BACKTESTING VOLATILITY ASSUMPTIONS USING OVERLAPPING OBSERVATIONS

MICHAEL A. CLAYTON

Abstract. In this paper the ability of a variety of backtesting experiments to identify a model with misspecified volatility is examined. This quantitative testing assumes five years of risk factor observations, considers overlapping and non-overlapping backtest observations with horizons out to a year, and make use of Kolmogorov-Smirnov, Anderson-Darling and a likelihood ratio test statistics.

In doing so the 'discriminatory power' of a test is defined, which is related to the average probability of correctly rejecting a model that is misspecified in a specific way, allowing tests to be quantitatively ranked in terms of this power. This is illustrated using a normal model with volatility misspecified by up to 25%, and it is shown the likelihood ratio test statistic is the most powerful of the considered test statistics for this purpose.

It is then demonstrated that test statistics that are adjusted for the correlation structure arising from the the use of overlapping return observations are more powerful than their unadjusted versions. The result of this analysis is that the adjusted version of the likelihood ratio test statistic is the most powerful statistic to identify misspecified volatility. These adjusted test statistics are shown to have comparable discriminatory power to the (non-overlapping) 1-day backtest experiments, whereas overlapping experiments with the unadjusted statistics have a discriminatory power that rapidly deteriorates with increasing overlap.

Contents

1. Introduction

In this paper we work within the Probability Integral Transform ("PIT") framework for backtesting risk factor forecasting models used in derivative counterparty credit risk models [10],[11]. When used in a regulatory capital setting such models are required to be quantitatively tested on an ongoing basis as part of a backtesting framework; see, for example [8] and references therein.

In this approach risk factor return observations from a defined time sequence are first standardized by using the cumulative distribution function of the assumed model to convert them into uniform variates. A test statistic is defined to be a function of these standardized observations, and the value of the test statistic from the observed return series is converted into the p-value of the test using the distribution of the test statistic under the assumed model. If the p-value is greater than a chosen confidence level then the model has failed the quantitative backtest. We consider a test (or 'backtest') to consists of a combination of:

Date: September 25, 2019; Revision: 214.
Key words and phrases. Derivative Counterparty credit Risk Backtesting; Overlapping Observations; Discriminatory Power.

Experiment: The temporal structure of the observations used for the backtest. For example, how often observations for backtesting are made and over what time horizons.

Test Statistic: The particular function used to convert a sequence of observations into the value of a statistic to be used for testing.

It will be demonstrated that the choice of both components of a test can have a material impact on the ability of the test to identify a particular model defect. It is also shown that for longer horizon tests, test statistics that are modified to account for the correlation structure of the overlapping observations can be materially more powerful in detecting volatility misspecification.

Two standard distributional tests will be considered here: the Kolmogorov-Smirnov ("KS") and Anderson-Darling ("AD") tests, as well as a particular type of Likelihood Ratio ("LR") test statistic defined in Section 2.3. For a volatility misspecification we will generally see that using the LR statistic results in the most powerful tests whereas the Kolmogorov-Smirnov statistic the least powerful. The same is true for the modified statistics used for overlapping observations.

We will specifically be considering the ability of such tests to identify cases where the assumed model underestimates volatility of a normal diffusion model. The setting is admittedly simplistic, but was chosen to keep the analysis as simple and transparent as possible while illustrating the adopted approach to backtesting. We nevertheless feel that it is relevant, it is difficult to see how one would argue that a well-designed backtesting process need not be effective in this setting as well.

It is a general problem when backtesting the forecasting models used in a counterparty credit risk setting, that a technical failure of a particular test does not necessarily indicate which feature of a model is misspecified. For example, if there is a technical failure of a KS test then what aspect of the model is incorrect? Is it a model parameter calibration issue? Is it a structurally incorrect model? Is it a data quality issue and therefore not indicative of a misspecified model? The main motivation for introducing discriminatory power testing is that we can use an alternative models with a specific defects to help identify the most powerful tests to identify that defect, and when a model is run against a suite of tests then the tests that have the largest technical failures suggest the type of model defect that is present.

In the remainder of this section we introduce the experiments that we will be considering (Section 1.1) as well as some required details of a normal diffusion model (Section 1.2). In Section 2 we then make explicit the standard test statistics that will be considered, and in Section 3 introduce a quantitative framework to determine a discriminatory power that indicates the power of a given experiment-test statistic combination to correctly reject a misspecified model. Having shown that the standard test statistics perform poorly for longer horizon tests (overlapping or non-overlapping), in Section 4 we introduce modified test statistics that are nearly as powerful when doing overlapping backtesting as for one day backtesting. A summary and discussion of possible extensions of this work is provided in Section 5.

1.1. **Backtest Experiments.** An 'experiment' consists of a specification of the temporal structure of risk factor observations that are used by the backtest: the number of risk factor observations (N_{obs}) in the observation window (the size of the historical window of risk factor observations used for the backtest), the number of days between forecast initialization dates (d) (i.e., restart frequency), the number of days in the backtest horizon (h), and as a consequence of these the number of backtest windows (N_{init})[1]. An illustration of an experiment with $N_{obs} = 10$, $N_{init}(2,5) = 3$, $d = 2$ and $h = 5$ is given in Figure 1. For the purposes of this document we will

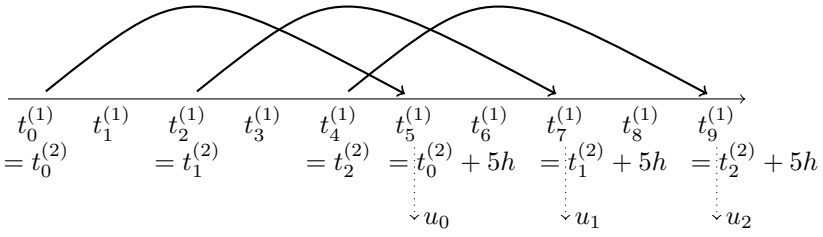

FIGURE 1. Illustration of an experiment with $N_{obs} = 10$, $N_{init}(2,5) = 3$, $d = 2$ and $h = 5$.

[1]As described here each day is given the same amount of market time. In practice this need not be the case, but for maximum transparency it is worthwhile to define the structure of an experiment in terms of whole numbers of a standard interval.

only consider experiments that have a fixed backtest horizon, with results being aggregated over initialization dates [8]. There is nothing preventing the same approach from being used for other types of experiments, but there appears to be little motivation for other aggregations at this time.

Generalizing this to forecast initializations spaced d-days apart and backtest window h days long, the backtest forecasts begin at time:

$$t_i^{(d)} = t_0 + (i-1)d\delta \tag{1.1}$$

and taking into account the backtest window size h, we can fit $N_{init}(d,h)$ windows into the observation window, where ($\lfloor x \rfloor$ is the largest integer less than or equal to x):

$$N_{init}(d,h) = \lfloor \frac{N_{obs} - 1 - h}{d} + 1 \rfloor \tag{1.2}$$

We omit the dependence on N_{obs} since the size of the observation window will be the same for all experiments considered.

The regulatory requirements specifically require testing out to (at least) a 1 year horizon, and also suggest that margin period of risk forecasts should be tested out to a year. This suggests considering the use of forward-starting backtest windows, where observations of the risk factor are made at the beginning and end of the backtest window, which is displaced forward in time from the start of the forecast. Such experiments would allow an explicit testing of the model forecasts for forward margin periods of risk; this is left to future work.

In this work we will report results for backtest experiments defined in Table **??**, which cover various horizons out to one year, as well as some that correspond to standard margin periods of risk consistent with the standardized approach requirements [9]. For all experiments considered here the backtest window will start at the initialization date, and backtests will either be initialized daily ($d = 1$) and will therefore be overlapping (for $h > 1$) or with initialization spacing equal to the backtest horizon ($d = h$) in which case the observations will be non-overlapping. For simplicity we adopt a 250-business day year, and define the year fraction for a single

Horizon			Non-Overlapping				Overlapping			
interval	h	Δt	N_{obs}	N_{init}	d	h	N_{obs}	N_{init}	d	h
1-day	1	1δ	1251	1250	1	1	1251	1250	1	1
5-day	5	5δ	1251	250	5	5	1251	1246	1	5
10-day	10	10δ	1251	125	10	10	1251	1241	1	10
14-day	14	14δ	1251	89	14	14	1251	1237	1	14
1-month	21	21δ	1251	59	21	21	1251	1230	1	21
3-month	62	62δ	1251	20	62	62	1251	1189	1	62
6-month	125	125δ	1251	10	125	125	1251	1126	1	125
1-year	250	250δ	1251	5	250	250	1251	1001	1	250

TABLE 1. The experiments used for testing. Also shown here is N_{init} which is the number of backtest windows in the resulting experiment.

business day as:

$$\delta = \frac{1}{250}. \tag{1.3}$$

The observation window will always correspond to 5 years of daily returns, so the number of 1-day returns is one less than the size of the observation window: $N_{init}(1,1) = 5 \times 250 = 1250 = N_{obs} - 1$. This corresponds to a standard error on a volatility estimate (using 1-day return observations) of: $\sigma/\sqrt{2N_{init}(1,1)} \sim 2\% \times \sigma$, so we therefore expect to be able to resolve volatility differences of a few percent. For all experiments we will fit as many backtest windows as possible into this fixed observations window, always aligning the first initialization date with the first risk factor observation.

