specific counterparties. For these reasons, our research objective is to study the statistical properties of the Anderson–Darling (AD) test (Anderson 2010; Anderson and Darling 1954), and our proposed approach for extending this test aims to overcome its limitations, given the AD’s purpose in risk management.

Our choice of the AD test is due to its wide use in the empirical literature, from biology to sociology; its well-known statistical properties for large samples (Daniel 1990; Gibbons and Chakraborti 2003); and its robustness in limited sample sizes. The issue of small sample sizes has been addressed by several authors when testing the residuals of linear regressions (see Bera et al 1984; Dufour et al 1998; Pfaffenberg and Dielman 1991; Stephens 1986). The effects of having a limited sample on statistical tests were first investigated by Ito (1969), Mardia (1970, 1971, 1975), Bera and John (1983) and Rincon-Gallardo et al (1979), all of whom studied the effects of nonnormality due to small samples on multivariate tests. Meanwhile, the issue of the sampling distribution of a test statistic if samples are small was first addressed by Clark (1996) on tests of covariance structures. Several authors, such as Rincon-Gallardo et al (1979) and Miller and Quesenberry (1979), have also searched for the effects of nonnormality due to small samples on multivariate tests. In these

\(^2\) We observe that, from a purely statistical point of view, the use of overlapping time windows to verify model performance cannot be considered a significant improvement toward the reduction of statistical uncertainty in the forecasted variables. In fact, it can be shown that, in our context, if the random variables are independent and identically distributed (iid), the additional information included in a statistical estimator based on overlapping time windows (compared with one based on nonoverlapping time windows) is not enough to significantly reduce the related statistical uncertainty. As a consequence, a backtesting methodology based on overlapping time windows would face similar statistical issues relating to small sample sizes as one based on nonoverlapping time windows.
studies, the authors addressed the issue of having a limited sample size by applying either the Edgeworth expansion methods, as computed by Phillips (1977), Kallenberg (1993), Sargan and Satchell (1986) and Hansen (2006), or the distance metric test with a Bartlett-type correction, as discussed in recent studies by Huang and Prokhorov (2014). The two cited methods enabled a comparison of the various test statistics through an estimation of the power of the tests, and the results are applied to verify either the normality of residuals after regression or the covariance structures of residuals, as studied by Clark (1996). We share with these studies the Monte Carlo approach that used the residuals of the regression models already cited, although we do not consider it to be directly applicable in such cases since test results should not depend on the estimated under- or over-volatility levels of residuals.

Finally, this paper focuses on the specific case of backtesting market risk using the CCR model instead of the value-at-risk (VaR) model. The aim of VaR backtesting is to quantify the goodness of market risk models through the conditional quantile of the distribution of returns on the bank’s portfolio trading book at a predefined probability $\alpha$, while the aim of CCR backtesting is to verify whether the entire distribution of simulated risk factors is able to capture future real exposures. As a consequence, studies and literature related to the VaR model focus on the goodness of backtesting a single percentile instead of the entire distribution. For this reason, a simple unconditional binomial test such as that of Kupiec (1995), the likelihood ratio of Wilks (1938) and the conditional tests of Christoffersen (1998) are not the right fit for the purposes of CCR, and a different approach is needed. Another dissimilarity is that market risk backtesting searches for both the magnitude and signs of the difference between model-generated measures and actual returns by indicating (using probability $\alpha$) whether the VaR model is the best choice for forecasting the underlying market risk. Both methods of backtesting share the goal of understanding whether failure is due to an institution’s overexposure or underexposure to risk. It is particularly relevant for risk management purposes as well as CCR backtesting validation that the test is able to detect volatility underestimation, which is more dangerous from a risk management point of view than volatility overestimation. This implies that a test’s rejection power should be higher when the forecasted distributions exhibit smaller variance than the empirical ones. In addition, this property should also hold in the case of a limited sample size in order to deal with practical situations in which data sets are typically small.

For these reasons, we propose a modified version of the AD test that may help risk managers to easily detect underestimation of volatility. We highlight that uniformity

---

3 The Edgeworth expansion methods extend the sampling density of test statistics around the asymptotic density in decreasing powers of $N^{-0.5}$, with $N$ being the sample size.
tests such as the AD, the Kolmogorov–Smirnov (KS) (see Smirnov 1948), the Jarque–Bera (see Jarque and Bera 1980) and the Cramér–von Mises (CM) (see Anderson 1962) help to statistically validate model forecasting values because, at each backtesting date, we can map the realized value to the corresponding percentile of the forecasted distribution. In particular, for a given risk factor $r$ (eg, interest rate, foreign exchange, commodity), backtesting date $t$ and time horizon $s$, the $F(r)$ corresponding to the realization $r(t + s)$ is computed according to the following algorithm:

$$F(r) = \begin{cases} 
\frac{1}{N + 2} & \text{for } r(t + s) < \hat{r}^{(1)}, \\
\frac{i + 1}{N + 2} & \text{for } \hat{r}^{(i)} < r(t + s) < \hat{r}^{(i+1)}, \\
\frac{N + 1}{N + 2} & \text{for } r(t + s) > \hat{r}^{(N)},
\end{cases}$$

where $\hat{r}^{(i)}$ represents the $i$-forecasted value and $N$ represents the total number of forecasted values. As a consequence, the collection of ordered, mapped values should be uniformly distributed if the model perfectly matches the realized values.

Examples of this approach can be found in the seminal work of the Basel Committee on Banking Supervision (2010) as well as a more recent paper by Anfuso et al (2014). The former offers an introduction on approaches related to how information should be collected in order to set up the prescribed regulatory backtesting program. However, it does not specify which types of tests are most appropriate, what the regulatory confidence level should be, what the risk factors are (eg, how many buckets of the interest rate curve should be used), or what the optimal method for aggregating the overall results of the backtesting program is. Finally, there are no suggestions related to managing a limited sample size. Anfuso et al (2014) explains how banks should set up the overall backtesting program in order to satisfy all regulatory requirements. Moreover, the authors suggest performing the AD test using an approach similar to that described in this work. However, there are no mentions of cases featuring limited sample sizes because the authors opt for overlapping data in the sample test, sacrificing the power of the uniformity test.

