iRuiz CONSULTING

Backtesting Counterparty Risk: How Good is your Model?

Ignacio Ruiz *

July 2012

Version 2.1.1

A version of this paper is being published in the Journal of Credit Risk

Abstract

Backtesting Counterparty Credit Risk models is anything but simple. Such backtesting is becoming increasingly important in the financial industry since both the CCR capital charge and CVA management have become even more central to banks. In spite of this, there are no clear guidelines by regulators 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. In this paper, the author explains a quantitative methodology to backtest counterparty risk 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 framework to assess the need for model enhancements and to manage model risk. The author has implemented this framework in Tier-1 a financial institution; the model report generated was sent to the regulators for IMM model approval. The model was approved a few months later.

Since the 2008 financial crisis, the world of banking is changing in a very fundamental way. One of the main driver of this transformation is the change in stand by governments, from a "loose" regulatory enviroment in the pre-2007 era to a much more hands-on approach now. In particular,

1. National regulators have substantially increased their scrutiny over the models used by banks to calculate risk and capital.

*Founding Director, iRuiz Consulting, London. Ignacio is a contractor and independent consultant in quantitative risk analytics and CVA. Prior to this, he was the head strategist 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

2. The amount of capital that banks need to hold against their balance sheet has increased substantially and, hence, the cost-benefit balance of investing in good accurate models has shifted substantially towards better models.

The Basel Committee on Banking Supervision states that banks using their internal model methods (IMM) for capital requirements must backtest their models on an on-going 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[1] profiles, that feed into the Counterparty Credit Risk (CCR) and CVA-VaR charge. The Basel Committee has stated very clear rules as to how to perform the VaR backtest, as well as to what are the boundaries discriminating good and bad models. The Committee is also clear about the consequences of a negative backtest for financial institutions [1].

However, at present, directives by the Basel Committee regarding EPE backtesting are not so strict. In fact, the Basel Committee has only provided *guidelines* in this respect; details are left to the national regulators to decide on [2]. This has created some degree of confusion between and within financial institutions, as they face a blend of (sometimes not clear) requirements from a number of national regulators. As a result, in the author's view, the global financial system is now exposed to regulatory "arbitrage" in this area.

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 bring to with regards to the backtesting, and, then, we 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 that illustrate the strengths and limits of the methodology.

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

Market Risk Backtesting

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

[1]Expected Positive Exposure: the average of portfolio values when floored at zero, called "EE" by the Basel Committee.

The VaR capital charge is based on 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 a result, "the backtesting framework ... involves the use of risk measurements calibrated to a one-day holding period"[2].

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 bands for the model:

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

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

The Probability Equivalent of Colour Bands

The original definition of those bands is driven by the estimation of the probability that the model is right or wrong. A green model means that the probability that the model is right is 95%, a yellow model means that that probability is 4.99% and a red band means that that probability is only 0.01%[3].

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

[2]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.

[3]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 bands.

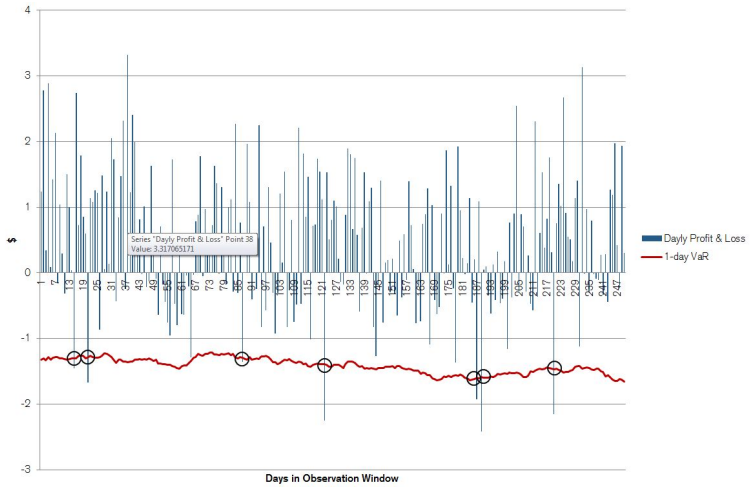

Figure 1: Illustrative example of backtesting exercise for a VaR model. Each circle constitutes an "exception".

assumption, we can use the binomial distribution to compute the probability P of number of exceptions (k) in a twelve month period that that model will give. That probability is given by

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

and is illustrated in Figure 2, left panel, with $N = 250$ and $p = 0.99$.

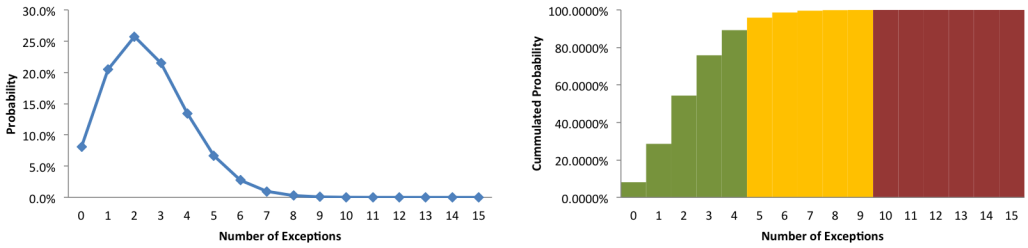

Figure 2: Probability distribution of exceptions, at 99% confidence, that a "perfect" model gives in a 12-month period. Each color marks the range of the corresponding band.

If we now draw a limit in the distribution of exceptions at the 95th and 99.99th percentiles, then the band 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 the Market Risk Measure (MRM) was 10-day VaR but 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), which accounts for the market risks which are not captured by the VaR model.

Regarding backtesting implications into the capital requirements, x is the number at stake. That number is given by the following table:

Band	**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 went beyond 10.

Proposed Framework for Counterparty Risk Backtesting

Regarding counterparty risk models, The Basel Committee has not provided a clear set of rules for backtesting as it has for market risk. Instead, all it has given is a set of guidelines for banks and national regulators [2]. In fact, 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". As a result, in the author's experience, backtesting methodologies in banks have become cumbersome, inconsistent and difficult to relate to each other.

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, that is scientifically sound, practical and that can be easily used by management. 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 for the quality of a model in a given time horizon.

3. Relate that single number measure to the three bands proposed by the Basel Committee, allowing one to classify a model to either the green, yellow or red band.

