Paper submitted for the:

Enterprise Risk Management Symposium

April 22-24 2013, Chicago, IL
A Quantitative Metric to Validate Risk Models

William Rearden\textsuperscript{1} M.A., M.Sc.
Chih-Kai, Chang\textsuperscript{2} Ph.D., CERA, FSA

Abstract
The paper applies a back-testing validation methodology of economic scenario generating models and introduces a new D-statistic to evaluate the robustness of the underlying model during a specified validation period. The statistic presented here can be used to identify the optimal model by repeating calibrations with changing initial parameters. It can compare between calibration methods, be used to rank between models, and provide a single concise reporting metric for on-going model monitoring. To illustrate this methodology and ranking between models, the closed form bond pricing solutions of the CIR 1- and 2-factor models are used. CIR model parameters were estimated using Matlab’s built-in least squares minimization routine. At each observation date during the validation period, a time-weighted point estimate of the error between the model and actual market term structure is calculated. Finally, the maximum of these time-weighted points across the validation duration is introduced as the D-statistic. The robustness of the D-statistic is improved by implementing a first-order autoregressive sampling bootstrapping algorithm, which generates an empirical distribution for calculating the standard error of the D-statistic.

Key Words: Economic Scenario Generators, Solvency II, Interest Rate Modeling, CIR 1- and 2- Factor Models, Calibration, Back-Testing, Bootstrapping Algorithm.

I. Introduction
If we cannot trust doctors when it comes to matters of health,\textsuperscript{3} then the validation of economic models becomes that much more important for matters of solvency capital. As European life insurance companies adopt risk-related models for Solvency II regulatory requirements, Economic Scenario

\textsuperscript{1} Contact: william.rearden@gmail.com, Managing Partner, 出來 Global, Canada.
\textsuperscript{2} Contact: ckchang@fcu.edu.tw, Department of Risk Management and Insurance, Feng Chia University, Taiwan.
\textsuperscript{3} See Abigail (2011)
Generating models have gained traction within the industry as a market consistent methodology for asset and liability valuation. By generating market efficient stochastic scenarios, an institution can demonstrate its solvency under worst-case tail scenarios to regulators. Continuous-time short-rate interest rate models provide the most efficient way to generate sufficient stochastic interest rate paths to fairly evaluate economic scenarios.

How can a risk manager who faces 20 stochastic interest rate models and 10 interest rate markets validate all 200 possible combinations of these candidate models and interest rate markets? According to CEIOP’s advice for the governance of internal models, in Article 120, risk managers are responsible for ensuring the ongoing appropriateness of the design and operations of the internal model. Furthermore, Article 116 also requires a robust governance system of the internal model operating properly on a continuous basis. Moreover, requirements described in Article 44 make the regular review or decision from various models and markets too time-consuming and exhaustive; adding to the difficulty is informing administration, management, or supervisory about the performance of the internal model.

Many concerns arise when trying to determine the most efficient set of model parameters for a particular type of model. Most calibration methods involve local minimization of model-to-market spot values at a specific point in time. Plenty of literature discusses calibration methods that attempt to return the most optimal set of model parameters. However, in practice it is impossible to determine whether the calibrated parameters found a global minimum market-to-model error. Once the parameters are determined, applying constant parameters to a model results in continuously changing modeling errors due to dynamically changing markets. With poorly calibrated results, these errors increase.

In this paper we introduce a validation methodology that back-tests the calibrated parameters before extending the application of locally calibrated models for macro risk management. In particular, we
introduce this metric as the D-statistic, calculated as the maximum of the time-weighted term structure market-to-model error from each date during the validation period. This D-statistic is similar to the Kolmogorov-Smirnov Statistic and provides a monitoring metric to efficiently communicate model consistency. As an additional advantage, the quantitative essence of an interest rate model, whether parametric or not, can easily be explained by this metric. Each application of the algorithm discussed in section IV returns a unique D-statistic. Ranking the D-statistics will aid in identify an optimal set of calibrated parameters between calibrations under different initial parameters, and serve as tool to compare and rank conservatism between models.