The structure of this paper is as follows. In Section 2, we briefly summarize the AD test and its reduced efficiency when the sample size is limited. In Section 3, we propose a modified version of the AD test in order to more quickly detect any underestimation of volatility. In Section 4, we conduct a numerical exercise that uses real data to generate a “fictitious” time series in order to compare the AD test and our modified version with respect to the KS test. In Section 5, we discuss the backtesting
of a Black–Karaskinski model as applied to the Euribor six-month interest rate as a case study. Our results are summarized in Section 6.

2 ANDERSON–DARLING TEST

Anderson and Darling (1954) designed a statistical test to determine whether a given sequence of random variables \( X = \{x_1, \ldots, x_n\} \) comes from a theoretical cumulative distribution function (CDF) \( F(x) \). The null hypothesis \( H_0 \) is that the data follows \( F(x) \), so this test should be used to prove that the data does not follow \( F(x) \) given a particular confidence level. The AD test is based on the estimation of the following random variable:

\[
W^2 = \int_{-\infty}^{+\infty} \frac{[F_n(x) - F(x)]^2}{(F(x)(1 - F(x)))} \, dF(x),
\]

(2.1)

where \( F(x) \) is the target CDF and \( F_n(x) \) is the empirical distribution derived from the data. The numerator of (2.1) represents the distance of the theoretical distribution from the empirical one, while the denominator represents the variance of the empirical estimation of \( F(x) \) when the central limit theorem holds: that is, when \( n \) is large enough. In other words, (2.1) represents the average of the squared errors between the two distributions (theoretical and empirical), weighted by the implicit uncertainty due to the estimation method of the empirical CDF (order statistics). As the CDF of a random variable is always distributed uniformly between zero and one (i.e., \( F(x) \in U(0, 1) \)), \( W^2 \) is a function of uniformly distributed random variables when \( H_0 \) and the central limit theorem hold. In particular, it does not depend on the distribution \( F(x) \).

In this context, we observe that when the variance of the uniform distribution \( (F(x)(1 - F(x))/n) \) is close to zero – that is, for rare events, \( F(x) \sim 0 \) or \( F(x) \sim 1 \) – the squared error is magnified by the small denominator. In this sense, we consider the AD test to be more sensitive with respect to the tails of the distribution. However, we should point out that if the number of observations used to perform the AD test is low, the denominator of the integral in (2.1) is large: that is, a large variance is associated with the difference between the theoretical and empirical distributions. As a consequence, a large difference between the distributions would also fall inside the variance amplitude, and the AD test would not be able to reject \( H_0 \), as the uncertainty in the measurement would be too large to support any conclusion. Accordingly, in order to reject \( H_0 \), the difference between the theoretical CDF and the empirical one has to be larger than their statistical uncertainty. Equation (2.1) can also be expressed as

\[
W^2 = -n - \sum_{k=1}^{n} \frac{2k - 1}{n} \ln(F(x_k)) + \ln(1 - F(x_{n+1-k})),
\]

(2.2)
where \( x_i \in X \) is the empirical data ordered from smallest to largest in value, and \( n \) is the size of the sample (i.e., the number of backtesting dates). The empirical distribution of \( W^2 \) was estimated using the AD test (Anderson and Darling 1954), and we report the percentiles of the \( W^2 \) distribution in Table 1, with the first line indicating the upper tail probabilities and the second line representing the corresponding percentiles.

### 2.1 Analytical result: the efficiency of the AD test

The AD test can be used for CCR backtesting purposes as a tool to verify whether the model-forecasted distributions are comparable to the empirical ones for a given confidence level. In this context, the null hypothesis is that the two distributions are equal. Therefore, the test gives a positive outcome when the null hypothesis is not rejected, meaning that the model is wrong if its distribution is not sufficiently different from the empirical one.

Given our backtesting approach, a large uncertainty in the empirical cumulative density function (cdf) estimation is due to the small number of observations. This negatively affects the rejection rate of our test-accepting distributions just because of the limited sample size. As a consequence, we consider the introduction of an efficient measure in such a test in order to increase its accuracy when the sample size is small.

In general, a measure is considered efficient and faithful if its uncertainty is much smaller than its expected value;\(^4\) in our case, the statistical uncertainty is related to the empirical estimation of the cdf, whereas the expected value is the theoretical cdf for each point of the empirical distribution. Unfortunately, the simple variance estimation for each point of the distribution is not enough, as the AD measure requires a sum over all of the probabilities \( F_n(x_i) \) for \( \{i = 1, \ldots, n\} \), and one must also consider the whole covariance matrix of the order statistics.

Accordingly, at each single point the expected value of our empirical estimation is given by \( F(x_i) = p_i \), and the covariance structure is given by \( p_{\min(i,j)} - p_i p_j \). We can define the coefficient of variation (CoV), or relative standard deviation, for each

---

\(^4\) The approach is not consistent if the expected value is equal to zero.
FIGURE 1  Decay of the coefficient of variation as a function of the number of observations.

point of the distribution as

\[ c = \frac{\sigma}{\mu} = \frac{1}{\sqrt{n}} \frac{\int_{1/n}^{1} \min(p, q) - pq \, dp \, dq}{\int_{1/n}^{1} p \, dp} = \frac{\sqrt{(1/n - 1)^2 (2/n + 3/4) / 3}}{\sqrt{n} (1/2 - 1/2n^2)} = \frac{1}{n + 1} \sqrt{n + \frac{8}{3}} \]

where \( \sigma \) and \( \mu \) represent the standard deviation and the expected value of the sum of the overestimated probabilities in a range between \( 1/n \) and 1: that is, \( \int_{1/n}^{1} F_n(x) \, dF(x) \). In Figure 1, we show the decay of the coefficient of variation as a function of the number of observations. In order to obtain a CoV below 10%, \( n = 50 \) observations are required. The CoV indicator gives us important information about the performance of the AD test when the sample size is small, and it can be used as a warning level when the AD test is applied. However, we should point out that the CoV is derived assuming the central limit theorem (CLT) holds; this hypothesis does not hold for very small sample sizes. Therefore, one should consider additional corrections to (2.3) that we do not take into consideration in this work.
3 ASYMMETRIC EXTENSION OF THE ANDERSON–DARLING TEST

In order to detect underestimation of the actual volatility when the number of observations is small, we derive an extension of the AD test that can be used for risk management purposes. The main idea behind this AD test extension comes from the observation that, when the sample size is small, it is easier to reject the null hypothesis when the empirical variance of the distribution is larger than the forecasted one. In contrast, when the empirical variance is smaller than the forecasted one, many observations fall inside the theoretical distribution, so it is more difficult to reject the \( H_0 \). This fact will be discussed further, with numerical examples, in the next sections.