1.2. **Normal Model and Volatility Misspecification.** Throughout we will refer to the currently adopted model that is being tested for possible misspecification as the "Null model", and the model it is being compared to as the "alternative model". To keep the setting as simple as possible, the Null model is assumed to be a

normal diffusion model with zero drift and constant volatility[2]:

$$dX_t = \sigma dW_t, \tag{1.4}$$

with dW_t a standard Wiener process. The 1-day returns are then:

$$X_i^{(1,1)} = X_{t_i^{(1)}} - X_{t_{i-1}^{(1)}} = \sigma\sqrt{\delta} W_i^{(1,1)}, \quad i = 1, \ldots, N_{init}(1,1), \tag{1.5}$$

where the 1-day standardized drivers are:

$$W_i^{(1,1)} = \frac{1}{\sqrt{\delta}} \int_{t_{i-1}}^{t_i} dW_s \sim N(0,1). \tag{1.6}$$

For a generic experiment the return for the i-th backtest window in a general experiment is:

$$X_i^{(d,h)} = X_{t_i^{(d)}+h\delta} - X_{t_i^{(d)}} = \sigma\sqrt{\delta h} W_i^{(d,h)}, \quad i = 1, \ldots, N_{init}(d,h), \tag{1.7}$$

where the standardized drivers can be written as a sum over the single day drivers as:

$$W_i^{(d,h)} = \frac{1}{\sqrt{h}} \sum_{j=(i-1)d+1}^{(i-1)d+h} W_j^{(1,1)} \sim N(0,1), \quad i = 1, \ldots, N_{init}(d,h). \tag{1.8}$$

Using this the correlation between two of these standard drivers can be determined:

$$\rho_{ij}^{(d,h)} = E[W_i^{(d,h)} W_j^{(d,h)}] = \begin{cases} 1 - \frac{|i-j|d}{h} & |i-j|d < h \\ 0 & \text{otherwise} \end{cases}. \tag{1.9}$$

From this we see that when the backtest windows are non-overlapping $h \leq d$ then the resulting observations are independent and the $W^{(d,h)}$ could equivalently be simulated as independent normal variates (or, equivalently, the u_i defined below as independent uniform variates). When they are correlated ($h > d$) then $W^{(d,h)}$ could be simulated directly as normal variates with correlation given by (1.9) (or, the u_i drawn from a normal copula with this correlation). When generating results produced for this paper we generate risk factor paths based on simulated single day drivers $W^{(1,1)}$. Doing so means that the same paths can be used for all experiments, and the test statistic disributions that result capture the correlation structure of the Null model as well.

The alternative model is a normal diffusion model with volatility that is larger by a multiplicative factr λ, that is: $\sigma \to \lambda\sigma$:

$$d\hat{X}_t = \lambda\sigma d\hat{W}_t, \tag{1.10}$$

Paths from this model are simulated in the same way as described above. Note that although it would be possible to use the same paths for both models in this simple setting, since this ii not generally possible we will be simulating independent paths fro the Null and alternative models.

For a given backtest window, the standardized observations from the time series are the uniform values u_i (or "u-values") that are defined as the cumulative probability that the risk factor would be less than or equal to the observed value conditional on the value of the risk factor at the beginning of the window, that is, using the Probability Integral Transform ("PIT")[3]:

$$u_i = \text{Prob}\left[X < X_{t_i^{(d)}+h\delta} | X_{t_i^{(d)}}; \sigma\right]. \tag{1.11}$$

Note that these u_i's are always calculated assuming that the Null model is correct, so the above is used to calculate standardized observations from both the Null and alternative model.

These standardized observations are used since often more direct observations of a risk factor (for example, a return or realized volatility) are dependent on the state of the market at the start of the observation window. In making use of the PIT to calculate the u_i's then much of this conditionality is removed in transforming the observation to a uniform distribution–assuming that the adopted model and parameters are correct. The observation sequence of a time series of the risk factor, either observed or simulated, is the collection of u_i

[2]The volatility is set equal to 10% when producing results, but since it cancels from all calculations the actual value is not relevant

[3]In an earlier version of this document these quantities were referred to as 'p-values'. Although they are defined similarly and could theoretically be used as p-values in some testing, we follow the the standard literature and reserve the term 'p-value' for the cumulative probability of finding a value less than equal to an observed value of a test statistic assuming the Null condition; see Section 2.

calculated from all backtest windows. Note that the u_i are ordered, although the in many cases the tests statistics may not depend on the order.

For the Null model this is calculated from the return observations $X_i^{(d,h)}$ as:

$$u_i = N\left(\frac{X_i^{(d,h)}}{\sigma\sqrt{h\delta}}\right) = N\left(W_i^{(d,h)}\right), \quad i = 1, \ldots, N_{init}(d,h). \tag{1.12}$$

When simulating a sequence of observations that would be consistent with a realization of a risk factor path from the Null model, for each path the daily drivers $W_i^{(1,1)}$ are simulated, converted into the standardized drivers as required using (1.8), which are then converted into the risk factor returns $X_i^{(d,h)}$ using (1.7), leading to the second form in the above equation.

For the alternative model, for each path a separate set of daily drivers $\hat{W}_i$ are simulated, converted into the standardized drivers as required using (1.8), these are turned into the risk factor returns $\hat{X}_i^{(d,h)}$ using the scaled volatility:

$$\hat{X}_i^{(d,h)} = \lambda\sigma\sqrt{\delta h}\hat{W}_i^{(d,h)}, \tag{1.13}$$

which are then turned into u-values $\hat{u}_i$ using the Null model distribution function:

$$\hat{u}_i = N\left(\frac{\hat{X}_i^{(d,h)}}{\sigma\sqrt{h\delta}}\right) = N\left(\lambda\hat{W}_i^{(d,h)}\right), \quad i = 1, \ldots, N_{init}(d,h). \tag{1.14}$$

A simulation of the distribution of the $\hat{u}_i$ from non-overlapping intervals from the correctly specified model ($\lambda = 1.0$) and with increasing misspecifications: $\lambda = 1.05, 1.10, 1.25$ are shown in Figure 2. From this it can be

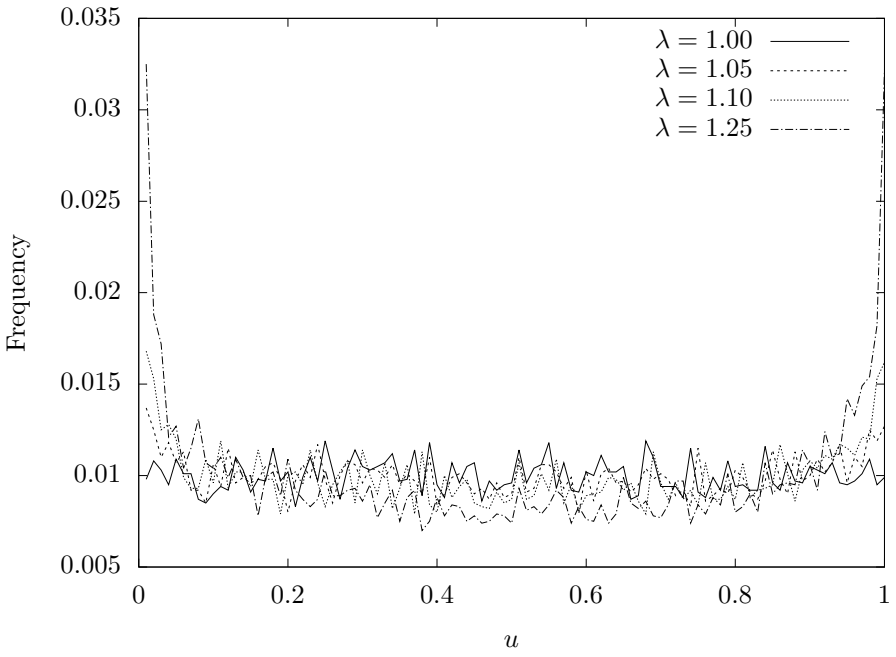

FIGURE 2. The simulated distribution of the u-values for several values of the misspecification. Histogram buckets are 1% wide and 10,000 simulations are used, therefore 100 values are expected per bucket with an expected statistical noise of $\sqrt{100} = 10$ corresponding to a frequency of 0.1%.

seen that when $\lambda > 1$ the $\hat{u}_i$'s will more often appear towards the ends of the unit interval, that is, there will be more observations near zero and one than would be expected if the they were uniformly distributed on the interval.

It is the degree of this non-uniformity that the standard distributional tests examine (see Section 2), and where it is sufficiently large then one perhaps concludes that the Null model is incorrect. Note though, that for

a small sample where the observations are correlated, it will more often be the case that the distribution of the u_i's will appear non-uniform even when the volatility is not misspecified since a single value near an endpoint will most likely be accompanied by others as a result of the correlation. It is exactly this correlation effect that we examine in Section 4, showing that a better choice of statistic can more easily distinguish between real and statistical non-uniformities.

2. Test Statistics and Distributions

In this setting a "confidence level" p_c is chosen ($p_c = 95\%$ or 99% are common) beyond which the assumed Null model (in this case a choice of forecasting model and parameter calibration scheme) is said to experience a technical failure. The expression 'technical failure' is used here since it is often the case that a 'failed' quantitative test does not ultimately indicate a problem with the model specification, and therefore we feel that it is worthwhile to distinguish between the outcome of the quantitative or technical testing from the resulting conclusions drawn about a particular model. For example, a typical review of backtesting results may conclude that although quantitative testing resulted in technical failures, these failures were the result of data-related artifacts (either integrity or features) and did not result from a poorly specified model.

The quantitative test is based on the comparison of the value of a test statistic calculated on observed data compared to the theoretical distribution of the test statistic. If the value of the test statistic calculated from the observed risk factor path is beyond the p_c-th percentile of the theoretical distribution then one concludes that there has been a technical failure of the model. That is, one decides that one is willing to accept a Type 2 error rate (the "false positive" rate or probability of rejecting a correct model) of $1 - p_c$, with the idea that such an unlikely observation is due to a misspecified model.

