WORKING PAPER

iRuiz CONSULTING

Backtesting Counterparty Risk: How Good is your Model?

Ignacio Ruiz *

July 2012

Version 1.1

Abstract

Backtesting Counterparty Credit Risk models is anything but simple. Such backtesting is becoming increasingly important in the financial industry since both CCR capital charge and CVA management has become even more central to banks. In spite of this, there are no clear guidelines by regulatory bodies as to how to perform this backtesting. This is in contrast to Market Risk models, where the Basel Committee set a strict set of rules in 1996 which are widely followed by national regulators. In this paper, the author explains a quantitative methodology to backtest EPE models. He expands the three-color Basel Committee scoring scheme from the Market Risk to the Counterparty Credit Risk framework. With this methodology, each model can be assigned a color score for each chosen time horizon. Financial institutions can then use this test to assess the need for model enhancements and to manage model risk.

Since the 2008 financial crisis, the world of banking is changing in a very fundamental way. One of the major changes has been the revising of previously "loose" regulation stands by governments. As a result,

1. National regulators have substantially increased their scrutiny over the models used by banks to calculate risk and capital.
2. The amount of capital that banks need to hold against their balance sheet has increased substantially and, hence, the cost-benefit balance between investing in good accurate models (as opposed to models which are too conservative) has shifted towards better models[1].

*Founding Director, iRuiz Consulting Ltd, London. Ignacio is a contractor and independent consultant in quantitative risk analytics and CVA. Prior to this, he was the head quant for Counterparty Risk, exposure measurement, at Credit Suisse, and Head of Market and Counterparty Risk Methodology for equity at BNP Paribas. Contact: ignacio@iruizconsulting.com

[1]Regulators allow the models used by banks to calculate risk and capital to be "conservative" (i.e. over-estimate risk) but never "aggressive" (i.e. under-estimate risk).

The Basel Committee on Banking Supervision states that banks using their internal model methods (IMM) for capital requirements must backtest their models on an ongoing basis. Here, "backtesting" refers to comparing of the model's output against realized values.

There are two major areas where backtesting applies: in the calculation of the Value at Risk (VaR), that later feeds into the Market Risk capital charge, and in the calculation of EPE[2] profiles, that feed into the Counterparty Credit Risk (CCR) charge. The Basel Committee has stated very clear rules as to how to perform the VaR backtest, as well as what the boundaries discriminating good and bad models are. The Committee is also clear about the consequences of a negative backtest for financial institutions (see [1]).

However, directives by the Basel Committee regarding CCR are at present not as strict with regard to both the backtesting methodology and the further actions banks need to follow should a model fail the backtest. In fact, the Basel Committee has only provided *guidelines* in this respect; details are left to the national regulators to decide on (see [2]). As a consequence, this can create confusion as international financial institutions face a blend of requirements from a number of national regulators. Furthermore, the global financial system may become exposed to potential regulatory "arbitrage".

In this paper, we first outline the backtesting framework set for market risk models by the Basel Committee. Thereafter, we explain the additional difficulties that counterparty risk models give rise to with regard the backtesting, and we then propose a methodology for expanding the Basel's VaR backtesting framework to the context of CCR, in a consistent way. This will be provided with a number of examples which illustrate the strengths and limits of the proposed methodology.

As mentioned, there is a quite limited literature in this topic, especially in the CCR context. This paper compiles information in references [1], [2], [3], and [4] and expands on it.

Market Risk Backtesting

In 1996, the Basel Committee set up very clear rules regarding backtesting of VaR models for IMM institutions[3] [1]. This section highlights a number of key features of that backtesting framework.

The VaR capital charge is based on a 10-day VaR. However, backtest is done in 1-day VaR. This is because, as stated in reference [1], "significant changes in the portfolio composition relative to the initial positions are common at major trading institutions". As

[2]Expected Positive Exposure: the average of portfolio values when floored at zero.

[3]IMM institution: a financial institution with Internal Model Methods approved by regulators for capital calculations.

a result, "the backtesting framework ... involves the use of risk measurements calibrated to a one-day holding period". However, the Basel Committee expresses concerns that "the overall one-day trading outcome is not a suitable point of comparison, because it reflects the effects of intra-day trading, possibly including fee income that is booked in connection with the sale of new products". Given this difficulty in dealing with this intra-day trading and fee income, it leaves it to the national regulator to manage this issue as found appropriate.

The backtesting should be done at least quarterly using the most recent twelve months of data. This yields approximately 250 daily observations. For each of those 250 days, the backtesting procedure will compare the bank profit&loss with the 1-day 99% VaR computed the day before. Each day for which the loss is greater than the VaR will create an "exception". The assessment of the quality of the VaR model will be based on the number of exceptions in the twelve month period under study.

The Basel Committee proposes three zones for the model:

- Green Zone: The backtesting suggests that the model is fit for purpose. The model is in this zone if the number of exceptions is between 0 and 4 (inclusive).
- Yellow Zone: The backtesting suggests potential problems with the model, but final conclusions "are not definitive". The model is in this zone if the number of exceptions is between 5 and 9 (inclusive). The market risk capital multiplier gets adjusted gradually.
- Red Zone: The backtesting suggests that, almost certainly, there is a problem with the model. The model is in this zone if the number of exceptions is 10 or greater. The market risk capital multiplier gets adjusted to the maximum.

An illustrative example of a backtesting exercise is shown in Figure 1.

The Probability Equivalent of Model Zones

In fact, the original definition of those zones is driven by the probability that the model is right or wrong. A green model means that the probability that the model is correct is 95%, a yellow model means that that probability is 4.99% and a red zone means that that probability is only 0.01%[4].

Let's assume that each of the 250 observations are independent from each other, and let's also assume that the model under study is "perfect"; that is, that the model will measure the 99th percentile of the profit & loss distribution accurately. Under that assumption, we can use the binomial distribution to compute the probability (P) of

[4]In fact, the Basel Committee was more fine than this. They considered both the probability that an accurate model is seen as inaccurate and vice versa, and came up with those 95% and 99.99% as the most appropriate limits for the zones.

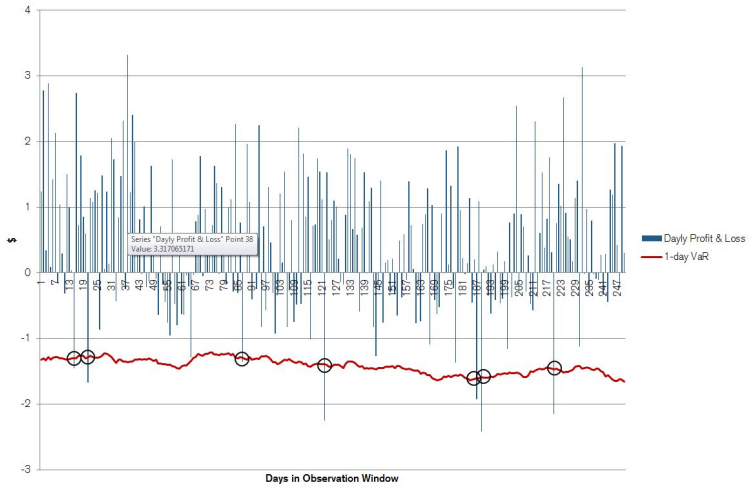

Figure 1: Illustrative example of backtesting exercise for a VaR model. Each circle constitutes an "exception".

number of exceptions (k) in a twelve month period that a "perfect" model will have. That probability is given by the equation

$$P(k) = \binom{N}{k} p^k (1-p)^{N-k} \tag{1}$$

and is illustrated in Figure 2, where $N = 250$ and $p = 0.99$.