Methodology

Context and Scope

It is general practice to refer to a CCR model backtest as the backtest 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 is properly modelled by the EPE model, which is the "aggregation" of 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, fee income and trade maturing. In practice, it appears that backtesting EPE numbers, as such, is not possible[4].

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. This is a consequence both of the "average" nature of the EPE and of the Mote Carlo noise. 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, both exact and proxies, tend to be well established as the quantitative finance community has been developing them for quite some time. There are already robust methodologies and testing procedures in place. So we are not going to cover it in this paper.
- The testing of collateral models tends to be done on a scenario by scenario basis. This is because collateral modelling is, in principle, quite mechanical[5] and also because, in practice, the data needed to backtest these models is usually not available.

[4]The author has known of some frameworks where artificial portfolios were constructed for this. However, in his view, the actual practical use of those frameworks is very limited.

[5]With the exceptions of approximations implemented in the algorithm.

- Finally, netting and aggregation models are simple to implement in a Monte Carlo simulator and do not usually need elaborate testing.

So, it appears that the most difficult challenge that EPE models have from a backtesting point of view relates to the RFEs.

Finally, 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 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. This makes the scale of the problem increase substantially compared to VaR models.

The Methodology

Backtesting an RFE means comparing the distribution of the risk factor given by the model over time with the distribution actually seen in the market. In other words, we want to check how the RFE measure and the observed "real" measure compare to each other.

To do this[6], we are going to consider the realized path (a time series) of the risk factor to be tested. That path is given by a collection of (typically daily) values x_{t_i}. 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, if Δ is the time horizon over which we want to test our model (e.g. $\Delta = 1$ month), we proceed as follows (the reader can see in Figure 3 an illustration of this process):

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[7].

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. We repeat continuously the above until $t_i + \Delta$ reaches t_{end}.

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

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

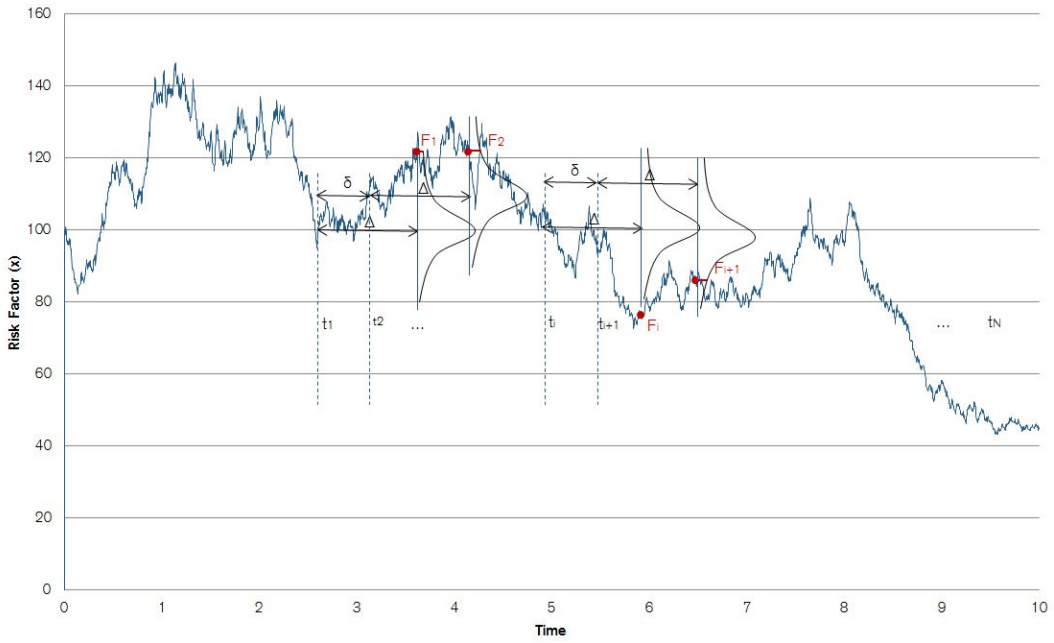

Figure 3: Illustration of backtesting methodology.

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), then $\{F_i\}_{i=1}^N$ will be 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 theoretical cumulative distribution function given by the model and by F_e the empirical cumulative distribution function obtained from $\{F_i\}_{i=1}^N$[8], then we can use, for example:

Anderson - Darling metric:

$$\begin{gathered} D_{AD} = \int_{\infty}^{-\infty} \left(F_e(x) - F(x)\right)^2 w\left(F(x)\right) dF(x) \\ w(F) = \frac{1}{F(1-F)} \end{gathered} \tag{3}$$

Cramer - von Mises metric:

$$\begin{gathered} D_{CM} = \int_{\infty}^{-\infty} \left(F_e(x) - F(x)\right)^2 w\left(F(x)\right) dF(x) \\ w(F) = 1 \end{gathered} \tag{4}$$

[8]For a time horizon Δ

Kolmogorov - Smirnov metric:

$$D_{KS} = \sup_x |F_e(x) - F(x)| \tag{5}$$

Each metric will deliver a different measurement of D. Which of them is the most appropriate depends on how the model being tested is actually used. This decision has some degree of subjectivity by the researcher and practitioner[9].

Having chosen a metric, we can compute now a value $\widetilde{D}$ that measures how good the model is.

The following questions arise 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 $\widetilde{D}$ will never be exactly zero even if the model were perfect[10]. How can we assess the validity of $\widetilde{D}$?

In order to answer those two questions, we can proceed as follows. Let's construct an artificial time series using the model being tested, and then apply our above procedure to it, yielding a value D[11]. The constructed time series will follow the model perfectly by definition, but D will not be exactly zero. This deviation will only be due to numerical "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 probability with respect to $\psi(D)$, then the model is likely to be accurate, and vice-versa[12].

The Three Bands

We are now in a position to extend the clear Basel framework established for market risk to the setting of counterparty risk. If we define D_y and D_r respectively as the

[9] 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.

[10] Zero is attained in the limit: $\lim_{N\to\infty} D = 0$.

[11] That artificial time series must have exactly the same time data as the empirical collection of values x_{t_i}.

[12] Strictly speaking, what we can say is that if $\widetilde{D}$ falls in a range with high probability in $\psi(D)$, then the model is compatible with a "perfect" models with high probability.