Starting from this observation, one could magnify the asymmetric effect by defining an AD-asymmetric (AD-Asym) function that introduces a more pronounced nonlinear behavior when the differences between the theoretical and the empirical cdf are large, as in the case of variance underestimation. We stress that the term “asymmetric” refers to the different behavior of the AD test when the variance of the forecasted distribution is over-/underestimated, and not to the analytical form of \( W^2 \).

Equation (2.1) is then generalized as

\[
W_{\text{Asym}}^2 = \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^{2\beta}}{((F(x)(1 - F(x)))^\beta/n)\beta} dF(x),
\]

where \( \beta \geq 1 \) is the parameter that controls the amplitude of the asymmetric effect, and by using \( \beta = 1 \) we recover (2.1). The reason we increase the power of the AD test instead of using another test, such as the CM, is due to the fact that the AD is the most suitable test for detecting the “wrong” model, where “wrong” means simulating the risk factors using a lower level of volatility than the real one (eg, the real volatility of the risk factor). To tackle this issue, the AD-Asym increases the nonlinearity of the AD test by enlarging by a factor of \( \beta \) the difference between the theoretical and empirical distributions (ie, the numerator of the test).

In this study, we focus our attention on the special case of \( \beta = 2 \); in this way, the small variance amplitude due to the small sample \( n \) is compensated for by the \( \beta \) exponent. In order to apply this new asymmetric formulation of the AD test in practical situations, we have to

(i) estimate the integral in (3.1),

(ii) obtain the distribution of the \( W^2 \) random variables assuming that \( H_0 \) is true, and

(iii) compare the empirical value of \( W^2 \) with the theoretical distribution obtained at the previous point and decide whether \( H_0 \) is rejected at a given confidence level.

www.risk.net/journals

Journal of Risk

Electronic copy available at: https://ssrn.com/abstract=3439413
Once a good estimation of (3.1) is obtained, the second and third steps can be overcome by conducting a numerical simulation of the $W^2$ random variable before giving consideration to the obtained cdf. Some care should be taken when executing the integral estimation, given the small sample size. In particular, along the same lines as the original work (Anderson and Darling 1954), we consider a sample of $\{x_1, \ldots, x_n\}$ observations, define $x_0 = 0$ and $u = F(x)$, and find a solution as follows:

$$W^2_{\text{Asym}} = \sum_{k=1}^{n} \int_{u_{k-1}}^{u_k} \frac{(u - ((k - 1)/n))^4}{((1-u)(u/n))^2} \, du$$

$$= \sum_{k=1}^{n} -\left( \frac{k-1}{n} \right)^4 \frac{1}{u} + 2 \left( \frac{k-1}{n} - 2 \right) \left( \frac{k-1}{n} \right)^3 \log(u)$$

$$- \frac{((k-1)/n - 1)^4}{u-1} - 2 \left( \frac{k-1}{n} + 1 \right) \left( 1 - \frac{k-1}{n} \right)^3 \log(1-u) + u \bigg|_{u_{k-1}}^{u_k}$$

$$= \gamma + \sum_{k=1}^{n-1} \frac{\alpha_1(k)}{u_k} + \alpha_2(k) \log(u_k) + \frac{\alpha_3(k)}{u_k-1} + \alpha_4(k) \log(1-u_k). \quad (3.2)$$

### 3.1 Choice of power magnitude

In this study, we chose to perform our numerical and calibration analyses using $\beta = 2$, to which the numerator and the denominator of the AD test are raised to obtain the AD-Asym test. This choice was not supported by quantitative analysis but by qualitative expert opinions from risk management expertise; the level of $\beta$ allows us to increase the rejection rate (lower error of type I) of models with volatilities lower than that currently in place in the financial markets. Hence, a higher level of $\beta$ increases the probability of recognizing whether the model approach is conservative. For this reason, we selected the AD test and decided to increase the power of the numerator (eg, the power of the AD-Asym test). However, the level of $\beta$ in the AD-Asym (3.2) might be due to different approaches, as follows.

**Number of observations.** It is possible to calibrate $\beta$ depending on the number of observations within the test sample. We believe the optimal $\beta$ should address to what extent the underestimation of the model volatility is a function of the sample size $n$ we are using to backtest our model. Using this approach, the risk manager fixes the amount of volatility underestimation that is considered suitable from a risk management point of view (perhaps $-50\%$) and then determines, via a theoretical exercise, the level of $\beta$ that allows the rejection of a wrong model.
at a given confidence level (perhaps 50%, at least). However, with this approach, there are two variables to set: namely, the level of volatility underestimation and the level of test rejection. Given such optimal levels, it is possible to determine the optimal level of $\beta$. As an example, Figure 2(a) shows the rejection rate of the AD and AD-Asym with five observations in the sample; the AD-Asym double-rejects the null hypothesis when the volatility of the forecasted distribution is half the realized distribution (e.g., AD-Asym rejects 40% instead of 2.2% when $\sigma = 1.5$). The result is similar when there are twenty observations, as shown in Figure 2(b). This leads us to conclude that, in the latter case, $\beta = 2$ is not enough to have at least a 50% chance of a model being rejected; on the contrary, when there are twenty observations, the rejection rate of the AD-Asym is 65%, which can be considered acceptable. We support the quantitative approach but recognize that an expert opinion is still necessary in choosing the level of volatility underestimation, which could be dangerous with regard to both using the model and selecting the level of rejection rate of the test.

**Model risk of pricing functions.** Consider $\beta$ to be the variable that estimates model risk in pricing functions. Following Rebonato (2001), “model risk” is the risk that there will be a significant difference between the mark-to-model value of a complex illiquid instrument and the price at which the same instrument is revealed to have traded in the market. As a consequence, a high level of difference between the mark-to-model price and the OTC price should provide a strong rationale for increasing the level of $\beta$ in order to avoid a situation in which backtesting provides good results when the mark-to-model is far from the real price. For example, we can consider the model risk of pricing swaps, caps, floors or swaptions as low and, as a consequence, set a level of $\beta < 2$, since the model risk is limited for such simple products.