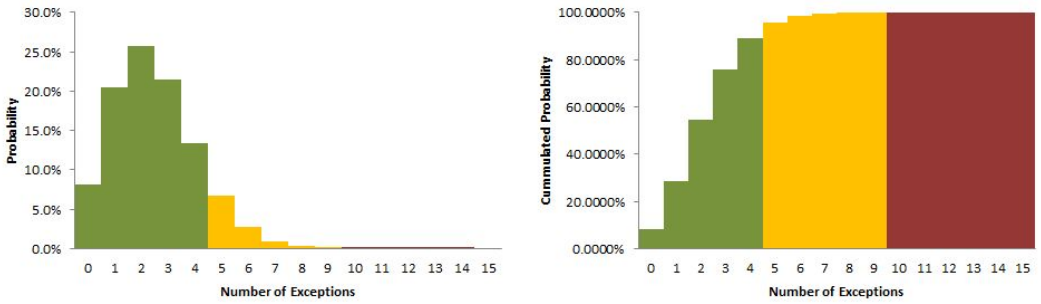

Figure 2: Probability distribution of exceptions in a 12-month period. Each color marks the range of the corresponding zone.

If we now draw a limit (NOTE: in N?) in the distribution of exceptions at the 95th and 99.99th percentiles, then the zone limits are set at 4 and 9 exceptions.

Consequences to Banks

A bank market risk capital charge is given by

$$\text{Market Risk Charge} = (3 + x + y) \cdot \text{MRM} \tag{2}$$

where x is given by the model performance, y is an add-on that national regulators can impose at their discretion and "MRM" (which stands for Market Risk Measure) was 10-day VaR under Basel I, it is now 10-day-VaR plus stress-10-day-VaR under Basel III. Also, some regulators add an additional component called Risks not in VaR (RniV) to MRM. This accounts for the market risks which are not captured model computing the VaR.

Regarding backtesting, x is the number at stake. That number is given by the following table under the Basel framework:

Zone	**Num. Exceptions**	x
Green	0 to 4	0.00
	5	0.40
	6	0.50
Yellow	7	0.65
	8	0.75
	9	0.85
Red	10+	1.00

After the large number of exceptions that all banks had in the 2008 financial crisis, some national regulators decided to remove the cap in x and increased it further as the number of exceptions goes beyond 10.

Proposed Framework for Counterparty Risk Backtesting

Regarding counterparty risk models, The Basel Committee does not have a clear set of rules for backtesting as it has for market risk. Instead, it has provided is a document with a set of guidelines for banks and national regulators [2]. Indeed, the Basel Committee states in that document that "It is not the intention of this paper to prescribe specific methodologies or statistical tests [for counterparty risk], nor to constrain firms' ability to develop their own validation techniques".

So, the goal of this section is to propose a methodology in the context of counterparty risk that can be related to the strict backtesting framework which is in place for market risk. In order to achieve this, we will

1. Define the context and scope in which backtesting can be done for counterparty risk models.

2. Define a single number measure of the quality of a model.
3. Relate that single number measure to the three zones proposed by the Basel Committee, allowing one to classify a model to either the green, yellow or red zone.

Methodology

Context and Scope

It is general practice to refer to a CCR model backtest as the backtesting of the models generating EPE profiles [2]. Those models can be decomposed into a number of sub-models: Risk Factor Evolution (RFE) models for the stress-testing of the underlying factors (e.g., yield curves, FX rates), pricing models for each derivative, collateral models for secured portfolios, and netting and aggregation models.

What we really want is that the value of the whole portfolio under consideration be properly modelled by the EPE model, with all its sub-models. However, running a backtest of the overall EPE model is most difficult, if not impossible. Typically, these calculations are done per counterparty. As a result, we would need to start with a long history of the composition and the value of the counterparty portfolio, which is usually not available to financial institutions. Even if it existed, that history of values will change not on only as a result of changes in the markets, but also from changes in its composition from trading activity and from natural trade expiries. In practice, it appears that backtesting EPE numbers, as such, is not possible.

So we are left with testing each of the sub-models.

The most important driver in an EPE profile tends to be the RFE model. Practitioners know that a 5% inacuracy in a pricer will typically change the EPE profile in a limited way. Sometimes the latter change is not even noticeable due to the Monte Carlo noise of the simulation. But a 5% change in the volatility of a Risk Factor tends to change the EPE profile significantly. So, a lot of care needs to be put into the design of an RFE, and a good backtesting framework is needed there.

Pricing algorithms tends to be well established and developed, as the quantitative finance community has been developing them for quite some time. There are robust methodologies and testing procedures in place, so we are not going to cover that in this paper.

Netting and aggregation models are simple to implement in a Monte Carlo simulator and do not need much testing.

Finally, the testing of collateral models tends to be done on a scenario by scenario basis. This is because collateral management is, in principle, quite mechanical. The complications in the models come either from the approximations needed in the implementation (e.g., how to model the collateral called or posted in a CSA with weekly re-margining

when the Monte Carlo simulation takes monthly time steps) or from modelling risky collateral (e.g., bonds), in which case we are lead again to an RFE backtest (e.g., risk-free rates and credit spreads). Another reason the test is typically done on a scenario by scenario basis is the fact that these models require a history of the portfolio data which is usually not available.

As a result, when banks and regulators talk about backtesting CCR models, the problem often focuses on the backtest of the RFE, because it is, by far, the most challenging part of the models and the one requiring most attention. In this paper, we hence focus on RFE backtesting.

Before we go ahead with the methodology, the reader must note that one of the critical aspects of CCR backtesting is the long time horizons under which we need to test the models. The following fact illustrates the scale of the problem: even though market risk capital charge is based on a 10-day VaR, Basel asks to perform backtesting of market risk models in a 1-day time horizon due to the complexity of dealing with portfolio changes in a 10-day time horizon. In contrast to the 10-day time horizon market risk models deal with, CCR EPE models measure risk in a *many*-year time horizon. As a result, the complexity of backtesting EPE models increases substantially.

The Methodology

Backtesting an RFE means comparing the distribution of the risk factor over time with the distribution seen in the market. In other words, we want to check how close the RFE measure is to the observed "real" measure.

To do this[5], we consider the realized path (a time series) of the risk factor to be tested. That path is given by a collection of values x_t. We will set a time point in that time series where the backtest starts (t_{start}), and a time point where it ends (t_{end}). The backtest time window is then $T = t_{end} - t_{start}$. Then:

1. The first time point of measurement is $t_1 = t_{start}$. At that point, we calculate the model risk factor distribution at a point $t_1 + \Delta$ subject to the realization of x_{t_1}; this can be done analytically if possible, or numerically otherwise. We then take the realized value $x_{t_1+\Delta}$ of the time series at $t_1 + \Delta$ and observe where that value falls in the risk factor cumulative distribution calculated previously. This yields a value F_1[6]. The reader can see in Figure 3 an illustration of this process.
2. We then move forward to $t_2 = t_1 + \delta$. We calculate the risk factor distribution at $t_2 + \Delta$ subject to realization of x_{t_2}, and proceed as before: we observe where in the model distribution function $x_{t_2+\Delta}$ falls and obtain F_2 from it.
3. ... Repeat the above until $t_i + \Delta$ reaches t_{end}.

[5] This part of the methodology follows Kenyon 2012 [4]
[6] For the sake of clarity, the reader should note that $F_i \in (0, 1)\ \forall i$.

So the outcome of this exercise is a collection $\{F_i\}_{i=1}^N$ where N is the number of time steps taken.