In the context of backtesting counterparty credit risk forecasting models, a test statistic can be thought of as a function of the sequence of standardized observations resulting from the experiment:

$$\mathrm{TS}(u) = \mathrm{TS}(u_1, u_2, \ldots, u_{N_{init}}), \tag{2.1}$$

which returns a single (real) value representing the value of the test statistic calculaed from the particular path. Then if $F(\cdot)$ is the cumulative distribution function of the test statistic, then the p-value of an observation TS_0 is the cumulative distribution function is:

$$p = F(\mathrm{TS}_0) = \mathrm{Prob}_F(\mathrm{TS} < \mathrm{TS}_0). \tag{2.2}$$

The quantitative test is simply whether the value of the test statistic on the observed path is larger than a chosen confidence threshold p_c, that is, a technical failure of a model has occurred if the reported confidence is greater than the threshold:

$$p = F(\mathrm{TS}(u)) \geq p_c. \tag{2.3}$$

Therefore in order to determine whether a technical failure has occurred the distribution of the test statistic under the assumed Null model is required. Although for standard statistics the distribution of the test statistic is known, it is often only know in the asymptotic limit that the number of observations is large, and in many cases of interest the number of observations is not sufficient to support the use of the asymptotic form. In addition, it is often the case that the test statistic distribution is not known analytically for a given choice of Null model, and may also depend on nuisance parameters from the structure of the model itself. For example, for a mean reverting model with overlapping observations the correlation structure – and therefore the distribution of the test statistic – depends on the mean reversion speed.

In order to avoid these issues, all results in this paper were generated from test statistic distributions generated through a Monte-Carlo simulation. The distribution of a test statistic from the Null or alternative model is generated using $N_{paths} = 10,000$ simulated paths from the misspecified model, each path is generated as follows:

(1) Simulate $N_{init}(1,1)$ independent normal variables $W_i^{(1,1)}$.
(2) Convert these simulated variables into the $N_{init}(d,h)$ variables $W_i^{(d,h)}$ required for the backtest experiment using (1.8)[4].
(3) Calculate the $N_{init}(d,h)$ risk factor returns: $X_i^{(d,h)}$ using (1.7) for the Null model, or (1.13) for the alternative model.
(4) Calculate the $N_{init}(d,h)$ u-values: u_i using (1.12).

[4]As an alternative to the first two steps one could generate correlated variables directly from a normal copula with correlation matrix given by (1.9).

(5) Calculate the value of the test statistic: $T(u_1, u_2, \ldots, u_{N_{init}(d,h)})$.

The resulting N_{paths} simulated values represents the distribution of the test statistic, which is then used to determine whether the reported p-value is above the chosen threshold (2.3).

In the following sub-sections we provide details of the standard test statistics we make use of in this paper. Note that for the calculation of many test statistics it is useful to define the 'order statistics' of the standardized observations, that is, a re-ordering of the u-values so that they are non-decreasing:

$$\tilde{u}_{i-1} \leq \tilde{u}_i. \tag{2.4}$$

It will also be convenient to define the sorted and unsorted 'z-values', which are the u-values converted into standard normal variables:

$$z_i = N^{-1}(u_i), \quad \tilde{z}_i = N^{-1}(\tilde{u}_i). \tag{2.5}$$

Clearly other distributions could be used for to define similar quantities.

2.1. **Kolmogorov-Smirnov (KS).** The Kolmogorov-Smirnov ("KS") statistic tests the u-values against a uniform distribution [4]:

$$\mathrm{KS}(u) = \max(D_+, D_-), \tag{2.6}$$

where:

$$D_+ = \max_i \left\{ \frac{i}{N_{init}} - \tilde{u}_i \right\}, \quad D_- = \max_i \left\{ \tilde{u}_i - \frac{i-1}{N_{init}} \right\}. \tag{2.7}$$

The simulated test statistic distribution using $N_{paths} = 10,000$ simulated paths for the 1-day backtest experiment is shown in Figure 3. Shown there are the distributions assuming that the model is correctly specified ($\lambda = 1.0$), as well as the distributions where the correct model has a volatility that is 5%, 10% and 25% larger that assumed. Note that while the distribution for $\lambda = 1.25$ is well-separated from that of the correctly specified model and therefore the test should easily identify cases with this large an overstatement, there remains significant overlap for the other cases.

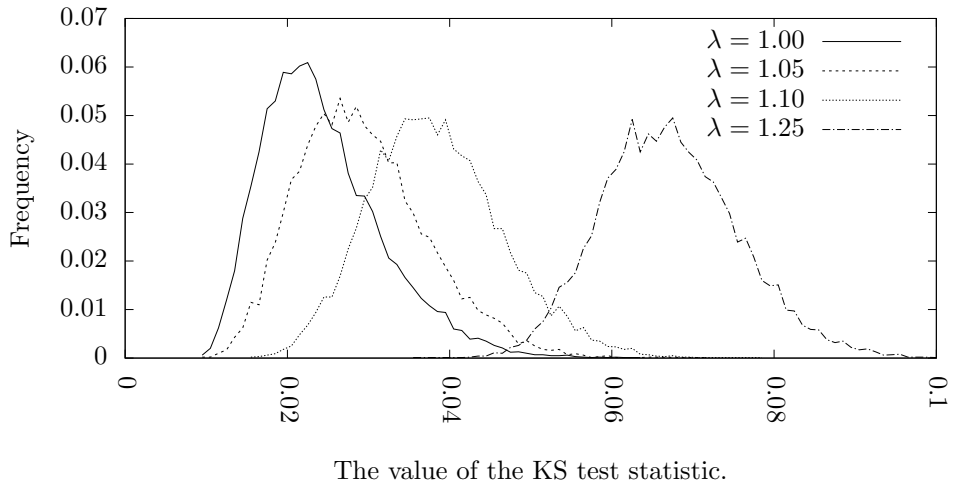

FIGURE 3. The distribution of the KS test statistic for the correctly specified model ($\lambda = 1.00$) as well as when the model has a volatility that is overstated by 5% ($\lambda = 1.05$), 10% ($\lambda = 1.10$) and 25% ($\lambda = 1.25$).

2.2. **Anderson-Darling (AD).** The Anderson-Darling ("AD") statistic [4] is:

$$\mathrm{AD}(p) = -N_{init} - \frac{1}{N_{init}} \sum_{i=1}^{N_{init}} \left[(2i-1) \ln(\tilde{u}_i) + (2(N_{init} - i) + 1) \ln(1 - \tilde{u}_i) \right]. \tag{2.8}$$

This test statistic is known to be more sensitive to differences near the endpoints of the distribution, and we therefore expect (and observe) that it will be a more powerful test against misspecified volatility. Note that one

must be a bit careful how one deals with extreme returns since a u-value that is numerically equal to zero or one will cause the value of the statistic to diverge.

The simulated test statistic distribution using $N_{paths} = 10,000$ simulated paths for the 1-day backtest experiment is shown in Figure 4[5]. Shown there are the distributions assuming that the model is correctly specified ($\lambda = 1.00$), as well as the distributions where the correct model has a volatility that is 5%, 10% and 25% larger that assumed. Note that while the distribution for $\lambda = 1.25$ is well-separated from that of the correctly specified model and therefore the test should easily identify cases with this large an overstatement, there remains significant overlap for the other cases.

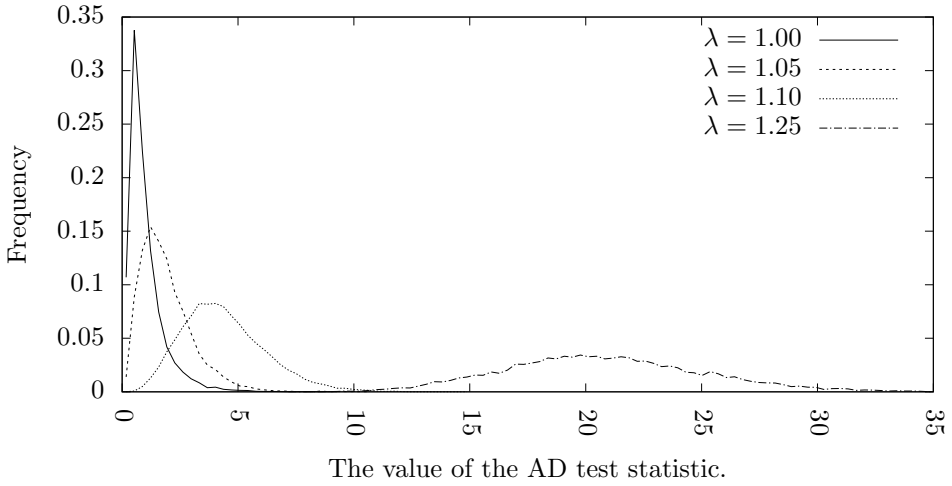

FIGURE 4. The distribution of the AD test statistic for the correctly specified model ($\lambda = 1.00$) as well as when the model has a volatility that is overstated by 5% ($\lambda = 1.05$), 10% ($\lambda = 1.10$) and 25% ($\lambda = 1.25$).

2.3. **Likelihood Ratio (LR).** This test statistic specifically tests whether the z-values come from a standard normal distribution in that it is the likelihood ratio from the misspecified model (standardized z-values) compared to the alternative that the volatility is not equal one. Although it is possible to account for the possibility of a misspecified mean in the test, as designed in (2.9) we are only testing whether the volatility is misspecified.

The statistic is a likelihood ratio test statistic [7] that specifically tests whether the u-values are related to standard normal z-values [1]:

$$\text{LR} = 2\ln P(\mu = \hat{\mu}, v = \hat{v}) - 2\ln P(\mu = \hat{\mu}, v = 1) = -N_{init}\left(1 - \hat{v} + \ln(\hat{v})\right), \tag{2.9}$$

where the variance is:

$$\hat{v} = \frac{1}{N_{init}}\sum_{i=1}^{N_{init}} z_i^2 - \left(\frac{1}{N_{init}}\sum_{i=1}^{N_{init}} z_i\right)^2. \tag{2.10}$$

The final form comes from the log-likelihood function for a normal process:

$$\ln P(\mu, v) = -\frac{1}{2v}\sum_{i=1}^{N_{init}}(z_i - \mu)^2 - \frac{N_{init}}{2}\ln(v) - \frac{N_{init}}{2}\ln(2\pi) \tag{2.11}$$

with resulting Maximum Likelihood estimators for the mean and variance:

$$\hat{\mu} = \frac{1}{N_{init}}\sum_{i=1}^{N_{init}} z_i, \quad \hat{v} = \frac{1}{N_{init}}\sum_{i=1}^{N_{init}} z_i^2 - \left(\frac{1}{N_{init}}\sum_{i=1}^{N_{init}} z_i\right)^2. \tag{2.12}$$

[5]An earlier version of this paper was showing results from an adjusted Anderson-Darling test statistic. Although the general results were not affected, the standard form of the statistic is now being used.