**Uncertainty in volatility estimation.** This approach stems from the fact that one uncertainty in CCR exposure is the level of volatility estimated using three years of historical data. This is due, for example, to a level or regime switch that can affect the estimation period (e.g., a day when pegged currencies are switched off). As a consequence, the level of $\beta$ should depend on such events. With this approach, risk managers may consider addressing the relevant missing assumptions of the model they use for pricing or simulation. For example, a model that does not consider the heteroscedasticity of returns in risk factor modeling could entail an error due to the “real” volatility, which is nonconstant. The $\beta$ level might take such a missing assumption into account, e.g., $\beta > 2$ might be used. In Section 4, we support the choice of $\beta = 2$ because the modeled interest rate volatility is estimated to be closer to the real interest rate volatility, since the average estimated volatility is $\sigma_{BK} = 0.01241$ while the real one is $\sigma_{Euribor\ 6M} = 0.0198$.  

www.risk.net/journals
FIGURE 2  Rejection rate as a function of the standard deviation: (a) sample size with five observations and (b) sample size with twenty observations.

“SS” stands for standard symmetric.

Risk appetite. Risk appetite can be defined as the variability in results, on both a short-term and a long-term basis, that the bank is prepared to accept in support of a stated strategy; in other words, an “outer boundary” for how much risk the bank can take. A bank’s risk appetite should be consistent with its desired rating target, which depends on the coherence of the relevant business strategies/revenue targets containing risk. As a consequence, a high (low) level of $\beta$ is consistent
with a strong (weak) perception of risk. Hence, the level of $\beta$ can be assessed using a qualitative approach that involves the overall framework of the bank’s risk appetite.

### 3.2 Testing the AD in a limited sample size

The work of Scholz and Stephens (1987) extended the AD test up to the $k$-sample dimension in order to establish any joint differences in the sampled populations without imposing constraints on their parametric assumptions. In particular, the authors studied the quality of the proposed large sample approximation for small samples via a Monte Carlo simulation, which can be considered an enhancement for the proposed AD-Asym test whenever there is a need to jointly test more samples of different size. For this reason, the proposed AD-Asym test cannot be considered a part of the statistical studies that aim at modifying the asymptotic properties of this test. There is, indeed, a generalization concerning the AD test ($\beta = 1$) that it aims to stress its own well-known asymptotic properties for research purposes. In this section, we compare the AD test and our proposed modification (3.2) to obtain some insights into AD-Asym performances when the number of observations is small.

For CCR management purposes, we still remark that our main goal is to reject more efficiently forecasted (theoretical) distributions with smaller variances than the actual (empirical) ones (ie, volatility underestimation). We address this issue by performing numerical simulations of the AD test assuming different theoretical probability density functions (PDFs). In particular, we compare a Gaussian distribution $N(0, 1)$, assumed to be our theoretical PDF, with other Gaussian distributions having the same mean but different variances. For a fixed number of observations, we estimate the rejection rate of the AD test as a function of the real standard deviation considered to generate the sample set. We expect that, for a large enough sample size, the AD test will reject the different distributions with a rejection rate equal to 1 in 100% of cases, the only exception being when the real variance is equal to the theoretical one (that is, equal to 1). In the latter case, the rejection rate depends on the confidence level required for the test, which in our case is set at 5%.

In Figure 3, we report the results of our analysis for a small sample size. As expected, the rejection rate is higher for larger standard deviations, although the overall performance is not considered to be satisfactory. We obtain a rejection rate of around 50% when considering five observations, and the standard deviation is twice the size of the empirical one. The asymmetry between the smaller and larger standard deviations becomes less evident as the number of observations increases; in fact, when the sample size is 100 observations, the rejection rate becomes symmetric, as demonstrated in the seminal work of Anderson and Darling (1954).
FIGURE 3 Comparing (a) the AD-SS and (b) the AD-Asym using different samples and standard deviations of a normal distribution ($N(0,1)$ versus $N(0,x)$).

In particular, the AD test performs better in our main area of interest: the most extreme cases, where models that underestimate the real volatility are rejected more, as shown on the right-hand side of Figure 3(a). In fact, the AD test has a higher rejection rate when the distribution of real data has the same mean as the forecasted distribution but a larger variance. The reason for this is, as already discussed in the previous section, that a realized PDF with a larger variance has a high probability of generating outliers (with respect to the forecasted PDF), even with only a few
observations. In contrast, if we consider the realized PDF with a lower variance than the forecasted one, all of the observations will fall inside the forecasted distribution, and it is more difficult to conclude whether the two distributions are different. We replicate this analysis when comparing the AD test with our AD-Asym version using a different sample size. Figure 3(b) shows that we have a similar rejection rate when the volatility is overestimated (σ < 1) and a higher rejection rate when the volatility is underestimated. Therefore, we conclude that if we need to avoid volatility underestimation, the AD-Asym test performs better – especially in a limited sample.

In order to extend our results, we compare the rejection rates of the different statistical tests that aim to identify whether a sample – always limited in our case – belongs to a given theoretical distribution. In particular, we consider four different statistical tests:

- the standard symmetric AD test, as described in the previous sections (AD-SS);
- the AD test with tail-sensitive indicators (AD-Asym);
- the Cramér–von Mises (CM) test; and
- the Kolmogorov–Smirnov (KS) test.

Figure 4 shows the results of our analyses as a function of \( n \); in particular, in each part, the rejection rate of different statistical tests is compared assuming a Gaussian \( N(0, 1) \) theoretical distribution and empirical Gaussian distributions with different means and standard deviations. As expected, the AD-Asym test shows a better overall performance (higher rejection rate) than the other tests when the volatility is underestimated.