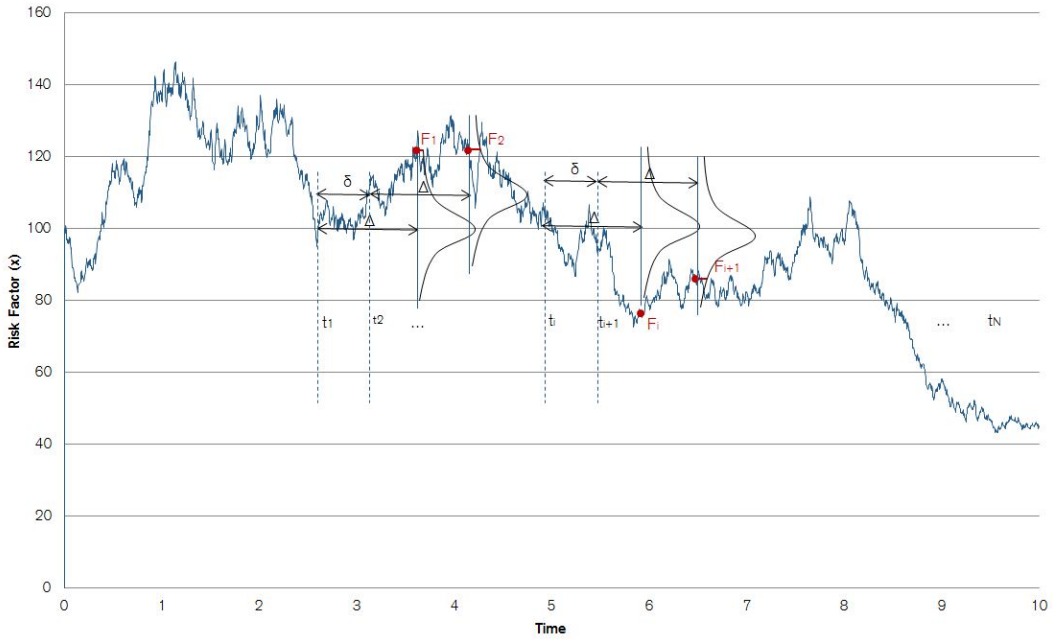

Figure 3: Illustration of backtesting methodology.

The key point in this methodology is the following: in the case of a "perfect" model (i.e., if the empirical distribution from the time series is the same as the distribution that the models predicts), the $\{F_i\}_{i=1}^N$ are uniformly distributed.

At this stage, we can define a metric of the difference, a "distance" D, between the empirical and the model distributions. If that distance is zero, then the model is "perfect".

There are a number of typical metrics for D. If we denote by F the cumulative distribution function given by the model (for a time horizon Δ) and by F_e the empirical cumulative distribution function obtained from the data, then:

Anderson - Darling metric:

$$D_{AD} = \int_{\infty}^{-\infty} (F_e(x) - F(x))^2 \, w\,(F(x))\, dF(x),\ w(F) = \frac{1}{F(1-F)}$$

Cramer - von Mises metric:

$$D_{CM} = \int_{\infty}^{-\infty} (F_e(x) - F(x))^2 \, w\,(F(x))\, dF(x),\ w(F) = 1$$

Kolmogorov - Smirnov metric:

$$D_{KS} = \sup_x |F_e(x) - F(x)|$$

Each metric will deliver a different measurement of D. Which of them is the most appropriate? That depends on how the model being tests is actually used, and this decision has some degree of subjectivity by the researcher and practitioner. For example, in risk management we are most interested in the quality of the models in the tails of the distribution, so we may want to use the Anderson - Darling metric. In capital calculations we are interested in the whole of the distribution function, so we may want to use Cramer - von Mises. If we are happy with small general deviations, but never large deviations, then we may want to use Kolmogorov - Smirnov.

Having chosen a metric, a "D" can now be calculated because we know $F(x)$ from the model and we can obtain $F_e(x)$ from the collection $\{F_i\}$. Hence, we have all we need to obtain a single number measure of the performance of the model: we pick one of those metrics and compute a value $\widetilde{D}$ measuring how good the model is for a given value of Δ.

So far so good. However, the following questions arises now

1. how large does $\widetilde{D}$ need to be to indicate that a model is bad? Or, equivalently, how close to zero must it be to indicate that our model is good?
2. N is a finite number, so even if the model were perfect, $\widetilde{D}$ would not be exactly zero[7]. This induces some stochastic noise. How can we assess the validity of $\widetilde{D}$?

We can to proceed as follows. Let's construct a time series using the model being tested, and then apply our above procedure to it, yielding a value D. The constructed time series will follow the model perfectly by construction, but D will not be exactly zero. This deviation will *only* be due to noise. If we repeat this exercise a large number of times (M), we will obtain a collection $\{D_k\}_{k=1}^{M}$, *all of them compatible with a "perfect" model*. That collection of D's will follow a certain probability distribution $\psi(D)$ that we can approximate numerically from $\{D_k\}_{k=1}^{M}$ by making M sufficiently large.

Now, having obtained $\psi(D)$, we can asses the validity of $\widetilde{D}$. If $\widetilde{D}$ falls in a range with high (low) probability with respect to $\psi(D)$, then the model is likely (unlikely) to be accurate.

The Three Zones

We are now in a position to extend the clear Basel framework established for Market Risk to the setting of counterparty risk backtesting. If we define D_y and D_r respectively as the 95th and 99.99th percentiles of $\psi(D)$, then we can define three zones for the model performance:

- Green zone if $\widetilde{D} \in [0, D_y)$

[7] That is, for a finite number N, $D \geq 0$ even if the model is "perfect"; zero is attained in the limit, i.e. $\lim_{N\to\infty} D = 0$.

- Yellow zone if $\widetilde{D} \in [D_y, D_r)$
- Red zone if $\widetilde{D} \in [D_r, \infty)$

This is illustrated in Figure 4.

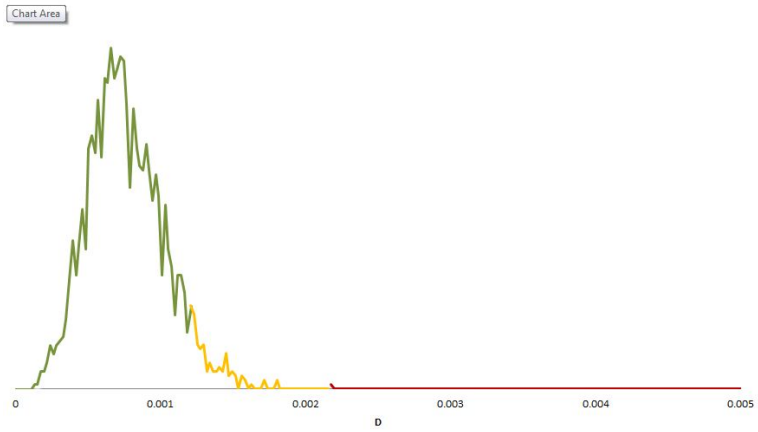

Figure 4: Illustrative example of the distribution of D's compatible with the model.

With this three-zone approach, a financial institution can score the model and proceed accordingly. If the score is "green", the model is most likely correct and no action is required. If the score is "yellow", the model could be wrong, so further investigations should take place. If the score is "red", the model is most likely wrong and hence needs to be revised in depth. In the latter case, the institution may perhaps want to temporarily replace the model with a more conservative one, or use it with extra add-ons which compensate for the lack in accuracy.

Key Strengths of this methodology

This methodology has the following key strengths:

Autocorrelation If $\delta < \Delta$, there is autocorrelation in the $\{F_i\}$ by construction. However, that autocorrelation also exists in the construction of $\psi(D)$. Hence the proposed methodology neutralizes this effect "automatically". As such, as long as the values of T, Δ, and δ are the same in the calculations of $\widetilde{D}$ and of $\psi(D)$, we do not need to do any further adjustments to compensate for the induced autocorrelation. This autocorrelation effect will be discussed in more detail later in the paper.

Numerical Approximations The calculation of D will be done numerically using a number of approximations. This will introduce some noise in this calculations. However,

similar noise will be introduced in the calculation of $\psi(D)$ and, hence, this noise is, also, "automatically" considered in the assessment of $\widetilde{D}$.

Maximum Utilization of Independent Information In practice, one of the big problems of counterparty risk backtesting is that historical data tends to be scarce and the time horizon Δ in which the models need to be tested is quite long. As such, there might be very few independent points in our test. For example, if we have 10 years of data with no auto-dependency and we want to measure a model's performance with a Δ of 2 years, we only have 5 independent points. So, the statistical relevance of the backtest can be quite limited by construction. However, this limitation is again "automatically" captured in this methodology. We will elaborate more on this later in the paper.