95th and 99.99th percentiles of $\psi(D)$, then we can define three bands for the model performance:

- Green band if $\widetilde{D} \in [0, D_y)$
- Yellow band if $\widetilde{D} \in [D_y, D_r)$
- Red band if $\widetilde{D} \in [D_r, \infty)$

This is illustrated in Figure 4 for a simple Geometric Brownian Motion model. With this three-band approach, a financial institution can easily score a model and proceed as found appropriate.

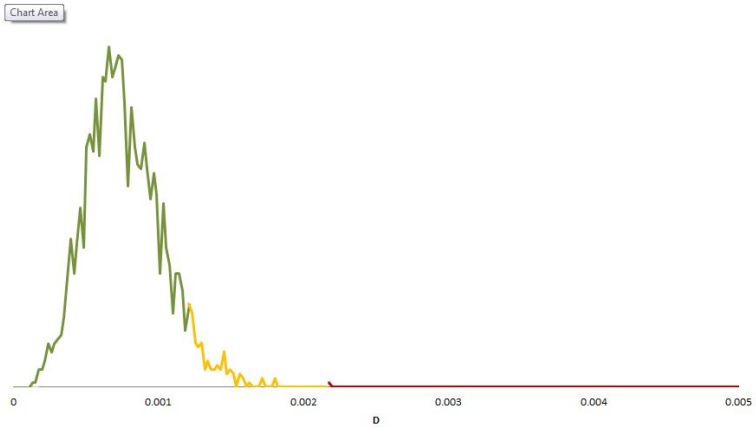

Figure 4: Illustrative example of the distribution of D's compatible with the model.

Examples

We now present a few examples where the mechanics of the methodology are illustrated. The Cramer - von Mises metric was used. Further examples can be found in the Appendix A.

Model with wrong skew

We want to see how the algorithm responds when the skew of the data and the model are different. For this, the author generated an empirical time series with a GBM plus 1-sided Poisson jump process that is taken as the empirical time series, and then it was tested against a simple GBM model with the same volatility as the time series[13].

[13]The empirical time series had a volatility of 34%, a skew of -0.35 and an excess kurtosis of 0.54. The GBM model had a volatility of 34%.

Figure 5 shows how the model responded[14]. 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).

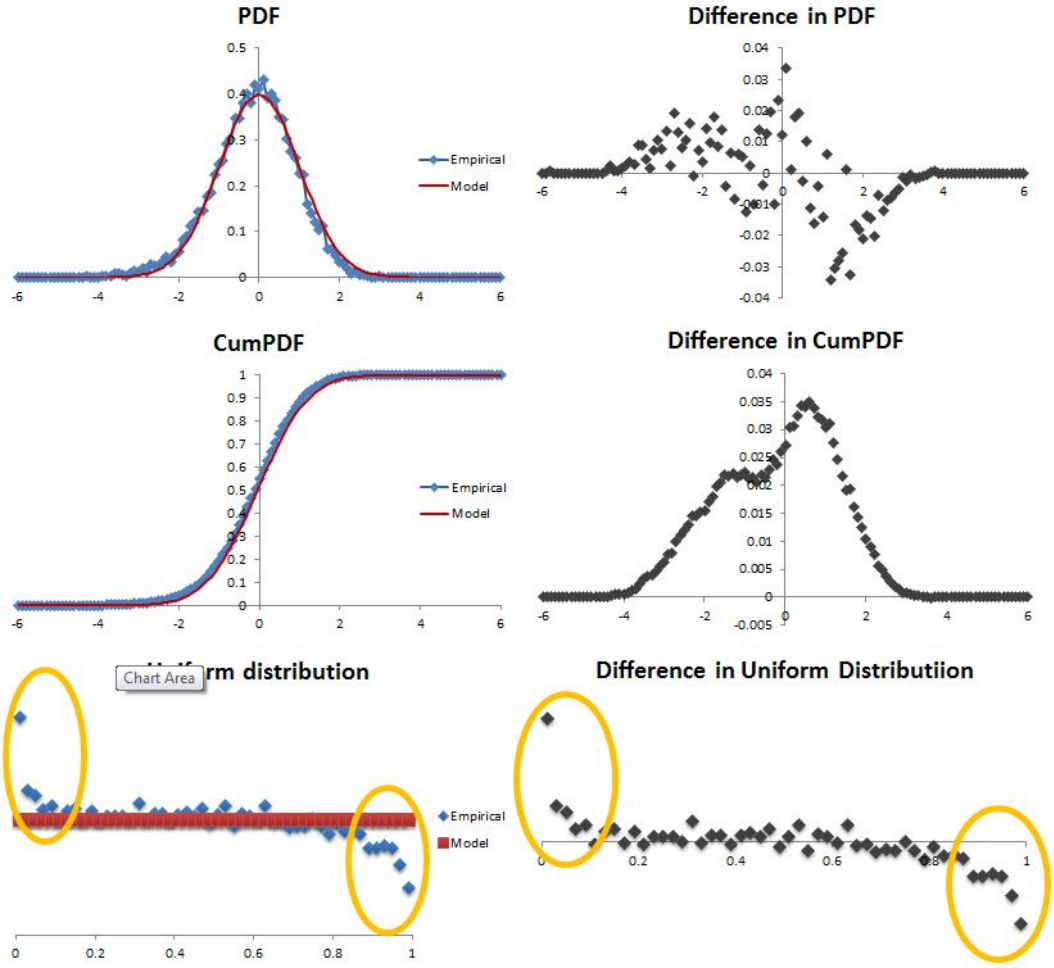

Figure 5: Illustrative example of how the backtest algorithm behaves when the model skew is different to the empirical skew.

Also, the explained colour test was run to assign a band to the time series. The results can be seen in Figure 6.

Real example: S&P500 time series vs GBM model historically calibrated

We now applied the backtest methodology to the S&P500 equity index against a GBM model.