It is also possible to design statistics that specifically test whether the volatility is greater or less than one rather than simply not equal to one, which can be useful in practice to test whether the volatility of a model is under- or over-specified.

The simulated test statistic distribution using $N_{paths} = 10,000$ simulated paths for the 1-day backtest experiment is shown in Figure 5. Shown there are the distributions assuming that the model is correctly specified ($\lambda = 1.00$), as well as the distributions where the correct model has a volatility that is 5%, 10% and 25% larger that assumed. In this case the distributions for both $\lambda = 1.25$ and $\lambda = 1.1$ are well-separated from that of the correctly specified model, with a less significant overlap for the other case. This suggests that this test statistic should have a higher power to discriminate between models with correctly specified volatility, and models with volatility parameter that is understated.

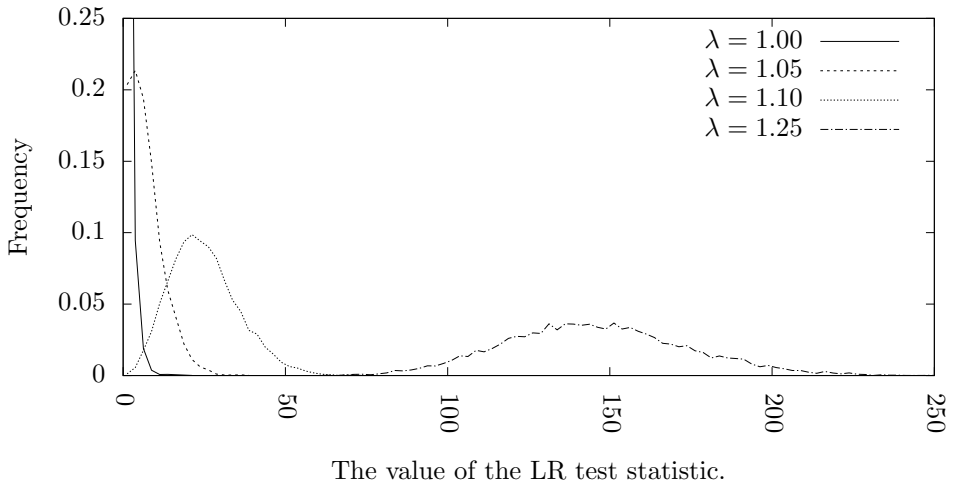

FIGURE 5. The distribution of the LR test statistic for the correctly specified model ($\lambda = 1.00$) as well as when the model has a volatility that is overstated by 5% ($\lambda = 1.05$), 10% ($\lambda = 1.10$) and 25% ($\lambda = 1.25$). Note that the y-axis has been truncated in this plot to make it easier to see the structure of the distributions..

3. Discriminatory Power (DP)

The purpose of this section is to introduce a quantitative framework for ranking the ability of different tests to discriminate between a correctly specified model and a misspecified model with a particular defect. Part of the motivation is to arrive at a relatively transparent framework to quantitatively validate some of the folk wisdom that seems to surround the backtesting of counterparty credit risk models, in particular the usefulness of experiments with overlapping observations of a risk factor. We also want to arrive at a quantitative ranking that is not dependent on a particular path (or realization from the model) to draw inferences from, since simply running a backtest on a particular path can give misleading results (for example, if the data generating process has a jump component that does not happen to have any jumps on the path in question).

This is accomplished through the use of the True Positive Rate ("TPR") curve, following a fairly well known approach [6]. This paper can be thought of as advocating for its use in backtesting counterparty credit risk forecasting models. Throughout we will be working with the normal model with misspecified volatility described in Section 1.2, but the same approach can be adopted for other models and defects as well.

Note that running multiple tests for a given model defect is, of course, also a problem since it is difficult to determine at what confidence a suite of tests is operating at if one specifies a confidence interval that is imposed separately on each test. For example, when running a large number of tests that are highly correlated and most likely to all experience a technical failure at the same time, then the confidence threshold p_c imposed on a single test is likely to be fairly close to the confidence of the entire suite of tests. However, when running a large number of nearly independent tests, then even when setting the threshold at a high level, the probability of some tests having technical failures purely for statistical reasons can be high. This will not be examined

further here [3], except to suggest that once the tests with the highest discriminatory power are identified, the suite of tests used for quantitative backtesting should be designed carefully.

As described here the framework implicitly assumes that the distribution of the test statistics are (in some sense) increasing functions of the misspecification [5, Chapter 34.11]. In particular the framework is designed to work for single-sided tests where the model is considered to have a potential defect when the value of the test statistic is in the right tail of the distribution, and requires some adjustment when a model is considered as defective when the value of the test statistic lies at either extreme of the test statistic distribution. This is often easily remedied by introducing separate tests for either tail, introducing truncated test statistics, or modifying them so that the negative and positive tail are combined. Ideally we would also like to guarantee that as the number of observations increases towards infinity, the test becomes perfectly distinguishing, that is: $TPR \overset{N\to\infty}{\to} 1$; this is left for future investigation.

We will start by defining the True Positive Rate ("TPR") curve (or Receiver Operating Characteristic ("ROC") curve [6]), which gives the probability that the test statistic from the misspecified model will be within the tail of distribution from the correctly specified model. To quantify this, start with the cumulative distribution function of the Null model that is to be tested for misspecification (Equation (2.2)):

$$F(x) = \text{Prob}_F(\text{TS} < x), \tag{3.1}$$

and assume that we also have from the distribution function from the alternative model as well:

$$\hat{F}(x) = \text{Prob}_{\hat{F}}(\text{TS} < x). \tag{3.2}$$

The True Positive Rate ("TPR") is defined as the mass of the distribution from alternative model that is above the chosen confidence from the Null model:

$$\text{TPR}(p) = 1 - \hat{F}\left(F^{-1}(p)\right). \tag{3.3}$$

note that this definition means that the TPR curve is non-decreasing as a function of $1 - p$.

The resulting TPR curve for the three test statistics described in Section 2 for the 1-day backtest experiment is shown in Figure 6. For example, the TPR at $p = 5\%$ for the KS statistic is $\sim 10\%$, indicating that the test would only correctly reject the misspecified Null model 10% of the time. On the other hand, the AD statistic has TPR(5%) $\sim 25\%$, so would correctly reject the misspecified Null model 25% of the time, and the LR statistic has TPR(5%) $\sim 70\%$ so is the most powerful test of the three, correctly rejecting the misspecified Null model 70% of the time. This ordering follows the intuition we developed in Section 2. On these plots is also shown the line of no discrimination ("No Disc."), on which the test correctly rejects a misspecified model with the same probability as it would incorrectly reject a correctly specified model (TPR$(p) = 1 - p$), that is, the test it has no discriminatory power. With a more materially under-stated volatility ($\lambda = 1.10$) then we see from Figure 7

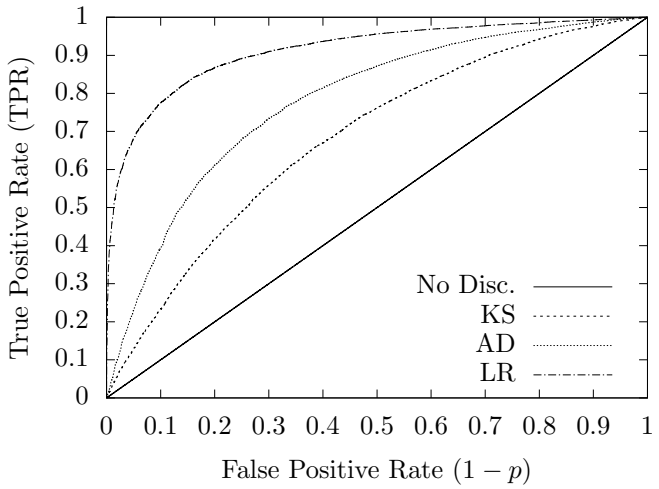

FIGURE 6. The TPR curve for the 1-day backtest experiment, with 5% understated volatility ($\lambda = 1.05$).

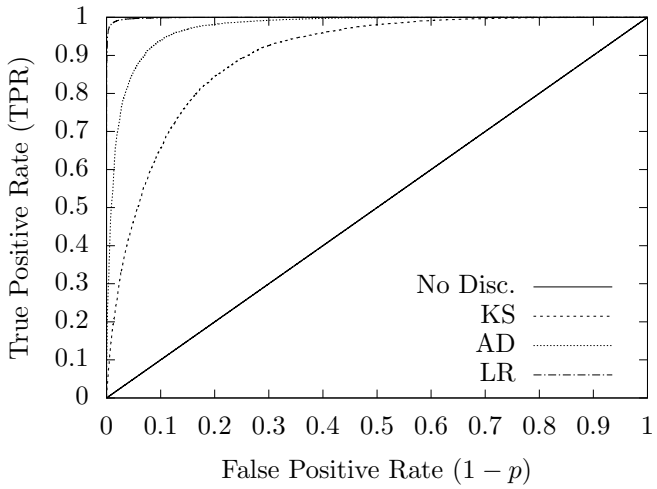

FIGURE 7. The TPR curve for the 1-day backtest experiment, with 10% understated volatility ($\lambda = 1.10$).

that all TPR curves show a larger DP, with the likelihood ratio statistic being the most powerful. When we move to a 10-day non-overlapping observations, then the discriminatory power declines, as shown in Figure 8, and the power of the tests increases moderately when using overlapping observations, as shown in Figure 9.