Asset-Class Agnostic The methodology can be applied to any asset class. It is asset class independent.

Expandable to dependency-structure backtest The methodology can be applied to individual risk factors or to sets of risk factors in parallel by extending the methodology to many-dimensions. In that case, the procedure will also test the dependency structure of the model against the "real" one existing in the data. We will elaborate on this later in the paper.

Global Test of Real Risk Measure Often, pure risk management uses a given percentile in of the distribution function to measure counterparty risk. For this reason, sometimes EPE models are tested by counting the number of exceptions outside of given percentile envelopes. However, that methodology is sub-optimal for regulatory capital models where the key measure of risk is the EPE[8]. This is because EPE is, loosely speaking, an average measure and, as such, to check its validity we need to test the quality of the whole distribution functions of exposures, not only exceptions above or below a given percentile.

Examples

We now present a few examples where the mechanics of the methodology are illustrated.

[8]The same could be applied to CVA if it was to be backtested.

Example 1: GBM time series vs GBM model with misspecified volatility

We want to gain some understanding as to how the algorithm performs when the empirical data has a different volatility to what the model assumes. In order to achieve this, we simulate a time series using a GBM process and treat it as the empirical data; hence, the inputs to the algorithm is well known and the output can be understood in detail.

Figure 5[9] shows the output of the backtest algorithm when the empirical volatility is lower than the model volatility. In those graphs, the data obtained from the model has been fitted to a standard normal distribution function, and the empirical data has also been normalized using the same normalization factors as for the model. As a result, the graphs illustrate the difference between the empirical and the model distribution functions. If the model were to fit the empirical data "perfectly", then both graphs (blue and red) should lie on top of each other.

Next, Figure 6 illustrates the classification procedure. Here, the "empirical" time series was again simulated numerically using a GBM process with a volatility of 35%, while the model being backtested was a GBM process with a volatility of 40%. The backtest procedure yielded $\widetilde{D} = 0.003$, while the red zone for this backtesting exercise starts at 0.0018; hence, the algorithm classifies this model into the Red zone for the time series, indicating that it is inaccurate[10].

Example 2: GBM + jumps time series vs GBM model with the same volatility as the time series

In this test, the author generated an empirical time series with a GBM plus 1-sided Poisson jump process, and tested the time series against a GBM model with the same volatility as the time series. On this way, we can observe how the algorithm responds when the skew of the data and the model are different.

Figure 7 shows how the model responded. In this specific example, the empirical time series has larger negative skew than the model. This is most visible in the Uniform Distribution graphs (bottom panels).

Also, a full Monte Carlo test was run to assign a color zone to the time series. The empirical series was numerically simulated from a sum of a GBM (with $\sigma = 30\%$) and a Poisson jump process. The obtained time series had a volatility of 34%, a skew of -0.35 and an excess kurtosis of 0.54. It was checked against a simple GBM model with a volatility of 34%. The results can be seen in Figure 8.

[9]PDF: Probability Distribution Function, CumPDF: Cumulative Probability Distribution Function.

[10]The author computed $\widetilde{D}$ for a number of GBM "empirical" time series with a volatility of 35%, the outcome was always around 0.003.

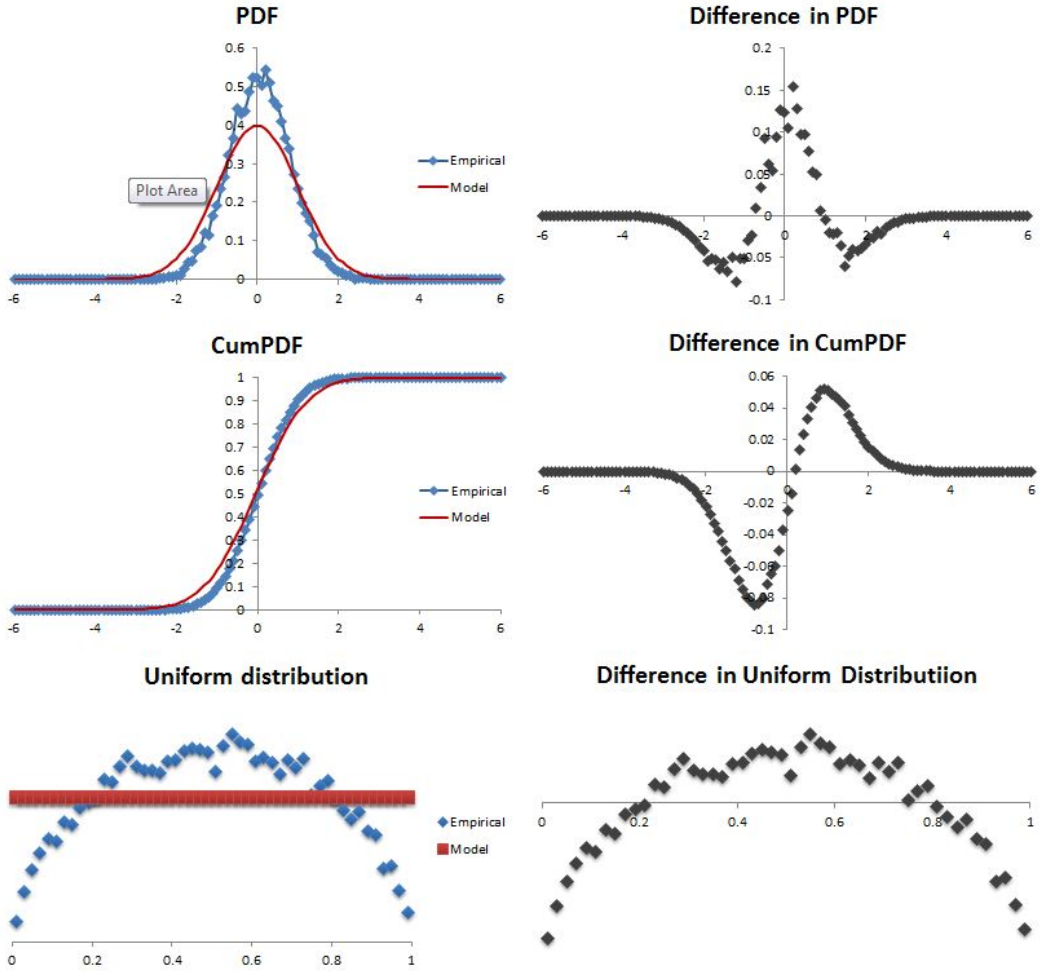

Figure 5: Illustrative example of how the backtest algorithm behaves when the model volatility is higher than the empirical volatility.

Example 3: GBM + double jump time series vs GBM model with the same volatility as the time series

In this case, we want to see how the algorithm behaves when the empirical time series has a different kurtosis to the model. For this, the author generated an empirical time series using a GBM plus a 2-sided and symmetric jump process, and tested the time series against a GBM model with the same volatility as the time series. The skew in both the model and the time series was zero.

The results can be seen in Figures 9 and 10. Again, the effect is best depicted in the Uniform Distribution graphs (Fig. 9): the symmetric wave we have in the blue dots indicates high kurtosis in the data.

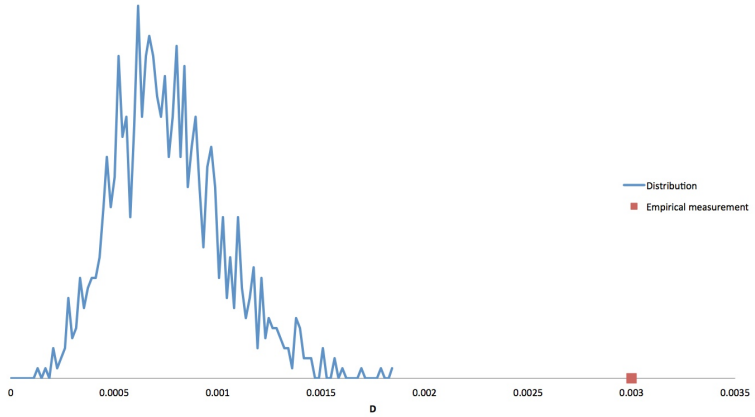

Figure 6: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-driven time series tested against a GBM model with a higher volatility.