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

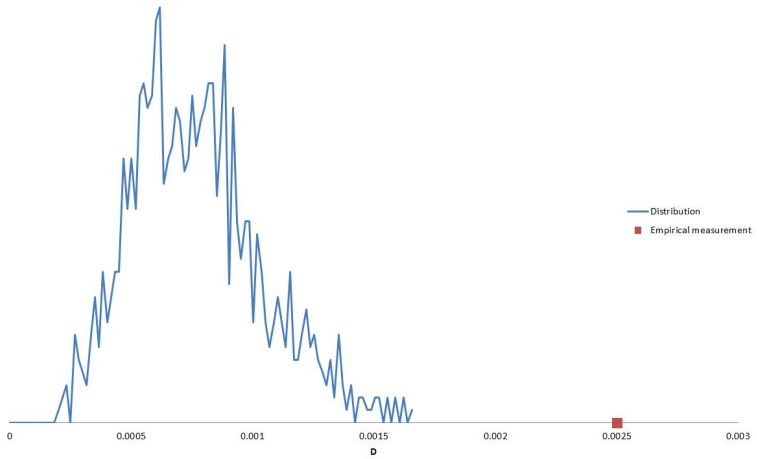

Figure 6: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-plus-jump time series tested against a GBM model with the same volatility as the time series.

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 volatility, implied volatility, etc) 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 same calibrating methodology and frequency in the backtesting exercise as we use in the live system.

This point was left out in the first explanation of the backtesting methodology for simplification of the explanation, but 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. The model we backtest is a simple GBM, with historical daily ($\delta_c = 1$ day) calibration and volatility equal to the annualized standard deviation of log-returns for different calibrating windows T_c. We want to test how the model performs for time horizons Δ of 10 days and one year, and for calibrating windows T_c of three months and three years.

The following table summarizes the scores in terms of color bands:

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

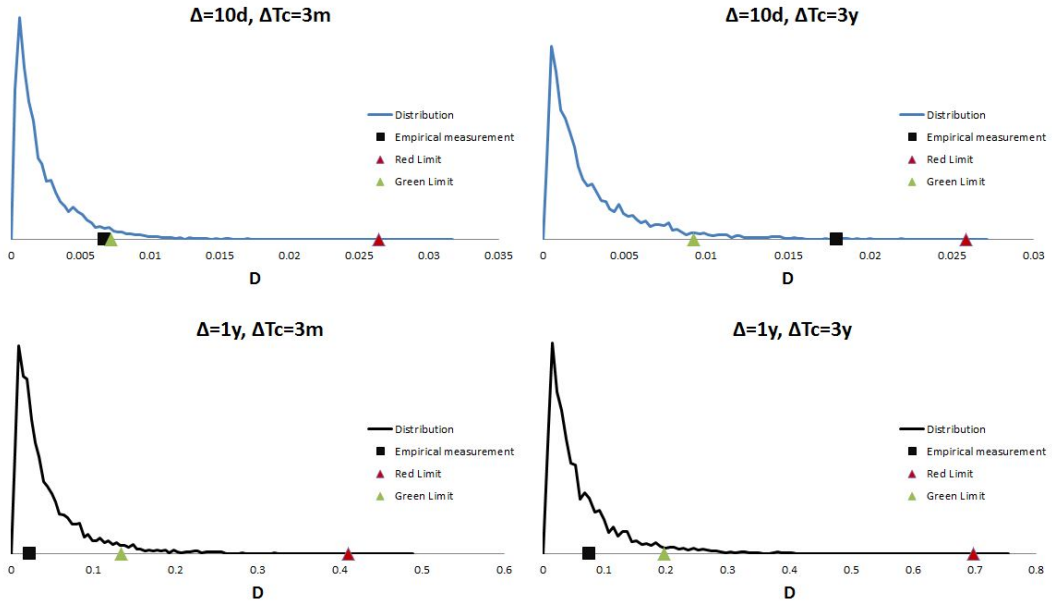

Figure 7: Backtest measurements for S&P500 from 2000 to 2011.

When the model time horizon Δ is short (10 days), a short calibration window ($T_c = 3$ months) makes the 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[15]. 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. This can be seen in Figure 7

Practical Considerations

Let's discuss some important practical considerations.

On autocorrelation - the role of δ

If $\delta < \Delta$, there is autocorrelation in the collection $\{F_i\}$ by construction. However, that autocorrelation also exists in the construction of $\psi(D)$. Hence the 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

[15]The problem is that 3-year historical volatility is not a good predictor of short-term future volatility.

adjustments to compensate for the induced autocorrelation. This autocorrelation effect is discussed in more detail in Appendix B.

Maximum utilisation of available information - the role of Δ and T

In practice, one of the big problems of counterparty risk backtesting is that historical data tends to be scarce: T is usually not long enough relative to the time horizon Δ in which the model needs to be tested. As a result, there may be too few independent points in our test. For example, if we have a backtesting window T of 10 years and we want to measure a model's performance with a Δ of 2 years, we only have 5 independent points[16]. So, the statistical relevance of the backtest can be quite limited by construction.

This is intrinsic to CCR backtesting and, in the author's opinion, there is no way to get away from it. However, this limitation is again captured in this methodology, as the width of the $\psi(D)$ "automatically" expands as T/Δ decreases. This is because, when we calculate $\psi(D)$ in a set up where T/Δ is small, the algorithm will provide a wide range of D's compatible with the model, as the independent information is relatively scarce. For this reason this methodology seems to be optimal as it "automatically" makes most of the available information, however abundant or scarce it may be. This effect is discussed in more detailed in Appendix C.

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 is firstly driven by the availability of good quality data. Once this is provided, 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 becomes mission impossible[17]. There is no set rule to define what is too much or too little; in practice, this can only be left to the experience and market knowledge of the researcher[18].

[16] Assuming no auto-dependency in the data.

[17] If there is sufficiently large amount of data, most models, if not all, will fail the backtest. See Appendix D for more details.

[18] In reality, the practitioner should not worry to much about this at first, as the constrain here tends to be lack of available good data, and so this problem is not even seen in most cases.

Δ will be determined from the typical tenor of the trades in the portfolio under testing, subject to availability of data. It should cover a range of values *up* to the typical maturity of the portfolio being tested - no need to test a Δ beyond 5 years, for example, if most of the portfolio affected by the RFE matures in 5 years. In this case the author would suggest Δ of one day, one week, twp weeks, one month, three months, six months, one year, two years, three years and five years, with special attention to those Δ where the risk of the portfolio tends to peak.

Regarding δ, the author recognises that it is not clear him what is ideal, but some anecdotal evidence[19] 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.

Further details on this subject can be found in the Appendix D.

Multiple-Δ scores V. Single-Δ score

In order to assess the validity of a model for CCR, we should test it not only for a number of representative time horizons Δ, as just explained, but also for a number of representative risk factors (e.g., several FX rates for an FX model). As a result we can easily end up with up to a few hundred colour scores.