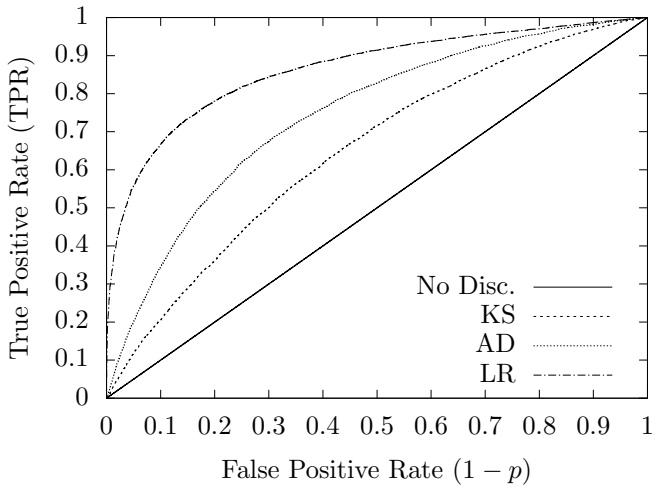

FIGURE 8. The TPR curve for the 10-day, non-overlapping backtest experiment, with 10% understated volatility ($\lambda = 1.10$).

From Table 3 we also see that testing at a higher confidence (99%) leads to testing that correctly rejects the misspecified Null model at lower rates, and therefore has a lower power to identify misspecified models. Although ultimately a technical failure occurrs when the reported p-value is above the chosen confidence level p_c, we are of the opinion that reporting the p-value is a good practice when backtesting, since near failures on some model can aid in identifying potential model misspecifications.

The Discretionary Power ("DP") of a test is defined to be twice the area of the TPR curve above the line with no discriminatory power. We normalize it so that a test with no discriminatory power has a value equal to zero, and perfect discriminatory power equal to one:

$$\mathrm{DP} = 2 \int_0^1 dp \, (\mathrm{TPR}(p) - (1-p)) \, . \tag{3.4}$$

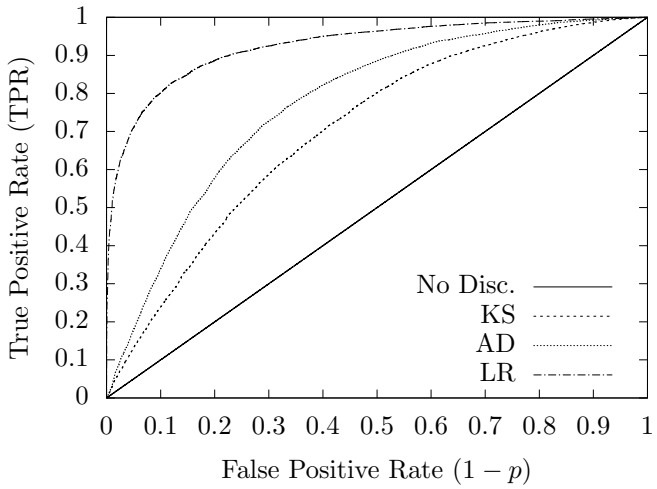

FIGURE 9. The TPR curve for the 10-day, overlapping backtest experiment, with 10% understated volatility ($\lambda = 1.10$).

With this definition the fact that the TPR curve is non-decreasing, if we know $\mathrm{TPR}(p)$ at confidence p then the discriminatory power satisfies[6]:

$$2\,p\,\mathrm{TPR}(p) - 1 \leq \mathrm{DP} \leq 1 - 2\,(1-p)(1-\mathrm{TPR}(p)). \tag{3.5}$$

This can be transformed into a condition on the true positive rate:

$$\max\left(0, 1 - \frac{1-\mathrm{DP}}{2(1-p)}\right) \leq \mathrm{TPR}(p) \leq \min\left(1, \frac{\mathrm{DP}+1}{2p}\right). \tag{3.6}$$

This range is fairly wide in general, however it does give us a reasonable target for the discriminatory power that will guarantee a large TPR: If we require that the DP $\geq 99\%$ then the above results in: $\mathrm{TPR}(5\%) \geq 90\%$, and therefore the test is guaranteed to correctly reject the Null model 90% of the time. This result is independent of the model or test statistic.

Quantitative results for the TPR at 95% and 99% are shown in Table 2 and Table 3, respectively, and the discriminatory power in Table 4. The general trends observed are as expected:

(1) A larger misspecification (corresponding to a larger value of λ here) results in a higher DP for a given experiment. This is as expected since it implies that a volatility that is misspecified by a larger amount will have a higher probability of being detected by the tests.
(2) For non-overlapping experiments the DP for a given statistic declines with increasing $d = h$, which makes sense intuitively since there are fewer observations of the process.
(3) For overlapping experiments the DP for a given statistic also declines with increasing h, but is generally moderately higher than the corresponding non-overlapping experiment. This is somewhat discouraging since there seems to be some feeling in practice that the use of overlapping observations should lead to stronger tests by overcoming the lower statistics that result when using non-overlapping observations. We will see in the following section that this can be the case, but different test statistics must be used to accomplish it.

Clearly one cannot always expect a test to have a large DP: for example, if the volatility is misspecified by a very small amount or the observation window is very short. In such a situation no tests may do a particularly good job at distinguishing the correct from the misspecified model, but there will nevertheless be a spectrum of tests of varying power, and one should be using tests with higher discriminatory power.

[6]Note that despite the examples seen in this paper, it is not always the case that: $\partial_p^2 \mathrm{TPM}(p) \leq 0$, so no obvious refinement of these bounds are available.

To gain some intuition on how this works, assume that under the Null assumption a test statistic is normally distributed with mean μ and standard deviation σ[7]:

$$T \sim \phi(T; \mu, \sigma). \tag{3.7}$$

This is not always a reasonable assumption, but Figure 3 suggests that it can be in some cases. The p-value of an observed value T is the cumulative distribution function:

$$p(T) = F(T) = N\left(\frac{T-\mu}{\sigma}\right). \tag{3.8}$$

If we likewise assume that under the alternative assumption the test statistic is normally distributed, but with different mean $\tilde{\mu}$ and standard deviation $\tilde{\sigma}$, then the cumulative distribution function has the same form:

$$\tilde{F}(T) = N\left(\frac{T-\tilde{\mu}}{\tilde{\sigma}}\right). \tag{3.9}$$

Using this the TPR can easily be found to be:

$$\mathrm{TPR}(p) = 1 - N\left[-\kappa + \lambda N^{-1}(p)\right], \quad \kappa = \frac{\tilde{\mu}-\mu}{\tilde{\sigma}}, \quad \gamma = \frac{\sigma}{\tilde{\sigma}}. \tag{3.10}$$

The TPR is plotted in Figure 10 for a selection of κ and γ, showing that test statistics that are shifted further under the alternative or are narrower will have a larger TPR.

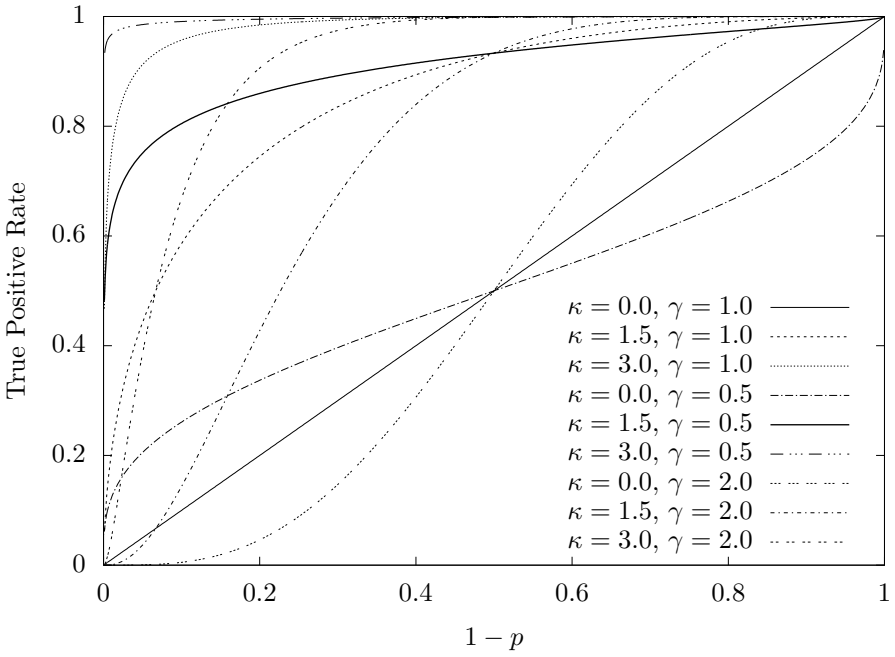

FIGURE 10. True positive rate for normally distributed test statistics; κ indicates the degree to which the misspecification shifts the mean of the distribution and γ the factor by which the width of the distribution has narrowed (see Equation (3.10)).