In particular, the time series, had a volatility of 58%. It was backtested against a GBM model with a $\sigma = 58\%$. The model fell in the Red zone, as shown in figure 10.

Example 4: S&P500 time series vs GBM model historically calibrated

Next, we applied the backtest methodology to the S&P500 equity index against a GBM model.

Before proceeding with the results, an important remark needs to be made:

Model Calibration A counterparty risk RFE model is more than a set of stochastic equations. It actually consists of (i) a set stochastic equations (e.g. GBM diffusion plus Poisson jumps), (ii) a calibration methodology (e.g. 3-year historic, or implied volatilities) and (iii) a calibration frequency. The backtesting algorithm must consider all those inputs. For example, if the RFE is a GBM model which is calibrated quarterly, with a volatility equal to the annualized standard deviation of the daily log-returns for the past 3 years, then the volatility of the GBM process must be recalibrated quarterly when we calculate $\widetilde{D}$ and when we generate each of the M paths leading to $\psi(D)$. It is very important to use the appropriate calibrating methodology and frequency in the backtesting exercise. This point was left out in the first outline of the backtesting methodology for simplification of the explanation. From now on, we will refer to the calibration period as T_c (if historically calibrated), and to the calibrating frequency as δ_c.

Having said that, we are now going to apply the backtesting algorithm to the S&P500 time series, using the daily time series from 2000 to 2011 ($T_c = 11$ years). The model we

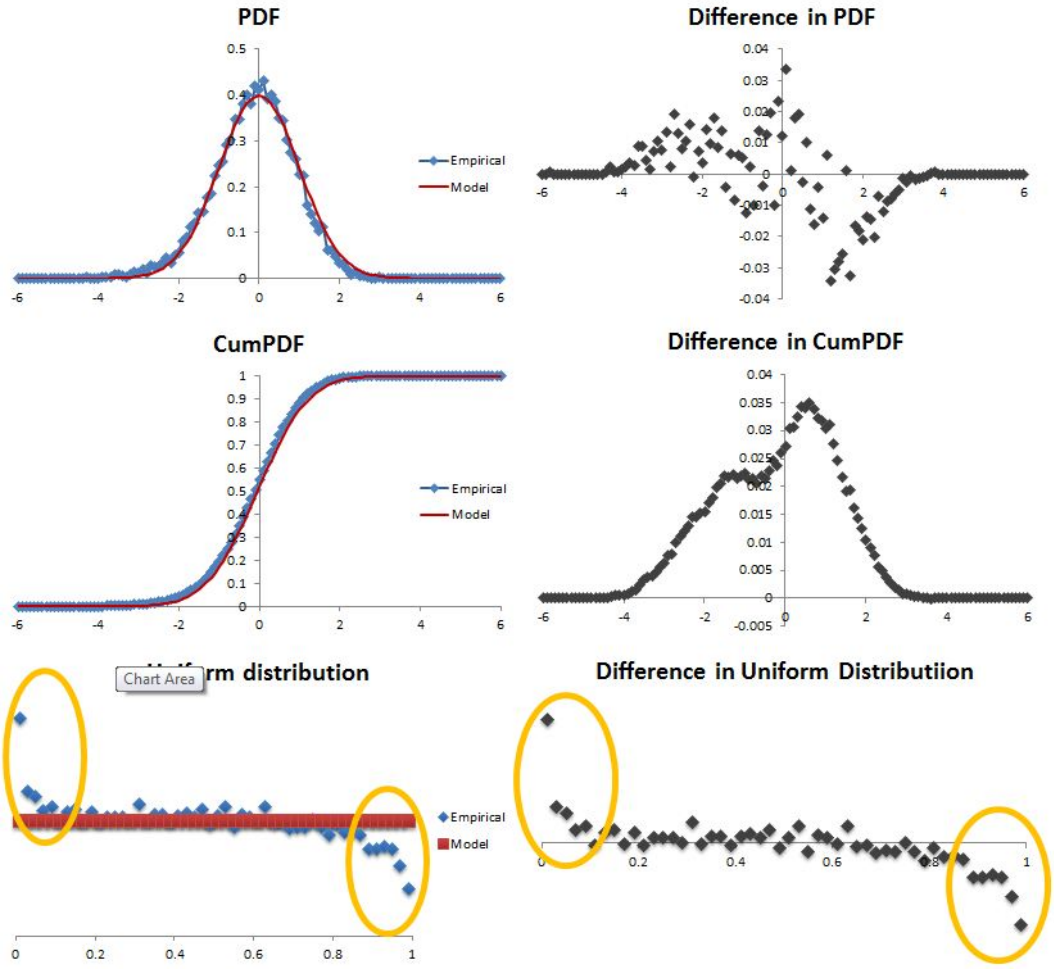

Figure 7: Illustrative example of how the backtest algorithm behaves when the model skew is different to the empirical skew.

backtest is a simple GBM, with historical daily ($\delta_c = 1$ day) calibration and volatility equal to the annualized standard deviation of log-returns. We want to test how the model performs for time horizons of 10 days and one year, and for calibrating windows of three months and three years.

The results can be seen in Figure 11. The following table summarizes the scores in terms of color zones:

	$T_c = 3$ months	$T_c = 3$ years
$\Delta = 10$ days	GREEN	YELLOW
$\Delta = 1$ year	GREEN	GREEN

The backtesting methodology output is as somewhat expected. When the model time horizon Δ is short (10 days), a short calibration window ($T_c = 3$ months) makes the

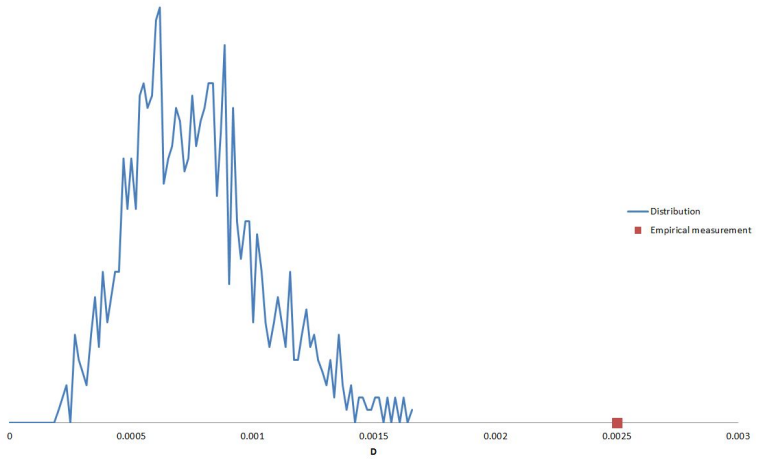

Figure 8: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-plus-jump time series tested against a GBM model with the same volatility as the time series.

model score a Green, but when the calibrating window is long ($T_c = 3$ years), then the algorithm captures that there is a problem with the methodology and the model scores a Yellow[11]. However, when the model time horizon Δ is long (1 year), then the backtest indicates that either calibration window of 3 months or 3 years is good; both have a Green score.

Some readers may find that last result quite remarkable, i.e. that for a time horizon of 1 year, the algorithm scores Green for both the 3 month and 3 year calibration windows. This will be further explored in the next section.

Practical Considerations

The Role of the Backtesting Time Window T

We previously mentioned that the statistical relevance of the backtest depends mainly on two parameters: Δ and T. For example, let's say that $\Delta = 1$ month and that $T = 1$ year. Assuming independence in the time series, we get 12 sample points which are independent[12]. If we now keep Δ at 1 month, but increase T to 10 years, then we will have 120 independent points in the backtesting procedure. A given model will be more likely to receive a Green score in the former case, because there are less points to match (12 vs 120) and, as such, there is more freedom in the model choice for obtaining a good fit to these data points; i.e., there will be a wider range of models that can accurately fit 12 points in a cumulative distribution function than 120 points.

