2025

FRM®

EXAM PART II

Market Risk Measurement and Management

GARP®

FRM® | Financial Risk Manager

GARP®
FRM® | Financial Risk Manager

2025

FRM®

EXAM PART II

Market Risk Measurement and Management

Pearson

Excerpts taken from:

Options, Futures, and Other Derivatives, Tenth Edition by John C. Hull
Copyright © 2017, 2015, 2012, 2009, 2006, 2003, 2000 by Pearson Education, Inc.
New York, New York 10013

Grateful acknowledgment is made to the following sources for permission to reprint material copyrighted or controlled by them:

"Estimating Market Risk Measures," "Non-Parametric Approaches," and "Parametric Approaches (III): Extreme Value" by Kevin Dowd, reprinted from *Measuring Market Risk*, Second Edition (2005), by permission of John Wiley & Sons, Inc. All rights reserved. Used under license from John Wiley & Sons, Inc.

"Back Testing VaR" and "VaR Mapping," by Philippe Jorion, reprinted from *Value at Risk: The New Benchmark for Managing Financial Risk*, Third Edition (2007), by permission of The McGraw Hill Companies.

"Messages from the Academic Literature on Risk Measurement for the Trading Book," Working Paper No. 19 January 2011, reprinted by permission of the Basel Committee on Banking Supervision.

"Some Correlation Basics: Definitions, Applications, and Terminology," "Empirical Properties of Correlation: How Do Correlations Behave in the Real World?," and "Financial Correlation Modeling—Bottom Up Approaches," by Gunter Meissner, reprinted from *Correlation Risk Modeling and Management*, Second Edition (2019), by permission of Risk Books/InfoPro Digital Services, Ltd.

"Empirical Approaches to Risk Metrics and Hedges," "The Science of Term Structure Models," "The Evolution of Short Rates and the Shape of the Term Structure," "The Art of Term Structure Models: Drift," and "The Art of Term Structure Models: Volatility and Distribution," by Bruce Tuckman and Angel Serrat, reprinted from *Fixed Income Securities: Tools for Today's Markets*, Third Edition (2012), by permission of John Wiley & Sons, Inc. All≈rights reserved. Used under license from John Wiley & Sons, Inc.

"Validating Bank Holding Companies' Value-at-Risk Models for Market Risk" from *Validation of Risk Management Models for Financial Institutions*, edited by David Lynch, Iftekhar Hasan, Akhtar Siddique. Used by permission of Cambridge University Press.

"Beyond Exceedance-Based Back testing of Value-at-Risk Models: Methods for Back testing the Entire Forecasting Distribution Using Probability Integral Transform" from *Validation of Risk Management Models for Financial Institutions*, edited by David Lynch, Iftekhar Hasan, Akhtar Siddique. Used by permission of Cambridge University Press.

"Fundamental Review of the Trading Book," by John C. Hull, reprinted from *Risk Management and Financial Institutions*, Fifth Edition (2018), by permission of John Wiley & Sons, Inc. All rights reserved. Used under license from John Wiley & Sons, Inc.

Learning Objectives provided by the Global Association of Risk Professionals.

Pearson Education, Inc., 221 River Street, Hoboken, NJ 07030
A Pearson Education Company
www.pearsoned.com

Printed in the United States of America

00040345-00000004 / A103001322219
EEB/KC

Pearson

ISBN 10: 0-13-538963-1
ISBN 13: 978-0-13-538963-8

Contents

PREFACE

On behalf of GARP's Board of Trustees, the FRM advisory committee, and GARP's FRM professional certification program staff. I want to thank you for your interest in and support of the FRM program.

The program's first offering in 1997 saw just over 100 candidates sit for the exam. During the past 27 years, hundreds of thousands of professionals have studied for and taken the FRM exam, with it now being the world's leading financial certification program.

The dynamic nature of the FRM program's curriculum means that it regularly and quickly responds to changes in the global financial marketplace. This ensures that its content and reach always address the risks and challenges of a fast-changing, complex, and globally connected financial system.

For example, for 2025, after much discussion and consideration, the FRM advisory committee made material changes to the program's 2025 market risk measurement and management content. The result is that about half of the subject readings in Market Risk Measurement and Management were updated.

But maintaining a current and highly relevant curriculum is not the sole focus of GARP's professional staff. GARP has focused considerable time and resources during the past year developing tools to assist a candidate in his or her exam program preparation. In addition to providing current content, a primary objective of ours is to ensure as much as possible that a candidate is making the best use of his or her valuable time in preparing for the exam.

In this regard, GARP offers FRM Part I candidates an electronic platform called GARP Learning. GARP Learning is a streamlined digital learning program that can be accessed via a mobile phone, tablet, or desktop computer. GARP Learning allows an FRM candidate to engage meaningfully in a self-directed fashion with the full FRM Part I curriculum. It provides the ability to monitor performance, identify strengths and weaknesses, and assists in creating a personalized study plan.

Supplemental to the support offered by the learning platform, candidates can also utilize end-of-chapter questions to test their understanding of the chapter's content immediately; and, importantly, take a full-length FRM Part I Practice Exam to gain familiarity with how topics are tested and how to pace oneself on the exam to ensure completion in the allotted time.

As you can readily see, we are committed to ensuring the FRM program retains its global reputation as being of the highest quality, and covering the concepts, issues, and challenges that financial risk management professionals must know, and in many cases master.

As always, we wish you the very best as you study for the FRM exams, and much success in your career as a risk management professional.

Yours truly,

Richard Apostolik

President & CEO

FRM® COMMITTEE

Chairperson

Nick Strange, FCA (Chair)
Senior Technical Advisor, Operational Risk & Resilience,
Prudential Regulation Authority, Bank of England

Members

Richard Apostolik
President and CEO, GARP

Richard Brandt
Managing Director, Operational Risk Management, Citigroup

Julian Chen, FRM
SVP, FRM Program Manager, GARP

Chris Donohue
Managing Director, GARP Benchmarking Initiative, GARP

Donald Edgar, FRM
Managing Director, Risk & Quantitative Analysis, BlackRock

Hervé Geny
Former Group Head of Internal Audit, London Stock Exchange Group

Aparna Gupta, Ph.D.
Professor of Quantitative Finance
Associate Dean, Academic Affairs
A.W. Lawrence Professional Excellence Fellow
Co-Director and Site Director, NSF IUCRC CRAFT,
Lally School of Management
Rensselaer Polytechnic Institute

John Hull, Ph.D.
Senior Advisor
Maple Financial Professor of Derivatives and Risk Management,
Joseph L. Rotman School of Management, University of Toronto

Keith Isaac, FRM
VP, Capital Markets Risk Management, TD Bank Group

William May
Managing Director, Global Head of Certifications and Educational Programs, GARP

Attilio Meucci, Ph.D., CFA
Founder, ARPM

Victor Ng, Ph.D.
Chairman, Audit and Risk Committee
Former Managing Director, Head of Risk Architecture,
Goldman Sachs

Matthew Pritsker, Ph.D.
Senior Financial Economist and Policy Advisor/Supervision,
Regulation, and Credit, Federal Reserve Bank of Boston

Samantha C. Roberts, Ph.D., FRM, SCR
Instructor and Consultant, Risk Modeling and Analytics

Til Schuermann, Ph.D.
Partner, Oliver Wyman

Evan Sekeris, Ph.D.
Head of Non-Financial Risk, MUFG

Sverrir Þorvaldsson, Ph.D., FRM
Senior Quant, SEB

1

Estimating Market Risk Measures

An Introduction and Overview

Learning Objectives

After completing this reading, you should be able to:

- Estimate VaR using a historical simulation approach.
- Estimate VaR using a parametric approach for both normal and lognormal return distributions.
- Estimate the expected shortfall given profit and loss (P&L) or return data.
- Estimate risk measures by estimating quantiles.
- Evaluate estimators of risk measures by estimating their standard errors.
- Interpret quantile-quantile (QQ) plots to identify the characteristics of a distribution.

Excerpt is Chapter 3 of Measuring Market Risk, *Second Edition, by Kevin Dowd.*

This chapter provides a brief introduction and overview of the main issues in market risk measurement. Our main concerns are:

- *Preliminary data issues*: How to deal with data in profit/loss form, rate-of-return form, and so on.
- *Basic methods of VaR estimation*: How to estimate simple VaRs, and how VaR estimation depends on assumptions about data distributions.
- How to estimate coherent risk measures.
- How to gauge the precision of our risk measure estimators by estimating their standard errors.
- *Overview*: An overview of the different approaches to market risk measurement, and of how they fit together.

We begin with the data issues.

1.1 DATA

Profit/Loss Data

Our data can come in various forms. Perhaps the simplest is in terms of profit/loss (or P/L). The P/L generated by an asset (or portfolio) over the period t, P/L_t, can be defined as the value of the asset (or portfolio) at the end of t plus any interim payments D_t minus the asset value at the end of $t - 1$:

$$P/L_t = P_t + D_t - P_{t-1} \tag{1.1}$$

If data are in P/L form, positive values indicate profits and negative values indicate losses.

If we wish to be strictly correct, we should evaluate all payments from the same point of time (i.e., we should take account of the time value of money). We can do so in one of two ways. The first way is to take the present value of P/L_t evaluated at the end of the previous period, $t - 1$:

$$\textit{Present Value}\,(P/L_t) = \frac{(P_t + D_t)}{(1 + d)} - P_{t-1} \tag{1.2}$$

where d is the discount rate and we assume for convenience that D_t is paid at the end of t. The alternative is to take the forward value of P/L_t evaluated at the end of period t:

$$\textit{Forward Value}\,(P/L_t) = P_t + D_t - (1 + d)P_{t-1} \tag{1.3}$$

which involves compounding P_{t-1} by d. The differences between these values depend on the discount rate d, and will be small if the periods themselves are short. We will ignore these differences to simplify the discussion, but they can make a difference in practice when dealing with longer periods.

Loss/Profit Data

When estimating VaR and ES, it is sometimes more convenient to deal with data in loss/profit (L/P) form. L/P data are a simple transformation of P/L data:

$$L/P_t = -P/L_t \tag{1.4}$$

L/P observations assign a positive value to losses and a negative value to profits, and we will call these L/P data 'losses' for short. Dealing with losses is sometimes a little more convenient for risk measurement purposes because the risk measures are themselves denominated in loss terms.

Arithmetic Return Data

Data can also come in the form of arithmetic (or simple) returns. The arithmetic return r_t is defined as:

$$r_t = \frac{P_t + D_t - P_{t-1}}{P_{t-1}} = \frac{P_t + D_t}{P_{t-1}} - 1 \tag{1.5}$$

which is the same as the P/L over period t divided by the value of the asset at the end of $t - 1$.

In using arithmetic returns, we implicitly assume that the interim payment D_t does not earn any return of its own. However, this assumption will seldom be appropriate over long periods because interim income is usually reinvested. Hence, arithmetic returns should not be used when we are concerned with long horizons.

Geometric Return Data

Returns can also be expressed in geometric (or compound) form. The geometric return R_t is

$$R_t = \ln\left(\frac{P_t + D_t}{P_{t-1}}\right) \tag{1.6}$$

The geometric return implicitly assumes that interim payments are continuously reinvested. The geometric return is often more economically meaningful than the arithmetic return, because it ensures that the asset price (or portfolio value) can never become negative regardless of how negative the returns might be. With arithmetic returns, on the other hand, a very low realized return—or a high loss—implies that the asset value P_t can become negative, and a negative asset price seldom makes economic sense.[1]

The geometric return is also more convenient. For example, if we are dealing with foreign currency positions, geometric returns will give us results that are independent of the reference

[1] This is mainly a point of principle rather than practice. In practice, any distribution we fit to returns is only likely to be an approximation, and many distributions are ill-suited to extreme returns anyway.

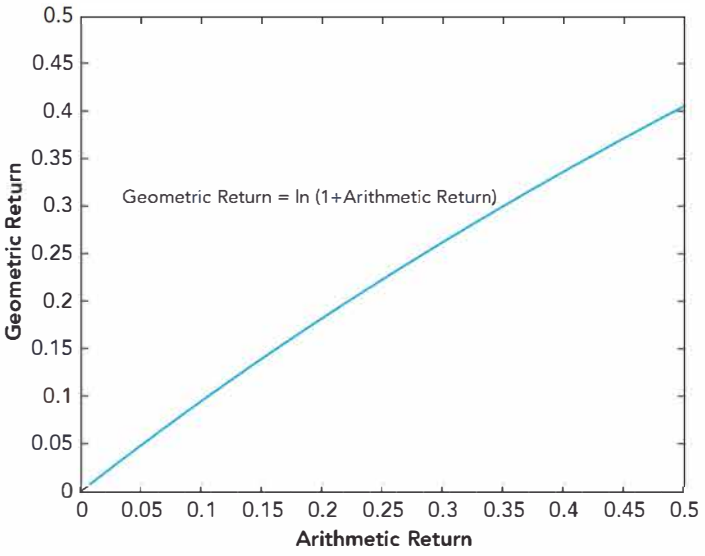


Figure 1.1 Geometric and arithmetic returns.

currency. Similarly, if we are dealing with multiple periods, the geometric return over those periods is the sum of the one-period geometric returns. Arithmetic returns have neither of these convenient properties.

The relationship of the two types of return can be seen by rewriting Equation (1.6) (using a Taylor's series expansion for the natural log) as:

$$R_t = \ln\left(\frac{P_t + D_t}{P_{t-1}}\right) = \ln(1 + r_t) = r_t - \frac{1}{2}r_t^2 + \frac{1}{3}r_t^3 - \cdots \quad (1.7)$$

from which we can see that $R_t \approx r_t$ provided that returns are 'small'. This conclusion is illustrated by Figure 1.1, which plots the geometric return R_t against its arithmetic counterpart r_t. The difference between the two returns is negligible when both returns are small, but the difference grows as the returns get bigger—which is to be expected, as the geometric return is a log function of the arithmetic return. Since we would expect returns to be low over short periods and higher over longer periods, the difference between the two types of return is negligible over short periods but potentially substantial over longer ones. And since the geometric return takes account of earnings on interim income, and the arithmetic return does not, we should always use the geometric return if we are dealing with returns over longer periods.

Example 1.1 Arithmetic and Geometric Returns

If arithmetic returns r_t over some period are 0.05, Equation (1.7) tells us that the corresponding geometric returns are $R_t = \ln(1 + r_t) = \ln(1.05) = 0.0488$. Similarly, if geometric returns R_t are 0.05, Equation (1.7) implies that arithmetic returns are $1 + r_t = \exp(R_t) \Rightarrow r_t = \exp(R_t) - 1 = \exp(0.05) - 1 = 0.0513$. In both cases the arithmetic return is close to, but a little higher than, the geometric return—and this makes intuitive sense when one considers that the geometric return compounds at a faster rate.

1.2 ESTIMATING HISTORICAL SIMULATION VaR

The simplest way to estimate VaR is by means of historical simulation (HS). The HS approach estimates VaR by means of ordered loss observations.

Suppose we have 1000 loss observations and are interested in the VaR at the 95% confidence level. Since the confidence level implies a 5% tail, we know that there are 50 observations in the tail, and we can take the VaR to be the 51st highest loss observation.[2]

We can estimate the VaR on a spreadsheet by ordering our data and reading off the 51st largest observation from the spreadsheet. We can also estimate it more directly by using the 'Large' command in Excel, which gives us the *k*th largest value in an array. Thus, if our data are an array called 'Loss_data', then our VaR is given by the Excel command 'Large(Loss_data,51)'. If we are using MATLAB, we first order the loss/profit data using the 'Sort()' command (i.e., by typing 'Loss_data = Sort(Loss_data)'); and then derive the VaR by typing in 'Loss_data(51)' at the command line.

More generally, if we have n observations, and our confidence level is α, we would want the $(1 - \alpha)$. $n + 1$th highest observation, and we would use the commands 'Large(Loss_data,(1 − *alpha*)*n + 1)' using Excel, or 'Loss_data((1 − *alpha*)*n + 1)' using MATLAB, provided in the latter case that our 'Loss_data' array is already sorted into ordered observations.[3]

[2] In theory, the VaR is the quantile that demarcates the tail region from the non-tail region, where the size of the tail is determined by the confidence level, but with finite samples there is a certain level of arbitrariness in how the ordered observations relate to the VaR itself—that is, do we take the VaR to be the 50th observation, the 51st observation, or some combination of them? However, this is just an issue of approximation, and taking the VaR to be the 51st highest observation is not unreasonable.

[3] We can also estimate HS VaR using percentile functions such as the 'Percentile' function in Excel or the 'prctile' function in MATLAB. However, such functions are less transparent (i.e., it is not obvious to the reader how the percentiles are calculated), and the Excel percentile function can be unreliable.

An example of an HS VaR is given in Figure 1.2. This figure shows the histogram of 1000 hypothetical loss observations and the 95%VaR. The figure is generated using the 'hsvarfigure' command in the MMR Toolbox. The VaR is 1.704 and separates the top 5% from the bottom 95% of loss observations.

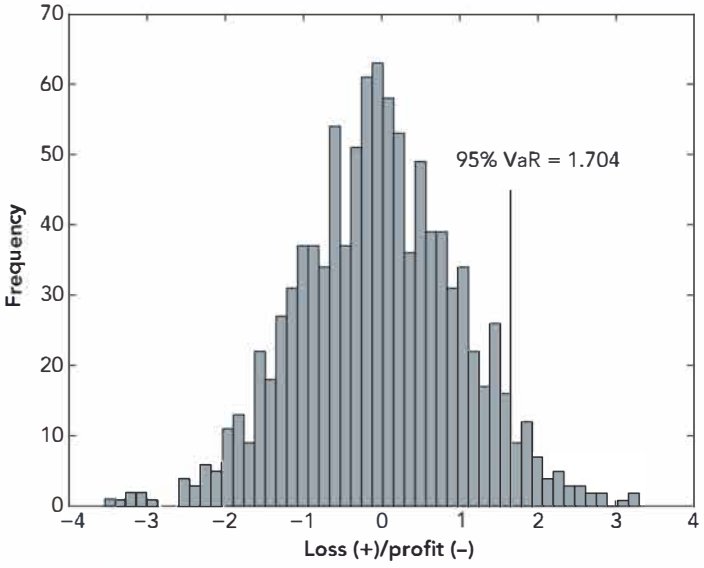


Figure 1.2 **Historical simulation VaR.**

Note: Based on 1000 random numbers drawn from a standard normal L/P distribution, and estimated with 'hsvarfigure' function.

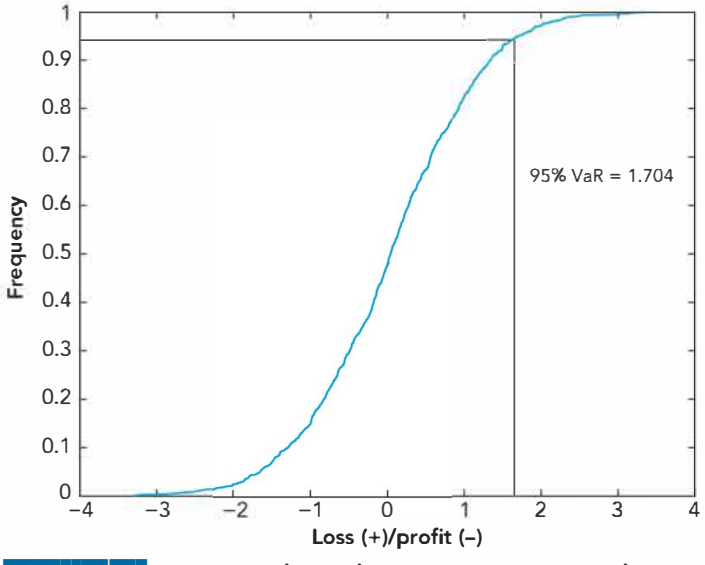


Figure 1.3 **Historical simulation via an empirical cumulative frequency function.**

Note: Based on the same data as Figure 1.2.

In practice, it is often helpful to obtain HS VaR estimates from a cumulative histogram, or empirical cumulative frequency function. This is a plot of the ordered loss observations against their empirical cumulative frequency (e.g., so if there are n observations in total, the empirical cumulative frequency of the ith such ordered observation is i/n). The empirical cumulative frequency function of our earlier data set is shown in Figure 1.3. The empirical frequency function makes it very easy to obtain the VaR: we simply move up the cumulative frequency axis to where the cumulative frequency equals our confidence level, draw a horizontal line along to the curve, and then draw a vertical line down to the x-axis, which gives us our VaR.

1.3 ESTIMATING PARAMETRIC VaR

We can also estimate VaR using parametric approaches, the distinguishing feature of which is that they require us to explicitly specify the statistical distribution from which our data observations are drawn. We can also think of parametric approaches as fitting curves through the data and then reading off the VaR from the fitted curve.

In making use of a parametric approach, we therefore need to take account of both the statistical distribution and the type of data to which it applies.

Estimating VaR with Normally Distributed Profits/Losses

Suppose that we wish to estimate VaR under the assumption that P/L is normally distributed. In this case our VaR at the confidence level α is:

$$\alpha VaR = -\mu_{P/L} + \sigma_{P/L} z_\alpha \tag{1.8}$$

where z_α is the standard normal variate corresponding to α, and $\mu_{P/L}$ and $\sigma_{P/L}$ are the mean and standard deviation of P/L. Thus, z_α is the value of the standard normal variate such that α of the probability density mass lies to its left, and $1 - \alpha$ of the probability density mass lies to its right. For example, if our confidence level is 95%, $z_\alpha = z_{0.95}$ will be 1.645.

In practice, $\mu_{P/L}$ and $\sigma_{P/L}$ would be unknown, and we would have to estimate VaR based on estimates of these parameters. Our VaR estimate, αVaR^e, would then be:

$$\alpha VaR^e = -m_{P/L} + s_{P/L} z_\alpha \tag{1.9}$$

where $m_{P/L}$ and $s_{P/L}$ are estimates of the mean and standard deviation of P/L.

Figure 1.4 shows the 95% VaR for a normally distributed P/L with mean 0 and standard deviation 1. Since the data are in P/L form, the VaR is indicated by the negative of the cut off point between the lower 5% and the upper 95% of P/L observations. The actual VaR is the negative of −1.645, and is therefore 1.645.

If we are working with normally distributed L/P data, then $\mu_{L/P} = -\mu_{P/L}$ and $\sigma_{L/P} = \sigma_{P/L}$, and it immediately follows that:

$$\alpha VaR = \mu_{L/P} + \sigma_{L/P} z_\alpha \tag{1.10a}$$

$$\alpha VaR^e = m_{L/P} + s_{L/P} z_\alpha \tag{1.10b}$$

Figure 1.5 illustrates the corresponding VaR. This figure gives the same information as Figure 1.4, but is a little more straightforward to interpret because the VaR is defined in units of losses (or 'lost money') rather than P/L. In this case, the VaR is given by the point on the *x*-axis that cuts off the top 5% of the pdf mass from the bottom 95% of pdf mass. If we prefer to work with the cumulative density function, the VaR is the *x*-value that corresponds to a cdf value of 95%. Either way, the VaR is again 1.645, as we would (hopefully) expect.

Example 1.2 VaR with Normal P/L

If P/L over some period is normally distributed with mean 10 and standard deviation 20, then (by Equation (1.8)) the 95% VaR is $-10 + 20z_{0.95} = -10 + 20 \times 1.645 = 22.9$. The corresponding 99% VaR is $-10 + 20z_{0.99} = -10 + 20 \times 2.326 = 36.52$.

Estimating VaR with Normally Distributed Arithmetic Returns

We can also estimate VaR making assumptions about returns rather than P/L. Suppose then that we assume that arithmetic returns are normally distributed with mean μ_r and standard deviation σ_r. To derive the VaR, we begin by obtaining the critical value of r_t, r^*, such that the probability that r_t exceeds r^* is equal to our confidence level α. r^* is therefore:

$$r^* = \mu_r - \sigma_r z_\alpha \tag{1.11}$$

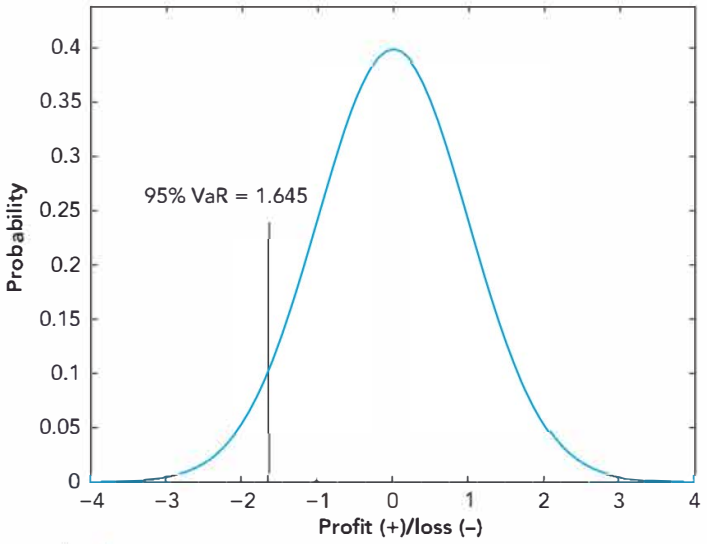


Figure 1.4 VaR with standard normally distributed profit/loss data.

Note: Obtained from Equation (1.9) with $\mu_{P/L} = 0$ and $\sigma_{P/L} = 1$. Estimated with the 'normalvarfigure' function.

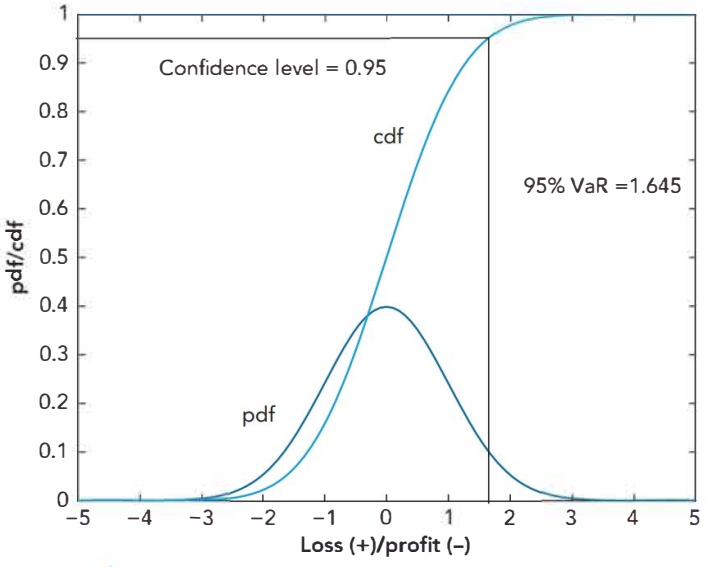


Figure 1.5 VaR with normally distributed loss/profit data.

Note: Obtained from Equation (1.10a) with $\mu_{L/P} = 0$ and $\sigma_{L/P} = 1$.

Since the actual return r_t is the loss/profit divided by the earlier asset value, P_{t-1}, it follows that:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} = -\frac{Loss_t}{P_{t-1}} \tag{1.12}$$

Substituting r^* for r_t then gives us the relationship between r^* and the VaR:

$$r_t^* = \frac{P_t^* - P_{t-1}}{P_{t-1}} = \frac{VaR}{P_{t-1}} \tag{1.13}$$

Substituting Equation (1.11) into Equation (1.13) and rearranging then gives us the VaR itself:

$$\alpha VaR = -(\mu_r - \sigma_r z_\alpha)P_{t-1} \tag{1.14}$$

Equation (1.14) will give us equivalent answers to our earlier VaR equations. For example, if we set $\alpha = 0.95$, $\mu_r = 0$, $\sigma_r = 1$ and $P_{t-1} = 1$, which correspond to our earlier illustrative P/L and L/P parameter assumptions, αVaR is 1.645: the three approaches give the same results, because all three sets of underlying assumptions are equivalent.

Example 1.3 VaR with Normally Distributed Arithmetic Returns

Suppose arithmetic returns r_t over some period are distributed as normal with mean 0.1 and standard deviation 0.25, and we have a portfolio currently worth 1. Then (by Equation (1.14)) the 95% VaR is $-0.1 + 0.25 \times 1.645 = 0.331$, and the 99% VaR is $-0.1 + 0.25 \times 2.326 = 0.482$.

Estimating Lognormal VaR

Each of the previous approaches assigns a positive probability of the asset value, P_t, becoming negative, but we can avoid this drawback by working with geometric returns. Now assume that geometric returns are normally distributed with mean μ_R and standard deviation σ_R. If D_t is zero or reinvested continually in the asset itself (e.g., as with profits reinvested in a mutual fund), this assumption implies that the natural logarithm of P_t is normally distributed, or that P_t itself is lognormally distributed. The lognormal distribution is explained in Box 1.1, and a lognormal asset price is shown in Figure 1.6: observe that the price is always non-negative, and its distribution is skewed with a long right-hand tail.

Since the VaR is a loss, and since the loss is the difference between P_t (which is random) and P_{t-1} (which we can take here as given), then the VaR itself has the same distribution as P_t. Normally distributed geometric returns imply that the VaR is lognormally distributed.

If we proceed as we did earlier with the arithmetic return, we begin by deriving the critical value of R, R^*, such that the probability that $R > R^*$ is α:

$$R^* = \mu_R - \sigma_R z_\alpha \tag{1.15}$$

BOX 1.1 THE LOGNORMAL DISTRIBUTION

A random variate X is said to be lognormally distributed if the natural log of X is normally distributed. The lognormal distribution can be specified in terms of the mean and standard deviation of ln X. Call these parameters μ and σ. The lognormal is often also represented in terms of m and σ, where m is the median of x, and $m = \exp(\mu)$.

The pdf of X can be written:

$$\phi(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2\right\}$$

for $x > 0$. Thus, the lognormal pdf is only defined for positive values of x and is skewed to the right as in Figure 1.6.

Let $\omega = \exp(\sigma^2)$ for convenience. The mean and variance of the lognormal can be written as:

$$\text{mean} = m\exp(\sigma^2/2) \quad \text{and} \quad \text{variance} = m^2\omega(\omega - 1)$$

Turning to higher moments, the skewness of the lognormal is

$$skewness = (\omega + 2)(\omega - 1)^{1/2}$$

and is always positive, which confirms the lognormal has a long right-hand tail. The kurtosis of the lognormal is

$$kurtosis = \omega^4 + 2\omega^3 + 3\omega^2 - 3$$

and therefore varies from a minimum of (just over) 3 to a potentially large value depending on the value of s.

We then use the definition of the geometric return to unravel the critical value of P^* (i.e., the value of P_t corresponding to a loss equal to our VaR), and thence infer our VaR:

$$\begin{aligned} R^* &= \ln(P^*/P_{t-1}) = \ln P^* - \ln P_{t-1} \\ &\Rightarrow \ln P^* = R^* + \ln P_{t-1} \\ &\Rightarrow P^* = P_{t-1}\exp[R^*] = P_{t-1}\exp[\mu_R - \sigma_R z_\alpha] \\ &\Rightarrow \alpha VaR = P_{t-1} - P^* = P_{t-1}(1 - \exp[\mu_R - \sigma_R z_\alpha]) \end{aligned} \tag{1.16}$$

This gives us the lognormal VaR, which is consistent with normally distributed geometric returns.

The lognormal VaR is illustrated in Figure 1.7, based on the standardised (but typically unrealistic) assumptions that $\mu_R = 0$, $\sigma_R = 1$, and $P_{t-1} = 1$. In this case, the VaR at the 95% confidence level is 0.807. The figure also shows that the distribution of L/P is a reflection of the distribution of P_t shown earlier in Figure 1.6.

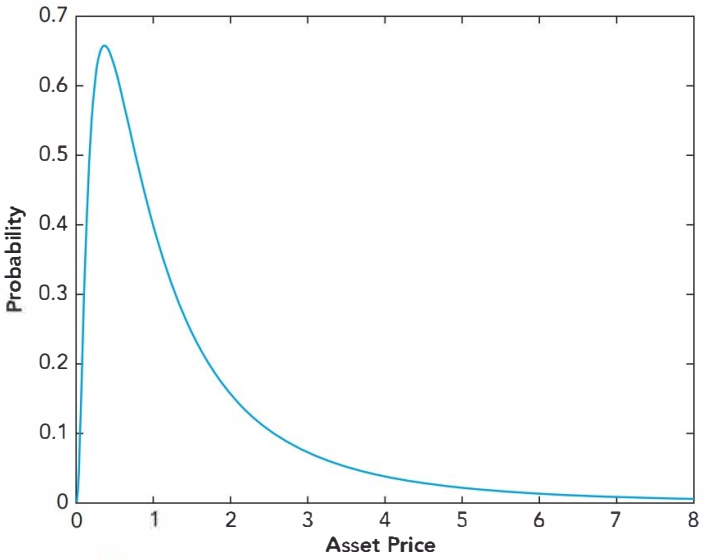


Figure 1.6 **A lognormally distributed asset price.**

Note: Estimated using the 'lognpdf' function in the Statistics Toolbox.

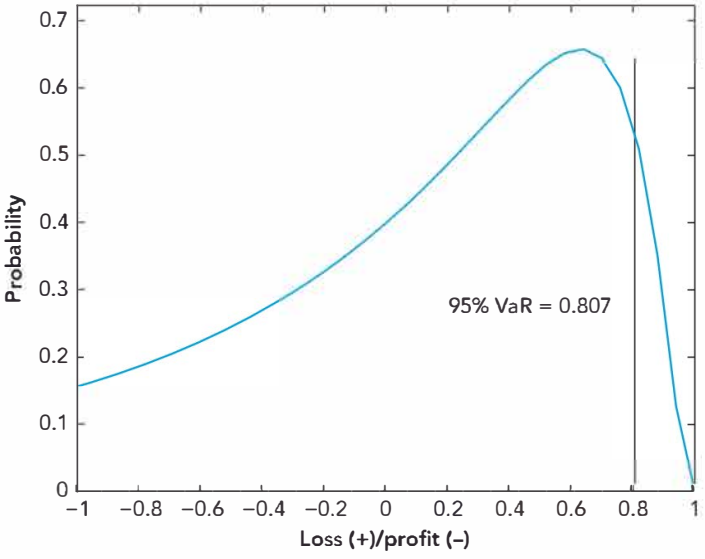


Figure 1.7 **Lognormal VaR.**

Note: Estimated assuming the mean and standard deviation of geometric returns are 0 and 1, and for an initial investment of 1. The figure is produced using the 'lognormalvarfigure' function.

Example 1.4 Lognormal VaR

Suppose that geometric returns R_t over some period are distributed as normal with mean 0.05, standard deviation 0.20, and we have a portfolio currently worth 1. Then (by Equation (1.16)) the 95% VaR is $1 - \exp(0.05 - 0.20 \times 1.645) = 0.244$. The corresponding 99% VaR is $1 - \exp(0.05 - 0.20 \times 2.326) = 0.340$. Observe that these VaRs are quite close to those obtained in Example 1.3, where the arithmetic return parameters were the same as the geometric return parameters assumed here.

Example 1.5 Lognormal VaR vs Normal VaR

Suppose that we make the empirically not too unrealistic assumptions that the mean and volatility of annualised returns are 0.10 and 0.40. We are interested in the 95% VaR at the 1-day holding period for a portfolio worth USD 1. Assuming 250 trading days to a year, the daily return has a mean $0.1/250 = 0.00040$ and standard deviation $0.40/\sqrt{250} = 0.0253$. The normal 95% VaR is $-0.0004 + 0.0253 \times 1.645 = 0.0412$. If we assume a lognormal, then the 95% VaR is $1 - \exp(0.0004 - 0.0253 \times 1.645) = 0.0404$. The normal VaR is 4.12% and the lognormal VaR is 4.04% of the value of the portfolio. This illustrates that normal and lognormal VaRs are much the same if we are dealing with short holding periods and realistic return parameters.

1.4 ESTIMATING COHERENT RISK MEASURES

Estimating Expected Shortfall

We turn now to the estimation of coherent risk measures, and the easiest of these to estimate is the expected shortfall (ES). The ES is the probability-weighted average of tail losses, and a normal ES is illustrated in Figure 1.8. In this case, the 95% ES is 2.063, corresponding to our earlier normal 95% VaR of 1.645.

The fact that the ES is a probability-weighted average of tail losses suggests that we can estimate ES as an average of 'tail VaRs'.[4] The easiest way to implement this approach is to slice the tail into a large number n of slices, each of which has the same probability mass, estimate the VaR associated with each slice, and take the ES as the average of these VaRs.

To illustrate the method, suppose we wish to estimate a 95% ES on the assumption that losses are normally distributed with mean 0 and standard deviation 1. In practice, we would use a

[4] The obvious alternative is to seek a 'closed-form' solution, which we could use to estimate the ES, but ES formulas seem to be known only for a limited number of parametric distributions (e.g., elliptical, including normal, and generalised Pareto distributions), whereas the 'average-tail-VaR' method is easy to implement and can be applied to any 'well-behaved' ESs that we might encounter, parametric or otherwise.

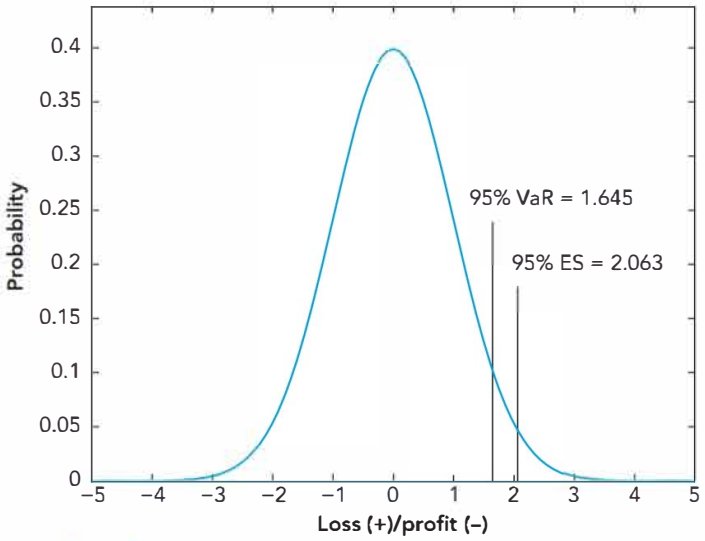


Figure 1.8 Normal VaR and ES.

Note: Estimated with the mean and standard deviation of P/L equal to 0 and 1 respectively, using the 'normalesfigure' function.

high value of n and carry out the calculations on a spreadsheet or using appropriate software. However, to show the procedure manually, let us work with a very small n value of 10. This value gives us 9 (i.e., $n - 1$) tail VaRs, or VaRs at confidence levels in excess of 95%. These VaRs are shown in Table 1.1, and vary from 1.6954 (for the 95.5% VaR) to 2.5758 (for the 99.5% VaR). Our estimated ES is the average of these VaRs, which is 2.0250.

Table 1.1 Estimating ES as a Weighted Average of Tail VaRs

Confidence Level	Tail VaR
95.5%	1.6954
96.0%	1.7507
96.5%	1.8119
97.0%	1.8808
97.5%	1.9600
98.0%	2.0537
98.5%	2.1701
99.0%	2.3263
99.5%	2.5738
Average of tail VaRs	2.0250

Note: VaRs estimated assuming the mean and standard deviation of losses are 0 and 1, using the 'normalvar' function in the MMR Toolbox.

Of course, in using this method for practical purposes, we would want a value of n large enough to give accurate results. To give some idea of what this might be, Table 1.2 reports some alternative ES estimates obtained using this procedure with varying values of n. These results show that the estimated ES rises with n, and gradually converges to the true value of 2.063. These results also show that our ES estimation procedure seems to be reasonably accurate even for quite small values of n. Any decent computer should therefore be able to produce accurate ES estimates quickly in real time.

Estimating Coherent Risk Measures

Other coherent risk measures can be estimated using modifications of this 'average VaR' method. Recall that a coherent risk measure is a weighted average of the quantiles (denoted by q_p) of our loss distribution:

$$M_\phi = \int_0^1 \phi(p) q_p dp \tag{1.17}$$

where the weighting function or risk-aversion function $\phi(p)$ is specified by the user. The ES gives all tail-loss quantiles an equal weight, and other quantiles a weight of 0. Thus the ES is a special case of M_ϕ obtained by setting $\phi(p)$ to the following:

$$\phi(p) = \begin{cases} 0 & \text{if } p < \alpha \\ 1/(1-\alpha) & \text{if } p \geq \alpha \end{cases} \tag{1.18}$$

Table 1.2 ES Estimates as a Function of the Number of Tail Slices

Number of Tail Slices (n)	ES
10	2.0250
25	2.0433
50	2.0513
100	2.0562
250	2.0597
500	2.0610
1000	2.0618
2500	2.0623
5000	2.0625
10 000	2.0626
True value	2.0630

Note: VaRs estimated assuming the mean and standard deviation of losses are 0 and 1.

The more general coherent risk measure, M_ϕ, involves a potentially more sophisticated weighting function $\phi(p)$. We can therefore estimate any of these measures by replacing the equal weights in the 'average VaR' algorithm with the $\phi(p)$ weights appropriate to the risk measure being estimated.

To show how this might be done, suppose we have the exponential weighting function:

$$\phi_\gamma(p) = \frac{e^{-(1-p)/\gamma}}{\gamma(1 - e^{-1/\gamma})} \tag{1.19}$$

and we believe that we can represent the degree of our risk-aversion by setting $\gamma = 0.05$. To illustrate the procedure manually, we continue to assume that losses are standard normally distributed and we set $n = 10$ (i.e., we divide the complete losses density function into 10 equal-probability slices). With $n = 10$, we have $n - 1 = 9$ (i.e., $n - 1$) loss quantiles or VaRs spanning confidence levels from 0.1 to 0.90. These VaRs are shown in the second column of Table 1.3, and vary from −1.2816 (for the 10% VaR) to 1.2816 (for the 90% VaR). The third column shows the $\phi(p)$ weights corresponding to each confidence level, and the fourth column shows the products of each VaR and corresponding weight. Our estimated exponential spectral risk measure is the $\phi(p)$-weighted average of the VaRs, and is therefore equal to 0.4228.

As when estimating the ES earlier, when using this type of routine in practice we would want a value of n large enough to give accurate results. Table 1.4 reports some alternative estimates obtained using this procedure with increasing values of n. These results show that the estimated risk measure rises with n, and gradually converges to a value in the region of about 1.854. The estimates in this table indicate that we may need a considerably larger value of n than we did earlier to get results of the same level of accuracy. Even so, a good computer should still be able to produce accurate estimates of spectral risk measures fairly quickly.

When estimating ES or more general coherent risk measures in practice, it also helps to have some guidance on how to choose the value of n. Granted that the estimate does eventually converge to the true value as n gets large, one useful approach is to start with some small value of n, and then double n repeatedly until we feel the estimates have settled down sufficiently. Each time we do so, we halve the width of the discrete slices, and we can monitor how this 'halving' process affects our estimates. This suggests that we look at the 'halving error' ε_n given by:

$$\varepsilon_n = \hat{M}^{(n)} - \hat{M}^{(n/2)} \tag{1.20}$$

where $\hat{M}^{(n)}$ is our estimated risk measure based on n slices. We stop doubling n when ε_n falls below some tolerance level that indicates an acceptable level of accuracy. The process is

Table 1.3 Estimating Exponential Spectral Risk Measure as a Weighted Average of VaRs

Confidence Level (α)	αVaR	Weight $\phi(\alpha)$	$\phi(\alpha) \times \alpha$VaR
10%	−1.2816	0	0.0000
20%	−0.8416	0	0.0000
30%	−0.5244	0	0.0000
40%	−0.2533	0.0001	0.0000
50%	0	0.0009	0.0000
60%	0.2533	0.0067	0.0017
70%	0.5244	0.0496	0.0260
80%	0.8416	0.3663	0.3083
90%	1.2816	2.7067	3.4689
Risk measure = mean ($\phi(\alpha)$ times αVaR) =			0.4226

Note: VaRs estimated assuming the mean and standard deviation of losses are 0 and 1, using the 'normalvar' function in the MMR Toolbox. The weights $\phi(\alpha)$ are given by the exponential function (Equation (1.19)) with $\gamma = 0.05$.

Table 1.4 Estimates of Exponential Spectral Coherent Risk Measure as a Function of the Number of Tail Slices

Number of Tail Slices	Estimate of Exponential Spectral Risk Measure
10	0.4227
50	1.3739
100	1.5853
250	1.7338
500	1.7896
1000	1.8197
2500	1.8392
5000	1.8461
10 000	1.8498
50 000	1.8529
100 000	1.8533
500 000	1.8536

Note: VaRs estimated assuming the mean and standard deviation of losses are 0 and 1, using the 'normalvar' function in the MMR Toolbox. The weights $\phi(\alpha)$ are given by the exponential function (Equation (1.19)) with $\gamma = 0.05$.

Table 1.5 Estimated Risk Measures and Halving Errors

Number of Tail Slices	Estimated Spectral Risk Measure	Halving Error
100	1.5853	0.2114
200	1.7074	0.1221
400	1.7751	0.0678
800	1.8120	0.0368
1600	1.8317	0.0197
3200	1.8422	0.0105
6400	1.8477	0.0055
12 800	1.8506	0.0029
25 600	1.8521	0.0015
51 200	1.8529	0.0008

Note: VaRs estimated assuming the mean and standard deviation of losses are 0 and 1, using the 'normalvar' function in the MMR Toolbox. The weights $\phi(\alpha)$ are given by the exponential function (Equation (1.19)) with $\gamma = 0.05$.

shown in Table 1.5. Starting from an arbitrary value of 100, we repeatedly double n (so it becomes 200, 400, 800, etc.). As we do so, the estimated risk measure gradually converges, and the halving error gradually falls. So, for example, for $n = 6400$, the estimated risk measure is 1.8477, and the halving error is 0.0055. If we double n to 12,800, the estimated risk measure becomes 1.8506, and the halving error falls to 0.0029, and so on.

However, this 'weighted average quantile' procedure is rather crude, and (bearing in mind that the risk measure (Equation (1.17)) involves an integral) we can in principle expect to get substantial improvements in accuracy if we resorted to more 'respectable' numerical integration or quadrature methods. This said, the crude 'weighted average quantile' method actually seems to perform well for spectral exponential risk measures when compared against some of these alternatives, so one is not necessarily better off with the more sophisticated methods.[5]

[5] There is an interesting reason for this: the spectral weights give the highest loss the highest weight, whereas the quadrature methods such as the trapezoidal and Simpson's rules involve algorithms in which the two most extreme quantiles have their weights specifically cut, and this undermines the accuracy of the algorithm relative to the crude approach. However, there are ways round these sorts of problems, and in principle versions of the sophisticated approaches should give better results.

Thus, the key to estimating any coherent risk measure is to be able to estimate quantiles or VaRs: the coherent risk measures can then be obtained as appropriately weighted averages of quantiles. From a practical point of view, this is extremely helpful as all the building blocks that go into quantile or VaR estimation—databases, calculation routines, etc.—are exactly what we need for the estimation of coherent risk measures as well. If an institution already has a VaR engine, then that engine needs only small adjustments to produce estimates of coherent risk measures: indeed, in many cases, all that needs changing is the last few lines of code in a long data processing system. The costs of switching from VaR to more sophisticated risk measures are therefore very low.

1.5 ESTIMATING THE STANDARD ERRORS OF RISK MEASURE ESTIMATORS

We should always bear in mind that any risk measure estimates that we produce are just that—estimates. We never know the true value of any risk measure, and an estimate is only as good as its precision: if a risk measure is very imprecisely estimated, then the estimator is virtually worthless, because its imprecision tells us that true value could be almost anything; on the other hand, if we know that an estimator is fairly precise, we can be confident that the true value is fairly close to the estimate, and the estimator has some value. Hence, we should always seek to supplement any risk estimates we produce with some indicator of their precision. This is a fundamental principle of good risk measurement practice.

We can evaluate the precision of estimators of risk measures by means of their standard errors, or (generally better) by producing confidence intervals for them. In this chapter we focus on the more basic indicator, the standard error of a risk measure estimator.

Standard Errors of Quantile Estimators

We first consider the standard errors of quantile (or VaR) estimators. Following Kendall and Stuart,[6] suppose we have a distribution (or cumulative density) function $F(x)$, which might be a parametric distribution function or an empirical

[6] Kendall and Stuart (1972), pp. 251–252.

distribution function (i.e., a cumulative histogram) estimated from real data. Its corresponding density or relative-frequency function is $f(x)$. Suppose also that we have a sample of size n, and we select a bin width h. Let dF be the probability that $(k-1)$ observations fall below some value $q - h/2$, that one observation falls in the range $q \pm h/2$, and that $(n-k)$ observations are greater than $q + h/2$. dF is proportional to

$$\{F(q)\}^{k-1} f(q)dq\{1 - F(q)\}^{n-k} \tag{1.21}$$

This gives us the frequency function for the quantile q not exceeded by a proportion k/n of our sample, i.e., the $100(k/n)$th percentile.

If this proportion is p, Kendall and Stuart show that Equation (1.21) is approximately equal to $p^{np}(1-p)^{n(1-p)}$ for large values of n. If ε is a very small increment to p, then

$$p^{np}(1-p)^{n(1-p)} \approx (p+\varepsilon)^{np}(1-p-\varepsilon)^{n(1-p)} \tag{1.22}$$

Taking logs and expanding, Equation (1.22) is itself approximately

$$(p+\varepsilon)^{np}(1-p-\varepsilon)^{n(1-p)} \approx \frac{n\varepsilon^2}{2p(1-p)} \tag{1.23}$$

which implies that the distribution function dF is approximately proportional to

$$\exp\left(\frac{-n\varepsilon^2}{2p(1-p)}\right) \tag{1.24}$$

Integrating this out,

$$dF = \frac{1}{\sqrt{2\pi}\sqrt{p(1-p)/n}} \exp\left(\frac{-n\varepsilon^2}{2p(1-p)}\right) d\varepsilon \tag{1.25}$$

which tells us that ε is normally distributed in the limit with variance $p(1-p)/n$. However, we know that $d\epsilon = dF(q) = f(q)dq$, so the variance of q is

$$var(q) \approx \frac{p(1-p)}{n[f(q)]^2} \tag{1.26}$$

This gives us an approximate expression for the variance, and hence its square root, the standard error, of a quantile estimator q.

This expression shows that the quantile standard error depends on p, the sample size n and the pdf value $f(q)$. The way in which the (normal) quantile standard errors depend on these parameters is apparent from Figure 1.9. This shows that:

- The standard error falls as the sample size n rises.
- The standard error rises as the probabilities become more extreme and we move further into the tail—hence, the more extreme the quantile, the less precise its estimator.

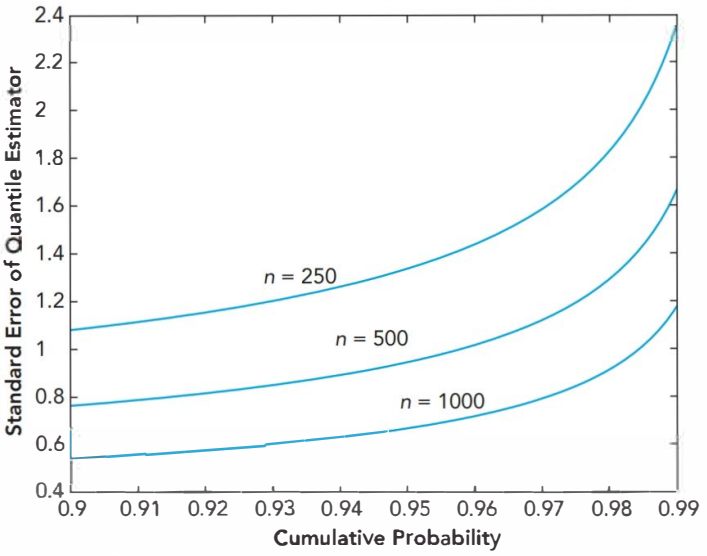


Figure 1.9 Standard errors of quantile estimators.

Note: Based on random samples of size n drawn from a standard normal distribution. The bin width h is set to 0.1.

In addition, the quantile standard error depends on the probability density function $f(.)$—so the choice of density function can make a difference to our estimates—and also on the bin width h, which is essentially arbitrary.

The standard error can be used to construct confidence intervals around our quantile estimates in the usual textbook way. For example, a 90% confidence interval for a quantile q is given by

$$[q - 1.645\ se(q),\ q + 1.645\ se(q)]$$
$$= \left[q - 1.645\frac{\sqrt{p(1-p)/n}}{f(q)},\ q + 1.645\frac{\sqrt{p(1-p)/n}}{f(q)}\right] \tag{1.27}$$

Example 1.6 Obtaining VaR Confidence Intervals Using Quantile Standard Errors

Suppose we wish to estimate the 90% confidence interval for a 95% VaR estimated on a sample of size of $n = 1000$ to be drawn from a standard normal distribution, based on an assumed bin width $h = 0.1$.

We know that the 95% VaR of a standard normal is 1.645. We take this to be q in Equation (1.27), and we know that q falls in the bin spanning $1.645 \pm 0.1/2 = [1.595, 1.695]$. The probability of a loss exceeding 1.695 is 0.045, and this is also equal to p, and the probability of profit or a loss less than 1.595 is 0.9446. Hence $f(q)$, the probability mass in the q range, is $1 - 0.0450 - 0.9446 = 0.0104$. We now plug the

relevant values into Equation (1.27) to obtain the 90% confidence interval for the VaR:

$$\left[1.645 - 1.645\frac{\sqrt{0.045(1-0.045)/1000}}{0.0104},\ 1.645 + 1.645\frac{\sqrt{0.045(1-0.045)/1000}}{0.0104}\right] = [0.6081, 2.6819]$$

This is a wide confidence interval, especially when compared to the OS and bootstrap confidence intervals.

The confidence interval narrows if we take a wider bin width, so suppose that we now repeat the exercise using a bin width $h = 0.2$, which is probably as wide as we can reasonably go with these data. q now falls into the range $1.645 \pm 0.2/2 = [1.545, 1.745]$. p, the probability of a loss exceeding 1.745, is 0.0405, and the probability of profit or a loss less than 1.545 is 0.9388. Hence $f(q) = 1 - 0.0405 - 0.9388 = 0.0207$. Plugging these values into Equation (1.27) now gives us a new estimate of the 90% confidence interval:

$$\left[1.645 - 1.645\frac{\sqrt{0.0405(1-0.0405)/1000}}{0.0207},\ 1.645 + 1.645\frac{\sqrt{0.0405(1-0.0405)/1000}}{0.0207}\right] = [1.1496, 2.1404]$$

This is still a rather wide confidence interval.

This example illustrates that although we can use quantile standard errors to estimate VaR confidence intervals, the intervals can be wide and also sensitive to the arbitrary choice of bin width.

The quantile-standard-error approach is easy to implement and has some plausibility with large sample sizes. However, it also has weaknesses relative to other methods of assessing the precision of quantile (or VaR) estimators—it relies on asymptotic theory and requires large sample sizes; it can produce imprecise estimators, or wide confidence intervals; it depends on the arbitrary choice of bin width; and the symmetric confidence intervals it produces are misleading for extreme quantiles whose 'true' confidence intervals are asymmetric reflecting the increasing sparsity of extreme observations as we move further out into the tail.

Standard Errors in Estimators of Coherent Risk Measures

We now consider standard errors in estimators of coherent risk measures. One of the first studies to examine this issue (Yamai and Yoshiba (2001b) did so by investigating the relative accuracy of VaR and ES estimators for Lévy distributions with varying α stability parameters. Their results suggested that VaR and ES estimators had comparable standard errors for near-normal Lévy distributions, but the ES estimators had much bigger standard errors for particularly heavy-tailed distributions. They explained this finding by saying that as tails became heavier, ES estimators became more prone to the effects of large but infrequent losses. This finding suggests the depressing conclusion that the presence of heavy tails might make ES estimators in general less accurate than VaR estimators.

Fortunately, there are grounds to think that such a conclusion might be overly pessimistic. For example, Inui and Kijima (2003) present alternative results showing that the application of a Richardson extrapolation method can produce ES estimators that are both unbiased and have comparable standard errors to VaR estimators.[7] Acerbi (2004) also looked at this issue and, although he confirmed that tail heaviness did increase the standard errors of ES estimators relative to VaR estimators, he concluded that the relative accuracies of VaR and ES estimators were roughly comparable in empirically realistic ranges.

However, the standard error of any estimator of a coherent risk measure will vary from one situation to another, and the best practical advice is to get into the habit of always estimating the standard error whenever one estimates the risk measure itself. Estimating the standard error of an estimator of a coherent risk measure is also relatively straightforward. One way to do so starts from recognition that a coherent risk measure is an L-estimator (i.e., a weighted average of order statistics), and L-estimators are asymptotically normal. If we take N discrete points in the density function, then as N gets large the variance of the estimator of the coherent risk measure (Equation (1.17)) is approximately

$$\sigma(\hat{M}_\phi^{(N)}) \to \frac{2}{N}\int_{p<q}\phi(p)\phi(q)\frac{p(1-q)}{f(F^{-1}(p))f(F^{-1}(q))}dpdq = \frac{2}{N}\int_{x<y}\phi(F(x))\phi(F(y))F(x)(1-F(y))dxdy \quad (1.28)$$

and this can be computed numerically using a suitable numerical integration procedure. Where the risk measure is the ES, the standard error becomes

$$\sigma(\hat{ES}^{(N)}) \to \frac{1}{N\alpha^2}\int_0^{F^{-1}(\alpha)}\int_0^{F^{-1}(\alpha)}[\min(F(x), F(y)) - F(x)F(y)]dxdy \quad (1.29)$$

and used in conjunction with a suitable numerical integration method, this gives good estimates even for relatively low values

[7] See Inui and Kijima (2003).

of N.[8] If we wish to obtain confidence intervals for our risk measure estimators, we can make use of the asymptotic normality of these estimators to apply textbook formulas (e.g., such as Equation (1.27)) based on the estimated standard errors and centred around a 'good' best estimate of the risk measure.

An alternative approach to the estimation of standard errors for estimators of coherent risk measures is to apply a bootstrap: we bootstrap a large number of estimators from the given distribution function (which might be parametric or non-parametric, e.g., historical); and we estimate the standard error of the sample of bootstrapped estimators. Even better, we can also use a bootstrapped sample of estimators to estimate a confidence interval for our risk measure.

1.6 THE CORE ISSUES: AN OVERVIEW

Before proceeding to more detailed issues, it might be helpful to pause for a moment to take an overview of the structure, as it were, of the subject matter itself. This is very useful, as it gives the reader a mental frame of reference within which the 'detailed' material that follows can be placed. Essentially, there are three core issues, and all the material that follows can be related to these. They also have a natural sequence, so we can think of them as providing a roadmap that leads us to where we want to be.

Which risk measure? The first and most important is to choose the type of risk measure: do we want to estimate VaR, ES, etc.? This is logically the first issue, because we need to know *what* we are trying to estimate before we start thinking about *how* we are going to estimate it.

Which level? The second issue is the *level* of analysis. Do we wish to estimate our risk measure at the level of the portfolio as a whole or at the level of the individual positions in it? The former would involve us taking the portfolio as our basic unit of analysis (i.e., we take the portfolio to have a specified composition, which is taken as given for the purposes of our analysis), and this will lead to a *univariate* stochastic analysis. The alternative is to work from the position level, and this has the advantage of allowing us to accommodate changes in the portfolio composition within the analysis itself. The disadvantage is that we then need a *multivariate* stochastic framework, and this is considerably more difficult to handle: we have to get to grips with the problems posed by variance–covariance matrices, copulas, and so on, all of which are avoided if we work at the portfolio level. There is thus a trade-off: working at the portfolio level is more limiting, but easier, while working at the position level gives us much more flexibility, but can involve much more work.

Which method? Having chosen our risk measure and level of analysis, we then choose a suitable estimation method. To decide on this, we would usually think in terms of the classic 'VaR trinity':

- Non-parametric methods
- Parametric methods
- Monte Carlo simulation methods

Each of these involves some complex issues.

1.7 APPENDIX

Preliminary Data Analysis

When confronted with a new data set, we should *never* proceed straight to estimation without some preliminary analysis to get to know our data. Preliminary data analysis is useful because it gives us a feel for our data, and because it can highlight problems with our data set. Remember that we never really know where our data come from, so we should always be a little wary of any new data set, regardless of how reputable the source might appear to be. For example, how do you know that a clerk hasn't made a mistake somewhere along the line in copying the data and, say, put a decimal point in the wrong place? The answer is that you don't, and never can. Even the most reputable data providers provide data with errors in them, however careful they are. Everyone who has ever done any empirical work will have encountered such problems at some time or other: the bottom line is that real data must always be viewed with a certain amount of suspicion.

Such preliminary analysis should consist of at least the first two and preferably all three of the following steps:

- The first and by far the most important step is to eyeball the data to see if they 'look right'—or, more to the point, we should eyeball the data to see if anything looks *wrong*. Does the pattern of observations look right? Do any observations stand out as questionable? And so on. The interocular trauma test is the most important test ever invented and also the easiest to carry out, and we should always perform it on any new data set.
- We should plot our data on a histogram and estimate the relevant summary statistics (i.e., mean, standard deviation, skewness, kurtosis, etc.). In risk measurement, we are particularly interested in any non-normal features of our data: skewness, excess kurtosis, out-liers in our data, and the like.

[8] See Acerbi (2004, pp. 200–201).

We should therefore be on the lookout for any evidence of non-normality, and we should take any such evidence into account when considering whether to fit any parametric distribution to the data.

- Having done this initial analysis, we should consider what kind of distribution might fit our data, and there are a number of useful diagnostic tools available for this purpose, the most popular of which are QQ plots—plots of empirical quantiles against their theoretical equivalents.

Plotting the Data and Evaluating Summary Statistics

To get to know our data, we should first obtain their histogram and see what might stand out. Do the data look normal, or non-normal? Do they show one pronounced peak, or more than one? Do they seem to be skewed? Do they have fat tails or thin tails? Are there outliers? And so on.

As an example, Figure 1.10 shows a histogram of 100 random observations. In practice, we would usually wish to work with considerably longer data sets, but a data set this small helps to highlight the uncertainties one often encounters in practice. These observations show a dominant peak in the centre, which suggests that they are probably drawn from a unimodal distribution. On the other hand, there may be a negative skew, and there are some large outlying observations on the extreme left of the distribution, which might indicate fat tails on at least the left-hand side. In fact, these particular observations are drawn from a Student-*t* distribution with 5 degrees of freedom, so in this case we know that the underlying true distribution is unimodal, symmetric and heavy tailed. However, we would not know this in a situation with 'real' data, and it is precisely because we do not know the distributions of real-world data sets that preliminary analysis is so important.

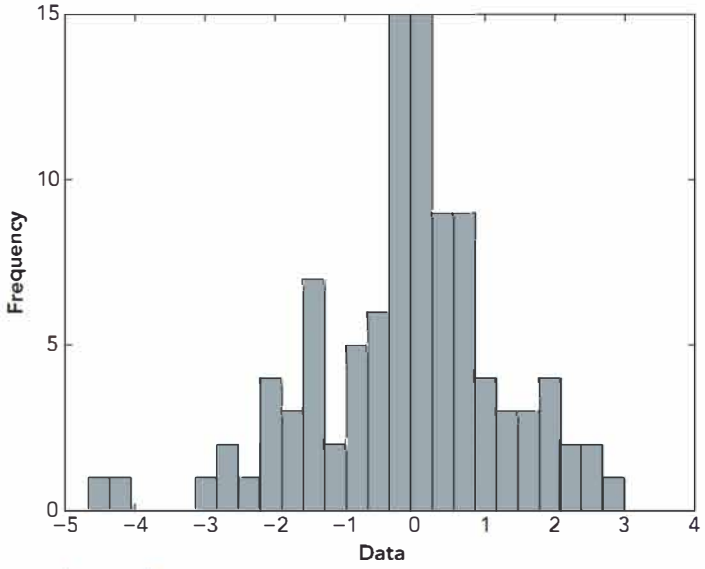


Figure 1.10 **A histogram.**

Note: Data are 100 observations randomly drawn from a Student-*t* with 5 degrees of freedom.

Some summary statistics for this data set are shown in Table 1.6. The sample mean (−0.099) and the sample mode differ somewhat (−0.030), but this difference is small relative to the sample standard deviation (1.363). However, the sample skew (−0.503) is somewhat negative and the sample kurtosis (3.985) is a little bigger than normal. The sample minimum (−4.660) and the sample maximum (3.010) are also not symmetric about the sample mean or mode, which is further evidence of asymmetry. If we encountered these results with 'real' data, we would be concerned about possible skewness and kurtosis. However, in this hypothetical case we know that the sample skewness is merely a product of sample variation, because we happen to know that the data are drawn from a symmetric distribution.

Depending on the context, we might also seriously consider carrying out some formal tests. For example, we might test whether the sample parameters (mean, standard deviation, etc.) are consistent with what we might expect under a null hypothesis (e.g., such as normality).

The underlying principle is very simple: since we *never* know the true distribution in practice, all we ever have to work with are *estimates* based on the *sample* at hand; it therefore behoves us to make the best use of the data we have, and to extract as much information as possible from them.

QQ Plots

Having done our initial analysis, it is often good practice to ask what distribution might fit our data, and a very useful device for identifying the distribution of our data is a quantile–quantile or QQ plot—a plot of the quantiles of the empirical distribution against those of some specified distribution. The shape of the QQ plot tells us a lot about how the empirical distribution compares to the specified one. In particular, if the QQ plot is linear, then the specified distribution fits the data, and we have identified the distribution to which our data belong. In addition, departures of the QQ from linearity in the tails can tell us whether the tails of our empirical distribution are fatter, or thinner, than the tails of the reference distribution to which it is being compared.

Table 1.6 Summary Statistics

Parameter	Value
Mean	−0.099
Mode	−0.030
Standard deviation	1.363
Skewness	−0.503
Kurtosis	3.985
Minimum	−4.660
Maximum	3.010
Number of observations	100

Note: Data are the same observations shown in Figure 1.10.

To illustrate, Figure 1.11 shows a QQ plot for a data sample drawn from a normal distribution, compared to a reference distribution that is also normal. The QQ plot is obviously close to linear: the central mass observations fit a linear QQ plot very closely, and the extreme tail observations somewhat less so. However, there is no denying that the overall plot is approximately linear. Figure 1.11 is a classic example of a QQ plot in which the empirical distribution matches the reference population.

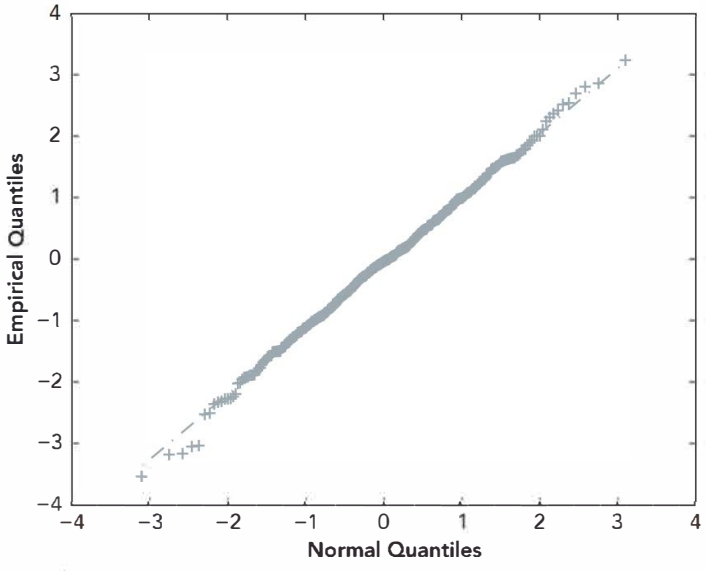


Figure 1.11 QQ plot: normal sample against normal reference distribution.

Note: The empirical sample is a random sample of 500 observations drawn from a standard normal. The reference distribution is standard normal.

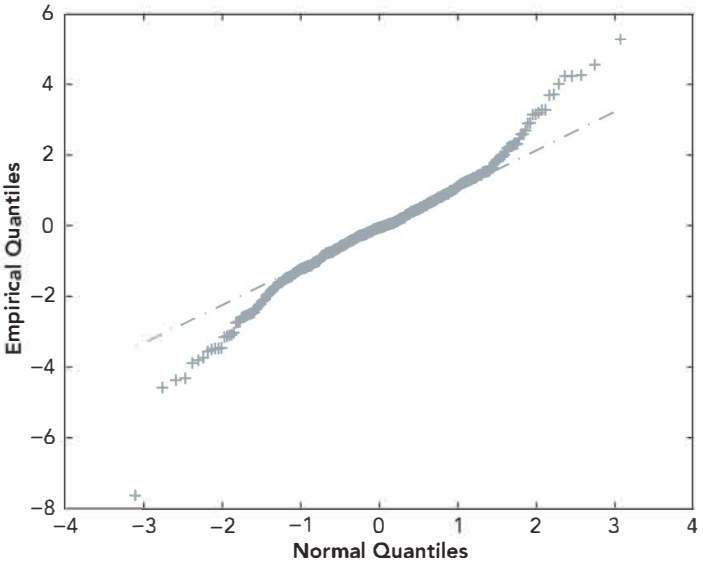


Figure 1.12 QQ plot: *t* sample against normal reference distribution.

Note: The empirical sample is a random sample of 500 observations drawn from Student-*t* with 5 degrees of freedom. The reference distribution is standard normal.

By contrast, Figure 1.12 shows a good example of a QQ plot where the empirical distribution does not match the reference population. In this case, the data are drawn from a Student-*t* with 5 degrees of freedom, but the reference distribution is standard normal. The QQ plot is now clearly non-linear: although the central mass observations are close to linear, the tails show steeper slopes indicative of the tails being heavier than those of the reference distribution.