An integration gives the DP:

$$DP = 1 - 2\int_0^1 dp N\left[-\kappa + \lambda N^{-1}(p)\right] = 1 - 2N\left[-\frac{\tilde{\mu}-\mu}{\sqrt{\sigma^2+\tilde{\sigma}^2}}\right] \tag{3.11}$$

The DP is shown (in Figure 11) to be an increasing function of the quantity: $(\tilde{\mu}-\mu)/\sqrt{\sigma^2+\tilde{\sigma}^2}$, which is the ratio of the mean to the standard deviation of the difference distribution. In order for the DP to reach the

[7]The notation: $x \sim \phi(x; \mu, \sigma)$ indicates that the variable x is normally distributed with mean μ and standard deviation σ.

heuristic target of 99%, we find that:

$$\frac{\tilde{\mu} - \mu}{\sqrt{\sigma^2 + \tilde{\sigma}^2}} \sim 2.6, \tag{3.12}$$

that is, the mean has to be shifted by 2.6 times the joint volatility of the test statistic distributions. Where the

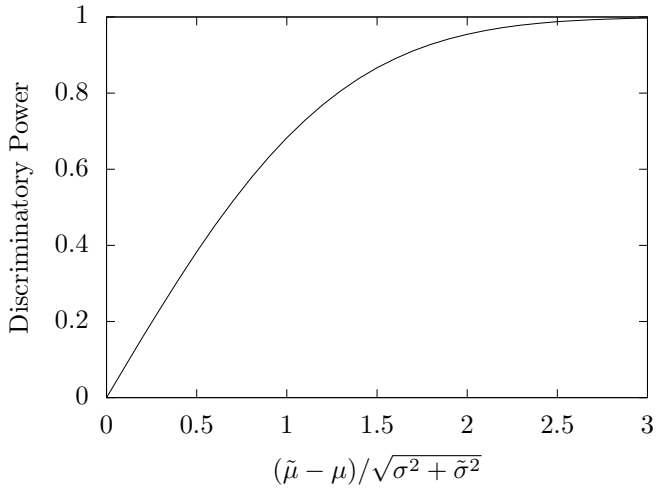

FIGURE 11. Discriminatory power for normally distributed test statistics; see Equation (3.11).

centre of the test statistic hasn't moved: $\tilde{\mu} = \mu$ then the test has zero discriminatory power (DP = 0). Note though that when $\tilde{\sigma} \neq \sigma$ then we can see from Figure 10 that for some values of p the TPR will be above or below the line of no discrimination, illustrating that the DP is an average discriminatroy power.

4. Modified Tests Statistics for Overlapping Observations

In this section we define modified versions of the test statistics defined in Section 2 that are more powerful in identifying models with misspecified volatility when observations are overlapping; the modified distributional statistics: KS_ρ and AD_ρ in Section 4.1 and the modified LR statistic: LR_ρ in Section 4.2. The TRP and DP of the tests using these modified statistics are given in Tables 2,3 and 4, from which we see that the power of these modified tests is comparable to that of 1-day experiments for all horizons, in stark contrast to the un-modified statistics which have power that declines rapidly as the horizon increases.

As a specific example of what this testing demonstrates, consider a model with volatility misspecified by 10% ($\lambda = 1.10$). The results of Table 2 show that if one uses 10-day, non-overlapping observations and is testing at 95%, then the standard KS test will only correctly reject the misspecified Null model 7.9% of the time and the standard AD test 12.9% of the time. Using the same standard tests with overlapping observations does not improve matters much, correctly rejecting the misspecified Null model 8.8% and 11.2% of the time, respectively. In contrast, the LR statistic correctly rejects the misspecified Null model 31.5% of the time for non-overlapping and 41.3% of the time for overlapping observations, a material improvement over the standard KS and AD tests. When all three of these statistics are adjusted for the correlation structure, the TPR materially increases, and the misspecified Null model is correctly rejected 47.0% of the time for KS, 86.7% for AD and 99.6% for LR. It is clear that using overlapping observations and adjusting the statistics for the correlation structure results in the strongest tests for identifying misspecified volatility out of those considered here. The discriminatory power shown in Table 4 for the same experiments shows a DP for the three overlapping, correlation-adjusted statistics of: 79.6% for KS, 94.9% for AD and 99.8% for LR.

From these results it should be clear that the LR_ρ test statistic is the most powerful of the test statistics considered in identifying a model with misspecified volatility. We should note that although the modified test statistics perform better, there is nothing inconsistent or 'incorrect' about using the unmodified statistics – they are simply less powerful for this particular model defect. This situation is similar to parameter estimators in that there are many estimators possible for a given parameter and it is generally preferable to use the most efficient estimator, that is, the one available with the smallest standard error.

4.1. **Modified Distributional Test Statistics (KS_ρ, AD_ρ).** The approach taken to modifying these test statistics is fairly simplistic: given the known correlation structure that results from the overlap structure (1.9), we can convert the correlated z_i's into de-correlated values, then calculate de-correlated u_i's from these, and finally use these values to compute the standard KS and AD test statistics. This is likely more sensible than using the un-modified test statistics with dependent u-values since known results for them assume that the observations are independent, which is very definitely not the case here (see, for example [5]). Heuristically we expect that the distribution of the test statistic will have lower volatility due to the lack of correlation between u-values.

To achieve this, note that an observation of the z-values can be written in terms of de-correlated z-values: $\bar{z}_n$ using the eigenmode decomposition of the correlation matrix as:

$$z_i = \sum_{n=1}^{N_{init}} \sqrt{\lambda^{(n)}} e_i^{(n)} \bar{z}_n, \tag{4.1}$$

where the $e_i^{(n)}$ are the eigenvectors and $\lambda^{(n)}$ the eigenvalues of the correlation matrix (1.9):

$$\sum_{j=1}^{N_{init}} \rho_{ij}^{(d,h)} e_j^{(n)} = \lambda^{(n)} e_i^{(n)}. \tag{4.2}$$

Using orthogonality of the eigenvectors:

$$\sum_{i=1}^{N_{init}} e_i^{(m)} e_i^{(n)} = \delta_{mn}, \tag{4.3}$$

we can convert the z-values into de-correlated z-values using[8]:

$$\bar{z}_n = \frac{1}{\sqrt{\lambda^{(n)}}} \sum_{i=1}^{N_{init}} e_i^{(n)} z_i, \tag{4.4}$$

and then the de-correlated standardized observations are determined from them:

$$\bar{u}_n = N(\bar{z}_n). \tag{4.5}$$

Note that the correlation structure is assumed to be known and given by (1.9) rather than estimated from the data, and therefore the eigenvalues and eigenvectors can be accurately determined once and reused for all simulated paths.

The modified test statistics KS_ρ and AD_ρ are then calculated using the same functions as described in Sections 2.1 and 2.2, but using the de-correlated $\bar{u}_i$'s. That is, the modified test statistics are computed as:

$$\mathrm{KS}_\rho(u) = \mathrm{KS}(\bar{u}), \quad \text{and} \quad \mathrm{AD}_\rho(u) = \mathrm{AD}(\bar{u}). \tag{4.6}$$

As can be seen from Figures 12 and 13, although for both the standard and modified statistics the maxima of the distributions are shifted by roughly the same amount as a result of the misspecification, the distributions of the modified test statistics are materially narrower, with the result that there is less overlap between the correctly specified and misspecified models.

Also see Figure 14 where the TPR curves are shown for the modified test statistics, which should be compared with Figures 8 and 9. From these it is clear that these modified test statistics will be materially more powerful for longer horizon tests. This is seen for the TPR for the modified statistics at 95% and 99% from the lower third of Table 2 and Table 3, respectively, as well as for the discriminatory power in Table 4. Although the DP of the modified tests still declines with increasing horizon (as a result of increasing statistical noise), it remains much closer to the DP of the 1-day backtest results, and is materially larger than the DP for the unmodified statistics on the corresponding experiments.

[8]We note that in some cases there are small eigenvalues, but to date we have not observed sufficient numerical instability to warrant any specific action be taken.

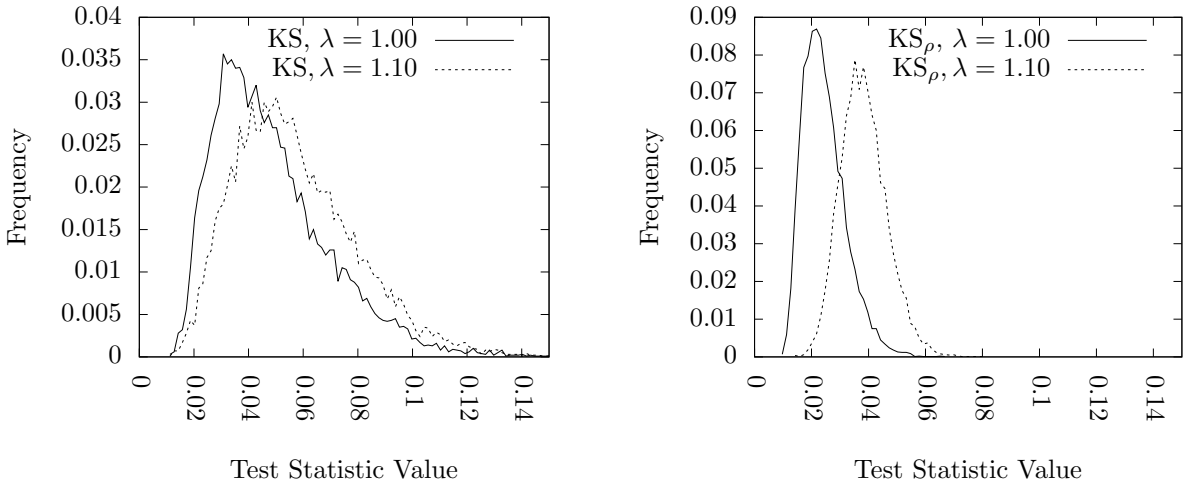

FIGURE 12. The distribution of the standard (KS) and modified (KS$_\rho$) Kolmogorov-Smirnov test statistic for a 10-day overlapping backtest and a 10% under-specified volatility ($\lambda = 1.10$).