[11] The problem is that 3-year historical volatility is not a good predictor of short-term future volatility.
[12] Let's forget about calibration details for now.

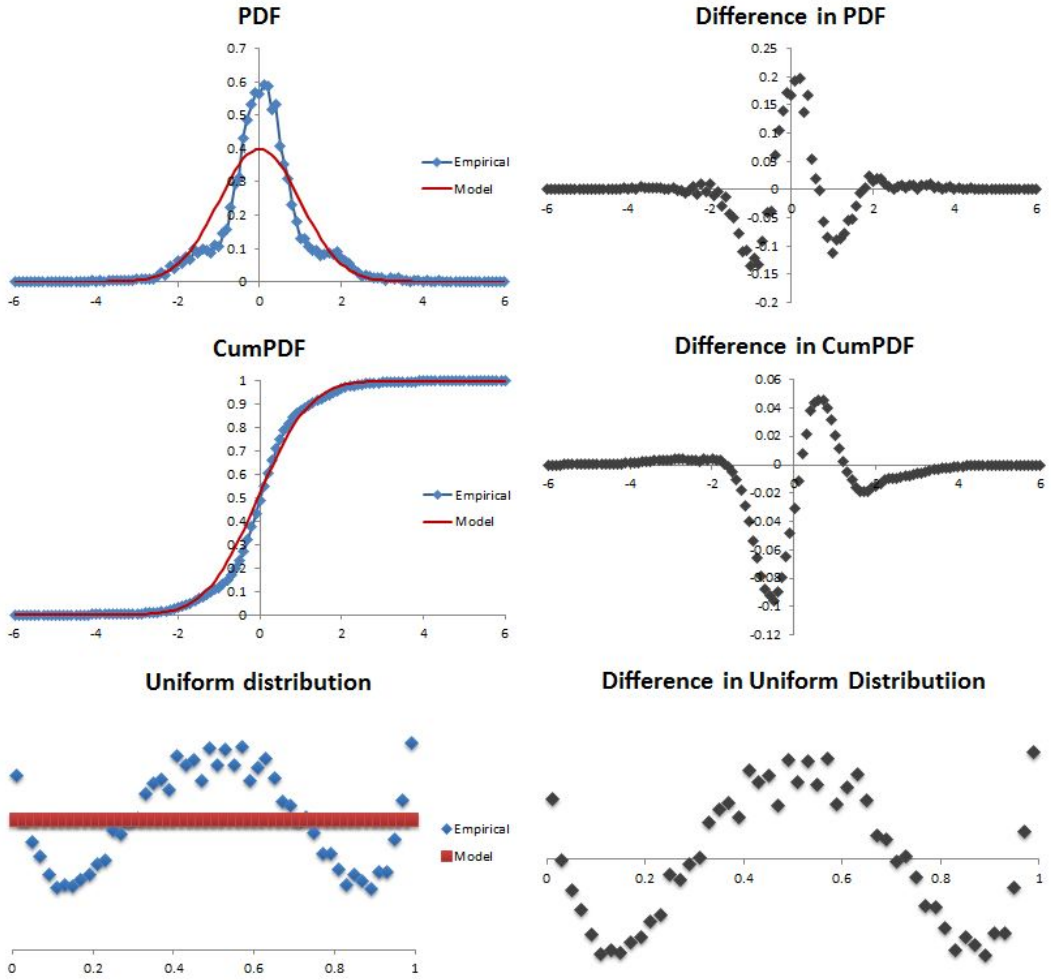

Figure 9: Illustrative example of how the backtest algorithm behaves when the model kurtosis is lower than that in the empirical data.

One of the key strengths of this backtesting methodology is that this is automatically captured. In order to illustrate this, let's re-run our S&P500 example with the time series starting in 1950; that is, we are expanding the backtesting window T from around 10 years to around 60 years. In particular, let's run the tests when $\Delta = 1$ year both for $T = 10$ years and $T = 60$ years. The results are shown in Figure 12.

This example illustrates how the backtesting algorithm accounts for the quality of the available information by contracting the width of $\psi(D)$ as the relevant information increases (e.g. as we go from 12 relevant points to 120 points). As the information regarding the "real measure" we are aiming for gets more granular, the width of $\psi(D)$ gets smaller, hence making it more difficult for a model to obtain a Green score. On the other hand, it is easier to find a model matching the limited information available about

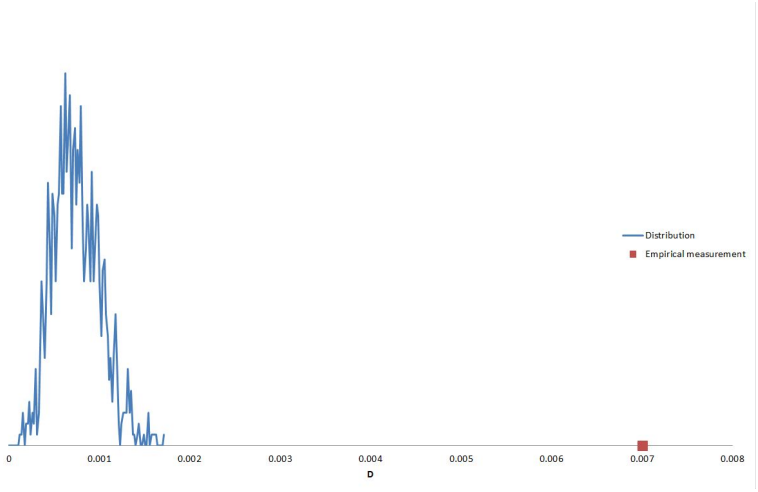

Figure 10: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-plus-double-jump time series tested against a GBM model with the same volatility as the time series.

the "real" measure when we have very little granularity.

That is why the model scores a Green in our S&P500 example when backtested against 10 years of data (i.e. only 10 independent points in the backtest), but it scores a Red when backtested against 60 years of data (i.e. 60 independent points in the backtest).

On Auto-Correlation

As previously mentioned, this methodology automatically incorporates a way to deal with the autocorrelation induced in the algorithm when $\delta < \Delta$. That automatism comes from the fact that the same autocorrelation is also induced in the calculation of $\psi(D)$. Given that the procedure scores a model based on a benchmark probability distribution which takes the autocorrelation effects into account, the outcome is neutral to that effect. For this reason, the empirical distribution goes to the uniform distribution also when $\delta < \Delta$ and auto-correlation exists.

However, this leads to the question: what is the optimal δ?

The author admits that he has no conclusive answer yet to this question. After discussing this subject with a number of colleagues, no clear conclusion was attained. The general consensus seems to be to make δ as small as possible, such that the calculation of D incorporates as many points as possible and, hence, would be more statistically relevant. However, it is not clear to the author why the statistical relevance of large set of numbers which are highly autocorrelated is better than a smaller set of numbers which are less autocorrelated.