We analyzed the performance of different statistical tests considering a Gaussian \( N(0, 1) \) theoretical distribution and an empirical Student \( t \) distribution with a variance equal to 1 and \( v \) degrees of freedom. This is due to the well-known fact that the distributions of financial returns frequently show fat-tailed behavior. Figure 5 shows the results of these analyses. The performance levels of the AD-Asym are higher than those related to the AD test, although we know that such tests are strongly influenced by the degrees of freedom considered. This can be explained by the fat-tailed nature of the Student \( t \) distribution, which signifies a higher risk than a Gaussian distribution with the same variance. For this reason, the use of a Gaussian distribution implies an underestimation of the risk that would be better identified by an AD-Asym test, as has already been pointed out in the previous section. In addition, we observe that the convergence of the rejection rate to 100% is not reached when \( n = 150 \), which is contrary to the Gaussian case. The reason for this behavior is from statistics: when the first two moments of the theoretical and empirical distributions are met exactly,
Tests assuming (a) volatility overestimation ($N(0, 1)$ versus $N(0, 0.5)$) and (b) volatility underestimation ($N(0, 1)$ versus $N(0, 1.5)$).

we need to estimate the higher moments in order to observe the differences between the PDFs, and for a fixed number of observations the uncertainty increases with the moment order. It is therefore reasonable to expect that the number of observations needed to obtain the convergence must be larger than that obtained using Gaussian PDFs.

3.3 Comparing the AD-Asym with other uniformity tests

The performance of the AD-Asym test can be compared with those of the AD and other uniformity tests, such as the CM or KS tests, for small sample sizes. Figure 6 shows the rejection rate for different uniformity tests when changing the tested
Tests assuming (c) mean overestimation and the correct volatility ($N(0, 1)$ versus $N(0.5, 1)$) and (d) mean overestimation and volatility underestimation ($N(0, 1)$ versus $N(0.5, 1.5)$).

distribution in terms of mean or volatility. The result is that the AD-Asym with $\beta = 4$ has a higher rejection rate than other tests when changing the parameters of the real distribution. In particular, the AD-Asym has a higher rejection rate when volatility is underestimated as well as when the mean is either correctly estimated or overestimated (see parts (b) and (d) of Figure 6). The reason lies in the mathematical definition of AD-Asym, described theoretically in Section 3 and shown empirically here: the proposed test aims to symmetrically increase the weight of large deviations between theoretical and empirical distributions by increasing the exponent of the integrand function of the standard AD test by a factor $\beta$. Because of the symmetry of
FIGURE 5  Rejection rate for the AD test (blue line with diamonds) and the AD-Asym test (red line with dots) assuming a Gaussian $\mathcal{N}(0, 1)$ theoretical distribution and a Student $t$ distribution with (a) $\nu = 2.8$ degrees of freedom, (b) $\nu = 3$ degrees of freedom and (c) $\nu = 3.5$ degrees of freedom.
the integrand function, an improvement in terms of the rejection rate is then expected in symmetric cases, ie, when the odd moments of the two distributions are similar. In these cases, the proposed test remarkably improves performance with respect to the standard AD test.

In Figure 6(b), the rejection rate of the proposed AD-Asym test when $N = 30$ is about 80%, while that for the AD test is about 60%. On the contrary, when the symmetry between the two distributions is broken and a mean effect is introduced, the performance of the two tests is then comparable. A similar behavior can be obtained with $\beta = 4$, where, as expected, we observe a slight improvement in performance for volatility underestimation; however, the performance is worse for volatility overestimation (Figure 6(c)). In these cases, test performance can be influenced by the introduction of the “mean” effect, which is magnified by the high $\beta$ factor.

These cases are shown for the sake of completeness, but they do not represent the aims of the proposed test, which are focused on under-/overestimation of the second moment. This aspect is left to future research. In conclusion, it is important to highlight that the low level of improvement computed by the AD-Asym test, shown in Figures 4 and 6, is due to the limited sample size and the low level of $\beta$ we have chosen in this work. The former would not facilitate any evidence of improvement: this was also demonstrated by Stephens (1974) in his early work, where the rejection power of different normality tests was compared for completely specified distributions, a normal distribution with its first two moments known and unknown, and a sample of more than five observations. The results of Stephens (1974) are similar to ours, with the AD and CM tests being the preferred statistics in terms of exhibiting better power with respect to KS, and there being no strong preference between the two. On the other hand, the slight improvement of the AD-Asym compared with the performance of the CM for very small sample sizes is questionable.

4 NUMERICAL EXERCISE

We consider the Euribor six-month interest rate, which provides us with a historical time series starting in 1999, to test the AD and AD-Asym tests in a realistic case. As a consequence, we have five/six observations in the backtesting sample if we are using three years of historical data for our estimation purposes and a forecasting horizon of two years. However, in order to address the properties of our AD test using a limited sample size but with the possibility of extending our analysis to a higher sample size, we detrend the interest rate time series using the Hodrick and Prescott (1997) filter applied to the five-day time series up to March 2012. In this way, we are able to obtain more robust calibration results using the interest rate model. We then estimate...
a Black and Karasinski (1991) interest rate model on the cycle component of the log-filtered time series in order to generate, using the estimated parameters, a “fictitious” time series with the desired sample to backtest. We use the Black–Karasinski short-rate model because of its analytical tractability and for the added benefit that rates cannot become negative under it. In fact, the Black–Karasinski model can model the logarithm of the cyclical component of the interest rate \( y(t) = \exp(r(t)) \) using the Ornstein–Uhlenbeck stochastic process:

\[
dy_t = a(\theta - y_t) \, dt + \sigma \, dW_t,
\]

(a) Underestimation of mean (\( N(0, 1) \) versus \( N(0, 0.5) \)). (b) Underestimation of mean and overestimation of variance (\( N(0, 1) \) versus \( N(0.5, 1.5) \)).
where $a$ captures the speed of the $y_t$ log cycle toward its long equilibrium value $\bar{\theta}$ (the mean level), $\sigma$ is the volatility of the process and $W_t$ is the Brownian motion. We estimate the following parameters using a moment-matching approximation formula on the overall data set:

$$\Theta \equiv [\tilde{\theta} = -0.0004871; \tilde{a} = 0.011853; \tilde{\sigma} = 0.018855].$$

The moment-matching method guarantees the level of mean reversion is positive, which is consistent with the values of the risk factor. In particular, this estimation method is based on a two-step formula, where the mean-reversion level is first
estimated and then plugged into the mean reversion rate:

\[
\hat{\theta} = \frac{1}{n} \sum_{k=1}^{n} y(t_k), \quad (4.2)
\]

\[
a = \frac{\sum_{k=1}^{n} (y(t_k) - y(t_{k-1}))^2}{2 \sum_{k=1}^{n} (y(t_k) - \bar{y}(t_k))^2}. \quad (4.3)
\]