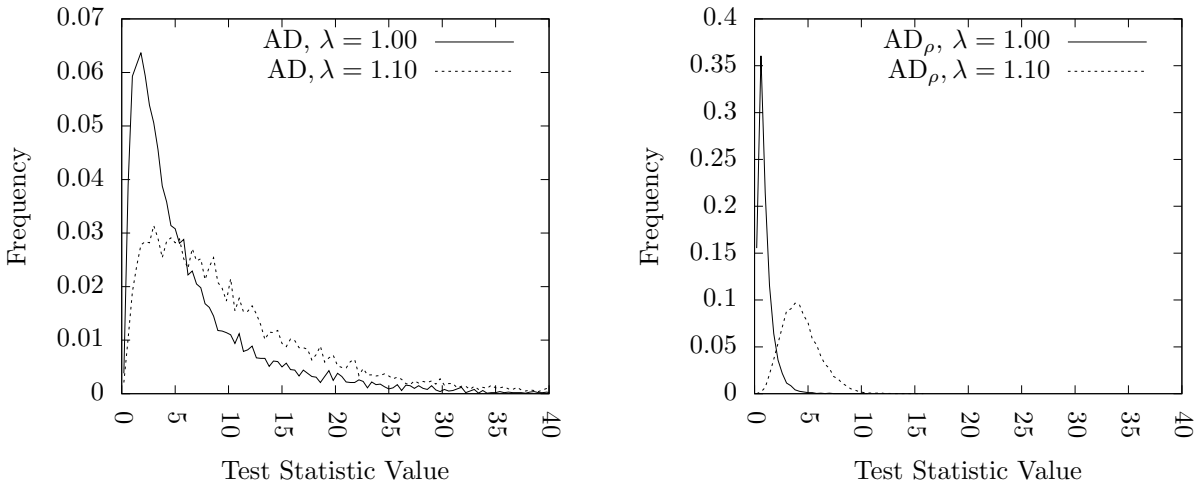

FIGURE 13. The distribution of the standard (AD) and modified (AD$_\rho$) Anderson-Darling test statistic for a 10-day overlapping backtest and a 10% under-specified volatility ($\lambda = 1.10$).

4.2. **Modified Likelihood Ratio Statistic (LR$_\rho$).** For this modified test statistic we simply note that the assumption that the z-values are uncorrelated that was made in Section 2.3 is incorrect. Instead we use a log-likelihood function that accounts for the known correlation structure resulting from an overlap of the observations, namely the log-probability for a sequence of $N_{init}(d, h)$ correlated observations:

$$\ln P_\rho(\mu, v) = -\frac{1}{2v}\sum_{i,j=1}^{N_{init}} \rho_{ij}^{-1}(Z_i - \mu)(Z_j - \mu) - \frac{N_{init}}{2}\ln(v) - \frac{1}{2}\ln(\det \rho_{ij}) - \frac{N_{init}}{2}\ln(2\pi). \tag{4.7}$$

With the correlation structure considered to be known, the maximum-likelihood estimators are straightforward to determine [2][9]; first the mean:

$$\hat{\mu}_\rho = \frac{\sum_{i,j=1}^{N_{init}} \rho_{ij}^{-1} z_j}{\sum_{i,j=1}^{N_{init}} \rho_{ij}^{-1}} \tag{4.8}$$

[9]Note that a closed form solution for ρ_{ij}^{-1} exists and appears in the same reference. In practice we simply solve $\rho x = z$ for $x = \rho^{-1} z$.

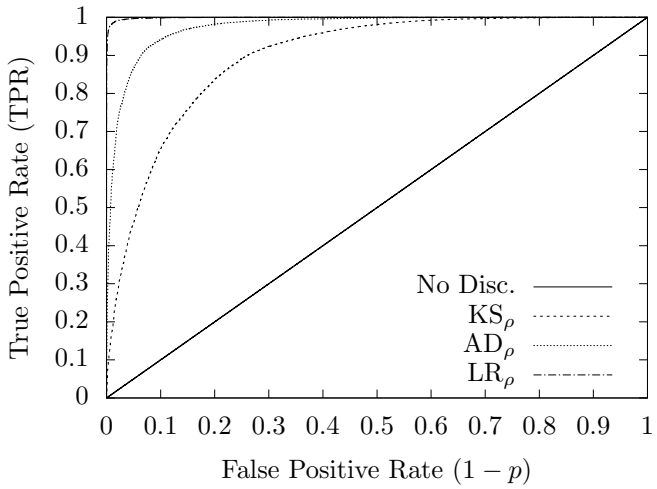

FIGURE 14. The TPR curves for the modified statistics for the 10-day, overlapping backtest experiment, with 10% understated volatility ($\lambda = 1.10$).

and the variance:

$$\hat{v}_\rho = \frac{1}{N_{init}} \sum_{i,j=1}^{N_{init}} \rho_{ij}^{-1} (z_i - \hat{\mu})(z_j - \hat{\mu}). \tag{4.9}$$

The modified test statistic is calculated in exactly the same way as before, using this log-likelihood function:

$$\mathrm{LR}_\rho(p) = 2 \ln P_\rho(\mu = \hat{\mu}_\rho, v = \hat{v}_\rho) - 2 \ln P_\rho(\mu = \hat{\mu}_\rho, v = 1) = -N_{init} \left(1 - \hat{v}_\rho + \ln(\hat{v}_\rho)\right). \tag{4.10}$$

Although looking at Figure 15 does not show that both the modified and un-modified LR statistics are shifted by a similar amount on average, it is nevertheless clear that there is a much smaller overlap for the adjusted statistic. Also see Figure 14 where the TPR curves are shown for the modified test statistics, which should be compared with Figures 8 and 9. The TPR for the modified statistics at 95% and 99% from the lower third of Table 2 and Table 3, respectively, as well as for the discriminatory power in Table 4.

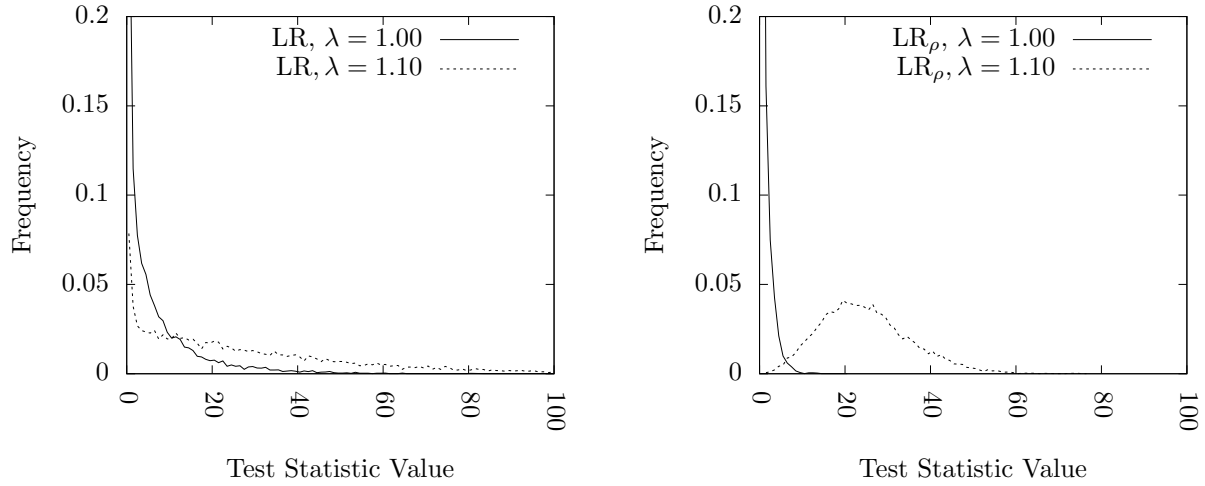

FIGURE 15. The distribution of the standard (LR) and modified (LR$_\rho$) Likelihood Ratio test statistic for a 10-day overlapping backtest and a 10% under-specified volatility ($\lambda = 1.10$). Note that the y-axis has been cut off to make the shape of the distribution easier to see.

5. Comments and Conclusions

This paper has made the following three observations: First, it has been demonstrated that different tests can have materially different sensitivities to model defects. This was shown in Section 2 by considering a normal model with misspecified volatility, with three fairly standard test statistics and a variety of backtest experiments. A poor choice of test and the resulting backtest can have a very low ability to test against even a fairly strongly misspecified volatility.

Second, we have introduced a framework for quantitatively measuring the discriminatory power of a tests (Section 3), which is related to the average probability of corectly rejecting the misspecified model. This can be used to rank tests for their ability identify models with a particular defect. Third, we have used this framework to show that test statistics that are modified to account for the correlation structure of overlapping observations are much more powerful in identifying a misspecified volatility in longer horizon tests (section 4).

It is admittedly not always the case that the correlation structure for the standardized observations is known, and therefore the modified test statistics may be difficult to motivate in some cases. However the real point to be taken away from this work is that using test statistics without examining their ability to test simple model defects can easily lead to the use of very weak tests. In particular, one should not be content with the use of low discriminatory power test statistics for long horizon, overlapping backtesting since there may well be alternative statistics that are materially more powerful.

There are a few extensions to this work:

(1) Where the correlation structure is known (or depends on model parameters in a known way, as for mean reverting models), then the approach described here can be used to identify more effective long horizon tests. Where the correlation structure is not known or known only approximately, then the test statistics may need to be generalized to determine the correlation structure. This may lead to a degradation of the power of the tests.
(2) The discriminatory power analysis can be used to identify more powerful tests for other types of model misspecification: drift, mean reversion, jumps, stochastic volatility, ... as well as calibration and data-related issues.
(3) The frequency and methodology chosen for model recalibration can have a material impact on the quality of forecasting models. It would be useful to design some testing to help identify best practices.
(4) A specific examination of the usefulness of forward starting testing to specifically test model performance in forward margin periods, compared to longer (overlapping) backtest horizons.
(5) Since it is inevitable that multiple tests on a single model will be required when backtesting counterparty credit risk forecasting models, it is important to control the family wide error rate. This issue will be examined by the author in an upcoming paper [3].

