A Quantitative Metric to Validate Risk Models

William Rearden and Chih-Kai Chang

Presented at the:
2013 Enterprise Risk Management Symposium
April 22-24, 2013

A Quantitative Metric to Validate Risk Models

William Rearden[1] and Chih-Kai Chang[2]

Abstract

The article 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 calibration methods, be used to rank models, and provide a single concise reporting metric for ongoing model monitoring. To illustrate this methodology and ranking of models, the closed-form bond-pricing solutions of the Cox, Ingersoll, Ross (CIR) one- and two-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.

1. Introduction

If we cannot trust doctors when it comes to matters of health,[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-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

[1] William Rearden MA, ASA is Managing Partner, 出來 Global, Canada, william.rearden@gmail.com.

[2] Chih-Kai Chang PhD, FSA, CERA is in the Department of Risk Management and Insurance, Feng Chia University, Taiwan, ckchang@fcu.edu.tw.

[3] See Abigail (2011).

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 CEIOPS' Advice for Level 2 Implementing Measures on Solvency II, in Article 120, risk managers are responsible for ensuring the ongoing appropriateness of the design and operations of the internal model. Furthermore, Article 116 of CEIOP 2009 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 officials 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. Many authors discuss 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 article we introduce a validation methodology that back-tests the calibrated parameters before extending the application of locally calibrated models for macrorisk 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 4 returns a unique D statistic. Ranking the D statistics will aid in identifying 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) (CIR) continuous-time short-rate interest rate model. 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.

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

This article 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.

2.1 CIR One-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

$$dr(t) = k\big(\theta - r(t)\big)dt + \sigma\sqrt{r(t)}dW(t),$$

$$r(0) = r_0,$$

with r_0, k, θ, σ 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 θ 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 one-factor a powerful risk management modeling tool for generating stochastic interest rate scenarios.

2.2 CIR Two-Factor Model

A one-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 two-factor model is justified based on principal component analysis, since two factors can explain more than 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 ρ. For the CIR two-factor model, however, we assume no correlation, $\rho = 0$, since the square root noncentral χ^2 process cannot maintain analytical tractability with nonzero instantaneous correlations.

The CIR two-factor model is then defined as

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

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

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

with instantaneous-correlated sources of randomness, $dW_1 dW_2 = 0$. Now the price at time t, of a unit amount of currency at time $T > t$, can be expressed as a generalization of the one-factor:

$$P(t,T) = \prod_{i=1}^{2} A_i(t,T)\, e^{-\sum_{j=1}^{2} B_j(t,T)x_j(t)},$$

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).

3. Data

The observation period for this article is two years from June 1, 2007, to June 6, 2009. For the purpose of analyzing the period during the financial crisis, the data are split into two sets: validation and monitoring. The validation data are from June 1, 2007, to June 6, 2008, and the monitoring data are from June 13, 2008, to June 5, 2009. The term structure data for this article are the middle of the week LIBOR and swap rates provided by Datastream, where one-, three-, and six-month rates are LIBOR, while years 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 one-month LIBOR rate by the following identity:

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

Table 1

Overnight Rate Summary Statistics and Principal Component Analysis[4]

	Validation Data	Monitoring Data	
Date	June 1, 2007–June 6, 2008	June 13, 2008–June 5, 2009	Difference
Number of observations	54	52	−2.00
Min	2.38%	0.31%	−2.07%
Max	5.81%	4.58%	−1.23%
Average	4.25%	1.54%	−2.71%
Standard deviation	1.17%	1.23%	0.06%
First principal component	97.03%	90.50%	−6.53%
Second principal component	1.99%	7.76%	5.77%
Third principal component	0.83%	0.95%	0.12%

[4] Principal component analysis is based on the following subset of yield rates from each data set: one-, three-, and six-month LIBOR rates, and 1-, 2-, 3-, 5-, 7-, 10-, 20-, and 30-year swap rates.

Principal component analysis and some observations between validation and monitoring periods are as follows:

- 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 first principal component the one-factor model explanatory power decreased by 6.53%, but still explained over 90% of the term structure variation during the monitoring period.
- From the second principal component the tilt variation of the yield curve increased by 5.77%.
- From the third principal component the convexity of the yield curve increased in variation by 0.12%.