In Figure 7, we plot the Euribor six-month interest rate and the filtered cyclical and trend components. We simulate interest rate scenarios using the estimated \( \Theta \) parameters and a constant number of backtested observations in order to verify the statistical properties of the tests at different time horizons. In Figure 8, we show the results of the AD and KS rejection rates, using a backtesting sample with five observations in every simulation. All of the simulations are performed 10 000 times using the same parameters.\(^5\) The rejection rate is plotted against the ratio of the average standard deviation of the simulated process (using \( \theta \) parameter) to the standard deviation at every backtesting date calibrated considering a time window of three years and keeping \( a \) constant;\(^6\) the variance of the process is estimated as shown in Appendix B:\(^7\)

\[\Delta = \tilde{\sigma}_{\text{sim}} / \tilde{\sigma}_{\text{bkt}}.\]

As a consequence, the condition \(\Delta > 1\) implies an underestimation of the volatility of the backtesting subsamples, and we expect a higher rejection rate for both tests. In turn, \(\Delta < 1\) implies that the backtested volatility is higher than that generated in the “fictitious” time series, so we expect a lower, or even zero, rejection rate. This numerical exercise confirms the theoretical analysis shown in Figures 1–3, where the AD test has a higher rejection rate than the KS test when the sample size is low.\(^8\) We control these results using ten, twenty-five and 100 observations and obtain similar results. Last, we note that the volatility estimated using three years of historical data is, on average, lower than that used when simulating the “fictitious” real interest rate. This is confirmed in Figure 9, where we plot the histogram of the distribution of the parameter \(\Delta\).

\(^5\) Note that the AD-Asym rejection rate is affected by the number of observations in the sample, so for \(\Delta > 1.2\) we might have a nonlinear rejection rate.

\(^6\) The variance of the process, as described in Appendix B, depends on the mean-reversion rate \(k\), so in order to avoid a misalignment between the volatility used to simulate the process (\(\theta\)) and those used for backtesting purposes, we opt to use a constant \(k\) with an overall effect that we verify as negligible for our purposes.

\(^7\) The variance of the Black–Karasinski process depends on the time horizon, so we expect this difference in the \(x\)-axis to be as shown in Figure 8.

\(^8\) The extreme boundaries of Figure 8 have few observations, so we discard these from our analysis.
5 EMPIRICAL APPLICATION

In this section, we give an example of a real risk factor backtest being applied to the AD and KS tests on interest rate-forecasted values obtained using the Black and Karasinski (1991) short-rate model. We computed it to verify the performance of the AD test in a real case, where we are interested in backtesting mark-to-the-future distributions at different horizons of the entire distribution of the interest rate. For simplicity, we apply the Black–Karasinski model as given by (4.1) to the Euribor six-month time series, where the parameters are calibrated using three years’ worth of historical data and the moment-matching method shown in (4.2).

We slightly modify the short-rate model (4.1) in order to perform tests with an alternative model setup, as follows:

\[ dy_t = a(\bar{\theta} - y_t) \, dt + \gamma \sigma \, dW_t, \]  

(5.1)

where the parameter \( \gamma \) is an adjustment to the volatility value. Figure 10 plots the Euribor six-month interest rate and the forecasted distribution at two years when setting \( \gamma = 1.9 \). We point out that the backtesting time window included the interest rate regime switch that occurred during the Lehman crisis, and, in general, interest

---

\[ \text{FIGURE 7} \quad \text{Euribor six-month interest rate, filtered with Hodrick–Prescott and simulated with parameters } [\hat{\theta}, \hat{k}, \ddot{\sigma}]. \]

(a) Euribor six-month interest rate. (b) Time series plot: cyclical component. (c) Time series plot: trend component.
rate models would have serious problems trying to correctly forecast the crisis and the following period. Figure 10 shows that

- only five backtesting dates are available for testing the model due to the long forecast horizons,
- the empirical data often falls into the extreme tail of the forecasted distribution, and
- we can capture the realized values only by increasing the model volatility or adjusting the mean volatility.
The efficiency of the AD test with a limited sample size

Table 2 reports the simulated rates and the corresponding real rates for the five backtesting dates. Table 3 reports the results of the AD, AD-Asym and KS tests for different levels of volatility assuming a 5% level of confidence: all of the tests reject the hypothesis that the model is correct (see the bold column). In order to compare the performance of the AD, AD-Asym and KS tests, we reparameterize the model by increasing the volatility. In particular, we increase $\gamma$ from one to up to three times the estimated volatility at each backtesting date, and we observe that, as expected, the tests accept the null hypothesis when the forecasted distribution is very large. In 10 A test rejects the null hypothesis when the $p$-value is lower than the confidence level.

---

**FIGURE 8** Continued.

(c) Time horizon: one year. (d) Time horizon: two years.

---

10 A test rejects the null hypothesis when the $p$-value is lower than the confidence level.
FIGURE 9 Distribution of $\Delta = (\sigma_{\text{sim}}/\sigma_{\text{bkt}})$ in the numerical exercise.

contrast, the AD-Asym test is less reactive to such adjustments than the AD and KS tests. We consider this exercise as evidence of the conservativeness of the AD and AD-Asym tests. Last, we check the robustness of our results with different buckets of the interest rate curve (i.e., one-year, five-year and ten-year buckets) and obtain similar results.

6 CONCLUSIONS

The European CRR/Capital Requirements Directive IV (CRD IV) regulations require the in-depth analysis of a model’s forecasts, which are used to compute CCR exposures. As a consequence, banks need to implement a backtesting program, and the results from this process are very important for the assessment of model weaknesses that affect counterparty exposures and risk-weighted assets. At the same time, banks commonly face a situation in which they are forecasting risk factors at long time horizons while satisfying a requirement to use three years of data in their model calibrations: a very limited sample data set. Thus, from a purely statistical point of view, the use of overlapping time windows to verify model performance cannot be considered a significant improvement in reducing the statistical uncertainty of the forecasted variables. A proper approach to CCR model backtesting should be developed, taking into consideration the strong constraints imposed by the limited statistical power of small data sets. It is worthwhile to observe that, in general, the statistical test should be able to easily detect a model’s volatility underestimation, which is more dangerous from a risk management point of view than volatility overestimation.
FIGURE 10  Simulated distribution of Euribor six-month interest rate at a two-year time horizon and three years of historical calibration.