To illustrate the validation approach we consider the classical Cox Ingersoll and Ross (1985) continuous-time short-rate interest rate model (CIR). The model’s preclusion to negative interest rates and mean reversion makes it an excellent model for generating several idealistic stochastic interest rate scenarios. We exploit the known model-to-market calibration error to emphasize the efficacy of the validation methodology introduced.

II. The Cox, Ingersoll and Ross (CIR) Model

This paper assumes the interest rate process can be represented as a diffusion process through time. The advantage of such a representation is that the entire zero-coupon bond curve can be conveniently described by the distributional properties of the instantaneous short-rate. The price at time $t$, of a unit amount of currency at time $T > t$, can be expressed as the expected present value of the interest rate diffusion:

$$P(t,T) = E_t \left\{ e^{-\int_t^T r(s) ds} \right\}$$

The disadvantage of a continuous-time representation implies that a poor model of the instantaneous short-rate will also produce a poor evolution of the term structure. It is assumed that, by increasing the flexibility of a model with more factors, we may improve a model’s accuracy to real-world observations.
CIR 1-Factor Model

We consider the short-rate process solution to the following CIR instantaneous short-rate diffusion model, under the risk-neutral measure \( Q \) as:

\[
dr(t) = k(\theta - r(t))dt + \sigma\sqrt{r(t)}dW(t) \quad \text{with} \quad r(0) = r_0
\]

with \( r_0, k, \theta, \sigma \) as positive constants. The condition \( 2k\theta > \sigma^2 \) is imposed to ensure that the origin is inaccessible to the process so that instantaneous short-rate \( r \) remains positive. By change of measure from the risk-neutral \( Q \) to the objective real-world measure \( Q_O \), the CIR model allows us to discount analytically at time \( t \) a unit amount of currency at time \( T > t \) as:

\[
P(t, T) = A(t, T)e^{-B(t, T)r(t)}
\]

where the derivation details of \( A(t, T) \) and \( B(t, T) \) for the affine CIR model is concisely presented in Brigo and Mercurio (2006). For brevity, the rigorous detailed derivations of the CIR model are omitted and only the model’s analytical results are reported.

The CIR model is widely known for precluding negative interest rates, offering a mean reverting expression of the short-rate process to mean \( \theta \) at speed \( k \), and being analytically tractable with many well established closed-form interest rate derivative formulas. These model advantages are not the reason for selecting this model, but rather to illustrate its shortcomings – namely, the model produces an endogenous term structure that does not match the real world, no matter how well the parameters are chosen. The problems are further amplified under poor calibration, rendering it pointless for longer term pricing. Later
sections will demonstrate the modeling error and the difference between the actual and calibrated term structures. The preclusion of negative interest rates and mean reversion makes the CIR 1-factor a powerful risk management modeling tool for generating stochastic interest rate scenarios.

**CIR 2-Factor Model**

A 1-factor model assumes that at every instant all maturities along the curve are perfectly correlated, so that a shock is equally transmitted across the curve. However, this is not empirically observed. A 2-factor model is justified based on principal component analysis, since two factors can explain over 95% of the total interest rate variation. The additional factor increases model flexibility by relaxing perfect correlation, so that the joint dynamics depends on the instantaneous correlation function $\rho$. For the CIR 2-factor model however we assume no correlation, $\rho = 0$, since the square-root non-central chi-square process cannot maintain analytical tractability with nonzero instantaneous correlations.