References

1. Jeremy Berkowitz, *Testing Density Forecasts with Applications to Risk Management*, (December 2000).
2. Michael A. Clayton, *Parameter Estimation from Overlapping Observations*, MACCI; SSRN-2968896 (2016), 42.
3. _____, *Multiple Testing in Credit Exposure Model Backtesting*, MACCI; TBD (2019).
4. Ralph B. D'Agostino and Michael A. Stephens (eds.), *Goodness–Of–Fit Techniques*, Statistics: Textbooks and Monographs, vol. 68, Marcel Dekker, Inc., New York, 1986.
5. Anirban DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer, 2008.
6. Tom Fawcett, *An Introduction to ROC Analysis*, Pattern Recognition Letters **27 (2006)** (2006), 861–874.
7. William H. Greene, *Econometric Analysis*, eighth ed., Pearson, 2018.
8. Basel Committee on Banking Supervision, *Sound practices for backtesting counterparty credit risk models*, Tech. report, Bank for International Settlements, December 2010.
9. _____, *The standardised approach for measuring counterparty credit risk exposures*, Tech. report, Bank for International Settlements, March 2014 (rev. April 2014).
10. Ignacio Ruiz, *Backtesting Counterparty Credit Risk: How Good is Your Model?*, iRuiz Consulting, version 2.0.1 (2012).
11. Mitsuo Tsumagari, *Backtesting for Counterparty Credit Risk*, SSRN-3154989 (2018).

Toronto, Canada
E-mail address: michael.clayton@sympatico.ca

Non-Overlapping ($d = h$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	12.9%	6.9%	5.7%	5.3%	5.8%	5.6%	5.3%	5.2%
AD	23.0%	8.4%	6.7%	6.3%	6.3%	5.8%	6.0%	6.0%
LR	68.2%	18.3%	11.2%	8.6%	7.1%	5.0%	4.5%	4.5%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	46.8%	11.9%	7.9%	7.1%	7.2%	6.4%	5.8%	5.8%
AD	85.7%	20.1%	12.9%	10.1%	9.7%	7.7%	7.7%	7.6%
LR	99.7%	55.8%	31.5%	23.0%	15.6%	6.9%	5.1%	4.3%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	100.0%	52.9%	26.3%	19.0%	14.3%	9.4%	7.7%	7.4%
AD	100.0%	91.4%	61.4%	45.2%	33.0%	17.3%	14.7%	12.8%
LR	100.0%	99.8%	93.0%	82.1%	64.6%	24.6%	12.0%	5.4%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	12.9%	6.9%	5.6%	5.4%	5.6%	5.3%	5.1%	5.3%
AD	23.0%	7.3%	6.3%	6.2%	5.8%	5.7%	5.7%	5.8%
LR	68.2%	23.9%	13.2%	10.8%	8.4%	5.1%	4.6%	3.8%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	46.8%	13.2%	8.8%	7.9%	7.2%	6.0%	5.6%	5.9%
AD	85.7%	18.0%	11.2%	9.4%	8.2%	7.1%	6.9%	6.8%
LR	99.7%	71.1%	41.3%	30.8%	21.3%	8.3%	5.4%	3.4%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	100.0%	70.6%	35.2%	25.1%	18.5%	9.7%	7.6%	7.3%
AD	100.0%	96.1%	64.1%	45.0%	29.5%	14.4%	11.5%	10.5%
LR	100.0%	100.0%	98.6%	93.6%	80.8%	33.2%	14.3%	5.4%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS_ρ	12.9%	12.5%	12.6%	12.6%	12.5%	13.0%	12.2%	10.7%
AD_ρ	23.0%	24.1%	24.1%	22.4%	22.8%	24.1%	22.5%	18.6%
LR_ρ	68.2%	67.9%	68.0%	67.7%	67.1%	65.8%	63.5%	57.8%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS_ρ	46.8%	46.0%	47.0%	47.0%	46.2%	45.9%	42.9%	36.7%
AD_ρ	85.7%	86.6%	86.7%	85.1%	85.7%	85.2%	83.1%	76.0%
LR_ρ	99.7%	99.7%	99.6%	99.6%	99.6%	99.5%	99.4%	98.5%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
AD_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
LR_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

TABLE 2. The True Positive Rate at 95% for non-overlapping and overlapping experiments.

Non-Overlapping ($d = h$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	3.1%	1.2%	1.0%	1.0%	1.1%	0.9%	1.1%	1.0%
AD	4.9%	1.5%	1.1%	1.1%	1.1%	1.0%	1.1%	0.9%
LR	46.2%	6.8%	3.4%	2.2%	1.6%	0.8%	1.2%	1.1%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	17.2%	2.4%	1.6%	1.4%	1.5%	1.1%	1.3%	1.2%
AD	52.2%	4.3%	2.4%	2.1%	1.9%	1.4%	1.6%	1.4%
LR	98.4%	33.5%	14.2%	8.6%	5.7%	1.7%	1.3%	1.0%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	99.8%	19.9%	6.5%	4.8%	4.1%	2.0%	1.9%	1.8%
AD	100.0%	66.8%	25.4%	16.8%	10.4%	4.3%	3.7%	3.3%
LR	100.0%	99.1%	82.7%	63.9%	43.5%	10.4%	4.2%	1.4%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	3.1%	1.3%	1.0%	1.1%	0.9%	0.8%	1.0%	1.0%
AD	4.9%	1.3%	1.0%	1.0%	1.0%	0.9%	1.0%	1.2%
LR	46.2%	10.4%	4.3%	3.4%	2.6%	1.2%	0.8%	0.6%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	17.2%	2.8%	1.7%	1.6%	1.4%	1.0%	1.1%	1.1%
AD	52.2%	3.4%	2.0%	1.6%	1.6%	1.2%	1.2%	1.5%
LR	98.4%	50.2%	21.3%	14.7%	9.1%	2.5%	1.4%	0.6%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	99.8%	29.3%	10.2%	7.2%	4.4%	2.0%	1.6%	1.7%
AD	100.0%	71.9%	22.5%	12.6%	7.6%	3.3%	3.0%	3.2%
LR	100.0%	100.0%	94.9%	84.5%	64.2%	17.6%	6.1%	2.2%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS_ρ	3.1%	3.1%	2.7%	3.1%	2.5%	3.2%	3.1%	2.5%
AD_ρ	4.9%	5.8%	5.8%	5.9%	5.1%	6.2%	5.3%	4.0%
LR_ρ	46.2%	45.7%	45.5%	45.5%	45.2%	44.1%	42.3%	36.5%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS_ρ	17.2%	17.4%	15.6%	16.5%	15.4%	16.6%	14.7%	12.3%
AD_ρ	52.2%	56.8%	56.7%	56.3%	54.6%	56.3%	50.5%	39.6%
LR_ρ	98.4%	98.3%	98.2%	98.2%	98.1%	98.0%	97.3%	95.3%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS_ρ	99.8%	99.8%	99.7%	99.8%	99.8%	99.6%	99.3%	98.2%
AD_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
LR_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

TABLE 3. The True Positive Rate at 99% for non-overlapping and overlapping experiments.

Non-Overlapping ($d = h$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	36.3%	9.4%	5.1%	2.6%	2.4%	1.5%	2.2%	0.5%
AD	55.9%	17.4%	11.1%	8.0%	6.8%	5.0%	5.5%	3.3%
LR	82.6%	28.8%	13.6%	9.2%	5.5%	0.4%	0.0%	-2.3%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	79.6%	29.7%	17.0%	11.6%	8.7%	4.5%	4.3%	2.1%
AD	94.7%	48.9%	31.3%	24.2%	19.2%	11.7%	10.6%	7.5%
LR	99.8%	72.8%	46.2%	34.8%	23.3%	5.4%	1.1%	-3.4%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	99.9%	82.3%	59.8%	47.8%	36.8%	17.5%	12.4%	7.6%
AD	100.0%	96.5%	83.7%	74.0%	62.1%	36.6%	28.2%	20.8%
LR	100.0%	99.9%	96.7%	90.7%	78.7%	35.5%	13.9%	-1.0%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS	36.3%	12.8%	6.8%	4.6%	3.2%	0.9%	0.1%	0.7%
AD	55.9%	19.4%	10.8%	8.1%	6.0%	3.1%	2.4%	2.4%
LR	82.6%	39.0%	19.9%	13.0%	7.0%	-1.5%	-3.1%	-3.1%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS	79.6%	40.8%	24.3%	18.0%	12.6%	4.4%	1.7%	1.4%
AD	94.7%	54.9%	34.2%	26.4%	19.4%	8.9%	6.0%	4.7%
LR	99.8%	85.1%	60.1%	46.6%	31.4%	5.2%	-3.0%	-6.4%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS	99.9%	91.2%	74.5%	63.9%	50.8%	22.2%	11.0%	5.6%
AD	100.0%	97.8%	87.8%	79.0%	66.5%	34.7%	21.6%	14.0%
LR	100.0%	100.0%	99.4%	97.3%	90.4%	46.1%	14.4%	-7.5%
Overlapping ($d = 1$)								
$\lambda = 1.05$	1	5	10	14	21	62	125	250
KS_ρ	36.3%	34.8%	35.9%	34.9%	34.7%	34.5%	33.1%	30.5%
AD_ρ	55.9%	55.3%	55.6%	55.0%	54.8%	54.2%	52.4%	48.7%
LR_ρ	82.6%	82.5%	82.4%	82.3%	82.2%	81.1%	79.7%	75.7%
$\lambda = 1.10$	1	5	10	14	21	62	125	250
KS_ρ	79.6%	79.3%	79.6%	79.1%	79.3%	78.1%	76.7%	72.8%
AD_ρ	94.7%	94.9%	94.9%	94.8%	94.6%	94.4%	93.5%	91.1%
LR_ρ	99.8%	99.8%	99.8%	99.8%	99.8%	99.7%	99.7%	99.4%
$\lambda = 1.25$	1	5	10	14	21	62	125	250
KS_ρ	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.7%
AD_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
LR_ρ	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

TABLE 4. Discriminatory Power for non-overlapping and overlapping experiments.