TABLE 2  Euribor six-month- and statistics-forecasted values at a two-year time horizon.

| Date       | 12/24/04 | 12/22/06 | 12/19/08 | 12/17/10 | 12/14/12 |
|------------|----------|----------|----------|----------|----------|
| Value      | 2.211    | 3.8291   | 2.7634   | 1.0449   | 0.1685   |
| Minimum    | 2.0609   | 1.2958   | 2.2502   | 1.8545   | 0.16382  |
| Average    | 3.3743   | 2.3433   | 3.4989   | 3.3177   | 1.2567   |
| Maximum    | 5.5148   | 3.7534   | 4.9848   | 5.357    | 4.8442   |

TABLE 3  AD, AD-Asym and KS results for Euribor six-month-forecasted values at a two-year time horizon.

| Test   | 1   | 2   | 2.5 | 2.75 | 3   |
|--------|-----|-----|-----|------|-----|
| AD     | Reject | Reject | Reject | Accept | Accept |
| AD-Asym | Reject | Reject | Reject | Reject | Accept |
| KS     | Reject | Reject | Accept | Accept |     |
For these reasons, in this study we focused on the AD test with the goal of understanding when the test is able to detect a volatility underestimation, and we derived an extension of the AD test that can be used for risk management purposes. The main idea behind this extension comes from the observation that the AD test rejection rate is higher when the real variance of the distribution is larger than the forecasted one; however, the rejection rate is quite low due to uncertainty related to the small sample size. As a consequence, we magnify this asymmetric effect by introducing a more pronounced nonlinear behavior, where the differences between the forecasted and real distributions are large when a model’s volatility is underestimated. This means the test’s rejection rate will be higher when the forecasted distribution has a smaller variance than the real one. This property should also hold in the case of a limited sample size in order to be conservative.

We verified the properties of our modified test in the case of a limited sample size, showing that its rejection performance is better overall than that of a standard uniformity test when the forecasting distribution is wrong. We checked this result by comparing it with the AD test and other uniformity tests, such as the KS test, using a numerical example that forecasted Euribor six-month interest rate values using a two-year time horizon with a Black–Karazinski short-rate model. We then used this model to backtest the real interest rate during the time window 2000–2014, which led to the chosen model being rejected.

**APPENDIX A. FUNCTIONS DEFINITION FOR THE ANDERSON–DARLING-ASYMMETRIC TEST**

In the following, we report the functions defined in (3.2):

\[
\gamma = u_n - \frac{(-1 + ((n - 1)/n))^4}{u_n - 1}
- \left(-1 + \frac{n - 1}{n}\right)^3 \left(2 + \frac{2(n - 1)}{n}\right) \log(-u_n + 1)
+ \frac{(-4 + 2((n - 1)/n))(n - 1)^3 \log(u_n)}{n^3} - \frac{(n - 1)^4}{n^4} - 1, \tag{A.1}
\]

\[
\alpha_1(k) = \frac{4k^3 - 6k^2 + 4k - 1}{n^4}, \tag{A.2}
\]

\[
\alpha_2(k) = \frac{2(6nk^2 - 6nk + 2n - 4k^3 + 6k^2 - 4k + 1)}{n^4}, \tag{A.3}
\]

\[
\alpha_3(k) = -\frac{(-n + k)^4 + (n - k + 1)^4}{n^4}, \tag{A.4}
\]

\[
\alpha_4(k) = \frac{2n^3 - 6nk^2 + 6nk - 2n + 4k^3 - 6k^2 + 4k - 1}{n^4}. \tag{A.5}
\]
APPENDIX B. VARIANCE OF THE BLACK–KARASINSKI PROCESS

\[
\begin{align*}
\text{Var}[X_t] &= E[(X_t)^2] - E[(X_t)]^2, \\
E[(X_t)^2] &= \exp \left\{ 2 \ln(X_0)e^{-kt} + 2a(1 - e^{-kt}) + \frac{\sigma^2}{k}(1 - e^{-2kt}) \right\}, \\
E[(X_t)] &= \exp \left\{ \ln(X_0)e^{-kt} + a(1 - e^{-kt}) + \frac{\sigma^2}{4k}(1 - e^{-2kt}) \right\}.
\end{align*}
\]

DECLARATION OF INTEREST

The views, thoughts and opinions expressed in this paper are those of the authors in their individual capacities and should not be attributed to the UniCredit Group or to the authors as representatives or employees of the UniCredit Group.

ACKNOWLEDGEMENTS

We truly appreciate the anonymous referees for their thoughtful and constructive comments. We thank the seminar participants of the 21st Annual Risk Mind Conference in Amsterdam, of the Global Derivative Conference in Budapest and of the Quantitative Finance Conference at the Scuola Normale Superiore di Pisa. We also thank our UniCredit Risk Management colleagues for their helpful comments on an earlier version of this paper. Finally, we thank the UniCredit Group and the UniQuant Community for supporting the research activity during the validation of the CCR internal model.

REFERENCES

Anderson, T. W. (1962). On the distribution of the two-sample Cramer–von Mises criterion. *Annals of Mathematical Statistics* 33(3), 1148–1159 (https://doi.org/10.1214/aoms/1177704477).

Anderson, T. W. (2010). Anderson–Darling tests of goodness-of-fit. Mimeo, Stanford University (https://doi.org/10.1007/978-3-642-04898-2_118).

Anderson, T. W., and Darling, D. A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association* 49, 765–769 (https://doi.org/10.1080/01621459.1954.10501232).

Anfuso, F., Karyampasz, D., and Nawroth, A. (2014). A sound Basel III compliant framework for backtesting credit exposure models. *Risk*, August (https://doi.org/10.2139/ssrn.2264620).

Basel Committee on Banking Supervision (1996a). Amendment to the capital accord to incorporate market risks. Report, Bank for International Settlements.

Basel Committee on Banking Supervision (1996b). Supervisory framework for the use of “backtesting” in conjunction with the internal models approach to market risk capital requirements. Report, Bank for International Settlements.
Basel Committee on Banking Supervision (2010). Sound practices for back-testing counterparty credit risk models. Report, Bank for International Settlements.

Bera, A. K., and John, S. (1983). Tests for multivariate normality with Pearson alternatives. *Communications in Statistics: Theory and Methods* 12(1), 103–111 (https://doi.org/10.1080/03610928308828444).