A QQ plot is useful in a number of ways. First, as noted already, if the data are drawn from the reference population, then the QQ plot should be linear. This remains true if the data are drawn from some linear transformation of the reference distribution (i.e., are drawn from the same distribution but with different location and scale parameters). We can therefore use a QQ plot to form a tentative view of the distribution from which our data might be drawn: we specify a variety of alternative distributions, and construct QQ plots for each. Any reference distributions that produce non-linear QQ plots can then be dismissed, and any distribution that produces a linear QQ plot is a good candidate distribution for our data.

Second, because a linear transformation in one of the distributions in a QQ plot merely changes the intercept and slope of the QQ plot, we can use the intercept and slope of a linear QQ plot to give us a rough idea of the location and scale parameters of our sample data. For example, the reference distribution in

Figure 1.11 is a standard normal. The linearity of the QQ plot in this figure suggests that the data are normal, as mentioned already, but Figure 1.11 also shows that the intercept and slope are approximately 0 and 1 respectively, and this indicates that the data are drawn from a standard normal, and not just any normal. Such rough approximations give us a helpful yardstick against which we can judge more 'sophisticated' estimates of location and scale, and also provide useful initial values for iterative algorithms.

Third, if the empirical distribution has heavier tails than the reference distribution, the QQ plot will have steeper slopes at its tails, even if the central mass of the empirical observations are approximately linear. Figure 1.12 is a good case in point. A QQ plot where the tails have slopes different than the central mass is therefore suggestive of the empirical distribution having heavier, or thinner, tails than the reference distribution.

Finally, a QQ plot is good for identifying outliers (e.g., observations contaminated by large errors): such observations will stand out in a QQ plot, even if the other observations are broadly consistent with the reference distribution.[9]

[9] Another useful tool, especially when dealing with the tails, is the mean excess function (MEF): the expected amount by which a random variable X exceeds some threshold u, given that $X > u$. The usefulness of the MEF stems from the fact that each distribution has its own distinctive MEF. A comparison of the empirical MEF with the theoretical MEF associated with some specified distribution function therefore gives us an indication of whether the chosen distribution fits the tails of our empirical distribution. However, the results of MEF plots need to be interpreted with some care, because data observations become more scarce as X gets larger. For more on these and how they can be used, see Embrechts et. al. (1997, Chapters 3.4 and 6.2).

2

Non-Parametric Approaches

Learning Objectives

After completing this reading, you should be able to:

- Apply the bootstrap historical simulation approach to estimate coherent risk measures.
- Describe historical simulation using non-parametric density estimation.
- Compare and contrast the age-weighted, the volatility-weighted, the correlation-weighted, and the filtered historical simulation approaches.
- Identify advantages and disadvantages of non-parametric estimation methods.

Excerpt is Chapter 4 of Measuring Market Risk, Second Edition, by Kevin Dowd.

This chapter looks at some of the most popular approaches to the estimation of risk measures—the non-parametric approaches, which seek to estimate risk measures without making strong assumptions about the relevant (e.g., P/L) distribution. The essence of these approaches is that we try to let the P/L data speak for themselves as much as possible, and use the recent *empirical* (or in some cases simulated) distribution of P/L—not some assumed theoretical distribution—to estimate our risk measures. All non-parametric approaches are based on the underlying assumption that the near future will be sufficiently like the recent past that we can use the data from the recent past to forecast risks over the near future—and this assumption may or may not be valid in any given context. In deciding whether to use any non-parametric approach, we must make a judgment about the extent to which data from the recent past are likely to give us a good guide about the risks we face over the horizon period we are concerned with.

To keep the discussion as clear as possible, we will focus on the estimation of non-parametric VaR and ES. However, the methods discussed here extend very naturally to the estimation of coherent and other risk measures as well. These can be estimated using an 'average quantile' approach along the lines discussed in Chapter 1: we would select our weighting function $\phi(p)$, decide on the number of probability 'slices' n to take, estimate the associated quantiles, and take the weighted average using an appropriate numerical algorithm (see Chapter 1).[1] We can then obtain standard errors or confidence intervals for our risk measures using suitably modified forms.

In this chapter we begin by discussing how to assemble the P/L data to be used for estimating risk measures. We then discuss the most popular non-parametric approach-historical simulation (HS). Loosely speaking, HS is a histogram-based approach: it is conceptually simple, easy to implement, very widely used, and has a fairly good historical record. We focus on the estimation of VaR and ES, but as explained in the previous chapter, more general coherent risk measures can be estimated using appropriately weighted averages of any non-parametric VaR estimates. We then discuss refinements to basic HS using bootstrap and kernel methods, and the estimation of VaR or ES curves and surfaces. We will discuss how we can estimate confidence intervals for HS VaR and ES. Then we will address weighted HS—how we might weight our data to capture the effects of observation age and changing market conditions. These methods introduce parametric formulas (such as GARCH volatility forecasting equations) into the picture, and in so doing convert hitherto non-parametric methods into what are best described as semi-parametric methods. Such methods are very useful because they allow us to retain the broad HS framework while also taking account of ways in which we think that the risks we face over the foreseeable horizon period might differ from those in our sample period. Finally we review the main advantages and disadvantages of non-parametric and semi-parametric approaches, and offer some conclusions.

2.1 COMPILING HISTORICAL SIMULATION DATA

The first task is to assemble a suitable P/L series for our portfolio, and this requires a set of historical P/L or return observations on the positions in our current portfolio. These P/Ls or returns will be measured over a particular frequency (e.g., a day), and we want a reasonably large set of historical P/L or return observations over the recent past. Suppose we have a portfolio of n assets, and for each asset i we have the observed return for each of T subperiods (e.g., daily subperiods) in our historical sample period. If $R_{i,t}$ is the (possibly mapped) return on asset i in subperiod t, and if w_i is the amount currently invested in asset i, then the historically simulated portfolio P/L over the subperiod t is:

$$P/L_t = \sum_{i=1}^{n} w_i R_{i,t} \tag{2.1}$$

Equation (2.1) gives us a historically simulated P/L series for our current portfolio, and is the basis of HS VaR and ES. This series will *not* generally be the same as the P/L *actually* earned on our portfolio—because our portfolio may have changed in composition over time or be subject to mapping approximations, and so on. Instead, the historical simulation P/L is the P/L we *would have* earned on our current portfolio had we held it throughout the historical sample period.[2]

As an aside, the fact that multiple positions collapse into one single HS P/L as given by Equation (2.1) implies that it is

[1] Nonetheless, there is an important caveat. This method was explained in Chapter 1 in an implicit context where the risk measurer could choose n, and this is sometimes not possible in a non-parametric context. For example, a risk measurer might be working with an n determined by the HS data set, and even where he/she has some freedom to select n, their range of choice might be limited by the data available. Such constraints can limit the degree of accuracy of any resulting estimated risk measures. However, a good solution to such problems is to increase the sample size by bootstrapping from the sample data. (The bootstrap is discussed further in Appendix 2 to this chapter).

[2] To be more precise, the historical simulation P/L is the P/L we would have earned over the sample period had we rearranged the portfolio at the end of each trading day to ensure that the amount left invested in each asset was the same as at the end of the previous trading day: we take out our profits, or make up for our losses, to keep the w_i constant from one end-of-day to the next.

very easy for non-parametric methods to accommodate high dimensions—unlike the case for some parametric methods. With non-parametric methods, there are no problems dealing with variance-covariance matrices, curses of dimensionality, and the like. This means that non-parametric methods will often be the most natural choice for high-dimension problems.

2.2 ESTIMATION OF HISTORICAL SIMULATION VaR AND ES

Basic Historical Simulation

Having obtained our historical simulation P/L data, we can estimate VaR by plotting the P/L (or L/P) on a simple histogram and then reading off the VaR from the histogram. To illustrate, suppose we have 1000 observations in our HS P/L series and we plot the L/P histogram shown in Figure 2.1. If these were daily data, this sample size would be equivalent to four years' daily data at 250 trading days to a year. If we take our confidence level to be 95%, our VaR is given by the *x*-value that cuts off the upper 5% of very high losses from the rest of the distribution. Given 1000 observations, we can take this value (i.e., our VaR) to be the 51st highest loss value, or 1.704.[3] The ES is then the average of the 50 highest losses, or 2.196.

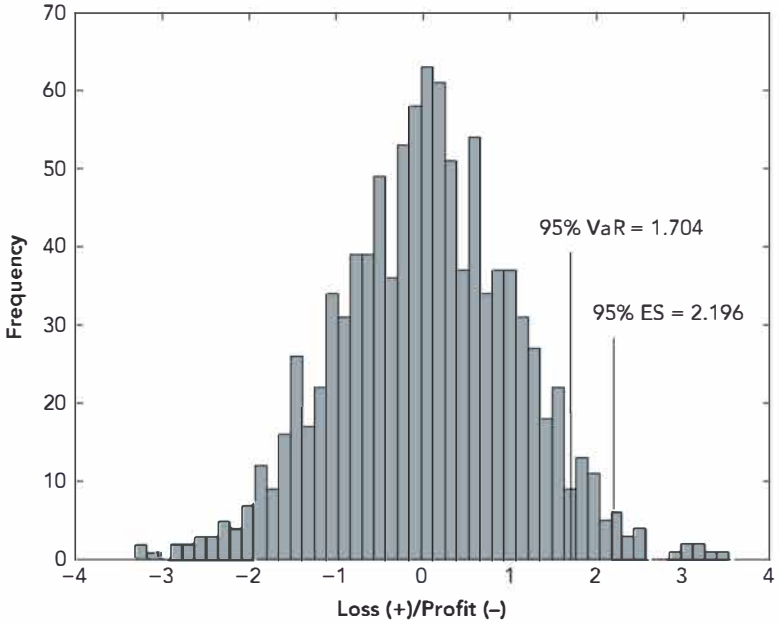


Figure 2.1 Basic historical simulation VaR and ES.

Note: This figure and associated VaR and ES estimates are obtained using the 'hsesfigure' function.

The imprecision of these estimates should be obvious when we consider that the sample data set was drawn from a standard normal distribution. In this case the 'true' underlying VaR and ES are 1.645 and 2.063, and Figure 2.1 should (ideally) be normal. Of course, this imprecision underlines the need to work with large sample sizes where practically feasible.

Bootstrapped Historical Simulation

One simple but powerful improvement over basic HS is to estimate VaR and ES from bootstrapped data. As explained in Appendix 2 to this chapter, a bootstrap procedure involves resampling from our existing data set with replacement. The bootstrap is very intuitive and easy to apply. A bootstrapped estimate will often be more accurate than a 'raw' sample estimate, and bootstraps are also useful for gauging the precision of our estimates. To apply the bootstrap, we create a large number of new samples, each observation of which is obtained by drawing at random from our original sample and replacing the observation after it has been drawn. Each new 'resampled' sample gives us a new VaR estimate, and we can take our 'best' estimate to be the mean of these resample-based estimates. The same approach can also be used to produce resample-based ES estimates—each one of which would be the average of the losses in each resample exceeding the resample VaR—and our 'best' ES estimate would be the mean of these estimates. In our particular case, if we take 1000 resamples, then our best VaR and ES estimates are (because of bootstrap sampling variation) about 1.669 and 2.114—and the fact that these are much closer to the known true values than our earlier basic HS estimates suggests that bootstraps estimates might be more accurate.

Historical Simulation Using Non-parametric Density Estimation

Another potential improvement over basic HS sometimes suggested is to make use of non parametric density estimation. To appreciate what this involves, we must recognise that basic HS does not make the best use of the information we have. It also has the practical drawback that it only allows us to

[3] We can also estimate the HS VaR more directly (i.e., without bothering with the histogram) by using a spreadsheet function that gives us the 51st highest loss value (e.g., the 'Large' command in Excel), or we can sort our losses data with highest losses ranked first, and then obtain the VaR as the 51st observation in our sorted loss data. We could also take VaR to be any point between the 50th and 51st largest losses (e.g., such as their mid-point), but with a reasonable sample size (as here) there will seldom be much difference between these losses anyway. For convenience, we will adhere throughout to this convention of taking the VaR to be the highest loss observation outside the tail.

estimate VaRs at discrete confidence intervals determined by the size of our data set. For example, if we have 100 HS P/L observations, basic HS allows us to estimate VaR at the 95% confidence level, but not the VaR at the 95.1% confidence level. The VaR at the 95% confidence level is given by the sixth largest loss, but the VaR at the 95.1% confidence level is a problem because there is no corresponding loss observation to go with it. We know that it should be greater than the sixth largest loss (or the 95% VaR), and smaller than the fifth largest loss (or the 96% VaR), but with only 100 observations there is no observation that corresponds to any VaR whose confidence level involves a fraction of 1%. With *n* observations, basic HS only allows us to estimate the VaRs associated with, at best, *n* different confidence levels.

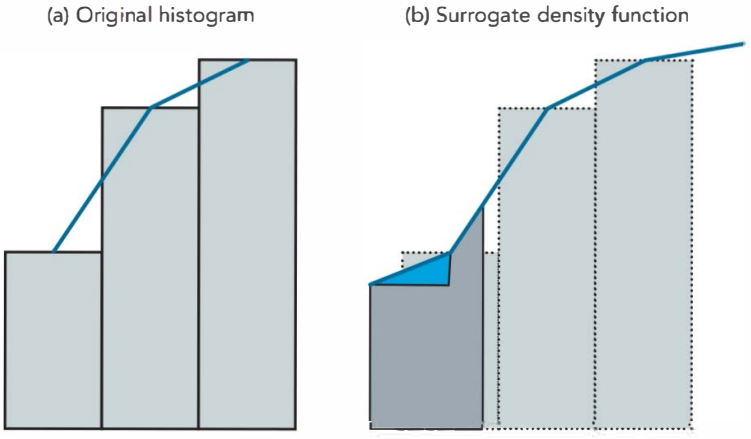


Figure 2.2 **Histograms and surrogate density functions.**

Non-parametric density estimation offers a potential solution to both these problems. The idea is to treat our data as if they were drawings from some unspecified or unknown empirical distribution function. This approach also encourages us to confront potentially important decisions about the width of bins and where bins should be centred, and these decisions can sometimes make a difference to our results. Besides using a histogram, we can also represent our data using naïve estimators or, more generally, kernels, and the literature tells us that kernels are (or ought to be) superior. So, having assembled our 'raw' HS data, we need to make decisions on the widths of bins and where they should be centred, and whether to use a histogram, a naïve estimator, or some form of kernel. If we make good decisions on these issues, we can hope to get better estimates of VaR and ES (and more general coherent measures).

Non-parametric density estimation also allows us to estimate VaRs and ESs for any confidence levels we like and so avoid constraints imposed by the size of our data set. In effect, it enables us to draw lines through points on or near the edges of the 'bars' of a histogram. We can then treat the areas under these lines as a surrogate pdf, and so proceed to estimate VaRs for arbitrary confidence levels. The idea is illustrated in Figure 2.2. The left-hand side of this figure shows three bars from a histogram (or naïve estimator) close up. Assuming that the height of the histogram (or naïve estimator) measures relative frequency, then one option is to treat the histogram itself as a pdf. Unfortunately, the resulting pdf would be a strange one—just look at the corners of each bar—and it makes more sense to approximate the pdf by drawing lines through the upper parts of the histogram.

A simple way to do this is to draw in straight lines connecting the mid-points at the top of each histogram bar, as illustrated in the figure. Once we draw these lines, we can forget about the histogram bars and treat the area under the lines as if it were a pdf. Treating the area under the lines as a pdf then enables us to estimate VaRs at any confidence level, regardless of the size of our data set. Each possible confidence level would correspond to its own tail similar to the shaded area shown in Figure 2.2(b), and we can then use a suitable calculation method to estimate the VaR (e.g., we can carry out the calculations on a spreadsheet or, more easily, by using a purpose-built function such as the 'hsvar' function in the MMR Toolbox).[4] Of course, drawing straight lines through the mid-points of the tops of histogram bars is not the best we can do: we could draw smooth curves that meet up nicely, and so on. This is exactly the point of non-parametric density estimation, the purpose of which is to give us some guidance on how 'best' to draw lines through the data points we have. Such methods are also straightforward to apply if we have suitable software.

Some empirical evidence by Butler and Schachter (1998) using real trading portfolios suggests that kernel-type methods produce VaR estimates that are a little different to those we would obtain under basic HS. However, their work also suggests that the different types of kernel methods produce quite similar VaR estimates, although to the extent that there are differences among them, they also found that the 'best' kernels were the adaptive Epanechinikov and adaptive Gaussian ones. To investigate these issues myself, I applied four standard kernel estimators—based on normal, box, triangular and Epanechinikov kernels—to the test data used in earlier examples, and found that each of these gave the same VaR estimate of 1.735. In this case, these different kernels produced the same VaR estimate, which is a little higher (and, curiously,

[4] The actual programming is a little tedious, but the gist of it is that if the confidence level is such that the VaR falls between two loss observations, then we take the VaR to be a weighted average of these two observations. The weights are chosen so that a vertical line drawn through the VaR demarcates the area under the 'curve' in the correct proportions, with α to one side and $1 - \alpha$ to the other. The details can be seen in the coding for the 'hsvar' and related functions.

a little less accurate) than the basic HS VaR estimate of 1.704 obtained earlier. Other results not reported here suggest that the different kernels can give somewhat different estimates with smaller samples, but again suggest that the exact kernel specification does not make a great deal of difference.

So although kernel methods are better in theory, they do not necessarily produce much better estimates in practice. There are also practical reasons why we might prefer simpler non-parametric density estimation methods over kernel ones. Although the kernel methods are theoretically better, crude methods like drawing straight-line 'curves' through the tops of histograms are more transparent and easier to check. We should also not forget that our results are subject to a number of sources of error (e.g., due to errors in P/L data, mapping approximations, and so on), so there is a natural limit to how much real fineness we can actually achieve.

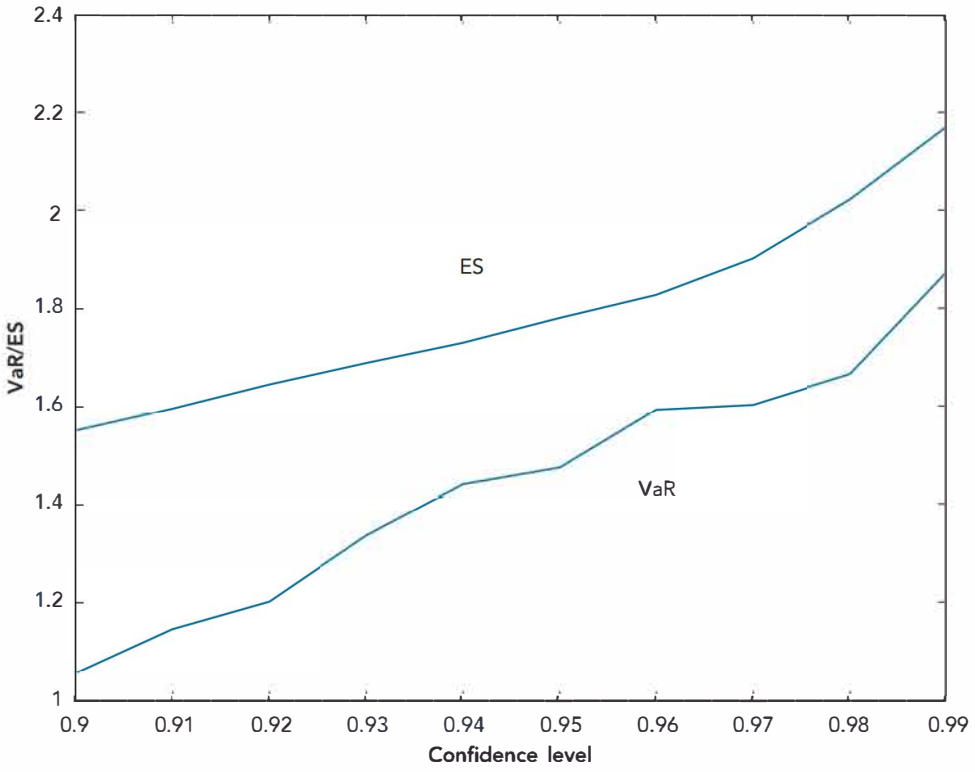


Figure 2.3 **Plots of HS VaR and ES against confidence level.**

Note: Obtained using the 'hsvaresplot2D_cl' function and the same hypothetical P/L data used in Figure 2.1.

Estimating Curves and Surfaces for VaR and ES

It is straightforward to produce plots of VaR or ES against the confidence level. For example, our earlier hypothetical P/L data yields the curves of VaR and ES against the confidence level shown in Figure 2.3. Note that the VaR curve is fairly unsteady, as it directly reflects the randomness of individual loss observations, but the ES curve is smoother, because each ES is an average of tail losses.

It is more difficult constructing curves that show how non-parametric VaR or ES changes with the holding period. The methods discussed so far enable us to estimate the VaR or ES at a single holding period equal to the frequency period over which our data are observed (e.g., they give us VaR or ES for a daily holding period if P/L is measured daily). In theory, we can then estimate VaRs or ESs for any other holding periods we wish by constructing a HS P/L series whose frequency matches our desired holding period: if we wanted to estimate VaR over a weekly holding period, say, we could construct a weekly P/L series and estimate the VaR from that. There is, in short, no theoretical problem as such with estimating HS VaR or ES over any holding period we like.

However, there is a major practical problem: as the holding period rises, the number of observations rapidly falls, and we soon find that we don't have enough data. To illustrate, if we have 1000 observations of daily P/L, corresponding to four years' worth of data at 250 trading days a year, then we have 1000 P/L observations if we use a daily holding period. If we have a weekly holding period, with five days to a week, each weekly P/L will be the sum of five daily P/Ls, and we end up with only 200 observations of weekly P/L; if we have a monthly holding period, we have only 50 observations of monthly P/L; and so on. Given our initial data, the number of effective observations rapidly falls as the holding period rises, and the size of the data set imposes a major constraint on how large the holding period can practically be. In any case, even if we had a very long run of data, the older observations might have very little relevance for current market conditions.

2.3 ESTIMATING CONFIDENCE INTERVALS FOR HISTORICAL SIMULATION VaR AND ES

The methods considered so far are good for giving point estimates of VaR or ES, but they don't give us any indication of the precision of these estimates or any indication of VaR or ES

confidence intervals. However, there are methods to get around this limitation and produce confidence intervals for our risk estimates.[5]

An Order Statistics Approach to the Estimation of Confidence Intervals for HS VaR and ES

One of the most promising methods is to apply the theory of order statistics, explained in Appendix 1 to this chapter. This approach gives us, not just a VaR (or ES) estimate, but a complete VaR (or ES) distribution function from which we can read off the VaR (or ES) confidence interval. (The central tendency parameters (mean, mode, median) also give us alternative point estimates of our VaR or ES, if we want them.) This approach is (relatively) easy to programme and very general in its application.

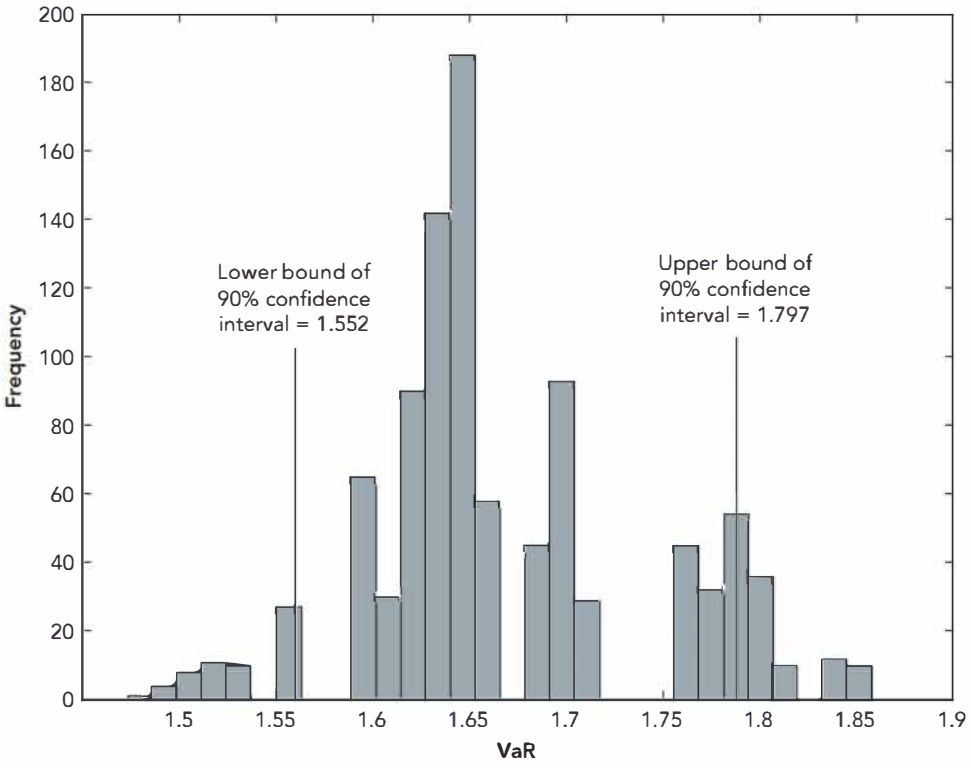


Figure 2.4 Bootstrapped VaR.

Note: Results obtained using the 'bootstrapvarfigure' function with 1000 resamples, and the same hypothetical data as in earlier figures.

Applied to our earlier P/L data, the OS approach gives us estimates (obtained using the 'hsvarpdfperc' function) of the 5% and 95% points of the 95% VaR distribution function—that is, the bounds of the 90% confidence interval for our VaR—of 1.552 and 1.797. This tells us we can be 90% confident that the 'true' VaR lies in the range [1.552, 1.797].

The corresponding points of the ES distribution function can be obtained (using the 'hsesdfperc' function) by mapping from the VaR to the ES: we take a point on the VaR distribution function, and estimate the corresponding percentile point on the ES distribution function. Doing this gives us an estimated 90% confidence interval of [2.021, 2.224].[6]

A Bootstrap Approach to the Estimation of Confidence Intervals for HS VaR and ES

We can also estimate confidence intervals using a bootstrap approach: we produce a bootstrapped histogram of resample-based VaR (or ES) estimates, and then read the confidence interval from the quantiles of this histogram. For example, if we take 1000 bootstrapped samples from our P/L data set, estimate the 95% VaR of each, and then plot them, we get the histogram shown in Figure 2.4. Using the basic percentile interval approach outlined in Appendix 2 to this chapter, the 90% confidence interval for our VaR is [1.554, 1.797]. The simulated histogram is surprisingly disjointed, although the bootstrap seems to give a relatively robust estimate of the confidence interval if we keep repeating the exercise.

We can also use the bootstrap to estimate ESs in much the same way: for each new resampled data set, we estimate the VaR, and then estimate the ES as the average of losses in excess of VaR. Doing this a large number of times gives us a large number of ES estimates, and we can plot them in the same way as the VaR estimates. The histogram of bootstrapped ES values is shown in Figure 2.5, and is better

[5] In addition to the methods considered in this section, we can also estimate confidence intervals for VaR using estimates of the quantile standard errors. However, as made clear there, such confidence intervals are subject to a number of problems, and the methods suggested here are usually preferable.

[6] Naturally, the order statistics approach can be combined with more sophisticated non-parametric density estimation approaches. Instead of applying the OS theory to the histogram or naive estimator, we could apply it to a more sophisticated kernel estimator, and thereby extract more information from our data. This approach has some merit and is developed in detail by Butler and Schachter (1998).

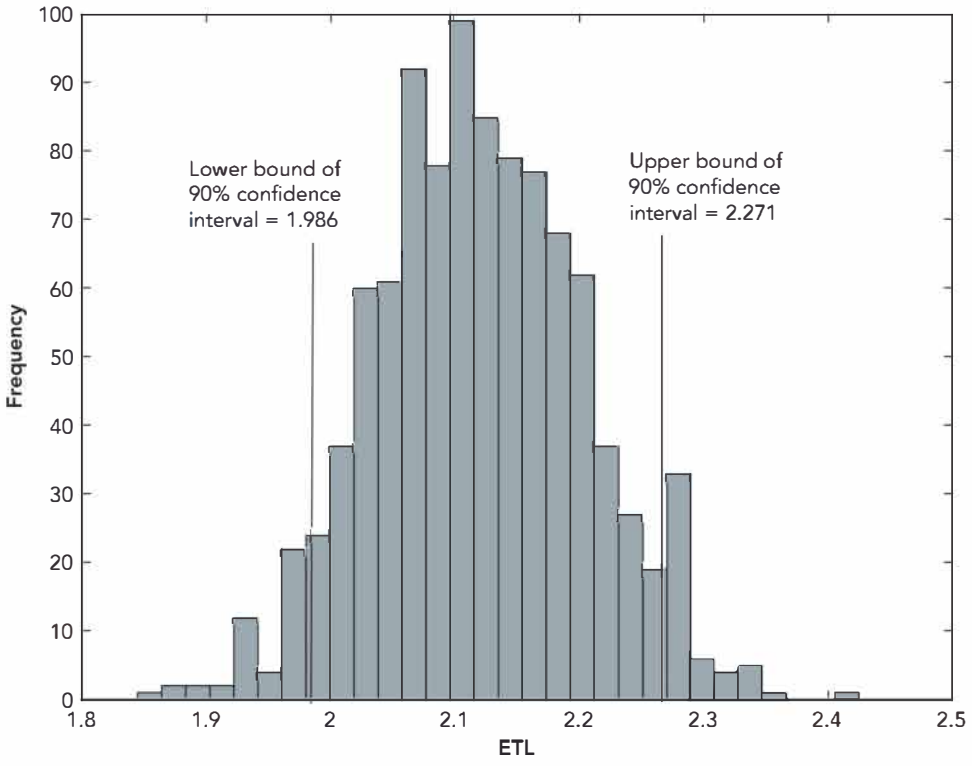


Figure 2.5 Bootstrapped ES.

Note: Results obtained using the 'bootstrapesfigure' function with 1000 resamples, and the same hypothetical data as in earlier figures.

Table 2.1 90% Confidence Intervals for Non-parametric VaR and ES

Approach	Lower bound	Upper bound
	95% VaR	
Order statistics	1.552	1.797
Bootstrap	1.554	1.797
	95% ES	
Order statistics	2.021	2.224
Bootstrap	1.986	2.271

Note: Bootstrap estimates based on 1000 resamples.

behaved than the VaR histogram in the last figure because the ES is an average of tail VaRs. The 90% confidence interval for our ES is [1.986, 2.271].

It is also interesting to compare the VaR and ES confidence intervals obtained by the two methods. These are summarised in Table 2.1, and we can see that the OS and bootstrap approaches give very similar results. This suggests that either approach is likely to be a reasonable one to use in practice.

2.4 WEIGHTED HISTORICAL SIMULATION

One of the most important features of traditional HS is the way it weights past observations. Recall that $R_{i,t}$ is the return on asset i in period t, and we are implementing HS using the past n observations. An observation $R_{i,t-j}$ will therefore belong to our data set if j takes any of the values $1, \ldots, t-n$, where j is the age of the observation (e.g., so $j = 1$ indicates that the observation is 1 day old, and so on). If we construct a new HS P/L series, P/L_t, each day, our observation $R_{i,t-j}$ will first affect P/L_t, then affect P/L_{t+1}, and so on, and finally affect P/L_{1+n}: our return observation will affect each of the next n observations in our P/L series. Also, other things (e.g., position weights) being equal, $R_{i,t-j}$ will affect each P/L in exactly the same way. But after n periods have passed, $R_{i,t-j}$ will fall out of the data set used to calculate the current HS P/L series, and will thereafter have no effect on P/L. In short, our HS P/L series is constructed in a way that gives any observation the same weight on P/L provided it is less than n periods old, and *no* weight (i.e., a zero weight) if it is older than that.

This weighting structure has a number of problems. One problem is that it is hard to justify giving each observation in our sample period the same weight, regardless of age, market volatility, or anything else. A good example of the difficulties this can create is given by Shimko et. al. (1998). It is well known that natural gas prices are usually more volatile in the winter than in the summer, so a raw HS approach that incorporates both summer and winter observations will tend to average the summer and winter observations together. As a result, treating all observations as having equal weight will tend to underestimate true risks in the winter, and overestimate them in the summer.[7]

The equal-weight approach can also make risk estimates unresponsive to major events. For instance, a stock market crash

[7] If we have data that show seasonal volatility changes, a solution—suggested by Shimko et. al. (1998)—is to weight the data to reflect seasonal volatility (e.g., so winter observations get more weight, if we are estimating a VaR in winter).

might have no effect on VaRs except at a very high confidence level, so we could have a situation where everyone might agree that risk had suddenly increased, and yet that increase in risk would be missed by most HS VaR estimates. The increase in risk would only show up later in VaR estimates if the stock market continued to fall in subsequent days—a case of the stable door closing only well after the horse had long since bolted. That said, the increase in risk *would* show up in ES estimates just after the first shock occurred—which is, incidentally, a good example of how ES can be a more informative risk measure than the VaR.[8]

The equal-weight structure also presumes that each observation in the sample period is equally likely and independent of the others over time. However, this 'iid' assumption is unrealistic because it is well known that volatilities vary over time, and that periods of high and low volatility tend to be clustered together. The natural gas example just considered is a good case in point.

It is also hard to justify why an observation should have a weight that suddenly goes to zero when it reaches age n. Why is it that an observation of age $n - 1$ is regarded as having a lot of value (and, indeed, the same value as any more recent observation), but an observation of age n is regarded as having no value at all? Even old observations usually have some information content, and giving them zero value tends to violate the old statistical adage that we should never throw information away.

This weighting structure also creates the potential for ghost effects—we can have a VaR that is unduly high (or low) because of a small cluster of high loss observations, or even just a single high loss, and the measured VaR will continue to be high (or low) until n days or so have passed and the observation has fallen out of the sample period. At that point, the VaR will fall again, but the fall in VaR is only a ghost effect created by the weighting structure and the length of sample period used.

We now address various ways in which we might 'adjust' our data to overcome some of these problems and take account of ways in which current market conditions might differ from those in our sample. These fall under the broad heading of 'weighted historical simulation' and can be regarded as semi-parametric methods because they combine features of both parametric and non-parametric methods.

Age-weighted Historical Simulation

One such approach is to weight the relative importance, of our observations by their age, as suggested by Boudoukh, Richardson and Whitelaw (BRW: 1998). Instead of treating each observation for asset i as having the same implied probability as any other (i.e., $1/n$), we could weight their probabilities to discount the older observations in favour of newer ones. Thus, if w(1) is the probability weight given to an observation 1 day old, then w(2), the probability given to an observation 2 days old, could be λw(1); w(3) could be λ^2w(1); and so on. The λ term is between 0 and 1, and reflects the exponential rate of decay in the weight or value given to an observation as it ages: a λ close to 1 indicates a slow rate of decay, and a λ far away from 1 indicates a high rate of decay. w(1) is set so that the sum of the weights is 1, and this is achieved if we set $w(1) = (1 - \lambda)/(1 - \lambda^n)$. The weight given to an observation i days old is therefore:

$$w(i) = \frac{\lambda^{i-1}(1 - \lambda)}{1 - \lambda^n} \tag{2.2}$$

and this corresponds to the weight of $1/n$ given to any in-sample observation under basic HS.

Our core information—the information inputted to the HS estimation process—is the paired set of P/L values and associated probability weights. To implement age-weighting, we merely replace the old equal weights 1/n with the age-dependent weights *w(i)* given by (2.4). For example, if we are using a spreadsheet, we can order our P/L observations in one column, put their weights *w(i)* in the next column, and go down that column until we reach our desired percentile. Our VaR is then the negative of the corresponding value in the first column. And if our desired percentile falls between two percentiles, we can take our VaR to be the (negative of the) interpolated value of the corresponding first-column observations.

This age-weighted approach has four major attractions. First, it provides a nice generalisation of traditional HS, because we can regard traditional HS as a special case with zero decay, or $\lambda \rightarrow 1$. If HS is like driving along a road looking only at the rear-view mirror, then traditional equal-weighted HS is only safe if the road is straight, and the age-weighted approach is safe if the road bends gently.

Second, a suitable choice of λ can make the VaR (or ES) estimates more responsive to large loss observations: a large loss event will receive a higher weight than under traditional HS, and

[8] However, both VaR and ES suffer from a related problem. As Pritsker (2001, p. 5) points out, HS fails to take account of useful information from the upper tail of the P/L distribution. If the stock experiences a series of large falls, then a position that was long the market would experience large losses that should show up, albeit later, in HS risk estimates. However, a position that was short the market would experience a series of large profits, and risk estimates at the usual confidence levels would be completely unresponsive. Once again, we could have a situation where risk had clearly increased—because the fall in the market signifies increased volatility, and therefore a significant chance of losses due to large rises in the stock market—and yet our risk estimates had failed to pick up this increase in risk.

the resulting next-day VaR would be higher than it would otherwise have been. This not only means that age-weighted VaR estimates are more responsive to large losses, but also makes them better at handling clusters of large losses.

Third, age-weighting helps to reduce distortions caused by events that are unlikely to recur, and helps to reduce ghost effects. As an observation ages, its probability weight gradually falls and its influence diminishes gradually over time. Furthermore, when it finally falls out of the sample period, its weight will fall from $\lambda^n w(1)$ to zero, instead of from $1/n$ to zero. Since $\lambda^n w(1)$ is less than $1/n$ for any reasonable values of λ and n, then the shock—the ghost effect—will be less than it would be under equal-weighted HS.

Finally, we can also modify age-weighting in a way that makes our risk estimates more efficient and effectively eliminates any remaining ghost effects. Since age-weighting allows the impact of past extreme events to decline as past events recede in time, it gives us the option of letting our sample size grow over time. (Why can't we do this under equal-weighted HS? Because we would be stuck with ancient observations whose information content was assumed never to date.) Age-weighting allows us to let our sample period grow with each new observation, so we never throw potentially valuable information away. This would improve efficiency and eliminate ghost effects, because there would no longer be any 'jumps' in our sample resulting from old observations being thrown away.

However, age-weighting also reduces the effective sample size, other things being equal, and a sequence of major profits or losses can produce major distortions in its implied risk profile. In addition, Pritsker shows that even with age-weighting, VaR estimates can still be insufficiently responsive to changes in underlying risk.[9] Furthermore, there is the disturbing point that the BRW approach is ad hoc, and that except for the special case where $\lambda = 1$ we cannot point to any asset-return process for which the BRW approach is theoretically correct.

Volatility-weighted Historical Simulation

We can also weight our data by volatility. The basic idea—suggested by Hull and White (HW; 1998b)—is to update return information to take account of recent changes in volatility. For example, if the current volatility in a market is 1.5% a day, and it was only 1% a day a month ago, then data a month old understate the changes we can expect to see tomorrow, and this suggests that historical returns would underestimate tomorrow's risks; on the other hand, if last month's volatility was 2% a day, month-old data will overstate the changes we can expect tomorrow, and historical returns would overestimate tomorrow's risks. We therefore adjust the historical returns to reflect how volatility tomorrow is believed to have changed from its past values.

Suppose we are interested in forecasting VaR for day T. Let $r_{t,i}$ be the historical return in asset i on day t in our historical sample, $\sigma_{t,i}$ be the historical GARCH (or EWMA) forecast of the volatility of the return on asset i for day t, made at the end of day $t-1$, and $\sigma_{T,i}$ be our most recent forecast of the volatility of asset i. We then replace the returns in our data set, $r_{t,i}$, with volatility-adjusted returns, given by:

$$r_{t,i}^{*} = \left(\frac{\sigma_{T,i}}{\sigma_{t,i}}\right) r_{t,i} \tag{2.3}$$

Actual returns in any period t are therefore increased (or decreased), depending on whether the current forecast of volatility is greater (or less than) the estimated volatility for period t. We now calculate the HS P/L using Equation (2.3) instead of the original data set $r_{t,i}$, and then proceed to estimate HS VaRs or ESs in the traditional way (i.e., with equal weights, etc.).[10]

The HW approach has a number of advantages relative to the traditional equal-weighted and/or the BRW age-weighted approaches:

- It takes account of volatility changes in a natural and direct way, whereas equal-weighted HS ignores volatility changes and the age-weighted approach treats volatility changes in a rather arbitrary and restrictive way.
- It produces risk estimates that are appropriately sensitive to current volatility estimates, and so enables us to incorporate information from GARCH forecasts into HS VaR and ES estimation.
- It allows us to obtain VaR and ES estimates that can exceed the maximum loss in our historical data set: in periods of high volatility, historical returns are scaled upwards, and the HS P/L series used in the HW procedure will have values that exceed actual historical losses. This is a major advantage over traditional HS, which prevents the VaR or ES from being any bigger than the losses in our historical data set.
- Empirical evidence presented by HW indicates that their approach produces superior VaR estimates to the BRW one.

[9] If VaR is estimated at the confidence level α, the probability of an HS estimate of VaR rising on any given day is equal to the probability of a loss in excess of VaR, which is of course $1-\alpha$. However, if we assume a standard GARCH(1,1) process and volatility is at its long-run mean value, then Pritsker's proposition 2 shows that the probability that HSVaR should increase is about 32% (Pritsker (2001, pp. 7–9)). In other words, most of the time HS VaR estimates should increase (i.e., when risk rises), they fail to.

[10] Naturally, volatility weighting presupposes that one has estimates of the current and past volatilities to work with.

The HW approach is also capable of various extensions. For instance, we can combine it with the age-weighted approach if we wished to increase the sensitivity of risk estimates to large losses, and to reduce the potential for distortions and ghost effects. We can also combine the HW approach with OS or bootstrap methods to estimate confidence intervals for our VaR or ES—that is, we would work with order statistics or resample with replacement from the HW-adjusted P/L, rather than from the traditional HS P/L.

Correlation-weighted Historical Simulation

We can also adjust our historical returns to reflect changes between historical and current correlations. Correlation-weighting is a little more involved than volatility-weighting. To see the principles involved, suppose for the sake of argument that we have already made any volatility-based adjustments to our HS returns along Hull-White lines, but also wish to adjust those returns to reflect changes in correlations.[11]

To make the discussion concrete, we have m positions and our (perhaps volatility adjusted) 1 × m vector of historical returns **R** for some period *t* reflects an m × m variance-covariance matrix Σ. Σ in turn can be decomposed into the product $\sigma\mathbf{C}\sigma^{\mathbf{T}}$, where σ is an m × m diagonal matrix of volatilities (i.e., so the *i*th element of σ is the *i*th volatility σ_i and the off-diagonal elements are zero), $\sigma^{\mathbf{T}}$ is its transpose, and **C** is the m × m matrix of historical correlations. **R** therefore reflects an historical correlation matrix **C**, and we wish to adjust **R** so that they become $\bar{\mathbf{R}}$ reflecting a current correlation matrix $\bar{\mathbf{C}}$. Now suppose for convenience that both correlation matrices are positive definite. This means that each correlation matrix has an m × m 'matrix square root', **A** and $\bar{\mathbf{A}}$ respectively, given by a Choleski decomposition (which also implies that they are easy to obtain). We can now write **R** and $\bar{\mathbf{R}}$ as matrix products of the relevant Choleski matrices and an uncorrelated noise process ε:

$$\mathbf{R} = \mathbf{A}\varepsilon \tag{2.4a}$$

$$\bar{\mathbf{R}} = \bar{\mathbf{A}}\varepsilon \tag{2.4b}$$

We then invert Equation (2.4a) to obtain $\varepsilon = \mathbf{A}^{-1}\mathbf{R}$, and substitute this into (Equation 2.4b) to obtain the correlation-adjusted series $\bar{\mathbf{R}}$ that we are seeking:

$$\bar{\mathbf{R}} = \bar{\mathbf{A}}\mathbf{A}^{-1}\mathbf{R} \tag{2.5}$$

The returns adjusted in this way will then have the currently prevailing correlation matrix **C** and, more generally, the currently prevailing covariance matrix $\bar{\Sigma}$.This approach is a major generalisation of the HW approach, because it gives us a weighting system that takes account of correlations as well as volatilities.

Example 2.1 Correlation-weighted HS

Suppose we have only two positions in our portfolio, so m = 2. The historical correlation between our two positions is 0.3, and we wish to adjust our historical returns **R** to reflect a current correlation of 0.9.

If a_{ij} is the *i*, *j*th element of the 2 × 2 matrix **A**, then applying the Choleski decomposition tells us that

$$a_{11} = 1, \quad a_{12} = 0, \quad a_{21} = \rho, \quad a_{22} = \sqrt{1 - \rho^2}$$

where $\rho = 0.3$. The matrix $\bar{\mathbf{A}}$ is similar except for having $\rho = 0.9$. Standard matrix theory also tells us that

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}}\begin{bmatrix} a_{22}, -a_{12} \\ -a_{21}, a_{11} \end{bmatrix}$$

Substituting these into Equation (2.5), we find that

$$\bar{\mathbf{R}} = \bar{\mathbf{A}}\mathbf{A}^{-1}\mathbf{R} = \begin{bmatrix} 1, 0 \\ 0.9, \sqrt{1 - 0.9^2} \end{bmatrix} \frac{1}{\sqrt{1 - 0.3^2}} \begin{bmatrix} \sqrt{1 - 0.3^2}, 0 \\ -0.3, 1 \end{bmatrix}^{\mathbf{R}}$$

$$= \frac{1}{\sqrt{1 - 0.3^2}} \begin{bmatrix} \sqrt{1 - 0.3^2} \\ 0.9\sqrt{1 - 0.3^2} - 0.3\sqrt{1 - 0.9^2}, \sqrt{1 - 0.9^2} \end{bmatrix}^{\mathbf{R}}$$

$$= \begin{bmatrix} 1, 0 \\ 0.7629, 0.4569 \end{bmatrix}^{\mathbf{R}}$$

Filtered Historical Simulation

Another promising approach is filtered historical simulation (FHS).[12] This is a form of semi-parametric bootstrap which aims to combine the benefits of HS with the power and flexibility of conditional volatility models such as GARCH. It does so by bootstrapping returns within a conditional volatility (e.g., GARCH) framework, where the bootstrap preserves the non-parametric nature of HS, and the volatility model gives us a sophisticated treatment of volatility.

Suppose we wish to use FHS to estimate the VaR of a single-asset portfolio over a 1-day holding period. The first step in FHS is to fit, say, a GARCH model to our portfolio-return data. We want a model that is rich enough to accommodate the key

[11] The correlation adjustment discussed here is based on a suggestion by Duffie and Pan (1997).

[12] This approach is suggested in Barone-Adesi et. al. (1998), Barone-Adesi et. al. (1999), Barone-Adesi aud Giannopoulos (2000) and in other papers by some of the same authors.

features of our data, and Barone-Adesi and colleagues recommend an asymmetric GARCH, or AGARCH, model. This not only accommodates conditionally changing volatility, volatility clustering, and so on, but also allows positive and negative returns to have differential impacts on volatility, a phenomenon known as the leverage effect. The AGARCH postulates that portfolio returns obey the following process:

$$r_t = \mu + \varepsilon_t \tag{2.6a}$$

$$\sigma_t^2 = \omega + \alpha(\varepsilon_{t-1} + \gamma)^2 + \beta\sigma_{t-1}^2 \tag{2.6b}$$

The daily return in Equation (2.6a) is the sum of a mean daily return (which can often be neglected in volatility estimation) and a random error ε_t. The volatility in Equation (2.6b) is the sum of a constant and terms reflecting last period's 'surprise' and last period's volatility, plus an additional term γ that allows for the surprise to have an asymmetric effect on volatility, depending on whether the surprise term is positive or negative.

The second step is to use the model to forecast volatility for each of the days in a sample period. These volatility forecasts are then divided into the realised returns to produce a set of standardised returns. These standardised returns should be independently and identically distributed (iid), and therefore be suitable for HS.

Assuming a 1-day VaR holding period, the third stage involves bootstrapping from our data set of standardised returns: we take a large number of drawings from this data set, which we now treat as a sample, replacing each one after it has been drawn, and multiply each random drawing by the AGARCH forecast of tomorrow's volatility. If we take *M* drawings, we therefore get *M* simulated returns, each of which reflects current market conditions because it is scaled by today's forecast of tomorrow's volatility.

Finally, each of these simulated returns gives us a possible end-of-tomorrow portfolio value, and a corresponding possible loss, and we take the VaR to be the loss corresponding to our chosen confidence level.[13]

We can easily modify this procedure to encompass the obvious complications of a multi asset portfolio or a longer holding period. If we have a multi-asset portfolio, we would fit a multivariate GARCH (or AGARCH) to the set or vector of asset returns, and we would standardise this vector of asset returns. The bootstrap would then select, not just a standardised portfolio return for some chosen past (daily) period, but the standardised vector of asset returns for the chosen past period. This is important because it means that our simulations would keep any correlation structure present in the raw returns. The bootstrap thus maintains existing correlations, without our having to specify an explicit multivariate pdf for asset returns.

The other obvious extension is to a longer holding period. If we have a longer holding period, we would first take a drawing and use Equation (2.6) to get a return for tomorrow; we would then use this drawing to update our volatility forecast for the day after tomorrow, and take a fresh drawing to determine the return for that day; and we would carry on in the same manner—taking a drawing, updating our volatility forecasts, taking another drawing for the next period, and so on—until we had reached the end of our holding period. At that point we would have enough information to produce a single simulated P/L observation; and we would repeat the process as many times as we wished in order to produce the histogram of simulated P/L observations from which we can estimate our VaR.

FHS has a number of attractions: (i) It enables us to combine the non-parametric attractions of HS with a sophisticated (e.g., GARCH) treatment of volatility, and so take account of changing market volatility conditions. (ii) It is fast, even for large portfolios. (iii) As with the earlier HW approach, FHS allows us to get VaR and ES estimates that can exceed the maximum historical loss in our data set. (iv) It maintains the correlation structure in our return data without relying on knowledge of the variance-covariance matrix or the conditional distribution of asset returns. (v) It can be modified to take account of autocorrelation or past cross-correlations in asset returns. (vi) It can be modified to produce estimates of VaR or ES confidence intervals by combining it with an OS or bootstrap approach to confidence interval estimation.[14] (vii) There is evidence that FHS works well.[15]

[13] The FHS approach can also be extended easily to allow for the estimation of ES as well as VaR. For more on how this might be done, see Giannopoulos and Tunaru (2004).

[14] The OS approach would require a set of paired P/L and associated probability observations, so we could apply this to FHS by using a P/L series that had been through the FHS filter. The bootstrap is even easier, since FHS already makes use of a bootstrap. If we want *B* bootstrapped estimates of VaR, we could produce, say, 100**B* or 1000**B* bootstrapped P/L values; each set of 100 (or 1000) P/L series would give us one HS VaR estimate, and the histogram of *M* such estimates would enable us to infer the bounds of the VaR confidence interval.

[15] Barone-Adesi and Giannopoulos (2000), p. 17. However, FHS does have problems. In his thorough simulation study of FHS, Pritsker (2001, pp. 22–24) comes to the tentative conclusions that FHS VaR might not pay enough attention to extreme observations or time-varying correlations, and Barone-Adesi and Giannopoulos (2000, p. 18) largely accept these points. A partial response to the first point would be to use ES instead of VaR as our preferred risk measure, and the natural response to the second concern is to develop FHS with a more sophisticated past cross-correlation structure. Pritsker (2001, p. 22) also presents simulation results that suggest that FHS-VaR tends to underestimate 'true' VaR over a 10-day holding period by about 10%, but this finding conflicts with results reported by Barone-Adesi et. al. (2000) based on real data. The evidence on FHS is thus mixed.

2.5 ADVANTAGES AND DISADVANTAGES OF NON-PARAMETRIC METHODS

Advantages

In drawing our discussion to a close, it is perhaps a good idea to summarise the main advantages and disadvantages of non-parametric approaches. The advantages include:

- Non-parametric approaches are intuitive and conceptually simple.
- Since they do not depend on parametric assumptions about P/L, they can accommodate fat tails, skewness, and any other non-normal features that can cause problems for parametric approaches.
- They can in theory accommodate any type of position, including derivatives positions.
- There is a widespread perception among risk practitioners that HS works quite well empirically, although formal empirical evidence on this issue is inevitably mixed.
- They are (in varying degrees, fairly) easy to implement on a spreadsheet.
- Non-parametric methods are free of the operational problems to which parametric methods are subject when applied to high-dimensional problems: no need for covariance matrices, no curses of dimensionality, etc.
- They use data that are (often) readily available, either from public sources (e.g., Bloomberg) or from in-house data sets (e.g., collected as a by-product of marking positions to market).
- They provide results that are easy to report and communicate to senior managers and interested outsiders (e.g., bank supervisors or rating agencies).
- It is easy to produce confidence intervals for non-parametric VaR and ES.
- Non-parametric approaches are capable of considerable refinement and potential improvement if we combine them with parametric 'add-ons' to make them semi-parametric: such refinements include age-weighting (as in BRW), volatility-weighting (as in HW and FHS), and correlation-weighting.

Disadvantages

Perhaps their biggest potential weakness is that their results are very (and in most cases, completely) dependent on the historical data set.[16] There are various other related problems:

- If our data period was unusually quiet, non-parametric methods will often produce VaR or ES estimates that are too low for the risks we are actually facing; and if our data period was unusually volatile, they will often produce VaR or ES estimates that are too high.
- Non-parametric approaches can have difficulty handling shifts that take place during our sample period. For example, if there is a permanent change in exchange rate risk, it will usually take time for the HS VaR or ES estimates to reflect the new exchange rate risk. Similarly, such approaches are sometimes slow to reflect major events, such as the increases in risk associated with sudden market turbulence.
- If our data set incorporates extreme losses that are unlikely to recur, these losses can dominate non-parametric risk estimates even though we don't expect them to recur.
- Most (if not all) non-parametric methods are subject (to a greater or lesser extent) to the phenomenon of ghost or shadow effects.
- In general, non-parametric estimates of VaR or ES make no allowance for plausible events that might occur, but did not actually occur, in our sample period.
- Non-parametric estimates of VaR and ES are to a greater or lesser extent constrained by the largest loss in our historical data set. In the simpler versions of HS, we cannot extrapolate from the largest historical loss to anything larger that might conceivably occur in the future. More sophisticated versions of HS can relax this constraint, but even so, the fact remains that non-parametric estimates of VaR or ES are still constrained by the largest loss in a way that parametric estimates are not. This means that such methods are not well suited to handling extremes, particularly with small- or medium-sized samples.

However, we can often ameliorate these problems by suitable refinements. For example, we can ameliorate volatility, market turbulence, correlation and other problems by semi-parametric adjustments, and we can ameliorate ghost effects by age-weighting our data and allowing our sample size to rise over time.

There can also be problems associated with the length of the sample window period. We need a reasonably long window

[16] There can also be problems getting the data set. We need time series data on all current positions, and such data are not always available (e.g., if the positions are in emerging markets). We also have to ensure that data are reliable, compatible, and delivered to the risk estimation system on a timely basis.

to have a sample size large enough to get risk estimates of acceptable precision, and as a broad rule of thumb, most experts believe that we usually need at least a couple of year's worth of daily observations (i.e., 500 observations, at 250 trading days to the year), and often more. On the other hand, a very long window can also create its own problems. The longer the window:

- the greater the problems with aged data;
- the longer the period over which results will be distorted by unlikely-to-recur past events, and the longer we will have to wait for ghost effects to disappear;
- the more the news in current market observations is likely to be drowned out by older observations—and the less responsive will be our risk estimates to current market conditions; and
- the greater the potential for data-collection problems. This is a particular concern with new or emerging market instruments, where long runs of historical data don't exist and are not necessarily easy to proxy.

CONCLUSIONS

Non-parametric methods are widely used and in many respects highly attractive approaches to the estimation of financial risk measures. They have a reasonable track record and are often superior to parametric approaches based on simplistic assumptions such as normality. They are also capable of considerable refinement to deal with some of the weaknesses of more basic non-parametric approaches. As a general rule, they work fairly well if market conditions remain reasonably stable, and are capable of considerable refinement. However, they have their limitations and it is often a good idea to supplement them with other approaches. Wherever possible, we should also complement non-parametric methods with stress testing to gauge our vulnerability to 'what if' events. We should never rely on non-parametric methods alone.

APPENDIX 1

Estimating Risk Measures with Order Statistics

The theory of order statistics is very useful for risk measurement because it gives us a practical and accurate means of estimating the distribution function for a risk measure—and this is useful because it enables us to estimate confidence intervals for them.

Using Order Statistics to Estimate Confidence Intervals for VaR

If we have a sample of n P/L observations, we can regard each observation as giving an estimate of VaR at an implied confidence level. For example, if $n = 1000$, we might take the 95% VaR as the negative of the 51st smallest P/L observation, we might take the 99% VaR as the negative of the 11th smallest, and so on. We therefore take the α VaR to be equal to the negative of the rth lowest observation, where r is equal to $100(1 - \alpha) + 1$. More generally, with n observations, we take the VaR as equal to the negative of the rth lowest observation, where $r = n(1 - \alpha) + 1$.

The rth order statistic is the rth lowest (or, alternatively, highest) in a sample of n observations, and the theory of order statistics is well established in the statistical literature. Suppose our observations $x_1, x_2, \ldots, x_n$ come from some known distribution (or cumulative density) function $F(x)$, with rth order statistic $x_{(r)}$. Now suppose that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. The probability that j of our n observations do not exceed a fixed value x must obey the following binomial distribution:

$$\Pr\{j \text{ observations} \leq x\} = \binom{n}{j}\{F(x)\}^j\{1 - F(x)\}^{n-j} \tag{2.7}$$

It follows that the probability that at least r observations in the sample do not exceed x is also a binomial:

$$G_r(x) = \sum_{j=r}^{n}\binom{n}{j}\{F(x)\}^j\{1 - F(x)\}^{n-j} \tag{2.8}$$

$G_r(x)$ is therefore the distribution function of our order statistic and, hence, of our quantile or VaR.[17]

This VaR distribution function provides us with estimates of our VaR *and* of its associated confidence intervals. The median (i.e., 50 percentile) of the estimated VaR distribution function gives us a natural 'best' estimate of our VaR, and estimates of the lower and upper percentiles of the VaR distribution function give us estimates of the bounds of our VaR confidence interval. This is useful, because the calculations are accurate and easy to carry out on a spreadsheet. Equation (2.8) is also very general and gives us confidence intervals for *any* distribution function $F(x)$, parametric (normal, t, etc.) or empirical.

To use this approach, all we need to do is specify $F(x)$ (as normal, t, etc.), set our parameter values, and use Equation (2.8) to estimate our VaR distribution function.

To illustrate, suppose we want to apply the order-statistics (OS) approach to estimate the distribution function of a standard

[17] See, e.g., Kendall and Stuart (1973), p. 348, or Reiss (1989), p. 20.

normal VaR. We then assume that $F(x)$ is standard normal and use Equation (2.8) to estimate three key parameters of the VaR distribution: the median or 50 percentile of the estimated VaR distribution, which can be interpreted as an OS estimate of normal VaR; and the 5 and 95 percentiles of the estimated VaR distribution, which can be interpreted as the OS estimates of the bounds of the 90% confidence interval for standard normal VaR.

Some illustrative estimates for the 95% VaR are given in Table 2.2. To facilitate comparison, the table also shows the estimates of standard normal VaR based on the conventional normal VaR formula as explained in Chapter 1. The main results are:

- The confidence interval—the gap between the 5 and 95 percentiles—is quite wide for low values of n, but narrows as n gets larger.
- As n rises, the median of the estimated VaR distribution converges to the conventional estimate.
- The confidence interval is (in this case, a little) wider for more extreme VaR confidence levels than it is for the more central ones.

The same approach can also be used to estimate the percentiles of other VaR distribution functions. If we wish to estimate the percentiles of a non-normal parametric VaR, we replace the normal distribution function $F(x)$ by the non-normal equivalent—the t-distribution function, the Gumbel distribution function, and so on. We can also use the same approach to estimate the confidence intervals for an empirical distribution function (i.e., for historical simulation VaR), where $F(x)$ is some empirical distribution function.

Conclusions

The OS approach provides an ideal method for estimating the confidence intervals for our VaRs and ESs. In particular, the OS approach is:

- Completely general, in that it can be applied to any parametric or non-parametric VaR or ES.
- Reasonable even for relatively small samples, because it is not based on asymptotic theory—although it is also the case that estimates based on small samples will also be less accurate, precisely because the samples are small.
- Easy to implement in practice.

The OS approach is also superior to confidence-interval estimation methods based on estimates of quantile standard errors (see Chapter 1), because it does not rely on asymptotic theory and/or force estimated confidence intervals to be symmetric (which can be a problem for extreme VaRs and ESs).

Table 2.2 Order Statistics Estimates of Standard Normal 95% VaRs and Associated Confidence Intervals

(a) As n varies						
No. of observations	100	500	1000	5000		10 000
Lower bound of confidence interval	1.267	1.482	1.531	1.595		1.610
Median of VaR distribution	1.585	1.632	1.639	1.644		1.644
Standard estimate of VaR	1.645	1.645	1.645	1.645		1.645
Upper bound of confidence interval	1.936	1.791	1.750	1.693		1.679
Width of interval/median	42.2%	18.9%	13.4%	6.0%		4.2%
(b) As VaR confidence level varies (with $n = 500$)						
VaR confidence level	0.90		0.95		0.99	
Lower bound of confidence interval	1.151		1.482		2.035	
Median of VaR distribution	1.274		1.632		2.279	
Standard estimate of VaR	1.282		1.645		2.326	
Upper bound of confidence interval	1.402		1.791		2.560	
Width of interval/median of interval	19.7%		18.9%		23.0%	

Notes: The confidence interval is specified at a 90% level of confidence, and the lower and upper bounds of the confidence interval are estimated as the 5 and 95 percentiles of the estimated VaR distribution (Equation (2.8)).

APPENDIX 2

The Bootstrap

The bootstrap is a simple and useful method for assessing uncertainty in estimation procedures. Its distinctive feature is that it replaces mathematical or statistical analysis with simulation-based resampling from a given data set. It therefore provides a means of assessing the accuracy of parameter estimators without having to resort to strong parametric assumptions or closed-form confidence-interval formulas. The roots of the bootstrap go back a couple of centuries, but the idea only took off in the last three decades after it was developed and popularised by the work of Bradley Efron. It was Efron, too, who first gave it its name, which refers to the phrase 'to pull oneself up by one's bootstraps'. The bootstrap is a form of statistical 'trick', and is therefore very aptly named.

The main purpose of the bootstrap is to assess the accuracy of parameter estimates. The bootstrap is ideally suited for this purpose, as it can provide such estimates without having to rely on potentially unreliable assumptions (e.g., assumptions of normality or large samples).[18] The bootstrap is also easy to use because it does not require the user to engage in any difficult mathematical or statistical analysis. In any case, such traditional methods only work in a limited number of cases, whereas the bootstrap can be applied more or less universally. So the bootstrap is easier to use, more powerful and (as a rule) more reliable than traditional means of estimating confidence intervals for parameters of interest. In addition, the bootstrap can be used to provide alternative 'point' estimates of parameters as well.[19]

Limitations of Conventional Sampling Approaches

The bootstrap is best appreciated by considering the limitations of conventional sampling approaches. Suppose we have a sample of size *n* drawn from a population. The parameters of the population distribution are unknown—and, more likely than not, so too is the distribution itself. We are interested in a particular parameter θ, where θ might be a mean, variance (or standard deviation), quantile, or some other parameter. The obvious approach is to estimate θ using a suitable sample estimator—so if θ is the mean, our estimator $\hat{\theta}$ would be the sample mean, if θ is the variance, our estimator $\hat{\theta}$ would be based on some sample variance, and so on. Obtaining an estimator for θ is therefore straightforward, but how do we obtain a confidence interval for it?

To estimate confidence intervals for θ using traditional closed-form approaches requires us to resort to statistical theory, and the theory available is of limited use. For example, suppose we wish to obtain a confidence interval for a variance. If we assume that the underlying distribution is normal, then we know that $(n-1)\hat{\sigma}^2/\sigma^2$ is distributed as χ^2 with $n-1$ degrees of freedom, and this allows us to obtain a confidence interval for σ^2. If we denote the α point of this distribution as $\chi^2_{\alpha,n-1}$, then the 90% confidence interval for $(n-1)\hat{\sigma}^2/\sigma^2$ is:

$$[\chi^2_{0.05,n-1}, \chi^2_{0.95,n-1}] \tag{2.9}$$

This implies that the 90% confidence interval for σ^2 is:

$$\left[\frac{(n-1)\hat{\sigma}^2}{\chi_{0.95,n-1}}, \frac{(n-1)\hat{\sigma}^2}{\chi_{0.05,n-1}}\right] \tag{2.10}$$

On the other hand, if we cannot assume that the underlying distribution is normal, then obtaining a confidence interval for σ^2 can become very difficult: the problem is that although we can estimate σ^2 itself, under more general conditions we would often not know the distribution of σ^2, or have expressions for standard errors, and we cannot usually obtain closed-form confidence intervals without them.

We can face similar problems with other parameters as well, such as medians, correlations, and tail probabilities.[20] So in general, closed-form confidence intervals are of limited applicability, and will not apply to many of the situations we are likely to meet in practice.

The Bootstrap and Its Implementation

The bootstrap frees us of this type of limitation, and is also much easier to implement. It enables us to estimate a confidence interval for any parameter that we can estimate, regardless of whether we have any formulas for the distribution function for that parameter or for the standard error of its estimator. The bootstrap also has the advantage that it comes with less baggage, in the sense that the assumptions needed

[18] The bootstrap is also superior to the jackknife, which was often used for similar purposes before the advent of powerful computers. The jackknife is a procedure in which we construct a large number of subsamples from an original sample by taking the original sample and leaving one observation out at a time. For each such subsample, we estimate the parameter of interest, and the jackknife estimator is the average of the subsample-based estimators. The jackknife can also be regarded as an approximation to the bootstrap, but it can provide a very poor approximation when the parameter estimator is a non-smooth function of the data. The bootstrap is therefore more reliable and easier to implement.

[19] The bootstrap also has other uses too. For example, it can be used to relax and check assumptions, to give quick approximations and to check the results obtained using other methods.

[20] However, in the case of quantiles, we can use order statistics to write down their distribution functions.

to implement the bootstrap are generally less demanding than the assumptions needed to estimate confidence intervals using more traditional (i.e., closed-form) methods.

The basic bootstrap procedure is very simple.[21] We start with a given original sample of size n.[22] We now draw a new random sample of the same size from this original sample, taking care to replace each chosen observation back in the sample pool after it has been drawn. (This random sampling, or resampling, is the very heart of the bootstrap. It requires that we have a uniform random number generator to select a random number between 1 and n, which determines the particular observation that is chosen each time.) When constructing the new sample, known as a resample, we would typically find that some observations get chosen more than once, and others don't get chosen at all: so the resample would typically be different from the original one, even though every observation included in it was drawn from the original sample. Once we have our resample, we use it to estimate the parameter we are interested in. This gives us a resample estimate of the parameter. We then repeat the 'resampling' process again and again, and obtain a set of B resample parameter estimates. This set of B resample estimates can also be regarded as a bootstrapped sample of parameter estimates.

We can then use the bootstrapped sample to estimate a confidence interval for our parameter θ. For example, if each resample i gives us a resample estimator $\hat{\theta}^B(i)$ we might construct a simulated density function from the distribution of our $\hat{\theta}^B(i)$ values and infer the confidence intervals from its percentile points. If our confidence interval spans the central $1 - 2\alpha$ of the probability mass, then it is given by:

$$\text{Confidence Interval} = [\hat{\theta}^B_{\alpha}, \hat{\theta}^B_{1-\alpha}] \tag{2.11}$$

where $\hat{\theta}^B_{\alpha}$ is the α quantile of the distribution of bootstrapped $\hat{\theta}^B(i)$ values. This 'percentile interval' approach is very easy to apply and does not rely on any parametric theory, asymptotic or otherwise.

Nonetheless, this basic percentile interval approach is limited itself, particularly if parameter estimators are biased. It is therefore often better to use more refined percentile approaches, and perhaps the best of these is the bias-corrected and accelerated (or BC_a) approach, which generates a substantial improvement in both theory and practice over the basic percentile interval approach. To use this approach we replace the α and $1 - \alpha$ subscripts in Equation (2.11) with α_1 and α_2, where

$$\alpha_1 = \Phi\left(\hat{z}^0 + \frac{\hat{z}^0 + z_{\alpha}}{1 - \hat{a}(\hat{z}^0 + z_{\alpha})}\right), \quad \alpha_2 = \Phi\left(\hat{z}^0 + \frac{\hat{z}^0 + z_{1-\alpha}}{1 - \hat{a}(\hat{z}^0 + z_{1-\alpha})}\right) \tag{2.12}$$

If the parameters $\hat{a}$ and $\hat{z}^0$ are zero, this BC_a confidence interval will coincide with the earlier percentile interval. However, in general, they will not be 0, and we can think of the BC_a method as correcting the end-points of the confidence interval. The parameter $\hat{a}$ refers to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter $\boldsymbol{\theta}$, and it can be regarded as a correction for skewness. This parameter can be estimated from the following, which would be based on an initial bootstrap or jackknife exercise:

$$\hat{a} = \frac{\sum_{i=1}^{M}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^B(i))^3}{6\left[\sum_{i=1}^{M}(\hat{\boldsymbol{\theta}} - \hat{\theta}^B(i))^2\right]^{3/2}} \tag{2.13}$$

The parameter $\hat{z}^0$ can be estimated as the standard normal inverse of the proportion of bootstrap replications that is less than the original estimate $\hat{\boldsymbol{\theta}}$. The BC_a method is therefore (relatively) straightforward to implement, and it has the theoretical advantages over the percentile interval approach of being both more accurate and of being transformation-respecting, the latter property meaning that if we take a transformation of θ (e.g., if θ is a variance, we might wish to take its square root to obtain the standard deviation), then the BC_a method will automatically correct the end-points of the confidence interval of the transformed parameter.[23]

We can also use a bootstrapped sample of parameter estimates to provide an alternative point estimator of a parameter that is often superior to the raw sample estimator $\hat{\boldsymbol{\theta}}$. Given that there are B resample estimators, we can take our bootstrapped point estimator $\hat{\theta}^B$ as the sample mean of our B $\hat{\theta}^B(i)$ values:[24]

$$\hat{\theta}^B = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}^B(i) \tag{2.14}$$

Relatedly, we can also use a bootstrap to estimate the bias in an estimator. The bias is the difference between the expectation of an estimator and the quantity estimated (i.e., the bias equals

[21] This application of the bootstrap can be described as a non-parametric one because we bootstrap from a given data sample. The bootstrap can also be implemented parametrically, where we bootstrap from the assumed distribution. When used in parametric mode, the bootstrap provides more accurate answers than textbook formulas usually do, and it can provide answers to problems for which no textbook formulas exist. The bootstrap can also be implemented semi-parametrically and a good example of this is the FRS approach.