For example, say that we are running a backtesting exercise with $T = 1$ year and $\Delta =$

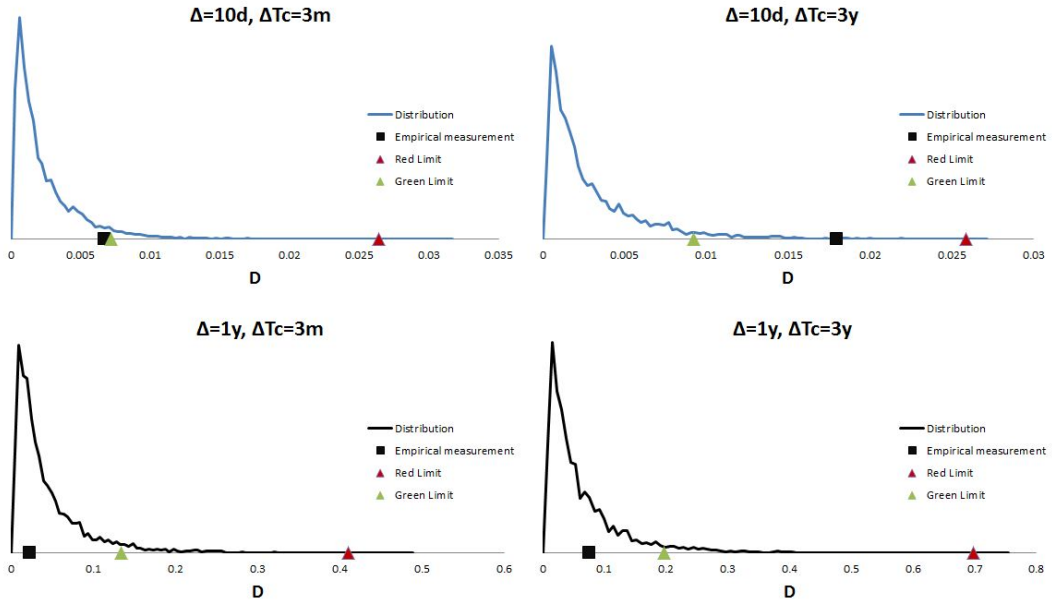

Figure 11: Backtest measurements for S&P500 from 2000 to 2011.

1 month. This means we only have 12 independent points in the measurement. If we choose δ of 1 month, then we will have 12 points in the measurement, and we know that they will be independent[13]. However, if we choose δ of 1 day we will have around 250 points in the measurement. Thus, the calculation of D may appear to have larger statistical relevance, but the quality of the information of those 250 points is poor.

In order to assess the impact of autocorrelation in the algorithm, we tried the following experiment. We applied the procedure to the S&P500 backtest described above, but with different values of δ, leaving everything else constant. We used a time window T from 2000 to 2011, a GBM model with daily 3-month historical calibration, and $\Delta = 1$ month. Values of δ ranging from 1 day (maximum autocorrelation) to 1 month (no autocorrelation) were used. The results are shown in Figure 13.

The obtained results suggest that the value of δ may be irrelevant. In all four cases, the models score in the Green zone, and in all cases $\widetilde{D}$ is in the vicinity of 65% of the Green Limit (the 95th percentile in $\psi(D)$). Needless to say, this constitutes by no means definitive evidence that the value of δ is irrelevant, but it does suggest that it may be irrelevant.

Perhaps the case in which it makes sense to have a small δ is when the number of independent measurement points is very small and a non-integer. For example, let's say that, for some reason, we only have 5 years of historical data but we are interested

[13] Assuming there is no autocorrelation in the time series nor in the construction of the model being tested.

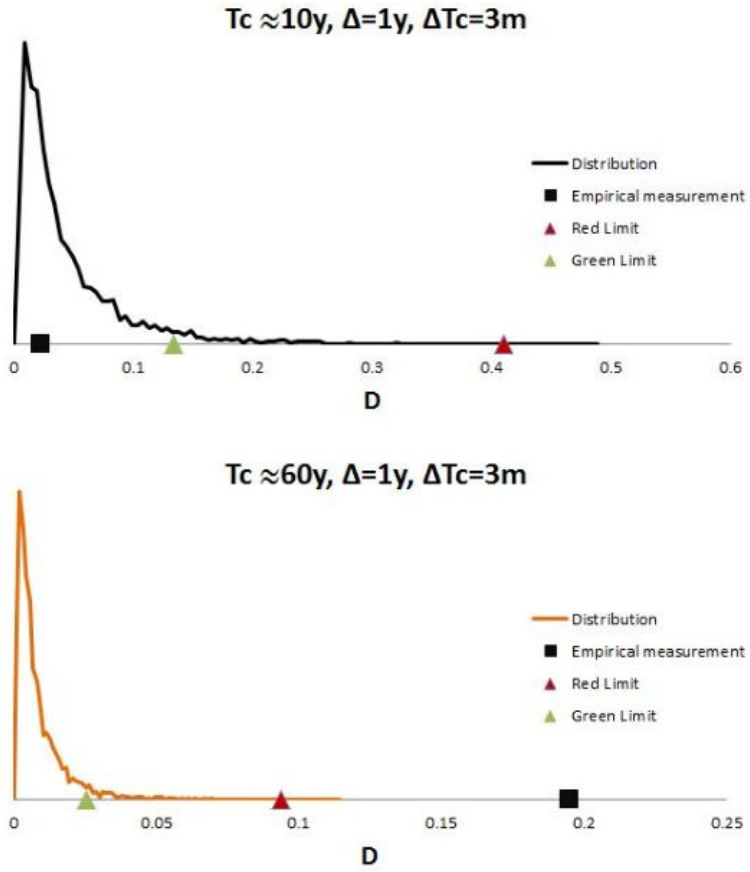

Figure 12: Backtest measurements for S&P500 from 2000 to 2011 (top) and from 1950 to 2011 (bottom).

in the model performance in a time horizon Δ of 2 years. In this case, we have "2.5" independent points. If we pick $\delta = 2$ years, then the measurement of D will be missing that extra bit of information coming from the ".5", and so by picking a $\delta = 1$ years, the calculation of D will maximize the use of the available information.

What Parameters to Use?

The proposed methodology has the following inputs. On the model side, we typically have a set of stochastic differential equations for the RFE, a calibration methodology, and a calibration frequency. On the backtesting side, we have a time window T, a time horizon Δ, and a step size δ. Given a model to be tested, how do we choose T,Δ, and δ?

T represents the time window from which we want to extract the "real" measure that

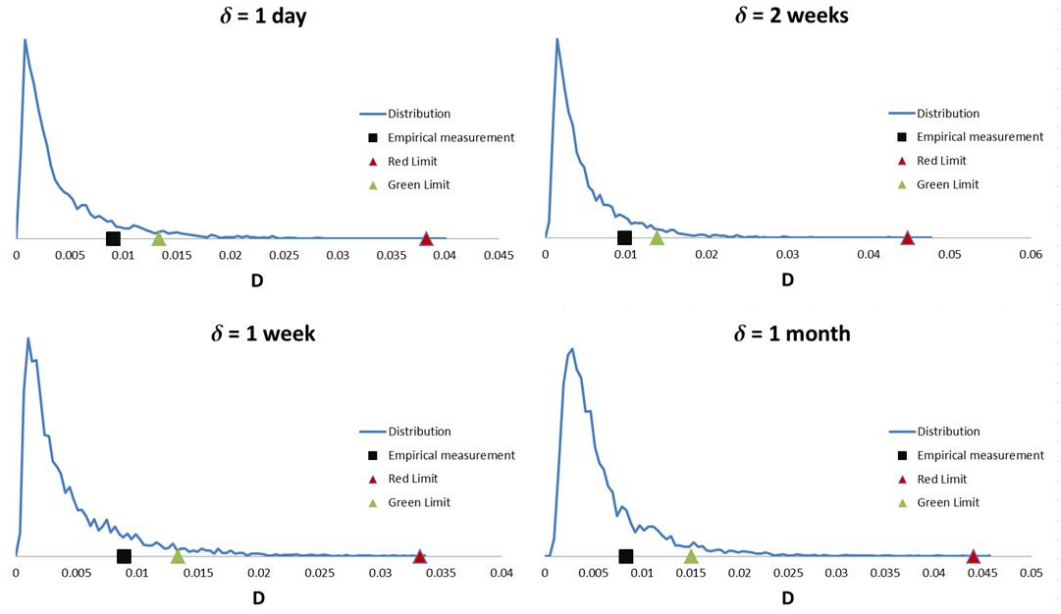

Figure 13: Illustration of backtesting algorithm behavior by changing δ

we want to test our model against. For that reason, as mentioned earlier, the larger this T, the more difficult it is to find a model which passes the backtest since there is less room to manoeuvre in the model. Hence, on the one hand, T should be long enough (relative to Δ) so that the backtest algorithm has statistical relevance. On the other hand, large values of T will make the task of building a model difficult to a degree that is impractical in a commercial banking environment. Also, another practical constraint in this regard is the availability of good quality data; often, data does not date as far back as we would like it to, and when it does, it usually lacks quality. So, in principle, there doesn't seem to be a set rule as to what the best value for T is[14].