We could have the temptation of building a final aggregated score based on those many sub-scores, but in the author's opinion and experience this will be difficult to manage in practice as the granularity of the analysis can get lost. In reality, he has found most practical to build a coloured table with the most representative scores, not more than one hundred, find patterns in it (e.g., most emerging market currencies fail for Δ greater than 2 year) and modify the model as needed until most scores are green and/or we can explain and manage those yellow and red scores. More details can be found in Appendix E.

Numerical approximations

The calculation of D will be done numerically using a number of approximations. This will introduce some noise in these calculations. However, the same noise will be introduced in the calculation of $\psi(D)$ and, hence, this noise is, also, "automatically" considered in the assessment of $\widetilde{D}$.

Asset Class Agnostic

This methodology can be applied to any asset class: interest rates, foreign exchange, equities, commodities, credit, etc. There are no limitations in this respect.

[19] See Appendix D.

Backtest of dependency structures

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. However, it must be noted that, in the authors experience, this should be done after all factors are scored independently, one-by-one, as explained in Appendix G. Otherwise the source of red scores will be very difficult to track. This is discussed in more detail in the Appendix F.

Testing the whole probability distribution

Often, risk management uses a given percentile in the distribution function to measure counterparty risk (e.g., 95% PFE[20]). For this reason, sometimes risk models are tested by counting the number of exceptions outside of a given percentile envelopes. However, that methodology is sub-optimal for regulatory capital models where the key measure of risk is the EPE[21]. This is because EPE is 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. In fact, the framework explained in this paper can easily be adapted to test specific parts of the probability distribution by changing the weight function $w(F)$.

Structural changes in the market

The explained methodology provides a score for a model during a time period T, but if that period contains a structural change in the market (e.g., 2008), it provides no information what-so-ever as to how the model reacted to capture those changes. A way to do this is buy performing a "rolling window" test. In this test, we can do the test with a relatively small window T_{rw} that we then roll over the period T. If we define a parameter Z as $\widetilde{D}$ over a critical D_c, where D_c can be the D value that marks the frontier between the green and yellow bands, then we can have a time series Z_t over the testing window T from where we can easily assess the model performance during structural changes in the markets: when $Z_t < 1$ the model is good, but when $Z_t > 1$ the model is having problems[22].

Further details can be found in the Appendix G.

[20] Potential Future Exposure.
[21] The same could be applied to CVA if it was to be backtested.
[22] If wanted, the same study can be done with D_c being the frontier between the yellow and red bands.

Historic V. implied model calibrations

Risk models tend to be calibrated historically, but pricing models (e.g., CVA) tend to have 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.

Conclusions

The author has proposed a backtesting framework for Counterparty Credit Risk models that provides a set of green/yellow/red scores to a model, resembling the widely used approach in the market risk area. On this way, a model receives a colour score for each time horizon Δ and relevant risk factor that enables proper good model management.

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 horizons Δ need to be chosen by considering the maturities of the portfolio being tested. Also the most significant risk factors for the portfolio need to be considered. 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 successfully implemented this methodology in a Tier-1 financial institution. The report generated was sent to the regulators seeking IMM model approval. The model was approved a few months later.

Acknowledgements

The author would like to thank Ahmed Aichi, Piero Del Boca, Chris Kenyon and Chris Morris for interesting comments and remarks to this piece of work.

Appendix

A Examples

Model with wrong volatility

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

Figure 8 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 9 illustrates the classification procedure. Here, the "empirical" time series had a volatility of 35%, while the model was a GBM process with a volatility of 40%. The backtest procedure yielded $\widetilde{D} = 0.003$, while the red band for this backtesting exercise starts at 0.0018; hence, the algorithm classifies this model into the red band for the time series, indicating that it is inaccurate.

Model with wrong kurtosis

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, we generated an empirical time series using a GBM plus a 2-sided and symmetric jump process, and tested this "artificial" data 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 10 and 11. The kurtosis effect is best depicted in Figure 10: the symmetric wave we have in the blue dots indicates high kurtosis in the data. In Figure 11 the time series had a volatility of 58%. and was backtested against a GBM model with the same volatility. The model fell in the red band.