The CIR 2-factor model is then defined as:

$$ r(t) = x_1(t) + x_2(t), \quad \text{where } x_1(0) = x_2(0) = 0 $$

$$ dx_1(t) = k_1(\theta_1 - x_1(t))dt + \sigma_1\sqrt{x_1(t)}dW_1(t) $$

$$ dx_2(t) = k_2(\theta_2 - x_2(t))dt + \sigma_2\sqrt{x_2(t)}dW_2(t) $$

with instantaneous-correlated sources of randomness, $dW_1dW_2 = 0$. Now the price at time $t$, of a unit amount of currency at time $T > t$ can be express as a generalization of the 1-factor:

$$ P(t,T) = \prod_{i=1}^{2} A_i(t,T) e^{-\Sigma_{j=1}^{2} B_j(t,T)x_j(t)} $$
where the well-known expressions for $A(t,T)$ and $B(t,T)$ for the affine CIR model is presented in Brigo and Mercurio (2006).

III. Data

The observation period for this paper is 2 years from June 1, 2007 to June 6, 2009. For the purpose of analyzing the period during the financial crisis, the data is split into 2 sets: validation and monitoring. The validation data is from June 1, 2007 to June 6, 2008 and the monitoring data is from June 13, 2008 to June 5, 2009. The term structure data for this paper is the middle of the week Libor and Swap rates provided by Data-Stream; where 1, 3 and 6 month rates are Libor, while year 1 through 10, 12, 15, 20, 25 and 30 years are the reported Swap rates. The annualized overnight rate, used as the instantaneous short-rate for CIR modeling and calibration, is imputed from the 1 month Libor rate by the following identity:

$$\left(1 + \frac{\text{OvernightRate}}{365}\right)^{365} = \left(1 + \frac{\text{1MonthLiborRate}}{12}\right)^{12}$$

Table 1: Overnight rate summary Statistics and Principal Component Analysis\(^4\)

<table>
<thead>
<tr>
<th></th>
<th>Validation Data</th>
<th>Monitoring Data</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observation</td>
<td>54</td>
<td>52</td>
<td>-2.00</td>
</tr>
<tr>
<td>Min</td>
<td>2.38%</td>
<td>0.31%</td>
<td>-2.07%</td>
</tr>
<tr>
<td>Max</td>
<td>5.81%</td>
<td>4.58%</td>
<td>-1.23%</td>
</tr>
<tr>
<td>Average</td>
<td>4.25%</td>
<td>1.54%</td>
<td>-2.71%</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1.17%</td>
<td>1.23%</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

\(^4\) Principal Component Analysis is based on the following subset of yield rates from each data set: 1, 3, and 6 month Libor rates, and 1, 2, 3, 5, 7, 10, 20 and 30 year swap rates.
1st Principal Component  |  97.03% |  90.50% |  -6.53%
2nd Principal Component  |  1.99% |  7.76%  |  5.77%
3rd Principal Component  |  0.83% |  0.95%  |  0.12%

Principal component analysis and some observations between validation and monitoring periods:

- The average short-rate between periods dropped by 2.71% from 4.251% to 1.543%, while the short-rate standard deviation remained relatively stable, increasing slightly by 0.06%.
- From the 1st principal component the 1-factor model explanatory power decreased by 6.53%, but still explained over 90% of the term structure variation during the monitoring period.
- From the 2nd principal component the tilt variation of the yield curve increased by 5.77%.
- From the 3rd principal component the convexity of the yield curve increased in variation by 0.12%.

IV. Model Selection and Validation Methodology

Step 1: Model Selection

From principal component analysis of the validation data, a 1-factor model can capture over 90% of the data’s interest rate variation, a 2-factor over 98%. Based on these results, we consider a CIR 1- and 2-factor model to illustrate the methodology of validating models.

Step 2: Calibration

Our calibration employs least squares error minimization to mid-week maturity date data. Based on the initial set of parameters, it continues to change the parameter vector, $\theta$, to minimize the error between model and market price for a given date, of the following objective function:

$$\text{ObjectiveFunction}(\theta) = \min \left\| \sum_{t=1/12}^{3} \left( \frac{\text{ModelPrice}_t(\theta) - \text{MarketPrice}_t}{\text{ModelPrice}_t(\theta)} \right) \right\|$$
where the term structure maturity dates, \( t = \frac{1}{12}, \frac{3}{12}, \frac{6}{12}, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20, 25, \text{and} \ 30 \).