[22] In practice, it might be possible to choose the value of n, but we will assume for the sake of argument that n is given.

[23] For more on BC_a and other refinements to the percentile interval approach, see Efron and Tibshirani (1993, Chapters 14 and 22) or Davison and Hinkley (1997, Chapter 5).

[24] This basic bootstrap estimation method can also be supplemented by variance-reduction methods (e.g., importance sampling) to improve accuracy at a given computational cost. See Efron and Tibshirani (1993, Chapter 23) or Davison and Hinkley (1997, Chapter 9).

$E[\hat{\theta}] - \theta$)), and can be estimated by plugging Equation (2.14) and a basic sample estimator $\hat{\theta}$ into the bias equation:

$$Bias = E[\hat{\theta}] - \theta \Rightarrow Estimated\ Bias = \hat{\theta}^B - \hat{\theta} \tag{2.15}$$

We can use an estimate of bias for various purposes (e.g., to correct a biased estimator, to correct prediction errors, etc.). However, the bias can have a (relatively) large standard error. In such cases, correcting for the bias is not always a good idea, because the bias-corrected estimate can have a larger standard error than the unadjusted, biased, estimate.

The programs to compute bootstrap statistics are easy to write and the most obvious price of the bootstrap, increased computation, is no longer a serious problem.[25]

Standard Errors of Bootstrap Estimators

Naturally, bootstrap estimates are themselves subject to error. Typically, bootstrap estimates have little bias, but they often have substantial variance. The latter comes from basic sampling variability (i.e., the fact that we have a sample of size n drawn from our population, rather than the population itself) and from resampling variability (i.e., the fact that we take only B bootstrap resamples rather than an infinite number of them). The estimated standard error for $\hat{\theta}$, $\hat{s}_B$, can be obtained from:

$$\hat{s}_B = \left(\frac{1}{B} \sum_{i=1}^{B} (\hat{\theta}^B(i) - \hat{\theta}^B)^2 \right)^{1/2} \tag{2.16}$$

where $\hat{\theta}^B = (1/B)\sum_{i=1}^{B} \hat{\theta}^B(i)$. $\hat{s}_B$ is of course also easy to estimate. However, $\hat{s}_B$ is itself variable, and the variance of $\hat{s}_B$ is:

$$\text{var}(\hat{s}_B) = \text{var}[E(\hat{s}_B)] + E[\text{var}(\hat{s}_B)] \tag{2.17}$$

Following Efron and Tibshirani (1993, Chapter 19), this can be rearranged as:

$$\text{var}(\hat{s}_B) = \text{var}[\hat{m}_2^{1/2}] + E\left[\frac{\hat{m}_2}{4B} \left(\frac{\hat{m}_4}{\hat{m}_2^2} - 1 \right) \right] \tag{2.18}$$

where $\hat{m}_i$ is the ith moment of the bootstrap distribution of the $\hat{\theta}^B(i)$. In the case where θ is the mean, Equation (2.18) reduces to:

$$\text{var}(\hat{s}_B) = \frac{\hat{m}_4/\hat{m}_2 - \hat{m}_2}{4n^2} + \frac{\sigma^2}{2nB} + \frac{\sigma^2(\hat{m}_4/\hat{m}_2^2 - 3)}{4n^2B} \tag{2.19}$$

If the distribution is normal, this further reduces to:

$$\text{var}(\hat{s}_B) = \frac{\sigma^2}{2n^2}\left(1 + \frac{n}{B}\right) \tag{2.20}$$

We can then set B to reduce $\text{var}(\hat{s}_B)$ to a desired level, and so achieve a target level of accuracy in our estimate of $\hat{s}_B$. However, these results are limited, because Equation (2.19) only applies to the mean and Equation (2.20) presupposes normality as well.

We therefore face two related questions: (a) how we can estimate $\text{var}(\hat{s}_B)$ in general? and (b) how can we choose B to achieve a given level of accuracy in our estimate of $\hat{s}_B$? One approach to these problems is to apply brute force: we can estimate $\text{var}(\hat{s}_B)$ using a jackknife-after-bootstrap (in which we first bootstrap the data and then estimate $\text{var}(\hat{s}_B)$ by jackknifing from the bootstrapped data), or by using a double bootstrap (in which we estimate a sample of bootstrapped $\hat{s}_B$ values and then estimate their variance). We can then experiment with different values of B to determine the values of these parameters needed to bring $\text{var}(\hat{s}_B)$ down to an acceptable level.

If we are more concerned about the second problem (i.e., how to choose B), a more elegant approach is the following, suggested by Andrews and Buchinsky (1997). Suppose we take as our 'ideal' the value of $\hat{s}_B$ associated with an infinite number of resamples, i.e., $\hat{s}_\infty$. Let τ be a target probability that is close to 1, and let *bound* be a chosen bound on the percentage deviation of $s\sigma_B$ from s_∞. We want to choose $B = B(bound, \tau)$ such that the probability that $\hat{s}_B$ is within the desired bound is τ:

$$\Pr\left[100 \left| \frac{\hat{s}_B - \hat{s}_\infty}{\hat{s}_B} \right| \leq bound \right] = \tau \tag{2.21}$$

If B is large, then the required number of resamples is approximately

$$B \approx \frac{2500(\kappa - 1)\chi_\tau^2}{bound^2} \tag{2.22}$$

However, this formula is not operational because κ, the kurtosis of the distribution of $\hat{\theta}^B$, is unknown. To get around this problem, we replace κ with a consistent estimator of κ, and this leads Andrews and Buchinsky to suggest the following three-step method to determine B:

- We initially assume that $\kappa = 3$, and plug this into Equation (2.22) to obtain a preliminary value of B, denoted by B_0, where

$$B_0 = \text{int}\left(\frac{5000\chi_\tau^2}{bound^2} \right) \tag{2.23}$$

and where int(a) refers to the smallest integer greater than or equal to a.

- We simulate B_0 resamples, and estimate the sample kurtosis of the bootstrapped $\hat{\theta}^B$ values, $\hat{\kappa}$.
- We take the desired number of bootstrap resamples as equal to $\max(B_0, B_1)$, where

$$B_1 \approx \frac{2500(\hat{\kappa} - 1)\chi_\tau^2}{bound^2} \tag{2.24}$$

[25] An example of the bootstrap approach applied to VaR is given earlier in this chapter discussing the bootstrap point estimator and bootstrapped confidence intervals for VaR.

- This method does not directly tell us what the variance of $\hat{s}_B$ might be, but we already know how to estimate this in any case. Instead, this method gives us something more useful: it tells us how to set B to achieve a target level of precision in our bootstrap estimators, and (unlike Equations (2.19) and (2.20)) it applies for any parameter u and applies however $\hat{\theta}^B$ is distributed.[26]

Time Dependency and the Bootstrap

Perhaps the main limitation of the bootstrap is that standard bootstrap procedures presuppose that observations are independent over time, and they can be unreliable if this assumption does not hold. Fortunately, there are various ways in which we can modify bootstraps to allow for such dependence:

- If we are prepared to make parametric assumptions, we can model the dependence parametrically (e.g., using a GARCH procedure). We can then bootstrap from the residuals, which should be independent. However, this solution requires us to identify the underlying stochastic model and estimate its parameters, and this exposes us to model and parameter risk.
- An alternative is to use a block approach: we divide sample data into non-overlapping blocks of equal length, and select a block at random. However, this approach can 'whiten' the data (as the joint observations spanning different blocks are taken to be independent), which can undermine our results. On the other hand, there are also various methods of dealing with this problem (e.g., making block lengths stochastic, etc.) but these refinements also make the block approach more difficult to implement.
- A third solution is to modify the probabilities with which individual observations are chosen. Instead of assuming that each observation is chosen with the same probability, we can make the probabilities of selection dependent on the time indices of recently selected observations: so, for example, if the sample data are in chronological order and observation i has just been chosen, then observation $i + 1$ is more likely to be chosen next than most other observations.

[26] This three-step method can also be improved and extended. For example, it can be improved by correcting for bias in the kurtosis estimator, and a similar (although more involved) three-step method can be used to achieve given levels of accuracy in estimates of confidence intervals as well. For more on these refinements, see Andrews and Buchinsky (1997).

3

Parametric Approaches (II): Extreme Value

Learning Objectives

After completing this reading, you should be able to:

- Explain the importance and challenges of extreme values in risk management.
- Describe extreme value theory (EVT) and its use in risk management.
- Describe the peaks-over-threshold (POT) approach.
- Compare and contrast the generalized extreme value (GEV) and POT approaches to estimating extreme risks.
- Discuss the application of the generalized Pareto (GP) distribution in the POT approach.
- Explain the multivariate EVT for risk management.

Excerpt is Chapter 7 of Measuring Market Risk, *Second Edition, by Kevin Dowd.*

There are many problems in risk management that deal with extreme events—events that are unlikely to occur, but can be very costly when they do. These events are often referred to as low-probability, high-impact events, and they include large market falls, the failures of major institutions, the outbreak of financial crises and natural catastrophes. Given the importance of such events, the estimation of extreme risk measures is a key concern for risk managers.

However, to estimate such risks we have to confront a difficult problem: extreme events are rare by definition, so we have relatively few extreme observations on which to base our estimates. Estimates of extreme risks must therefore be very uncertain, and this uncertainty is especially pronounced if we are interested in extreme risks not only *within* the range of observed data, but *well beyond* it—as might be the case if we were interested in the risks associated with events more extreme than any in our historical data set (e.g., an unprecedented stock market fall).

Practitioners can only respond by relying on assumptions to make up for lack of data. Unfortunately, the assumptions they make are often questionable. Typically, a distribution is selected arbitrarily, and then fitted to the whole data set. However, this means that the fitted distribution will tend to accommodate the more central observations, because there are so many of them, rather than the extreme observations, which are much sparser. Hence, this type of approach is often good if we are interested in the central part of the distribution, but is ill-suited to handling extremes.

When dealing with extremes, we need an approach that comes to terms with the basic problem posed by extreme-value estimation: that the estimation of the risks associated with low-frequency events with limited data is inevitably problematic, and that these difficulties increase as the events concerned become rarer. Such problems are not unique to risk management, but also occur in other disciplines as well. The standard example is hydrology, where engineers have long struggled with the question of how high dikes, sea walls and similar barriers should be to contain the probabilities of floods within reasonable limits. They have had to do so with even less data than financial risk practitioners usually have, and their quantile estimates—the flood water levels they were contending with—were also typically well out of the range of their sample data. So they have had to grapple with comparable problems to those faced by insurers and risk managers, but have had to do so with even less data and potentially much more at stake.

The result of their efforts is extreme-value theory (EVT)—a branch of applied statistics that is tailor-made to these problems.[1] EVT focuses on the distinctiveness of extreme values and makes as much use as possible of what theory has to offer. Not surprisingly, EVT is quite different from the more familiar 'central tendency' statistics that most of us have grown up with. The underlying reason for this is that central tendency statistics are governed by central limit theorems, but central limit theorems do not apply to extremes. Instead, extremes are governed, appropriately enough, by extreme-value theorems. EVT uses these theorems to tell us what distributions we should (and should not!) fit to our extremes data, and also guides us on how we should estimate the parameters involved. These EV distributions are quite different from the more familiar distributions of central tendency statistics. Their parameters are also different, and the estimation of these parameters is more difficult.

This chapter provides an overview of EV theory, and of how it can be used to estimate measures of financial risk. We will focus mainly on the VaR (and to a lesser extent, the ES) to keep the discussion brief, but the approaches considered here extend naturally to the estimation of other coherent risk measures as well.

The chapter itself is divided into four sections. The first two discuss the two main branches of univariate EV theory, the next discusses some extensions to, including multivariate EVT, and the last concludes.

3.1 GENERALISED EXTREME-VALUE THEORY

Theory

Suppose we have a random loss variable X, and we assume to begin with that X is independent and identically distributed (iid) from some unknown distribution $F(x) = \text{Prob}(X \leq x)$. We wish to estimate the extreme risks (e.g., extreme VaR) associated with the distribution of X. Clearly, this poses a problem because we don't know what $F(x)$ actually is.

This is where EVT comes to our rescue. Consider a sample of size n drawn from $F(x)$, and let the maximum of this sample be M_n.[2] If n is large, we can regard M_n as an extreme value. Under relatively general conditions, the celebrated Fisher–Tippett theorem

[1] The literature on EVT is vast. However, some standard book references on EVT and its finance applications are Embrechts et. al. (1997), Reiss and Thomas (1997) and Beirlant et. al. (2004). There is also a plethora of good articles on the subject, e.g., Bassi et. al. (1998), Longin (1996, 1999), Danielsson and de Vries (1997a,b), McNeil (1998), McNeil and Saladin (1997), Cotter (2001, 2004), and many others.

[2] The same theory also works for extremes that are the minima rather than the maxima of a (large) sample: to apply the theory to minima extremes, we simply apply the maxima extremes results but multiply our data by −1.

(1928) then tells us that as n gets large, the distribution of extremes (i.e., M_n) converges to the following generalised extreme-value (GEV) distribution:

$$H_{\xi,\mu,\sigma} = \begin{cases} \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right] & \text{if } \xi \neq 0 \\ \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right] & \xi = 0 \end{cases} \quad (3.1)$$

where x satisfies the condition $1 + \xi(x - \mu)/\sigma > 0$.[3] This distribution has three parameters. The first two are μ, the location parameter of the limiting distribution, which is a measure of the central tendency of M_n, and σ, the scale parameter of the limiting distribution, which is a measure of the dispersion of M_n. These are related to, but distinct from, the more familiar mean and standard deviation, and we will return to these presently. The third parameter, ξ, the tail index, gives an indication of the shape (or heaviness) of the tail of the limiting distribution.

The GEV Equation (3.1) has three special cases:

- If $\xi > 0$, the GEV becomes the Fréchet distribution. This case applies where the tail of $F(x)$ obeys a power function and is therefore heavy (e.g., as would be the case if $F(x)$ were a Lévy distribution, a t-distribution, a Pareto distribution, etc.). This case is particularly useful for financial returns because they are typically heavy-tailed, and we often find that estimates of ξ for financial return data are positive but less than 0.35.
- If $\xi = 0$, the GEV becomes the Gumbel distribution, corresponding to the case where $F(x)$ has exponential tails. These are relatively light tails such as those we would get with normal or lognormal distributions.
- If $\xi < 0$, the GEV becomes the Weibull distribution, corresponding to the case where $F(x)$ has lighter than normal tails. However, the Weibull distribution is not particularly useful for modelling financial returns, because few empirical financial returns series are so light-tailed.[4]

The standardised (i.e., $\mu = 0, \sigma = 1$) Fréchet and Gumbel probability density functions are illustrated in Figure 3.1. Both are skewed to the right, but the Fréchet is more skewed than the Gumbel and has a noticeably longer right-hand tail. This means that the Fréchet has considerably higher probabilities of producing very large X-values.

Observe that most of the probability mass is located between x values of -2 and $+6$. More generally, this means most of the probability mass will lie between x values of $\mu - 2\sigma$ and $\mu + 6\sigma$.

To obtain the quantiles associated with the GEV distribution, we set the left-hand side of Equation (3.1) to p, take logs of both sides of Equation (3.1) and rearrange to get:

$$\ln(p) = \begin{cases} -\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi} & \text{if } \xi \neq 0 \\ -\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\} & \xi = 0 \end{cases} \quad (3.2)$$

We then unravel the x-values to get the quantiles associated with any chosen (cumulative) probability p:[5]

$$x = \mu - \frac{\sigma}{\xi}[1 - (-\ln(p))^{-\xi}] \quad \text{(Fréchet, } \xi > 0) \quad (3.3a)$$

$$x = \mu - \sigma \ln[-\ln(p)] \quad \text{(Gumbel, } \xi = 0) \quad (3.3b)$$

Example 3.1 Gumbel quantiles

For the standardised Gumbel, the 5% quantile is $-\ln[-\ln(0.05)] = -1.0972$ and the 95% quantile is $-\ln[-\ln(0.95)] = 2.9702$.

Example 3.2 Fréchet quantiles

For the standardised Fréchet with $\xi = 0.2$, the 5% quantile is $-(1/0.2)[1 - (-\ln(0.05))^{-0.2}] = -0.9851$ and the 95% quantile is $-(1/0.2)[1 - (-\ln(0.95))^{-0.2}] = 4.0564$. For $\xi = 0.3$, the 5% quantile is $-(1/0.3)[1 - (-\ln(0.05))^{-0.3}] = -0.9349$ and the 95% quantile is $-(1/0.3)[1 - (-\ln(0.95))^{-0.3}] = 4.7924$. Thus, Fréchet quantiles are sensitive to the value of the tail index ξ, and tend to rise with ξ. Conversely, as $\xi \to 0$, the Fréchet quantiles tend to their Gumbel equivalents.

We need to remember that the probabilities in Equations (3.1)–(3.3) refer to the probabilities associated with the extreme loss distribution, not to those associated with the distribution of the 'parent' loss distribution from which the extreme losses are drawn. For example, a 5th percentile in Equation (3.3) is the cut-off point between the lowest 5% of extreme (high) losses

[3] See, e.g., Embrechts et. al. (1997), p. 316.

[4] We can also explain these three cases in terms of domains of attraction. Extremes drawn from Lévy or t-distributions fall in the domain of attraction of the Fréchet distribution, and so obey a Fréchet distribution as n gets large; extremes drawn from normal and lognormal distributions fall in the domain of attraction of the Gumbel, and obey the Gumbel as n gets large, and so on.

[5] We can obtain estimates of EV VaR over longer time periods by using appropriately scaled parameters, bearing in mind that the mean scales proportionately with the holding period h, the standard deviation scales with the square root of h, and (subject to certain conditions) the tail index does not scale at all. In general, we find that the VaR scales with a parameter κ (i.e., so $VaR(h) = VaR(1)(h)^{\kappa}$, where h is the holding period), and empirical evidence reported by Hauksson et. al. (2001, p. 93) suggests an average value for κ of about 0.45. The square-root scaling rule (i.e., $\kappa = 0.5$) is therefore usually inappropriate for EV distributions.

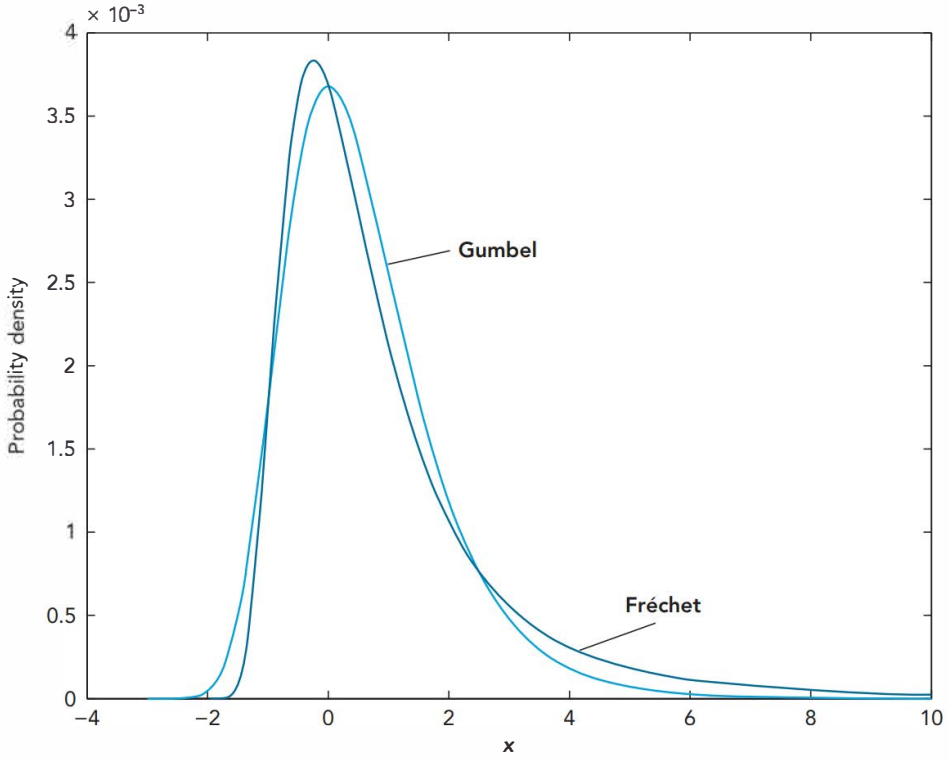


Figure 3.1 Standardised Gumbel and Fréchet probability density functions.

and the highest 95% of extreme (high) losses; it is *not* the 5th percentile point of the parent distribution. The 5th percentile of the extreme loss distribution is therefore on the *left*-hand side of the distribution of extreme losses (because it is a *small* extreme loss), but on the *right*-hand tail of the original loss distribution (because it *is* an extreme loss).

To see the connection between the probabilities associated with the distribution of M_n and those associated with the distribution of X, we now let M_n^* be some extreme threshold value. It then follows that:

$$\Pr[M_n < M_n^*] = p = \{\Pr[X < M_n^*]\}^n = [\alpha]^n \tag{3.4}$$

where α is the VaR confidence level associated with the threshold M_n^*. To obtain the α VaR, we now use Equation (3.4) to substitute $[\alpha]^n$ for p in Equation (3.3), and this gives us:

$$VaR = \mu_n - \frac{\sigma_n}{\xi_n}[1 - (-n\ln(\alpha))^{-\xi n}] \text{ (Fréchet, } \xi > 0) \tag{3.5a}$$

$$VaR = \mu_n - \sigma_n \ln[-n\ln(\alpha)] \text{ (Gumbel, } \xi = 0) \tag{3.5b}$$

(Since n is now explicit, we have also subscripted the parameters with n to make explicit that in practice these would refer to the parameters associated with maxima drawn from samples of size n. This helps to avoid errors with the limiting VaRs as n gets large.) Given values for the extreme-loss distribution parameters μ_n, σ_n and (where needed) ξ_n, Equation (3.5) allows us to estimate the relevant VaRs. Of course, the VaR formulas given by Equation (3.5) are meant only for extremely high confidence levels, and we cannot expect them to provide accurate estimates for VaRs at low confidence levels.

Example 3.3 Gumbel VaR

For the standardised Gumbel and $n = 100$, the 99.5% VaR is $-\ln[-100 \times \ln(0.995)] = 0.6906$, and the 99.9% VaR is $-\ln[-100 \times \ln(0.999)] = 2.3021$.

Example 3.4 Fréchet VaR

For the standardised Fréchet with $\xi = 0.2$ and $n = 100$, the 99.5% VaR is $-(1/0.2)[1 - (-100 \times \ln(0.995))^{-0.2}] = 0.7406$ and the 99.9% VaR is $-(1/0.2)[1 - (-100 \times \ln(0.999))^{-0.2}] = 2.9237$. For $\xi = 0.3$, the 99.5% VaR is $-(1/0.3)[1 - (-100 \times \ln(0.995))^{-0.3}] = 0.7674$ and the 99.9% VaR is $-(1/0.3)[1 - (-100 \times \ln(0.999))^{-0.3} = 3.3165$.

These results tell us that EV-VaRs (and, by implication, other EV risk measures) are sensitive to the value of the tail index ξ_n, which highlights the importance of getting a good estimate of ξ_n when applying EVT. This applies even if we use a Gumbel, because we should use the Gumbel only if we think ξ_n is insignificantly different from zero.

Example 3.5 Realistic Fréchet VaR

Suppose we wish to estimate Fréchet VaRs with more realistic parameters. For US stock markets, some fairly plausible parameters are $\mu = 2\%$, $\sigma = 0.7\%$ and $\xi = 0.3\%$. If we put these into our Fréchet VaR formula Equation (3.5a) and retain the earlier n value, the estimated 99.5% VaR (in %) is $2 - (0.7/0.3)[1 - (-100 \times \ln(0.995))^{-0.3}] = 2.537$, and the estimated 99.9% VaR (in %) is $2 - (0.7/0.3)[1 - (-100 \times \ln(0.999))^{-0.3}] = 4.322$. For the next trading day, these estimates tell us that we can be 99.5% confident of not making a loss in excess of 2.537% of the value of our portfolio, and so on.

It is also interesting to note that had we assumed a Gumbel (i.e., $\xi = 0$) we would have estimated these VaRs (again in %) to be $2 - 0.7 \times \ln[-100 \times \ln(0.995)] = 2.483$ and $2 - 0.7 \times \ln[-100 \times \ln(0.999)] = 3.612$. These are lower than the Fréchet VaRs, which underlines the importance of getting the ξ right.

How do we choose between the Gumbel and the Fréchet? There are various ways we can decide which EV distribution to use:

- If we are confident that we can identify the parent loss distribution, we can choose the EV distribution in whose domain of attraction the parent distribution resides. For example, if we are confident that the original distribution is a t, then we would choose the Fréchet distribution because the t belongs in the domain of attraction of the Fréchet. In plain English, we choose the EV distribution to which the extremes from the parent distribution will tend.
- We could test the significance of the tail index, and we might choose the Gumbel if the tail index was insignificant and the Fréchet otherwise. However, this leaves us open to the danger that we might incorrectly conclude that ξ is 0, and this could lead us to underestimate our extreme risk measures.
- Given the dangers of model risk and bearing in mind that the estimated risk measure increases with the tail index, a safer option is always to choose the Fréchet.

A Short-Cut EV Method

There are also short-cut ways to estimate VaR (or ES) using EV theory. These are based on the idea that if $\xi > 0$, the tail of an extreme loss distribution follows a power-law times a slowly varying function:

$$F(x) = k(x)x^{-1/\xi} \tag{3.6}$$

where $k(x)$ varies slowly with x. For example, if we assume for convenience that $k(x)$ is approximately constant, then Equation (3.6) becomes:

$$F(x) \approx kx^{-1/\xi} \tag{3.7}$$

Now consider two probabilities, a first, 'in-sample' probability $p_{in\text{-}sample}$, and a second, smaller and typically out-of-sample probability $p_{out\text{-}of\text{-}sample}$. Equation (3.7) implies:

$$p_{in\text{-}sample} \approx kx_{in\text{-}sample}^{-1/\xi} \text{ and}$$
$$p_{out\text{-}of\text{-}sample} \approx kx_{out\text{-}of\text{-}sample}^{-1/\xi} \tag{3.8}$$

which in turn implies:

$$\frac{p_{in\text{-}sample}}{p_{out\text{-}of\text{-}sample}} \approx \left(\frac{x_{in\text{-}sample}}{x_{out\text{-}of\text{-}sample}}\right)^{-1/\xi}$$
$$\Rightarrow x_{out\text{-}of\text{-}sample} \approx x_{in\text{-}sample}\left(\frac{p_{in\text{-}sample}}{p_{out\text{-}of\text{-}sample}}\right)^{\xi} \tag{3.9}$$

This allows us to estimate one quantile (denoted here as $x_{out\text{-}of\text{-}sample}$) based on a known in-sample quantile $x_{in\text{-}sample}$, a known out-of-sample probability $p_{out\text{-}of\text{-}sample}$ (which is known because it comes directly from our VaR confidence level), and an unknown in-sample probability $p_{in\text{-}sample}$.

The latter can easily be proxied by its empirical counterpart, t/n, where n is the sample size and t the number of observations higher than $x_{in\text{-}sample}$. Using this proxy then gives us:

$$x_{out\text{-}of\text{-}sample} \approx x_{in\text{-}sample}\left(\frac{np_{out\text{-}of\text{-}sample}}{t}\right)^{-\xi} \tag{3.10}$$

which is easy to estimate using readily available information.

To use this approach, we take an arbitrarily chosen in-sample quantile, $x_{in\text{-}sample}$, and determine its counterpart empirical probability, t/n. We then determine our out-of-sample probability from our chosen confidence level, estimate our tail index using a suitable method, and our out-of-sample quantile estimator immediately follows from Equation (3.10).[6]

Estimation of EV Parameters

To estimate EV risk measures, we need to estimate the relevant EV parameters—μ, σ and, in the case of the Fréchet, the tail index ξ, so we can insert their values into our quantile formulas

[6] An alternative short-cut is suggested by Diebold et. al. (2000). They suggest that we take logs of Equation (3.7) and estimate the log-transformed relationship using regression methods. However, their method is still relatively untried, and its reliability is doubtful because there is no easy way to ensure that the regression procedure will produce a 'sensible' estimate of the tail index.

(i.e., Equation (3.5)). We can obtain estimators using maximum likelihood (ML) methods, regression methods, moment-based or semi-parametric methods.

ML Estimation Methods

ML methods derive the most probable parameter estimators given the data, and are obtained by maximising the likelihood function. To apply an ML approach, we begin by constructing the likelihood or log-likelihood function. In the case of the Gumbel ($\xi = 0$) and with m observations for M_n, the log-likelihood function is:

$$l(\mu_n, \sigma_n) = -m \ln(\sigma_n) - \sum_{i=1}^{m} \exp\left(-\frac{M_n - \mu_n}{\sigma_n}\right) - \sum_{i=1}^{m} \frac{M_n - \mu_n}{\sigma_n} \quad (3.11)$$

Where $\xi \neq 0$ the log-likelihood function is:

$$l(\mu_n, \sigma_n, \xi_n) = -m \ln(\sigma_n) - (1 + 1/\xi_n) \sum_{i=1}^{m} \ln\left[1 + \xi_n\left(\frac{M_n - \mu_n}{\sigma_n}\right)\right] - \sum_{i=1}^{m} \ln\left[1 + \xi_n\left(\frac{M_n - \mu_n}{\sigma_n}\right)\right]^{-\frac{1}{\xi_n}} \quad (3.12)$$

which would be maximised subject to the constraint that any observation M_n^i satisfies $1 + \xi(M_n^i - \mu)/\sigma > 0$. The ML approach has some attractive properties (e.g., it is statistically well grounded, parameter estimators are consistent and asymptotically normal if $\xi_n > -1/2$, we can easily test hypotheses about parameters using likelihood ratio statistics, etc.). However, it also lacks closed-form solutions for the parameters, so the ML approach requires the use of an appropriate numerical solution method. This requires suitable software, and there is the danger that ML estimators might not be robust. In addition, because the underlying theory is asymptotic, there is also the potential for problems arising from smallness of samples.

Regression Methods

An easier method to apply is a regression method due to Gumbel (1958).[7] To see how the method works, we begin by ordering our sample of M_n^i values from lowest to highest, so $M_n^1 \leq M_n^2 \leq \ldots \leq M_n^m$. Because these are order statistics, it follows that, for large n:

$$E[H(M_n^i)] = \frac{i}{1+m} \Rightarrow H(M_n^i) \approx \frac{i}{1+m} \quad (3.13)$$

where $H(M_n^i)$ is the cumulative density function of maxima, and we drop all redundant scripts for convenience. (See Equation (3.1) above.) In the case where $\xi \neq 0$, Equations (3.1) and (3.13) together give

$$\frac{i}{1+m} \approx \exp[-(1 + \xi_n(M_n^i - \mu_n)/\sigma_n)^{-1/\xi}] \quad (3.14)$$

Taking logs twice of both sides yields:

$$\log\left[-\log\left(\frac{i}{1+m}\right)\right] \approx -\frac{1}{\xi_n} \log\left[1 + \xi_n\left(\frac{M_n - \mu_n}{\sigma_n}\right)\right] \quad (3.15)$$

and we can obtain least squares estimates of μ_n, σ_n and ξ_n from a regression of $\log[-\log(i/(1+m))]$ against $[1 + \xi_n(M_n - \mu_n)/\sigma_n]$. When $\xi = 0$, then the equivalent of Equation (3.14) is:

$$\log\left[-\log\left(\frac{i}{1+m}\right)\right] \approx \left(\frac{M_n - \mu_n}{\sigma_n}\right) \quad (3.16)$$

and the recovery of parameter estimates from a regression is straightforward.

Semi-Parametric Estimation Methods

We can also estimate parameters using semi-parametric methods. These are typically used to estimate the tail index ξ, and the most popular of these is the Hill estimator. This estimator is directly applied to the ordered parent loss observations. Denoting these from highest to lowest by $X_1, X_2, \ldots, X_n$, the Hill $\hat{\xi}_{n,k}^{(H)}$ is:

$$\hat{\xi}_{n,k}^{(H)} = \frac{1}{k} \sum_{i=1}^{k} \ln X_i - \ln X_{k+1} \quad (3.17)$$

where k, the tail threshold used to estimate the Hill estimator, has to be chosen in an appropriate way. The Hill estimator is the average of the k most extreme (i.e., tail) observations, minus the $k+1$th observation, or the one next to the tail. The Hill estimator is known to be consistent and asymptotically normally distributed, but its properties in finite samples are not well understood, and there are concerns in the literature about its small-sample properties and its sensitivity to the choice of threshold k. However, these (and other) reservations notwithstanding, many EVT practitioners regard the Hill estimator as being as good as any other.[8]

The main problem in practice is choosing a cut-off value for k. We know that our tail index estimates can be sensitive to the choice of k, but theory gives us little guidance on what the value of k should be. A suggestion often given to this problem is that

[7] See Gumbel (1958), pp. 226, 260, 296.

[8] An alternative is the Pickands estimator (see, e.g., Bassi et. al. (1998), p. 125 or Longin (1996), p. 389). This estimator does not require a positive tail index (unlike the Hill estimator) and is asymptotically normal and weakly consistent under reasonable conditions, but is otherwise less efficient than the Hill estimator.

BOX 3.1 MOMENT-BASED ESTIMATORS OF EV PARAMETERS

An alternative approach is to estimate EV parameters using empirical moments. Let m_i be the ith empirical moment of our extremes data set. Assuming $\xi \neq 0$, we can adapt Embrechts et. al. (1997, pp. 293–295) to show that:

$$m_1 = \mu - \frac{\sigma}{\xi}(1 - \Gamma(1 - \xi))$$

$$2m_2 - m_1 = \frac{\sigma}{\xi}\Gamma(1 - \xi)(2^{\xi} - 1)$$

$$3m_3 - m_1 = \frac{\sigma}{\xi}\Gamma(1 - \xi)(3^{\xi} - 1)$$

where the $\Gamma(.)$ is a gamma function. Dividing the last of these into the preceding one gives us an implied estimator $\hat{\xi}$ of ξ. The first two equations can then be rearranged to give us estimators for μ and σ in terms of $\hat{\xi}$ and sample moments m_1 and m_2:

$$\hat{\sigma} = \frac{(2m_2 - m_1)\hat{\xi}}{\Gamma(1 - \hat{\xi})(2^{\hat{\xi}} - 1)}$$

$$\hat{\mu} = m_1 + \frac{\hat{\sigma}}{\hat{\xi}}(1 - \Gamma(1 - \hat{\xi}))$$

The Gumbel equivalents are obtained by taking the limit as $\xi \to 0$. In this case

$$\hat{\sigma} = \sqrt{\frac{6m_2}{\pi}}$$

$$\hat{\mu} = m_1 + \Gamma(1)\hat{\sigma} = m_1 - 0.57722\hat{\sigma}$$

This moment-based approach is easy to apply, but, it is unreliable because of the poor sampling properties of the second- and higher-order moments.

However, following Hosking et. al. (1985), we can obtain estimates with superior sampling properties if we replace the m_i in the above expressions with their probability-weighted counterparts w_i, where $w_i = E[X(F(X)^{i-1})]$ for $i = 1, 2, \ldots$ If we wish to, we can also replace the m_i with more general probability-weighted moments $w_{i,r,s}$, where $w_{i,r,s} = E[X^r(F(X)^{i-1}(1 - F(X))^{s-1}]$ for i, r, $s = 1, 2, \ldots$

we estimate Hill (or Pickands) estimators for a range of k values, and go for k values where the plot of estimators against k-values (hopefully) becomes more or less horizontal: if the plot stabilises and flattens out, then the plateau value should give a reasonable estimate of our tail index. This suggestion tries to extract the maximum possible information from all our data, albeit in an informal way.

To show how this might be done, Figure 3.2 shows a 'Hill plot'—a plot of the values of the Hill estimator against k, the tail threshold size used to estimate the Hill estimator, based on 1000 simulated observations from an underlying distribution. As we can see, the Hill estimates are a little unsteady for low values of k, but they become more stable and settle down as k gets larger, and one might suppose that the 'true' value of the tail index lies in the region of 0.18–0.20. Such a value would be plausible for many real situations, so if we met such a situation in practice we could easily persuade ourselves that this was a fair estimate.

However, in coming to such a conclusion, we are implicitly presuming that the values of the Hill estimator do indeed settle down for values of k bigger than 40. Is this assumption justified? The answer, sadly, is that it is often not. We can see why when we extend the same plot for higher values of k: despite the fact that the values of the Hill estimator looked like they were settling down as k approached 40, it turns out that they were doing nothing of the sort. This comes as a shock. In fact, the values of the Hill estimator show no sign of settling down at all. The Hill plot becomes a 'Hill horror plot' and gives us no real guidance on how we might choose k—and this means that it does not help us to determine what the value of the Hill estimator might be.[9] The Hill horror plot is shown in Figure 3.3.

Hill horror plots can be a real problem, and it is sometimes suggested that the best practical response when meeting them is to 'patch' up the estimator and hope for the best. To illustrate this in the present context, I played around a little with the above data and soon discovered that I could obtain a fairly nice Hill plot by making a very small adjustment to the Hill formula.[10] The resulting 'Hill happy plot' is shown in Figure 3.4. In this case, the values of the Hill estimator do settle down as k gets larger, and the plot suggests that we can take the best value of the tail index to be somewhere in the region of 0.15. We have therefore 'solved' the problem of the Hill horror plot. However, this

[9] Purists might point out that we might expect a badly behaved Hill estimator when using data drawn from a normal distribution. This may be true, but it misses the main point of the exercise: Hill horror plots are all too common, and occur with many non-normal distributions as well.

[10] For those who are curious about it, the adjustment used is to add in the extra term $-0.0015\,k$ to the Hill formula Equation (3.17).

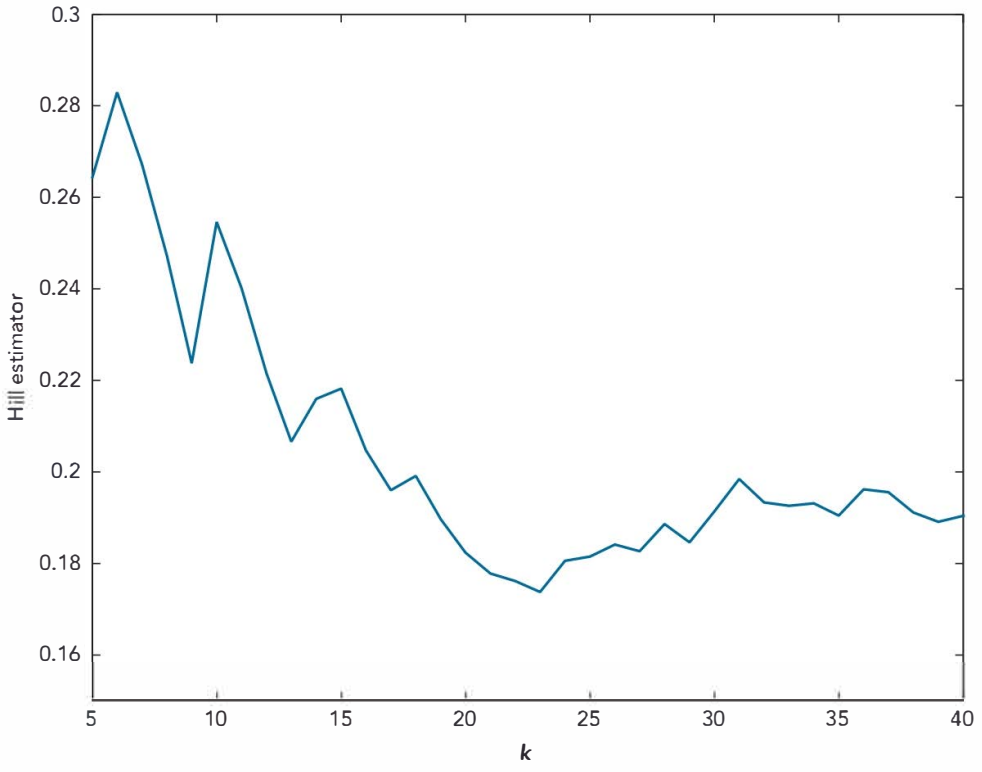


Figure 3.2 Hill plot.

Note: Based on 1000 simulated drawings from a standard normal distribution.

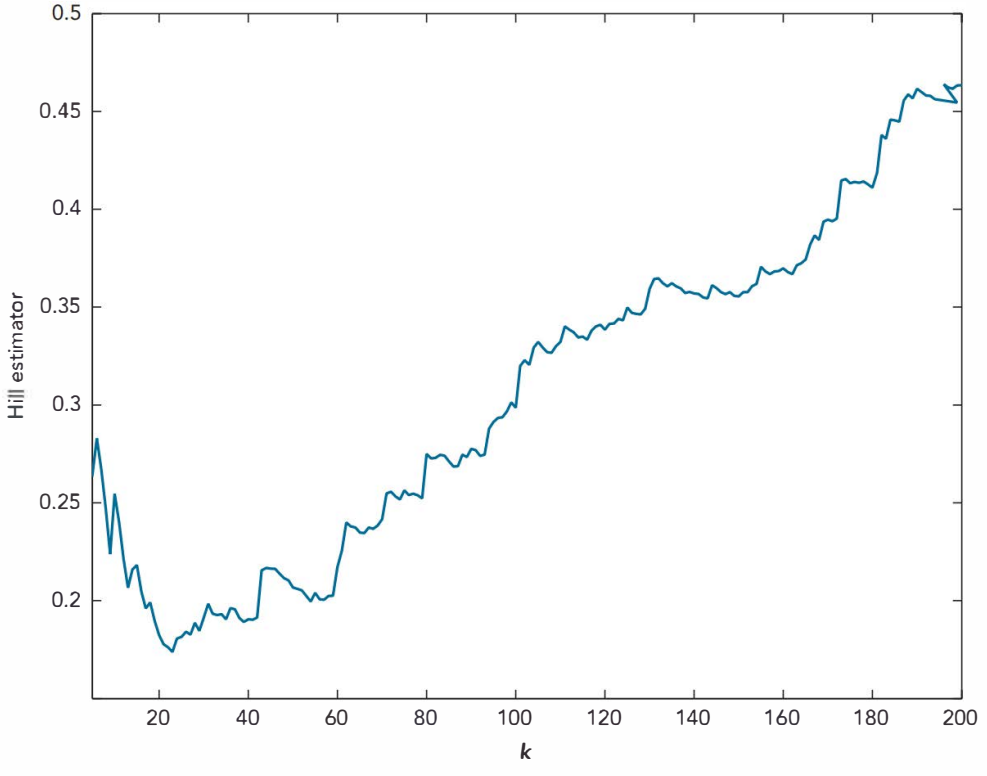


Figure 3.3 Hill horror plot.

Note: Based on 1000 simulated drawings from a standard normal distribution.

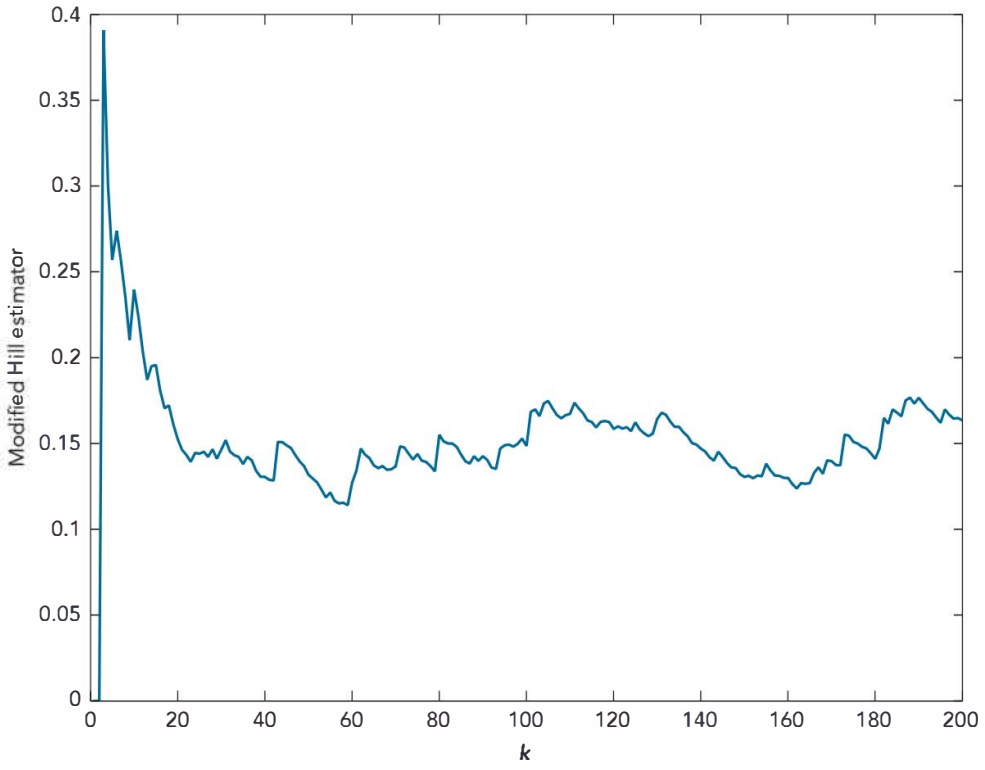


Figure 3.4 **'Hill happy plot.'**

Note: Based on 1000 simulated drawings from a standard normal distribution.

'solution' comes at a big price: the adjustment is completely ad hoc and has no theoretical basis whatever. More to the point, we don't even know whether the answer it gives us is any good: all we really know is that we have managed to patch up the estimator to stabilise the Hill plot, but whether this actually helps us is quite a different matter.

If we have a very large sample size, we can also use an alternative method of gauging the 'right' value of *k*.

Danielsson and de Vries (1997b) have suggested an ingenious (though rather involved) procedure based on the fact that the choice of *k* implies a trade-off between bias and variance. If we increase *k*, we get more data and so move to the centre of the distribution. This increases the precision of our estimator (and therefore reduces its variance), but also increases the bias of the tail estimator by placing relatively more weight on observations closer to the centre of our distribution. Alternatively, if we decrease *k* and move further out along the tail, we decrease bias but have fewer data to work with and get a higher variance. These authors suggest that we choose the value of *k* to minimise a mean-squared-error (MSE) loss function, which reflects an optimal trade-off, in an MSE sense, between bias and variance. The idea is that we take a second-order approximation to the tail of the distribution function $F(x)$, and exploit the point that the tail size is optimal in an asymptotic mean-squared-error sense where bias and variance disappear at the same rate. This optimal size can be found by a subsample bootstrap procedure. However, this approach requires a large sample size—at least 1500 observations—and is therefore impractical with small sample sizes. In addition, any automatic procedure for selecting *k* tends to ignore other, softer, but nonetheless often very useful, information, and this leads some writers to be somewhat sceptical of such methods.

3.2 THE PEAKS-OVER-THRESHOLD APPROACH: THE GENERALISED PARETO DISTRIBUTION

Theory

We turn now to the second strand of the EV literature, which deals with the application of EVT to the distribution of excess losses over a (high) threshold. This gives rise to the peaks-over-threshold (POT) or generalised Pareto approach, which

BOX 3.2 ESTIMATING VaR UNDER MAXIMUM DOMAIN OF ATTRACTION CONDITIONS

We have assumed so far that our maxima data were drawn exactly from the GEV. But what happens if our data are only approximately GEV distributed (i.e., are drawn from the maximum domain of attraction of the GEV)? The answer is that the analysis becomes somewhat more involved. Consider the Fréchet case where $\xi = 1/\alpha > 0$. The far-right tail $\bar{F}(x) = 1 - F(x)$ is now $\bar{F}(x) = x^{-\alpha}L(x)$ for some slowly varying function L. However, the fact that the data are drawn from the maximum domain of attraction of the Fréchet also means that

$$\lim_{n\to\infty} n\,F(c_n x + d_n) = -\log H_\xi(x)$$

where $H_\xi(x)$ is the standardised (0 location, unit scale) Fréchet, and c_n and d_n are appropriate norming (or scaling) parameters. Invoking Equation (3.1), it follows for large $u = c_n x + d_n$ that

$$\bar{F}(u) \approx \frac{1}{n}\left(1 + \xi\frac{u - d_n}{c_n}\right)^{-\frac{1}{\xi}}$$

This leads to the quantile estimator

$$\hat{X}_p = \hat{d}_n + \frac{\hat{c}_n}{\hat{\xi}}([n(1-p)]^{-\hat{\xi}} - 1)$$

for appropriate parameter estimators $\hat{c}_n$, $\hat{d}_n$ and $\hat{\xi}$, and some high probability (confidence level) p. The problem is then to estimate p-quantiles outside the range of the data where the empirical tail $\hat{F}() = 0$. The standard approach to this problem is a subsequence trick: in effect, we replace n with n/k. This yields the quantile estimator

$$\hat{X}_p = \hat{d}_{n/k} + \frac{\hat{c}_{n/k}}{\hat{\xi}}\left(\left[\frac{n}{k}(1-p)\right]^{-\xi} - 1\right)$$

$\hat{c}_{n/k}$ and $\hat{d}_{n/k}$ can be obtained using suitable semi-parametric methods, and $\hat{\xi}$ can be obtained using the usual Hill or other tail index approaches.[11]

(generally) requires fewer parameters than EV approaches based on the generalised extreme value theorem. The POT approach provides *the* natural way to model exceedances over a high threshold, in the same way that GEV theory provides the natural way to model the maxima or minima of a large sample.

If X is a random iid loss with distribution function $F(x)$, and u is a threshold value of X, we can define the distribution of excess losses over our threshold u as:

$$F_u(x) = \Pr\{X - u \le x | X > u\} = \frac{F(x+u) - F(u)}{1 - F(u)} \qquad (3.18)$$

for $x > 0$. This gives the probability that a loss exceeds the threshold u by at most x, given that it does exceed the threshold. The distribution of X itself can be any of the commonly used distributions: normal, lognormal, t, etc., and will usually be unknown to us. However, as u gets large, the Gnedenko–Pickands–Balkema–deHaan (GPBdH) theorem states that the distribution $F_u(x)$ converges to a generalised Pareto distribution, given by:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp(-x/\beta) & \text{if } \xi = 0 \end{cases} \qquad (3.19)$$

defined for $x \ge 0$ for $\xi \ge 0$ and $0 \le x \le -\beta/\xi$ for $\xi < 0$. This distribution has only two parameters: a positive scale parameter, β, and a shape or tail index parameter, ξ, that can be positive, zero or negative. This latter parameter is the same as the tail index encountered already with GEV theory. The cases that usually interest us are the first two, and particularly the first (i.e., $\xi > 0$), as this corresponds to data being heavy tailed.

The GPBdH theorem is a very useful result, because it tells us that the distribution of excess losses always has the same form (in the limit, as the threshold gets high), pretty much regardless of the distribution of the losses themselves. Provided the threshold is high enough, we should therefore regard the GP distribution as *the* natural model for excess losses.

To apply the GP distribution, we need to choose a reasonable threshold u, which determines the number of observations, N_u, in excess of the threshold value. Choosing u involves a trade-off: we want a threshold u to be sufficiently high for the GPBdH theorem to apply reasonably closely; but if u is too high, we won't have enough excess-threshold observations on which to make reliable estimates. We also need to estimate the parameters ξ and β. As with the GEV distributions, we can estimate these using maximum likelihood approaches or semi-parametric approaches.

We now rearrange the right-hand side of Equation (3.18) and move from the distribution of exceedances over the threshold to the parent distribution $F(x)$ defined over 'ordinary' losses:

$$F(x) = (1 - F(u))G_{\xi,\beta}(x - u) + F(u) \qquad (3.20)$$

[11] For more on estimation under maximum domain of attraction conditions, see Embrechts et. al. (1997, section 6.4).

where $x > u$. To make use of this equation, we need an estimate of $F(u)$, the proportion of observations that do not exceed the threshold, and the most natural estimator is the observed proportion of below-threshold observations, $(n - N_u)/n$. We then substitute this for $F(u)$, and plug Equation (3.19) into Equation (3.20):

$$F(x) = 1 - \frac{N_u}{n}\left[1 + \xi\left(\frac{x - u}{\beta}\right)\right]^{-1/\xi} \tag{3.21}$$

The VaR is given by the x-value in Equation (3.21), which can be recovered by inverting Equation (3.21) and rearranging to get:

$$VaR = u + \frac{\beta}{\xi}\left\{\left[\frac{n}{N_u}(1 - \alpha)\right]^{-\xi} - 1\right\} \tag{3.22}$$

where α, naturally, is the VaR confidence level.

The ES is then equal to the VaR plus the mean–excess loss over VaR. Provided $\xi < 1$, our ES is:

$$ES = \frac{VaR}{1 - \xi} + \frac{\beta - \xi u}{1 - \xi} \tag{3.23}$$

Example 3.6 POT risk measures

Suppose we set our parameters at some empirically plausible values denominated in % terms (i.e., $\beta = 0.8$, $\xi = 0.15$, $u = 2\%$ and $N_u/n = 4\%$; these are based on the empirical values associated with contracts on futures clearinghouses). The 99.5% VaR (in %) is therefore

$$VaR = 2 + \frac{0.8}{0.15}\left\{\left[\frac{1}{0.04}(1 - 0.995)\right]^{-0.15} - 1\right\} = 3.952$$

The corresponding ES (in %) is

$$ES = \frac{3.952}{1 - 0.15} + \frac{0.8 - 0.15 \times 2}{1 - 0.15} = 5.238$$

If we change the confidence level to 99.9%, the VaR and ES are easily shown to be 5.942 and 17.578.

Estimation

To obtain estimates, we need to choose a reasonable threshold u, which then determines the number of excess-threshold observations, N_u. The choice of threshold is the weak spot of POT theory: it is inevitably arbitrary and therefore judgmental. Choosing u also involves a trade-off: we want the threshold u to be sufficiently high for the GPBdH theorem to apply reasonably closely; but if u is too high, we will not have enough excess-threshold observations from which to obtain reliable estimates. This threshold problem is very much akin to the problem of choosing k to estimate the tail index. We can also (if we are lucky!) deal with it in a similar way. In this case, we would plot the mean–excess function, and choose a threshold where the MEF becomes horizontal. We also need to estimate the parameters ξ and β and, as with the earlier GEV approaches, we can estimate these using maximum likelihood or other appropriate methods.[12] Perhaps the most reliable are the ML approaches, which involve the maximisation of the following log-likelihood:

$$l(\xi, \beta) = \begin{cases} -m \ln \beta - (1 + 1/\xi)\sum_{i=1}^{m} \ln(1 + \xi X_i/\beta) & \xi \neq 0 \\ -m \ln \beta - (1/\beta)\sum_{i=1}^{m} X_i & \xi = 0 \end{cases} \text{ if } \tag{3.24}$$

subject to the conditions on which $G_{\xi,\beta}(x)$ is defined. Provided $\xi > -0.5$, ML estimators are asymptotically normal, and therefore (relatively) well behaved.

GEV vs POT

Both GEV and POT approaches are different manifestations of the same underlying EV theory, although one is geared towards the distribution of extremes as such, whereas the other is geared towards the distribution of exceedances over a high threshold. In theory, there is therefore not too much to choose between them, but in practice there may sometimes be reasons to prefer one over the other:

- One might be more natural in a given context than the other (e.g., we may have limited data that would make one preferable).
- The GEV typically involves an additional parameter relative to the POT, and the most popular GEV approach, the block maxima approach (which we have implicitly assumed so far), can involve some loss of useful data relative to the POT approach, because some blocks might have more than one extreme in them. Both of these are disadvantages of the GEV relative to the POT.
- On the other hand, the POT approach requires us to grapple with the problem of choosing the threshold, and this problem does not arise with the GEV.

However, at the end of the day, either approach is usually reasonable, and one should choose the one that seems to best suit the problem at hand.

[12] We can also estimate these parameters using moment-based methods, as for the GEV parameters (see Box 8-2). For the GPD, the parameter estimators are $\beta = 2m_1m_2/(m_1 - 2m_2)$ and $\xi = 2 - m_1/(m_1 - 2m_2)$ (see, e.g., Embrechts et. al. (1997), p. 358). However, as with their GEV equivalents, moment-based estimators can be unreliable, and the probability-weighted or ML ones are usually to be preferred.

3.3 REFINEMENTS TO EV APPROACHES

Having outlined the basics of EVT and its implementation, we now consider some refinements to it. These fall under three headings:

- Conditional EV.
- Dealing with dependent (or non-iid) data.
- Multivariate EVT.

Conditional EV

The EVT procedures described above are all unconditional: they are applied directly (i.e., without any adjustment) to the random variable of interest, *X*. As with other unconditional applications, unconditional EVT is particularly useful when forecasting VaR or ES over a long horizon period. However, it will sometimes be the case that we wish to apply EVT to *X* adjusted for (i.e., conditional on) some dynamic structure, and this involves distinguishing between *X* and the random factors driving it. This conditional or dynamic EVT is most useful when we are dealing with a short horizon period, and where *X* has a dynamic structure that we can model. A good example is where *X* might be governed by a GARCH process. In such circumstances we might want to take account of the GARCH process and apply EVT not to the raw return process itself, but to the random innovations that drive it.

One way to take account of this dynamic structure is to estimate the GARCH process and apply EVT to its residuals. This suggests the following two-step procedure:[13]

- We estimate a GARCH-type process (e.g., a simple GARCH, etc.) by some appropriate econometric method and extract its residuals. These should turn out to be iid. The GARCH-type model can then be used to make one-step ahead predictions of next period's location and scale parameters, μ_{t+1} and σ_{t+1}.
- We apply EVT to these residuals, and then derive VaR estimates taking account of both the dynamic (i.e., GARCH) structure and the residual process.

Dealing with Dependent (or Non-iid) Data

We have assumed so far that the stochastic process driving our data is iid, but most financial returns exhibit some form of time dependency (or pattern over time). This time dependency usually takes the form of clustering, where high/low observations are clustered together. Clustering matters for a number of reasons:

- It violates an important premise on which the earlier results depend, and the statistical implications of clustering are not well understood.
- There is evidence that data dependence can produce very poor estimator performance.[14]
- Clustering alters the interpretation of our results. For example, we might say that there is a certain quantile or VaR value that we would expect to be exceeded, on average, only once every so often. But if data are clustered, we do not know how many times to expect this value to be breached in any given period: how frequently it is breached will depend on the tendency of the breaches to be clustered.[15] Clustering therefore has an important effect on the interpretation of our results.

There are two simple methods of dealing with time dependency in our data. Perhaps the most common (and certainly the easiest) is just to apply GEV distributions to block maxima. This is the simplest and most widely used approach. It exploits the point that maxima are usually less clustered than the underlying data from which they are drawn, and become even less clustered as the periods of time from which they are drawn get longer. We can therefore completely eliminate time dependence if we choose long enough block periods. This block maxima approach is very easy to use, but involves some efficiency loss, because we throw away extreme observations that are not block maxima. There is also the drawback that there is no clear guide about how long the block periods should be, which leads to a new bandwidth problem comparable to the earlier problem of how to select *k*.

A second solution to the problem of clustering is to estimate the tail of the conditional distribution rather than the unconditional one: we would first estimate the conditional volatility model (e.g., via a GARCH procedure), and then estimate the tail index of conditional standardized data. The time dependency in our data is then picked up by the deterministic part of our model, and we can treat the random process as independent.[16]

[13] This procedure is developed in more detail by McNeil and Frey (2000).

[14] See, e.g., Kearns and Pagan (1997).

[15] See McNeil (1998), p. 13.

[16] There is also a third, more advanced but also more difficult, solution. This is to estimate an extremal index—a measure of clustering—and use this index to adjust our quantiles for clustering. For more details on the extremal index and how to use it, see, e.g., Embrechts et. al. (1997, Chapter 8.1).

Multivariate EVT

We have been dealing so far with univariate EVT, but there also exists multivariate extreme value theory (MEVT), which can be used to model the tails of multivariate distributions in a theoretically appropriate way. The key issue here is how to model the dependence structure of extreme events. To appreciate this issue, it is again important to recognise how EV theory differs from more familiar central-value theory. As we all know, when dealing with central values, we often rely on the central limit theorem to justify the assumption of a normal (or more broadly, elliptical) distribution. When we have such a distribution, the dependence structure can then be captured by the (linear) correlations between the different variables. Given our distributional assumptions, knowledge of variances and correlations (or, if we like, covariances) suffices to specify the multivariate distribution. This is why correlations are so important in central-value theory.

However, this logic does not carry over to extremes. When we go beyond elliptical distributions, correlation no longer suffices to describe the dependence structure. Instead, the modeling of multivariate extremes requires us to make use of copulas. MEVT tells us that the limiting distribution of multivariate extreme values will be a member of the family of EV copulas, and we can model multivariate EV dependence by assuming one of these EV copulas. In theory, our copulas can also have as many dimensions as we like, reflecting the number of random variables to be considered. However, there is a curse of dimensionality here. For example, if we have two independent variables and classify univariate extreme events as those that occur one time in a 100, then we should expect to see one multivariate extreme event (i.e., both variables taking extreme values) only one time in 100^2, or one time in 10 000 observations; with three independent variables, we should expect to see a multivariate extreme event one time in 100^3, or one time in 1 000 000 observations, and so on. As the dimensionality rises, our multivariate EV events rapidly become much rarer: we have fewer multivariate extreme observations to work with, and more parameters to estimate. There is clearly a limit to how many dimensions we can handle.

One might be tempted to conclude from this example that multivariate extremes are sufficiently rare that we need not worry about them. However, this would be a big mistake. Even in theory, the occurrence of multivariate extreme events depends on their joint distribution, and extreme events cannot be assumed to be independent. Instead the occurrence of such events is governed by the tail dependence of the multivariate distribution. Indeed, it is for exactly this reason that tail dependence is the central focus of MEVT. And, as a matter of empirical fact, it is manifestly obvious that (at least some) extreme events are not independent: a major earthquake can trigger other natural or financial disasters (e.g., tsunamis or market crashes). We all know that disasters are often related. It is therefore important for risk managers to have some awareness of multivariate extreme risks.

3.4 CONCLUSIONS

EVT provides a tailor-made approach to the estimation of extreme probabilities and quantiles. It is intuitive and plausible; and it is relatively easy to apply, at least in its more basic forms. It also gives us considerable practical guidance on what we should estimate and how we should do it; and it has a good track record. It therefore provides *the* ideal, tailor-made, way to estimate extreme risk measures.

EVT is also important in what it tells us *not* to do, and the most important point is not to use distributions justified by central limit theory—most particularly, the normal or Gaussian distribution—for extreme-value estimation. If we wish to estimate extreme risks, we should do so using the distributions suggested by EVT, not arbitrary distributions (such as the normal) that go against what EVT tells us.

But we should not lose sight of the limitations of EV approaches, and certain limitations stand out:

- EV problems are intrinsically difficult, because by definition we always have relatively few extreme-value observations to work with. This means that any EV estimates will necessarily be very uncertain, relative to any estimates we might make of more central quantiles or probabilities. EV estimates will therefore have relatively wide confidence intervals attached to them. This uncertainty is not a fault of EVT as such, but an inevitable consequence of our paucity of data.
- EV estimates are subject to considerable model risk. We have to make various assumptions in order to carry out extreme-value estimations, and our results will often be very sensitive to the precise assumptions we make. At the same time, the veracity or otherwise of these assumptions can be difficult to verify in practice. Hence, our estimates are often critically dependent on assumptions that are effectively unverifiable. EVT also requires us to make ancillary decisions about threshold values and the like, and there are no easy ways to make those decisions: the application of EV methods involves a lot of subjective 'judgment'. Because of this uncertainty, it is especially important with extremes to estimate confidence

intervals for our estimated risk measures and to subject the latter to stress testing.

- Because we have so little data and the theory we have is (mostly) asymptotic, EV estimates can be very sensitive to small sample effects, biases, non-linearities, and other unpleasant problems.

In the final analysis, we need to make the best use of theory while acknowledging that the paucity of our data inevitably limits the reliability of our results. To quote McNeil,

> We are working in the tail . . . and we have only a limited amount of data which can help us. The uncertainty in our analyses is often high, as reflected by large confidence intervals. . . . However, if we wish to quantify rare events we are better off using the theoretically supported methods of EVT than other ad hoc approaches. EVT gives the best estimates of extreme events and represents the most honest approach to measuring the uncertainty inherent in the problem.[17]

Thus EVT has a very useful, albeit limited, role to play in risk measurement. As Diebold et. al. nicely put it:

> EVT is here to stay, but we believe that best-practice applications of EVT to financial risk management will benefit from awareness of its limitations—as well as its strengths. When the smoke clears, the contribution of EVT remains basic and useful: It helps us to draw smooth curves through the extreme tails of empirical survival functions in a way that is guided by powerful theory. . . . [But] we shouldn't ask more of the theory than it can deliver.[18]

---

[17] McNeil (1998, p. 18).

[18] Diebold et. al. (2000), p. 34.

4

Backtesting VaR

Learning Objectives

After completing this reading, you should be able to:

- Describe backtesting and exceptions and explain the importance of backtesting VaR models.
- Explain the significant difficulties in backtesting a VaR model.
- Evaluate the accuracy of a VaR model based on exceptions or failure rates by using a model verification test.
- Identify and describe Type I and Type II errors in the context of a backtesting process.
- Explain the need to consider conditional coverage in the backtesting framework.
- Describe the Basel rules for backtesting.

Excerpt is Chapter 6 of Value at Risk: The New Benchmark for Managing Financial Risk, *Third Edition, by Philippe Jorion.*

> Disclosure of quantitative measures of market risk, such as value-at-risk, is enlightening only when accompanied by a thorough discussion of how the risk measures were calculated and how they related to actual performance.
>
> —Alan Greenspan (1996)

Value-at-risk (VaR) models are only useful insofar as they predict risk reasonably well. This is why the application of these models always should be accompanied by validation. *Model validation* is the general process of checking whether a model is adequate. This can be done with a set of tools, including backtesting, stress testing, and independent review and oversight.

This chapter turns to backtesting techniques for verifying the accuracy of VaR models. *Backtesting* is a formal statistical framework that consists of verifying that actual losses are in line with projected losses. This involves systematically comparing the history of VaR forecasts with their associated portfolio returns.

These procedures, sometimes called *reality checks*, are essential for VaR users and risk managers, who need to check that their VaR forecasts are well calibrated. If not, the models should be reexamined for faulty assumptions, wrong parameters, or inaccurate modeling. This process also provides ideas for improvement and as a result should be an integral part of all VaR systems.

Backtesting is also central to the Basel Committee's groundbreaking decision to allow internal VaR models for capital requirements. It is unlikely the Basel Committee would have done so without the discipline of a rigorous backtesting mechanism. Otherwise, banks may have an incentive to understate their risk. This is why the backtesting framework should be designed to maximize the probability of catching banks that willfully understate their risk. On the other hand, the system also should avoid unduly penalizing banks whose VaR is exceeded simply because of bad luck. This delicate choice is at the heart of statistical decision procedures for backtesting.

This chapter first provides an actual example of model verification and discusses important data issues for the setup of VaR backtesting, then presents the main method for backtesting, which consists of counting deviations from the VaR model. It also describes the supervisory framework by the Basel Committee for backtesting the internal-models approach. Finally, practical uses of VaR backtesting are illustrated.

4.1 SETUP FOR BACKTESTING

VaR models are only useful insofar as they can be demonstrated to be reasonably accurate. To do this, users must check systematically the validity of the underlying valuation and risk models through comparison of predicted and actual loss levels.

When the model is perfectly calibrated, the number of observations falling outside VaR should be in line with the confidence level. The number of exceedences is also known as the number of *exceptions*. With too many exceptions, the model underestimates risk. This is a major problem because too little capital may be allocated to risk-taking units; penalties also may be imposed by the regulator. Too few exceptions are also a problem because they lead to excess, or inefficient, allocation of capital across units.

An Example

An example of model calibration is described in Figure 4.1, which displays the fit between actual and forecast daily VaR numbers for Bankers Trust. The diagram shows the absolute value of the daily profit and loss (P&L) against the 99 percent VaR, defined here as the *daily price volatility*.[1] The graph shows substantial time variation in the VaR measures, which reflects changes in the risk profile of the bank. Observations that lie above the diagonal line indicate days when the absolute value of the P&L exceeded the VaR.

Assuming symmetry in the P&L distribution, about 2 percent of the daily observations (both positive and negative) should lie above the diagonal, or about 5 data points in a year. Here we observe four exceptions. Thus the model seems to be well calibrated. We could have observed, however, a greater number of deviations simply owing to bad luck. The question is: At what point do we reject the model?

Which Return?

Before we even start addressing the statistical issue, a serious data problem needs to be recognized. VaR measures assume that the current portfolio is "frozen" over the horizon. In

[1] Note that the graph does not differentiate losses from gains. This is typically the case because companies usually are reluctant to divulge the extent of their trading losses. This illustrates one of the benefits of VaR relative to other methods, namely, that by taking the absolute value, it hides the direction of the positions.