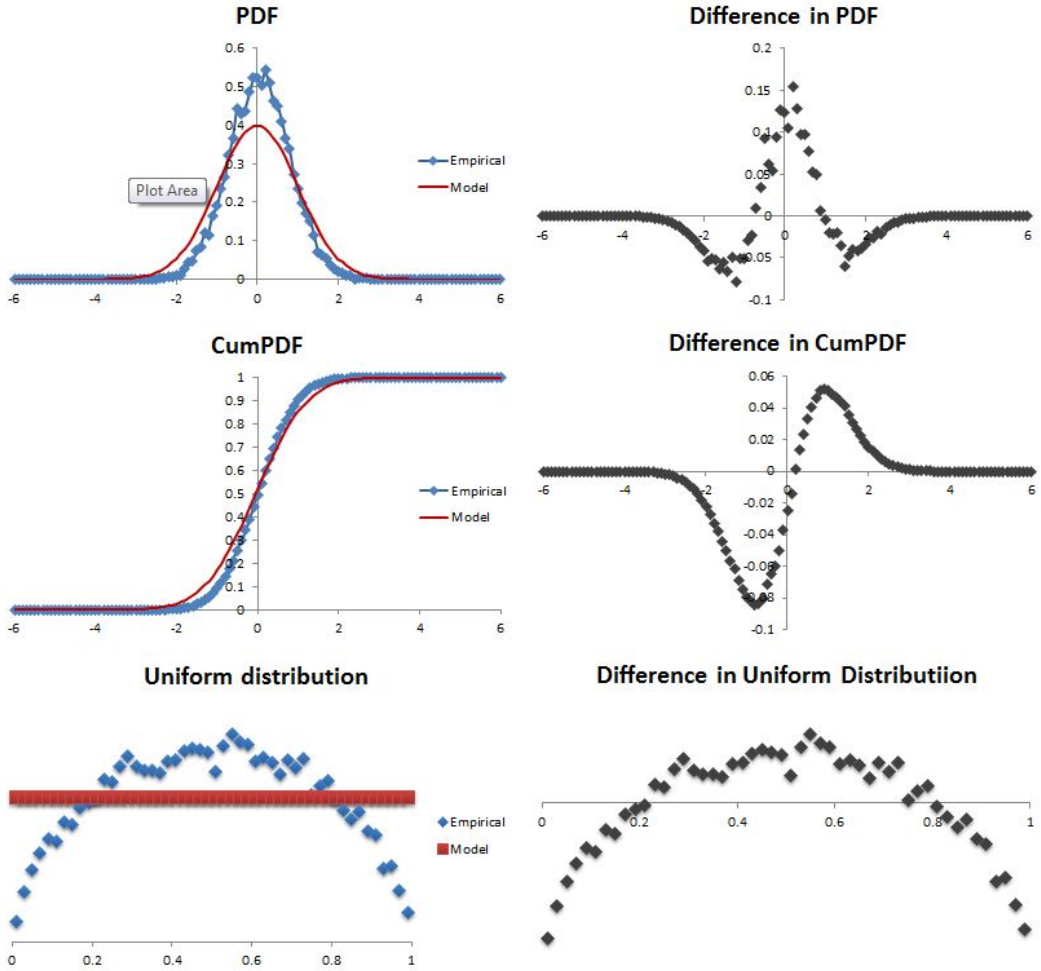

Figure 8: Illustrative example of how the backtest algorithm behaves when the model volatility is higher than the empirical volatility.

B On Autocorrelation

As said, 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.

This leads to one clear question: what is the optimal δ?

The author admits that he has no conclusive answer yet to this question. After discussing this problem with a number of colleagues, no clear conclusion was attained. The general

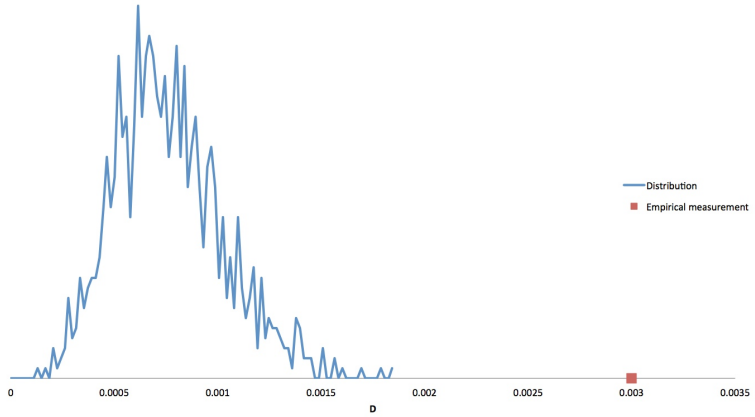

Figure 9: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-driven time series tested against a GBM model with a higher volatility.

consensus seemed 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.

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 in the text, 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 12. The obtained results suggest that the value of δ may be irrelevant. In all four cases, the models score in the green band, and in all cases $\widetilde{D}$ is in the vicinity of 65% of the green limit.

It must be said that this constitutes by no means definitive evidence that the value of δ is irrelevant, but it is a healthy sanity check.

Further to that, perhaps the case in which it makes sense to have a δ smaler than Δ is when the number of independent measurement points is very small and a non-integer. For example, let's say that, $T = 5$ years but we are interested in the model performance in a time horizon $\Delta = 2$ 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" year, and so by picking a $\delta = 1$ years, the calculation of D will maximize the use of the available information.

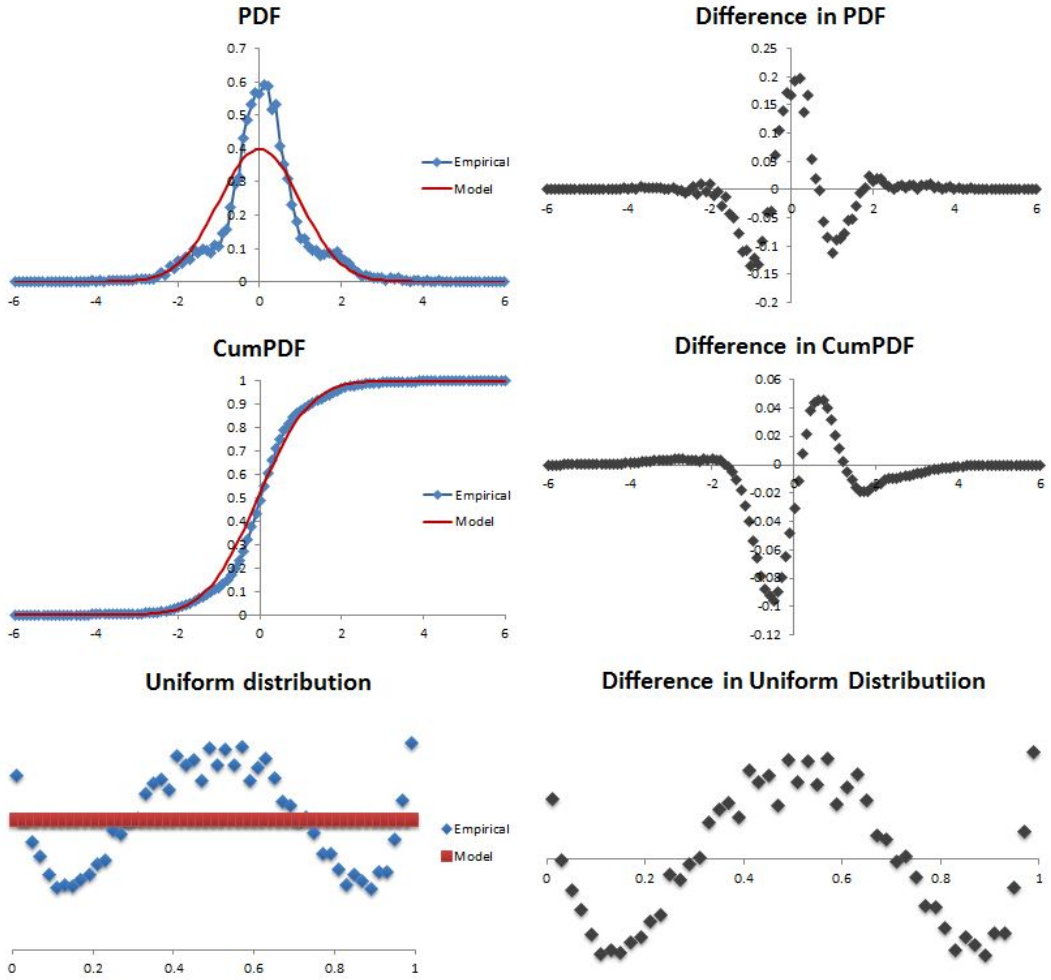

Figure 10: Illustrative example of how the backtest algorithm behaves when the model kurtosis is lower than that in the empirical data.