For illustration throughout this paper we consider the mid-week term structure data of date June 6, 2008 for calibration; for this date the CIR 1- and 2-factor model yields the following parameters, 
\[
\hat{\theta}^1(k, \vartheta, \sigma, \gamma) = (0.0585, 1.378, 0.2419, -0.0998) \quad \text{and} \quad \hat{\theta}^2(k_1, k_2, \vartheta_1, \vartheta_2, \sigma_1, \sigma_2, \gamma_1, \gamma_2) = (0.1671, 1.9436, 10.39, 0.0010, 0.0081, 0.8958, 1.593, 0.9826),
\]
respectively. Based on the calibrated results Table 2 reports the error between the model price and market price for the selected maturities for June 6, 2008.

Table 2: Comparing Selected Modeling Errors between Models for Calibration Date June 6

<table>
<thead>
<tr>
<th></th>
<th>CIR 1-Factor</th>
<th>CIR 2-Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Month</td>
<td>0.004%</td>
<td>0.191%</td>
</tr>
<tr>
<td>2 Year</td>
<td>0.186%</td>
<td>1.070%</td>
</tr>
<tr>
<td>5 Year</td>
<td>0.103%</td>
<td>0.392%</td>
</tr>
<tr>
<td>10 Year</td>
<td>0.036%</td>
<td>0.148%</td>
</tr>
<tr>
<td>20 Year</td>
<td>0.552%</td>
<td>0.661%</td>
</tr>
<tr>
<td>30 Year</td>
<td>0.427%</td>
<td>0.599%</td>
</tr>
<tr>
<td>Overall Time-Weighted Average Model Error(^5)</td>
<td>0.26%</td>
<td>0.40%</td>
</tr>
</tbody>
</table>

On this particular date, June 6, 2008, the 1-factor overall time-weighted average modeling error of .26% is less than .40% of the 2-factor model.

**Step 3: Model Validation**

Based on the calibration parameters estimated in step 2, these calibrated parameters are used to calculate the modeling error between the model and market prices to each historical data during the validation.

\[
^5 Average = \frac{\sum_{t=1/12}^{30} t \cdot \text{ModelError}_t}{\sum_{t=1/12}^{30} t}
\]
period. At each date there are several maturities. The modeling error differs for each point on the yield curve; this error often increases with the yield to maturity. To circumvent this multidimensional modeling error problem, a time-series is generated by calculating the time-weighted average modeling error, as in Table 2 for each date in the validation period. Assuming a fixed set of model parameters from calibration, Figure 1 illustrates how this time-weighted average modeling error changes with weekly changes of the short-term interest rate. The maximum modeling error is defined as the D-statistic to mimic the Kolmogorov-Smirnov statistic. This D-statistic improves the vetting of risk models by comparing the maximum historical modeling error, similar to how the Kolmogorov-Smirnov statistic is used to compare the maximum difference between distributions. Calculating the D-statistic and back-testing the model steps away from a snapshot of current model error analysis at calibration, and moves toward a more macro level model efficiency for risk management. Furthermore, the D-statistic illustrates how well the modeling error behaved with validation data without the need to fully understand the esoteric rigor of the underlying models. Moreover, bootstrapping of the modeling error time-series provides an empirical distribution, which improves the robustness of the D-statistic.

**Figure 1: Validation Time-Weighted Average Modeling Error**

Step 4: Model Evaluation
Table 3 highlights the key modeling error statistic information, which can be used as a model evaluation criterion. By evaluating D – the maximum error over the model validation period, the statistic can be used to address the following modeling concerns.