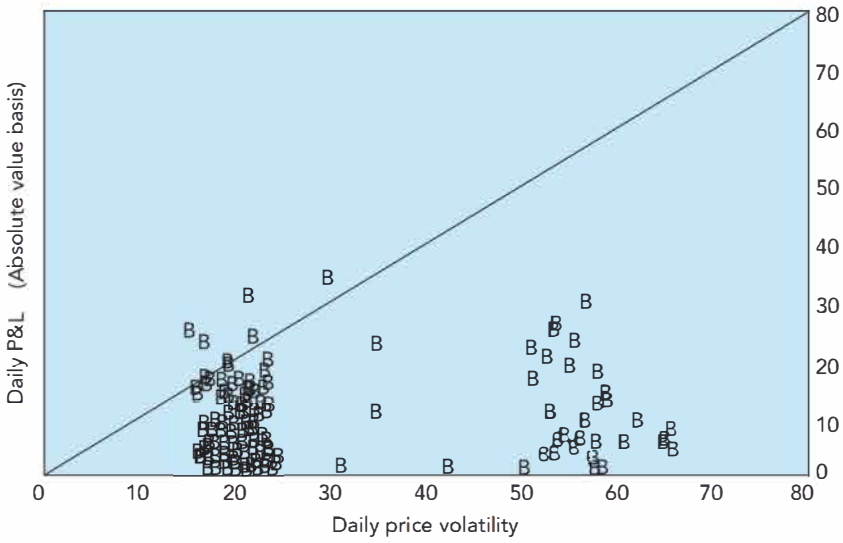


Figure 4.1 **Model evaluation: Bankers trust.**

practice, the trading portfolio evolves dynamically during the day. Thus the actual portfolio is "contaminated" by changes in its composition. The *actual return* corresponds to the actual P&L, taking into account intraday trades and other profit items such as fees, commissions, spreads, and net interest income.

This contamination will be minimized if the horizon is relatively short, which explains why backtesting usually is conducted on daily returns. Even so, intraday trading generally will increase the volatility of revenues because positions tend to be cut down toward the end of the trading day. Counterbalancing this is the effect of fee income, which generates steady profits that may not enter the VaR measure.

For verification to be meaningful, the risk manager should track both the actual portfolio return R_t and the hypothetical return R_t^* that most closely matches the VaR forecast. The *hypothetical return* R_t^* represents a frozen portfolio, obtained from fixed positions applied to the actual returns on all securities, measured from close to close.

Sometimes an approximation is obtained by using a *cleaned return*, which is the actual return minus all non-mark-to-market items, such as fees, commissions, and net interest income. Under the latest update to the *market-risk amendment*, supervisors will have the choice to use either hypothetical or cleaned returns.[2]

[2] See BCBS (2005b).

Since the VaR forecast really pertains to R^*, backtesting ideally should be done with these hypothetical returns. Actual returns do matter, though, because they entail real profits and losses and are scrutinized by bank regulators. They also reflect the true ex post volatility of trading returns, which is also informative. Ideally, both actual and hypothetical returns should be used for backtesting because both sets of numbers yield informative comparisons. If, for instance, the model passes backtesting with hypothetical but not actual returns, then the problem lies with intraday trading. In contrast, if the model does not pass backtesting with hypothetical returns, then the modeling methodology should be reexamined.

4.2 MODEL BACKTESTING WITH EXCEPTIONS

Model backtesting involves systematically comparing historical VaR measures with the subsequent returns. The problem is that since VaR is reported only at a specified confidence level, we expect the figure to be exceeded in some instances, for example, in 5 percent of the observations at the 95 percent confidence level. But surely we will not observe exactly 5 percent exceptions. A greater percentage could occur because of bad luck, perhaps 8 percent. At some point, if the frequency of deviations becomes too large, say, 20 percent, the user must conclude that the problem lies with the model, not bad luck, and undertake corrective action. The issue is how to make this decision. This *accept or reject decision* is a classic statistical decision problem.

At the outset, it should be noted that this decision must be made at some confidence level. The choice of this level for the *test*, however, is not related to the quantitative level p selected for VaR. The decision rule may involve, for instance, a 95 percent confidence level for backtesting VaR numbers, which are themselves constructed at some confidence level, say, 99 percent for the Basel rules.

Model Verification Based on Failure Rates

The simplest method to verify the accuracy of the model is to record the *failure rate*, which gives the proportion of times VaR is exceeded in a given sample. Suppose a bank provides a VaR figure at the 1 percent left-tail level ($p = 1 - c$) for a total of T days. The user then counts how many times the actual loss

exceeds the previous day's VaR. Define N as the number of exceptions and N/T as the failure rate. Ideally, the failure rate should give an *unbiased* measure of p, that is, should converge to p as the sample size increases.

We want to know, at a given confidence level, whether N is too small or too large under the null hypothesis that $p = 0.01$ in a sample of size T. Note that this test makes no assumption about the return distribution. The distribution could be normal, or skewed, or with heavy tails, or time-varying. We simply count the number of exceptions. As a result, this approach is fully *nonparametric*.

The setup for this test is the classic testing framework for a sequence of success and failures, also called *Bernoulli trials*. Under the null hypothesis that the model is correctly calibrated, the number of exceptions x follows a *binomial* probability distribution:

$$f(x) = \binom{T}{x} p^x (1-p)^{T-x} \tag{4.1}$$

We also know that x has expected value of $E(x) = pT$ and variance $V(x) = p(1-p)T$. When T is large, we can use the central limit theorem and approximate the binomial distribution by the normal distribution

$$Z = \frac{x - pT}{\sqrt{p(1-p)T}} \approx N(0, 1) \tag{4.2}$$

which provides a convenient shortcut. If the decision rule is defined at the two-tailed 95 percent test confidence level, then the cutoff value of $|z|$ is 1.96. Box 4.1 illustrates how this can be used in practice.

This binomial distribution can be used to test whether the number of exceptions is acceptably small. Figure 4.2 describes the distribution when the model is calibrated correctly, that is, when $p = 0.01$ and with 1 year of data, $T = 250$. The graph shows that under the null, we would observe more than four exceptions 10.8 percent of the time. The 10.8 percent number describes the probability of committing a *type 1* error, that is, rejecting a correct model.

Next, Figure 4.3 describes the distribution of number of exceptions when the model is calibrated incorrectly, that is, when $p = 0.03$ instead of 0.01. The graph shows that we will not reject the incorrect model more than 12.8 percent of the time. This describes the probability of committing a *type 2* error, that is, not rejecting an incorrect model.

BOX 4.1 J.P. MORGAN'S EXCEPTIONS

In its 1998 annual report, the U.S. commercial bank J.P. Morgan (JPM) explained that

> In 1998, daily revenue fell short of the downside (95 percent VaR) band . . . on 20 days, or more than 5 percent of the time. Nine of these 20 occurrences fell within the August to October period.

We can test whether this was bad luck or a faulty model, assuming 252 days in the year. Based on Equation (4.2), we have $z = (x - pT)/\sqrt{p(1-p)T} = (20 - 0.05 \times 252)/\sqrt{0.05(0.95)252} = 214$. This is larger than the cutoff value of 1.96. Therefore, we reject the hypothesis that the VaR model is unbiased. It is unlikely (at the 95 percent test confidence level) that this was bad luck.

The bank suffered too many exceptions, which must have led to a search for a better model. The flaw probably was due to the assumption of a normal distribution, which does not model tail risk adequately. Indeed, during the fourth quarter of 1998, the bank reported having switched to a "historical simulation" model that better accounts for fat tails. This episode illustrates how backtesting can lead to improved models.

When designing a verification test, the user faces a trade-off between these two types of error. Table 4.1 summarizes the two states of the world, correct versus incorrect model, and the decision. For backtesting purposes, users of VaR models need to balance type 1 errors against type 2 errors. Ideally, one would want to set a low type 1 error rate and then have a test that

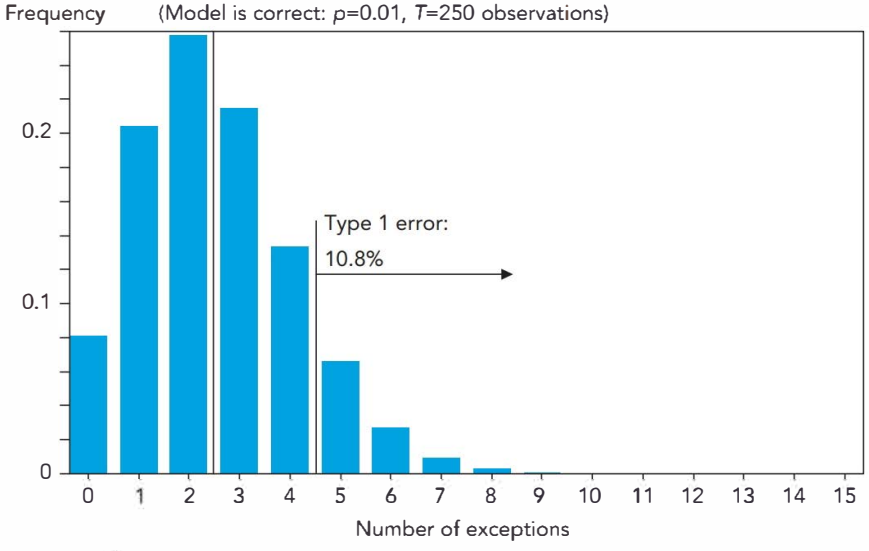


Figure 4.2 Distribution of exceptions when model is correct.

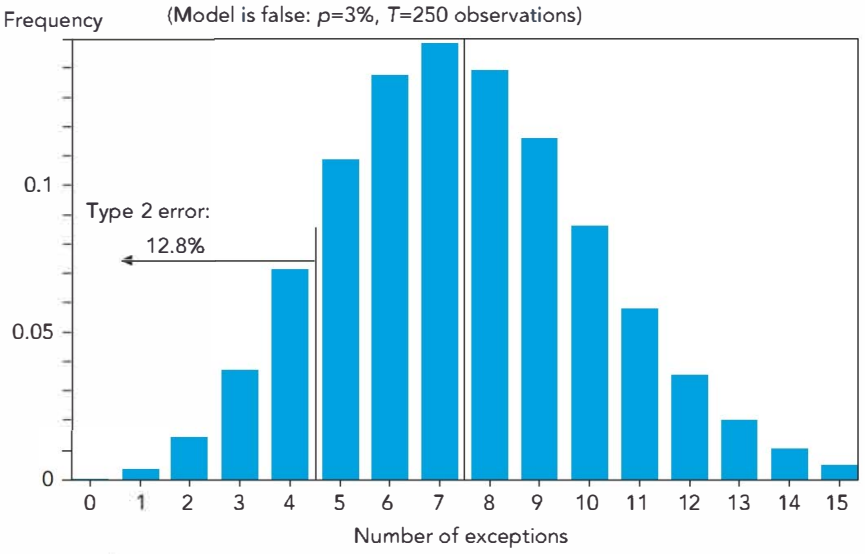


Figure 4.3 Distribution of exceptions when model is incorrect.

creates a very low type 2 error rate, in which case the test is said to be *powerful*. It should be noted that the choice of the confidence level for the decision rule is not related to the quantitative level p selected for VaR. This confidence level refers to the decision rule to reject the model.

Table 4.1 Decision Errors

	Model	
Decision	**Correct**	**Incorrect**
Accept	OK	Type 2 error
Reject	Type 1 error	OK

Kupiec (1995) develops approximate 95 percent confidence regions for such a test, which are reported in Table 4.2. These regions are defined by the tail points of the log-likelihood ratio:

$$LR_{uc} = -2\ln[(1-p)^{T-N}p^N] + 2\ln\{[1-(N/T)]^{T-N}(N/T)^N\} \quad (4.3)$$

which is asymptotically (i.e., when T is large) distributed chi-square with one degree of freedom under the null hypothesis that p is the true probability. Thus we would reject the null hypothesis if $LR > 3.841$. This test is equivalent to Equation (4.2) because a chi-square variable is the square of a normal variable.

In the JPM example, we had $N = 20$ exceptions over $T = 252$ days, using $p = 95$ percent VaR confidence level. Setting these numbers into Equation (4.3) gives $LR_{uc} = 3.91$. Therefore, we reject unconditional coverage, as expected.

For instance, with 2 years of data ($T = 510$), we would expect to observe $N = pT = 1$ percent times $510 = 5$ exceptions. But the VaR user will not be able to reject the null hypothesis as long as N is within the $[1 < N < 11]$ confidence interval. Values of N greater than or equal to 11 indicate that the VaR is too low or that the model understates the probability of large losses. Values of N less than or equal to 1 indicate that the VaR model is overly conservative.

The table also shows that this interval, expressed as a proportion N/T, shrinks as the sample size increases. Select, for instance, the $p = 0.05$ row. The interval for $T = 252$ is $[6/252 = 0.024, 20/252 = 0.079]$; for $T = 1000$, it is $[37/1000 = 0.037, 65/1000 = 0.065]$. Note how the interval

Table 4.2 Model Backtesting, 95 Percent Nonrejection Test Confidence Regions

		Nonrejection Region for Number of Failures N		
Probability level P	**VAR Confidence Level c**	**$T = 252$ Days**	**$T = 510$ Days**	**$T = 1000$ Days**
0.01	99%	$N < 7$	$1 < N < 11$	$4 < N < 17$
0.025	97.5%	$2 < N < 12$	$6 < N < 21$	$15 < N < 36$
0.05	95%	$6 < N < 20$	$16 < N < 36$	$37 < N < 65$
0.075	92.5%	$11 < N < 28$	$27 < N < 51$	$59 < N < 92$
0.10	90%	$16 < N < 36$	$38 < N < 65$	$81 < N < 120$

Note: N is the number of failures that could be observed in a sample size T without rejecting the null hypothesis that p is the correct probability at the 95 percent level of test confidence.

Source: Adapted from Kupiec (1995).

shrinks as the sample size extends. With more data, we should be able to reject the model more easily if it is false.

The table, however, points to a disturbing fact. For small values of the VaR parameter p, it becomes increasingly difficult to confirm deviations. For instance, the nonrejection region under $p = 0.01$ and $T = 252$ is $[N < 7]$. Therefore, there is no way to tell if N is abnormally small or whether the model systematically overestimates risk. Intuitively, detection of systematic biases becomes increasingly difficult for low values of p because the exceptions in these cases are very rare events.

This explains why some banks prefer to choose a higher VaR confidence level, such as c = 95 percent, in order to be able to observe sufficient numbers of deviations to validate the model. A multiplicative factor then is applied to translate the VaR figure into a safe capital cushion number. Too often, however, the choice of the confidence level appears to be made without regard for the issue of VaR backtesting.

The Basel Rules

This section now turns to a detailed analysis of the Basel Committee rules for backtesting. While we can learn much from the Basel framework, it is important to recognize that regulators operate under different constraints from financial institutions. Since they do not have access to every component of the models, the approach is perforce implemented at a broader level. Regulators are also responsible for constructing rules that are comparable across institutions.

The Basel (1996a) rules for backtesting the internal-models approach are derived directly from this failure rate test. To design such a test, one has to choose first the type 1 error rate, which is the probability of rejecting the model when it is correct. When this happens, the bank simply suffers bad luck and should not be penalized unduly. Hence one should pick a test with a low type 1 error rate, say, 5 percent (depending on its cost). The heart of the conflict is that, inevitably, the supervisor also will commit type 2 errors for a bank that willfully cheats on its VaR reporting.

The current verification procedure consists of recording daily exceptions of the 99 percent VaR over the last year. One would expect, on average, 1 percent of 250, or 2.5 instances of exceptions over the last year.

The Basel Committee has decided that up to four exceptions are acceptable, which defines a "green light" zone for the bank. If the number of exceptions is five or more, the bank falls into a "yellow" or "red" zone and incurs a progressive penalty whereby the multiplicative factor k is increased from 3 to 4, as described in Table 4.3. An incursion into the "red" zone generates an automatic penalty.

Within the "yellow" zone, the penalty is up to the supervisor, depending on the reason for the exception. The Basel Committee uses the following categories:

- *Basic integrity of the model.* The deviation occurred because the positions were reported incorrectly or because of an error in the program code.
- *Model accuracy could be improved.* The deviation occurred because the model does not measure risk with enough precision (e.g., has too few maturity buckets).
- *Intraday trading.* Positions changed during the day.
- *Bad luck.* Markets were particularly volatile or correlations changed.

The description of the applicable penalty is suitably vague. When exceptions are due to the first two reasons, the penalty "should" apply. With the third reason, a penalty "should be considered." When the deviation is traced to the fourth reason, the Basel document gives no guidance except that these exceptions should "be expected to occur at least some of the time." These exceptions may be excluded if they are the "result of such occurrences as sudden abnormal changes in interest rates or exchange rates, major political events, or natural disasters." In other words, bank supervisors want to keep the flexibility to adjust the rules in turbulent times as they see fit.

The crux of the backtesting problem is separating bad luck from a faulty model, or balancing type 1 errors against type 2 errors. Table 4.4 displays the probabilities of obtaining a given number of exceptions for a correct model (with 99 percent coverage) and incorrect model (with only 97 percent coverage). With five exceptions or more, the cumulative probability, or type 1 error rate, is 10.8 percent. This is rather high to start with. In the current framework, one bank out of 10 could be penalized even with a correct model.

Even worse, the type 2 error rate is also very high. Assuming a true 97 percent coverage, the supervisor will give passing grades

Table 4.3 The Basel Penalty Zones

Zone	Number of Exceptions	Increase in k
Green	0 to 4	0.00
Yellow	5	0.40
	6	0.50
	7	0.65
	8	0.75
	9	0.85
Red	10+	1.00

Table 4.4 Basel Rules for Backtesting, Probabilities of Obtaining Exceptions (T = 250)

Zone	Number of Exceptions N	Coverage = 99% Model Is Correct		Coverage = 97% Model Is Incorrect		
		Probability $P(X = N)$	Cumulative (Type 1) (Reject) $P(X \geq N)$	Probability $P(X = N)$	Cumulative (Type 2) (Do not reject) $P(X < N)$	Power (Reject) $P(X \geq N)$
Green	0	8.1	100.0	0.0	0.0	100.0
	1	20.5	91.9	0.4	0.0	100.0
	2	25.7	71.4	1.5	0.4	99.6
	3	21.5	45.7	3.8	1.9	98.1
Green	4	13.4	24.2	7.2	5.7	94.3
Yellow	5	6.7	10.8	10.9	12.8	87.2
	6	2.7	4.1	13.8	23.7	76.3
	7	1.0	1.4	14.9	37.5	62.5
	8	0.3	0.4	14.0	52.4	47.6
Yellow	9	0.1	0.1	11.6	66.3	33.7
Red	10	0.0	0.0	8.6	77.9	21.1
	11	0.0	0.0	5.8	86.6	13.4

to 12.8 percent of banks that have an incorrect model. The framework therefore is not very powerful. And this 99 versus 97 percent difference in VaR coverage is economically significant. Assuming a normal distribution, the true VaR would be 23.7 percent times greater than officially reported, which is substantial.

The lack of power of this framework is due to the choice of the high VaR confidence level (99 percent) that generates too few exceptions for a reliable test. Consider instead the effect of a 95 percent VaR confidence level. (To ensure that the amount of capital is not affected, we could use a larger multiplier k.) We now have to decide on the cutoff number of exceptions to have a type 1 error rate similar to the Basel framework. With an average of 13 exceptions per year, we choose to reject the model if the number of exceptions exceeds 17, which corresponds to a type 1 error of 12.5 percent. Here we controlled the error rate so that it is close to the 10.8 percent for the Basel framework. But now the probability of a type 2 error is lower, at 7.4 percent only.[3] Thus, simply changing the VaR confidence level from 99 to 95 percent sharply reduces the probability of not catching an erroneous model.

[3] Assuming again a normal distribution and a true VaR that is 23.7 percent greater than the reported VaR, for an alternative coverage of 90.8 percent.

Another method to increase the power of the test would be to increase the number of observations. With $T = 1000$, for instance, we would choose a cutoff of 14 exceptions, for a type 1 error rate of 13.4 percent and a type 2 error rate of 0.03 percent, which is now very small. Increasing the number of observations drastically improves the test.

Conditional Coverage Models

So far the framework focuses on *unconditional coverage* because it ignores conditioning, or time variation in the data. The observed exceptions, however, could cluster or "bunch" closely in time, which also should invalidate the model.

With a 95 percent VaR confidence level, we would expect to have about 13 exceptions every year. In theory, these occurrences should be evenly spread over time. If, instead, we observed that 10 of these exceptions occurred over the last 2 weeks, this should raise a red flag. The market, for instance, could experience increased volatility that is not captured by VaR. Or traders could have moved into unusual positions or risk "holes." Whatever the explanation, a verification system should be designed to measure proper *conditional coverage*, that is, conditional on current conditions. Management then can take the appropriate action.

Table 4.5 Building an Exception Table: Expected Number of Exceptions

	Conditional		
	Day Before		
	No Exception	Exception	Unconditional
Current day			
No exception	$T_{00} = T_0(1 - \pi_0)$	$T_{10} = T_1(1 - \pi_1)$	$T(1 - \pi)$
Exception	$T_{01} = T_0(\pi_0)$	$T_{11} = T_1(\pi_1)$	$T(\pi)$
Total	T_0	T_1	$T = T_0 + T_1$

Such a test has been developed by Christoffersen (1998), who extends the LR_{uc} statistic to specify that the deviations must be serially independent. The test is set up as follows: Each day we set a deviation indicator to 0 if VaR is not exceeded and to 1 otherwise. We then define T_{ij} as the number of days in which state j occurred in one day while it was at i the previous day and π_i as the probability of observing an exception conditional on state i the previous day. Table 4.5 shows how to construct a table of conditional exceptions.

If today's occurrence of an exception is independent of what happened the previous day, the entries in the second and third columns should be identical. The relevant test statistic is

$$LR_{ind} = -2 \ln [(1 - \pi)^{(T_{00}+T_{10})} \pi^{(T_{01}+T_{11})}] + 2 \ln [(1 - \pi_0)^{T_{00}} \pi_0^{T_{01}} (1 - \pi_1)^{T_{10}} \pi_1^{T_{11}}] \quad (4.4)$$

Here, the first term represents the maximized likelihood under hypothesis that exceptions are independent across days, or $\pi = \pi_0 = \pi_1 = (T_{01} + T_{11})/T$. The second term is the maximized likelihood for the observed data.

The combined test statistic for conditional coverage

then is

$$LR_{cc} = LR_{uc} + LR_{ind} \quad (4.5)$$

Each component is independently distributed as $\chi^2(1)$. asymptotically. The sum is distributed as $\chi^2(2)$. Thus we would reject at the 95 percent test confidence level if $LR > 5.991$. We would reject independence alone if $LR_{ind} > 3.841$.

As an example, assume that JPM observed the following pattern of exceptions during 1998. Of 252 days, we have 20 exceptions, which is a fraction of $\pi = 7.9$ percent. Of these, 6 exceptions occurred following an exception the previous day. Alternatively, 14 exceptions occurred when there was none the previous day. This defines conditional probability ratios of $\pi_0 = 14/232 = 6.0$ percent and $\pi_1 = 6/20 = 30.0$ percent. We seem to have a much higher probability of having an exception following another one. Setting these numbers into Equation (4.4), we find $LR_{ind} = 9.53$. Because this is higher than the cutoff value of 3.84, we reject independence. Exceptions do seem to cluster abnormally. As a result, the risk manager may want to explore models that allow for time variation in risk.

Extensions

We have seen that the standard exception tests often lack power, especially when the VaR confidence level is high and when the number of observations is low. This has led to a search for improved tests.

The problem, however, is that statistical decision theory has shown that this exception test is the most powerful among its class. More effective tests would have to focus on a different hypothesis or use more information.

For example, Crnkovic and Drachman (1996) developed a test focusing on the entire probability distribution, based on the *Kuiper statistic*. This test is still nonparametric but is more powerful. However, it uses other information than the VaR forecast at a given confidence level. Another approach is to focus on the time period between exceptions, called *duration*. Christoffersen and Pelletier (2004) show that duration-based tests can be more powerful than the standard test when risk is time-varying.

Finally, backtests could use parametric information instead. If the VaR is obtained from a multiple of the standard deviation, the risk manager could test the fit between the realized and forecast volatility. This would lead to more powerful tests because more information is used. Another useful avenue would be to backtest the portfolio components as well. From the viewpoint of the regulator, however, the only information provided is the daily VaR, which explains why exception tests are used most commonly nowadays.

4.3 APPLICATIONS

Berkowitz and O'Brien (2002) provide the first empirical study of the accuracy of internal VaR models, using data reported to U.S. regulators. They describe the distributions of P&L, which are compared with the VaR forecasts. Generally, the P&L distributions are symmetric, although they display fatter tails than the normal. Stahl et. al. (2006) also report that, although the components of a trading portfolio could be strongly

nonnormal, aggregation to the highest level of a bank typically produces symmetric distributions that resemble the normal.

Figure 4.4 plots the time series of P&L along with the daily VaR (the lower lines) for a sample of six U.S. commercial banks. With approximately 600 observations, we should observe on average 6 violations, given a VaR confidence level of 99 percent.

It is striking to see the abnormally small number of exceptions, even though the sample includes the turbulent 1998 period. Bank 4, for example, has zero exceptions over this sample. Its VaR is several times greater than the magnitude of extreme fluctuations in its P&L. Indeed, for banks 3 to 6, the average VaR is at least 60 percent higher than the actual 99th percentile of the P&L distribution. Thus banks report VaR measures that are *conservative*, or too large relative to their actual risks. These results are surprising because they imply that the banks' VaR and hence their market-risk charges are too high. Banks therefore allocate too much regulatory capital to their trading activities. Box 4.2 describes a potential explanation, which is simplistic.

Perhaps these observations could be explained by the use of actual instead of hypothetical returns.[4] Or maybe the models are too simple, for example failing to account for diversification effects. Yet another explanation is that capital requirements are currently not binding. The amount of economic capital U.S. banks currently hold is in excess of their regulatory capital. As a result, banks may prefer to report high VaR numbers to avoid the possibility of regulatory intrusion. Still, these practices impoverish the informational content of VaR numbers.

	Conditional		
	Day Before		
	No Exception	Exception	Unconditional
Current day			
No exception	218	14	232
Exception	14	6	20
Total	232	20	252

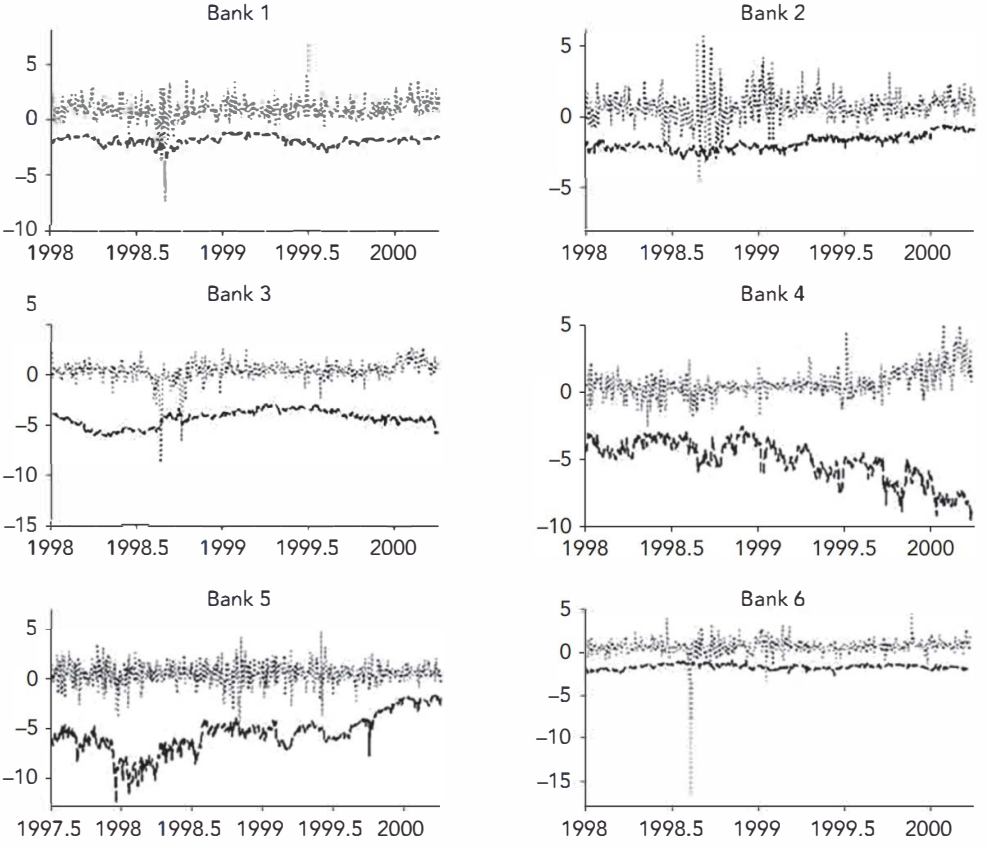


Figure 4.4 Bank VaR and trading profits.

4.4 CONCLUSIONS

Model verification is an integral component of the risk management process. Backtesting VaR numbers provides valuable feedback to users about the accuracy of their models. The procedure also can be used to search for possible improvements.

[4] Including fees increases the P&L, reducing the number of violations. Using hypothetical income, as currently prescribed in the European Union, could reduce this effect. Jaschke, Stahl, and Stehle (2003) compare the VaRs for 13 German banks and find that VaR measures are, on average, less conservative than for U.S. banks. Even so, VaR forecasts are still too high.

BOX 4.2 NO EXCEPTIONS

The CEO of a large bank receives a daily report of the bank's VaR and P&L. Whenever there is an exception, the CEO calls in the risk officer for an explanation.

Initially, the risk officer explained that a 99 percent VaR confidence level implies an average of 2 to 3 exceptions per year. The CEO is never quite satisfied, however. Later, tired of going "upstairs," the risk officer simply increases the confidence level to cut down on the number of exceptions.

Annual reports suggest that this is frequently the case. Financial institutions routinely produce plots of P&L that show no violation of their 99 percent confidence VaR over long periods, proclaiming that this supports their risk model.

Due thought should be given to the choice of VaR quantitative parameters for backtesting purposes. First, the horizon should be as short as possible in order to increase the number of observations and to mitigate the effect of changes in the portfolio composition. Second, the confidence level should not be too high because this decreases the effectiveness, or power, of the statistical tests.

Verification tests usually are based on "exception" counts, defined as the number of exceedences of the VaR measure. The goal is to check if this count is in line with the selected VaR confidence level. The method also can be modified to pick up bunching of deviations.

Backtesting involves balancing two types of errors: rejecting a correct model versus accepting an incorrect model. Ideally, one would want a framework that has very high power, or high probability of rejecting an incorrect model. The problem is that the power of exception-based tests is low. The current framework could be improved by choosing a lower VaR confidence level or by increasing the number of data observations.

Adding to these statistical difficulties, we have to recognize other practical problems. Trading portfolios do change over the horizon. Models do evolve over time as risk managers improve their risk modeling techniques. All this may cause further structural instability.

Despite all these issues, backtesting has become a central component of risk management systems. The methodology allows risk managers to improve their models constantly. Perhaps most important, backtesting should ensure that risk models do not go astray.

5

VaR Mapping

Learning Objectives

After completing this reading, you should be able to:

- Explain the principles underlying VaR mapping and describe the mapping process.
- Explain how the mapping process captures general and specific risks, and calculate these risks in a portfolio given a set of primitive risk factors.
- Differentiate among the three methods of mapping portfolios of fixed-income securities.
- Summarize how to map a fixed-income portfolio into positions of standard instruments.
- Describe how mapping of risk factors can support stress testing.
- Explain how VaR can be calculated and used relative to a performance benchmark.
- Describe the method of mapping forwards, forward rate agreements, interest rate swaps, and options.

Excerpt is Chapter 11 of Value at Risk: The New Benchmark for Managing Financial Risk, *Third Edition, by Philippe Jorion.*

The second [principle], to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution.

—René Descartes

Whichever value-at-risk (VaR) method is used, the risk measurement process needs to simplify the portfolio by *mapping* the positions on the selected risk factors. Mapping is the process by which the current values of the portfolio positions are replaced by exposures on the risk factors.

Mapping arises because of the fundamental nature of VaR, which is portfolio measurement at the highest level. As a result, this is usually a very large-scale aggregation problem. It would be too complex and time-consuming to model all positions individually as risk factors. Furthermore, this is unnecessary because many positions are driven by the same set of risk factors and can be aggregated into a small set of exposures without loss of risk information.

This chapter illustrates the mapping process for major financial instruments. First we review the basic principles behind mapping for VaR. We then proceed to illustrate cases where instruments are broken down into their constituent components. We will see that the mapping process is instructive because it reveals useful insights into the risk drivers of derivatives. The next sections deal with fixed-income securities and linear derivatives. We cover the most important instruments, forward contracts, forward rate agreements, and interest-rate swaps. Then we describe nonlinear derivatives, or options.

5.1 MAPPING FOR RISK MEASUREMENT

Why Mapping?

The essence of VaR is aggregation at the highest level. This generally involves a very large number of positions, including bonds, stocks, currencies, commodities, and their derivatives. As a result, it would be impractical to consider each position separately (see Box 5.1). Too many computations would be required, and the time needed to measure risk would slow to a crawl.

Fortunately, mapping provides a shortcut. Many positions can be simplified to a smaller number of positions on a set of elementary, or *primitive*, risk factors. Consider, for instance, a trader's desk with thousands of open dollar/euro forward contracts. The positions may differ owing to different maturities and delivery prices. It is unnecessary, however, to model all these positions individually. Basically, the positions are exposed to a single major risk factor, which is the dollar/euro spot exchange rate. Thus they could be summarized by a single aggregate exposure on this risk factor. Such aggregation, of course, is not appropriate for the pricing of the portfolio. For risk measurement purposes, however, it is perfectly acceptable. This is why risk management methods can differ from pricing methods.

Mapping is also the only solution when the characteristics of the instrument change over time. The risk profile of bonds, for instance, changes as they age. One cannot use the history of prices on a bond directly. Instead, the bond must be mapped on yields that best represent its current profile. Similarly, the risk profile of options changes very quickly. Options must be mapped on their primary risk factors. Mapping provides a way to tackle these practical problems.

BOX 5.1 WHY MAPPING?

"J.P. Morgan Chase's VaR calculation is highly granular, comprising more than 2.1 million positions and 240,000 pricing series (e.g., securities prices, interest rates, foreign exchange rates)." (Annual report, 2004)

Mapping as a Solution to Data Problems

Mapping is also required in many common situations. Often a complete history of all securities may not exist or may not be relevant. Consider a mutual fund with a strategy of investing in *initial public offerings* (IPOs) of common stock. By definition, these stocks have no history. They certainly cannot be ignored in the risk system, however. The risk manager would have to replace these positions by exposures on similar risk factors already in the system.

Another common problem with global markets is the time at which prices are recorded. Consider, for instance, a portfolio or mutual funds invested in international stocks. As much as 15 hours can elapse from the time the market closes in Tokyo at 1:00 A.M. EST (3:00 P.M. in Japan) to the time it closes in the United States at 4:00 P.M. As a result, prices from the Tokyo close ignore intervening information and are said to be *stale*. This led to the mutual-fund scandal of 2003, which is described in Box 5.2.

For risk managers, stale prices cause problems. Because returns are not synchronous, daily correlations across markets are too low, which will affect the measurement of portfolio risk.

BOX 5.2 MARKET TIMING AND STALE PRICES

In September 2003, New York Attorney General Eliot Spitzer accused a number of investment companies of allowing *market timing* into their funds. Market timing is a short-term trading strategy of buying and selling the same funds.

Consider, for example, our portfolio of Japanese and U.S. stocks, for which prices are set in different time zones. The problem is that U.S. investors can trade up to the close of the U.S. market. *Market timers* could take advantage of this discrepancy by rapid trading. For instance, if the U.S. market moves up following good news, it is likely the Japanese market will move up as well the following day. Market timers would buy the fund at the stale price and resell it the next day. Such trading, however, creates transactions costs that are borne by the other investors in the fund. As a result, fund companies usually state in their prospectus that this practice is not allowed. In practice, Eliot Spitzer found out that many mutual-fund companies had encouraged market timers, which he argued was fraudulent. Eventually, a number of funds settled by paying more than USD 2 billion.

This practice can be stopped in a number of ways. Many mutual funds now impose short-term redemption fees, which make market timing uneconomical. Alternatively, the cutoff time for placing trades can be moved earlier.

One possible solution is mapping. For instance, prices at the close of the U.S. market can be estimated from a regression of Japanese returns on U.S. returns and using the forecast value conditional on the latest U.S. information. Alternatively, correlations can be measured from returns taken over longer time intervals, such as weekly. In practice, the risk manager needs to make sure that the data-collection process will lead to meaningful risk estimates.

The Mapping Process

Figure 5.1 illustrates a simple mapping process, where six instruments are mapped on three risk factors. The first step in the analysis is marking all positions to market in current dollars or whatever reference currency is used. The market value for each instrument then is allocated to the three risk factors.

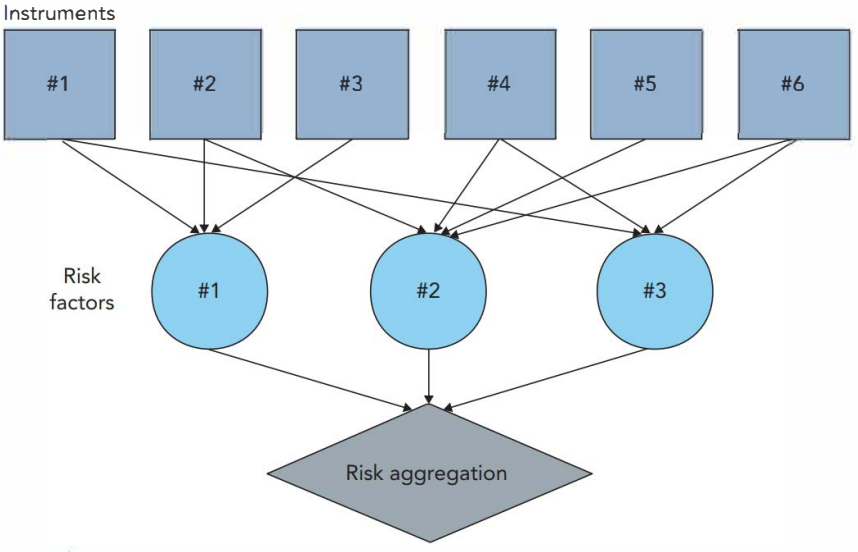


Figure 5.1 Mapping instruments on risk factors.

Table 5.1 shows that the first instrument has a market value of V_1, which is allocated to three exposures, x_{11}, x_{12}, and x_{13}. If the current market value is not fully allocated to the risk factors, it must mean that the remainder is allocated to cash, which is not a risk factor because it has no risk.

Table 5.1 Mapping Exposures

		Exposure on Risk Factor		
	Market Value	**1**	**2**	**3**
Instrument 1	V_1	x_{11}	x_{12}	x_{13}
Instrument 2	V_2	x_{21}	x_{22}	x_{23}
⋮	⋮	⋮	⋮	⋮
Instrument 6	V_6	x_{61}	x_{62}	x_{63}
Total portfolio	V	$x_1 = \sum_{i=1}^{6} x_{i1}$	$x_2 = \sum_{i=1}^{6} x_{i2}$	$x_3 = \sum_{i=1}^{6} x_{i3}$

Next, the system allocates the position for instrument 2 and so on. At the end of the process, positions are summed for each risk factor. For the first risk factor, the dollar exposure is $x_1 = \sum_{i=1}^{6} x_{i1}$. This creates a vector x of three exposures that can be fed into the risk measurement system.

Mapping can be of two kinds. The first provides an exact allocation of exposures on the risk factors. This is obtained for derivatives, for instance, when the price is an exact function of the risk factors. As we shall see in the rest of this chapter, the partial derivatives of the price function generate *analytical* measures of exposures on the risk factors.

Alternatively, exposures may have to be estimated. This occurs, for instance, when a stock is replaced by a position in the stock index. The exposure then is estimated by the slope coefficient from a regression of the stock return on the index return.

General and Specific Risk

This brings us to the issue of the choice of the set of primitive risk factors. This choice should reflect the trade-off between better quality of the approximation and faster processing. More factors lead to tighter risk measurement but also require more time devoted to the modeling process and risk computation.

The choice of primitive risk factors also influences the size of specific risks. *Specific risk* can be defined as risk that is due to issuer-specific price movements, after accounting for general market factors. Hence the definition of specific risk depends on that of general market risk. The Basel rules have a separate charge for specific risk.[1]

To illustrate this decomposition, consider a portfolio of N stocks. We are mapping each stock on a position in the stock market index, which is our primitive risk factor. The return on a stock Ri is regressed on the return on the stock market index R_m, that is,

$$R_i = \alpha_i + \beta_i R_m + \epsilon_i \tag{5.1}$$

which gives the exposure β_i. In what follows, ignore α, which does not contribute to risk. We assume that the specific risk owing to ϵ is not correlated across stocks or with the market. The relative weight of each stock in the portfolio is given by w_i. Thus the portfolio return is

$$R_p = \sum_{i=1}^{N} w_i R_i = \sum_{i=1}^{N} w_i \beta_i R_m + \sum_{i=1}^{N} w_i \epsilon_i \tag{5.2}$$

These exposures are aggregated across all the stocks in the portfolio. This gives

$$\beta_p = \sum_{i=1}^{N} w_i \beta_i \tag{5.3}$$

If the portfolio value is W, the mapping on the index is $x = W\beta_p$.

Next, we decompose the variance R_p in Equation (5.2) and find

$$V(R_p) = (\beta_p^2)V(R_m) + \sum_{i=1}^{N} w_i^2 \sigma_{\epsilon i}^2 \tag{5.4}$$

The first component is the general market risk. The second component is the aggregate of specific risk for the entire portfolio. This decomposition shows that with more detail on the primitive or general-market risk factors, there will be less specific risk for a fixed amount of total risk $V(R_p)$.

As another example, consider a corporate bond portfolio. Bond positions describe the distribution of money flows over time by their amount, timing, and credit quality of the issuer. This creates a continuum of risk factors, going from overnight to long maturities for various credit risks.

In practice, we have to restrict the number of risk factors to a small set. For some portfolios, one risk factor may be sufficient. For others, 15 maturities may be necessary. For portfolios with options, we need to model movements not only in yields but also in their implied volatilities.

Our primitive risk factors could be movements in a set of J government bond yields z_j and in a set of K credit spreads s_k sorted by credit rating. We model the movement in each corporate bond yield dy_i by a movement in z at the closest maturity and in s for the same credit rating. The remaining component is ϵ_i.

The movement in value W then is

$$dW = \sum_{i=1}^{N} \text{DVBP}_i dy_i = \sum_{j=1}^{J} \text{DVBP}_j dz_j + \sum_{k=1}^{K} \text{DVBP}_k ds_k + \sum_{i=1}^{N} \text{DVBP}_i d\epsilon_i \tag{5.5}$$

where DVBP is the total dollar value of a basis point for the associated risk factor. The values for DVBP_j then represent the summation of the DVBP across all individual bonds for each maturity.

This leads to a total risk decomposition of

$$V(dW) = \text{general risk} + \sum_{i=1}^{N} \text{DVBP}_i^2 V(d\epsilon_i) \tag{5.6}$$

A greater number of general risk factors should create less residual risk. Even so, we need to ascertain the size of the second, specific risk term. In practice, there may not be sufficient history to measure the specific risk of individual bonds, which is why it is often assumed that all issuers within the same risk class have the same risk.

[1] Typically, the charge is 4 percent of the position value for equities and unrated debt, assuming that the banks' models do not incorporate specific risks.

5.2 MAPPING FIXED-INCOME PORTFOLIOS

Mapping Approaches

Once the risk factors have been selected, the question is how to map the portfolio positions into exposures on these risk factors. We can distinguish three mapping systems for fixed-income portfolios: principal, duration, and cash flows. With *principal mapping*, one risk factor is chosen that corresponds to the average portfolio maturity. With *duration mapping*, one risk factor is chosen that corresponds to the portfolio duration. With *cash-flow mapping*, the portfolio cash flows are grouped into maturity buckets. Mapping should preserve the market value of the position. Ideally, it also should preserve its market risk.

As an example, Table 5.2 describes a two-bond portfolio consisting of a USD 100 million 5-year 6 percent issue and a USD 100 million 1-year 4 percent issue. Both issues are selling at par, implying a market value of USD 200 million.

The portfolio has an average maturity of 3 years and a duration of 2.733 years. The table lays out the present value of all portfolio cash flows discounted at the appropriate zero-coupon rate.

Principal mapping considers the timing of redemption payments only. Since the average maturity of this portfolio is 3 years, the VaR can be found from the risk of a 3-year maturity, which is 1.484 percent from Table 5.3. VaR then is USD 200 × 1.484/100 = USD 2.97 million. The only positive aspect of this method is its simplicity. This approach overstates the true risk because it ignores intervening coupon payments.

The next step in precision is duration mapping. We replace the portfolio by a zero-coupon bond with maturity equal to the duration of the portfolio, which is 2.733 years. Table 5.3 shows VaRs of 0.987 and 1.484 for these maturities, respectively. Using a linear interpolation, we find a risk of 0.987 + (1.484 − 0.987) × (2.733 − 2) = 1.351 percent for this hypothetical zero. With a USD 200 million portfolio, the duration-based VaR is USD 200 × 1.351/100 = USD 2.70 million, slightly less than before.

Table 5.2 Mapping for a Bond Portfolio (USD millions)

	Cash Flows			Mapping (PV)		
Term (Year)	5-Year	1-Year	Spot Rate	Principal	Duration	Cash Flow
1	USD 6	USD 104	4.000%	0.00	0.00	USD 105.77
2	USD 6	0	4.618%	0.00	0.00	USD 5.48
2.733	—	—		—	USD 200.00	—
3	USD 6	0	5.192%	USD 200.00	0.00	USD 5.15
4	USD 6	0	5.716%	0.00	0.00	USD 4.80
5	USD 106	0	6.112%	0.00	0.00	USD 78.79
Total				USD 200.00	USD 200.00	USD 200.00

Table 5.3 Computing VaR from Change in Prices of Zeroes

Term (Year)	Cash Flows	Old Zero Value	Old PV of Flows	Risk (%)	New Zero Value	New PV of Flows
1	USD 110	0.9615	USD 105.77	0.4696	0.9570	USD 105.27
2	USD 6	0.9136	USD 5.48	0.9868	0.9046	USD 5.43
3	USD 6	0.8591	USD 5.15	1.4841	0.8463	USD 5.08
4	USD 6	0.8006	USD 4.80	1.9714	0.7848	USD 4.71
5	USD 106	0.7433	USD 78.79	2.4261	0.7252	USD 76.88
Total			USD 200.00			USD 197.37
Loss						USD 2.63

Finally, the cash-flow mapping method consists of grouping all cash flows on term-structure "vertices" that correspond to maturities for which volatilities are provided. Each cash flow is represented by the present value of the cash payment, discounted at the appropriate zero-coupon rate.

The diversified VaR is computed as

$$\text{VaR} = \alpha\sqrt{x'\Sigma x} = \sqrt{(x \times V)'R(x \times V)} \qquad (5.7)$$

where $V = \alpha\sigma$ is the vector of VaR for zero-coupon bond returns, and R is the correlation matrix.

Table 5.4 shows how to compute the portfolio VaR using cash-flow mapping. The second column reports the cash flows x from Table 5.2. Note that the current value of USD 200 million is fully allocated to the five risk factors. The third column presents the product of these cash flows with the risk of each vertex $x \times V$, which represents the individual VaRs.

With perfect correlation across all zeros, the VaR of the portfolio is

$$\text{Undiversified VaR} = \sum_{i=1}^{N} |x_i| V_i$$

which is USD 2.63 million. This number is close to the VaR obtained from the duration approximation, which was USD 2.70 million.

The right side of the table presents the correlation matrix of zeroes for maturities ranging from 1 to 5 years. To obtain the portfolio VaR, we premultiply and postmultiply the matrix by the dollar amounts (xV) at each vertex. Taking the square root, we find a diversified VaR measure of USD 2.57 million.

Note that this is slightly less than the duration VaR of USD 2.70 million. This difference is due to two factors. First, risk measures are not perfectly linear with maturity, as we have seen in a previous section. Second, correlations are below unity, which reduces risk even further. Thus, of the USD 130,000 difference in these measures, (USD 2.70 − USD 2.57 million), USD 70,000 is due to differences in yield volatility, and (USD 2.70 − USD 2.63 million), USD 60,000 is due to imperfect correlations. The last column presents the component VaR using computations as explained earlier.

Stress Test

Table 5.3 presents another approach to VaR that is directly derived from movements in the value of zeroes. This is an example of stress testing.

Assume that all zeroes are perfectly correlated. Then we could decrease all zeroes' values by their VaR. For instance, the 1-year zero is worth 0.9615. Given the VaR in Table 5.3 of 0.4696, a 95 percent probability move would be for the zero to fall to $0.9615 \times (1 - 0.4696/100) = 0.9570$. If all zeroes are perfectly correlated, they should all fall by their respective VaR. This generates a new distribution of present-value factors that can be used to price the portfolio. Table 5.3 shows that the new value is USD 197.37 million, which is exactly USD 2.63 million below the original value. This number is exactly the same as the undiversified VaR just computed.

The two approaches illustrate the link between computing VaR through matrix multiplication and through movements in underlying prices. Computing VaR through matrix multiplication is much more direct, however, and more appropriate because it allows nonperfect correlations across different sectors of the yield curve.

Benchmarking

Next, we provide a practical fixed-income compute VaR in relative terms, that is, relative to a performance benchmark. Table 5.5 presents the cash-flow decomposition of the J.P. Morgan U.S. bond index, which has a duration of 4.62 years.

Table 5.4 Computing the VaR of a USD 200 Million Bond Portfolio (Monthly VaR at 95 Percent Level)

	PV Cash Flows	Individual VaR	Correlation Matrix R					Component VaR
Term (Year)	x	$x \times V$	1Y	2Y	3Y	4Y	5Y	$x\Delta$VaR
1	USD 105.77	0.4966	1					USD 0.45
2	USD 5.48	0.0540	0.897	1				USD 0.05
3	USD 5.15	0.0765	0.886	0.991	1			USD 0.08
4	USD 4.80	0.0947	0.866	0.976	0.994	1		USD 0.09
5	USD 78.79	1.9115	0.855	0.966	0.988	0.998	1	USD 1.90
Total	USD 200.00	2.6335						
Undiversified VaR		USD 2.63						
Diversified VaR								USD 2.57

Table 5.5 Benchmarking a USD 100 Million Bond Index (Monthly Tracking Error VaR at 95 Percent Level)

			Position: Portfolio				
Vertex	**Risk (%)**	**Position: Index (USD)**	**1 (USD)**	**2 (USD)**	**3 (USD)**	**4 (USD)**	**5 (USD)**
≤1m	0.022	1.05	0.0	0.0	0.0	0.0	84.8
3m	0.065	1.35	0.0	0.0	0.0	0.0	0.0
6m	0.163	2.49	0.0	0.0	0.0	0.0	0.0
1Y	0.470	13.96	0.0	0.0	0.0	59.8	0.0
2Y	0.987	24.83	0.0	0.0	62.6	0.0	0.0
3Y	1.484	15.40	0.0	59.5	0.0	0.0	0.0
4Y	1.971	11.57	38.0	0.0	0.0	0.0	0.0
5Y	2.426	7.62	62.0	0.0	0.0	0.0	0.0
7Y	3.192	6.43	0.0	40.5	0.0	0.0	0.0
9Y	3.913	4.51	0.0	0.0	37.4	0.0	0.0
10Y	4.250	3.34	0.0	0.0	0.0	40.2	0.0
15Y	6.234	3.00	0.0	0.0	0.0	0.0	0.0
20Y	8.146	3.15	0.0	0.0	0.0	0.0	0.0
30Y	11.119	1.31	0.0	0.0	0.0	0.0	15.2
Total		100.00	100.0	100.0	100.0	100.0	100.0
Duration		4.62	4.62	4.62	4.62	4.62	4.62
Absolute VaR		USD 1.99	USD 2.25	USD 2.16	USD 2.04	USD 1.94	USD 1.71
Tracking error VaR		USD 0.00	USD 0.43	USD 0.29	USD 0.16	USD 0.20	USD 0.81

Assume that we are trying to benchmark a portfolio of USD 100 million. Over a monthly horizon, the VaR of the index at the 95 percent confidence level is USD 1.99 million. This is about equivalent to the risk of a 4-year note.

Next, we try to match the index with two bonds. The rightmost columns in the table display the positions of two-bond portfolios with duration matched to that of the index. Since no zero-coupon has a maturity of exactly 4.62 years, the closest portfolio consists of two positions, each in a 4- and a 5-year zero. The respective weights for this portfolio are USD 38 million and USD 62 million.

Define the new vector of positions for this portfolio as x and for the index as x_0. The VaR of the deviation relative to the benchmark is

$$\text{Tracking Error VaR} = \alpha\sqrt{(x - x_0)'\Sigma(x - x_0)} \qquad (5.8)$$

After performing the necessary calculations, we find that the *tracking error* VaR (TE-VaR) of this duration-hedged portfolio is USD 0.43 million. Thus the maximum deviation between the index and the portfolio is at most USD 0.43 million under normal market conditions. This potential shortfall is much less than the USD 1.99 million absolute risk of the index. The remaining tracking error is due to nonparallel moves in the term structure.

Relative to the original index, the tracking error can be measured in terms of variance reduction, similar to an R^2 in a regression. The variance improvement is

$$1 - \left(\frac{0.43}{1.99}\right)^2 = 95.4 \text{ percent}$$

which is in line with the explanatory power of the first factor in the variance decomposition.

Next, we explore the effect of altering the composition of the tracking portfolio. Portfolio 2 widens the bracket of cash flows in years 3 and 7. The TE-VaR is USD 0.29 million, which is an improvement over the previous number. Next, portfolio 3 has positions in years 2 and 9. This comes the closest to approximating the cash-flow positions in the index, which has the greatest weight on the 2-year vertex. The TE-VaR is reduced further to USD 0.16 million. Portfolio 4 has positions in years 1 and 10. Now the TE-VaR increases to USD 0.20 million. This mistracking is even more pronounced for a portfolio consisting of 1-month bills and 30-year zeroes, for which the TE-VaR increases to USD 0.81 million.

Among the portfolios considered here, the lowest tracking error is obtained with portfolio 3. Note that the absolute risk of these portfolios is lowest for portfolio 5. As correlations decrease for more distant maturities, we should expect that a duration-matched portfolio should have the lowest absolute risk for the combination of most distant maturities, such as a *barbell* portfolio of cash and a 30-year zero. However, minimizing absolute market risk is not the same as minimizing relative market risk.

This example demonstrates that duration hedging only provides a first approximation to interest-rate risk management. If the goal is to minimize tracking error relative to an index, it is essential to use a fine decomposition of the index by maturity.

5.3 MAPPING LINEAR DERIVATIVES

Forward Contracts

Forward and futures contracts are the simplest types of derivatives. Since their value is linear in the underlying spot rates, their risk can be constructed easily from basic building blocks. Assume, for instance, that we are dealing with a forward contract on a foreign currency. The basic valuation formula can be derived from an arbitrage argument.

To establish notations, define

S_t = spot price of one unit of the underlying cash asset

K = contracted forward price

r = domestic risk-free rate

y = income flow on the asset

τ = time to maturity.

When the asset is a foreign currency, y represents the foreign risk-free rate r^*. We will use these two notations interchangeably. For convenience, we assume that all rates are compounded continuously.

We seek to find the current value of a forward contract f_t to buy one unit of foreign currency at K after time τ. To do this, we consider the fact that investors have two alternatives that are economically equivalent: (1) Buy $e^{-y\tau}$ units of the asset at the price S_t and hold for one period, or (2) enter a forward contract to buy one unit of the asset in one period. Under alternative 1, the investment will grow, with reinvestment of dividend, to exactly one unit of the asset after one period. Under alternative 2, the contract costs f_t upfront, and we need to set aside enough cash to pay K in the future, which is $Ke^{-r\tau}$. After 1 year, the two alternatives lead to the same position, one unit of the asset. Therefore, their initial cost must be identical. This leads to the following valuation formula for outstanding forward contracts:

$$f_t = S_t e^{-y\tau} - Ke^{-r\tau} \tag{5.9}$$

Note that we can repeat the preceding reasoning to find the current forward rate F_t that would set the value of the contract to zero. Setting $K = F_t$ and $f_t = 0$ in Equation (5.9), we have

$$F_t = (S_t e^{-y\tau})e^{r\tau} \tag{5.10}$$

This allows us to rewrite Equation (5.9) as

$$f_t = F_t e^{-r\tau} - Ke^{-r\tau} = (F_t - K)e^{-r\tau} \tag{5.11}$$

In other words, the current value of the forward contract is the present value of the difference between the current forward rate and the locked-in delivery rate. If we are long a forward contract with contracted rate K, we can liquidate the contract by entering a new contract to sell at the current rate F_t. This will lock in a profit of $(F_t - K)$, which we need to discount to the present time to find f_t.

Let us examine the risk of a 1-year forward contract to purchase 100 million euros in exchange for USD 130.086 million. Table 5.6 displays pricing information for the contract (current spot, forward, and interest rates), risk, and correlations. The first step is to find the market value of the contract. We can use Equation (5.9), accounting for the fact that the quoted interest rates are discretely compounded, as

$$f_t = \text{USD } 1.2877 \frac{1}{(1 + 2.2810/100)} - \text{USD } 1.3009 \frac{1}{(1 + 3.3304/100)}$$

$$= \text{USD } 1.2589 - \text{USD } 1.2589 = 0$$

Table 5.6 Risk and Correlations for Forward Contract Risk Factors (Monthly VaR at 95 Percent Level)

			Correlations		
Risk Factor	**Price or Rate**	**VaR (%)**	**EUR Spot**	**EUR 1Y**	**USD 1Y**
EUR spot	USD 1.2877	4.5381	1	0.1289	0.0400
Long EUR bill	2.2810%	0.1396	0.1289	1	−0.0583
Short USD bill	3.3304%	0.2121	0.0400	−0.0583	1
EUR forward	USD 1.3009				

Thus the initial value of the contract is zero. This value, however, may change, creating market risk.

Among the three sources of risk, the volatility of the spot contract is the highest by far, with a 4.54 percent VaR (corresponding to 1.65 standard deviations over a month for a 95 percent confidence level). This is much greater than the 0.14 percent VaR for the EUR 1-year bill or even the 0.21 percent VaR for the USD bill. Thus most of the risk of the forward contract is driven by the cash EUR position.

But risk is also affected by correlations. The positive correlation of 0.13 between the EUR spot and bill positions indicates that when the EUR goes up in value against the dollar, the value of a 1-year EUR investment is likely to appreciate. Therefore, higher values of the EUR are associated with lower EUR interest rates.

This positive correlation increases the risk of the combined position. On the other hand, the position is also short a 1-year USD bill, which is correlated with the other two legs of the transaction. The issue is, what will be the net effect on the risk of the forward contract?

VaR provides an exact answer to this question, which is displayed in Table 5.7. But first we have to compute the positions x on each of the three building blocks of the contract. By taking the partial derivative of Equation (5.9) with respect to the risk factors, we have

$$\begin{aligned} df &= \frac{\partial f}{\partial S}dS + \frac{\partial f}{\partial r^*}dr^* + \frac{\partial f}{\partial r}dr \\ &= e^{-r^*\tau}dS - Se^{-r^*\tau}\tau dr^* + Ke^{-r\tau}\tau dr \end{aligned} \tag{5.12}$$

Here, the building blocks consist of the spot rate and interest rates. Alternatively, we can replace interest rates by the price of bills. Define these as $P = e^{-r\tau}$ and $P^* = e^{-r^*\tau}$. We then replace dr with dP using $dP = (-\tau)e^{-r\tau}\,dr$ and $dP^* = (-\tau)e^{-r^*\tau}\,dr^*$. The risk of the forward contract becomes

$$df = (Se^{-r^*\tau})\frac{dS}{S} + (Se^{-r^*\tau})\frac{dP^*}{P^*} - (Ke^{-r\tau})\frac{dP}{P} \tag{5.13}$$

This shows that the forward position can be separated into three cash flows: (1) a long spot position in EUR, worth EUR 100 million = USD 130.09 million in a year, or $(Se^{-r^*\tau})$ = USD 125.89 million now, (2) a long position in a EUR investment, also worth USD 125.89 million now, and (3) a short position in a USD investment, worth USD 130.09 million in a year, or $(Ke^{-r\tau})$ = USD 125.89 million now. Thus a position in the forward contract has three building blocks:

Long forward contract = long foreign currency spot + long foreign currency bill + short U.S.dollar bill

Considering only the spot position, the VaR is USD 125.89 million times the risk of 4.538 percent, which is USD 5.713 million. To compute the diversified VaR, we use the risk matrix from the data in Table 5.7 and pre- and postmultiply by the vector of positions (PV of flows column in the table). The total VaR for the forward contract is USD 5.735 million. This number is about the same size as that of the spot contract because exchange-rate volatility dominates the volatility of 1-year bonds.

More generally, the same methodology can be used for long-term currency swaps, which are equivalent to portfolios of forward contracts. For instance, a 10-year contract to pay dollars and receive euros is equivalent to a series of 10 forward contracts to exchange a set amount of dollars into euros. To compute the VaR, the contract must be broken down into a currency-risk component and a string of USD and EUR fixed-income components. As before, the total VaR will be driven primarily by the currency component.

Commodity Forwards

The valuation of forward or futures contracts on commodities is substantially more complex than for financial assets such as currencies, bonds, or stock indices. Such financial assets have a well-defined income flow y, which is the foreign interest rate, the coupon payment, or the dividend yield, respectively.

Table 5.7 Computing VaR for a EUR 100 Million Forward Contract (Monthly VaR at 95 Percent Level)

Position	Present-Value Factor	Cash Flows (CF)	PV of Flows, x	Individual VaR, $\lvert x \rvert V$	Component VaR, $x\Delta$VaR
EUR spot			USD 125.89	USD 5.713	USD 5.704
Long EUR bill	0.977698	EUR100.00	USD 125.89	USD 0.176	USD 0.029
Short USD bill	0.967769	− USD 130.09	−USD 125.89	USD 0.267	USD 0.002
Undiversified VaR				USD 6.156	
Diversified VaR					USD 5.735

Table 5.8 Risk of Commodity Contracts (Monthly VaR at 95 Percent Level)

	Energy Products			
Maturity	**Natural Gas**	**Heating Oil**	**Unleaded Gasoline**	**Crude Oil–WTI**
1 month	28.77	22.07	20.17	19.20
3 months	22.79	20.60	18.29	17.46
6 months	16.01	16.67	16.26	15.87
12 months	12.68	14.61	—	14.05
	Base Metals			
Maturity	**Aluminum**	**Copper**	**Nickel**	**Zinc**
Cash	11.34	13.09	18.97	13.49
3 months	11.01	12.34	18.41	13.18
15 months	8.99	10.51	15.44	11.95
27 months	7.27	9.57	—	11.59
	Precious Metals			
Maturity	**Gold**	**Silver**	**Platinum**	
Cash	6.18	14.97	7.70	

Things are not so simple for commodities, such as metals, agricultural products, or energy products. Most products do not make monetary payments but instead are consumed, thus creating an implied benefit. This flow of benefit, net of storage cost, is loosely called *convenience yield* to represent the benefit from holding the cash product. This convenience yield, however, is not tied to another financial variable, such as the foreign interest rate for currency futures. It is also highly variable, creating its own source of risk.

As a result, the risk measurement of commodity futures uses Equation (5.11) directly, where the main driver of the value of the contract is the current forward price for this commodity. Table 5.8 illustrates the term structure of volatilities for selected energy products and base metals. First, we note that monthly VaR measures are very high, reaching 29 percent for near contracts. In contrast, currency and equity market VaRs are typically around 6 percent. Thus commodities are much more volatile than typical financial assets.

Second, we observe that volatilities decrease with maturity. The effect is strongest for less storable products such as energy products and less so for base metals. It is actually imperceptible for precious metals, which have low storage costs and no convenience yield. For financial assets, volatilities are driven primarily by spot prices, which implies basically constant volatilities across contract maturities.

Let us now say that we wish to compute the VaR for a 12-month forward position on 1 million barrels of oil priced at USD 45.2 per barrel. Using a present-value factor of 0.967769, this translates into a current position of USD 43,743,000.

Differentiating Equation (5.11), we have

$$df = \frac{\partial f}{\partial F} dF = e^{-r\tau} dF = (e^{-r\tau} F) \frac{dF}{F} \tag{5.14}$$

The term between parentheses therefore represents the exposure. The contract VaR is

$$\text{VaR} = \text{USD } 43{,}743{,}000 \times 14.05/100 = \text{USD } 6{,}146{,}000$$

In general, the contract cash flows will fall between the maturities of the risk factors, and present values must be apportioned accordingly.

Forward Rate Agreements

Forward rate agreements (FRAs) are forward contracts that allow users to lock in an interest rate at some future date. The buyer of an FRA locks in a borrowing rate; the seller locks in a lending rate. In other words, the "long" receives a payment if the spot rate is above the forward rate.

Define the timing of the short leg as τ_1 and of the long leg as τ_2, both expressed in years. Assume linear compounding for simplicity. The forward rate can be defined as the implied rate that equalizes the return on a τ_2-period investment with a τ_1-period investment rolled over, that is,

$$(1 + R_2\tau_2) = (1 + R_1\tau_1)[1 + F_{1,2}(\tau_2 - \tau_1)] \tag{5.15}$$

Table 5.9 Computing the VaR of a USD 100 Million FRA (Monthly VaR at 95 Percent Level)

Position	PV of Flows, x	Risk (%), V	Correlation Matrix, R		Individual VaR, $\|x\| V$	Component VaR, $x\Delta$VaR
180 days	−USD 97.264	0.1629	1	0.8738	USD 0.158	−USD 0.116
360 days	USD 97.264	0.4696	0.8738	1	USD 0.457	USD 0.444
Undiversified VaR					USD 0.615	
Diversified VaR						USD 0.327

For instance, suppose that you sold a 6 × 12 FRA on USD 100 million. This is equivalent to borrowing USD 100 million for 6 months and investing the proceeds for 12 months. When the FRA expires in 6 months, assume that the prevailing 6-month spot rate is higher than the locked-in forward rate. The seller then pays the buyer the difference between the spot and forward rates applied to the principal. In effect, this payment offsets the higher return that the investor otherwise would receive, thus guaranteeing a return equal to the forward rate. Therefore, an FRA can be decomposed into two zero-coupon building blocks.

$$\text{Long } 6 \times 12 \text{ FRA} = \text{long 6-month bill} + \text{short 12-month bill}$$

Table 5.9 provides a worked-out example. If the 360-day spot rate is 5.8125 percent and the 180-day rate is 5.6250 percent, the forward rate must be such that

$$(1 + F_{1,2}/2) = \frac{(1 + 5.8125/100)}{(1 + 5.6250/200)}$$

or $F = 5.836$ percent. The present value of the notional USD 100 million in 6 months is $x = \text{USD } 100/(1 + 5.625/200) = \text{USD } 97.264$ million. This amount is invested for 12 months. In the meantime, what is the risk of this FRA?

Table 5.9 displays the computation of VaR for the FRA. The VaRsof 6- and 12-month zeroes are 0.1629 and 0.4696, respectively, with a correlation of 0.8738. Applied to the principal of USD 97.26 million, the individual VaRs are USD 0.158 million and USD 0.457 million, which gives an undiversified VaR of USD 0.615 million. Fortunately, the correlation substantially lowers the FRA risk. The largest amount the position can lose over a month at the 95 percent level is USD 0.327 million.

Interest-Rate Swaps

Interest-rate swaps are the most actively used derivatives. They create exchanges of interest-rate flows from fixed to floating or vice versa. Swaps can be decomposed into two legs, a fixed leg and a floating leg. The fixed leg can be priced as a coupon-paying bond; the floating leg is equivalent to a floating-rate note (FRN).

To illustrate, let us compute the VaR of a USD 100 million 5-year interest-rate swap. We enter a dollar swap that pays 6.195 percent annually for 5 years in exchange for floating-rate payments indexed to London Interbank Offer Rate (LIBOR). Initially, we consider a situation where the floating-rate note is about to be reset. Just before the reset period, we know that the coupon will be set at the prevailing market rate. Therefore, the note carries no market risk, and its value can be mapped on cash only. Right after the reset, however, the note becomes similar to a bill with maturity equal to the next reset period.

Interest-rate swaps can be viewed in two different ways: as (1) a combined position in a fixed-rate bond and in a floating-rate bond or (2) a portfolio of forward contracts. We first value the swap as a position in two bonds using risk data from Table 5.4. The analysis is detailed in Table 5.10.

The second and third columns lay out the payments on both legs. Assuming that this is an at-the-market swap, that is, that its coupon is equal to prevailing swap rates, the short position in the fixed-rate bond is worth USD 100 million. Just before reset, the long position in the FRN is also worth USD 100 million, so the market value of the swap is zero. To clarify the allocation of current values, the FRN is allocated to cash, with a zero maturity. This has no risk.

The next column lists the zero-coupon swap rates for maturities going from 1 to 5 years. The fifth column reports the present value of the net cash flows, fixed minus floating. The last column presents the component VaR, which adds up to a total diversified VaR of USD 2.152 million. The undiversified VaR is obtained from summing all individual VaRs. As usual, the USD 2.160 million value somewhat overestimates risk.

This swap can be viewed as the sum of five forward contracts, as shown in Table 5.11. The 1-year contract promises payment of USD 100 million plus the coupon of 6.195 percent; discounted at the spot rate of 5.813 percent, this yields a present value of −USD 100.36 million. This is in exchange for USD 100 million now, which has no risk.