Δ represents the time horizon in which we want to test the model. In this case, there is a clear optimal value. It will be determined by the maturity of the portfolios affected by the RFE under testing. For example, if we want to test a foreign exchange model for a portfolio with most of the trades with a maturity of less than 3 years, then it is not necessary to test the RFE beyond 3 years. This can cause some practical problems, as there are some asset classes where the typical length of a portfolio can be much larger than the available past data (e.g., inflation, where trade tenors tend to be at around 25 years and can even go up to 50 years). In such cases, it is impossible to do a good backtest and all we can can do is to test the model for shorter time horizons and make

[14]The practitioner and researcher should note the fact that when the model is calibrated historically, the first section of the time series used for calibration is used up by the calibration. For example, if we have 10 years of historical data, and the model is calibrated using 4 years of historical data, then we only have a T of 6 years for the actual backtest.

sure that the long-term RFE behavior is reasonable.

δ represents the granularity in the backtesting calculation. As discussed, it is not clear to the author whether a more granular calculation is better. If T is sufficiently large to guarantee statistical relevance with independence between time points in the time series, then the author suggests to use the smallest δ which guarantees independence. If there is no autocorrelation in the time series, this means that $\delta = \Delta$; otherwise, $\delta = \Delta + \tau$, where τ is the minimum time needed for the autocorrelation effects to disappear. However, as explained, the case in which the algorithm will benefit from a smaller δ is when T/Δ is small and non-integer.

To sum up, given a model to be tested, T is first driven by the availability of good quality data. It should be large enough so we can have some statistical relevance in the algorithm, but not too large so that the search for a good model does not become impossible. There is no set rule for defining "large enough" or "too large"; in practice, this can only be left to the experience and market knowledge of the researcher. Δ will be determined from the typical tenor of the trades in the portfolio under testing, subject to availability of data. Regarding δ, it is not clear to the author what is ideal, but preliminary and non-rigorous evidence suggests that there is no reason to make the value of δ smaller than Δ, hence making the algorithm slower, unless T/Δ is small and non-integer.

Some Important Remarks

Structural Changes in the Market

The markets undergo deep fundamental changes from time to time. The most recent one started in the credit asset class in 2007 and, arguably, spread to other asset classes. These phenomena occurs for a number of reasons which tend to gravitate around markets drying out, important imbalances in the size and/or price of a market with respect to its economic fundamentals, human "manias", etc. Interestingly, after these events occur, it is "easy" to make sense of them, but very few people are able to foresee them. In some cases, it is impossible by definition[15].

Stochastic models cannot capture these events, but we ideally want models which react quickly to changing market conditions and which are valid for a wide range of market conditions[16]. Hence, assessing the quality of an EPE model requires assessing how the model behaves under structural changes in the markets.

[15] These are the now called Black Swan events[5].

[16] We can certainly build a model which gives a certain probability to certain stress events in the future. However all we will achieve with this is that some scenarios in our Monte Carlo simulation will follow those stress events, but the impact in the "average" scenario will be limited in general. For this reason, risk management uses Potential Future Exposure or Expected Shortfall risk measures to monitor the so-called tail risk.

The proposed backtesting methodology can help in that respect by implementing a rolling window test, in which T is kept constant and is rolled over time.

For example, let's say we have a credit model which we are testing in the 2000 to 2012 window. The model receives a Green score. The proposed methodology, with a time window of 12 years, gives a Green to the overall performance of the model in that period, but gives no indication as to how the model performed during the crisis. In order to assess this, we can implement a window T of 2 years starting in 2000, roll it forward up to 2010, obtaining a rolling $\widetilde{D}_t$. This way, the researcher can assess the quality of the response of the model to changing market conditions.

In fact, the author has seen precisely this kind of behaviour in the backtesting work he has done for a client. There, a model was backtested for the 2001 to 2011 period. Overall, the model scored a Green. A rolling window test, with $T = 2$ years and $\Delta = 1$ month, showed how the model started in the Green zone, went to the Yellow or Red zones when the backtesting period overlapped with the the peak of the credit crunch, and then went back to Green when the (historical) calibration window started to overlap with the credit crunch. This test showed that the model was not good at coping with structural changes in the market (as expected as it was historically calibrated) but it was good both for quiet and stress regimes once the calibration accounted for it. As a result, the institution implemented a special stress process to cover the risk of sudden changes in the market. However, it was proven that there was not need for a different model for stressed market conditions.

Historic vs Implied Calibrations

Risk models tend to be calibrated historically, but pricing models (e.g., CVA) tend to have risk-neutral market-implied calibrations. The proposed backtesting methodology is agnostic to the type of calibration in the model. The only thing that the researcher must bear in mind is that the *same* calibration methodology must be applied to compute both $\widetilde{D}$ and $\psi(D)$. In the case of market-implied calibration, this will require a joint model for the RFE and the calibration variables.

Extrapolating to the multidimensional case

So far, we have explained how to apply this procedure in the context of one risk factor. However, this methodology can be extended to a multi-dimensional case in which the dependency between risk factors is also tested. This can be particularly relevant when the risk factors apply to various curves like the interest rates, credit spreads, commodities, etc. In that case, we ideally want to test not only whether the model of each particular tenor point in the curve is correct, but also whether the dependency between different points in the curve is correctly captured.

Say that we have a model for a curve consisting of a number of tenor points which have a certain dependency structure. We have historical time series for each of those points. Testing the RFE of each point in the curve, in isolation to the rest, will provide no information whatsoever regarding the quality of the model with regard to the dependency structure between points.

However, this methodology can test the dependency structure by expanding the back-testing exercise to a multi-dimensional case. The color scheme can be applied in exactly the same way as in the one-dimensional case. This way, the backtest can include, for example, the validity of a copula structure.

For example, we can expand the Anderson - Darling metric to two dimensions as follows

$$D'_{AD} = \int_{\infty}^{-\infty} \int_{\infty}^{-\infty} \left(F_e(x,y) - F(x,y)\right)^2 w\left(F(x,y)\right) dF(x,y),\ w(F) = \frac{1}{F(1-F)}$$

In practice, this calculation can become quite convoluted, so it must be made with a lot of care. The author suggests that models be first tested in isolation. Then, multi-dimensional tests can be implemented. This way, should the final joint-RFE test be non-satisfactory, potential problems can be better identified and managed.

Conclusions

The author has proposed a backtesting algorithm for Counterparty Credit Risk models that provides a Green/Yellow/Red score to a model, resembling the widely used approach in in Market Risk framework. On this way, a model can receive a color score for each time horizon Δ and management and regulators can easily monitor model quality.

There are a number of important factors to consider. First, a "distance metric" D between the empirical and the model distribution functions needs to be chosen. Then, a time window T needs to be picked; in most cases, this time window will be determined by the availability of data. The time horizon Δ needs to be chosen by taking the maturities of the portfolio being tested into account. Finally, δ should arguably be chosen equal to Δ, except when T/Δ is small and non-integer.

We have seen how this methodology maximizes the utilization of available information and it can be applied to any asset-class and calibrating methodology.

The author has implemented this methodology in tier-one banking environments that is leading to model implementation with regulatory approval.

References

[1] *Supervisory framework for the use of "backtesting" in conjunction with the internal models approach to market risk captial requirements*, tech. report, Basel Committee of Banking Supervision, 1996.

[2] *Sound practices for backtesting counterparty credit risk models*, tech. report, Basel Committee of Banking Supervision, 2010.

[3] E. Canabarro, *Counterparty Credit Risk*, Risk Books, first ed., 2009.

[4] C. Kenyon, *Model risk in risk factor evolution*, in Measuring and Controlling Model Risk, London, 2011.

[5] N. N. Taleb, *The Back Swah: the Impact of the Highly Improbable*, Penguin Books, first ed., 2008.