- The statistic provides a quantitative point estimate to compare the trade-off between modeling error and model complexity. In our illustration the lower modeling error (D=10.45% with standard error 1.24%) of the 2-factor model is too small to justify its complexity over the simpler 1-factor model (D=10.54% with standard error 1.58%).
- The statistic also provides a ranking between model calibration methods, for example between least-squares minimization versus Kalman filtering. Since both methods are likely to calculate different sets of parameters, the D-statistic can be used for comparing.
- A more optimal set of parameters will be selected by re-calculating the D-statistic for each calibration trial that uses a different set of initial parameters. The trail with the smallest D-statistic then provides additional robustness for the final estimated parameters selected.
- Finally, the D-statistic can highlight differences between similar and different models to different market data, for example, comparing the differences between models with monthly, weekly, and daily calibration.

### Table 3: Comparing Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Date of Maximum Model Error</th>
<th>D – Maximum Modeling Error</th>
<th>Standard Error (D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIR 1-factor</td>
<td>July 6, 2007</td>
<td>10.54%</td>
<td>1.58%</td>
</tr>
<tr>
<td>CIR 2-factor</td>
<td>March 31, 2008</td>
<td>10.45%</td>
<td>1.24%</td>
</tr>
</tbody>
</table>

---

6 Caution must be exercised using this statistic to compare different models calibrated to different markets.

7 See appendix for more detail – bootstrapped 200 samples assuming modeling error time series is AR(1)
V. Model Monitoring

Due to the difficulty in frequently re-calculating economic capital based on the scenarios generated from the underlying model, the D-statistic can provide ongoing support to the reported economic capital by actively reporting the modeling error associated between the models and market. A hedging policy can be implemented and monitored to enforce capital solvency based on the observed modeling error. Going forward with this methodology, the modeling error can be actively monitored to changing market data, which can be used to ensure adequacy of the hedging policy.

Figure 2 illustrates monitoring the evolution of the modeling error relative to the D-statistic. A negative ratio between current time-weighted modeling error and the D-statistic implies that current modeling error is below the maximum error observed during validation. During the financial crisis, changes in the weekly short-rate dramatically increased; this increased volatility is also captured by the gradual increase
from negative to positive of the modeling error relative to the D-statistic. Although the modeling error of the 2-factor CIR model is consistently less, both models have relatively the same time-weighted modeling error throughout the financial crisis monitoring period. Initially the time-weighted modeling errors of both models were below their respective D-statistics, as the financial crisis approached, the time-weighted average modeling error increased. During the crisis, both model errors broached their respective D-statistic errors of approximately 10% to almost an additional 20%. This implies that during the financial crisis economic capital calculated using these calibrated parameters was subject not only to a maximum of approximately 10% modeling error, but an additional error of almost 20%.

The CIR 1 and 2-factor models in this situation illustrate a gradual increase in modeling error, allowing time for re-evaluation of the solvency hedging policy without having to re-update economic capital calculations. A lower value for the D-statistic corresponds to greater accuracy of the model during validation. Future studies will explore whether there is a tradeoff between model accuracy during validation and the response time during the monitoring period.

VI. Conclusion

This paper presents a simple and concise validation methodology for monitoring the efficacy of Economic Scenario Generating models used to generate multiple scenarios for solvency capital requirements. For a particular point in time, each currency has many interest rates depending on duration. Here the multidimensional model-to-market error of the term structure is collapsed to a concise time-weighted estimate, and the maximum estimate across the validation data is defined as the D-statistic. By applying the closed form analytical bond pricing solutions to the CIR 1 and 2-factor models, this paper illustrates the application of the validation methodology introduced, and uses the 2008 financial crisis period to demonstrate the effectiveness of the D-statistic as a monitoring metric. The D-statistic introduced here is very similar in application with the well-known Kolmogorov-Smirnov Statistic. The methodology first
involved calibrating the model parameters, and then a time-series of the modeling error was generated by
calculating the time-weighted modeling error for each term-structure in the validation data. The maximum
of the time-weighted modeling error across the validation period was defined the D-statistic, and finally
an AR(1) bootstrapping algorithm improved the robustness by generating an empirical distribution to
calculate the standard error for the D-statistic. By introducing this metric, the D-statistic serves as a basis
with which to compare regular monitoring of current model-to-market error. A complete breakdown of
model efficacy can be detected in advance when the modeling error from unexpected market behavior
exceeds the D-statistic threshold. The D-statistic can be used to improve analysis of finding an optimal
set of model parameters, compare between calibration algorithms, and succinctly rank between different
models.