The next contract is a 1 × 2 forward contract that promises to pay the principal plus the fixed coupon in 2 years, or

Table 5.10 Computing the VaR of a USD 100 Million Interest-Rate Swap (Monthly VaR at 95 Percent Level)

	Cash Flows					
Term (Year)	**Fixed**	**Float**	**Spot Rate**	**PV of Net Cash Flows**	**Individual VaR**	**Component VaR**
0	USD 0	+USD 100		+USD 100.000	USD 0	USD 0
1	−USD 6.195	USD 0	5.813%	−USD 5.855	USD 0.027	USD 0.024
2	−USD 6.195	USD 0	5.929%	−USD 5.521	USD 0.054	USD 0.053
3	−USD 6.195	USD 0	6.034%	−USD 5.196	USD 0.077	USD 0.075
4	−USD 6.195	USD 0	6.130%	−USD 4.883	USD 0.096	USD 0.096
5	−USD 106.195	USD 0	6.217%	−USD 78.546	USD 1.905	USD 1.905
Total				USD 0.000		
Undiversified VaR					USD 2.160	
Diversified VaR						USD 2.152

Table 5.11 An Interest-Rate Swap Viewed as Forward Contracts (Monthly VaR at 95 Percent Level)

	PV of Flows: Contract					
Term (Year)	**1**	**1 × 2**	**2 × 3**	**3 × 4**	**4 × 5**	**VaR**
1	−USD 100.36	USD 94.50				
2		−USD 94.64	USD 89.11			
3			−USD 89.08	USD 83.88		
4				−USD 83.70	USD 78.82	
5					−USD 78.55	
VaR	USD 0.471	USD 0.571	USD 0.488	USD 0.446	USD 0.425	
Undiversified VaR						USD 2.401
Diversified VaR						USD 2.152

−USD 106.195 million; discounted at the 2-year spot rate, this yields −USD 94.64 million. This is in exchange for USD 100 million in 1 year, which is also USD 94.50 million when discounted at the 1-year spot rate. And so on until the fifth contract, a 4 × 5 forward contract.

Table 5.11 shows the VaR of each contract. The undiversified VaR of USD 2.401 million is the result of a simple summation of the five VaRs. The fully diversified VaR is USD 2.152 million, exactly the same as in the preceding table. This demonstrates the equivalence of the two approaches.

Finally, we examine the change in risk after the first payment has just been set on the floating-rate leg. The FRN then becomes a 1-year bond initially valued at par but subject to fluctuations in rates. The only change in the pattern of cash flows in Table 5.10 is to add USD 100 million to the position on year 1 (from −USD 5.855 to USD 94.145). The resulting VaR then decreases from USD 2.152 million to USD 1.763 million. More generally, the swap's VaR will converge to zero as the swap matures, dipping each time a coupon is set.

5.4 MAPPING OPTIONS

We now consider the mapping process for nonlinear derivatives, or options. Obviously, this nonlinearity may create problems for risk measurement systems based on the delta-normal approach, which is fundamentally linear.

To simplify, consider the Black-Scholes (BS) model for European options.[2] The model assumes, in addition to

[2] For a systematic approach to pricing derivatives, see the excellent book by Hull (2005).

Table 5.12 Derivatives for a European Call

Parameters: S = USD 100, σ = 20%, r = 5%, r^* = 3%, τ = 3 months					
			Exercise Price		
	Variable	Unit	K = 90	K = 100	K = 110
c		*Dollars*	11.01	4.20	1.04
		Change per			
Δ	Spot price	Dollar	0.869	0.536	0.195
Γ	Spot price	Dollar	0.020	0.039	0.028
Λ	Volatility	(% pa)	0.102	0.197	0.138
ρ	Interest rate	(% pa)	0.190	0.123	0.046
ρ^*	Asset yield	(% pa)	−0.217	−0.133	−0.049
θ	Time	Day	−0.014	−0.024	−0.016

perfect capital markets, that the underlying spot price follows a continuous *geometric brownian motion* with constant volatility $\sigma(dS/S)$. Based on these assumptions, the Black-Scholes (1973) model, as expanded by Merton (1973), gives the value of a European call as

$$c = c(S,K,\tau,r,r^*,\sigma) = Se^{-r^*\tau}N(d_1) - Ke^{-r\tau}N(d_2) \qquad (5.16)$$

where $N(d)$ is the cumulative normal distribution function with arguments

$$d_1 = \frac{\ln(Se^{-r^*\tau}/Ke^{-r\tau})}{\sigma\sqrt{\tau}} + \frac{\sigma\sqrt{\tau}}{2}, \quad d_2 = d_1 - \sigma\sqrt{\tau}$$

where K is now the *exercise price* at which the option holder can, but is not obligated to, buy the asset.

Changes in the value of the option can be approximated by taking partial derivatives, that is,

$$\begin{aligned} dc &= \frac{\partial c}{\partial S}dS + \frac{1}{2}\frac{\partial^2 c}{\partial S^2}dS^2 + \frac{\partial c}{\partial r^*}dr^* + \frac{\partial c}{\partial r}dr + \frac{\partial c}{\partial \sigma}d\sigma + \frac{\partial c}{\partial t}dt \qquad (5.17)\\ &= \Delta dS + \tfrac{1}{2}\Gamma dS^2 + \rho^* dr^* + \rho dr + \Lambda d\sigma + \Theta dt \end{aligned}$$

The advantage of the BS model is that it leads to closed-form solutions for all these partial derivatives. Table 5.12 gives typical values for 3-month European call options with various exercise prices.

The first partial derivative, or *delta*, is particularly important. For a European call, this is

$$\Delta = e^{-r^*\tau}N(d_1) \qquad (5.18)$$

This is related to the cumulative normal density function. Figure 5.2 displays its behavior as a function of the underlying spot price and for various maturities.

The figure shows that delta is not a constant, which may make linear methods inappropriate for measuring the risk of options. Delta increases with the underlying spot price. The relationship becomes more nonlinear for short-term options, for example, with an option maturity of 10 days. Linear methods approximate delta by a constant value over the risk horizon. The quality of this approximation depends on parameter values.

For instance, if the risk horizon is 1 day, the worst down move in the spot price is $-\alpha S\sigma\sqrt{T} = -1.645 \times \text{USD } 100 \times 0.20\sqrt{1/252} = -\text{USD } 2.08$, leading to a worst price of USD97.92. With a 90-day option, delta changes from 0.536 to 0.452 only. With such a small change, the linear effect will dominate the nonlinear effect. Thus linear approximations may be acceptable for options with long maturities when the risk horizon is short.

It is instructive to consider only the linear effects of the spot rate and two interest rates, that is,

$$\begin{aligned} dc &= \Delta dS + \rho^* dr^* + \rho dr\\ &= [e^{-r^*\tau}N(d_1)]dS + [-Se^{-r^*\tau}\tau N(d_1)]dr^* + [Ke^{-r\tau}\tau N(d_2)]dr\\ &= [Se^{-r^*\tau}N(d_1)]\frac{dS}{S} + [Se^{-r^*\tau}N(d_1)]\frac{dP^*}{P^*} - [Ke^{-r\tau}N(d_2)]\frac{dP}{P}\\ &= x_1\frac{dS}{S} + x_2\frac{dP^*}{P^*} + x_3\frac{dP}{P} \qquad (5.19) \end{aligned}$$

This formula bears a striking resemblance to that for foreign currency forwards, as in Equation (5.13). The only difference is that the position on the spot foreign currency and on the foreign currency bill $x_1 = x_2$ now involves $N(d_1)$, and the position on the dollar bill x_3 involves $N(d_2)$. In the extreme case, where the option is deep in the money, both $N(d_1)$ and $N(d_2)$ are equal to unity, and the option behaves exactly like a position in a forward contract. In this case, the BS model reduces to $c = Se^{-r^*\tau} - Ke^{-r\tau}$, which is indeed the valuation formula for a forward contract, as in Equation (5.9).

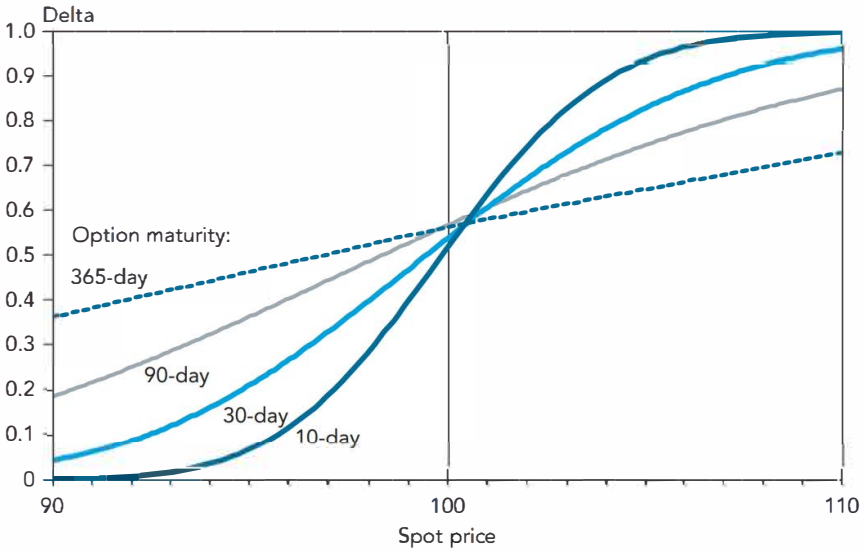


Figure 5.2 **Delta as a function of the risk factor.**

Also note that the position on the dollar bill $Ke^{-r\tau}N(d_2)$ is equivalent to $Se^{-r^*\tau}N(d_1) - c = S\Delta - c$. This shows that the call option is equivalent to a position of Δ in the underlying asset plus a short position of $(\Delta S - c)$ in a dollar bill, that is

$$\text{Long option} = \text{long}\Delta\text{asset} + \text{short}(\Delta S - c)\text{bill}$$

For instance, assume that the delta for an at-the-money call option on an asset worth USD 100 is $\Delta = 0.536$. The option itself is worth USD 4.20. This option is equivalent to a $\Delta S =$ USD 53.60 position in the underlying asset financed by a loan of $\Delta S - c =$ USD 53.60 − USD 4.20 = USD 49.40.

The next step in the risk measurement process is the aggregation of exposures across the portfolio. Thus all options on the same underlying risk factor are decomposed into their delta equivalents, which are summed across the portfolio. This generalizes to movements in the implied volatility, if necessary. The option portfolio would be characterized by its net vega, or Λ. This decomposition also can take into account second-order derivatives using the net gamma, or Γ. These exposures can be combined with simulations of the underlying risk factors to generate a risk distribution.

5.5 CONCLUSIONS

Risk measurement at financial institutions is a top-level aggregation problem involving too many positions to be modeled individually. As a result, instruments have to be mapped on a smaller set of primitive risk factors.

Choosing the appropriate set of risk factors, however, is part of the art of risk management. Too many risk factors would be unnecessary, slow, and wasteful. Too few risk factors, in contrast, could create blind spots in the risk measurement system.

The mapping process consists of replacing the current values of all instruments by their exposures on these risk factors. Next, exposures are aggregated across the portfolio to create a net exposure to each risk factor. The risk engine then combines these exposures with the distribution of risk factors to generate a distribution of portfolio values.

For some instruments, the allocation into general-market risk factors is exhaustive. In other words, there is no specific risk left. This is typically the case with derivatives, which are tightly priced in relation to their underlying risk factor. For others positions, such as individual stocks or corporate bonds, there remains some risk, called *specific risk*. In large, well-diversified portfolios, this remaining risk tends to wash away. Otherwise, specific risk needs to be taken into account.

6

Validating Bank Holding Companies' Value-at-Risk Models for Market Risk

DAVID LYNCH*

Learning Objectives

After completing this reading, you should be able to:

- Describe some important considerations for a bank in assessing the conceptual soundness of a VaR model during the validation process.
- Explain how to conduct sensitivity analysis for a VaR model, and describe the potential benefits and challenges of performing such an analysis.
- Describe the challenges a financial institution could face when calculating confidence intervals for VaR.
- Discuss the challenges in benchmarking VaR models and various approaches proposed to overcome them.

* Federal Reserve Board. The views expressed in this paper are the author's own and do not represent the views of the Board of Governors or the Federal Reserve System.

Excerpt is Chapter 2 of Validation of Risk Management Models for Financial Institutions, *edited by David Lynch, Iftekhar Hasan, Akhtar Siddique.*

6.1 INTRODUCTION

The Basel Committee on Banking Supervision established the use of Value-at-Risk (VaR) models for capitalizing banks' trading activities in 1996. At the same time as they introduced VaR models for regulatory capital, they required that banks backrest their VaR models using a stoplight test based on the Kupiec (1995) testing procedure. If a bank's loss exceeds their VaR forecast made the previous day, then there is an exception. Banks must apply a multiplier to their VaR, which increases with the number of exceptions they experienced over the past year, in order to determine their capital requirements. For this reason, backtesting has played an especially important role in validating VaR models for market risk. More recently, the Basel Committee on Banking Supervision has switched to the use of expected shortfall (BIS 2019) with expected implementation in 2025.

This emphasis on the role of backtesting VaR models in capital determinations has been an important element of model validation for market risk models, and there is no doubt that it will continue to play an important role. However, there are other elements of validating VaR models that financial institutions should pay attention to, but have not emphasized given the importance of backtesting in the regulatory framework. Academic analysis has improved the tools available for model validation of trading models, including backtesting, and these tools should be applied within financial institutions.

In addition to the Basel backtesting requirements, the US banking agencies have issued guidance on model validation that contains three key components.[1] This guidance applies to all models, including trading models. The Basel requirements on backtesting are simply more specific on this single aspect regarding the validation of capital models for trading. The guidance covers three types of testing that should be performed by banks. First, banks must assess the conceptual soundness of the model. This includes reviewing data, assumptions and techniques used in the model, and performing sensitivity analysis on those elements of the model where warranted. Second, the banks must perform an outcomes analysis, which could include backtesting. The third aspect is the benchmarking of models or a comparison of the model to other models. Banks have traditionally emphasized backtesting for their trading activities, but the other aspects of model validation are applicable to trading as well.

In this chapter, we provide a short overview of the VaR models commonly used at banks. We then review how these three aspects of validation can be applied to VaR models of banks' trading activities. In the case of backtesting and benchmarking we show how banks' VaR models fare under some of the backtesting and benchmarking tests.

[1] Note FRB OCC and EU requirements.

6.2 VaR MODELS

There are many treatments that describe the implementation of VaR models. Jorion (2006), Christoffersen (2012), Andersen et al. (2006) are examples of how to construct a VaR model. Nieto and Ruiz (2016) provide an overview of both VaR models and their testing. The summary description of VaR models here is to provide background for understanding the validation of those models at financial institutions. For these purposes VaR is defined as follows:

$$P(\Delta V_{P,t+N} \leq -VaR_{t+N}^{1-c} | \mathcal{F}_t) = 1 - c.$$

The probability that the change in the value of portfolio P, $\Delta V_{P,t+N}$, from t to $t + N$ is less than or equal to the negative of the VaR using the information $\mathcal{F}_t$ available at time t is equal to 1−c, where c is the coverage level of the VaR.[2] Under the Basel requirements from 1996, N is 10 days and c is 99% (thus the superscript to the VaR represents the probability that a loss exceeds the negative value of VaR, in this case 1%). The change in the value of the portfolio is the sum of the changes in the value of the portfolio components or positions:[3]

$$\Delta V_{P,t+N} = \sum_{i \in P} \Delta V_{i,t+N}.$$

In most types of VaR models, a pseudo history of changes in the value of the portfolio is necessary. Generally, that history is constructed of one-day changes in value of the portfolio so that $N = 1$ and the history is constructed based on the composition of the current portfolio V, at time T. In this case the pseudo history is described by

$$\{\Delta V_t^T\}_{t=1}^T \equiv \left\{ \sum_{i \in P} V_{iT} r_{it} \right\}_{t=1}^T,$$

[2] VaR models are usually defined by their coverage ratio, the probability that the financial institution will not experience a loss larger than the VAR rather than the probability of seeing an exception 1−c. Furthermore, VaR is usually reported as a positive number even though it represents a loss.

[3] Academic papers usually perform analysis on the returns or log returns of a portfolio. Practitioners usually perform analysis on the change in value of the portfolio – its profit or loss. We emphasize the analysis on the profit or loss in this chapter rather than returns so that it is more directly applicable to practitioners at financial firms and can be applied to data reported by banks. This does not come without difficulties, notably all tests require that the variable of interest is i.i.d or that the variable be transformed to i.i.d. When P&L is used directly, and banks change their trading portfolio frequently, the use of P&L without transformation violates this condition.

where r_{it} are the returns on each date in the history for the positions. The construction of this pseudo history is not a trivial task and more will be said about this later. In fact, supervisory reviews spend a lot of effort ensuring that this pseudo history is accurate.

The most straightforward way to turn this pseudo history into a VaR model is to use it to generate a distribution of returns and select the return representing the appropriate quantile for the VaR model. This is the historic simulation method:

$$HS\text{–}VaR_{T+1}^{1-c} = -Q_{1-c}(\{\Delta V_t^T\}_{t=1}^T).$$

One orders the valuation change of the pseudo history from lowest to highest and selects the c^*T ranked return, interpolating if necessary. Pritsker (1997) provides a critique of the use of historic simulation, noting that it is not very responsive to recent volatility in portfolio valuations, and Escanciano and Pei (2012) provide a critique of unconditional backtesting of historic simulation. Nonetheless, historic simulation remains the most widely used method of computing VaR at commercial banks.

To address the lack of responsiveness, GARCH models are often used to make the volatility of the portfolio valuation changes depend on recent changes in the pseudo history. In this case

$$\Delta V_{P,t+1} = \mu_t + \sigma_t z_t \quad z_t \sim g(0,1) \tag{6.1}$$

and

$$\sigma_t^2 = \omega + \sum_{j=1}^{p} \alpha_j \Delta V_{t-j}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2, \tag{6.2}$$

with the parameters estimated from the pseudo history. g is the general probability density function with mean 0 and variance of 1, which is not necessarily the normal distribution. The Garch VaR is then determined by

$$G_{p,q}\text{–}VaR_{T+1}^{1-c} = \sigma_{T+1} G_c^{-1} - \mu_t. \tag{6.3}$$

Oftentimes, μ_t is assumed to be zero. G_c^{-1} is the inverse cumulative density function of g evaluated at c, or, equivalently, the cth quantile of G. An important special case of the GARCH VaR is riskmetrics VaR (Riskmetrics 1997), where g is assumed to be a standard normal distribution and a restricted ARCH (1,1) is used to estimate the variance of the pseudo history of changes in the portfolio value.

$$\sigma_t^2 = \lambda \sigma_{t-1}^2 + (1-\lambda)\Delta V_{t-1}^{T^2},$$
$$RM\text{–}VaR_{T+1}^{c} = \sigma_{T+1} \Phi_c^{-1},$$

where Φ_c^{-1} represents the cth quantile of the standard normal distribution. A drawback of the GARCH approach is that one has to specify the distribution of z_t. To avoid making this distributional assumption one can get the cth quantile from the pseudo history of shocks z_t. One orders the shocks and selects the 1 – cth smallest. This is the filtered historic simulation (FHS) method of Barone-Adesi et al. (1999). The FHS Value-at-Risk is given by

$$F_{p,q}\text{–}VaR_{T+1}^{1-c} = -\sigma_{T+1} Q_{1-c}(\{z_t\}) - \mu_{T+1} \tag{6.4}$$

which is essentially taking the $(1-C)^*T$ lowest z_t and multiplying it by the current volatility from the GARCH model to calculate the VaR. The FHS method allows the model to capture the dependence in volatility without making a distributional assumption regarding the error terms.

These are the main methods to estimate a VaR using a univariate approach. In many cases a more disaggregated approach is desireable and rather than estimate a univariate model, a multivariate model is estimated, with the pseudo history decomposed into the pseudo history of the portfolios' individual positions. This requires estimating the volatility of each individual position as well as the correlation between the positions in the portfolio. This can become intractable quite quickly as the number of positons in the portfolio grows. See Andersen et al. (2006) or Christoffersen (2012) for an exposition of extending the univariate approaches described here to a multivariate model.

6.3 CONCEPTUAL SOUNDNESS

The banking guidance on model validation requires a review of conceptual soundness of the model. This requires the validator to determine whether the model assumptions, techniques and data used are appropriate. In most cases this is accomplished through a narrative provided by the validator. As one builds a large-scale VaR model, numerous modeling decisions are made. Model Validation is a review of those decisions and an assessment of those choices and how realistic they are. At its core, a very basic decision is whether the model is suitable for the purpose it is developed for. In the case of VaR models, their basic purpose is to help the firm manage the risk of its positions; they have other uses such as for capital calculations, but these are perhaps ancillary to the risk management purpose.

In an early evaluation of VaR models, Berkowitz and O'Brien (2002) compared banks' VaR models to a simple model based on a GARCH model of the actual profit and loss (P&L) of firms. They found that the GARCH model of actual P&L outperformed the VaR model used by the commercial banks. It would be tempting to conclude that the GARCH on actual P&L should be used in favor of the banks' VaR models. Lo (2001) and Jorion (2007) have an exchange regarding VaR models that illustrates the danger of using a P&L-based VaR for a dynamic portfolio. Lo describes the profitability and risk of a hypothetical hedge fund

and capital decimation partners, and shows that it appears to make good returns on low risk based on those profitability numbers. The dynamic strategy consists of selling out of the money puts. The risk measures do not recognize the dynamic nature of the portfolio. Jorion shows how the risk measures would change if the VaR incorporated positional information, essentially recomputing the pseudo history of the change in the portfolio values whenever the portfolio changes to reflect the change in positons. Doing this more accurately reflects the riskiness of selling deep out of the money puts and reflects when the portfolio becomes more risky due to a change in positions.

For a risk manager and for traders in general, it is important to reflect how a change in the firm's positions is reflected in VaR. Consider a risk manager who has told a trader to reduce the risk of their portfolio measured by VaR. The trader dutifully reduces their positons. If the risk manager bases the risk on the actual P&L history, the reduction in positions does not result in a reduction in VaR since the actual P&L history is unaffected by the change in positions. On the other hand, a VaR model that is based on the pseudo history of P&L described above would show a reduction of risk as the positions (V_T) would have changed. This highlights a crucial aspect of VaR modeling; its key purpose is to show how risk changes when positions change. VaR models that cannot be used to show how risk changes when positions change are not "fit for purpose" and thus fail a crucial conceptual soundness test in the model validation process.

Beyond the consideration of how the model will be used, the conceptual soundness of the model can be considered using more quantitative tests. These include sensitivity analysis that can show the effect of data limitations or choices and provide statistical confidence intervals around VaR estimates. These can answer fundamental questions regarding the performance of the VaR model and can also assess the severity of data problems and estimation errors. These two tests shed light on the degree to which the model can be used to help manage positions.

6.4 SENSITIVITY ANALYSIS

Since the model is designed to show risk changes when positions change, it would be important to check the sensitivity of the VaR model to changes in positions. Garman (1997), Hollerbach (2003), Tasche (2000) and Gourieroux et al. (2000) describe methods to decompose and perform sensitivity analysis for VaR models. In the context of the models described above, this provides a way to check the assumptions or simplifications made in developing the pseudo history of the portfolios' value changes. This is especially true for the omissions in the pseudo history.

One would like to know that the simplification or omission did not materially affect the VaR model's output. In fact, supervisors around the world have begun to have financial institutions track and quantify the risks that are not included in their VaR model and to track their use of data proxies. Generally the methods to quantify and track these omitted risks are ad hoc. However, sensitivity analysis provides a consistent framework for making the assessment.

Value-at-Risk is linear homogeneous in positions. It therefore satisfies the Euler equation

$$VaR(V_{PT}) = \sum_{i \in P} \frac{\partial\ VaR}{\partial\ V_{iT}} V_{iT}.$$

Each term on the right-hand side is known as the component VaR and the derivative, $\partial VaR/\partial V_{it}$, is known as the marginal VaR. The marginal VaR shows how VaR would change due to a small change in the size of the position and is related to the regression coefficient of a regression of the change in value of the whole portfolio on the change in the value of the position:

$$\Delta V_{it} = \alpha + \beta_i \Delta V_{Pt} + \varepsilon_t, \tag{6.5}$$

which leads to a formula for each component VaR based on the position

$$component_i VaR(V_{PT}) = VaR(V_{PT}) w_i \beta_i.$$

Thus, the sensitivity of VaR to the position is dependent on the share of the position in the portfolio, w_i, and the sensitivity of the position value to the overall portfolio value. If the financial institution has a concern regarding some choice it has made regarding the valuation of a particular position, this analysis can help show how sensitive the VaR is to the valuation choice.

This framework provides some mechanism to assess the importance of a position in the VaR. However, it can be difficult to apply in circumstances that frequently occur at trading institutions. First, Equation 6.5 requires that you have a full time series of the change in value of the positions. Often the financial institution wants to know the impact of the position because this valuation data is missing or scarce. Equation 6.5 may be difficult to estimate due to the data scarcity. A proxy may be used, but a proxy may not be as volatile as the original position or may be less correlated with the portfolio than the original position and thus understates the contribution to VaR. With scarce data, the regression approach described is difficult to implement. Tasche (2000) and Hallerbach (2003) provide a method for when the regression approach will not work that is especially easy to implement when using historic simulation.

Tasche and Hollerbach show that linear homogeneity implies

$$VaR_{PT} = \Delta V^*_{PT} = \sum_{i \in P} E(\Delta V_{iT} | \Delta V_{PT} = \Delta V^*_{PT}),$$

when the VaR is determined by a scenario where ΔV_{PT}^{*} is the change in portfolio value. This change in portfolio value can be decomposed into the sum of the expected changes in value of the positions in the portfolio, conditional on realizing the change in value of the whole portfolio equal to the VaR. The component VaR for each position can be estimated by the change in value each position would have experienced on the day that determines the historic simulation VaR scenario. In the case of a "missing" position this reduces the problem of estimating the component VaR to observing how much it would have lost on the day that determines the VaR. This single observation is less reliable than an estimate based on multiple observations, so the average of the ordered observations near the HS-VAR may be used instead.

This is the nature of estimating the omitted risks in VaR. If you have sufficient data to estimate it accurately, it is likely that the omitted risk would be included already. If you don't have much data, it isn't in the VaR and it is difficult to make a precise estimate of how sensitive the overall VaR is to the omission. The process is quite general and can be applied in many cases where there is an omission from the valuation of the portfolio. For example, the portfolio valuation may rely on a Taylor series expansion with only the first few terms used. The omission of the "next" higher order term may be estimated in an analogous fashion.

In practice, the sensitivity depends on the portfolio composition. The portfolio composition of a large trading organization may change rapidly. For this reason many supervisors ask for the risk not in VaR to be estimated based on both the component VaR and the stand alone VaR so that they get a sense of the effect of the omission that is independent of the portfolio.

This process of examining the pseudo history of changes in the portfolio value can be viewed as an application of the validation process described by Jarrow (2011) whereby the model is examined to see if it is sensitive to changes in assumptions. In this case the assumption that particular positions or risk factors being omitted is immaterial is being tested by this sensitivity analysis. The process also provides a more rigorous way of prioritizing changes to the model to include more risk factors or higher order terms.

6.5 CONFIDENCE INTERVALS FOR VaR

When estimating a Value-at-Risk figure for a portfolio it is natural to ask about the accuracy of the estimate. The statistical way to answer this question is to place confidence intervals around the estimate whereby the true value of the VaR estimate should be within the confidence interval a stated high percentage of the time. This provides an assessment of the estimation risk in the VaR. In many contexts an assessment of the estimation risk is standard and widely expected. However, despite VaR being an inherently statistical framework, most financial institutions do not calculate the confidence levels for their VaR estimates. This is strange and it seems that it would be an important aspect of model validation of VaR models and a test that should be carried out routinely. Jorion (1996) describes a method for estimating the confidence interval of a VaR model based on the asymptotic standard error of a quantile.

$$SE(VaR_{PT}^{1-c}) = \sqrt{\frac{c(1-c)}{Tf(VaR_{PT}^{1-c})^2}}$$

This is a seemingly straightforward calculation with VaR, T, and c known. The main issue is evaluating the probability density function at the VaR estimate. In most cases one would impose a distributional assumption on the change in portfolio values and proceed to calculate confidence intervals. However, it is a stylized fact that financial returns are non-normal so imposing a normal distribution would probably be a mistake. Furthermore, as Jorion notes, it may be desirable to base the confidence interval on more than the single point of the quantile estimate and instead estimate the confidence interval based on a calculation of the variance of the distribution of pseudo returns. The trick in applying a formula like the one proposed by Jorion is clearly in determining *f()* and one could proceed along those lines. Alternatively, one could apply a more non-parametric approach. Two approaches in the academic literature are the use of order statistics (Dowd 2006) or the use of bootstrap techniques (Christoffersen and Goncalves 2006).

The use of order statistics starts with the pseudo history of portfolio value changes, $\{\Delta V_t^T\}_{t=1}^T$ and reorders them from lowest to highest so that $\Delta V_{(1)}^T \leq \Delta V_{(2)}^T \leq \Delta V_{(3)}^T \ldots \leq \Delta V_{(T)}^T$. The probability that at least r of the observations in the sample do not exceed a specified ΔV is given by the distribution

$$G_r(\Delta V) = \sum_{j=r}^{T} \binom{T}{j} [F(\Delta V)]^j [1-F(\Delta V)]^{n-j}$$

where $F(\Delta V)$ is the cumulative density function of the observations of the change in portfolio value. Since VaR is one particular value of ΔV the formula applies to the negative of VaR. To create 90% confidence intervals around VaR one sets G_r(VaR) equal to 0.05 and 0.95 to create the lower and upper bound of the confidence interval, with r set to 0.05T and 0.95T respectively. One then numerically solves for F_l for the lower bound and F_u for the upper bound.

How one proceeds from that point depends on the type of VaR model. For a historical simulation VaR, one takes the confidence

interval as $\left(\Delta V^T_{(F_l\Delta\cdot T)}, \Delta V^T_{(F_u\Delta\cdot T)}\right)$ from the ordered pseudo history.

For a GARCH model one uses the estimated variance and distributional assumptions to specify $F(\Delta V)$. The ΔV that corresponds to the upper and lower bound is then taken from that function.

Christoffersen and Goncalves (2005) propose to use bootstrap procedures to calculate confidence intervals for VaR estimates and provide methods for historic simulation, GARCH(1,1), and filtered historic simulation methods. These procedures are fairly straightforward to apply for univariate estimates, although potentially computationally expensive.

To bootstrap 90% confidence intervals for a historic simulation VaR, one draws observations from the pseudo history $\{\Delta V^T_t\}^T_{t=1}$ with replacement to generate a bootstrapped pseudo history $\{\Delta V^{S_1}_t\}^T_{t=1}$. The VaR is then calculated from the bootstrapped sample,

$$HS\text{-}VaR^{1-c}_{S_1+1} = -Q_{1-c}(\{\Delta V^{S_1}_t\}^T_{t=1}).$$

Repeat this procedure generating a bootstrap sample S_i and a VaR for the bootstrap sample, B times to generate $HS\text{-}VaR^{1-c}_{S_i+1}$ for $i = 1$ to B. The distribution of VaRs is the sampling distribution and the confidence interval can be calculated directly from the ordered VaR values as the fifth percentile and ninety-fifth percentile of the bootstrapped VaRs. More generally, to calculate a confidence interval of CI, the confidence interval will be given by:

$$\left[Q_{CI/2}\left(\left\{HS\text{-}VaR^{1-c}_{(i)}\right\}^B_{i=1}\right), Q_{1-CI/2}\left(\left\{HS\text{-}VaR^{1-c}_{(i)}\right\}^B_{i=1}\right)\right] \quad (6.6)$$

This approach to calculating confidence intervals is nonparametric like the calculation of historic simulation VaR itself. However, it does assume independence of the changes in portfolio value over time and thus does not reflect possible dependence in returns over time.

The GARCH VaR accounts for the dependence of changes in portfolio values. The bootstrap approach must be designed to mimic the dependence properties of these changes and it is easiest to resample from an IID series to generate the bootstrap sample. In the case of a GARCH VaR model, z_t in Eq. 6.1 is IID and Eqs. 6.1 and 6.2 are used to generate samples of $\Delta V_{P,s_i}$ and σ_{s_i}.

It would be tempting to calculate the bootstrapped VaR for the sample from Eq. 6.3 directly. However, Eq. 6.3 is subject to estimation risk and we would also like to account for that in the bootstrapped confidence level. To do this, Eq. 6.2 is re-estimated for the bootstrapped sample using $\Delta V_{P,s_i}$. New estimates of $\Delta V^*_{P,s_i}$ and $\sigma^*_{s_i}$ are generated based on a re-estimated equation. The GARCH VaR for the samples are then re-calculated based on this.

$$G_{p,q}\text{-}VaR^{1-c}_{T+1} = \sigma^*_{T+1}G^{-1}_c - \mu \quad (6.7)$$

This process is repeated B times and the confidence interval CI as it was for historic simulation

$$\left[Q_{CI/2}\left(\left\{G_{p,q}\text{-}VaR^{1-c}_{(i)}\right\}^B_{i=1}\right), Q_{1-CI/2}\left(\left\{G_{p,q}\text{-}VaR^{1-c}_{(i)}\right\}^B_{i=1}\right)\right] \quad (6.8)$$

In the case of a Filtered Historic Simulation model, one does not make the parametric assumption in Eq. 6.7. Instead, one also keeps re-estimated residuals $z^*_{s_i}$ and recalculates the VaR for each sample according to Eq. 6.4. The confidence interval is then calculated similarly to Eqs. 6.6 and 6.9.

More recently, methods of providing confidence intervals for VaR models using the empirical likelihood have been proposed. Chan et al. (2007) provide a method for GARCH models. Gong Li and Peng (2009) provide a method for ARCH models.

As a basis of comparison, in Table 6.1 we calculate confidence intervals for one-day 99% VaR based on the returns of the S&P 500. We look at different methods to compute the VaR (historic simulation, GARCH, and FHS) and different methods to compute the confidence intervals over different time intervals to compare the methods. The time periods used are a recent one-year time period, a longer time period, a one-year stress period, and a three-year stress period.

The results indicate first that confidence intervals are not symmetric and, as expected, that more data used to estimate VaR leads to tighter confidence intervals. Second for Historic simulation VAR, while there is a difference in the confidence intervals calculated using order statistics and the bootstrap, there does not appear to be convincing evidence that one method or the other produces tighter confidence intervals on a consistent basis. Historic simulation VaR has particularly wide confidence intervals during stress periods. Most notable are the narrow confidence intervals for filtered historic simulation, suggesting that it provides more efficient estimates of VaR.

Few Bank Holding Companies routinely calculate confidence intervals for their VaR estimates. This is somewhat puzzling since the methods for computing them are well established and there is significant value in determining the accuracy of a VaR model. We do not have the pseudo history of banks' own P&L used to calculate their VaRs. Following Berkowitz and O'Brien, we can use the VaR calculated on the profit and loss of the banks' trading portfolio (the change in value of the portfolio of trading positions held at the end of the previous day) as a comparison rather than use the pseudo history of the P&L. As discussed earlier, this VaR is not suitable for risk management, but can be useful as a description of the accuracy of Bank Holding companies' VaR estimates based on P&L. Table 6.2 shows a comparison of

Table 6.1 Confidence intervals for one-day 99% VaR under different methods.

VaR method	CI method	N	Date range	VaR	95% CI		CI as a percentage of VaR	
HS	Bootstrap	250	5/2016–5/2017	1.725	1.347	1.98	21.9	14.8
HS	Bootstrap	9424	1/1980–5/2017	2.878	2.727	3.138	5.2	9.0
HS	Bootstrap	253	8/2008–8/2009	8.439	5.536	10.58	34.4	25.4
HS	Bootstrap	755	8/2006–8/2009	5.662	4.103	6.271	27.5	10.8
HS	Order	250	5/2016–5/2017	1.725	1.358	1.761	21.3	2.1
HS	Order	9424	1/1980–5/2017	2.878	2.729	3.137	5.2	9.0
HS	Order	253	8/2008–8/2009	8.439	5.008	10.24	40.7	21.3
HS	Order	755	8/2006–8/2009	5.662	4.001	6.171	29.3	9.0
Garch	Order	250	5/2016–5/2017	1.109	0.846	1.152	23.7	3.9
Garch	Order	9424	1/1980–5/2017	1.11	0.903	1.252	18.6	12.8
Garch	Order	253	8/2008–8/2009	2.91	2.443	3.3	16.0	13.4
Garch	Order	755	8/2006–8/2009	2.815	2.579	3.101	8.4	10.2
FHS	Order	250	5/2016–5/2017	1.205	1.188	1.222	1.4	1.4
FHS	Order	9424	1/1980–5/2017	1.324	1.314	1.334	0.8	0.8
FHS	Order	253	8/2008–8/2009	3.109	3.001	3.136	3.5	0.9
FHS	Order	755	8/2006–8/2009	3.239	3.158	3.284	2.5	1.4

Table 6.2 One-day 99% VaR confidence intervals on Hypothetical P&L as a percentage of the VaR estimate.

	Lower bound			Upper bound		
VaR method	Minimum (%)	Median (%)	Maximum (%)	Minimum (%)	Median (%)	Maximum (%)
HS	8.50	19.60	50.90	7.20	15.92	34.00
GARCH(1,1)	0.01	1.20	7.40	0.10	2.30	9.70

confidence interval estimates for two VaRs calculated on banks' history of actual P&L. The results indicate much narrower confidence intervals for the Garch VaR than for Historic simulation VaR. While not conclusive, these results suggest that filtered historic simulation would provide improved estimates of VaR for banks that do not use that technique.

6.6 BACKTESTING

The second aspect of model validation is to compare predicted outcomes to realized outcomes, otherwise known as backtesting. Banks have been backtesting their VaR models for their trading portfolios since 1996 when backtesting was included by the Basel Committee in regulatory capital requirements.

Berkowitz and O'Brien (2002) summarize the early results of backtesting VaR models by banks. They found the VaR modeling to be lacking, notably that banks' losses that exceeded VaR occurred infrequently and occurred in clusters, demonstrating dependence. Perignon and Smith (2006, 2010a, 2010b) find similar results and provide possible explanations for the conservative VaR models. Szerszen and O'Brien (2017) find that Banks VaR models were conservative both before and after the financial crisis, but were not conservative during the financial crisis.

This testing is predicated on banks reporting their backtesting results either to their supervisors or publicly. In most cases, banks have reported their actual profit and loss values. In making the comparison of *ex ante* VaR to *ex post* profit and loss, the profit and loss includes components of profit

and loss that perhaps should not be included in the backtest. Fresard, Perignon and Wilhelmsson (2011) describe the effect of including fee income, commissions and other components on backtesting results in a simulation model and include this as a possible explanation for the backrests showing the VaR models are conservative. More recently US bank holding companies operating under the Basel 2.5 trading requirements have reported at the subportfolio level comparsons of VaR to the profit or loss that banks would experience if they had held their portfolios at the end of the previous day fixed and held it for one day. In this regard the test has improved in that the profit and loss used for the comparison is closer to what banks VaR models are designed to capture. For example, fees and commissions are not expected to be part of the VaR model and should not be part of the profit and loss that a bank backrests against either if one is testing the accuracy of the model. Much of the research conducted on actual backtesting include these components in the banks' profit and loss.

Kupiec (1995) and Christoffersen (1998) provide the early work on backtesting procedures. At its simplest, backtesting starts with the observation that a VaR with 99% coverage should on average see a violation of VaR 1% of the time. This is the basis of the unconditional coverage test of Kupiec and the BCBS backtesting procedures. Christoffersen includes a test of independence and conditional coverage. The test for conditional coverage is a test that a (conditional) VaR with 99% coverage if properly specified should have a probability of a violation of VaR 1% of the time at any point in time, not just on average.

To create the test for VaR evaluation, one needs to count the number of times that a loss exceeds the VaR threshold. Specifically, one creates the indicator variable for exceedances that takes a value of 1 when VaR is exceeded and is zero otherwise.

$$I_{t+1} = \begin{cases} 1 \text{ if } \Delta V_{P,t+1} < -VaR_t^{1-c} \\ 0 \text{ otherwise} \end{cases}$$

The test for unconditional coverage is a test of whether the probability of violation is equal to 1 minus the coverage rate of the VaR.

$$P(I_{t+1} = 1) = 1 - c$$

It does not depend on the information at time *t*. In contrast, the test for conditional coverage is a test of whether the probability of a violation is one minus the coverage rate taking into account all information available at time *t*.

$$P(I_{t+1} = 1 | \mathcal{F}_t) = 1 - c$$

The null hypothesis for the unconditional backtest is that the indicator variable is independently and identically distributed over time as a Bernoulli variable. We denote the number of times that the indicator takes a value of zero and one as T_0 and T_1 (which sum to the backtesting sample T) and then the likelihood function for an observed fraction of violations π in the unconditional test is

$$L(\pi) = (1 - \pi)^{T_0}(\pi)^{T_1}.$$

The fraction of violations is estimated from the backtesting sample as T_1/T and this is compared to the expected number of violations, 1 − c from the VaR model in a likelihood ratio test.

$$LR_{uc} = -2\ln\left[\frac{L(1-c)}{L\left(\frac{T_1}{T}\right)}\right] \sim \chi_1^2$$

This provides a test of the whether the model has the correct coverage on average. However, if the violations cluster around a point in time, then the risk of extreme losses during a time when the exceedences are more likely would be problematic. If the exceedences were independent over time, then there would be no clustering and no concern over the effect of clustering.

Christoffersen (1998) provides a test of independence that serves as a bridge to a test of conditional coverage, which is a joint test of proper coverage and independence. Christoffersen considers a first order markov chain as the model of the dependence structure. In this case, the violations can be described by a simple transition probability matrix where π_{ij} is the probability of being in state *j* today conditional on being in state *i* the previous day.

$$\Pi_1 = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix} = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}$$

The test for independence is that $\pi_{01} = \pi_{11} = T_1/T$ where T_1 and T have the same meaning as in the unconditional backtest, so that past exceedences do not affect the current probability of an exceedance. Extending the notation from the unconditional backtest, let T_{ij} where *i, j* = 0, 1 represent the number of observations in the sample with a *j* followed by an *i*. The maximum likelihood estimates for π_{01} and π_{11} and the likelihood function are:

$$\hat{\pi}_{01} = \frac{T_{01}}{(T_{00}+T_{01})}$$
$$\hat{\pi}_{11} = \frac{T_{11}}{(T_{10}+T_{11})}$$
$$L_c(\Pi_1) = (1 - \pi_{01})^{T_{00}}\pi_{01}^{T_{01}}(1 - \pi_{11})^{T_{10}}\pi_{11}^{T_{11}}.$$

The likelihood ratio test for independence is then given by

$$LR_{ind} = -2ln\left[\frac{L\left(\frac{T_1}{T}\right)}{L_c(\hat{\pi}_1)}\right] \sim \chi_1^2,$$

where $L_c(\hat{\Pi}_1)$ uses the maximum likelihood estimates described above. The test does not consider whether the coverage level is

correct. The test for conditional coverage combines the unconditional coverage test and the independence test and conducts the joint test of whether the number of violations is correct and if the violations are independent. In this case we test whether $\pi_{01} = \pi_{11} = 1 - c$. The likelihood ratio test for conditional coverage is

$$LR_{cc} = -2ln\left[\frac{L(1-c)}{L_c(\hat{\Pi}_1)}\right] \sim \chi^2_2,$$

and we note that $LR_{cc} = LR_{uc} + LR_{ind}$.

Pajhede (2017) provides a discussion of generalizing the conditional coverage test. Notably the idea of allowing for greater than first order dependence in the transition probability matrix is explored. Testing for *Kth* order dependence would entail increasing the number of estimated parameters that would quickly become intractable without some restrictions. To overcome this issue, Pajhede expands the window over which the previous exceedances affect today's probability of an exceedance. Instead of just counting the previous day the exception counts over the previous *K* days. In this case π_{01} represents the probability of an exceedance given that there were no exceedances over the previous *K* days and π_{11} represents the probability of an exceedance conditional on there being at least one exceedance in the previous *K* days. This is one particular way of expanding the dependence structure and others could be envisioned. More generally if one is considering testing for higher order dependence a Ljung–Box test could be run. Although this only tests for independence not for coverage.

Engle and Manganelli (2004) provide a dynamic quantile (DQ) test for a linear probability model that is fairly simple to implement, allows for dependence in the exceedances and allows tests of whether other information might significantly improve the VaR model. The test statistic is

$$DQ = \frac{(Hit_t' x_t [x_t' x_t]^{-1} x_t' Hit_t)}{T(1-c)c} \sim \chi^2_q,$$

where the de-meaned indicator function $Hit_t = I_t - (1 - c)$ and X_t is a vector of variables available at time t − 1 including lagged values of the Hit function.

Clements and Taylor (2003), Patton (2006) and Berkowitz et al. (2011) convert the DQ test to logistic regression-based test to account for the fact that the indicator function is a limited dependent variable. We call this the logistic dynamic quantile test (LDQ). A model of the nth-order auto regression is given by:

$$I_t = \alpha + \sum_{k=1}^{n} \beta_{1k} I_{t-k} + \sum_{k=1}^{n} \beta_{2k} g(I_{t-k}, I_{t-k-1}, \ldots, \Delta V_{t-k}, \Delta V_{t-k-1}, \ldots,) + u_t.$$

Other variables known at time t − 1 may also be included. A likelihood ratio test can then be constructed to test the significance of the beta coefficients while the significance of each beta parameter can be tested via Wald test. A test of independence (that could include tests of greater than first order dependence) and that the VaR model has been properly conditioned on available information is a test of whether

$$P(I_t = 1) = \frac{e^{\alpha}}{1+e^{\alpha}}.$$

To test whether the model has both correct conditional coverage and is independent we test whether

$$P(I_t = 1) = \frac{e^{\alpha}}{1+e^{\alpha}} = 1 - c.$$

These regression-based approaches are especially easy to implement and Berkowitz et al. (2011) find that the LDQ test has significant power advantages over other tests that they compared it to.

Gaglianone et al. (2011) propose a test based on a quantile regression. It can be thought of as an extension of a Mincer Zarnowitz (1969) regression test to predicted quantiles where VaR is regressed on the change in portfolio value. To test a VaR model the regression setup is:

$$\Delta V_t^t = \alpha_0^{1-c} + \alpha_1^{1-c} VaR_{t-1}^{1-c} + \varepsilon_t.$$

In this quantile regression the hypothesis of a good VaR model is a test that α_0 is equal to zero and α_1 is equal to one. A value of α_0 different than zero indicates a biased VaR estimate and also that the VaR is consistently either too high or too low. A value of α_1 greater than one indicates that high values of VaR underpredict the quantile of the change in portfolio value. One could augment this equation with additional variables known at time t−1, all of which should have coefficients equal to zero.

In performing these tests, the choice of significance level is driven by the tradeoff of the possibility of making a type I error (rejecting a correct model) or making a type II error (failing to reject an incorrect model) Increasing the significance level increases a type I error but reduces the type II error. The case can be made that type II errors are expensive in risk management so accepting a lower confidence level than in academic work would be appropriate.

It is rare to have a large number of observations when performing backtesting. Having a large number of violations is even rarer. Christoffersen (2012) and Berkowitz et al. (2011) advocate the use of Dufour's (2006) Monte Carlo simulated P-values rather than using p-values from the asymptotic (generally chi-squared) distribution. This procedure ensures a correctly sized test in small samples and addresses the small sample issue as

well as it can be addressed. Kupiec (1995) and others describe the low power of backrests and how large samples are needed to reject incorrect models. Bringing in more information to assess the model as is done in the QR, LQR and VQR tests helps address this power issue.

6.7 RESULTS OF THE BACKTESTS

We run these tests on the trading portfolios of bank holding companies that are subject to the market risk rule in the United States. Prior to 2013, banks compared VaR to actual P&L, where actual profit and loss would include fees, commissions, and intraday trading revenue, among other items. This tended to increase P&L above the simple change in value of the portfolio that was held at the end of the previous trading day, although in a few instances it would reduce the profit and loss. Indeed, a problem that can be encountered at trading banks is where a trading desk is losing money on its portfolio, but the fees and commissions obscure the losses on the portfolio and make it appear that the desk inventory is profitable. Current rules for backtesting do not include these items in the reported profit and loss. Banks now report the change in value of the portfolio they held at the end of the previous trading day, which is what the bank's VaR model is trying to capture. In this sense this is the appropriate test of the VaR model itself, although it could be argued that the appropriate test from a supervisory standpoint would include the banks' ability to earn other sources of revenue as well.

The data used cover the period from the second half of 2013 to 2016. Banks in the sample were subject to the market risk rule over substantially the whole time period, but a few became subject to the market risk rule at a slightly later date. The data we present covers twenty Bank Holding companies who reported backtesting data for at least 499 trading days and up to 647 trading days. The banks provide their 1 day 99% VaR and their profit and loss for each trading day.

Table 6.4 shows the results of the unconditional coverage test, the conditional coverage test, the DQ test, the LDQ test, and the VQR test.

The aggregate results are summarized in Table 6.3. The exceedance rate over the time period shows that firms are on average conservative in their VaR estimates with an average exceedance rate of 0.4%, below the 1% exceedance rate that would be expected. Since 2013–2016 was a benign period for markets, overall this is consistent with the observation that VaR models are generally conservative during benign periods, although clearly some firms were not conservative with the highest exceedance rate being 2.1%, double the expected rate.

Table 6.4 presents the results of specific tests described above using a two-sided confidence interval so that banks may fail because their model is either too conservative or too aggressive. No additional explanatory variables are included. The unconditional coverage test, the conditional coverage test, the dynamic quantile test, and the logistic dynamic quantile test all performed similarly, with a small number of firms failing the test, and just one or two failing because the model was aggressive. The VaR quantile regression test failed nineteen of the Bank holding companies' VaR models. It appears to be a much more stringent test, perhaps also affected more by the non-stationarity of the portfolio.

The tests described summarize the tests that use the indicator variable for an exceedance since most testing in practice uses this indicator. Both because of their simplicity and because of their use in regulations. The more stringent quantile test is included for comparison. The quantile test shows rather poor performance. In some ways the use of the indicator function as a regulatory evaluation of the model could have changed the behavior of Bank Holding Companies, causing them to alter their models to pass the test, perhaps at the expense of performance on other tests.

Table 6.3 **Summary statistics on backtesting data for 99% VaR.**

Number of firms	20
Firms with zero exceedences	5
Average exceedence rate	0.4%
Maximum exceedance rate	2.1%

Table 6.4 **Results of backtesting.**

	UC	CC	DQ	LDQ	VQR
Number of firms that fail at 90% confidence	4	3	2	3	19
Number of non-conservative fails	1	1	1	2	8
Minimum P-value	0.0003	0.0015	0.0000	0.0011	
Maximum P-value	0.5972	0.8354	0.9999	0.9004	
Average P-value	0.1627	0.2605	0.6945	0.4472	

Other types of tests have been proposed, for example, the duration based tests (Christofferson and Pellitier, 2004) but these are largely unused for evaluating models in practice. More recently, Gordy and McNeil (2020) propose extensions using probability integral transforms that weights parts of the distribution to allow the user to choose what parts of the distribution to weight more heavily in the backrest.

6.8 BENCHMARKING

Perhaps the most neglected aspect of model validation in market risk is benchmarking. While many banks will run a new model in parallel with an old model for a short period of time to check if it is well behaved, there is rarely any formal comparison between a bank's model and an alternative model. In some sense this is understandable; it is time-consuming and resource-intensive to build a single VaR model at a bank. Building two models simply compounds the problem. However, during the time a bank aims to replace a VaR mode or at least upgrade part of a VaR mode there is a period of time during which the bank can perform this comparison at low cost. The most common practice is simply to plot the two VaR models over time. There rarely is anything useful to say about the models from these plots other than one model seems more conservative than the other or that they seem to behave about the same. Turning this comparison into more formal tests would be an improvement in model validation. The literature for comparing models on their predictive ability is well developed. Komunjer (2013) provides a comprehensive overview of all forms of evaluating quantile forecasts including VaR models.

Two aspects of benchmarking hinder the application of statistical tests comparing two models. The first is that trading portfolios change frequently so that errors are not independent and identically distributed. This seems to be particularly problematic for regression-based tests. Christoffersen et al. (2001) propose methods to overcome this but restrict this to distributions that are location scale models so that the quantile is a linear function of the volatility. The second aspect has already been mentioned, that banks rarely have two VaR models available to compare. Berkowitz and O'Brien (2002) overcome this issue by estimating a GARCH (1,1) VaR model on the profit and loss from trading by the bank. This data is available to all banks. As indicated in the section on conceptual soundness, there are reasons why a bank would not use a VaR model based on GARCH on the profit and loss for risk management, but it is a ready source of comparison for banks, and, as has been shown, it is a difficult benchmark to beat.

In comparisons of forecast accuracy, the starting point is Diebold and Mariano (1995). This paper did not explicitly consider the case of quantile forecasts but set the groundwork for comparisons of all types of forecasts. An important point made in the paper is that evaluation should depend on the loss function of the forecaster. This may or may not take on the typical mean squared error loss evaluation that is used for point forecasts. For example, Lopez (1996) introduces the regulatory loss function. Since capital is based on the VaR model, the regulator is more concerned about cases where the loss exceeds the VaR rather than cases where the bank's VaR exceeds the P&L. More concretely for any observation the loss, l_{t+1}, (since VaR is a positive number even though it represents a loss) can be described as:

$$l_{t+1} = \begin{cases} (\Delta V_{P,t+1} + VaR_t^{1-c})^2 \text{ if } \Delta V_{P,t+1} < -VaR_t^{1-c} \\ 0 \text{ otherwise} \end{cases}$$

This regulatory criterion clearly favors conservatism in the estimate of VaR since only exceedances are penalized. Alternatively, the quantile could be evaluated based on accuracy using the same "check" function as is used to estimate a quantile regression. In this case the loss is:

$$l'_{t+1} = ((1-c) - 1(\Delta V_{P,t+1} < -VaR_t^{1-c})) \cdot (\Delta V_{P,t+1} + VaR_t^{1-c})$$

Sarma et al. (2003) uses the sign test described in Diebold and Mariano to evaluate VaR models estimated on the S&P 500 and India's NSE-50 index. The sign test is a test of the median of the distribution of the loss differential between two competing models. The loss differential between two models, i and j, is $z_t = l_t^i - l_t^j$. The sign test is then given by:

$$\frac{\sum_{i=t+1}^{t+N} 1(z_t > 0) - 0.5N}{\sqrt{0.25\,N}} \sim N(0,1)$$

Rejecting the null hypothesis would indicate that model i is a significantly better model under the loss function considered. The sign test is easy to implement so there is little reason for banks not to make their comparisons more formal using this test.

The backtesting data described above allows a comparison of a bank's VaR model (positional VaR) to a VaR model based on running a GARCH (1,1) model on the bank's P&L (P&L VaR). Both are calculated as a one-day VaR at 99% coverage. The sign test is used to compare the models based on the regulatory loss function and based on the check loss function. When the regulatory loss function is used the positional VaR model significantly outperforms the P&L VaR in every case. This demonstrates the conservative nature of banks' positional VaR models, often attributed to the effect of regulatory oversight (Perignon Deng and Wang 2008).

When the positional VaR and P&L VaR are compared on accuracy using the check loss function, the P&L VaR outperforms the positional VaR at sixteen out of nineteen banks. Only one bank's positional VaR outperformed the P&L VaR; for the remaining two banks the difference was insignificant. It is clear that banks' VaR models are designed to be conservative; the conservative nature of the positional VaR models may hinder their ability to outperform the P&L VaR models on accuracy.

6.9 CONCLUSION

The tests described above provide a demonstration of how bank holding companies may make concrete the types of validation that regulators seek. US regulators expect tests of conceptual soundness, outcomes analysis and benchmarking. This chapter makes concrete how these tests can be done in the context of VaR models.

The fundamental review of the trading book BIS (2019) replaces the use of VaR models to determine regulatory capital with expected shortfall models. In many cases there is a direct translation of tests used for VaR models to those used for expected shortfall models. Both Dowd and Christoffersen and Goncalves describe how confidence intervals for expected shortfall can be provided. Hollerbach and Tasche provide examples of how to estimate the effect of missing risk factors in expected shortfall models. A relatively simple method for backtesting for expected shortfall models is described in Du and Escanciano (2016). The sign test for comparing models can also be extended to expected shortfall models.

References

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X. (2006). Volatility and correlation forecasting. In Elliott, G., Granger, C. W. J., and Timmerman A. (Eds.), *Handbook of Economic Forecasting*, Vol. 1, Amsterdam: North-Holland, 777–878.

Barone-Adesi, G., Giannopoulos, K. and Vosper, L. (1999). VaR without correlations for portfolios of derivative securities. *Journal of Future Markets*, 19(5), 583–602.

Berkowitz, J., Christoffersen, P. and Pelletier, D. (2011). Evaluating Value-at-Risk models with desk-level data. *Management Science*, 57(2), 2213–2227.

Berkowitz, J. and O'Brien, J. (2002). How accurate are Value-at-Risk models at commercial banks? *Journal of Finance*, 57, 1093–1111.

BIS (2019). Minimum Capital Requirements for Market Risk. Basel Committee on Banking Supervision.

Chan, N. H., Deng, S.-J., Peng, L. and Xia, Z. (2007). Interval estimation of Value-at-Risk based on GARCH models with heavy-tailed innovations. *Journal of Econometrics*, 137(2), 556–576.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39, 841–862.

(2012). *Elements of Financial Risk Management* (2nd ed.). Amsterdam: Academic Press.

Christoffersen, P. F. and Goncalves, S. (2005). Estimation risk in financial risk management. *Journal of Risk*, 7, 1–28.

Clements, M. P. and Taylor, N. (2003). Evaluating interval forecasts of high frequency financial data. *Journal of Applied Econometrics*, 18, 445–456.

Dowd, K. (2006). Using order statistics to estimate confidence intervals for probabilistic risk measures. *Journal of Derivatives*, 14(2), 77–81.

Du, Z. and Escanciano, J. C. (2016). Backtesting expected shortfall: Accounting for tail risk. *Management Science*, 63(4), 940–958.

Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics*, 133(2), 443–477.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, 22(4), 367–381.

Escanciano, J. C. and Pei, P. (2012). Pitfalls in backtesting historical simulation VaR models. *Journal of Banking and Finance*, 36, 2233–2244.

Frésard, L., Pérignon, C. and Wilhelmsson, A. (2011). The pernicious effects of contaminated data in risk management. *Journal of Banking and Finance*, 35(10), 2569–2583.

Gaglianone, W. P., Lima, L. R., Linton, O. and Smith, D. R. (2011). Evaluating Value-at-Risk models via quantile regression. *Journal of Business and Economic Statistics*, 29(1), 150–160.

Garman, M. (1997). Taking VaR to pieces. *Risk*, 10(10), 70–71.

Gong, Y., Li, Z. and Peng, L. (2010). Empirical likelihood intervals for conditional Value-at-Risk in ARCH-GARCH models. *Journal of Time Series Analysis*, 31(2), 65–75.

Gordy, M. B. and McNeil, A. J. (2020). Spectral backrests of forecast distributions with application to risk management. *Journal of Banking and Finance*, 116, 1–13.

Gourieroux, C., Laurent, J. P. and Scaillet, O. (2000). Sensitivity analysis of Values at Risk. *Journal of Empirical Finance,* 7(3), 225–245.

Hallerbach, W. (2003). Decomposing portfolio Value-at-Risk: A general analysis. *Journal of Risk,* 2(5), 1–18.

Jarrow, R. A. (2011). Risk management models: Construction, testing usage. *Journal of Derivatives,* 18(4), 89–98.

Jorion, P. (1996). Risk2: Measuring the risk in value at risk. *Financial Analysts Journal,* 52, 47–56.

(2006). *Value-at-Risk: The new benchmark for managing financial risk* (3rd ed.). New York: McGraw-Hill.

(2007). Risk Management for Hedge Funds with Position Information. *Journal of Portfolio Management,* 34(1), 127–134.

(2009). Risk management lessons from the credit crisis. *European Financial Management,* 15(5), 923–933.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives,* 2, 173–184.

Lo, A.W. (2001). Risk management for hedge funds: Introduction and overview. *Financial Analysts Journal,* 57(6), 16–33.

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts and expectations. In J. Mincer (Ed.), *Economic Forecasts and Expectations.* New York: National Bureau of Economic Research.

Nieto, M. R. and Ruiz, E. (2016). Frontiers in VaR forecasting and backtesting. *International Journal of Forecasting* 32, 475–501.

O'Brien, James and Szerszen, Pawel J. (2017). An evaluation of bank measures for market risk before, during and after the financial crisis. *Journal of Banking & Finance, Elsevier,* vol. 80(C), 215–234.

Pajhede T. (2017). Backtesting Value at Risk: A generalized Markov test. *Journal of Forecasting,* 36(5), 597–613.

Patton, A. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review,* 47, 527–556.

Pérignon, C., Deng, Z. Y. and Wang, Z. Y. (2008). Do banks overstate their Value-at-Risk? *Journal of Banking and Finance,* 32, 783–794.

Pérignon, C. and Smith, D.R. (2008). A new approach to comparing VaR estimation method. *The Journal of Derivatives,* 16(2), 54–66.

(2010a). Diversification and Value-at-Risk. *Journal of Banking and Finance,* 34, 55–66.

(2010b). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking and Finance,* 34, 362–377.

Pritsker, M. (1997). The hidden dangers of historical simulation. *Journal of Banking and Finance,* 30(2), 561–582.

Riskmetrics (1997). *Riskmetrics: Technical Document* (4th ed). J.P. Morgan/Reuters.

Tasche, D. (2000). *Risk Contributions and Performance Measurement. Working paper,* Munich University of Technology.

7

Beyond Exceedance-Based Backtesting of Value-at-Risk Models: Methods for Backtesting the Entire Forecasting Distribution Using Probability Integral Transform

DIANA IERCOSAN, ALYSA SHCHERBAKOVA, DAVID MCARTHUR AND REBECCA ALPER

Learning Objectives

After completing this reading, you should be able to:

- Identify the properties of an exceedance-based backtest that indicate a VaR model is accurate, and describe how these properties are reflected in a PIT-based backtest.
- Explain how to derive probability integral transforms (PITs) in the context of validating a VaR model.
- Describe how the shape of the distribution of PITs can be used as an indicator of the quality of a VaR model.
- Describe backtesting using PITs, and compare the various goodness-of-fit tests that can be used to evaluate the distribution of PITs: he Kolmogorov-Smirnov test, the Anderson-Darling test, and the Cramér-von Mises test.

"Once more unto the breach, dear friends, once more;"

William Shakespeare (Henry V)

Excerpt is Chapter 4 of Validation of Risk Management Models for Financial Institutions, *edited by David Lynch, Iftekhar Hasan, Akhtar Siddique.*

7.1 INTRODUCTION

Banks are required to develop sophisticated and reliable models of market risk inherent in their trading portfolios. These models have multiple uses, including determination of regulatory capital requirements, monitoring and limiting of trader risk taking, as well as banks' internal management of risk. Many of these uses, and in particular determination of regulatory capital requirements, focus on the calculation of the portfolio's Value-at-Risk (VaR), which quantifies the portfolio's downside risk. More precisely, VaR is a mathematical concept that measures the loss that a given portfolio is expected not to exceed over a specific time interval and at a predetermined confidence level.

Regulatory capital requirements depend on the one-day VaR with a 99% confidence level. Simply put, a one percent VaR of $1,000,000 over a one-day horizon implies that a one-day realized loss on a portfolio is expected to exceed $1,000,000 one percent of the time and remain below the $1,000,000 threshold the remaining 99% of the time. In the context of regulatory oversight, assuming a one-day horizon with a 99% confidence level, if the risk model is accurate, then, each day, there will be an exactly 1% probability that the next day's profit and loss (P&L) will be a realized loss greater than the VaR measure estimated by the model.

There is a vast literature on the statistical properties of VaR models. Jorion (2002) provided a comprehensive survey and analysis of various VaR model specifications in the first edition of his book which described VaR as the "the new benchmark for controlling market risk." Other analysis showing forecast performance used hypothetical portfolios Marshall and Siegel (1997) and Pritsker (1997). Moreover, Berkowitz and O'Brien (2002) showed empirical results on the performance of banks' actual trading risk models by examining the statistical accuracy of their regulatory VaR forecasts for a sample of large trading banks. Berkowitz et al. (2016) showed acuracy of forecasts of VaR models for a small sample of desks. The contribution of this chapter is twofold: it describes how banks' VaR models fare in backtesting at 99% and introduces tests of the whole distrubution; and it shows how banks fare on those tests at a firm-wide aggregated level, but also at a disaggregated portfolio level.

Backtesting has proven to be a key tool in the validation of risk models by comparing realized outcomes to the model's forecast for those outcomes. Strictly speaking, backtesting is a statistical procedure where the precision of a portfolio's VaR estimates are systematically compared to corresponding realized P&L outcomes. As such, a disciplined backtesting regime ensures that models remain properly constructed for internal risk management purposes and calculation of regulatory capital.

Since the mid 1990s various methods for backtesting have been proposed. A standard backtesting practice is to count instances when daily P&L is lower than the *ex ante* VaR (i.e., portfolio loss exceeds its estimated VaR). These instances are also known as VaR exceedances, exceptions or breaches.

In addition to exceedance counts, recent quantitative literature has placed significant emphasis on the calculation of the probability integral transform (PIT), often called the "p-value", associated with the VaR model P&L as an integral component of a robust backtesting process. The PIT was introduced by Diebold et al. (1998) for backtesting of density forecasts and represents the cumulative probability of observing a loss greater than the current P&L realization based on the previous day's VaR model forecast. When PITs are well estimated they provide information on the accuracy of the risk model at any percentile of the forecast distribution.

According to Christoffersen (1998), any backtesting of a VaR model that accurately reflects the actual distribution of the P&L is expected to have two distinct properties:

1. unconditional coverage;
2. and independence.

The unconditional coverage property restricts the number of exceedances which may be observed in a given time period at a determined statistical significance level (again, in the regulatory context, this is a one-day horizon with a 99% confidence level). This property was investigated by Kupiec (1995) who defined a statistical test. Unconditional coverage is analogous to the uniformity distributional property of a series of PITs. If the risk is adequately modeled, the exceedances at 1% VaR should be observed 1% of the time, exceedances at 3% VaR should be observed 3% of the time, etc. In other words, the series of probabilities of observing each P&L outcome in relation to VaR should be uniformly distributed over a zero to one interval, $U(0,1)$.

The independence property asserts that the observed exceedances should be independent from one another, and each observed exceedance should not be informative of future exceedances. Operationally, if a risk model is perfectly accurate, a series of PITs are *i.i.d.* $U(0,1)$, Rosenblatt (1952).

Deviations from these properties indicate that the model is likely to be misspecified, with that misspecification taking on a conservative or an aggressive manner. Conservative misspecification implies that the model distribution is too wide, and P&L observed is small, clustering in the middle of the distribution (i.e., the model parameters are too conservative to accurately model market dynamics). Aggressive misspecification implies that the model distribution is too narrow and we often see realized P&L in the tails of the distribution.

A number of statistical tests have been developed to analyze the performance of VaR models. These tests are based on evaluating the degree to which the model exhibits the two properties described above either individually or jointly. This chapter provides a comprehensive overview of the range of tests available to assess VaR model fit and performance.

First, we investigate results from a set of tests used to assess unconditional coverage, conditional coverage, and independence properties of the realized VaR exceptions. Second, we present a comprehensive overview of tests used to assess the uniformity and independence properties of a series of PIT estimates generated from real-world risk models. The analysis includes tests based on the empirical CDF (e.g. Kolmogorov–Smirnov; Cramér–Von Mises; and Anderson–Darling) as well tests of dependence based on regression analysis of the observed PITs. In this chapter we assess the accuracy and possible misspecification of VaR models, and offer a comparison of backtesting results using PITs over exceedances for the same sample of real portfolios.

7.2 DATA

Under Basel III Subpart F (Market Risk) Section 205, paragraphs (c)(1)–(c)(3), US financial institutions are required to submit backtesting information for each subportfolio, for each business day, on an ongoing basis beginning January 2013. We apply the tests catalogued in this chapter to a sample of these backtesting results. Our data starts on January 1, 2013 and ends on December 31, 2015. The data used in the analysis was collected as part of the ongoing Basel III regulatory reporting for market risk. The VaR model performance results are aggregated in order to preserve anonymity of individual firms, but are also disaggregated by the type and geographical region of the financial products being modeled. The sample consists of 20 US financial firms with significant asset concentrations in the trading book, a total of 597 distinct, non-overlapping subportfolios, with a mean of 591 days and median of 707 days of data for each subportfolio. Each firm defined its subportfolios based on the firm's internal risk management practice and aligned subportfolio definitions with existing business practices. For each day and each subportfolio, firms reported to their supervisors a one-day regulatory VaR, calibrated to a 99% confidence level, one-day clean P&L (i.e. the net change in the price of the positions held in the subportfolio at the end of the previous business day), and the corresponding PIT. Clean P&L is a hypothetical P&L excluding new trades and fees that are not accounted for in the VaR model.

The analysis is based on the premise that if the risk model produces an accurate daily forecast distribution for the portfolio P&L, $F_t()$, then after observing the daily realization of the portfolio P&L, PL_{t+1}, we can calculate the risk model's probability of observing a loss below the actual P&L, known as Probability Integral Transform denoted by p_{t+1},

$$p_{t+1} = F_t(PL_{t+1}). \tag{7.1}$$

From the daily VaR and P&L series we can infer exceedances and assess model performance using these observations. Mathematically, the sequence of VaR_t exceedances is defined as,

$$I_{t+1} = \begin{cases} 1, & \text{if } R_{t+1} < VaR_t(p) \\ 0, & \text{else} \end{cases} \tag{7.2}$$

where 1 indicates a VaR breach or hit on day $t + 1$.