4. Model Selection and Validation Methodology

4.1 Step1: Model Selection

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

4.2 Step 2: Calibration

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

$$ObjectiveFunction(\theta) = min \left\| \sum_{t=1/12}^{30} \left(\frac{ModelPrice_t(\theta) - MarketPrice_t}{ModelPrice_t(\theta)} \right) \right\|$$

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 article we consider the midweek term structure data of date June 6, 2008, for calibration; for this date the CIR one- and two-factor model yields the following parameters: $\hat{\theta}^1(k,\vartheta,\sigma,\gamma) = (0.0585, 0.1378\,, 0.2419, -0.0998)$ and $\hat{\theta}^2(k_1,k_2,\vartheta_1,\vartheta_2,\sigma_1,\sigma_2,\gamma_{1,}\gamma_2) = (0.1671, 0.9436, 0.1039, 0.0010, 0.0081, 0.8958, 0.1593, 0.9826)$, 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

	CIR One-Factor	CIR Two-Factor
1 month	0.004%	0.191%
2 year	0.186	1.070
5 year	0.103	0.392
10 year	0.036	0.148
20 year	0.552	0.661
30 year	0.427	0.599
Overall time-weighted average model error[a]	0.26	0.40

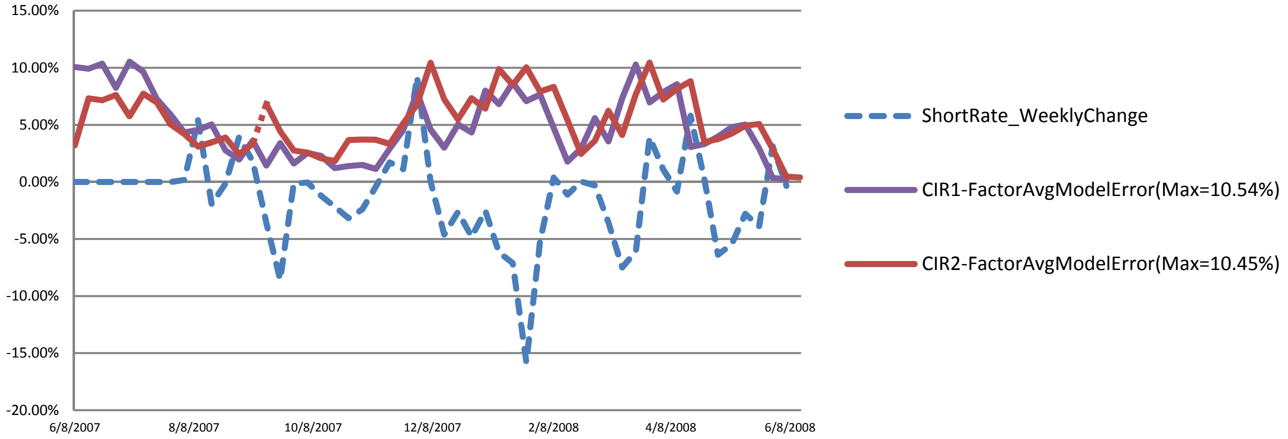
a. $Average = \frac{\sum_{t=1/12}^{30} t\cdot ModelError_t}{\sum_{t=1/12}^{30} t}$

On this particular date, June 6, 2008, the one-factor overall time-weighted average modeling error of 0.26% is less than 0.40% of the two-factor model.

4.3 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 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 snap shot of current model error analysis at calibration and moves toward a more macrolevel 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



4.4 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 two-factor model is too small to justify its complexity over the simpler one-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 recalculating 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[5] between models with monthly, weekly, and daily calibration.

Table 3

Comparing Models

Model	Date of Maximum Model Error	*D*: Maximum Modeling Error	Standard Error (*D*)[a]
CIR one-factor	July 6, 2007	10.54%	1.58%

[5] Caution must be exercised using this statistic to compare different models calibrated to different markets.

CIR two-factor	March 31, 2008	10.45	1.24

a. See Appendix for more detail: bootstrapped 200 samples assuming modeling error time series is AR(1).

5. Model Monitoring

Due to the difficulty in frequently recalculating 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: Monitoring Modeling Error *Relative* to Validation Max Error