C Maximum Utilisation of available information

We have mentioned that the statistical relevance (i.e., the amount of independent information available) of the backtest is determined by T/Δ. For example, let's say that $\Delta = 1$ month and that $T = 1$ year. Assuming independence in the time series, we have 12 sample points which are independent[23]. 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. In principle, we should expect that it should be easier to find a model that scores a green in the former than in the later case. This is because, in the later case, we will have more flexibility in the model side to match 12 independent points compared to the 120 points

[23]Let's forget about calibration details for now.

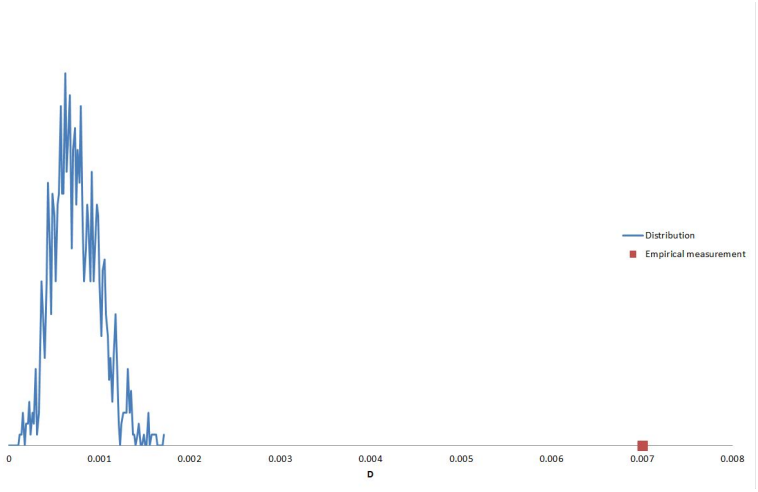

Figure 11: $\psi(D)$ (blue) and $\widetilde{D}$ (red dot) for a GBM-plus-double-jump time series tested against a GBM model with the same volatility as the time series.

in the later case.

One of the key strengths of this backtesting methodology is that it accounts for this accounts for this and it makes, by construction, most of the available information. This is done by automatically expanding or shrinking the width of $\psi(D)$ as needed.

Best to illustrate this with a practical example. Let's re-run our S&P500 example comparing results with T of 10 years and of 60 years. That is, we are expanding the backtesting window T to 1950. In particular, we are running tests when $\Delta = 1$ year for both $T = 10$ years and $T = 60$ years. The results are shown in Figure 13. 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.

D What Parameters to Use

T is the time window from which we want to extract the "real" measure. For that reason, as explained in Appendix C, the larger this T is the more difficult it is to find a model that passes the backtest, as there is less room to manoeuvre in the model side. So... what is the optimal T?

First of all, a 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 often lacks quality. If that problem is solved, on the one hand, T should be long enough (relative to Δ) so that the backtest algorithm has statistical relevance. On the other

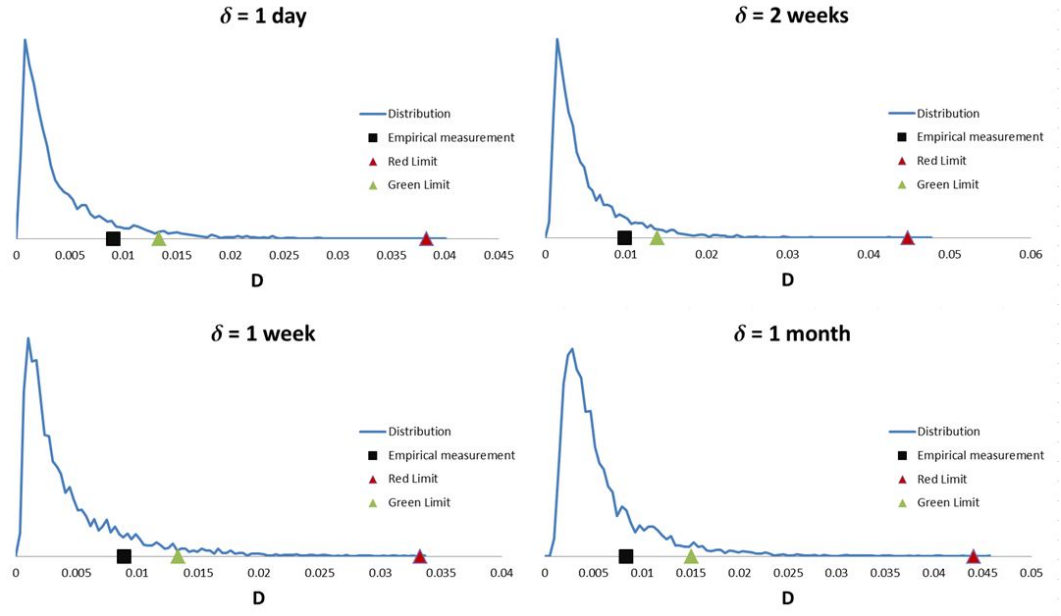

Figure 12: Illustration of backtesting algorithm behavior by changing δ

hand, large values of T will make the task of building a model difficult to a degree that could be impractical.

In the author's view, unfortunately there is no set rule for the optimal T. In practice, the researcher and practitioner will have to balance out all these constrains to come up with a view on the optimal T.

Δ represents the time horizon in which we want to test the model. In this case, there is a clear optimal value for the maximum Δ. 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 maturing in 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 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 sure that the long-term RFE behavior is reasonable. Given a maximum Δ, we should test our model in a representative range of values from one day to that maximum Δ.