The analysis of the series of breaches or PIT is based on data aggregated across subportfolios grouped by, (i) single product and (ii) multiple product, according to subportfolio characteristics. "Single" product refers to subportfolios identified as consisting predominantly of a single product type. "Multiple" product refers to all portfolios in which a given product was present, regardless of materiality of its representation among other products in that portfolio. Table 7.1 illustrates the composition of the sample and summarizes the count of single (i.e., unique) and multiple product subportfolios.

Table 7.1 Subportfolio Count by Product Composition.

Product	Count single	Count multiple
Sovereign Bonds	6	266
Corporate Bonds	8	177
Muni Bonds	3	44
Agency MBS	1	47
Non Agency MBS	–	44
CMO	–	40
Interest Rates	30	382
FX	43	305
Commodities	23	64
Other Product Type	15	190

7.3 GRAPHICS OF THE EXCEEDANCE COUNT AND DISTRIBUTION OF PITs

Analysis of the VaR model performance is done on the subportfolio level using the backtesting results from the most granular and distinct subportfolios. Analysis is also performed on the

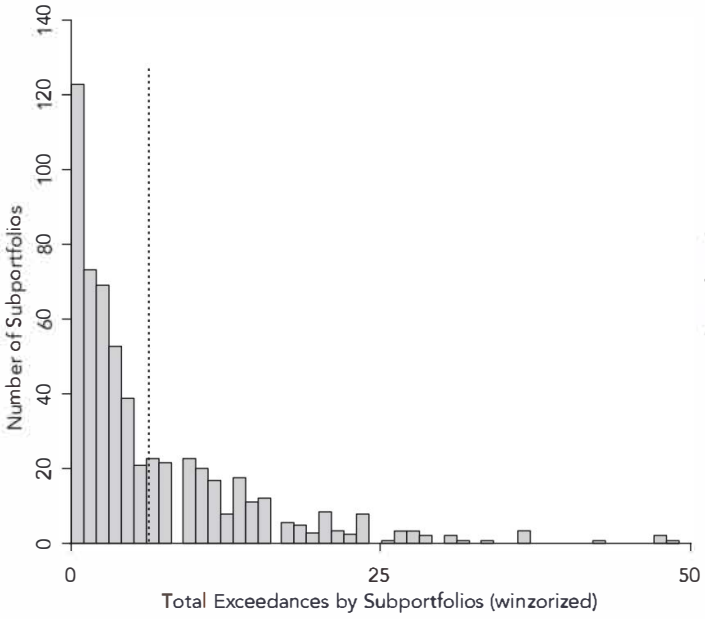


Figure 7.1a **Total exceedances by subportfolio (winzorized).**

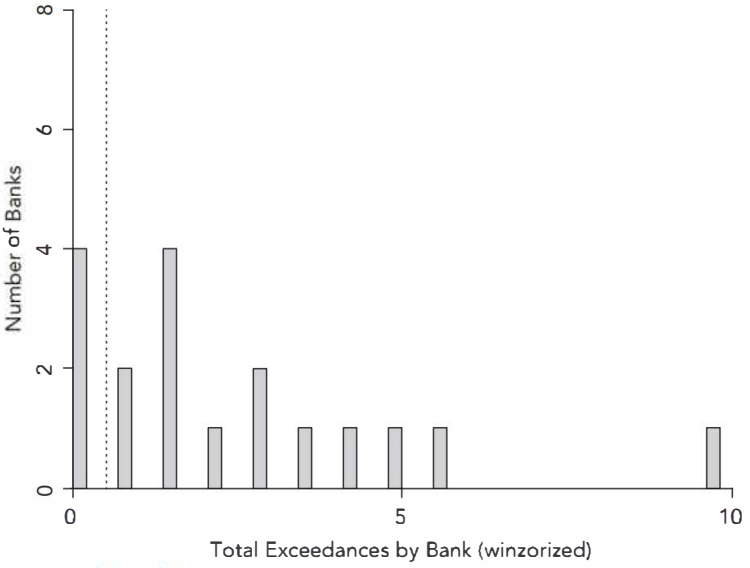


Figure 7.1b **Top of the house exceedances by bank (winzorized).**

combined top-of-the house backtesting data reported by firms. The analysis is based on the distribution of exceedance counts, and on the distributional properties of PITs associated with subportfolio backtests. As we discussed in the introduction, realized exceedances provide information on the performance of the VaR model at a single percentile, while reported PITs provide information on the quality and robustness of the model used for the daily forecasting of the full distribution of portfolio P&L.

As we are assessing regulatory VaR models, a priori, we expect to observe 1% exceedances, and we expect the series of PITs to be uniformly distributed. In what follows, we rely on a qualitative assessment of how closely a firm's methodology supports these expectations. We expect that the degree to which actual distribution of PITs differs from uniform may provide information regarding the model's ability to accurately and reliably assess risk of a given portfolio of instruments.

The charts presented in this section were produced using distinct subportfolios submitted by a number of financial

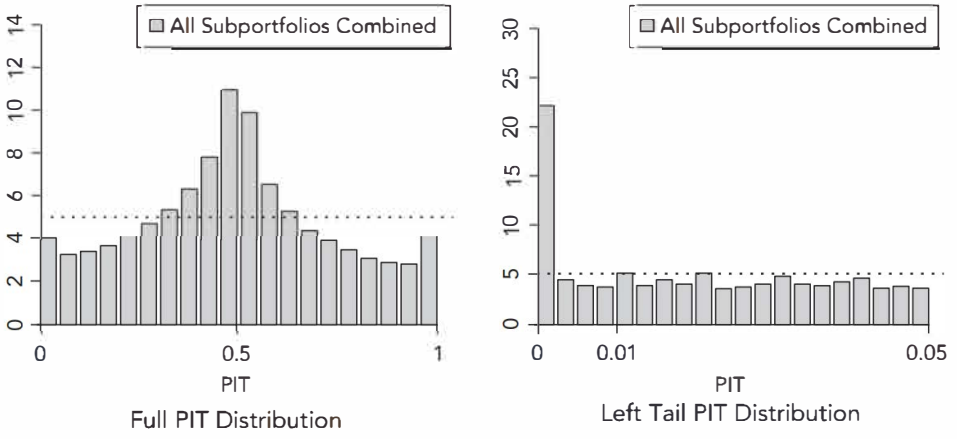


Figure 7.2a **PIT distribution.**

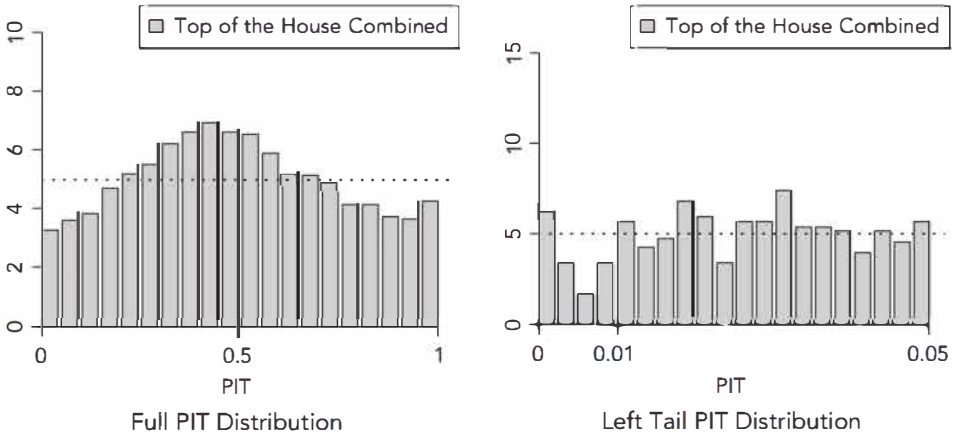


Figure 7.2b **Top of the house PIT distribution.**

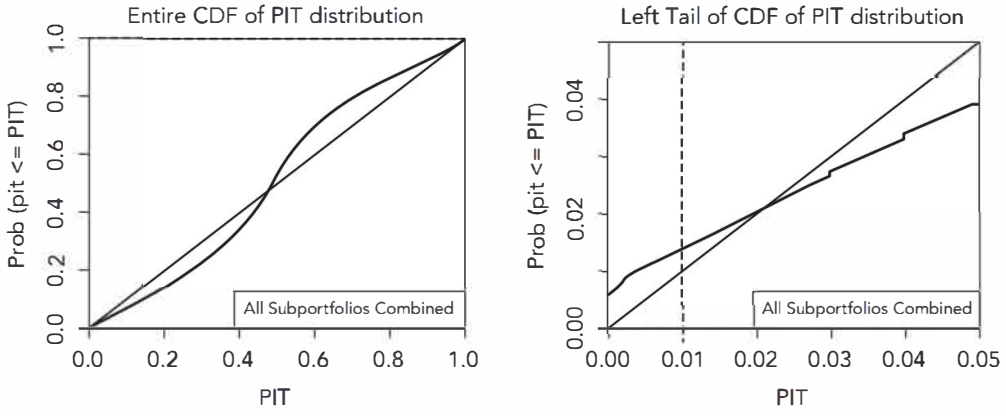


Figure 7.3a CDF of PIT distribution.

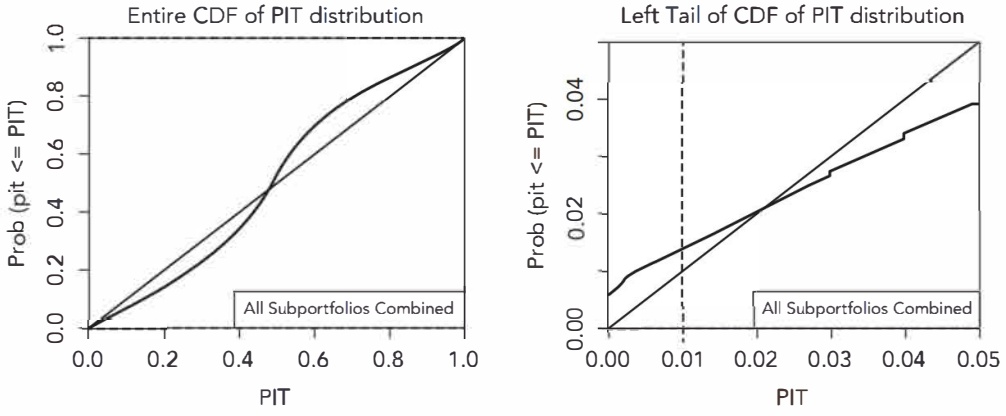


Figure 7.3b Top of the house CDF of PIT distribution.

institutions in the sample. The first histogram (Figure 7.1) illustrates risk model performance in terms of the number of exceedances. The vertical line indicates the expected number of exceedances based on one percentile confidence level (i.e., about 6 exceedances given our time series of approximately 625 daily observations for each subportfolio). Subportfolios to the left of the line produced reasonable backtesting results (i.e., demonstrating fewer than 6 exceedances), while subportfolios to the right of the line have an excess number of exceedances, thus failing VaR backtesting. Note, ten subportfolios exhibited more than fifty exceptions each in the 625 day time interval and were excluded from the histogram for illustration purposes.

The next histograms in Figure 7.2a and 7.2b illustrate the deviation from uniformity observed in the distribution PITs aggregated across all subportfolios and across all entire firm portfolios, respectively. That is, for the purpose of this illustration, all PITs were combined into a single series and graphed. The distorted shape of the distributions indicates that when considered in aggregate, models are fairly conservative in the body of the forecast distribution (i.e., humped shape). Figure 7.2a depicts the distribution of subportfolio PITs and shows that the tails of these risk models are thinner than is required by the realizations of P&L (i.e., visible spikes at each end of the distribution). The conditional histogram on the right focuses on the 5% of the left tail, demonstrating that the estimation at 1% is conservative across all subportfolios, but the tail is understated for extreme realizations. Nevertheless Figure 7.2b shows less deviation from uniformity in the loss tail for firm-wide risk models, indicating more adequate modeling of losses at the top of the house.

The quantile–quantile (Q–Q) plots of the series of PITs are provided above in Figure 7.3a and 7.3b. The first Q–Q plot of PITs highlights that risk models are built to be conservative generally. However, the second Q–Q plot in Figure 7.3a highlights the fact that risk models underestimate losses in the far left tail at subportfolio level, which is not confirmed in Figure 7.3b that shows results for aggregate firm level.

7.4 QUANTIFYING DEVIATIONS FROM UNIFORMITY OF DISTRIBUTION OF PITs

Numerical methods quantify deviations from uniformity in terms of mean, median, standard deviation, skewness and kurtosis of the distribution of PITs. The histograms in Figure 7.4a illustrate the statistical properties of subportfolios in our sample. Similarly, moments of the distribution of PITs for each firm are plotted in Figure 7.4b.

The vertical line in each histogram indicates an expected mean, median, standard deviation, skewness and kurtosis of a Uniform [0, 1] distribution. The histograms in Figure 7.4a clearly illustrate that VaR models generally produce distributions that overstate variance and understate kurtosis. Further, we observe deviations from the expected mean, as well as evidence of skewness in either positive or negative direction. Note, nine subportfolios exhibited kurtosis in excess of fifteen and, for illustration purposes, were excluded from histograms provided in this section. Figure 7.4b shows less

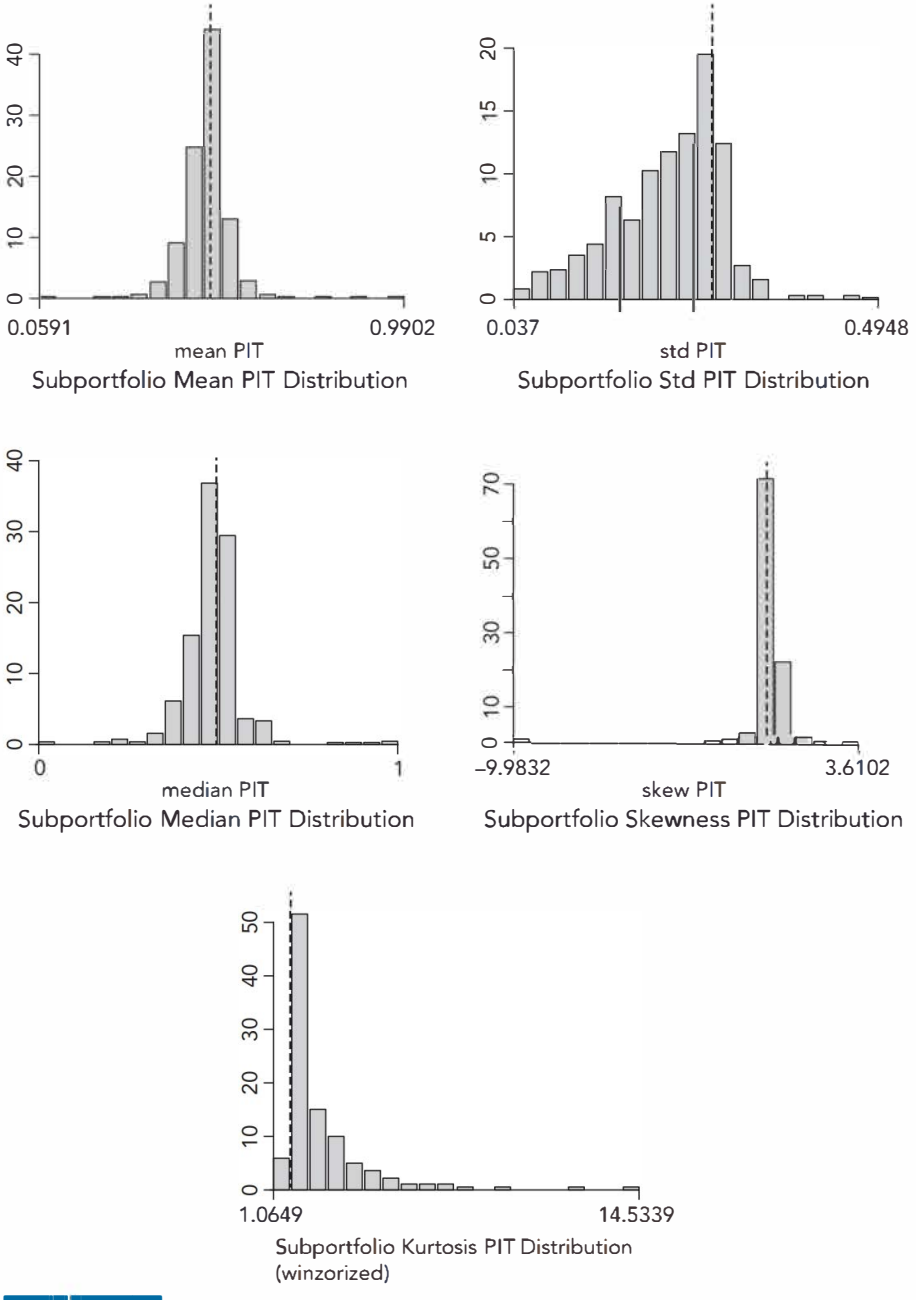


Figure 7.4a Subportfolio moments of the distribution of PITs.

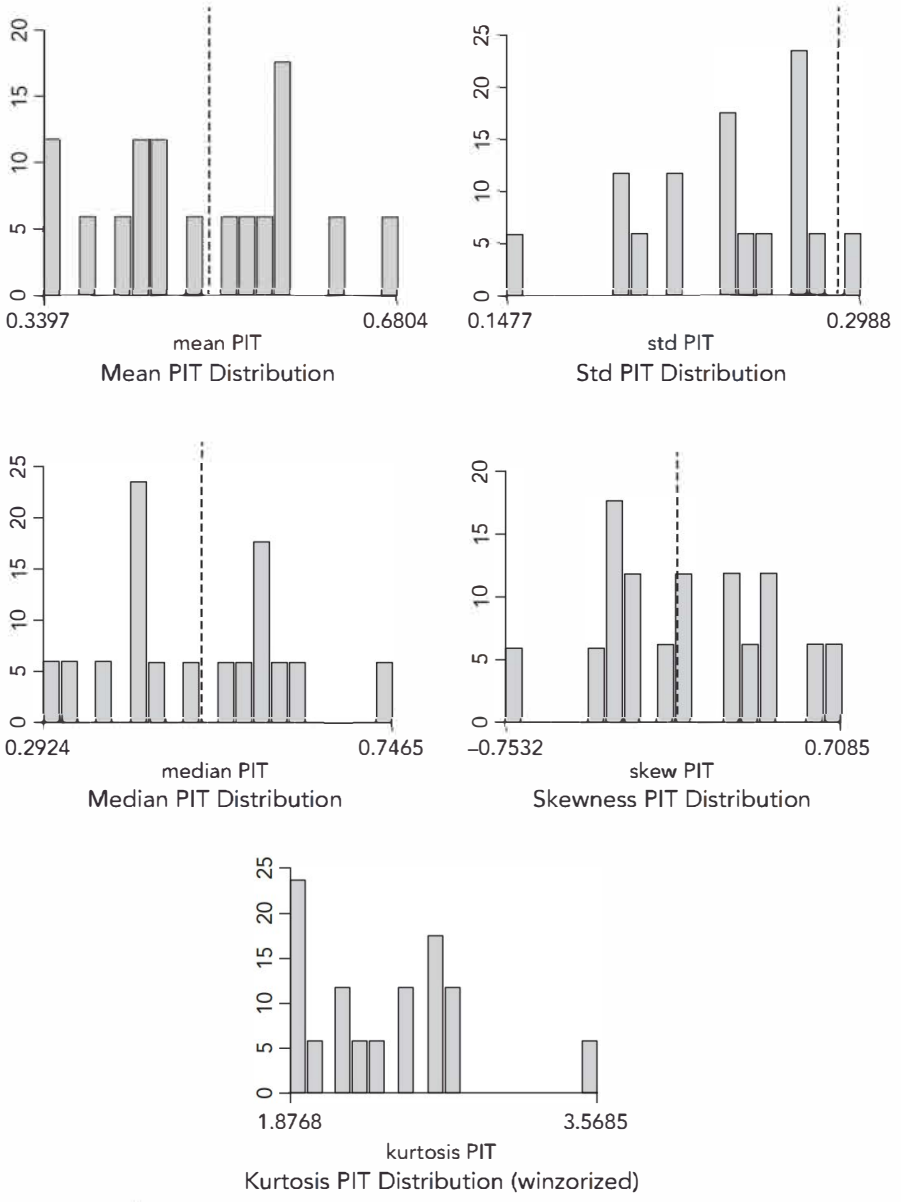


Figure 7.4b **Top of the house moments of the distribution of PITs.**

dispersion in moments of the firm-wide PITs than for the sub-portfolio level, and with considerably smaller kurtosis.

7.5 MISSPECIFICATION TESTS BASED ON EXCEPTIONS

Standard backtesting practices are rooted in tests of unconditional coverage and independence of VaR exceedances. These tests rely on the assumption that exceedances should be independent and occur precisely the number of times corresponding to the confidence level of the calculated VaR (i.e., for a 99% VaR, exceedances are expected to occur 1% of the time).

The *unconditional coverage test* introduced by Kupiec (1995) examines whether the theoretical confidence level matches the empirical probability of an exceedance. In other words, given a 1% VaR backtest, we expect to observe a VaR violation 1% of the time. If, instead, exceedances are observed more frequently, our conclusion is that the VaR systematically underestimates risk at the 1% level.

In the Kupiec proportion of failures test, the null hypothesis assumes that the series of observed exceptions I_t (also referred

to as the "hit" series) follows a Bernoulli distribution with a success probability parameter p such that,

$$I_t \sim \text{i.i.d. } Bernoulli(p)$$

This null is tested against an alternative that the observed success probability of the Bernoulli distribution is different from the assumed probability p,

$$I_t \sim Bernoulli(\pi)$$

Mathematically, the test for unconditional coverage (denoted as *(uc)*) is stated as follows,

$$H_{0,uc} : \pi = p$$

The likelihood function for a sample of T i.i.d. observations from a Bernoulli variable, I_t, with known probability p is written as

$$L(I,p) = p^{T_1}(1 - p)^{T-T_1}. \quad (7.3)$$

where T_1 is the number of ones in the sample. The likelihood function for an i.i.d. Bernoulli with unknown probability parameter, π_1, to be estimated is

$$L(I,\pi_1) = \pi_1^{T_1}(1 - \pi_1)^{T-T_1}. \quad (7.4)$$

The maximum-likelihood (ML) estimate of π_1 is

$$\hat{\pi}_1 = T_1/T \quad (7.5)$$

and we can thus write a likelihood ratio test of unconditional coverage as

$$LR_{uc} = 2(ln\ L(I,\hat{\pi}_1) - ln\ L(I,p)). \quad (7.6)$$

The number of subportfolios for which we do not reject the null at the 1% confidence level is 401, which represents 67% of the sample. Out of the 196 subportfolios that failed the unconditional coverage test, there are 83 subportfolios with zero exceedances. Table 7.2 summarizes results of the Kupiec test by providing counts and shares of subportfolios which pass (fail to reject the null hypothesis) categorizing subportfolios by product. VaR models of muni bonds, single-name credit default swaps (CDS), and tranched products appear to perform very well relative to other products both, for single product portfolios as well as for combinations of products. However, the small sample size of these categories reduces the statistical significance of this conclusion. Instead, if we consider products consisting of larger subportfolio groups, we note that linear equities, interest rate products, FX and commodities have passage rates of 70% for single and in excess of 64% for multiple product categories, implying that, relatively speaking, these products lend themselves to more "accurate" modeling.

It's important to point out the limitations of the unconditional coverage test, most notable of which is that the test ignores

Table 7.2 Count and Percent of Subportfolios Pass the Kupiec Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	6	175	100	66
Corporate Bonds	6	118	75	67
Muni Bonds	3	32	100	73
Agency MBS	0	25	0	53
Non Agency MBS	–	26	–	59
CMO	–	21	–	52
Index CDS	1	118	50	71
Single Name CDS	2	131	100	71
Tranched	1	43	100	70
Linear Equities	14	109	70	64
Nonlinear Equities	1	90	100	66
Exotic Equities	–	31	–	60
Interest Rates	21	256	70	67
FX	31	213	72	70
Commodities	17	43	74	67
Other Product Type	9	128	60	67

the time variation in the data. This limitation is mitigated through the independence test proposed by Christoffersen (1998) and noted by Berkowitz and O'Brien (2002). Rather than focusing on the expected vs observed fraction of violations, Christoffersen focuses on whether the observed violations cluster or evenly spread out over time. The *independence test* assumes that the sequence of exceptions follows a first-order Markov process with a switching probability matrix such that,

$$\Pi = \begin{bmatrix} 1 - \pi_{01} & \pi_{01} \\ 1 - \pi_{11} & \pi_{11} \end{bmatrix}$$

where π_{ij} is the probability of an i on day $t - 1$ being followed by a j on day t. The test of independence (denoted as *ind*) is specified as,

$$H_{0,ind} : \pi_{01} = \pi_{11}.$$

The likelihood ratio for the test of the null independence hypothesis can be computed using the likelihood function

$$L(I,\pi_{01},\pi_{11}) = (1 - \pi_{01})^{T_0 - T_{01}}\pi_{01}^{T_{01}}(1 - \pi_{11})^{T_1 - T_{11}}\pi_{11}^{T_{11}}, \quad (7.7)$$

where T_{ij} denotes the number of observations with a j following an i, and T_i is the number of i (i.e., is the number of ones or zeros in the sample).

The ML estimates are

$$\hat{\pi}_{01} = T_{01}/T_0 \tag{7.8}$$

$$\hat{\pi}_{11} = T_{11}/T_1 \tag{7.9}$$

and the independence test statistic is

$$LR_{ind} = 2(ln\ L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - ln\ L(I, \hat{\pi}_1)). \tag{7.10}$$

The number of subportfolios for which we fail to reject the null of the Christoffersen independence test at the 1% confidence level is 551, which represents 92% of the total number of subportfolios. Table 7.3 summarizes the results of the test (i.e., counts of subportfolios which fail to reject the null) categorizing the subportfolios by product. Here too, linear equities, interest rate products, FX and commodities appear to have high pass rates, in excess of 85% for single products and exceeding 94% for subportfolios containing multiple products.

We previously mentioned that an accurate VaR measure must exhibit the properties of unconditional coverage AND independence. It follows that in order to assess accuracy, a model must be evaluated against these tests jointly. The unconditional coverage and independence tests can be combined to test a sequence for *conditional coverage* (denoted cc). Empirically, this is represented as,

$$H_{0,cc} : \pi_{01} = \pi_{11} = p.$$

The likelihood ratio for the test of the null of conditional coverage hypothesis can be computed using the previous likelihood function and adding the coverage restriction:

$$LR_{cc} = 2(ln\ L(I, \hat{\pi}_{01}, \hat{\pi}_{11}) - ln\ L(I, p)). \tag{7.11}$$

The independence and conditional coverage tests are carried out conditioning on the first observation. The tests are asymptotically distributed as chi-square with degree of freedom one for the *uc* and *ind* tests and two for the cc test. But we will rely on finite sample *p*-values below.

Practically, the test checks whether the proportion of violations following an exceedance is the same as the proportion of violations following a non-exceedance, while simultaneously determining whether each of these proportions is significantly different from the pre-determined confidence level (here, 1%).

The number of subportfolios for which we fail to reject the null of conditional coverage at the 1% confidence level is 422, which represents 71% of the sample. Table 7.4 summarizes the test results (i.e., fail to reject the null) categorizing the subportfolios by product. As with individual tests of unconditional coverage and independence, linear equities, interest rate products, FX and commodities continue to outperform other products when it comes to passage rates.

Table 7.3 Count and Percent of Subportfolios That Pass the Independence Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	6	251	100	94
Corporate Bonds	8	163	100	92
Muni Bonds	0	34	0	77
Agency MBS	0	43	0	91
Non Agency MBS	–	41	–	93
CMO	–	36	–	90
Index CDS	2	161	100	96
Single Name CDS	2	174	100	95
Tranched	1	59	100	97
Linear Equities	15	156	75	92
Nonlinear Equities	1	127	100	93
Exotic Equities	–	50	–	96
Interest Rates	29	357	97	93
FX	42	287	98	94
Commodities	22	63	96	98
Other Product Type	13	175	87	92

While joint tests may perform well at identifying instances where both properties have been violated, they have significant limitations in their ability to detect misspecifications when only one of the two properties is violated.

While independence and conditional coverage tests are considered to have greater discriminatory power than unconditional coverage tests, it is important to recognize that they too have significant drawbacks. Specifically, the Christoffersen test does not allow for the possibility that independence may be violated at a distance greater than one lag. While it may be true that yesterday's exceedance has no information about the probability of observing an exceedance today, the test provides no information about the dependence of breaches one week apart. Or whether exceedances occurring one month apart are independent from one another.

In addition to Kupiec and Christoffersen, other tests have been emphasized for their ability to identify misspecified

Table 7.4 Count and Percent of Subportfolios That Pass the Conditional Coverage Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	6	190	100	71
Corporate Bonds	7	125	88	71
Muni Bonds	0	27	0	61
Agency MBS	0	30	0	64
Non Agency MBS	–	29	–	66
CMO	–	26	–	65
Index CDS	2	127	100	76
Single Name CDS	2	138	100	75
Tranched	1	50	100	82
Linear Equities	13	118	65	69
Nonlinear Equities	1	91	100	66
Exotic Equities	–	35	–	67
Interest Rates	20	274	67	72
FX	34	228	79	75
Commodities	18	49	78	77
Other Product Type	9	137	60	72

Table 7.5 Count and Percent of Subportfolios That Pass the Ljung–Box Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	5	161	83	83
Corporate Bonds	6	110	75	81
Muni Bonds	0	27	0	71
Agency MBS	0	20	0	71
Non Agency MBS	–	16	–	73
CMO	–	16	–	70
Index CDS	1	99	100	88
Single Name CDS	2	112	100	85
Tranched	1	38	100	90
Linear Equities	11	122	58	81
Nonlinear Equities	1	96	100	82
Exotic Equities	–	35	–	85
Interest Rates	22	245	92	83
FX	36	226	92	85
Commodities	18	52	90	93
Other Product Type	9	120	69	81

models. Among others, these include the *Ljung–Box test* (1978) which examines the autocorrelation properties of hits and Chrisoffersen and Pelletier (2004) test of independence of time between exceptions.

Ljung–Box test allows us to jointly test whether m number of lags are zero. To do so, we need to calculate the following test statistic,

$$Q = T(T+2)\sum_{k=1}^{m}\frac{r_k^2}{T-k} \tag{7.12}$$

which is asymptotically chi-square with m degrees of freedom.

The null of Ljung–Box test assumes no autocorrelation of exceedances. The number of subportfolios for which we fail to reject the null of autocorrelation at a single lag at 1% confidence level is 387, which represents 82% of the total number of subportfolios. Table 7.5 summarizes the test results (i.e., fail to reject the null) categorizing the subportfolios by product. FX appears to demonstrate a strongest performance relative to other products for both, single as well as multiple product categories.

The Ljung–Box test can be extended as in Berkowitz et al. (2011) to include other available information as predictors of exceedances, providing further evidence that the risk model has been misspecified, in this case by not incorporating all relevant past information. As in Berkowitz et al., we estimate a linear probability model (LPM) and logit regressions of the exceedances on their lags and, additionally, lagged VaR or lagged P&L. Both regressions use a binary outcome variable based on exceedances. For linear probability model regressions, errors are assumed to have zero conditional mean, but there is no explicit parametric assumption on their distribution.

While LPM is simple to estimate and constitutes a robust check of the independence property, it has non-trivial drawbacks. For instance, the error term of a binary variable has a Bernoulli structure, the LPM can yield values of predicted values of the dependent variable outside of the [0, 1] interval, and it imposes linearity on the relationship between the dependent variable and the right-hand side variables. In light of these limitations, a more appropriate model specification for the error structure is the logistic distribution where expectation of an exceedance is,

$$E[I_{t+1}|X_t] = p_t = logit^{-1}(\beta \cdot X_t) = \frac{1}{1+e^{-\beta \cdot X_t}} \tag{7.13}$$

where I_{t+1} is the indicator variable if the subportfolio observed a breach on date $t + 1$, while X_t is the information available

to risk manager at the time of the model forecast, i.e., past exceedances I_t, I_{t-1}, etc., past portfolio returns PL_t, PL_{t-1}, etc., and past information sets that were included in previous VaR models VaR_t, VaR_{t-1}.

In our analysis, we focused on estimating models with a single lagged explanatory variable and tested whether the β coefficients are statistically significantly different from zero. Observing significant coefficients indicates that the forecasting model for losses does not respond sufficiently fast to the observed changes in market environment. The results of these tests are summarized in Table 7.6, which provides the counts of subportfolios with significant coefficients at a 1% confidence level for various specifications of the Linear Probability and Logit models. For example, the first logit specification includes lag exceedances and lag VaR as explanatory variables. Out of 597 logit regressions, fifty-four regressions had a significant coefficient on lag exceedances and thirty-six had a significant coefficient on lag VaR. For seven regressions both lag exceedence and lag VaR were significant.

In a seminal work, Christoffersen and Pelletier (2004) demonstrated that the tests described thus far can lack power. In order to increase the robustness of results, Christoffersen and Pelletier proposed a duration test, which assesses the information content of time between observed exceedances. Broadly, the intuition behind the duration-based tests is that the clustering of violations will result in an excessive number of relatively short and relatively long durations, where no exceptions are observed, corresponding to market turbulence and market calm, respectively. Ideally, the time duration between VaR exceedances should not cluster, implying that in a correctly specified risk model, the no-exception duration should have no memory. This Christoffersen and Pelletier duration test will be able to identify models with dependent exceptions, but also models that do not producce a sufficient number of expected exceptions.

Duration between exceedances is defined as

$$D_i = t_i - t_{i-1}. \tag{7.14}$$

Using the Bernoulli distribution property, the probability of a breach in the next period is exactly equal to

$$Pr(D_i = 1) = Pr(I_{t+1} = 1) = p. \tag{7.15}$$

It follows that the probability of a breach in d periods can be defined as,

$$Pr(D_i = d) = Pr(I_{t+1} = 0, I_{t+2} = 0, \ldots, I_{t+d} = 1), \tag{7.16}$$

and since

$$I_{t+1} \sim \text{iid}(p, p(1-p)), \tag{7.17}$$

we have

$$Pr(D_i = d) = (1-p)\ldots(1-p)(p) = (1-p)^{d-1}p. \tag{7.18}$$

For a correctly specified model, with no dependence between exceedances, the distribution for duration is exponential,

$$f_{exp}(D; p) = pe^{-pD}. \tag{7.19}$$

We test the null against the alternative hypothesis that duration has a Weibull distribution,

$$f_w(D;a,b) = a^b bD^{b-1}\, exp^{-(aD)^b}. \tag{7.20}$$

Statistically, the independence of the duration test can be summarized as follows,

$$H_o : b = 1 \text{ and } a = p;\ H_a : b \neq 1 \text{ or } a \neq p. \tag{7.21}$$

A total of 77 subportfolios fail the duration test at the 1% confidence level, implying that in effect, the conditional expected duraton between violations is not constant. The fraction of failed supportfolios represents 13% of the overall sample.

Table 7.6 **Logit and LPM Regressions of Exceedances on Lagged VaR or Lagged P&L.**

	Logit (1)	Logit (2)	Logit (3)	LPM (1)	LPM (2)	LPM (3)
Lag Exceedance	54	45	68	83	84	87
Lag VaR	36	–	–	45	–	–
Lag PnL	–	33	–	–	54	–
Both Lags	7	4	–	17	18	–
Neither Lags	514	523	529	361	352	385
N	597	597	597	472	472	472

7.6 MISSPECIFICATION TESTS BASED ON THE DISTRIBUTION OF PITS

In the previous section, we discussed the backtesting of VaR models using the "hit" series of exceedances. While this approach to backtesting allows a simplistic assessment of model quality, we can enhance the precision of our estimates of model accuracy and fit by using more sophisticated backtesting approaches. We are

referring to the practice of backtesting with PITs and assessing their entire distribution for properties of independence and uniformity.

A wide range of goodness-of-fit tests (e.g., Kolmogorov–Smirnov, Anderson–Darling, Cramér–von Mises) can be used to assess the uniformity of the distribution of PITs. *Kolmogorov–Smirnov test* (KS) is a widely used nonparametric test that evaluates the goodness-of-fit by comparing the empirical distribution function of a sample with some reference distribution function. Operationally, the KS statistic quantifies the distance between the empirical CDF of the sample and the CDF of the reference distribution.

As in the previous section, the reference distribution function against which PIT distribution is compared is the continuous uniform distribution over the [0, 1] interval and the KS test allows us to determine whether the data is in fact drawn from the distribution in question. The KS test statistic is defined as,

$$D = \max_j\{\text{abs}(z_j - A_j)\} \tag{7.22}$$

where z_j is the theoretical CDF under the null, A_j the empirical CDF and D is equal to zero under the null hypothesis.

The number of subportfolios for which we fail to reject null of the Kolmogorov–Smirnov test at 1% confidence level is 123, which is approximately 21% of the entire sample. The low number of subportfolios that "pass" the KS test is not surprising given the deviation from uniformity that we observed by representing the PIT statistics as histograms in Section 7.3. Table 7.7 summarizes the count and percentage of subportfolios for which we fail to reject the null of uniformity.

We should note that a limitation of the KS test is its lack of sensitivity at the extremes of the distribution. That is, while the test is effective in detecting differences in central masses of the two distributions, it is not reliable when it comes to comparison of the tails. *Anderson–Darling test* (A-D) is a modification of the Kolmogorov–Smirnov test that corrects for this limitation. Unlike the KS test, A-D puts more weight on the tails of the distribution, yielding more power in the presence of distributional biases. A majority of financial institutions use a tail metric as a risk measure, e.g. 99% VaR. Therefore, any misspecification in the tail of the distribution of the risk model will be better identified by the A-D test.

Statistically speaking, Anderson–Darling tests are to assess if the data are independently and identically distributed from a specific distribution. Here, we are interested in assessing whether the subportfolio PITs may be from a uniform [0, 1] distribution, which is reflected in the null hypothesis. Our alternative hypothesis is that the empirical CDF of PITs is a function other than uniform.

Table 7.7 Count and Percent of Subportfolios That Pass the Kolmogorov–Smirnov Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	1	42	17	16
Corporate Bonds	0	22	0	12
Muni Bonds	1	9	33	20
Agency MBS	0	1	0	2
Non Agency MBS	–	2	–	5
CMO	–	1	–	2
Index CDS	1	25	50	15
Single Name CDS	1	31	50	17
Tranched	0	10	0	16
Linear Equities	7	36	35	21
Nonlinear Equities	0	26	0	19
Exotic Equities	–	11	–	21
Interest Rates	11	75	37	20
FX	16	73	37	24
Commodities	8	14	35	22
Other Product Type	1	33	7	17

The A-D test statistic is calculated as

$$A^2 = -n - \frac{1}{n}\sum_{j=1}^{n}(2j - 1)[log(z_j) + log(1 - z_{-j})] \tag{7.23}$$

where n is the number of observations. As with the KS, the test statistic is equal to zero under the null hypothesis.

The number of subportfolios for which we fail to reject null of the Anderson–Darling test at 1% confidence level is 65, a mere 11% of the total number of subportfolios. It is clear that as the power of the test increases, we are able to find fewer subportfolios which exhibit the desired properties, summarized also by product in Table 7.8.

Cramér–von Mises test is another variation of the Kolmogorov–Smirnov test. However, rather than using the supremum, it relies on the mean squared deviation of the distribution. Cramér–von Mises tests the goodness-of-fit relative to the continuous uniform distribution [0, 1]. Here, too, we assume that the values are independent, identically distributed random values. The null hypothesis is that the empirical CDF is cumulative uniform, while the alternative hypothesis is that the distribution is some other function.

Table 7.8 Count and Percent of Subportfolios That Pass the Anderson–Darling Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	1	26	17	10
Corporate Bonds	0	13	0	7
Muni Bonds	0	3	0	7
Agency MBS	0	1	0	2
Non Agency MBS	–	2	–	5
CMO	–	1	–	2
Index CDS	0	13	0	8
Single Name CDS	0	17	0	9
Tranched	0	7	0	11
Linear Equities	6	16	30	9
Nonlinear Equities	0	10	0	7
Exotic Equities	–	5	–	10
Interest Rates	7	40	23	10
FX	8	36	19	12
Commodities	5	7	22	11
Other Product Type	1	19	7	10

Table 7.9 Count and Percent of Subportfolios That Pass the Cramér–Von Mises Test.

Product	Count single	Count multiple	Percent single	Percent multiple
Sovereign Bonds	1	48	17	18
Corporate Bonds	0	22	0	12
Muni Bonds	1	10	33	23
Agency MBS	1	3	100	6
Non Agency MBS	–	3	–	7
CMO	–	3	–	8
Index CDS	1	27	50	16
Single Name CDS	1	32	50	17
Tranched	0	11	0	18
Linear Equities	7	36	35	21
Nonlinear Equities	0	27	0	20
Exotic Equities	–	11	–	21
Interest Rates	13	83	43	22
FX	16	78	37	26
Commodities	8	13	35	20
Other Product Type	1	34	7	18

Cramér–von Mises test statistic is calculated as follows,

$$W^2 = \sum_{j=1}^{n} [z_j - (2j - 1)/2n]^2 + (1/12n). \qquad (7.24)$$

As with the KS, it is equal to zero under the null hypothesis.

The number of subportfolios for which we fail to reject null of CVM test at 1% confidence level is 123, 21% of the total number of subportfolios. Table 7.9 shows the results with single name CDS, interest rates and FX exhibiting stronger passing rates than other products.

To parallel the independence test of exceedances, we proceed by assessing the *independence* of the PIT realizations. For this, we transform PITs into standard normal variables to avoid problems of inference due to bounded support, see Berkowitz (2001). Their independence is then assessed by fitting a linear regression of transformed PITs on their lags as well as lagged values of the P&L. Table 7.10 summarized the counts of subportfolios with statistically significant coefficients on the lagged variables at a 1% confidence level.

Table 7.10 Linear Regression of Transformed PITs on Lagged Transformed PIT and Lagged P&L.

	AR1	AR conditioned
Lag Trans PIT	218	154
Lag PnL	–	108
Both Lags	–	79
Neither Lags	379	414
N	*597*	*597*

Conditional uniformity of the distribution of PITs can be visually assessed by plotting pairs of PITs and lagged PIT values for all subportfolios. The scatter plot, in Figure 7.5, provides graphical evidence for the goodness-of-fit and identifies pairs of PIT and lagged PIT values that do not follow the independence and uniformity properties of an accurate model. We observe significantly more mass in the center and very extremes of the distribution, suggesting dependence and lack of uniformity in the bivariate distribution.

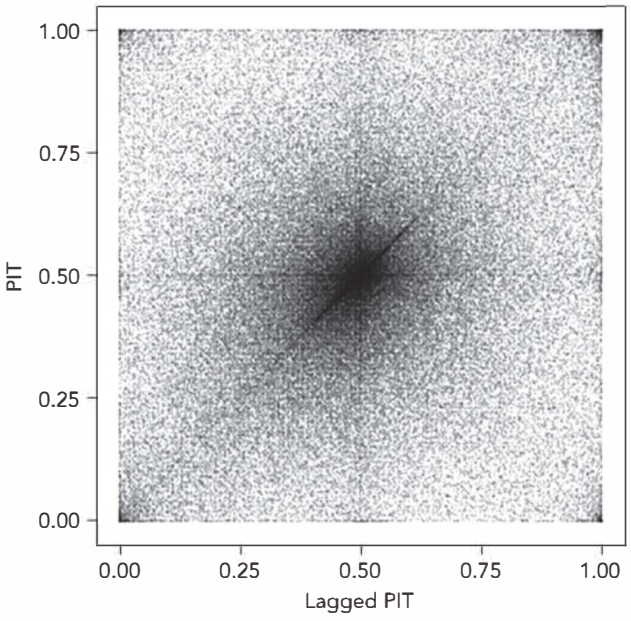


Figure 7.5 PITs vs lagged PITs by subportfolios.

7.7 CONCLUSION

Exceedance-based backtesting of risk models and, in particular, VaR models, is a well established methodology for which the financial firms have appropriately adapted their models. For example, we found that 67% of the sample passes the unconditional coverage test and moreover, if we are to disregard subportfolios with zero exceedances 78% subportfolios "pass" the unconditional coverage test. When performing an independence test based on logit regressions of exceedances, 11% of subportfolios had auto correlated exceedances. However, pure exceedance analysis fails to identify the more nuanced model misspecifications, as highlighted by the enhanced backtesting based on PITs. When performing distributional tests for uniformity we find that only 21% of the entire sample pass the Kolmogorov–Smirnov test. Moreover, when checking for independence of PITs based on autoregressive estimations, we found a larger portion of subportfolios, 37% that exhibited dependence in PITs. Overall, banks perform better on exceedence tests than on tests of distribution. This could be because risk models are designed to be more conservative and contain regulatory add ons.

References

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics,* 19(4), 465--474.

Berkowitz, J. and O'Brien, J. (2002). How Accurate Are Value-at-Risk Models at Commercial Banks.*Journal of Finance,* 57(3), 1093–1111.

Berkowitz, J., Christoffersen, P. and Pelletier, D. (2016). Evaluating Valueat-Risk models with desk-level data. *Management Science,* 57(12), 2213–2227.

Campbell, S. (2006). A review of backtesting and backtesting procedures. *Journal of Risk,* 9(2), 1–17.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review,* 39, 841–862.

Christoffersen, P., Hahn, J. and Inoue A. (2001). Testing and comparing Value-at-Risk measures. *Journal of Empirical Finance,* 8, 325–342.

Christoffersen, P. and Pelletier, D. (2004). Backtesting Value-at-Risk: A duration-based approach. *Journal of Financial Econometrics,* 2(1), 84–108.

Diebold, F., Gunther, T., and Tay, A. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review,* 39(4), 863–883.

Federal Register (2013). Market Risk Capital Rule; Vol. 78 No. 198.

J. orion, P. (2002). How informative are Value-at-Risk disclosures. *The Accouting Review,* 77(4), 911–931.

Kupiec P. (1995). Techniques for verifying the accuracy of risk management models. *Journal of Derivatives,* 3, 73–84.

Ljung, G. and Box, G. (1978). On a measure of a lack of fit in time series models. *Biometrika,* 65(2), 297–303.

Marshall, C. and Siegel, M. (1997). Value-at-Risk: Implementing a risk measurement standard. *Journal of Derivatives,* 4, 91–111.

Noceti, P., Smith, J. and Hodges, S. (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting,* 22(6–7), 447–455.

Pritsker, M. (1997). Evaluating Value-at-Risk methodologies: Accuracy versus computational time. *Journal of Financial Services Research,* 12, 201–242.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics,* 23, 470–472.

8

Correlation Basics: Definitions, Applications, and Terminology

Learning Objectives

After completing this reading, you should be able to:

- Describe financial correlation risk and the areas in which it appears in finance.
- Explain how correlation contributed to the global financial crisis of 2007–2009.
- Describe how correlation impacts the price of quanto options as well as other multi-asset exotic options.
- Describe the structure, uses, and payoffs of a correlation swap.
- Estimate the impact of different correlations between assets in the trading book on the VaR capital charge.
- Explain the role of correlation risk in market risk and credit risk.
- Explain how correlation risk relates to systemic and concentration risk.

Excerpt is Chapter 1 of Correlation Risk Modeling and Management, *2nd Edition, by Gunter Meissner.*

"Behold the fool saith, 'Put not all thine eggs in the one basket'"

—Mark Twain

In this introductory chapter, we define correlation and correlation risk, and show that correlations are critical in many areas of finance such as investments, trading, and risk management, as well as in financial crises and in financial regulation. We also show how correlation risk relates to other risks in finance such as market risk, credit risk, systemic risk, and concentration risk. Before we do, let's see how it all started.

8.1 A SHORT HISTORY OF CORRELATION

As with many groundbreaking discoveries, there is a bit of a controversy as to who the creator of the concept of correlation is. Foundations on the behaviour of error terms were laid in 1846 by the French mathematician Auguste Bravais, who essentially derived what is today termed the "regression line". However, Helen Walker (1929) describes Bravais nicely as "a kind of Columbus, discovering correlation without fully realising that he had done so". Further significant theoretical and empirical work on correlation was done by Sir Walter Galton in 1886, who created a simple linear regression and interestingly also discovered the statistical property of "Regression to Mediocrity", which today we call "Mean-Reversion".

A student of Walter Galton, Karl Pearson, whose work on relativity, antimatter and the fourth dimension inspired Albert Einstein, expanded the theory of correlation significantly. Starting in 1900, Pearson defined the correlation coefficient as a product moment coefficient, introduced the method of moments and principal component analysis, and founded the concept of statistical hypothesis testing, applying P-Values and Chi-squared distances.

8.2 WHAT ARE FINANCIAL CORRELATIONS?

Heuristically (meaning non-mathematically), we can define two types of financial correlations, static and dynamic:

(a) Definition: static financial correlations measure how two or more financial assets are associated at a certain point in time or within a certain time period.

Examples are:

1. Correlating bond prices and their respective yields at a certain point in time, which will result in a negative association.
2. The classic VaR (value-at-risk) model, which answers the question: what is the maximum loss of correlated assets in a portfolio with a certain probability for a given time period (see "Risk management and correlation" below).
3. The copula approach for CDOs (collateralised debt obligations). It measures the default correlations between all assets in the CDO, typically 125, for a certain time period.
4. The binomial default correlation model of Lucas (1995), which is a special case of the Pearson correlation model. It measures the probability of two assets defaulting together within a short time period.

Besides the static correlation concept, there are dynamic correlations:

(b) Definition: dynamic financial correlations measure how two or more financial assets move together in time.

Examples are:

1. In practice, "pairs trading" – where one asset is purchased and another is sold – is performed. Let's assume that the asset returns x and y have moved highly correlated in time. If now asset X performs poorly with respect to Y, then asset X is bought and asset Y is sold with the expectation that the gap will narrow.
2. Within the deterministic correlation approaches, the Heston model (1993) correlates the Brownian motions dz_1 and dz_2 of assets 1 and 2. The core equation is $dz_1(t) = \rho dz_2(t) + \sqrt{(1 - \rho^2)}dz_3(t)$ where dz_1 and dz_2 are correlated in time with correlation parameter ρ.
3. Correlations behave random and unpredictable. Therefore, it is a good idea to model them as a stochastic process. Stochastic correlation processes are by construction time-dependent and can replicate correlation properties well.

"Suddenly everything was highly correlated"

Financial Times, April 2009

8.3 WHAT IS FINANCIAL CORRELATION RISK?

Financial correlation risk is defined as the risk of financial loss due to adverse movements in correlation between two or more variables. These variables can comprise any financial variables. For example, the positive correlation between Mexican bonds

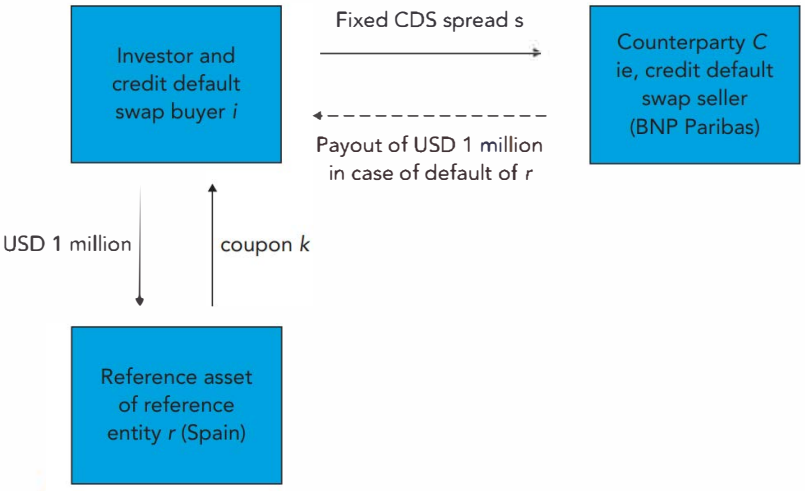


Figure 8.1 An investor hedging their Spanish bond exposure with a CDS.

and Greek bonds can hurt Mexican bond investors, if Greek bond prices decrease, which happened in 2012 during the Greek crisis. Or the negative correlation between commodity prices and interest rates can hurt commodity investors if interest rates rise. A further example is the correlation between a bond issuer and a bond insurer, which can hurt the bond investor (see the example displayed in Figure 8.1).

Correlation risk is especially critical in risk management. An increase in the correlation of asset returns increases the risk of financial loss, which is often measured by the VaR concept. For details see "Risk management and correlation" below. An increase in correlation is typical in a severe, systemic crisis. For example, during the great recession from 2007 to 2009, financial assets and financial markets worldwide became highly correlated. Risk managers who had negatively or low correlated assets in their portfolio suddenly witnessed many of them decline together, hence asset correlations increased sharply. For more on systemic risk, see "The global financial crises 2007 to 2009 and correlation" below as well as Chapter 9, which displays empirical findings of correlations.

Correlation risk can also involve variables that are non-financial as economic or political events. For example, the correlation between the increasing sovereign debt and currency value can hurt an exporter, as in Europe in 2012, where a decreasing euro hurt US exporters. Geopolitical tensions, as for example in the Middle East, can hurt airline companies due to the increasing oil price, or a slowing GDP in the US can hurt Asian and European exporters and investors, since economies and financial markets are correlated worldwide.

Let's look at correlation risk via an example of a credit default swap (CDS). A CDS is a financial product in which the credit risk is transferred from the investor (or CDS buyer) to a counterparty (CDS seller). Let's assume an investor has bought USD 1 million in a bond from Spain. They are now worried about Spain defaulting and have purchased a CDS from a French bank, BNP Paribas. Graphically this is displayed in Figure 8.1.

The investor is protected against a default from Spain since, in case of default, the counterparty BNP Paribas will pay the originally invested USD 1 million to the investor. For simplicity, let's assume the recovery rate and accrued interest are zero.

The value of the CDS, ie, the fixed CDS spread s,[1] is mainly determined by the default probability of the reference entity Spain. However, the spread s is also determined by the joint default correlation of BNP Paribas and Spain. If the correlation between Spain and BNP Paribas increases, the present value of the CDS for the investor will decrease and they will suffer a paper loss. Worst-case scenario is the joint default of Spain and BNP Paribas, in which case the investor will lose their entire investment in the Spanish bond of USD 1 million.

In other words, the investor is exposed to default correlation risk between the reference asset r (Spain) and the counterparty c (BNP Paribas). Since both Spain and BNP Paribas are in Europe, let's assume that there is a positive default correlation between the two. In this case, the investor has "Wrong-Way Correlation Risk" or, for short, "Wrong-Way Risk" (WWR). Let's assume the default probabilities of Spain and BNP Paribas both increase. This means that the credit exposure to the reference entity Spain increases (since the CDS has a higher present value for the investor) *and* the credit risk increases, since it is more unlikely that the counterparty BNP Paribas can pay the default insurance.

The magnitude of the correlation risk is expressed graphically in Figure 8.2.

From Figure 8.2, we observe that for a correlation of -0.3 and higher, the higher the correlation, the lower is the CDS spread. This is because an increasing ρ means a higher probability of the reference asset and the counterparty defaulting together. In the extreme case of a perfect correlation of 1, the CDS is worthless. This is because, if Spain defaults, so will the insurance seller BNP Paribas.

We also observe from Figure 8.2 that, for a correlation from about -0.3 to -1, the CDS spread decreases slightly. This seems counterintuitive at first. However, an increase in the negative

[1] The CDS spread s is the premium or fee that the CDS buyer pays for getting protection. It is called a spread since it is approximately the spread between the yield of the risky bond (the bond of Spain in Figure 8.1) in the CDS minus the yield of a riskless bond. See Meissner 2005, for details.

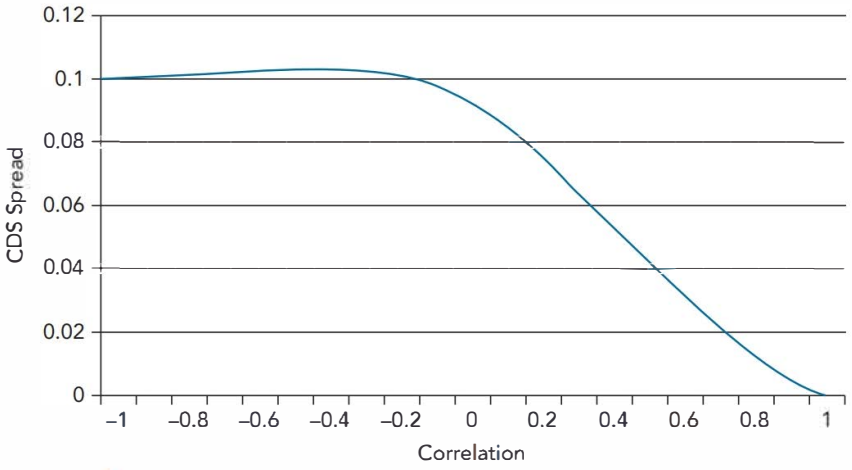


Figure 8.2 **CDS spread s of a hedged[2] bond purchase (as displayed in Figure 8.1) with respect to the default correlation between the reference entity *r* and the counterparty *c*.**

correlation means a higher probability of either Spain or BNP Paribas defaulting. Hence we have two scenarios: (a) in the case of Spain defaulting (and BNP Paribas surviving) the CDS buyer will get compensated by BNP Paribas; (b) if the insurance seller BNP Paribas defaults (and Spain survives), the CDS buyer will lose his insurance and will have to repurchase it. This may have to be done at a higher cost. The cost will be higher if the credit quality of Spain has decreased since inception of the original CDS. For example, the CDS spread may have been 3% in the original CDS, but may have increased to 6% due to a credit deterioration of Spain. The scenarios (a) and (b) combined lead to a slight decrease of the CDS spread. For more details on pricing CDSs with counterparty risk and the reference asset – counterparty correlation – see Kettunen and Meissner (2006).

We observe from Figure 8.2 that the dependencies between a variable (here the CDS spread) and correlation may be non-monotonic, ie, the CDS spread sometimes increases and sometime decreases if correlation increases.

8.4 MOTIVATION: CORRELATIONS AND CORRELATION RISK ARE EVERYWHERE IN FINANCE

Why study financial correlations? That's an easy one. Financial correlations appear in many areas in finance. We will briefly discuss five areas: (1) investments, (2) trading, (3) risk management, (4) the global financial crisis and (5) regulation. Naturally, if an entity is exposed to correlation, this means that the entity has correlation risk, ie, the risk of a change in the correlation.

Investments and Correlation

From our studies of the Nobel Prize-rewarded Capital Asset Pricing Model (Markowitz (1952), Sharpe (1964)), we remember that an increase in diversification increases the return/risk ratio. Importantly, high diversification is related to low correlation. Let's show this in an example. Let's assume we have a portfolio of two assets, X and Y. They have performed as in Table 8.1.

Let's define the return of asset X at time t as x_t, and the return of asset Y at time t as y_t. A return is calculated as a percentage change, $(S_t - S_{t-1})/S_{t-1}$, where S is a price or a rate. The average return of asset X for the timeframe 2014 to 2018 is $\mu_X = 29.03\%$; for asset Y the average return is $\mu_Y = 20.07\%$. If we assign a weight to asset X, w_X, and a weight to asset Y, w_Y, the portfolio return is:

$$\mu_P = w_X\mu_X + w_Y\mu_Y \tag{8.1}$$

where $w_X + w_Y = 1$

The standard deviation of returns, called *volatility*, is derived for asset X with equation:

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \mu_X)^2} \tag{8.2}$$

where x_t is the return of asset X at time t and n is the number of observed points in time. The volatility of asset Y is derived accordingly. Equation 8.2 can be computed with = stdev in Excel and std in MATLAB. From our example in Table 8.1, we find that $\sigma_X = 44.51\%$ and $\sigma_Y = 47.58\%$.

Table 8.1 **Performance of a Portfolio with Two Assets**

Year	Asset X	Asset Y	Return of Asset X	Return of Asset Y
2013	100	200		
2014	120	230	20.00%	15.00%
2015	108	460	−10.00%	100.00%
2016	190	410	75.93%	−10.87%
2017	160	480	−15.79%	17.07%
2018	280	380	75.00%	−20.83%
		Average	29.03%	20.07%

[2] To hedge means to protect More precisely, hedging means to enter into a second trade to protect against the risk of an orginal trade.

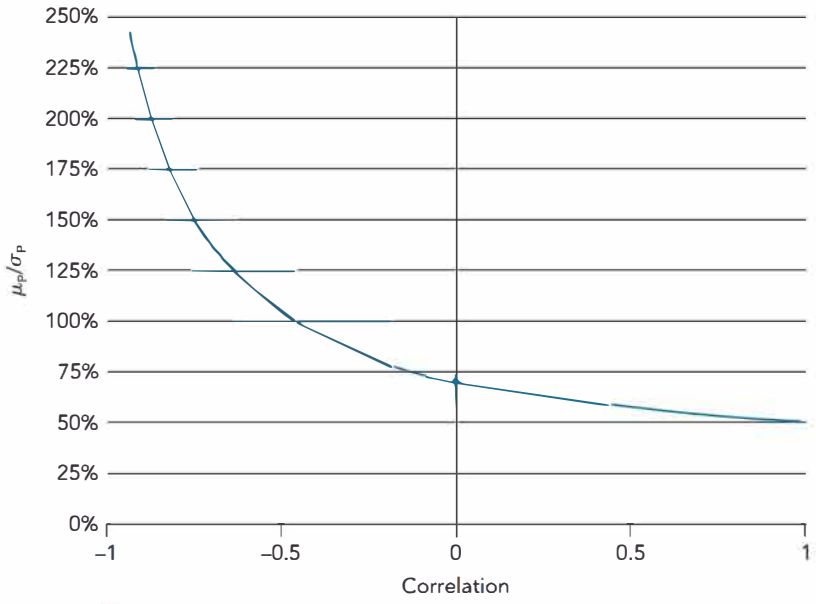


Figure 8.3 **The negative relationship of the portfolio return/portfolio risk ratio μ_P/σ_P with respect to the correlation ρ of the assets in the portfolio (input data are from Table 8.1).**

Let's now look at the covariance. The covariance measures how two variables "co-vary", ie, move together. More precisely, the covariance measures the strength of the linear relationship between two variables. The covariance of returns for assets X and Y is derived with equation:

$$COV_{XY} = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \mu_X)(y_t - \mu_Y) \tag{8.3}$$

For our example in Table 8.1 we derive $COV_{XY} = -0.1567$. Equation (8.3) is = Covariance. S in Excel and cov in MATLAB. The covariance is not easy to interpret, since it takes values between $-\infty$ and $+\infty$. Therefore, it is more convenient to use the Pearson correlation coefficient ρ_{XY}, which is a standardised covariance, ie, it takes values between -1 and $+1$. The Pearson correlation coefficient is:

$$\rho_{XY} = \frac{COV_{XY}}{\sigma_X \sigma_Y} \tag{8.4}$$

For our example in Table 1, $\rho_{XY} = -0.7403$, showing that the returns of assets X and Y are highly negatively correlated. Equation (8.4) is "correl" in Excel and "corrcoef" in MATLAB. For the derivation of the numerical examples of equations (8.2) to (8.4) and more information on the covariances see the appendix of Chapter 8 and www.dersoft.com/matrixprimer.xlsx, sheet "Covariance Matrix".

We can calculate the standard deviation for our two-asset portfolio P as:

$$\sigma_P = \sqrt{w_X^2\sigma_X^2 + w_Y^2\sigma_Y^2 + 2w_X w_Y COV_{XY}} \tag{8.5}$$

With equal weights, ie, $w_X = w_Y = 0.5$, the example in Table 8.1 results in $\sigma_P = 16.66\%$.

Importantly, the standard deviation (or its square, the variance) is interpreted in finance as risk. The higher the standard deviation, the higher the risk of an asset or a portfolio. Is standard deviation a good measure of risk? The answer is: it's not great, but it's one of the best there are. A high standard deviation may mean high upside potential of the asset in question! So it penalises possible profits! But high standard deviation naturally also means high downside risk. In particular, risk-averse investors will not like a high standard deviation, ie, high fluctuation of their returns.

An informative performance measure of an asset or a portfolio is the risk-adjusted return, ie, the return/risk ratio. For a portfolio it is μ_P/σ_P, which we derived in Equations (8.1) and (8.5). In Figure 8.3 we observe one of the few "free lunches" in finance: the lower (preferably negative) the correlation of the assets in a portfolio, the higher the return/risk ratio. For a rigorous proof, see Markowitz (1952) and Sharpe (1964).

Figure 8.3 shows the high impact of correlation on the portfolio return/risk ratio. A high negative correlation results in a return/risk ratio of close to 250%, whereas a high positive correlation results in a 50% ratio. The equations (8.1) to (8.5) are derived within the framework of the Pearson correlation approach.

"Only by great risks can great results be achieved"

Xerxes

8.5 TRADING AND CORRELATION

In finance every risk is also an opportunity. Therefore, at every major investment bank and hedge fund, correlation desks exist. The traders try to forecast changes in correlation and try to financially gain from these changes in correlation. We already mentioned the correlation strategy "pairs trading" above. Generally, correlation trading means trading assets, whose price is determined at least in part by the co-movement of one asset or more in time. Many types of correlation assets exist.

Many different types of multi-asset options, also called rainbow options or mountain-range options, are traded. S_1 is the price of asset 1 and S_2 is the price of asset 2 at option maturity. K is the strike price, or the price determined at option start at which the underlying asset can be bought in case of a call, or the price at which the underlying asset can be sold in case of a put.

- Option on the better of two. Payoff = max (S_1, S_2).
- Option on the worse of two. Payoff = min (S_1, S_2).
- Call on the maximum of two. Payoff = max [0, max(S_1, S_2) − K].
- Exchange option (such as a convertible bond). Payoff = max (0, S_2 − S_1).
- Spread call option. Payoff = max [0, (S_2 − S_1) − K].
- Option on the better of two or cash. Payoff = max (S_1, S_2, cash).
- Dual strike call option. Payoff = max (0, S_1 − K_1, S_2 − K_2).
- Basket option.

$$\left[\sum_{i=1}^{n} n_i S_i - K, 0\right]$$

where n_i is the weight of assets i.

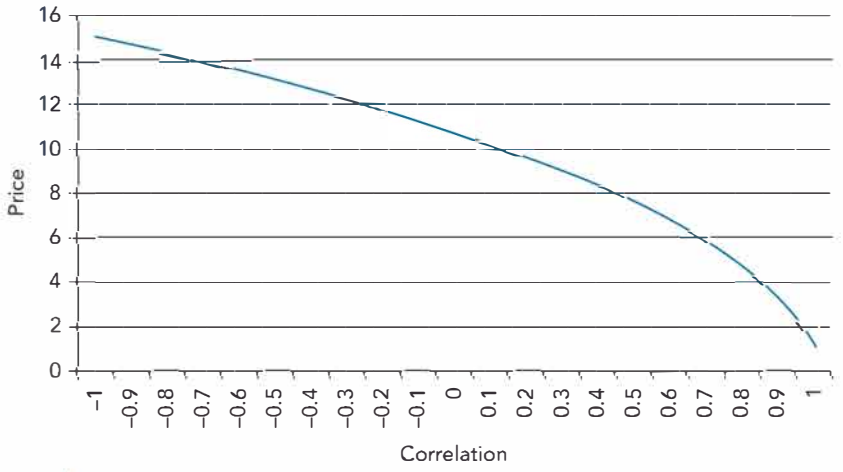


Figure 8.4 **Exchange option price with respect to correlation of the assets in the portfolio.**

Importantly, the price of these correlation options is highly sensitive to the correlation between the asset prices S_1 and S_2. In the list above, except for the option on the worse of two, and the basket option, the lower the correlation, the higher is the option price. This makes sense since a low, preferably negative correlation means that, if one asset decreases (on average), the other increases. So one of the two assets is likely to result in a high price and therefore in a high payoff. Multi-asset options can be conveniently priced analytically with extensions of the Black–Scholes–Merton option model (1973).

Let's look at the evaluation of an exchange option with a payoff of max(0, S_2 − S_1). The payoff shows that the option buyer has the right to give away Asset 1 and receive Asset 2 at option maturity. Hence, the option buyer will exercise their right, if $S_2 > S_1$. The price of the exchange option can be easily derived. We first rewrite the payoff equation max(0, S_2 − S_1) as: S_1 max(0, (S_2/S_1) − 1). We then input the covariance between asset S_1 and S_2 into the implied volatility function of the exchange option using a variation of equation (8.5):

$$\sigma_E = \sqrt{\sigma_A^2 + \sigma_B^2 - 2COV_{AB}} \qquad (8.5a)$$

where σ_E is the implied volatility of S_2/S_1, which is input into the standard Black–Scholes–Merton option pricing model (1973). For an exchange option pricing model and further discussion, see the model at www.dersoft.com/exchangeoption.xlsm.

Importantly, the exchange option price is highly sensitive to the correlation between the asset prices S_1 and S_2, as seen in Figure 8.4.

From Figure 8.4 we observe the strong impact of the correlation on the exchange option price. The price is close to 0 for high correlation and USD 15.08 for a negative correlation of −1. As in Figures 8.2 and 8.3, the correlation approach underlying Figure 8.4 is the Pearson correlation model.

Another interesting correlation option is the quanto option. This is an option that allows a domestic investor to exchange their potential option payoff in a foreign currency back into their home currency at a fixed exchange rate. A quanto option therefore protects an investor against currency risk. Let's assume an American believes the Nikkei will increase, but they are worried about a decreasing yen, which would reduce or eliminate her profits from the Nikkei call option. The investor can buy a quanto call on the Nikkei, with the yen payoff being converted into dollars at a fixed (usually the spot) exchange rate.

Originally, the term quanto comes from the word "quantity", meaning that the amount that is re-exchanged to the home currency is unknown, because it depends on the future payoff of the option. Therefore the financial institution that sells a quanto call, does not know two things:

1. How deep will the call be in the money at option maturity, ie, which yen amount has to be converted into dollars?
2. What is the exchange rate at option maturity at which the stochastic yen payoff will be converted into dollars?