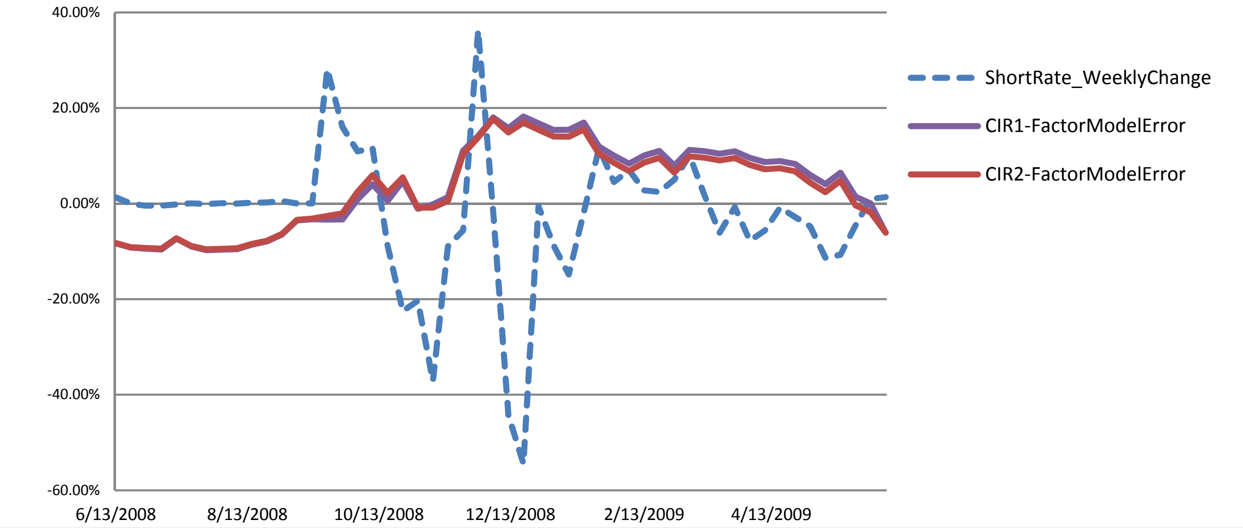


Figure 2 illustrates monitoring the evolution of the modeling error relative to the *D* statistic. A negative ratio between current time-weighted modeling error and the *D* statistic implies that current modeling error is below the maximum error observed during validation. During the financial crisis, 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 two-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, but as the financial crisis approached, the time-weighted average modeling error increased. During the crisis, both model errors breached 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 also an additional error of almost 20%.

The CIR one- and two-factor models in this situation illustrate a gradual increase in modeling error, allowing time for reevaluation of the solvency hedging policy without having to reupdate 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.

6. Conclusion

This article 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 one- and two-factor models, this article 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.

Acknowledgments

We acknowledge a Li Xiang Research Grant and thank Sarah Abigail of the 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 $t_i = date_i$ be from the modeling error time-series data $\{t_1, t_2, \dots, t_n\}$ 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_t is the weighted average model error time series. The Kolmogorov-Smirnov–like statistic is $\hat{D} = Max(y_1, y_2, \dots, y_\mathrm{n})$.

(2) Define $z_\mathrm{t} = y_\mathrm{t} - \mu$ as the centered measurements, then all of the z_t have expectation $\mathrm{E}[z_t] = 0$. μ is estimated by the observed average $\bar{y}$.

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

(4A) Estimate $\hat{\beta}$, then calculate $\hat{\varepsilon_t} = z_t - \hat{\beta} z_{t-1}$. Generate empirical error distribution F → $\{\hat{\varepsilon_2}, \hat{\varepsilon_3}, \dots, \hat{\varepsilon_n}\}$ where each $\hat{\varepsilon_t}$ has probability 1/(*n*−1).

(5A) Bootstrap Algorithm

i. Generate $\hat{\mathrm{F}} \rightarrow \{\varepsilon_2^*, \varepsilon_3^*, \dots, \varepsilon_n^*\}$ by sampling with replacement from F

ii. $z_1^* = z_1$

iii. Compute $z_t^* = \hat{\beta}\, z_{t-1}^* + \varepsilon_t^*$, for *t* = 2, 3, … , *n*

iv. Estimate $\hat{D}^* = \mathrm{Max}(y_1^* = z_1^* - \mu, y_2^* = z_2^* - \mu, \dots, \mathrm{y}_n^* = z_n^* - u)$ 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:

Underlying Model	CIR One-Factor	CIR Two-Factor
Samples	200	200
$\hat{\beta}$	0.7131	0.6422
Standard error ($\hat{\beta}$)	0.1016	0.1076
$\widehat{D}$	10.54%	10.45%
Standard error ($\widehat{D}$)	1.58%	1.24%

References

Abigail, S. 2011. The Obstacle of Therapeutic Privilege in Healthcare Mediation. *American Journal of Mediation,* 5th ed., pg. 1–8.

Brigo, D., and F. Mercurio. 2006. *Interest Rate Models—Theory and Practice with Smile, Inflation and Credit.* New York: Springer.

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

Cox, J. C., J. E. Ingersoll, and S. A. Ross. 1985. A Theory of the Term Structure of Interest Rates. *Econometrica* 53: 385–407.

Efron, B., and R. Tibshirani. 1993. *An Introduction to the Bootstrap.* New York: Chapman and Hall/CRC.