Acknowledgements

Li Xiang Research Grant. Thanks to both Sarah Abigail of University of Pennsylvania and Sarah Rich of
Harvard University for their editorial comments.
Appendix

AR(1) Bootstrapping Algorithm:

The following bootstrap algorithm is taken from Efron and Tibshirani (1993) for determining the standard error of the linear coefficient. This methodology is extended here to also produce an empirical distribution of the D-statistic for determining its standard error.

Let \( t_i = \text{date}_i \) be from the modeling error time-series data \( \{t_1, t_2, ..., t_n\} \) with \( n \) observations.

(1) The model parameters \( \hat{\theta} \) are estimated by least squares calibration. At time \( t_i \), for each bond there is error between the model and market price. For each \( t_i \), let \( y \) be the weighted average model error between model and market prices for all quoted maturities, so that \( y_i \) is the weighted average model error time-series. The Kolmogorov-Smirnov like Statistic is \( \hat{D} = \text{Max}(y_1, y_2, ..., y_n) \).

(2) Define \( z_i = y_i - \mu \) as the centered measurements, then all of the \( z_i \) have expectation \( E[z_i] = 0 \). \( \mu \) is estimated by the observed average \( \bar{y} \).

(3A) Assume \( z_i \) is AR(1) process, \( z_t = \beta z_{t-1} + \epsilon_t \)

(4A) Estimate \( \hat{\beta} \), then calculate \( \hat{\epsilon}_t = z_t - \hat{\beta} z_{t-1} \). Generate empirical error distribution \( F \rightarrow \{\hat{\epsilon}_2, \hat{\epsilon}_3, ..., \hat{\epsilon}_n\} \) where each \( \hat{\epsilon}_t \) has probability \( 1/(n-1) \).

(5A) Bootstrap Algorithm

i. Generate \( \hat{F} \rightarrow \{\hat{\epsilon}_2, \hat{\epsilon}_3, ..., \hat{\epsilon}_n\} \) by sampling with replacement from \( F \)

ii. \( z^*_t = z_t \)

iii. Compute \( z^*_t = \hat{\beta} z^*_{t-1} + \hat{\epsilon}^*_t \), for \( t = 2, 3, ..., n \)

iv. Estimate \( \hat{D}^* = \text{Max}(y^*_1 = z^*_1 - \mu, y^*_2 = z^*_2 - \mu, ..., y^*_n = z^*_n - \mu) \) and \( \hat{\beta}^* \)

v. Repeat 200 times

(6A) Algorithm Generates empirical distribution for \( \hat{D} \) and \( \hat{\beta} \) with 200 samples; the specific results are summarized below:
<table>
<thead>
<tr>
<th>Underlying Model</th>
<th>CIR 1-Factor</th>
<th>CIR 2-Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>( \hat{\beta} )</td>
<td>0.7131</td>
<td>0.6422</td>
</tr>
<tr>
<td>Standard Error (( \hat{\beta} ))</td>
<td>0.1016</td>
<td>0.1076</td>
</tr>
<tr>
<td>( \hat{D} )</td>
<td>10.54%</td>
<td>10.45%</td>
</tr>
<tr>
<td>Standard Error (( \hat{D} ))</td>
<td>1.58%</td>
<td>1.24%</td>
</tr>
</tbody>
</table>
References


CEIOPS’ Advice for Level 2 Implementing Measures on Solvency II: Articles 120 to 126 Tests and Standards for Internal Model Approval, October 2009.