The correlation between (1) and (2), ie, the price of the underlying S′ and the exchange rate X, significantly influences the quanto call option price. Let's consider a call on the Nikkei S′ and an exchange rate X defined as domestic currency per unit foreign currency (so USD/1 yen for a domestic American) at maturity.

If the correlation is positive, an increasing Nikkei will also mean an increasing yen. That is in favour of the call seller. They have to

settle the payoff, but need only a small yen amount to achieve the dollar payment. Therefore, the more positive the correlation coefficient, the lower is the price for the quanto option. If the correlation coefficient is negative, the opposite applies: if the Nikkei increases, the yen decreases in value. Therefore more yen are needed to meet the dollar payment. As a consequence, the lower the correlation coefficient, the more expensive is the quanto option. Hence we have a similar negative relationship between the option price and correlation, as in Figure 8.4.

Quanto options can be conveniently priced analytically with an extension of the Black–Scholes–Merton model (1973). For a pricing model and a more detailed discussion on a quanto option, see www.dersoft.com/quanto.xls.

The correlation between assets can also be traded directly with a correlation swap. In a correlation swap, a fixed (ie, known) correlation is exchanged with the correlation that will actually occur, called realised or stochastic (ie, unknown) correlation, as seen in Figure 8.5.

Paying a fixed rate in a correlation swap is also called "buying correlation". This is because the present value of the correlation swap will increase for the correlation buyer if the realised correlation increases. Naturally the fixed-rate receiver is "selling correlation".

The realised correlation ρ in Figure 8.5 is the correlation between the assets that actually occur during the time of the swap. It is calculated as:

$$\rho_{realised} = \frac{2}{n^2 - n}\sum_{i>j}\rho_{i,j} \tag{8.6}$$

where $\rho_{i,j}$ is the Pearson correlation between asset i and j, and n is the number of assets in the portfolio. The payoff of a correlation swap for the correlation fixed rate payer at maturity is:

$$N(\rho_{realised} - \rho_{fixed}) \tag{8.7}$$

where N is the notional amount. Let's look at an example of a correlation swap.

Example 8.1

What is the payoff of a correlation swap with three assets, a fixed rate of 10%, a notional amount of USD 1,000,000 and a 1-year maturity?

First, the daily log-returns $\ln(S_t/S_{t-1})$ of the three assets are calculated for one year.[3] Let's assume the realised pairwise correlations of the log-returns at maturity are as displayed in Table 8.2.

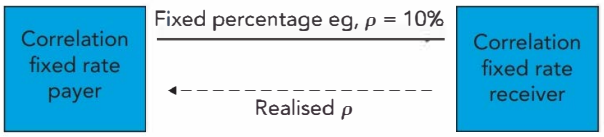


Figure 8.5 **A correlation swap with a fixed 10% correlation rate.**

The average correlation between the three assets is derived by equation (8.6). We only apply the correlations in the shaded area from Table 8.2, since these satisfy $i > j$. Hence we have

$$\rho_{realised} = \frac{2}{3^2 - 3}(0.5 + 0.3 + 0.1) = 0.3$$

Following equation (8.7), the payoff for the correlation fixed-rate payer at swap maturity is USD 1,000,000 X (0.3 − 0.1) = USD 200,000.

Correlation swaps can indirectly protect against decreasing stock prices. As we will see in this chapter in "How does correlation risk fit into the broader picture of risks in finance?", Figure 8.8, as well as in Chapter 9, when stock prices decrease, typically the correlation between the stocks increases. Hence a fixed correlation payer protects themselves indirectly against a stock market decline.

At the time of writing there is no industry-standard valuation model for correlation swaps. Traders often use historical data to anticipate $\rho_{realised}$. To apply swap valuation techniques, we require a term structure of correlation in time. However, no correlation term structure currently exists. We can also apply stochastic correlation models to value a correlation swap. Stochastic correlation models are currently emerging.

Another way of buying correlation (ie, benefiting from an increase in correlation) is to buy put options on an index such as the Dow Jones Industrial Average (Dow) and sell put options on individual stock of the Dow. As we will see in Chapter 9, there is a positive relationship between correlation and volatility.

Table 8.2 **Pairwise Pearson Correlation Coefficient at Swap Maturity**

	$S_{j=1}$	$S_{j=2}$	$S_{j=3}$
$S_{i=1}$	1	0.5	0.1
$S_{i=2}$	0.5	1	0.3
$S_{i=3}$	0.1	0.3	1

[3] Log-returns $\ln(S_1/S_0)$ are an approximation of percentage returns $(S_1 - S_0)/S_0$. We typically use log-returns in finance since they are additive in time, whereas percentage returns are not. For details see Appendix A2.

Therefore, if the correlation between the stocks of the Dow increases, for example in a market downturn, so will the implied volatility[4] of the put on the Dow. This increase is expected to outperform the potential loss from the increase in the short put positions on the individual stocks.

Creating exposure on an index and hedging with exposure on individual components is exactly what the "London whale", JP Morgan's London trader Bruno Iksil, did in 2012. Iksil was called the London whale because of his enormous positions in CDSs.[5] He had sold CDSs on an index of bonds, the CDX.NA.IG.9, and "hedged" it with buying CDSs on individual bonds. In a recovering economy this is a promising trade: volatility and correlation typically decrease in a recovering economy. Therefore, the sold CDSs on the index should outperform (decrease more than) the losses on the CDSs of the individual bonds.

But what can be a good trade in the medium and long terms can be disastrous in the short term. The positions of the London whale were so large, that hedge funds "short squeezed" Iksil: they started to aggressively buy the CDS index CDX.NA.IG.9. This increased the CDS values in the index and created a huge (paper) loss for the whale. JP Morgan was forced to buy back the CDS index positions at a loss of over USD 2 billion.

Risk Management and Correlation

Since the global financial crises of 2007 to 2009, financial markets have become more risk-averse. Commercial banks and investment banks as well as nonfinancial institutions have increased their risk-management efforts. As in the investment and trading environment, correlation plays a vital part in risk management. Let's first clarify what risk management means in finance.

Definition: Financial risk management is the process of identifying, quantifying and, if desired, reducing financial risk.

The main types of financial risk are:

1. market risk;
2. credit risk; and
3. operational risk.

Additional types of risk may include systemic risk, liquidity risk, volatility risk and correlation risk. We will concentrate in this chapter on market risk. Market risk consists of four types of risk: (1) equity risk, (2) interest-rate risk, (3) currency risk and (4) commodity risk.

There are several concepts to measure the market risk of a portfolio such as VaR, expected shortfall (ES), enterprise risk management (ERM) and more. VaR is currently (year 2018) the most widely applied risk-management measure. Let's show the impact of asset correlation on VaR.[6]

First, what is value-at-risk (VaR)? VaR measures the maximum loss of a portfolio with respect to market risk for a certain probability level and for a certain time frame. The equation for VaR is:

$$VaR_P = \sigma_P \alpha \sqrt{x} \tag{8.8}$$

where VaR_P is the value-at-risk for portfolio P, and

α: Abscise value of a standard normal distribution, corresponding to a certain confidence level. It can be derived as = normsinv(confidence level) in Excel or norminv(confidence level) in MATLAB; α takes the values $-\infty < \alpha < +\infty$;

x: Time horizon for the VaR, typically measured in days;

σ_P: Volatility of the portfolio P, which includes the correlation between the assets in the portfolio. We calculate σ_P via:

$$\sigma_P = \sqrt{\beta_h C \beta_v} \tag{8.9}$$

where β_h is the horizontal β vector of invested amounts (price time quantity); β_v is the vertical β vector of invested amounts (also price time quantity);[7] C is the covariance matrix of the returns of the assets.

Let's calculate VaR for a two-asset portfolio and then analyse the impact of different correlations between the two assets on VaR.

Example 8.2

What is the 10-day VaR for a two-asset portfolio with a correlation coefficient of 0.7, daily standard deviation of returns of asset 1 of 2%, asset 2 of 1%, and USD 10 million invested in asset 1 and USD 5 million invested in asset 2, on a 99% confidence level?

[4] Implied volatility is volatility derived (implied) by option prices. The higher the implied volatility, the higher the option price.

[5] Simply put, a CDS is an insurance against default of an underlying (eg, a bond). However, if the underlying is not owned, a long CDS is a speculative instrument on the default of the underlying (just like a naked put on a stock is a speculative position on the stock going down). See Meissner (2005) for more.

[6] We will use a "variance-covariance VaR" approach in this book to derive VaR. Another way to derive VaR is the "non-parametric VaR". This approach derives VaR from simulated historical data. See Markovich (2007) for details.

[7] More mathematically, the vector β_h is the transpose of the vector β_v and vice versa: ${\beta_h}^T = \beta_v$ and ${\beta_v}^T = \beta_h$. Hence we can also write Equation (8.9) as $\sigma_P = \sqrt{\beta_h C \beta_v^T}$. See www.dersoft.com/matrixprimer.xlsx sheet "Matrix Transpose" for more.

First, we derive the covariances Cov:

$$\text{Cov}_{11} = \rho_{11}\sigma_1\sigma_1 = 1 \times 0.02 \times 0.02 = 0.0004^{8} \quad (8.10)$$

$$\text{Cov}_{12} = \rho_{12}\sigma_1\sigma_2 = 0.7 \times 0.02 \times 0.01 = 0.00014$$

$$\text{Cov}_{21} = \rho_{21}\sigma_2\sigma_1 = 0.7 \times 0.01 \times 0.02 = 0.00014$$

$$\text{Cov}_{22} = \rho_{22}\sigma_2\sigma_2 = 1 \times 0.01 \times 0.01 = 0.0001$$

Hence our covariance matrix is

$$C = \begin{pmatrix} 0.0004 & 0.00014 \\ 0.00014 & 0.0001 \end{pmatrix}$$

Let's calculate σ_P following equation (8.9). We first derive $\beta_h C$

$$(10 \;\; 5) \begin{pmatrix} 0.0004 & 0.00014 \\ 0.00014 & 0.0001 \end{pmatrix}$$

$$= (10 \times 0.0004 + 5 \times 0.00014 \;\; 10 \times 0.00014 + 5 \times 0.0001)$$
$$= (0.0047 \quad 0.0019)$$

and then

$$(\beta_h C)\beta_v = (0.0047 \quad 0.0019)\begin{pmatrix} 10 \\ 5 \end{pmatrix}$$
$$= 10 \times 0.0047 + 5 \times 0.0019 = 5.65\%$$

Hence we have

$$\sigma_P = \sqrt{\beta_h C \beta_v} = \sqrt{5.65\%} = 23.77\%$$

We find the value for α in equation (8.8) from Excel as =normsinv(0.99) = 2.3264, or MATLAB as norminv(0.99) = 2.3264.

Following equation (8.8), we now calculate the VaR_P as $0.2377 \times 2.3264 \times \sqrt{10} = 1.7486$.[9]

Interpretation: We are 99% certain that we will not lose more than USD 1.7486 million in the next 10 days due to correlated market price changes of asset 1 and 2.

The number USD 1.7486 million is the 10-day VaR on a 99% confidence level. This means that on average once in a hundred 10-day periods (so once every 1,000 days), this VaR number of USD 1.7486 million will be exceeded. If we have roughly 250 trading days in a year, the company is expected to exceed the VaR about once every four years.

[8] The attentive reader realises that we calculated the covariance differently in Equation (8.3). In Equation (8.3) we derived the covariance "from scratch", inputting the return values and means. In Equation (8.10) we are assuming that we already know the correlation coefficient ρ and the standard deviation σ.

[9] This calculation, including Excel matrix multiplication, can be found at www.dersoft.com/2assetVaR.xlsx.

Let's now analyse the impact of different correlations between the asset 1 and asset 2 on VaR. Figure 8.6 shows the impact.

As expected, we observe from Figure 8.6 that the lower the correlation, the lower is the risk, measured by VaR. Preferably the correlation is negative. In this case, if one asset decreases, the other asset on average increases, hence reducing the overall risk. The impact of correlation on VaR is strong: for a perfect negative correlation of −1, VaR is USD 1.1 million; for a perfect positive correlation, VaR is close to USD 1.9 million. A spreadsheet for calculating two-asset VaRs can be found at www.dersoft.com/2assetVaR.xlsx (case-sensitive).

"There are no toxic assets, just toxic people."

The Global Financial Crises 2007 to 2009 and Correlation

Currently, in 2018, the global financial crisis of 2007 to 2009 seems like a distant memory. The Dow Jones Industrial Average has recovered from its low in March 2009 of 6,547 points and has almost quadrupled to over 25,000 as of October 2018. World economic growth is at a moderate 2.5%. The US unemployment rate as of October 2018 is historically low at 3.7%. However, to fight the crisis, governments engaged in huge stimulus packages to revive their faltering economies. As a result, enormous sovereign deficits are plaguing the world economy. The US debt is also far from benign with a total gross-debt-to-GDP ratio of about 107%. One of the few nations that are enjoying these enormous debt levels is China, which is happy buying the debt and taking in the proceeds.

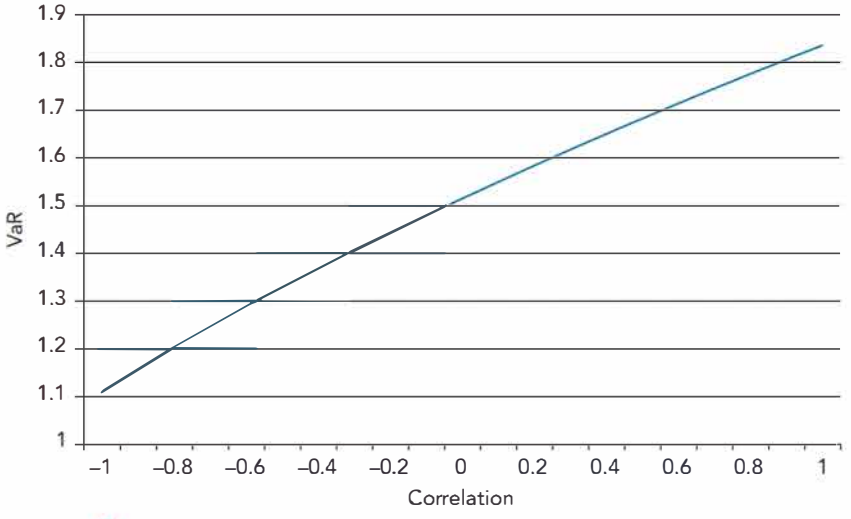


Figure 8.6 **VaR of the two-asset portfolio of Example 8.2 with respect to correlation ρ.**

A crisis that brought the financial and economic system worldwide to a standstill is naturally not mono-causal, but has many reasons. Here are the main ones.

(a) An extremely benign economic and risk environment from 2003 to 2006 with record low credit spreads, low volatility and low interest rates.

(b) Increasing risk-taking and speculation of traders and investors who tried to benefit in these presumably calm times. This led to a bubble in virtually every market segment like the housing market, the mortgage market (especially the subprime mortgage market), the stock market and the commodity market. In 2007, US investors had borrowed 470% of the US national income to invest and speculate in the real-estate, financial and commodity markets.

(c) A new class of structured investment products such as CDOs, CDO squared, CPDOs (constant-proportion debt obligations) and CPPI (constant proportion portfolio insurance), as well as new products such as options on CDSs, credit indices etc.

(d) The new copula correlation model, which was trusted naively by many investors and which could presumably correlate the $n(n-1)/2$ assets in a structured product. Most CDOs contained 125 assets. Hence there are $125(125-1)/2 = 7{,}750$ asset correlation pairs to be quantified and managed.

(e) A moral hazard of rating agencies, who were paid by the same companies whose assets they rated. As a consequence, many structured products received AAA ratings and gave the illusion of low price and default risk.

(f) Risk managers and regulators who lowered their standards in light of the greed and profit frenzy. We recommend an excellent – anonymous – paper in *The Economist*: "Confessions of a Risk Manager".

The topic of this book is correlation risk, so let's concentrate on the correlation aspect of the crisis. Around 2003, two years after the Internet bubble burst, the risk appetite of the financial markets increased and investment banks, hedge funds, and private investors began to speculate and invest in the stock markets, commodities and especially in the real-estate market.

In particular, residential mortgages became an investment object. The mortgages were packaged in CDOs and then sold off to investors locally and globally. The CDOs typically consist of several tranches, ie, the investor can choose a particular degree of default risk. The equity tranche holder is exposed to the first 3% of mortgage defaults, the mezzanine tranche holder is exposed to the 3–7% of defaults and so on. The new copula correlation model, derived by Abe Sklar in 1959 and transferred to finance by David Li in 2000, could presumably manage the default correlations in the CDOs.

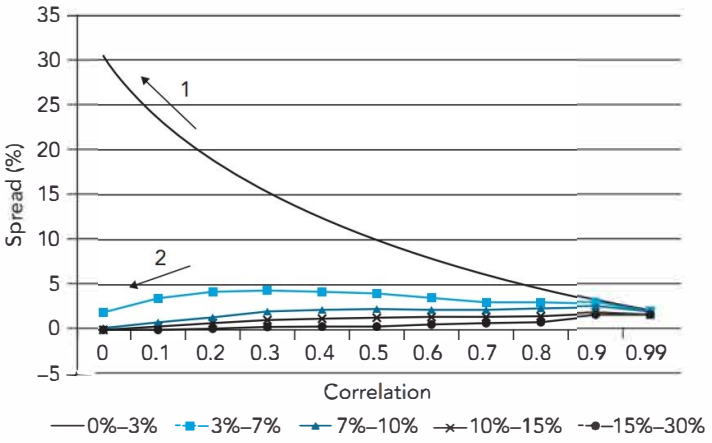


Figure 8.7 **CDO tranche spreads with respect to correlation between the assets in the CDO.**

A first correlation-related crisis, which was a forerunner of the major one to come in 2007 to 2009, occurred in May 2005. General Motors was downgraded to BB and Ford was downgraded to BB+, so both companies were now in "junk status". A downgrade to junk typically leads to a sharp bond price decline, since many mutual funds and pension funds are not allowed to hold junk bonds.

Importantly, the correlation of the bonds in CDOs (which originally were only investment-grade bonds) decreased, since bonds of different credit qualities are typically lower-correlated. This led to huge losses of hedge funds, which had put on a strategy where they were long the equity tranche of the CDO and short the mezzanine tranche of the CDO. Figure 8.7 shows the dilemma. Hedge funds had invested in the equity tranche[10] (0% to 3% in Figure 8.7) to collect the high-equity tranche spread. They had then presumably hedged[11] the risk by going short the mezzanine tranche[12] (3% to 7% in Figure 8.7). However, as we can see from Figure 8.7, this "hedge" is flawed.

When the correlations between the assets in the CDO decreased, the hedge funds lost on both positions.

[10] Investing in the equity tranche means "assuming credit risk" since a credit deterioration hurts the investor. This is similar to a bond, where the investor assumes the credit risk. Investors in the equity tranche receive the high equity tranche contract spread.

[11] To hedge means to protect or to reduce risk.

[12] Going short the mezzanine tranche means being "short credit", ie, benefiting from a credit deterioration. Going short the mezzanine tranche means paying the (fairly low) mezzanine tranche contract spread.

1. The equity tranche spread increased sharply (see Arrow 1). Hence the spread that the hedge fund received in the original transaction was now significantly lower than the current market spread, resulting in a paper loss.
2. In addition, the hedge funds lost on their short mezzanine tranche position, since a lower correlation lowers the mezzanine tranche spread (see Arrow 2). Hence the spread that the hedge fund paid in the original transactions was now higher than the market spread, resulting in another paper loss.

As a result of the huge losses, several hedge funds such as Marin Capital, Aman Capital and Baily Coates Cromwell filed for bankruptcy. It is important to point out that the losses resulted from a lack of understanding of the correlation properties of the tranches in the CDO. The CDOs themselves can hardly be blamed or called toxic for their correlation properties.

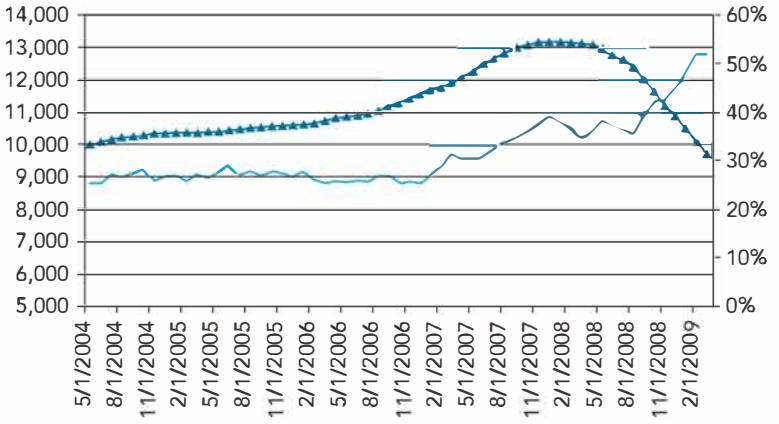


Figure 8.8 **Relationship between the Dow (graph with triangles, numerical values on left axis) and correlation between the stocks in the Dow (numerical values on right axis) before and during the systemic 2007–09 global financial crisis; one-year moving average of monthly correlations.**

From 2003 to 2006 the CDO market, mainly referencing residential mortgages, had exploded and increased from USD 64 billion to USD 455 billion. To fuel the CDOs, more and more questionable subprime mortgages were given, named NINJA loans, standing for "no income, no job or assets". When housing prices started levelling off in 2006, the first mortgages started to default. In 2007 more and more mortgages defaulted, finally leading to a real-estate market collapse. With it the huge CDO market collapsed, leading to the stock market and commodity market crash and a freeze in the credit markets. The financial crisis spread to the world economies, creating a global severe recession now called the "great recession".

In a systemic crash like this, naturally many types of correlations increase; see also Figure 8.8. From 2007 to 2009, default correlations between the mortgages in the CDOs increased. This actually helped equity tranche investors, as we can see from Figure 8.7. If default correlations between the assets in the CDO increase, the equity tranche spread decreases, leading to an increase in the value of the equity tranche. However, this increase was overcompensated by a strong increase in default probability of the mortgages; as a consequence, tranche spreads increased sharply, resulting in a huge loss of the equity tranche investors as well investors in the other tranches.

Correlations between the tranches of the CDOs also increased during the crisis. This had a devastating effect on the super-senior tranches. In normal times, these tranches were considered extremely safe since (a) there were AAA-rated and (b) they were protected by the lower tranches. But, with the increased tranche correlation and the generally deteriorating credit market, these super-senior tranches were suddenly considered risky and lost up to 20% of their value.

To make things worse, many investors had leveraged the super-senior tranches, termed LSS (leveraged super-senior tranche) to receive a higher spread. This leverage was typically 10 or 20 times, meaning an investor paid USD 10,000,000 but had risk exposure of USD 100,000,000 or USD 200,000,000. What made things technically even worse, was that these LSSs came with an option for the investors to unwind the super-senior tranche if the spread had widened (increased). So many investors started to sell the LSS at low prices, realising a loss and increasing the LSS tranche spread even further.

In addition to the overinvestment in CDOs, the CDS market also exploded from its beginnings in the mid-1990s from about USD 8 trillion in 2004 to almost USD 60 trillion in 2007. CDSs are typically used as insurance to protect against default of a debtor, as we discussed in Figure 8.1. No one will argue that an insurance contract is toxic. On the contrary, it is the principle of insurance to spread the risk to a wider audience and hence reduce individual risk, as we can see from health insurance or life insurance contracts.

CDSs, though, can also be used as a speculative instrument. For example, the CDS seller (ie, the insurance seller) hopes that the insured event (eg, default or credit deterioration of the company) will not occur. In this case, the CDS seller keeps the CDS spread (ie, the insurance premium), as income as AIG tried to do in the crisis. A CDS buyer, when they do not own the underlying asset, speculates on the credit deterioration of the underlying asset, just like a naked put option holder speculates on the decline of the underlying asset.

So who can we blame for the 2007--09 global financial crises? The quants, who created the new products such as CDSs and CDOs

and the models to value them? The upper management and the traders, who authorised and conducted the overinvesting and extreme risk-taking? The rating agencies, who gave an AAA rating to many CDOs? The regulators, who approved the overinvestments? The risk managers, who allowed the excessive risk taking?

The answer is: All of them. The whole global financial crisis can be summed up in one word: greed! It was the upper management, the traders and investors who engaged in excessive trading and irresponsible risk taking to receive high returns, huge salaries and generous bonuses. And most risk managers and regulators turned a blind eye.

For example, the London unit of the insurance company AIG had sold close to USD 500 billion in CDSs without much reinsurance! Their main hedging strategy seemed to have been: pray that the insured contracts don't deteriorate. The investment banks of Iceland, a small country in Northern Europe, had borrowed 10 times Iceland's national GDP and invested it. With this leverage, Iceland naturally went *de facto* into bankruptcy in 2008, when the credit markets deteriorated. Lehman Brothers, before filing for bankruptcy in September 2008, reported a leverage of 30.7, ie, USD 691 billion in assets and only USD 22 billion in stockholders' equity. The true leverage was even higher, since Lehman tried to hide their leverage with materially misleading repo transactions.[13] In addition, Lehman had 1.5 million derivatives transactions with 8,000 different counterparties on their books.

Did the upper management and traders of hedge funds and investment banks admit to their irresponsible leverage, excessive trading and risk taking? No. Instead they created the myth of the "toxic asset", which is absurd. It is like a murderer saying: "I did not shoot that person – it was my gun!" Toxic are not the financial products, but humans and their greed.

Most traders were well aware of the risks that they were taking. In the few cases where traders did not understand the risks, the asset itself cannot be blamed, rather the incompetence of the trader is the reason for the loss. While it is ethically disappointing that the investors and traders did not admit to their wrongdoing, at the same time it is understandable. If they would admit to irresponsible trading and risk taking, they would immediately be prosecuted.

Naturally risk managers and regulators have to take part of the blame to allow the irresponsible risk taking. The moral hazard of the rating agencies, being paid by the same companies whose assets they rate, needs to also be addressed.

[13] Repo stands for repurchase transaction. It can be viewed as a short-term collateralised loan.

Regulation and Correlation

Correlations are critical inputs in regulatory frameworks such as the Basel accords, especially in regulations for market risk and credit risk. We will discuss the correlation approaches of the Basel accords in this book. First, let's clarify.

What are Basel I, II and III?

Basel I, implemented in 1988, Basel II, implemented in 2006, and Basel III, which is currently being developed and implemented until 2019, are regulatory guidelines to ensure the stability of the banking system.

The term Basel comes from the beautiful city of Basel in Switzerland, where the honourable regulators meet. None of the Basel accords has legal authority. However, most countries (about 100 for Basel II) have created legislation to enforce the Basel accords for their banks.

Why Basel I, II and III?

The objective of the Basel accords is to provide incentives for banks to enhance their risk measurement and management systems and to contribute to a higher level of safety and soundness in the banking system. In particular, Basel III addresses the deficiencies of the banking system during the financial crisis 2007 to 2009. Basel III introduces many new ratios to ensure liquidity and adequate leverage of banks. In addition, new correlation models are implemented that deal with double defaults in insured risk transactions as displayed in Figure 8.1. Correlated defaults in a multi-asset portfolio quantified with the Gaussian copula, correlations in derivatives transactions termed credit value adjustment (CVA) and correlations in what is called "wrong-way risk" (WWR) have been proposed.

8.6 HOW DOES CORRELATION RISK FIT INTO THE BROADER PICTURE OF RISKS IN FINANCE?

As already mentioned, we differentiate three main types of risks in finance: market risk, credit risk and operational risk. Additional types of risk may include systemic risk, concentration risk, liquidity risk, volatility risk, legal risk, reputational risk and more. Correlation risk plays an important part in market risk and credit risk and is closely related to systemic risk and concentration risk. Let's discuss it.

Correlation Risk and Market Risk

Correlation risk is an integral part of market risk. Market risk comprises equity risk, interest-rate risk, currency risk and commodity risk. Market risk is typically measured with the VaR concept. Since VaR has a covariance matrix of the assets in the portfolio as an input, VaR implicitly incorporates correlation risk, ie, the risk that the correlations in the covariance matrix change. We have already studied the impact of different correlations on VaR in "Risk management and correlation" above.

Market risk is also quantified with expected shortfall (ES), also termed "conditional VaR" or "tail risk". Expected shortfall measures market risk for extreme events, typically for the worst 0.1%, 1% or 5% of possible future scenarios. A rigorous valuation of expected shortfall naturally includes the correlation between the asset returns in the portfolio, as VaR does.[14]

Correlation Risk and Credit Risk

Correlation risk is also a critical part of credit risk. Credit risk comprises (a) migration risk and (b) default risk. Migration risk is the risk that the credit quality of a debtor decreases, ie, migrates to a lower credit state. A lower credit state typically results in a lower asset price, so a paper loss for the creditor occurs. We already studied the effect of correlation risk of an investor, who has hedged their bond exposure with a CDS earlier in the section titled, "What is financial correlation risk?". We derived that the investor is exposed to changes in the correlation between the reference asset and the counterparty, ie, the CDS seller. The higher the default correlation, the higher is the CDS paper loss for the investor and, importantly, the higher is the probability of a total loss of their investment.

The degree to which defaults occur together (ie, default correlation) is critical for financial lenders such as commercial banks, credit unions, mortgage lenders and trusts, which give many types of loans to companies and individuals. Default correlations are also critical for insurance companies, which are exposed to credit risk of numerous debtors. Naturally, a low default correlation of debtors is desired to diversify the credit risk. Table 8.3 shows the default correlation from 1981 to 2001 of 6,907 companies, of which 674 defaulted.

The default correlations in Table 8.3 are one-year default correlations averaged over the time period 1981 to 2001. For example, the number 3.8% in the upper left corner means that, if a certain bond in the auto industry defaulted, there is a 3.8% probability that another bond in the auto industry will default. The number −2.5% in the column named "Fin" in the fourth row means that, if a bond in the energy sector defaulted, this actually decreases the probability that a bond in the financial sector defaults by 2.5% and vice versa.

From Table 8.3 we also observe that default correlations between industries are mostly positive, with the exception of the energy sector. This sector is typically viewed as a recession-resistant, stable sector with no or low correlation to other sectors. We also observe that the default correlation within sectors is higher than between sectors. This suggests that systematic factors (such as a recession or structural weakness as the general decline of a sector) impact on defaults more than idiosyncratic factors. Hence if General Motors defaults, it is more likely that Ford defaults, rather than Ford benefiting from the default of its rival GM.

Since the intra-sector default correlations are higher than inter-sector default correlations, a lender is advised to have a sector-diversified loan portfolio to reduce default correlation risk.

Defaults are binomial events: either default or no default. Therefore, to model defaults, often a simple binomial model is applied. However, we can also analyse defaults in more detail and look at term structure of defaults. Let's assume a creditor has given loans to two debtors. One debtor is A-rated and one is CC-rated. A historical default term structure of these bonds is displayed in Table 8.4.

To clarify, the number 0.15% in the column corresponding to the fifth year and second row means that an A-rated bond has a 0.15% probability to default in year 5. For most investment-grade bonds, the term structure of default probabilities increases in time, as we see from Table 8.4 for the A-rated bond. This is because the longer the time horizon, the higher the probability of adverse internal events as mismanagement, or external events as increased competition or a recession. For bonds in distress, however, the default term structure is typically inverse, as seen for the CC-rated bond in Table 8.4. This is because for a distressed company, the immediate future is critical. If the company survives the coming problematic years, the probability of default decreases.

For a creditor, the default correlation of his debtors is critical. As mentioned, a creditor will benefit from a low default correlation of their debtors, which spreads the default correlation risk. We can correlate the default term structures in Table 8.4 with the famous (now infamous) copula model. This will allow us to

[14] See the original ES paper by Artzner (1997), an educational paper by Yamai and Yoshia (2002), as well as Acerbi and Tasche (2001), and McNeil *et al* (2005).

Table 8.3 Default Correlation of 674 Defaulted Companies by Industry

	Auto	Cons	Ener	Fin	Build	Chem	HiTec	Insur	Leis	Tele	Trans	Util
Auto	3.8%	1.3%	1.2%	0.4%	1.1%	1.6%	2.8%	−0.5%	1.0%	3.9%	1.3%	0.5%
Cons	1.3%	2.8%	−1.4%	1.2%	2.8%	1.6%	1.8%	1.1%	1.3%	3.2%	2.7%	1.9%
Ener	1.2%	−1.4%	**6.4%**	−2.5%	−0.5%	0.4%	−0.1%	−1.6%	−1.0%	−1.4%	−0.1%	0.7%
Fin	0.4%	1.2%	−2.5%	**5.2%**	2.6%	0.1%	0.4%	3.0%	1.6%	3.7%	1.5%	4.5%
Build	1.1%	2.8%	−0.5%	2.6%	**6.1%**	1.2%	2.3%	1.8%	2.3%	**6.5%**	4.2%	1.3%
Chem	1.6%	1.6%	0.4%	0.1%	1.2%	3.2%	1.4%	−1.1%	1.1%	2.8%	1.1%	1.0%
HiTec	2.8%	1.8%	−0.1%	0.4%	2.3%	1.4%	3.3%	0.0%	1.4%	4.7%	1.9%	1.0%
Insur	−0.5%	1.1%	−1.6%	3.0%	1.8%	−1.1%	0.0%	**5.6%**	1.2%	−2.6%	2.3%	1.4%
Leis	1.0%	1.3%	−1.0%	1.6%	2.3%	1.1%	1.4%	1.2%	2.3%	4.0%	2.3%	0.6%
Tele	3.9%	3.2%	−1.4%	3.7%	**6.5%**	2.8%	4.7%	−2.6%	4.0%	**10.7%**	3.2%	0.8%
Trans	1.3%	2.7%	−0.1%	1.5%	4.2%	1.1%	1.9%	2.3%	2.3%	3.2%	4.3%	0.2%
Util	0.5%	1.9%	0.7%	4.5%	1.3%	1.0%	1.0%	1.4%	0.6%	−0.8%	−0.2%	**9.4%**

Correlations above 5% are in bold.

Note: One year US default correlations – non – investment grade bonds 1981–2001.

Table 8.4 Term Structure of Default Probabilities for an A-rated Bond and a CC-Rated Bond in 2002

Year										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0.02%	0.07%	0.13%	0.14%	0.15%	0.17%	0.18%	0.21%	0.24%	0.25%
CC	23.83%	13.29%	10.31%	7.62%	5.04%	5.13%	4.04%	4.62%	2.62%	2.04%

Source: Moody's.

answer questions as "What is the joint probability of Debtor 1 defaulting in Year 3 and Debtor 2 defaulting in Year 5?"

"Correlations always increase in stressed markets"

John Hull

8.7 CORRELATION RISK AND SYSTEMIC RISK

So far, we have analysed correlation risk with respect to market risk and credit risk and have concluded that correlations are a critical input when quantifying market risk and credit risk. Correlations are also closely related to systemic risk, which we define as the risk that a financial market or an entire financial system collapses.

An example of systemic risk is the collapse of the entire credit market in 2008. At the height of the crisis in September 2008, when Lehman Brothers filed for bankruptcy, the credit markets were virtually frozen with essentially no lending activities. Even as the Federal Reserve guaranteed interbank loans, lending resumed only very gradually and slowly.

The stock market crash starting in October 2007, with the Dow (Dow Jones Industrial Average) at 14,093 points and then falling by 53.54% to 6,547 points by March 2009, is also a systemic market collapse. All but one of the Dow 30 stocks had declined. Walmart was the lone stock, which was up during the crisis. Of the S&P 500 stocks, 489 declined during this timeframe. The 11 stocks that were up were:

- Apollo Group (APOL), educational sector; provides educational programmes for working adults and is a subsidiary of the University of Phoenix;
- Autozone (AZ0), auto industry; provides auto replacement parts;
- CF Industries (CF), agricultural industry; provides fertiliser;
- DeVry Inc. (DV), educational sector; holding company of several universities;
- Edward Lifesciences (EW), pharmaceutical-industry; provides products to treat cardiovascular diseases;
- Family Dollar (FDO), consumer staples;
- Gilead Pharmaceuticals (GILD), pharmaceutical industry; provides HIV, hepatitis medication;
- Netflix (NFLX), entertainment industry; provides Internet subscription service;
- Ross Stores (ROST), consumer staples;
- Southwestern Energy (SWN), energy sector; and
- Walmart (WMT), consumer staples.

From this list we can see that the consumer staples sector (which provides basic necessities as food and basic household items) fared well during the crisis. The educational sector also typically thrives in a crisis, since many unemployed seek to further their education.

Importantly, systemic financial failures such as the one from 2007 to 2009 typically spread to the economy with a decreasing GDP, increasing unemployment and, therefore, a decrease in the standard of living.

Systemic risk and correlation risk are highly dependent. Since a systemic decline in stocks involves almost the entire stock market, correlations between the stocks increase sharply. Figure 8.8 shows the relationship between the percentage change of the Dow and the correlation between the stocks in the Dow before the crisis from May 2004 to October 2007 and during the crisis from October 2007 to March 2009.

In Figure 8.8 we downloaded daily closing prices of all 30 stocks in the Dow and put them into monthly bins. We then derived monthly 30×30 correlation matrices using the Pearson correlation measure and averaged the matrices. We then smoothed the graph by taking the one-year moving average.

From Figure 8.8 we can observe a somewhat stable correlation from 2004 to 2006, when the Dow increased moderately. In the time period from January 2007 to February 2008 we observe that the correlation in the Dow increases when the Dow increases more strongly. Importantly, in the time of the severe decline of the Dow from August 2008 to March 2009 we observe a sharp increase in the correlation from non-crisis levels of an average 27% to over 50%. In Chapter 9, we will observe empirical correlations in detail and we will find that, at the height of the crisis in February 2009, the correlation of the stocks in the Dow reached a high of 96.97%. Hence, portfolios that were considered well diversified in benign times experienced a sharp increase in correlation and hence unexpected losses due to the combined, highly correlated decline of many stocks during the crisis.

8.8 CORRELATION RISK AND CONCENTRATION RISK

Concentration risk is a fairly new risk category and therefore not yet uniquely defined. A sensible definition is the risk of financial loss due to a concentrated exposure to a specific group of counterparties.

Concentration risk can be quantified with the concentration ratio. For example, if a creditor has 10 loans of equal size, the concentration ratio would be $1/10 = 0.1$. If a creditor has

only one loan to one counterparty, the concentration ratio would be 1. Naturally, the lower the concentration ratio, the more diversified is the default risk of the creditor, assuming the default correlation between the counterparties is smaller than 1.

We can also categorise counterparties into groups – for example, sectors. We can then analyse sector concentration risk. The higher the number of different sectors a creditor has lent to, the higher is their sector diversification. High sector diversification reduces default risk, since intra-sector defaults are higher correlated than counterparties in different sectors, as seen in Table 8.3.

Naturally, concentration and correlation risk are closely related. Let's verify this in an example.

Example 8.3

Case (a) The commercial bank C has lent USD 10,000,000 to a single company W. So C's concentration ratio is 1. Company W has a default probability P_W of 10%. Hence the expected loss (EL) for bank C is USD 10,000,000 × 0.1 = USD 1,000,000. Graphically, we have Figure 8.9.

Case (b) The commercial bank C has lent USD 5,000,000 to company X and USD 5,000,000 to company Y. Both X and Y have a 10% default probability. So C's concentration ratio is reduced to 1/2.

If the default correlation between X and Y is bigger than 0 and smaller than 1, we derive that the worst-case scenario, ie, the default of X and Y, $P(X \cap Y)$, with a loss of USD 1,000,000 is reduced, as seen in Figure 8.10.

The exact joint default probability depends on the correlation model and correlation parameter values. For any model, though, if default correlation between X and Y is 1, then there is no benefit from the lower concentration ratio. The probability space would be as in Figure 8.9.

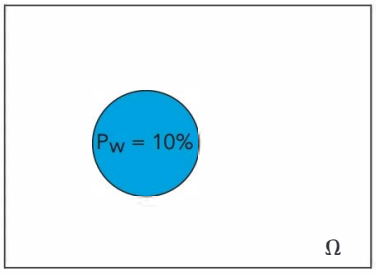


Figure 8.9 Probability space for the default probability of a single loan to W.

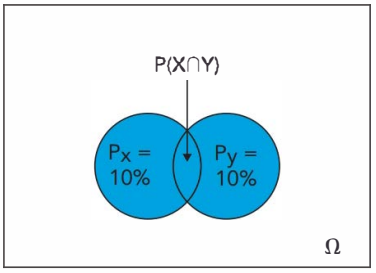


Figure 8.10 Probability space for loans to companies X and Y.

Case (c) If we further decrease the concentration ratio, the worst-case scenario, ie, the expected loss of 10% decreases further. Let's assume the lender C gives loans to three companies X, Y and Z, of USD 3.33 million each. The default probability of X, Y and Z is 10% each. Therefore, the concentration ratio decreases to a third. The probabilities are displayed in Figure 8.11.

Hence, from Figures 8.9 to 8.11 we observe the benefits of a lower concentration ratio. The worst-case scenario, an expected loss of USD 1,000,000, reduces with a decreasing concentration ratio.

A decreasing concentration ratio is closely related to a decreasing correlation coefficient. Let's show this. The defaults of companies *X* and *Y* are expressed as two binomial variables that take the value 1 if in default, and 0 otherwise. Equation (8.11) gives the joint probability of default for the two binomial events:

$$P(X \cap Y) = \rho_{XY}\sqrt{P_X(1 - P_X)\,P_Y(1 - P_Y)} + P_X P_Y \quad (8.11)$$

where ρ_{XY} is the correlation coefficient and

$$\sqrt{P_X(1 - P_X)} \quad (8.12)$$

is the standard deviation of the binomially distributed variable *X*.

Let's assume again that the lender C has given loans to X and Y of USD 5,000,000 each. Both X and Y have a default probability of 10%. Following equation (8.12), this means that the standard deviation for X and Y is $\sqrt{0.1x\,(1 - 0.1)} = 0.3$.

Let's first look at the case where the default correlation is $\rho_{XY} = 1$. This means that X and Y cannot default individually. They can only default together or survive together. The probability that they default together is 10%. Hence the expected loss is the same as in case a) EL = (USD 5,000,000 + USD 5,000,000) × 0.1 = USD 1,000,000. We can verify this with equation (8.11) for the joint probability of two binomial events, $P(X \cap Y) = 1 \times \sqrt{0.1(1 - 0.1) \times 0.1(1 - 0.1)} + 0.1 \times 0.1 = 10\%$. The probability space is graphically the same as Figure 8.9 with $P_X = P_Y = 10\%$ as the probability event.

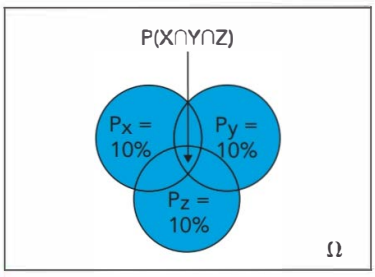


Figure 8.11 **Probability space for loans to companies X, Y and Z.**

If we now decrease the correlation coefficient, we can see from equation (8.11) that the worst-case scenario, the joint default probability of X and Y, $P(X \cap Y)$, will decrease. For example, $\rho_{XY} = 0.5$ results in $P(X \cap Y) = 5.5\%$, $\rho_{XY} = 0$ results in $P(X \cap Y) = 1\%$. Interestingly, even a slightly negative correlation coefficient can result in a positive joint default probability if the standard deviation of the binomial events is fairly low and the default probabilities are high. In our example, the standard deviation of both entities is 30% and a default probability of both entities is 10%. Together with a negative correlation coefficient of −0.1, following equation (8.11) leads to a joint default probability of 0.1%.

In conclusion, we have shown the beneficial aspect of a lower concentration ratio that is closely related to a lower correlation coefficient. In particular, both a lower concentration ratio and a lower correlation coefficient reduce the worst-case scenario for a creditor, the joint probability of default of his debtors.

We will verify this result and find that a higher (copula) correlation between assets results in a higher credit value-at-risk (CVaR). CVaR measures the maximum loss of a portfolio of correlated debt with a certain probability for a certain timeframe. Hence CVaR measures correlated default risk and is analogous to the VaR concept for correlated market risk, which we discussed earlier.

8.9 A WORD ON TERMINOLOGY

As mentioned in the section "Trading and correlation" above, we find the terms "correlation desks" and "correlation trading" in trading practice. Correlation trading means that traders trade assets or execute trading strategies, whose value is at least in part determined by the co-movement of two or more assets in time. We already mentioned the strategy "pairs trading", the exchange option and the quanto option as examples of correlation trading. In trading practice, the term "correlation" is typically applied quite broadly, referring to any co-movement of asset prices in time.

However, in financial theory, especially in recent publications, the term "correlation" is often defined more narrowly, referring only to the linear Pearson correlation model, as in Cherubini et al (2004), Nelsen (2006) and Gregory (2010). These authors refer to other than Pearson correlation coefficients as dependence measures or measures of association. However, in financial theory the term "correlation" is also often applied to generally describe dependencies, as in the terms "credit correlation", "default correlation" and "volatility–asset return correlation", which are quantified by non-Pearson models as in Heston (1993), Lucas (1995) and Li (2000).

In this book, we will refer to the Pearson coefficient as correlation coefficient and the coefficients derived by non-Pearson models as dependency coefficients. In accordance with most literature, we will refer to all methodologies that measure some form of dependency as correlation models or dependency models.

SUMMARY

There are two types of financial correlations: (1) static correlations, which measure how two or more financial assets are associated within a certain time period, for example a year; (2) dynamic financial correlations, which measure how two or more financial assets move together in time.

Correlation risk can be defined as the risk of financial loss due to adverse movements in correlation between two or more variables. These variables can be financial variables such as correlated defaults between two debtors or nonfinancial such as the correlation between political tensions and an exchange rate. Correlation risk can be non-monotonic, meaning that the dependent variable, for example the CDS spread, can increase or decrease when the correlation parameter value increases.

Correlations and correlation risk are critical in many areas in finance such as investments, trading and especially risk management, where different correlations result in very different degrees of risk. Correlations also play a key role in a systemic crisis, where correlations typically increase and can lead to high unexpected losses. As a result, the Basel III accord has introduced several correlation concepts and measures to reduce correlation risk.

Correlation risk can be categorised as its own type of risk. However, correlation parameters and correlation matrices are critical

inputs and hence a part of market risk and credit risk. Market risk and credit risk are highly sensitive to changing correlations. Correlation risk is also closely related to concentration risk, as well as systemic risk, since correlations typically increase in a systemic crisis.

The term "correlation" is not uniquely defined. In trading practice "correlation" is applied quite broadly and refers to the co-movements of assets in time, which may be measured by different correlation concepts. In financial theory, the term "correlation" is often defined more narrowly, referring only to the linear Pearson correlation coefficient. Non-Pearson correlation measures are termed "dependence measures" or "measures of association".

APPENDIX A1

Dependence and Correlation

Dependence

In statistics, two events are considered dependent if the occurrence of one affects the probability of another. Conversely, two events are considered independent if the occurrence of one does not affect the probability of another. Formally, two events A and B are independent if and only if the joint probability equals the product of the individual probabilities:

$$P(A \cap B) = P(A)P(B) \tag{A1}$$

Solving equation (A1) for $P(A)$, we get

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

Following the Kolmogorov definition

$$\frac{P(A \cap B)}{P(B)} \equiv P(A|B)$$

we derive

$$P(A) = \frac{P(A \cap B)}{P(B)} = P(A|B) \tag{A2}$$

where $P(A|B)$ is the conditional probability of A with respect to B. $P(A|B)$ reads "probability of A given B". In equation (A2) the probability of A, $P(A)$, is not affected by B, since $P(A) = P(A|B)$, hence the event A is independent from B.

From equation (A2), we also derive

$$P(B) = \frac{P(A \cap B)}{P(A)} = P(B|A) \tag{A3}$$

Hence from equation (A1) it follows that A is independent from B and B is independent from A.

Example A1: Statistical Independence

The historical default probability of company A, P(A) = 3%, the historical default probability of company B, P(B) = 4%, and the historical joint probability of default is 3% × 4% = 0.12%. In this case P(A) and P(B) are independent. This is because, from equation (A2), we have

$$P(A) = \frac{P(A \cap B)}{P(B)} = P(A|B) = 3\% = \frac{3\% \times 4\%}{4\%} = 3\%$$

Since $P(A) = P(A|B)$, the event A is independent from the event B. Using equation (A3), we can do the same exercise for event B, which is independent from event A.

Correlation

As mentioned in the section on terminology above, the term "correlation" is not uniquely defined. In trading practice, the term "correlation" is used quite broadly, referring to any co-movement of asset prices in time. In statistics, correlation is typically defined more narrowly and typically referred to as the linear dependency derived in the Pearson correlation model. Let's look at the Pearson covariance and relate it to the dependence discussed above.

A covariance measures how strong the linear relationship between two variables is. These variables can be deterministic (which means their outcome is known), as the historical default probabilities in example Al above. For random variables (variables with an unknown outcome such as flipping a coin), the Pearson covariance is derived with expectation values:

$$COV(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) \tag{A4}$$

where $E(X)$ and $E(Y)$ are the expected values of X and Y respectively, also known as the mean. $E(XY)$ is the expected value of the product of the random variables X and Y. The covariance in equation (A4) is not easy to interpret. Therefore, often a normalised covariance, the correlation coefficient is used. The Pearson correlation coefficient $\rho(XY)$ is defined as

$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)} \tag{A5}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y respectively. While the covariance takes value between $-\infty$ and $+\infty$, the correlation coefficient conveniently takes values between -1 and $+1$.

Independence and Uncorrelatedness

From equation (A1) above we find that the condition for independence for two random variables is $E(XY) = E(X)\,E(Y)$. From

equation (A4) we see that $E(XY) = E(X)\,E(Y)$ is equal to a covariance of zero. Therefore, if two variables are independent, their covariance is zero.

Is the reverse also true? Does a zero covariance mean independence? The answer is no. Two variables can have a zero covariance even when they are dependent! Let's show this with an example. For the parabola $Y = X^2$, Y is clearly dependent on X, since Y changes when X changes. However, the correlation of the function $Y = X^2$ derived by equations (A4) or (A5) is zero! This can be shown numerically and algebraically. For a numerical derivation, see the simple spreadsheet www.dersoft.com/dependenceandcorrelation.xlsm, sheet 1. Algebraically, we have from equation (A4):

$$COV(X, Y) = E(XY) - E(X)E(Y)$$

Inputting $Y = X^2$, we derive

$$= E(X\,X^2) - E(X)\,E(X^2)$$
$$= E(X^3) - E(X)\,E(X^2)$$

Let X be a uniform variable bounded in $[-1, +1]$. Then the mean $E(X)$ and $E(X^3)$ are zero and we have

$$= 0 - 0\,E(X^2)$$
$$= 0$$

For a numerical example, see www.dersoft.com/dependenceandcorrelation.xlsm, sheet 2.

In conclusion, the Pearson covariance or correlation coefficient can give values of zero, ie, tells us the variables are uncorrelated, even if the variables are dependent! This is because the Pearson correlation concept measures only linear dependence. It fails to capture nonlinear relationships. This shows the limitation of the Pearson correlation concept for finance, since most financial relationships are nonlinear.

APPENDIX A2

On Percentage and Logarithmic Changes

In finance, growth rates are expressed as relative changes, $(S_t - S_{t-1})/S_{t-1}$, where S_t and S_{t-1} are the prices of an asset at time t and $t-1$, respectively. For example, if $S_t = 110$, and $S_{t-1} = 100$, the relative change is $(110 - 100)/100 = 0.1 = 10\%$.

We often approximate relative changes with the help of the natural logarithm:

$$(S_t - S_{t-1})/S_{t-1} \approx \ln(S_t/S_{t-1}) \qquad \text{(A6)}$$

This is a good approximation for small differences between S_t and S_{t-1}. $\text{Ln}(S_t/S_{t-1})$ is called a log-return. The advantage of using log-returns is that they can be added over time. Relative changes are not additive over time. Let's show this in two examples.

Example 1

A stock price at t_0 is USD 100. From t_0 to t_1, the stock increases by 10%. Hence the stock increases to USD 110. From t_1 to t_2, the stock increases again by 10%. So the stock price increases to USD 110 × 0.1 = USD 121. This increase of 21% higher than adding the percentage increases of 10% + 10% = 20%. Hence percentage changes are not additive over time.

Let's look at the log-returns. The log-return from t_0 to t_1 is $\ln(110/100) = 9.531\%$. From t_1 to t_2 the log-return is $\ln(121/110) = 9.531\%$. When adding these returns, we get 9.531% + 9.531% = 19.062%. This is the same as the log-return from t_0 to t_2, ie, $\ln(121/100) = 19.062\%$. Hence log-returns are additive in time.[15]

Let's now look at another, more extreme example.

Example 2

A stock price in t_0 is USD 100. It moves to USD 200 in t_1 and back to USD 100 in t_2. The percentage change from t_0 to t_1 is (USD 200 − USD 100)/USD 100 = 100%. The percentage change from t_1 to t_2 is (USD 100 − USD 200)/(USD 200) = −50%. Adding the percentage changes, we derive +100% − 50% = +50%, although the stock has not increased from t_0 to t_2! Naturally this type of performance measure is incorrect and not allowed in accounting.

Log-returns give the correct answer: the log-return from t_0 to t_1 is $\ln(200/100) = 69.31\%$. The log-return from t_1 to t_2 is $\ln(100/200) = -69.31\%$. Adding these log-returns in time, we get the correct return of the stock price from t_0 to t_2 of 69.31% − 69.31% = 0%.

These examples are displayed in a simple spreadsheet at www.dersoft.com/logreturns.xlsx.

[15] We could have also solved for the absolute value 121, which matches a logarithmic growth rate of 9.531%: $\ln(x/110) = 9.531\%$, or, $\ln(x) - \ln(110) = 9.531\%$, or, $\ln(x) = \ln(110) + 9.531\%$. Taking the power of e we get, $e^{(\ln(x))} = X = e^{(\ln(110)+0.09531)} = 121$.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

8.1 What two types of financial correlations exist?

8.2 What is "wrong-way correlation risk" or for short "wrong-way risk"?

8.3 Correlations can be non-monotonous. What does this mean?

8.4 Correlations are critical in many areas in finance. Name five.

8.5 High diversification is related with low correlation. Why is this considered one of the few "free lunches" in finance?

8.6 Create a numerical example and show why a lower correlation results in a higher return/risk ratio.

8.7 What is "correlation trading"?

8.8 What is "pairs trading"?

8.9 Name three correlation options, in which a lower correlation results in a higher option price.

8.10 Name one correlation option where a lower correlation results in a lower option price.

8.11 Create a numerical example of a two-asset portfolio and show that lower correlation coefficient leads to a lower VaR number.

8.12 Why do correlations typically increase in a systemic market crash?

8.13 In 2005, a correlation crisis with respect to CDOs occurred that led to the default of several hedge funds. What happened?

8.14 In the global financial crisis 2007–09, many investors in the presumably safe super-senior tranches got hurt. What exactly happened?

8.15 What is the main objective of the Basel III accord?

8.16 The Basel accords have no legal authority. So why do most developed countries implement them?

8.17 How is correlation risk related to market risk and credit risk?

8.18 How is correlation risk related to systemic risk and concentration risk?

8.19 How can we measure the joint probability of occurrence of a binomial event as default or no-default?

8.20 Can it be that two binomial events are negatively correlated but they have a positive probability of joint default?

8.21 What is value-at-risk (VaR) and credit value-at-risk (CVaR)? How are they related?

8.22 Correlation risk is quite broadly defined in trading practice, referring to any co-movement of assets in time. How is the term "correlation" defined in statistics?

8.23 What do the terms "measure of association" and "measure of dependence" refer to in statistics?

9

Empirical Properties of Correlation: How Do Correlations Behave in the Real World?

Learning Objectives

After completing this reading, you should be able to:

- Describe how equity correlations and correlation volatilities behave throughout various economic states.
- Calculate a mean reversion rate using standard regression and calculate the corresponding autocorrelation.
- Identify the best-fit distribution for equity, bond, and default correlations.

Excerpt is Chapter 2 of Correlation Risk Modeling and Management, *2nd Edition, by Gunter Meissner.*

"Anything that relies on correlation, is charlatanism"

— Nassim Taleb

In this chapter we show that, contrary to common beliefs, financial correlations display statistically significant and expected properties.

9.1 HOW DO EQUITY CORRELATIONS BEHAVE IN A RECESSION, NORMAL ECONOMIC PERIOD OR STRONG EXPANSION?

In our study, we observed daily closing prices of the 30 stocks in the Dow Jones Industrial Average (Dow) from January 1972 to July 2017. This resulted in 11,214 daily observations of the Dow stocks and hence 11,214 × 30 = 336,420 closing prices. We built monthly bins and derived 900 correlation values (30 × 30) for each month, applying the Pearson correlation approach. Since we had 534 months in the study, altogether we derived 534 × 900 = 480,600 correlation values.

The composition of the Dow is changing in time, with successful stocks being input into the Dow and unsuccessful stocks being removed. Our study comprises the Dow stocks that represent the Dow at each particular point in time.

Figure 9.1 shows the 534 monthly averaged correlation levels: we created monthly 30 by 30 bins of the Dow stock returns from 1972 to 2017, derived the Pearson correlation between each Dow stock returns, eliminated the unit correlation on the diagonal and averaged the remaining correlation values. We then differentiated the three states: an expansionary period with GDP (gross domestic product) growth rates of 3.5% or higher, a normal economic period with growth rates between 0% and 3.49% and a recession with two consecutive quarters of negative growth rates.

Figure 9.2 shows the volatility of the averaged monthly correlations. For the calculation of volatility, see Chapter 8.

From Figures 9.1 and 9.2, we observe the somewhat erratic behaviour of Dow correlation levels and volatility. However, Table 9.1 reveals some expected results:

From Table 9.1, we observe that correlation levels are lowest in strong economic growth times. The reason may be that in strong growth periods equity prices react primarily to idiosyncratic, not to macroeconomic, factors. In recessions, correlation levels are typically high as shown in Table 9.1. Correlation levels increased sharply in the great recession from 2007 to 2009.

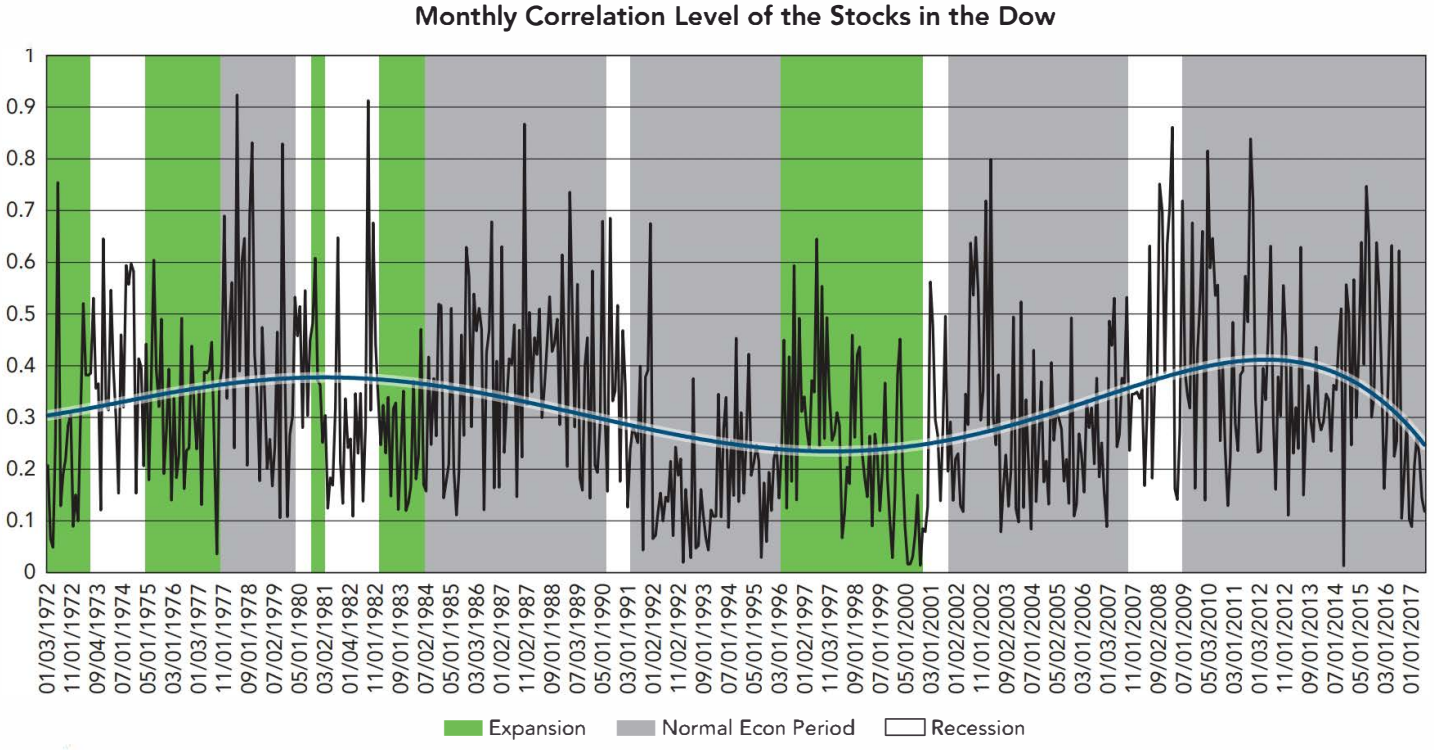


Figure 9.1 **Average correlation of monthly 30 × 30 Dow stock return bins. The light grey background displays an expansionary economic period, the medium gray background a normal economic period and the white background represents a recession. The horizontal line shows the polynomial trendline of order 4.**

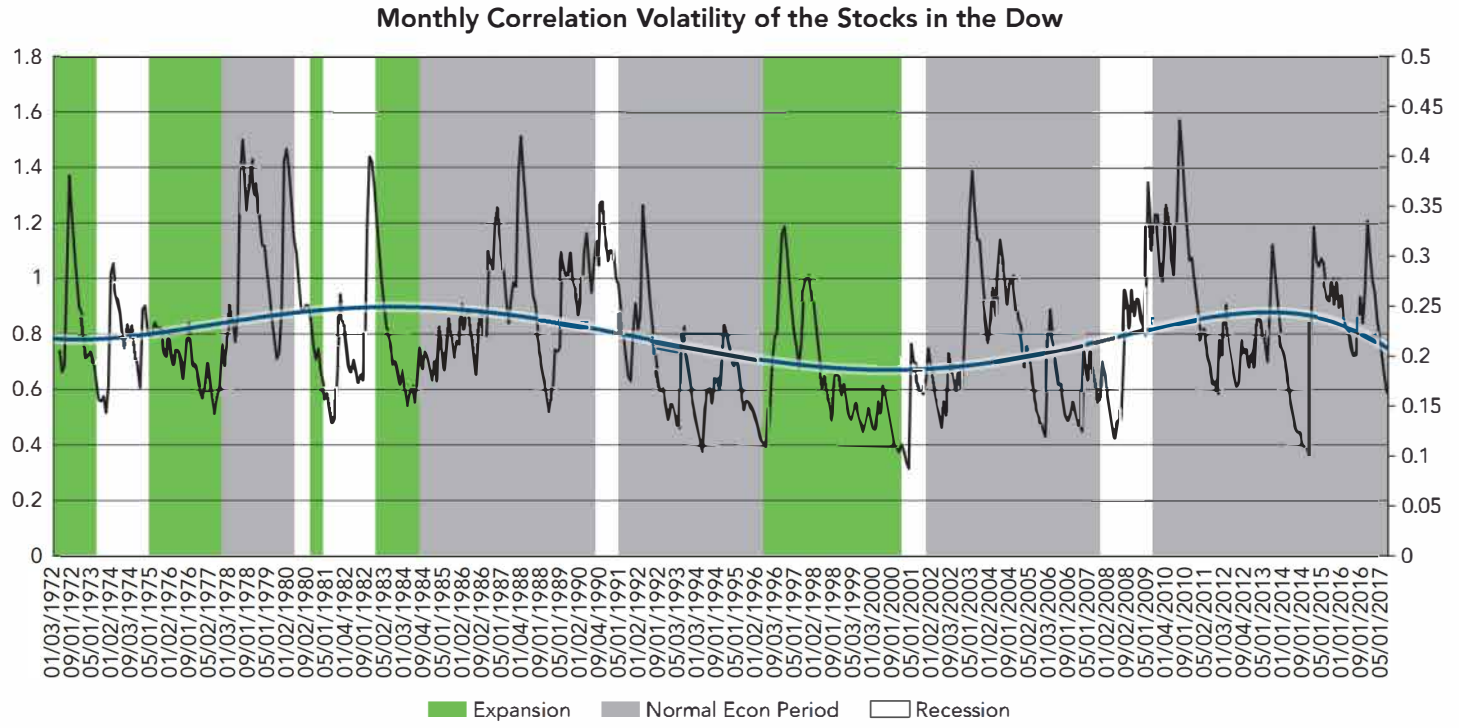


Figure 9.2 Correlation volatility of the average correlation of monthly 30 × 30 Dow stock return bins with respect to the state of the economy. The horizontal line shows the polynomial trendline of order 4.

Table 9.1 Correlation Level and Correlation Volatility with Respect to the State of the Economy

	Correlation Level	Correlation Volatility
Expansionary period	27.46%	71.17%
Normal economic period	33.06%	83.06%
Recession	36.96%	80.48%

In a recession, macroeconomic factors seem to dominate idiosyncratic factors, leading to a downturn across multiple stocks.

A further expected result in Table 9.1 is that correlation volatility is lowest in an economic expansion and higher in worse economic states. We did expect a higher correlation volatility in a recession compared with a normal economic state. However, it seems that high correlation levels in a recession remain high without much additional volatility. We will analyse whether the correlation volatility is an indicator for future recessions below. Altogether, Table 9.1 displays the higher correlation risk in bad economic times, which traders and risk managers should consider in their trading and risk management.

From Table 9.1, we observe a generally positive relationship between correlation level and correlation volatility. This is verified in more detail in Figure 9.3.

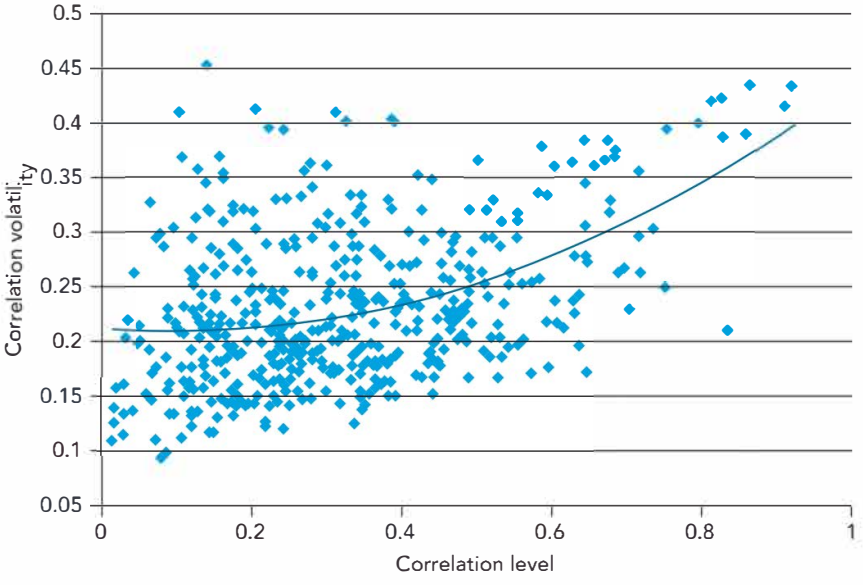


Figure 9.3 Positive relationship between correlation level and correlation volatility with a polynomial trendline of order 2 (data from 1972 to 2017).

9.2 DO EQUITY CORRELATIONS EXHIBIT MEAN REVERSION?

Mean reversion is the tendency of a variable to be pulled back to its long-term mean. In finance, many variables, such as bonds, interest rates, volatilities, credit spreads and more, are assumed to exhibit mean reversion. Fixed coupon bonds, which do not default, exhibit strong mean reversion: a bond is typically issued at par – for example, at USD 100. If the bond does not default, at maturity it will revert to exactly that price of USD 100, which is typically close to its long term mean.

Interest rates are also assumed to be mean-reverting: in an economic expansion, typically, demand for capital is high and interest rates rise. These high interest rates will eventually lead to cooling off of the economy, possibly leading to a recession. In this process capital demand decreases and interest rates decline from their high levels towards their long-term mean, eventually falling below it. Being in a recession, economic activity eventually increases again, often supported by monetary and fiscal policy. In this reviving economy, demand for capital increases, in turn increasing interest rates to their long term mean.

How Can We Quantify Mean Reversion?

Mean reversion is present if there is a negative relationship between the change of a variable, $S_t - S_{t-1}$, and the variable at $t - 1$, S_{t-1}. Formally, mean reversion exists if

$$\frac{\partial(S_t - S_{t-1})}{\partial S_{t-1}} < 0 \tag{9.1}$$

where

S_t: Price at time t

S_{t-1}: Price at the previous point in time $t - 1$

∂: Partial derivative coefficient

Equation (9.1) tells us: If S_{t-1} increases by a very small amount, $S_t - S_{t-1}$ will decrease by a certain amount and vice versa. In particular, if S_{t-1} has decreased (in the denominator), then at the next point in time t, mean reversion will "pull up" S_{t-1} to S_t, and therefore increasing $S_t - S_{t-1}$. Conversely, if S_{t-1} has increased (in the denominator) and is high in $t - 1$, then at the next point in time t, mean reversion will "pull down" S_{t-1} to S_t and therefore decreasing $S_t - S_{t-1}$. The degree of the "pull" is the degree of the mean reversion, also called mean reversion rate, mean reversion speed, or gravity.