Bera, A. K., Jarque, C. M., and Lee, L. F. (1984). Testing the normality assumption in limited dependent variable models. *International Economic Review* 25, 563–578 (https://doi.org/10.2307/2526219).

Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts Journal* 47(4), 52–59 (https://doi.org/10.2469/faj.v47.n4.52).

Blinnikov, S., and Moessner, R. (1998). Expansions for nearly Gaussian distributions. *Astronomy and Astrophysics Supplement Series* 130, 193–205 (https://doi.org/10.1051/aas:1998221).

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review* 39(4), 841–862 (https://doi.org/10.2307/2527341).

Clark, T. E. (1996). Small-sample properties of estimators of nonlinear models of covariance structure. *Journal of Business and Economic Statistics* 14, 36–73 (https://doi.org/10.2307/1392448).

Daniel, W. W. (1990). *Applied Nonparametric Statistics*. Duxbury Thomson Learning, Pacific Grove, CA.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* 25, 631–650 (https://doi.org/10.1214/aoms/1177728652).

Dufour, J.-M., Farhat, A., Gardiol, L., and Khalaf, L. (1998). Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal* 1, 154–173 (https://doi.org/10.1111/1368-423X.11009).

Gibbons, J. D., and Chakraborti, S. (2003). *Non-parametric Statistical Inference*. Marcel Dekker, New York.

Giles, D. E. A. (2000). A saddle-point approximation to the distribution function of the Anderson–Darling test statistic. Econometrics Working Paper 0005, Department of Economics, University of Victoria.

Hansen, B. E. (2006). Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions. In *Econometric Theory And Practice: Frontiers of Analysis and Applied Research*, Corbae, D., Durlauf, S. N., and Hansen, B. E. (eds), pp. 9–35. Cambridge University Press (https://doi.org/10.1017/CBO9781139164863.003).

Hodrick, R., and Prescott, E. (1997). Postwar US business cycles: an empirical investigation. *Journal of Money, Credit, and Banking* 29(1), 1–19 (https://doi.org/10.2307/2953682).

Huang, W., and Prokhorov, A. (2014). Bartlett-type correction of distance metric test. In *Festschrift in Honor of Peter Schmidt: Econometric Methods and Applications*, pp. 371–403. Springer (https://doi.org/10.1007/978-1-4899-8008-3_13).

Ito, K. (1969). On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. *Multivariate Analysis* 2, 87–120.

Jarque, C. M., and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6(3), 255–259 (https://doi.org/10.1016/0165-1765(80)90024-5).

*Journal of Risk* www.risk.net/journals

Electronic copy available at: https://ssrn.com/abstract=3439413
The efficiency of the AD test with a limited sample size

Kallenberg, W. C. M. (1993). Interpretation and manipulation of Edgeworth expansions. *Annals of the Institute of Statistical Mathematics* **45**, 341–351 (https://doi.org/10.1007/BF00775819).

Kupiec, P. (1995). Techniques for verifying the accuracy of risk management models. *Journal of Derivatives* **3**, 73–84 (https://doi.org/10.3905/jod.1995.407942).

Lewis, P. A. W. (1961). Distribution of the Anderson–Darling statistic. *Annals of Mathematical Statistics* **32**, 1118–1124 (https://doi.org/10.1214/aoms/1177704850).

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–530 (https://doi.org/10.1093/biomet/57.3.519).

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika* **58**, 105–121 (https://doi.org/10.1093/biomet/58.1.105).

Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling’s T2 test. *Applied Statistics* **24**, 163–171 (https://doi.org/10.2307/2346563).

Miller, F. L., and Quesenberry, C. P. (1979). Power studies of tests for uniformity. *Communications in Statistics: Simulation and Computation* **8**(3), 271–290 (https://doi.org/10.1080/03610917908812119).

Pfaffenberg, R. C., and Dielman, T. E. (1991). Testing normality of regression disturbances: a Monte Carlo study of the Filiben tests. *Computational Statistics* **86**, 359–373.

Phillips, P. C. B. (1977). A general theorem in the theory of asymptotic expansions as approximations to the finite sample distributions of econometric estimators. *Econometrica* **45**, 1517–1534 (https://doi.org/10.2307/1912315).

Rahman, M., and Chakrobarty, S. (2004). Tests for uniformity: a comparative study. *Journal of the Korean Data & Information Science Society* **15**(1), 211–218.

Rahman, M., Pearson, L. M., and Heien, H. C. (2006). A modified Anderson–Darling test for uniformity. *Bulletin of the Malaysian Mathematical Sciences Society* **29**(1), 11–16.

Rebonato, R. (2001). *Managing Model Risk: Handbook of Risk Management*. FT/Prentice Hill.

Rincon-Gallardo, S., Quesenberry, C. P., and O’Reilly, F. J. (1979). Conditional probability, integral transformations and goodness-of-fit tests for multivariate normal distributions. *Annals of Statistics* **7**(5), 1052–1057 (https://doi.org/10.1214/aos/1176344788).

Sargan, J. D., and Satchell, S. E. (1986). A theorem of validity for Edgeworth expansions. *Econometrica* **54**, 189–213 (https://doi.org/10.2307/1914165).

Scholz, F. W., and Stephens, M. A. (1987). K-sample Anderson–Darling tests. *Journal of the American Statistical Association* **82**, 918–924 (https://doi.org/10.1080/01621459.1987.10478517).

Shervish, M. (1997). *Goodness-of-Fit Tests on a Circle: Theory of Statistics*. Springer Series in Statistics, Volume 48. Springer.

Smirnov, N. V. (1948). Tables for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**, 279 (https://doi.org/10.1214/aoms/1177730256).

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* **69**, 730–737 (https://doi.org/10.1080/01621459.1974.10480196).

www.risk.net/journals
Stephens, M. A. (1986). Tests based on EDF statistics. In Goodness-of-Fit Techniques, d’Agostino, R. B., and Stephens, M. A. (eds), pp. 97–193. Marcel Dekker, New York (https://doi.org/10.1201/9780203753064-4).

Tsumagari, M. (2018). Backtesting for counterparty credit risk. Working Paper, Social Science Research Network (https://doi.org/10.2139/ssrn.3154989).

Watson, G. S. (1961). Goodness-of-fit tests on a circle. Biometrika 48, 109–114 (https://doi.org/10.1093/biomet/48.1-2.109).

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9, 60–62 (https://doi.org/10.1214/aoms/1177732360).