δ represents the granularity in the backtesting calculation. As discussed in Appendix B, it is not clear to the author whether a small δ is better "per se". 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

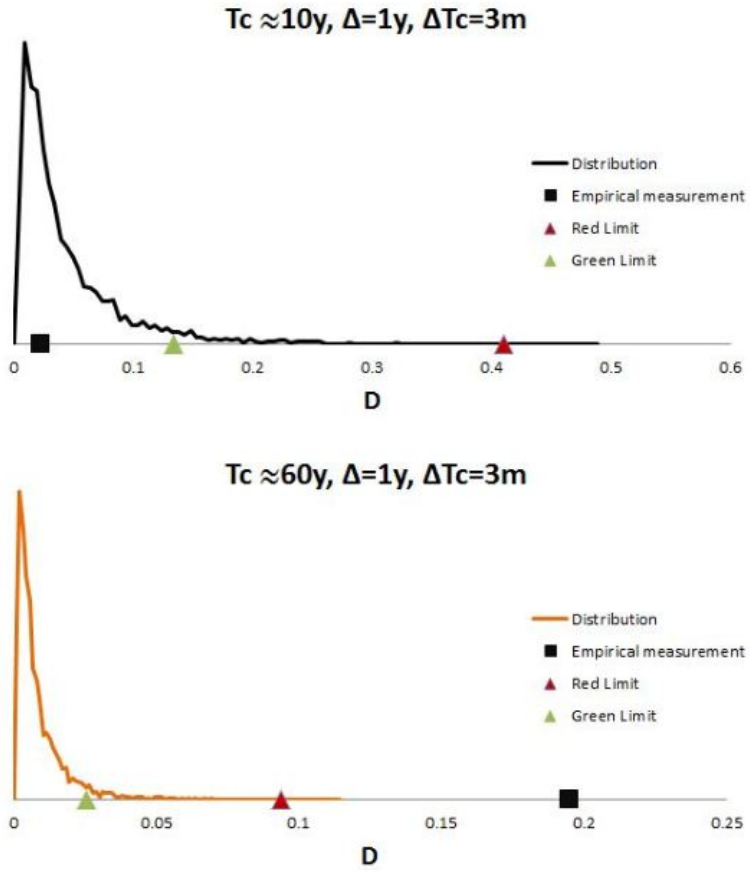

Figure 13: Backtest measurements for S&P500 from 2000 to 2011 (top) and from 1950 to 2011 (bottom).

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 in Appendix B, the case in which the algorithm will benefit from a smaller δ is when T/Δ is small and non-integer.

E Multiple-Δ scores V. single-Δ score

A practical consequence of the mutiple Δ's that we need to use in the test is that a model will not have one score, but a collection of them, one per Δ. Typically, we would have between 5 to 10 Δ's.

Further to that, in reality we should have a number of those collections, as an RFE should be tested against several relevant risk factors. For example, an FX RFE will typically need to be tested against all major currencies (USD, GBP, EUR, CHF, JPY)

and against all other currencies significantly relevant to the financial institution. In general, we may have ten to thirty risk factors to test.

As a result, we will end up with between 50 and 30 scores for the model. How could be proceed to make these measurements useful?

We could define an overall score, using some formula that agregate all those scores into a single one. However, in the author's view and experience, this will have quite a limited use, as the granular information provided by the analysis will be too hidden. In his view, the best way to proceed is as follows:

1. Produce one single sheet with a table showing, with colours, scores for all Δ's and risk factors. Ideally, we do not want more than 100 scores. In this table, we will have some green, some yellow and some red boxes. This gives a very good eye-ball snap shot of the model performance. We are aiming at lots of green, some yellow and perhaps, a few red boxes.
2. Find patterns in the yellow and red boxes. For example, "red scores tend to happen in emerging market currencies for Δ greater than 2 years".
3. Based on those patterns, try to modify the model so that those yellow and red boxes tend towards the green.
4. Proceed in a loop in the previous steps until we have mostly green boxes.
5. Finally, find the reason behind the remaining non-green scores, if they exist; e.g., poor quality data, a currency linked to a government default, etc. Then estimate the impact of those exceptional cases in your portfolio and assess the need of a special process for those cases in the organisation.

In the author's experience, this is the optimal way to develop an RFE model that backtests well.

F Backtest of Dependency Structures

Let's say that we have a model for a curve consisting of a number of tenor points, which have a certain dependency structure (e.g., a yield curve). 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 what-so-ever as to the quality of the model with regards to the dependency structure between those points.

In this paper we have seen how to apply the explained 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 done by expanding the calculation of D to multi-dimensions. 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 check for 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 one dimension. Then, multi-dimensional tests can be implemented. This way, should the final joint-RFE test be non-satisfactory, potential problems can be better isolated, identified and managed.

G Structural Changes in the Market

The markets undergo periods of stress from time to time. The most recent one started in the credit asset class in 2007. These phenomena occurs for a number of reasons which tend to gravitate around wrong economic fundamentals, markets drying out, important imbalances in the size and/or price of markets, 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[24].

Stochastic models cannot capture these events, they are not build for that. But we ideally want models which react quickly to changing market conditions and which are valid for a wide range of market environments[25]. Because of that, assessing the quality of an EPE model requires also assessing how the model behaves under structural changes in the markets.

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 and that receives a green score. What that means is that the explained methodology provides a green score to the overall performance of the model in the period T, but it gives no indication as to how the model performed during the 2007-09 crisis. In order to assess this, we can

1. Implement a window T of 2 years starting in 2000 and roll it forward up to 2010, obtaining a rolling $\widetilde{D}_t$.

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

[25] 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.

2. In each of those rolling tests, record the critical value $D^c_{g,t}$ that sets the frontier between the green and the yellow band, and $D^c_{r,t}$ that sets the frontier between the yellow and the rend band.
3. Finally, divide each $\widetilde{D}_t$ by each D^c to obtain a time series of rolling scores, that we are going to call Z-scores: $Z_{g,t}$ and $Z_{r,t}$.

If $Z_{g,t} < 1$, the model is in the green band, but when it goes over that value, the model goes into the yellow or red band. A parallel analysis can be done with $Z_{r,t}$. On this way, by plotting each time series Z_t, the researcher can assess the quality of the model response to changing market conditions and, when the quality decreases, have an idea of how bad the problem is.

In fact, the author has seen precisely this kind of behaviour in the backtesting work he has done for one of his clients. 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 band, went to the yellow/red bands when the backtesting period overlapped with the peak of the credit crunch, and then went back to green when the (historical) calibration window overlapped sufficiently 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 but it was shown that there was not need for a different model for stressed market conditions.

References

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

[2] *Sound practices for backtesting counterparty credit risk models*, tech. rep., 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.