Let's quantify the degree of mean reversion. Let's start with the discrete Vasicek 1987 process, which goes back to Ornstein–Uhlenbeck 1930:

$$S_t - S_{t-1} = a(\mu_s - S_{t-1})\Delta t + \sigma_s \varepsilon \sqrt{\Delta t} \tag{9.2}$$

where

S_t: Price at time t

S_{t-1}: Price at the previous point in time $t - 1$

a: Degree of mean reversion, also called mean reversion rate or gravity, $0 \leq a \leq 1$

μ_s: Long term mean of S

σ_s: Volatility of S

ε: Random drawing from a standardised normal distribution at time t, $\varepsilon(t) = n \sim (0, 1)$. We can compute ε as =normsinv(rand()) in Excel/VBA and norminv(rand) in MATLAB. See www.dersoft.com/epsilon.xlsx for details.

We are currently interested only in mean reversion, so for now we will ignore the stochastic part in equation (9.2), $\sigma_s \varepsilon \sqrt{\Delta t}$.

For ease of explanation, let's assume $\Delta t = 1$. Then, from equation (9.2), we see that a mean reversion parameter of $a = 1$ will pull S_{t-1} to the long-term mean μ_s completely at every time step, assuming S_{t-1} was below the mean. For example if S_{t-1} is 80 and u_s is 100, then $=1 \times (100 - 80) = 20$ so the S_{t-1} of 80 is "mean-reverted up" to its long-term mean of 100 in one time step. Naturally, a mean-reversion parameter "a" of 0.5 will lead to a mean reversion of 50% at each time step, and a mean-reversion parameter "a" of 0, will result in no mean reversion.

Let's now quantify mean reversion. Setting $\Delta t = 1$, equation (9.2) without stochasticity reduces to

$$S_t - S_{t-1} = a(\mu_s - S_{t-1}) \tag{9.3}$$

or

$$S_t - S_{t-1} = a\mu_s - aS_{t-1} \tag{9.4}$$

To find the mean reversion rate "a", we can run a standard regression analysis of the form

$$Y = \alpha + \beta X$$

Following equation (9.4) we are regressing $S_t - S_{t-1}$ with respect to S_{t-1}:

$$\underbrace{S_t - S_{t-1}}_{\gamma} = \underbrace{a\mu_s}_{\alpha} - \underbrace{aS_{t-1}}_{\beta X} \tag{9.5}$$

Importantly, from equation (9.5), we observe that the regression coefficient β is equal to the negative mean-reversion parameter "a".

We now run a regression of equation (9.5) to find the empirical mean reversion of our correlation data. Hence S represents the 30×30 Dow stock monthly average correlations from 1972 to 2017. The regression analysis is displayed in Figure 9.4.

The regression function in Figure 9.4 displays a strong mean reversion of 79.03%. This means that, on average in every month, a deviation from the long-term correlation mean (32.38%

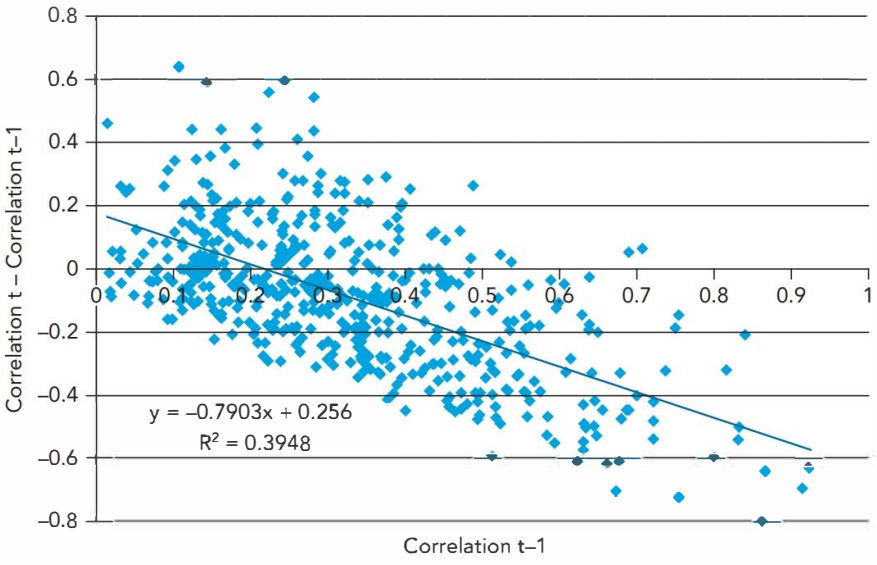


Figure 9.4 **Regression function (2.5) for 534 monthly average Dow stock return correlations from 1972 to 2017.**

in our study) is pulled back to that long-term mean by 79.03%. We can observe this strong mean reversion also by looking at Figure 9.1. An upward spike in correlation is typically followed by a sharp decline in the next time period, and vice versa.

Let's look at an example of modelling correlation with mean reversion.

Example 9.1: The long-term mean of the correlation data is 32.38%. In February 2017, the averaged correlation of the 30 × 30 Dow correlation matrices was 26.15%. From the regression function from 1972 to 2017, we find that the average mean reversion is 79.03%. What is the expected correlation for March 2017 following equation (9.3) or (9.4)?

Solving equation (9.3) for S_t, we have $S_t = a(u_s - S_{t-1}) + S_{t-1}$. Hence the expected correlation in March is

$$S_t = 0.7903 \times (0.3238 - 0.2615) + 0.2615 = 0.3107$$

As a result, when applying equation (9.3) with the mean reversion rate of 79.03%, we expect the correlation in March 2017 to be 31.07%.[1]

9.3 DO EQUITY CORRELATIONS EXHIBIT AUTOCORRELATION?

Autocorrelation is the degree to which a variable is correlated to its past values. Autocorrelation can be quantified with the Nobel prize-rewarded ARCH (Autoregressive Conditional Heterosce-dasticity) model of Robert Engle (1982) or its extension GARCH (Generalized Autoregressive Conditional Heteroscedasticity) by Tim Bollerslev (1988). However, we can also regress the time series of a variable to its past time series values to derive autocorrelation. This is the approach we will take here.

In finance, positive autocorrelation is also termed "persistence". In mutual-fund or hedge-fund performance analysis, an investor typically wants to know if an above-market performance of a fund has persisted for some time, ie, is positively correlated to its past strong performance.

Autocorrelation is the "reverse property" to mean reversion: the stronger the mean reversion, ie, the stronger a variable is pulled back to its long-term mean, the lower is the autocorrelation, ie, the lower is its correlation to its past values, and vice versa.

For our empirical correlation analysis, we derive the autocorrelation AC for a time lag of one period with the equation

$$AC(\rho_t, \rho_{t-1}) = \frac{COV(\rho_t, \rho_{t-1})}{\sigma(\rho_t)\sigma(\rho_{t-1})} \tag{9.6}$$

where

AC: Autocorrelation

ρ_t: Correlation values for time period t (in our study, the monthly average of the 30 × 30 Dow stock return correlation matrices from 1972 to 2017, after eliminating the unity correlation on the diagonal)

ρ_{t-1}: Correlation values for time period $t - 1$ (ie, the monthly correlation values starting and ending one month prior than period t

COV: Covariance, see equation (1.3) for details

Equation (9.6) is algebraically identical with the Pearson correlation coefficient equation (1.4). The autocorrelation just uses the correlation values of time period t and time period $t - 1$ as inputs.

Following equation (9.6), we find the one-period lag autocorrelation of the correlation values from 1972 to 2017 to be 20.97%. As mentioned above, autocorrelation is the "opposite property" of mean reversion. Therefore, not surprisingly, the autocorrelation of 20.97% and the mean reversion is our study of 79.03% (see the above section "Do equity correlations exhibit mean reversion?") add up to 1.

[1] Note that we have omitted any stochasticity, which is typically included when modelling financial variables, as shown in equation (9.2).

Figure 9.5 shows the autocorrelation with respect to different time lags.

From Figure 9.5, we observe that 2-month lag autocorrelation, so autocorrelation with respect to two months prior, produces the highest autocorrelation. Altogether we observe the expected decay in autocorrelation with respect to time lags of earlier periods.

9.4 HOW ARE EQUITY CORRELATIONS DISTRIBUTED?

The input data of our distribution tests are daily correlation values between all 30 Dow stocks from 1972 to 2017. This resulted in 464,580 correlation values. The distribution is shown in Figure 9.6.

From Figure 9.6, we observe that most correlations between the stocks in the Dow are positive. In fact, 77.23% of all 464,580 correlation values were positive.

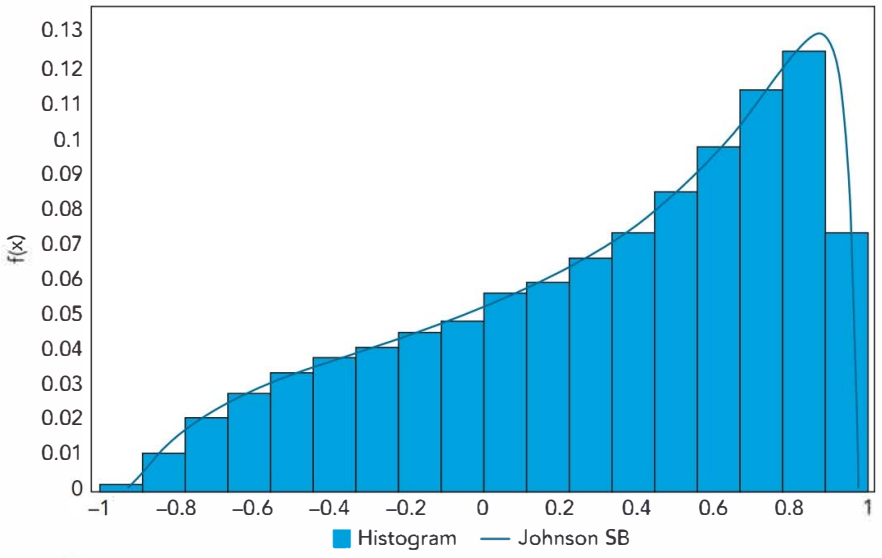


Figure 9.6 **Histogram of 464,580 correlations between the Dow 30 stocks from 1972 to 2017; the continuous line shows the Johnson SB distribution, which provided the best fit.**

We tested 61 distributions for fitting the histogram in Figure 9.6, applying three standard fitting tests: (a) Kolmogorov–Smirnov, (b) Anderson–Darling and (c) Chi-Squared. Not surprisingly, the versatile Johnson SB distribution with four parameters $-\gamma$ and δ for the shape, μ for location and σ for scale–provided the best fit.

Standard distributions such as normal distribution, lognormal distribution or beta distribution provided a poor fit.

We also tested the correlation distribution between the Dow stocks for different states of the economy. The results were slightly but not significantly different; see www.dersoft.com/correlationfitting.docx.

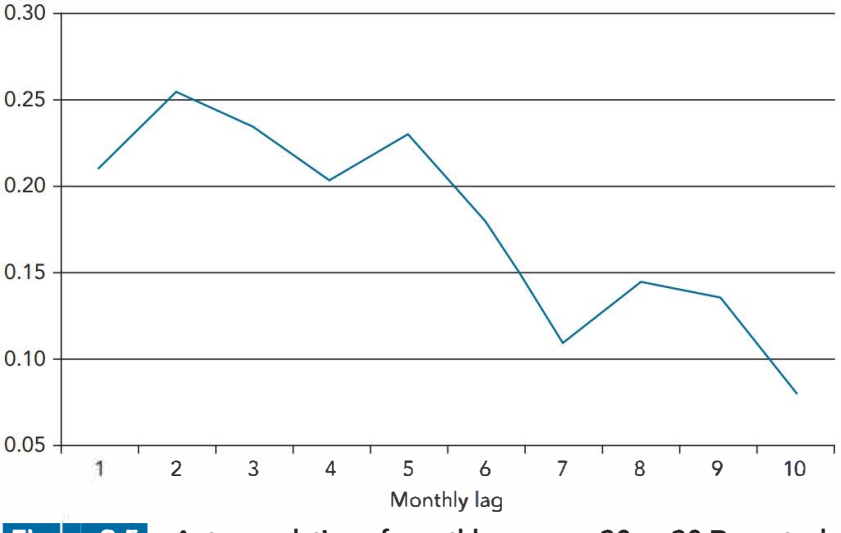


Figure 9.5 **Autocorrelation of monthly average 30 × 30 Dow stock correlations from 1972 to 2017; the time period of the lags is months.**

9.5 IS EQUITY CORRELATION VOLATILITY AN INDICATOR FOR FUTURE RECESSIONS?

In our study from 1972 to 2017, six recessions occurred: (1) a severe recession in 1973–74 following the first oil price shock, (2) a short recession in 1980, (3) a severe recession in 1981–82 following the second oil price shock, (4) a mild recession in 1990–91, (5) a mild recession in 2001 after the Internet bubble burst and (6) the "great recession" 2007–09, following the global financial crisis. Table 9.2 displays the relationship of a change in the correlation volatility preceding the start of a recession.

From Table 9.2, we observe the severity of the 2007–09 "great recession", which exceeded the severity of the oil price shock induced recessions in 1973–74 and 1981–82.

From Table 9.2, we also notice that, except for the mild recession in 1990–91, before every recession a downturn in correlation volatility occurred. This coincides with the fact that correlation volatility is low in an expansionary period (see Table 9.1), which often precedes a recession. However, the relationship between a decline in volatility and the severity of the

Table 9.2 Decrease in Correlation Volatility, Preceding a Recession. The Decrease in Correlation Volatility is Measured as a 6-Month Change of 6-Month Moving Average Correlation Volatility. The Severity of the Recession is Measured as the Total GDP Decline During the Recession

	% Change in Correlation Volatility Before Recession	Severity of Recession (% change of GDP)
1973–74	−7.22%	−11.93%
1980	−10.12%	−6.53%
1981–82	−4.65%	−12.00%
1990–91	0.06%	−4.05%
2001	−5.55%	−1.80%
2007–09	−2.64%	−14.75%

recession is statistically non-significant. The regression function is almost horizontal and the R^2 is close to zero. Studies with more data, going back to 1920, are currently being conducted.

9.6 PROPERTIES OF BOND CORRELATIONS AND DEFAULT PROBABILITY CORRELATIONS

Our preliminary studies of 7,645 bond correlations and 4,655 default probability correlations display similar properties as equity correlations. Correlation levels were higher for bonds (41.67%) and slightly lower for default probabilities (30.43%) compared with equity correlation levels (34.83%). Correlation volatility was lower for bonds (63.74%) and slightly higher for default probabilities (87.74%) compared with equity correlation volatility (79.73%).

Mean reversion was present in bond correlations (25.79%) and in default probability correlations (29.97%). These levels were lower than the very high equity correlation mean reversion of 77.51%.

The default probability correlation distribution is similar to equity correlation distribution (see Figure 9.4) and can be replicated best by the Johnson SB distribution. However, the bond correlation distribution shows a more normal shape and can be best fitted with the generalised extreme value distribution and quite well with the normal distribution. Some fitting results are at www.dersoft.com/correlationfitting.docx. The bond correlation and default probability results are currently being verified with a larger sample data base.

SUMMARY

The following are the main findings of our empirical analysis:

(a) Our study confirmed that the worse the state of the economy the higher are equity correlations. Equity correlations were extremely high during the great recession of 2007-09 and reached 96.97% in February 2009.

(b) Equity correlation volatility is lowest in an expansionary period and higher in normal and recessionary economic periods. Traders and risk managers should take these higher correlation levels and higher correlation volatility that markets exhibit during economic distress into consideration.

(c) Equity correlation levels and equity correlation volatility are positively related.

(d) Equity correlations show very strong mean reversion. The Dow correlations from 1972 to 2017 showed a monthly mean reversion of 79.03%. Hence, when modelling correlation, mean reversion should be included in the model.

(e) Since equity correlations display strong mean reversion, they display low autocorrelation. The degree of autocorrelations shows the typical decrease with respect to time (ie, the autocorrelation is higher for more recent time lags).

(f) The equity correlation distribution showed a distribution, which can be replicated well with the Johnson SB distribution. Other distributions such as normal, lognormal and beta distribution do not provide a good fit.

(g) First results show that bond correlations display similar properties as equity correlations. Bond correlation levels and bond correlation volatilities are generally higher in economic bad times. In addition, bond correlations exhibit mean reversion, although lower mean reversion than equity correlations exhibit.

(h) First results show that default correlations also exhibit properties seen in equity correlations. Default probability correlation levels are slightly lower than equity correlation levels, and default probability correlation volatilities are slightly higher than equity correlations. Studies with more data are currently being conducted.

The following questions are intended to help candidates understand the material. They are *not* actual FRM exam questions.

QUESTIONS

9.1 In which state of the economy are equity correlations the highest?

9.2 In which state of the economy is equity correlation volatility high?

9.3 What follows from Questions 1 and 2 for risk management?

9.4 What is mean reversion?

9.5 How can we quantify mean reversion?

9.6 What is autocorrelation? Name two approaches for how to quantify autocorrelation.

9.7 For equity correlations, we see the typical decrease of autocorrelation with respect to time lags. What does that mean?

9.8 How are mean reversion and autocorrelation related?

9.9 What is the distribution of equity correlations?

9.10 When modelling stocks, bonds, commodities, exchange rates, volatilities and other financial variables, we typically assume a normal or lognormal distribution. Can we do this for equity correlations?

10

Financial Correlation Modeling—Bottom-Up Approaches

Learning Objectives

After completing this reading, you should be able to:

- Explain the purpose of copula functions and how they are applied in finance.
- Describe the Gaussian copula and explain how to use it to derive the joint probability of default of two assets.
- Summarize the process of finding the default time of an asset correlated to all other assets in a portfolio using the Gaussian copula.

Excerpt is taken from Chapter 5 of Correlation Risk Modeling and Management, *2nd Edition, by Gunter Meissner.*

10.1 COPULA CORRELATIONS

A fairly recent and famous as well as infamous correlation approach applied in finance is the copula approach. Copulas go back to Abe Sklar (1959). Extensions are provided by Schweizer and Wolff (1981) and Schweizer and Sklar (1983). One-factor copulas were introduced to finance by Oldrich Vasicek in 1987. More versatile, multivariate copulas were applied to finance by David Li in 2000.

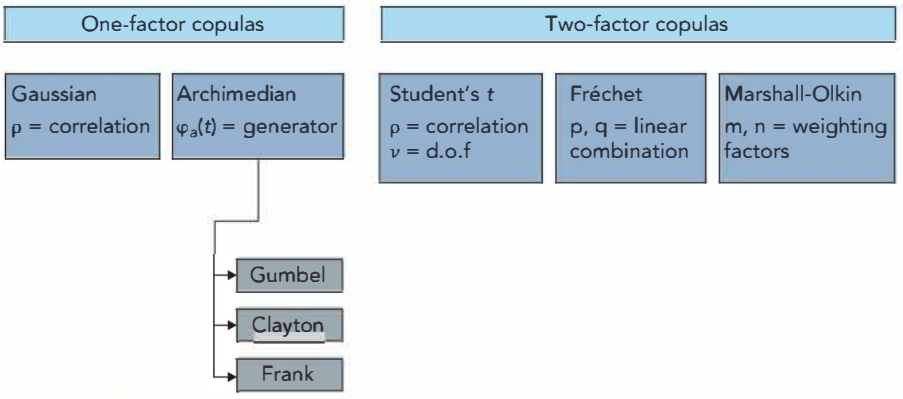


Figure 10.1 Popular copula functions in finance.

When flexible copula functions were introduced to finance in 2000, they were enthusiastically embraced but then fell into disgrace when the global financial crisis hit in 2007. Copulas became popular because they could presumably solve a complex problem in an easy way: it was assumed that copulas could correlate multiple assets; for example, the 125 assets in a CDO, with a single, although multi-dimensional, function. Let's first look at the maths of the copula correlation concept.

Copula functions are designed to simplify statistical problems. They allow the joining of multiple univariate distributions to a single multivariate distribution. Formally, a copula function C transforms an n-dimensional function on the interval [0,1] into a unit-dimensional one:

$$C : [0, 1]^n \rightarrow [0, 1] \tag{10.1}$$

More explicitly, let $G_i(u_i)$ be a univariate, uniform distribution with $u_i = u_1, \ldots, u_n$, and $i \in N$. Then there exists a copula function C such that

$$C[G_1(u_1), \ldots, G_n(u_n)] = F_n[F_1^{-1}(G_1(u_1)), \ldots, F_n^{-1}(G_n(u_n)); \rho_F] \tag{10.2}$$

where $G_i(u_i)$ are called marginal distributions and F_n is the joint cumulative distribution function. F_i^{-1} is the inverse of F_i. ρ_F is the correlation structure of F_n.

Equation (10.2) reads: given are the marginal distributions $G_1(u_1)$ to $G_n(u_n)$. There exists a copula function that allows the mapping of the marginal distributions $G_1(u_1)$ to $G_n(u_n)$ via F^{-1} and the joining of the (abscise values) $F^{-1}(G_i(u_i))$ to a single, n-variate function $F_n[F^{-1}(G_1(u_1)), \ldots, F_n^{-1}(G_n(u_n))]$ with correlation structure of ρ_F.

If the mapped values $F_i^{-1}(G_i(u_i))$ are continuous, it follows that C is unique. For detailed properties and proofs of equation (10.2), see Sklar (1959) and Nelsen (2006).

Numerous types of copula functions exist. They can be broadly categorised in one-parameter copulas as the Gaussian copula;[1] and the Archimedean copula family, the most popular being Gumbel, Clayton and Frank copulas. Often cited two-parameter copulas are student-t, Frechet, and Marshall-Olkin. Figure 10.1 shows an overview of popular copula functions.

The Gaussian Copula

Due to its convenient properties, the Gaussian copula C_G is among the most applied copulas in finance. In the n-variate case, it is defined

$$C_G[G_1(u_1), \ldots, G_n(u_n)] = M_n[N^{-1}(G_1(u_1)), \ldots, N^{-1}(G_n(u_n)); \rho_M] \tag{10.3}$$

where M_n is the joint, n-variate cumulative standard normal distribution with ρ_M, the $n \times n$ symmetric, positive-definite correlation matrix of the n-variate normal distribution M_n. N^{-1} is the inverse of a univariate standard normal distribution.

If the $G_x(u_x)$ are uniform, then the $N^{-1}(G_x(u_x))$ are standard normal and M_n is standard multivariate normal. For a proof, see Cherubini *et al* 2005.

It was David Li (2000), who transferred the copula approach of equation (10.3) to finance. He defined the cumulative default probabilities Q for entity i at a fixed time t, $Q_i(t)$ as marginal distributions. Hence we derive the Gaussian default time copula C_{GD},

$$C_{GD}[Q_i(t), \ldots, Q_n(t)] = M_n[N^{-1}(Q_1(t)), \ldots N^{-1}(Q_n(t)); \rho_M] \tag{10.4}$$

[1] Strictly speaking, only the bivariate Gaussian copula is a one-parameter copula, the parameter being the copula correlation coefficient. A multivariate Gaussian copula may incorporate a correlation *matrix*, containing various correlation coefficients.

Equation (10.4) reads: given are the marginal distributions, ie, the cumulative default probabilities Q of entities $i = 1$ to n at times t, $Q_i(t)$. There exists a Gaussian copula function C_{GD}, which allows the mapping of the marginal distributions $Q_i(t)$ via N^{-1} to standard normal and the joining of the (abscise values) $N^{-1}Q_i(t)$ to a single n-variate standard normal distribution M_n with the correlation structure ρ_M.

More precisely, in equation (10.4) the term N^{-1} maps the cumulative default probabilities Q of asset i for time t, $Q_i(t)$, percentile to percentile to a univariate standard normal distribution. So the 5th percentile of $Q_i(t)$ is mapped to the 5th percentile of the standard normal distribution; the 10th percentile of $Q_i(t)$ is mapped to the 10th percentile of the standard normal distribution, etc. As a result, the $N^{-1}(Q_i(t))$ in equation (10.4) are abscise (x-axis) values of the standard normal distribution. For a numerical example see example 10.1 and Figure 10.2 below. The $N_i^{-1}(Q_i(t))$ are then joined to a single n-variate distribution M_n by applying the correlation structure of the multivariate normal distribution with correlation matrix ρ_M. The probability of n correlated defaults at time t is given by M_n.

We will now look at the Gaussian copula in an example.

Example 10.1 Let's assume we have two companies, B and Caa, with their estimated default probabilities for years 1 to 10 as displayed in Table 10.1.

Default probabilities for investment-grade companies typically increase in time, since uncertainty increases with time. However, in Table 10.1 both companies are in distress. For these companies the next years are the most difficult. If they survive these next years, their default probability decreases.

Let's now find the joint default probabilities of the companies B and Caa for any time t with the Gaussian copula function (10.4). First, we map the cumulative default probabilities $Q(t)$, which are in columns 3 and 5 in Table 10.1, to the standard normal distribution via $N^{-1}(Q(T))$. Computationally, this can be done with = normsin-v(Q(t)) in Excel or norminv(Q(t)) in MATLAB. Graphically the mapping can be represented in two steps, which are displayed in Figure 10.2. In the lower graph of Figure 10.2, the cumulative default probability of asset B, $Q_B(t)$, is displayed. We first map these cumulative probabilities percentile to percentile to a cumulative standard normal distribution in the upper graphs of Figure 10.1 (up arrows). In a second step the abscise (x-axis) values of the cumulative normal distribution are found (down arrows).

The same mapping procedure is done for company Caa, ie, the cumulative default probabilities of company Caa, which are displayed in Table 10.1 in column 5 are mapped percentile to percentile to a cumulative standard normal distribution via $N^{-1}(Q_{Caa}(t))$.

We have now derived the percentile to percentile mapped cumulative default probability values of our companies to a cumulative standard normal distribution. These values are displayed in Table 10.2, columns 3 and 5.

We can now use the derived $N^{-1}(Q_B(t))$ and $N^{-1}(Q_{Caa}(t))$ and apply them to equation (10.4).

Since we have only $n = 2$ companies B and Caa in our example, equation (10.4) reduces to

$$M_2[N^{-1}(Q_B(t)), N^{-1}(Q_{Caa}(t)); \rho] \tag{10.5}$$

From equation (10.5) we see that since we have only two assets in our example, we have only one correlation coefficient ρ, not a correlation matrix ρ_M.

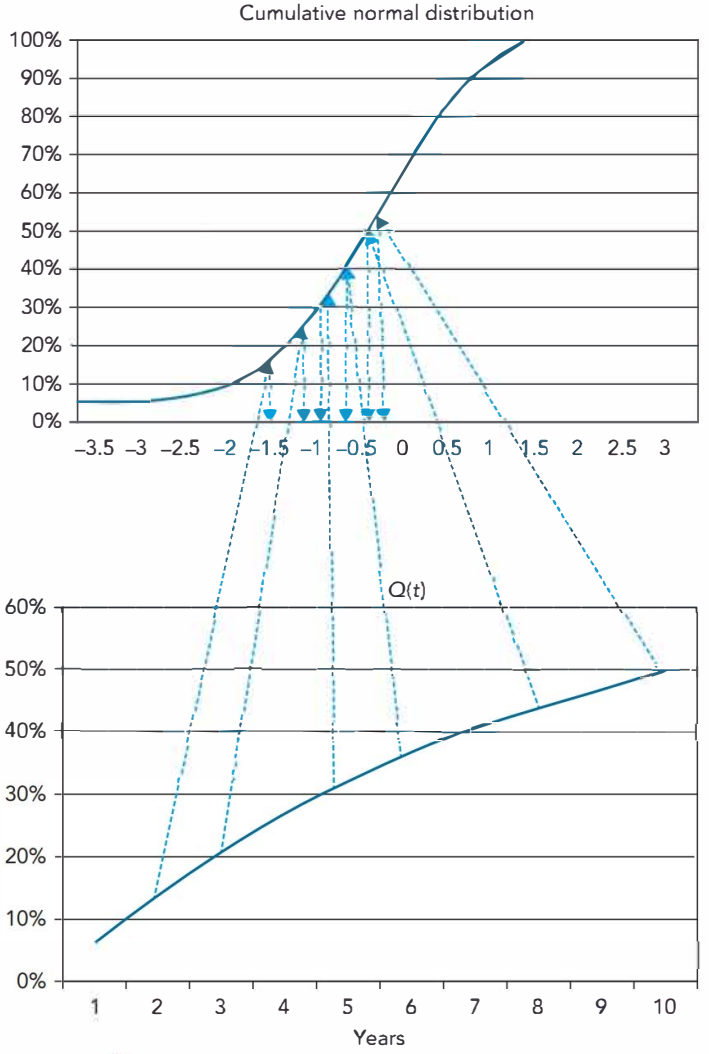


Figure 10.2 Graphical representation of the copula mapping $N^{-1}(Q(t))$.

Table 10.1 Default Probability and Cumulative Default Probability of Companies B and Caa

Default Time t	Company B Default Probability	Company B Cumulative Default Probability $Q_B(t)$	Company Caa Default Probability	Company Caa Cumulative Default Probability $Q_{Caa}(t)$
1	6.51%	6.51%	23.83%	23.83%
2	7.65%	14.16%	13.29%	37.12%
3	6.87%	21.03%	10.31%	47.43%
4	6.01%	27.04%	7.62%	55.05%
5	5.27%	32.31%	5.04%	60.09%
6	4.42%	36.73%	5.13%	65.22%
7	4.24%	40.97%	4.04%	69.26%
8	3.36%	44.33%	4.62%	73.88%
9	2.84%	47.17%	2.62%	76.50%
10	2.84%	50.01%	2.04%	78.54%

Table 10.2 Cumulative Default Probabilities Mapped Percentile to Percentile to Standard Normal. For Example, Using Excel, the Value −1.5133 is Derived using = normsinv(0.0651) = −1.5133

Default Time t	Company B Cumulative Default Probability $Q_B(t)$	Company B Cumulative Standard Normal Percentiles $N^{-1}(Q_B(t))$	Company Caa Cumulative Default Probability $Q_{Caa}(t)$	Company Caa Cumulative Standard Normal Percentiles $N^{-1}(Q_{Caa}(t))$
1	6.51%	−1.5133	23.83%	−0.7118
2	14.16%	−1.0732	37.12%	−0.3287
3	21.03%	−0.8054	47.43%	−0.0645
4	27.04%	−0.6116	55.05%	0.1269
5	32.31%	−0.4590	60.09%	0.2557
6	36.73%	−0.3390	65.22%	0.3913
7	40.97%	−0.2283	69.26%	0.5032
8	44.33%	−0.1426	73.88%	0.6397
9	47.17%	−0.0710	76.50%	0.7225
10	50.01%	0.0003	78.54%	0.7906

Importantly, the copula model now assumes that we can apply the correlation structure ρ_M or ρ of the multivariate distribution (in our case the Gaussian multivariate distribution M), to the transformed marginal distributions $N^{-1}(Q_B(t))$ and $N^{-1}(Q_{Caa}(t))$. This is done for mathematical and computational convenience.

The bivariate normal distribution M_2 is displayed in Figure 10.3.

The code for the bivariate cumulative normal distribution M can be found on the Internet. It is also displayed at www.dersoft.com/2assetdefaulttimecopula.xlsm in Module 1.

We now have all necessary ingredients to find the joint default probabilities of our companies B and Caa. For example, we

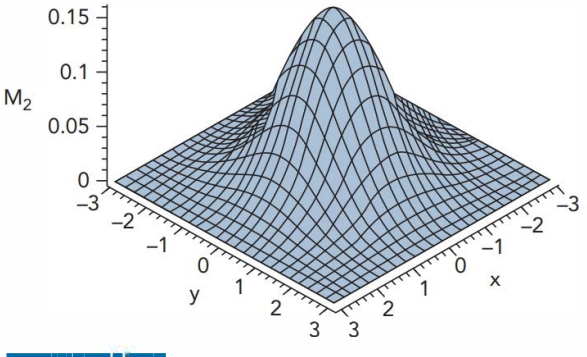


Figure 10.3 Bivariate (non-cumulative) normal distribution.

can answer the question: what is the joint default probability Q of companies B and Caa in the next year assuming a one-year Gaussian default correlation of 0.4? The solution is

$$Q(t_b \le 1 \cap t_{Caa} \le 1)$$
$$\equiv M(x_B \le -1.5133 \cap x_{Caa} \le -0.7118, \rho = 0.4) = 3.44\% \quad (10.6)$$

where t_B is the default time of company B and t_{Caa} is the default time of company Caa. x_B and x_{Caa} are the mapped abscise values of the bivariate normal distribution, which are derived from Table 10.2.

In another example, we can answer the question: what is the joint probability of company B defaulting in year 3 and company Caa defaulting in year 5? It is

$$Q(t_B \le 3 \cap t_{Caa} \le 5)$$
$$\equiv M(x_B \le -0.8054 \cap x_{Caa} \le 0.2557, \rho = 0.4) = 16.93\% \quad (10.7)$$

Equations (10.6) and (10.7) show why this type of copula is also called "default-time copula". We are correlating the default times of two or more assets t_i. A spreadsheet that correlates the default times of two assets can be found at www.dersoft.com/2assetdefaulttimecopula.xlsm. The numerical value of 3.44% of equation (10.6) is in cell Q17.

Simulating the Correlated Default Time for Multiple Assets

The preceding example considers only two assets. We will now find the default time for an asset that is correlated to the default times of all other assets in a portfolio using the Gaussian copula. To derive the default time τ of asset i, τ_i, which is correlated to the default times of all other assets $i = 1, \ldots, n$, we first derive a sample $M_n(\cdot)$ from a multivariate copula (r.h.s. of equation (10.5) in the Gaussian case), $M_n(\cdot) \in [0, 1]$. This is done via Cholesky decomposition. The sample includes the default correlation via the default correlation matrix ρ_M of the n-variate standard normal distribution M_n. We equate the sample $(\cdot)$ from M_n, $M_n(\cdot)$ with the cumulative individual default probability Q of asset i at time τ, $Q_i(\tau_i)$. Therefore,

$$M_n(\cdot) = Q_i(\tau_i) \text{ or} \quad (10.8)$$
$$\tau_i = Q_i^{-1}(M_n(\cdot)) \quad (10.9)$$

There is no closed-form solution for equation (10.8) or (10.9). To find the solution, we first take the sample $M_n(\cdot)$ and use equation (10.8) to equate it to $Q_i(\tau_i)$. This can be done with a search procedure such as Newton-Raphson. We can also use a simple lookup function in Excel.

Let's assume the random drawing from $M_n(\cdot)$ was 35%. We now equate 35% with the market-given function $Q_i(\tau_i)$ and find the expected default time of asset i, τ_i. This is displayed in Figure 10.4, where $\tau_i = 5.5$ years. We repeat this procedure numerous times, for example 100,000 times and average each τ_i of every simulation to find our estimate for τ_i. Importantly, the estimated default time of asset i, τ_i, includes the default correlation with the other assets in the portfolio, since the correlation matrix is an input of the n-variate standard normal distribution M_n.

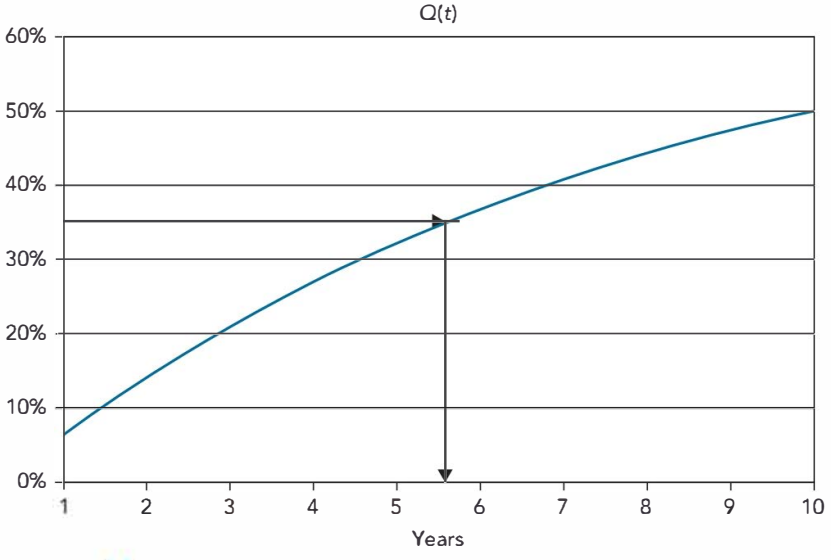


Figure 10.4 **Finding the default time τ of 5.5 years from equation (10.8) for a random sample of the n-variate normal distribution $M_n(\cdot)$ of 35%.**

11

Regression Hedging and Principal Component Analysis

Learning Objectives

After completing this reading, you should be able to:

- Explain the drawbacks to using a DV01-neutral hedge for a bond position.
- Describe a regression hedge and explain how it can improve a standard DV01-neutral hedge.
- Calculate the regression hedge adjustment factor, beta.
- Calculate the face value of an offsetting position needed to carry out a regression hedge.
- Calculate the face value of multiple offsetting swap positions needed to carry out a two-variable regression hedge.
- Compare and contrast level and change regressions.
- Explain why and how a regression hedge differs from a hedge based on a reverse regression.
- Describe principal component analysis and explain how it is applied to constructing a hedging portfolio.

Excerpt is Chapter 6 of Fixed Income Securities: Tools for Today's Markets, *Fourth Edition, by Bruce Tuckman and Angel Serrat.*

The risk metrics and hedges assume relationships across rates of different terms. The assumptions are typically motivated by a combination of economic theory, empirical analysis of historical data, and views about future economic and financial developments. This chapter introduces approaches that rely explicitly on historical data. It would be an oversimplification, however, to categorize the techniques as subjective, while categorizing the approaches of this chapter as objective. Empirical methods also require assumptions, such as the number of rates or instruments used in the analysis, the particular rates or instruments chosen, and the historical time period selected for estimation.

The first section of this chapter describes single-variable regression hedging in the context of hedging a 40-year Johnson & Johnson (JNJ) bond with a 30-year Treasury. The second section describes two-variable regression hedging in the context of a relative value trade of 20-year versus 10- and 30-year Treasuries. The third and fourth sections discuss two other issues that arise in the context of regression hedging, namely, the choice between *level* and *change* regressions, and *reverse* regressions.

One conceptual problem with using regression hedging in practice is that each regression is essentially a different model of the term structure of interest rates with different underlying assumptions. Consider, for example, the manager of a trading desk in which some traders are using single-variable regressions and some two-variable regressions, or some are estimating regressions from one month of historical data and others from six months.

The final section of the chapter introduces a unified empirical description of how the entire term structure evolves, namely, *principal component analysis (PCA)*. PCA provides both an empirical hedging methodology, which can be used consistently across the term structure, along with easily interpreted descriptions of how the term structure fluctuates over time. While presentations of PCA tend to be highly mathematical, great effort has been made in this chapter to make the material more broadly accessible.

11.1 SINGLE-VARIABLE REGRESSION HEDGING

This section considers the problem of a market maker or a relative value trader on May 14, 2021, who purchases $100 million face amount of the JNJ 2.450s of 09/01/2060 and hedges the resulting interest rate risk by selling the US Treasury 1.625s of 11/15/2050. Because there is no 40-year Treasury bond outstanding, the 1.625s of 11/15/2050, with about 30 years to maturity, are selected as the best alternative. Table 11.1 gives the coupons, maturity dates, yields, and DV01s of these two bonds, along with those of two other bonds that are referenced in the next section.

Table 11.1 **Yields and Yield-Based DV01s for the JNJ 2.450s of 09/01/2060 and Selected US Treasury Bonds, as of May 14, 2021. Yields Are in Percent.**

Issuer	Bond	Yield	DV01
JNJ	2.450s of 09/01/2060	2.962	0.2124
Treasury	0.875s of 11/15/2030	1.601	0.0847
Treasury	1.375s of 11/15/2040	2.246	0.1446
Treasury	1.625s of 11/15/2050	2.364	0.1910

The trader can choose the face amount of the Treasury bond in the hedge using the ratio of DV01s. In this case, the trader sells $100 million × 0.2124/0.1910, or $111.2 million. This hedge assumes that the yields of the JNJ and Treasury bonds move up or down in parallel. But because the JNJ bonds sell at a changing corporate spread to Treasuries, and because 40- and 30-year rates are not perfectly correlated, there is good reason to question the assumption of parallel yield shifts in this case.

The scatter plot in Figure 11.1 shows the daily changes in yield of the JNJ bonds against those of the Treasury bond from January 19, 2021, to May 14, 2021, which is a window of about four months before the date of the hedge. The line in the figure is the regression line fitted through the data, which is discussed presently. The figure teaches two lessons. First, there is a lot of variation in the relationship between these changes. Some days,

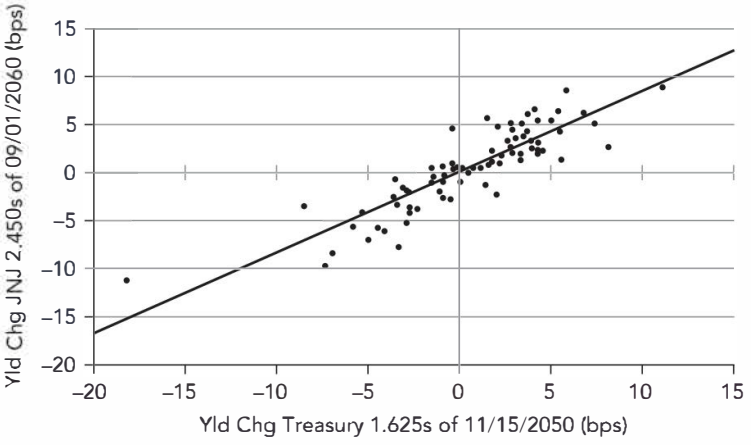


Figure 11.1 **Regression of Daily Changes in Yields of the JNJ 2.450s of 09/01/2060 on Daily Changes in Yields of the Treasury 1.625s of 11/15/2050, from January 19, 2021, to May 14, 2021.**

the Treasury yield changes by more (e.g., the point on the graph at (−18.2, −11.1)); some days by less (e.g., the point (−3.3, −7.7)); and some days even in opposite directions (e.g., the point (2.0, −2.2)). Second, from the slope of the line, the average relationship between yield changes is less than one-to-one; that is, the change in the yield of the JNJ bonds tends to be less than the change in the yield of the Treasury bonds.

Figure 11.1 implicitly assumes that the most relevant period for designing an empirical hedge is the recent, immediate past. This assumption is often reasonable, but there are times and situations in which some earlier period seems more relevant. For example, if the Federal Reserve is expected to raise short-term rates in the immediate future, a similar episode in the past might be more relevant to the future than the recent past, in which the Federal Reserve left rates unchanged or lowered them. Choosing the length of the observation or estimation window is also part of the art of regression hedging. Too short a window might fail to furnish statistically reliable estimates, but too long a window might include less relevant historical data.

In light of the empirical evidence in Figure 11.1, the trader might very well choose to: i) adjust the hedge ratio to account for the less than one-to-one relationships between changes in yields; and ii) measure the variation around the average relationship to gain a better understanding of the risk of the hedged position. Regression analysis is a tool with which to achieve both of these objectives.

Let Δy_t^{JNJ} and Δy_t^{30} be the changes in yields of the JNJ and 30-year Treasury bonds on date t, respectively. A regression model linking these changes is,

$$\Delta y_t^{JNJ} = \alpha + \beta \Delta y_t^{30} + \varepsilon_t \tag{11.1}$$

Equation 11.1 says that the *dependent variable,* here the change in the yield of the JNJ bond, equals: a constant or intercept, α; plus a slope, β, times the *independent* variable, here the change in the yield of the 30-year Treasury bond; plus an error term, ε_t. The unknown constant and slope parameters are estimated from the data, in a manner explained presently. These estimated parameters, denoted $\hat{\alpha}$ and $\hat{\beta}$, respectively, can then be used for prediction. Given the change in the Treasury bond yield on date t, the predicted change in the yield of the JNJ bonds on that date, denoted $\Delta \hat{y}_t^{JNJ}$, is,

$$\Delta \hat{y}_t^{JNJ} = \hat{\alpha} + \hat{\beta} \Delta y_t^{30} \tag{11.2}$$

and the realized error or *residual* on that day, $\hat{\varepsilon}_t$, is given by,

$$\hat{\varepsilon}_t = \Delta y_t^{JNJ} - \hat{\alpha} - \hat{\beta} \Delta y_t^{30} \tag{11.3}$$

$$= \Delta y_t^{JNJ} - \Delta \hat{y}_t^{JNJ} \tag{11.4}$$

For example, say that the estimated constant and slope parameters are 0 and 0.84, respectively, and that the Treasury bond yield changes by −18.2 basis points. Then, by Equation (11.2), the predicted change in the yield of the JNJ bond is 0 + 0.84 × (−18.2) = −15.3 basis points. If, furthermore, the actual change in the JNJ bond is −11.1 basis points, then, by (11.3) or (11.4), the realized error or residual is −11.1 − (−15.3) or 4.2 basis points. In Figure 11.1, this residual can be thought of as a vertical line dropped from the data point, (−18.2, −11.1), to the regression line.

Least-squares estimation of the unknown parameters finds $\hat{\alpha}$ and $\hat{\beta}$ to minimize the sum of the squares of the residuals over the observation period,

$$\sum_t \hat{\varepsilon}_t^2 = \sum_t \left(\Delta y_t^{JNJ} - \hat{\alpha} - \hat{\beta} \Delta y_t^{30} \right)^2 \tag{11.5}$$

where the equality follows from Equation (11.3). Squaring of the errors ensures that offsetting positive and negative errors are not considered as acceptable as zero errors, and that large errors in absolute value are penalized heavily relative to smaller errors.

Least-squares estimation assumes that the regression model is a true description of the dynamics of the dependent and independent variables, that the errors across time have the same probability distribution, that they are independent of each other, and that they are uncorrelated with the independent variable. Under these assumptions, least-squares parameter estimates are linear, unbiased, consistent, and efficient.[1]

Least-squares estimation is available in many statistical packages and spreadsheets. Table 11.2 gives a typical summary output from estimating Equation (11.1) using the data shown in Figure 11.1. The estimate of the slope coefficient, $\hat{\beta}$, is 0.842, which says that, on average, the change in the yield of the JNJ bond is only 0.842 times the change in the yield of the Treasury bond, which is very different from a parallel shift. The estimate of the constant, $\hat{\alpha}$ is not very different from zero, which is typically the case in regressions of this sort. From an economic perspective, it would be odd if, over an extended period of time, changes in the yield of the JNJ bond tended to be positive or negative when there is no change in the yield of the Treasury bond. The line in Figure 11.1 is the *fitted regression line,* which is Equation (11.2) with its estimated coefficients,

$$\Delta \hat{y}_t^{JNJ} = 0.060 + 0.842 \Delta y_t^{30} \tag{11.6}$$

[1] A linear estimator is linear in the observations of the dependent variable. The expectation of an unbiased estimator of a parameter equals the true value of that parameter. A consistent estimator of a parameter, with enough data, becomes arbitrarily close to the true value of the parameter. And an efficient estimator has the minimum possible variance among linear estimators.

Table 11.2 also gives the standard errors of the constant and slope coefficients, which provide confidence intervals around the estimates: the interval of each estimate plus or minus two standard errors falls around the true parameter values approximately 95% of the time. In this regression, the confidence intervals are 0.060 plus or minus 2 times 0.223, or (–0.386, 0.446), and 0.842 plus or minus 2 times 0.051, or (0.740, 0.944). Hence, because the confidence interval around the estimated constant includes zero, the hypothesis that $\alpha = 0$ cannot be rejected with 95% confidence. But, because the confidence interval for the slope coefficient does not include one, the hypothesis that $\beta = 1$ can be rejected with 95% confidence. Hence, the hypothesis of parallel shifts in the yields of the two bonds is rejected by the data.

Table 11.2 reports that the R-squared of the regression is 77.5%, meaning that 77.5% of the variance of changes in the JNJ yield can be explained by the model, that is, by changes in the yield of the 30-year Treasury bond. In a one-variable regression, the R-squared is just the square of the correlation between the dependent and independent variables, which gives a correlation here of $\sqrt{77.5\%}$, or 88.0%. That these statistics are well below 1.0 indicates that hedging in this case does not come close to eliminating all interest rate risk.

The remaining statistic to be discussed in Table 11.2 is the standard error of the regression, which is essentially the standard deviation of the realized errors or residuals, as defined in Equations (11.3) and (11.4).[2] This standard error is a measure of how well the model fits the data and is in the same units as the dependent variable, in this case, basis points. Roughly speaking, then, the standard deviation of the errors in explaining daily changes in the yield of the JNJ bond with daily changes in the yield of the Treasury bond is two basis points. This statistic is particularly useful in describing the risk of a regression-based hedge, as discussed presently.

Table 11.2 **Regression of Daily Changes in Yields of the JNJ 2.450s of 09/01/2060 on Daily Changes in Yields of the Treasury 1.625s of 11/15/2050, from January 29, 2021, to May 14, 2021.**

No. of Observations	**82**	
R-Squared	**77.5%**	
Standard Error	**2.00**	
Regression Coefficients	**Value**	**Std. Error**
Constant ($\hat{\alpha}$)	0.060	0.223
Chg 30yr Treasury Yield ($\hat{\beta}$)	0.842	0.051

All the results in Table 11.2 are *in-sample;* that is, they are computed from the particular data sample used to estimate the regression model. Relying on these results for hedging assumes that the future will be sufficiently like this particular historical period. The success or failure of this assumption is discussed at the end of the section.

Turning now to regression hedging, assume for the moment that the yield of the JNJ bonds changes by exactly $\hat{\beta}$ basis points for every one-basis-point change in the yield of the Treasury bonds. Let F^{JNJ}, $DV01^{JNJ}$, F^{30}, and $DV01^{30}$ be the face amounts and DV01s of the JNJ and 30-year Treasury bonds, respectively. Then, the position is hedged against changes in yields if,

$$F^{JNJ}\frac{DV01^{JNJ}}{100}\hat{\beta} + F^{30}\frac{DV01^{30}}{100} = 0 \qquad (11.7)$$

$$F^{30} = -F^{JNJ}\frac{DV01^{JNJ}}{DV01^{30}}\hat{\beta} \qquad (11.8)$$

Plugging in numbers, \$100 million for the face amount of the JNJ bonds to be hedged, DV01s from Table 11.1, and $\hat{\beta}$ from Table 11.2,

$$F^{30} = -\$100\text{mm}\frac{0.2124}{0.1910}0.842 = -\$93.6\text{mm} \qquad (11.9)$$

The yield-based DV01 hedge for the JNJ bonds, which is given by Equation (11.9), is given earlier without the slope coefficient of 0.842 as \$111.2 million. The regression hedge of (11.9) sells only \$93.6 million as the yield of the 30-year Treasury bond. Hence, fewer Treasury bonds need be sold to hedge the interest rate risk of the JNJ bonds.

Regression hedges are sometimes described in terms of *risk weights.* Rearranging terms in Equation (11.7) or (11.8),

$$\frac{-F^{30} \times DV01^{30}/100}{F^{JNJ} \times DV01^{JNJ}/100} = \hat{\beta} = 84.2\% \qquad (11.10)$$

In words, the left-hand side of the equation is the DV01 of the hedge as a fraction of the DV01 of the bonds being hedged. The risk weight of a yield-based DV01 hedge is always 100% – the DV01s of the two sides of the trades are, by construction, equal. In this regression hedge, however, the DV01 of the Treasury bonds is only 84.2 % of the DV01 of the JNJ bonds. In general, as Equation (11.10) shows, the risk weight of a regression hedge exactly equals the estimated slope coefficient, $\hat{\beta}$.

The best argument for the regression hedge is actually not the earlier assumption that bond yields change exactly according to the regression model. Write the P&L of the position as,

$$P\&L = -F^{JNJ}\frac{DV01^{JNJ}}{100}\Delta y_t^{JNJ} - F^{30}\frac{DV01^{30}}{100}\Delta y_t^{30} \qquad (11.11)$$

[2] A standard error is actually the sum of the squared residuals divided by the number of observations minus two, while a standard deviation divides by the number of observations minus one. Note than, in a regression with a constant, the average of the residuals is zero by construction.

where the negative signs reflect that a positive face amount with a positive change (i.e., increase) in yield lowers P&L. It can then be shown that the regression hedge in Equation (11.8) minimizes the variance (11.11). (See Appendix A6.1.) In other words, to the extent that P&L variance is an appropriate measure of risk, the regression hedge minimizes the risk of the hedged position.

Appendix A6.1 also derives the standard deviation of the regression-hedged P&L. Denote this standard deviation by $\sigma_{P\&L}$ and the standard deviation of the residuals by σ_ε. Then,

$$\sigma_{P\&L} = \left| \frac{F^{JNJ} DV01^{JNJ}}{100} \right| \sigma_\varepsilon \qquad (11.12)$$

where $|\cdot|$ is the symbol for absolute value, so that the standard deviation is positive whether the original position is long (positive F^{JNJ}) or short (negative F^{JNJ}). In words, the P&L of the hedged position is the DV01 of the position being hedged times the standard error of the regression residuals. Intuitively, the hedged P&L on any given day is exactly zero if the yield of the JNJ bonds moves by 0.842 basis points times the change in the Treasury yield. But if the residual is one basis point, so that the yield of the JNJ bond increases by one basis point more than that, the hedged position loses the DV01 of the JNJ bonds; and if the residual is minus two basis points, then the hedged position gains twice the DV01 of the JNJ bonds; etc. Hence, the volatility of the hedge is proportional to the variability of the residuals.

Applying Equation (11.12) to the case at hand, the DV01 of the JNJ bond position is \$100 million × 0.2124/100, or \$212,400, and the standard error of the regression, reported in Table 11.2, is two basis points per day. Therefore, the standard deviation of the hedged P&L in the sample is \$212,400 × 2 or \$424,800 per day. Whether this is too much risk or not depends on how much and how fast the trader is making money buying the JNJ bonds and hedging them. If the trader is making five basis points on the position and holding it for a day, then a standard deviation of two basis points per day likely represents a reasonable risk–return trade-off. If, on the other hand, the trader is making 1.5 basis points and holding the position for a week, a standard deviation of two basis points per day likely ruins the trade from a risk–return perspective.

This section concludes with an *out-of-sample* analysis of the estimated regression model. Figure 11.2 shows the same regression line as estimated in Table 11.2 and graphed in Figure 11.1. The plus signs, however, are the changes in yields over the period May 17, 2021, to July 19, 2021. The regression model, estimated over the earlier period, January 29, 2021, to May 15, 2021, holds up quite well. In fact, the standard error of the residuals of

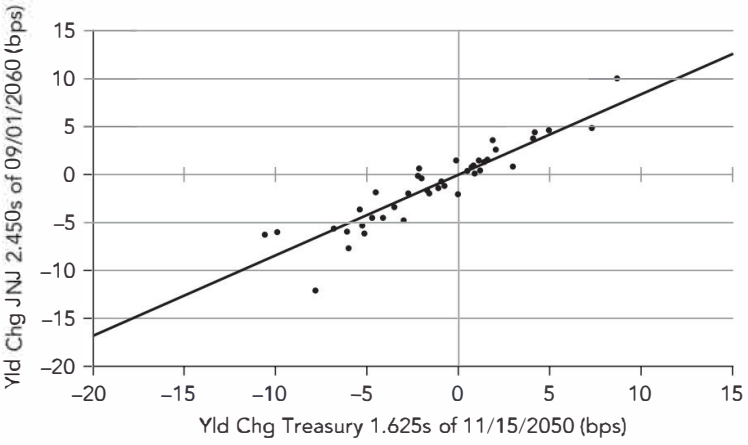


Figure 11.2 **Yield Changes of the JNJ 2.450s of 09/01/2060 and the Treasury 1.625s of 11/15/2050 from May 17, 2021, to July 19, 2021, and the Regression Line Estimated over the Period January 19, 2021, to May 14, 2021.**

the out-of-sample data against the original regression line is 1.5 basis points, which is actually smaller than the in-sample equivalent. The trader hedging as of May 14, 2021, cannot, of course, run this analysis. But other out-of-sample tests can be informative. A trader might see how a regression model performed over a period before the estimation period, perhaps a period right before that window or perhaps an even earlier period that might be more likely to resemble the future. In any case, poor out-of-sample performance should raise questions about the stability of the regression coefficients over time and, therefore, the reliability of the resulting hedge.

11.2 TWO-VARIABLE REGRESSION HEDGING

Hedging approaches generalize to account for the fact that rates across the term structure are not perfectly correlated. Similarly, two-variable regression hedges account for the fact that changes in a bond's yield might be better explained by changes in the yields of two other bonds, rather than just one, as in a one-variable regression.

To illustrate two-variable regression hedging, consider a relative value trader who believes that yields in the 20-year US Treasury bond sector are too high – or prices too low – relative to the 10- and 30-year sectors. Implementing this trade idea by buying a 20-year bond outright is too risky: if rates increase across the board, the trade loses money even if the trader is right, that is, even if the 20-year bond does outperform 10- and 30-year bonds. But buying a 20-year bond and hedging the interest

rate risk by selling a 10-year bond also bears too much risk that is unrelated to the trade idea: if the curve steepens (e.g., 30-year yields increase more than 20-year yields, which increase more than 10-year yields), the trade may lose money even if the 20-year bonds outperform. And, finally, buying a 20-year bond and hedging with a 30-year bond can lose money if the curve flattens even if the 20-year bonds outperform. In practice then, this trade idea is typically implemented with a *butterfly:* buy 20-year bonds and sell both 10- and 30-year bonds: both shorts defend against general rate increases; the 10-year short defends against flattening; and the 30-year short defends against steepening. The trader's problem then becomes to choose the face amount of the 10- and 30-year bonds to sell against, say, $100 million face amount of the 20-year bond.

In this illustration, the trader chooses the three Treasury bonds listed in Table 11.1: the 1.375s of 11/15/2040 as the 20-year; the 0.875s of 11/15/2030 as the 10-year; and the 1.625s of 11/15/2050 as the 30-year. The two-variable regression model of changes in yields of these bonds is,

$$\Delta y_t^{20} = \alpha + \beta^{10}\Delta y_t^{10} + \beta^{30}\Delta y_t^{30} + \varepsilon_t \qquad (11.13)$$

where the notation is analogous to that of the one-variable regression. Here there are two slope coefficients, describing how changes in the 20-year yield are related to changes in each of the 10-year and 30-year yields.

Continuing as in the case of one-variable regression, least-squares estimation finds the regression coefficients so as to minimize the sum of the squared residuals,

$$\sum_t (\Delta y_t^{20} - \hat{\alpha} + \hat{\beta}^{10}\Delta y_t^{10} + \hat{\beta}^{30}\Delta y_t^{30})^2 \qquad (11.14)$$

And, with these estimated coefficients, the predicted change of the 20-year rate is,

$$\Delta \hat{y}_t^{20} = \hat{\alpha} + \hat{\beta}^{10}\Delta y_t^{10} + \hat{\beta}^{30}\Delta y_t^{30} \qquad (11.15)$$

Table 11.3 gives the results of the regression, estimated with data from January 29, 2021, to May 14, 2021. The R-squared is quite high relative to that of the single-variable regression, in Table 11.2, in part because two explanatory variables are used, rather than one, and in part because all of the bonds in this regression are Treasuries, whereas the single-variable regression mixes a corporate bond with a Treasury bond. The standard error is also significantly lower here, at 1.15 basis points per day. Again, however, as usual for regressions of this sort, the estimate of $\hat{\alpha}$ is small and not significantly different from zero.

The slope coefficients say that a one-basis-point increase in the 10-year yield increases the 20-year yield by 0.465 basis points, while a one-basis-point increase in the 30-year yield increases

Table 11.3 **Regression of Daily Changes in Yields of the Treasury 1.375s of 11/15/2040 on Daily Changes in Yields of the Treasury 0.875s of 11/15/2030 and l.625s of 11/15/2050, from January 29, 2021, to May 14, 2021.**

No. of Observations	**82**	
R-Squared	**94.7%**	
Standard Error	**1.15**	
Regression Coefficients	**Value**	**Std. Error**
Constant ($\hat{\alpha}$)	0.019	0.129
Chg 10yr Treasury Yield ($\hat{\beta}^{30}$)	0.465	0.068
Chg 30yr Treasury Yield ($\hat{\beta}^{10}$)	0.669	0.067

the 20-year yield by 0.669 basis points. With 95% confidence intervals for these coefficients of (0.329, 0.601) and (0.535, 0.803), respectively, both coefficients are significantly different from zero; that is, changes in the yields of both bonds are useful in explaining changes in the yield of the 20-year bond.

To derive the hedge based on these regression results, write the P&L of the hedged position as,

$$P\&L = -\frac{F^{20}DV01^{20}}{100}\Delta y_t^{20} - \frac{F^{10}DV01^{10}}{100}\Delta y_t^{10} - \frac{F^{30}DV01^{30}}{100}\Delta y_t^{30} \qquad (11.16)$$

and then substitute for Δy_t^{20} from (11.15) to see that,

$$P\&L = \left[-\frac{F^{20}DV01^{20}}{100}\hat{\beta}^{10} - \frac{F^{10}DV01^{10}}{100}\right]\Delta y_t^{10} + \left[-\frac{F^{20}DV01^{20}}{100}\hat{\beta}^{30} - \frac{F^{30}DV01^{30}}{100}\right]\Delta y_t^{30} \qquad (11.17)$$

Next, to ensure that P&L is zero, under the assumption that the change in the 20-year rate follows the regression model, set each of the terms in brackets in Equation (11.17) equal to zero. Solving,

$$F^{10} = -\frac{F^{20}DV01^{20}}{DV01^{10}}\hat{\beta}^{10} \qquad (11.18)$$

$$F^{30} = -F^{20}\frac{DV01^{20}}{DV01^{30}}\hat{\beta}^{30} \qquad (11.19)$$

or, in terms of risk weights,

$$\frac{-F^{10} \times DV01^{10}}{F^{20}DV01^{20}} = \hat{\beta}^{10} \qquad (11.20)$$

$$\frac{-F^{30} \times DV01^{30}}{F^{20}DV01^{20}} = \hat{\beta}^{30} \qquad (11.21)$$

Assuming a trade size of $100 million face amount in the 20-year Treasury, substituting the DV01s of the bonds from Table 11.1 and the results of the regression from Table 11.3, the hedging

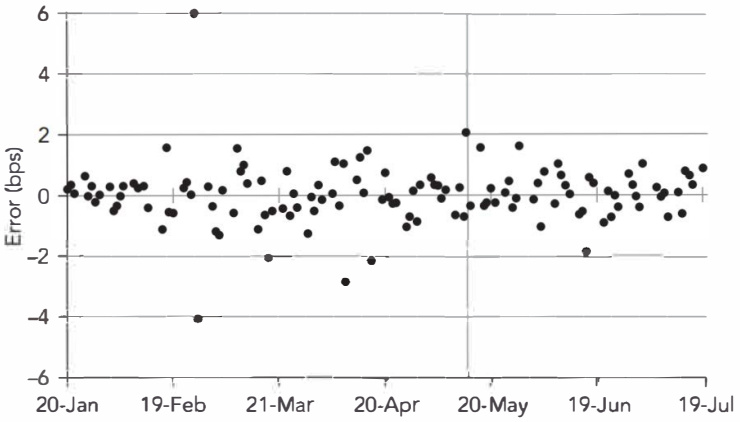


Figure 11.3 **Residuals Using the Regression Coefficients in Table 11.3, in-Sample – from January 19, 2021, to May 14, 2021 – and Out-of-Sample – from May 17, 2021, to July 19, 2021.**

face amounts and risk weights are $79.44 million and 46.5% for the 10-year, along with $50.69 million and 66.9% for the 30-year. Note that the sum of the risk weights is 113.4%, which means that the DV01 of the hedging position is 13.4% greater than the DV01 of the position being hedged. This follows immediately from the slope coefficients of the regression: a simultaneous one-basis-point change in both the 10- and 30-year yields is associated with a 1.134-basis-point increase in the 20-year yield. Hence, the hedging portfolio requires an extra 13.4% in DV01.

Figure 11.3 compares in-sample and out-of-sample residuals from the regression in Table 11.3. The out-of-sample residuals are very well behaved, in fact, better behaved than the in-sample residuals: the standard error is 1.15 in-sample, and only 0.70 out-of-sample. Traders are not always so fortunate!

11.3 LEVEL VERSUS CHANGE REGRESSIONS

When estimating regression-based hedges, most practitioners regress changes in yields on changes in yields, as in the previous sections, but some regress yields on yields. Mathematically, in the single-variable case, the level-on-level regression with dependent variable y and independent variable x is,

$$y_t = \alpha + \beta x_t + \varepsilon_t \tag{11.22}$$

while the change-on-change regression is,[3]

$$y_t - y_{t-1} = \beta(x_t - x_{t-1}) + \varepsilon_t - \varepsilon_{t-1} \tag{11.23}$$

$$\Delta y_t = \beta \Delta x_t + \Delta \varepsilon_t \tag{11.24}$$

[3] While it is usual to include a constant term in the change-on-change regression, the constant is omitted here for expositional purposes.

If the assumptions of least-squares estimation, mentioned earlier, hold true for the level model (11.22), then they also hold for the change model (11.24), and least-squares estimates from both specifications are unbiased, consistent, and efficient. If, however, the assumption about the independence of the error terms is violated in either specification, then the least-squares estimates from that specification may not be efficient, but they are still unbiased and consistent.

To discuss the economics behind the assumption that error terms are independent, say that $\alpha = 0$, that $\beta = 1$, and that y and x are the yields of two different bonds. Say further that the yield of the x-bond is constant at 5%, while the yield of the y-bond was 1% yesterday. The level regression, in Equation (11.22), predicts that the yield of the y-bond will be 5% today, despite its having been 1% yesterday. It is more likely, however, that the yield of the y-bond today will be closer to 1% than to 5%, and that the model error today will be closer to its value yesterday, of −4%, than to zero. In other words, the error terms of the level regression are not likely to be independent of each other, but rather persistent, correlated over time, or *serially correlated.*

In this same scenario, because the change in the yield of the x-bond is zero, the change-on-change regression in Equation (11.24) predicts that the change in the yield of the y-bond is zero as well and that its yield remains at 1%. While more plausible than the level-on-level prediction that the yield of the y-bond suddenly jumps to 5%, it is more likely that the yield of the y-bond will gradually trend from its current value of 1% to its model value of 5%. Hence, the error terms in the change-on-change regression are likely to be positive for some time, and, as such, serially correlated.

This discussion suggests an alternate model, which would capture, in the scenario just discussed, that the yield of the y-bond moves gradually from 1% to 5%. In particular, assume the level-on-level model, but with error dynamics,

$$\varepsilon_t = \rho \varepsilon_{t-1} + \nu_t \tag{11.25}$$

for some $\rho < 1$. In this model, with say, $\rho = 75\%$, yesterday's error of −4% would fall, on average, to an error of 75% times −4%, or −3% today, thus giving an expected y-bond yield today of 2%. In this way, the error structure in Equation (11.25) gradually pushes the yield of the y-bond up from its starting point of 1% to its model value, that is, the 5% yield of the x-bond. The procedure for estimating (11.22) with the error structure in (11.25) is given in many statistical texts.

11.4 REVERSE REGRESSIONS

In Section 11.1, a trader regresses changes in yields of the JNJ 2.450s of 09/01/2060 – with a DV01 of 0.2124 – on changes in yields of the Treasury 1.625s of 11/15/2050 – with a DV01 of 0.1910;

obtains a regression coefficient of 0.842; and, against $100 million of the JNJ bonds, calculates a regression hedge to sell $100 million × (0.2124/0.1910) × 0.842, or $93.6 million Treasury bonds.

What if another trader runs the *reverse* regression, that is, regresses changes in yield of the Treasury bond on changes in yield of the JNJ bond? Table 11.4 compares the slope coefficients and standard errors of the original regression and the reverse regression. With a reverse regression $\hat{\beta}$ of 0.921, this second trader hedges $93.6 million Treasury bonds with $93.6 million × (0.1910/0.2124) × 0.921, or $82.8 million JNJ bonds.

These hedges are clearly different. The same $93.6 million of Treasuries are hedged with a different amount of JNJ bonds. Or, in terms of risk weights, both quoted as the DV01 of the Treasury bond position as a percent of the DV01 of the JNJ position, the risk weight of the regression is 84.2%, while the risk weight of the reverse regression is 1/0.921, or 108.6%. Is one of these hedges right and the other wrong?

This question actually reveals the importance of the trader's decision in Section 11.1 to hedge $100 million face amount of JNJ bonds. Choosing this face amount actually sets the risk of the trade. As shown earlier, the volatility of the hedged position is $100 million × 0.2124/100 × the two-basis-point standard error of the regression, or about $425,000. However, the risk of the reverse regression, which sets the face amount of the Treasury bonds at $93.6 million, is $93.6 million × 0.1910/100 × the 2.09 standard error of the reverse regression, or about $374,000. These trades, therefore, are not comparable.

Choosing to hedge $100 million face amount of the JNJ bonds, however, is not just about the risk of the hedged position. There are many combinations of positions in the JNJ and Treasury bonds that have the same volatility.[4] For example, a scaled-up reverse regression hedge, with $106.37 million Treasury bonds and $88 million JNJ bonds (i.e., $106.37 million × (0.1910/0.2124) × 0.921), has the same $425,000 volatility as the regression hedge ($106.37 million × 0.1910/100 × 2.09). But while this and other positions might have the same volatility, because they do not hold $100 million in JNJ bonds, they are different trades. Most obviously, they do not satisfy the objective of buying $100 million of JNJ bonds from a client. Less obviously, to the extent that the return profile of the JNJ and Treasury bonds differ, different portfolios of the two bonds have different return characteristics as well.

In short, the regression hedge in Section 11.1 minimizes the variance of hedging $100 million face amount of the JNJ bonds. Trades with other objectives, like holding a fixed amount of Treasuries or holding a fixed amount of volatility risk with particular return characteristics, are constructed differently.

Table 11.4 **Regression: Daily Changes in Yields of the JNJ 2.450s of 09/01/2060 on Daily Changes in Yields of the Treasury 1.625s of 11/15/2050. Reverse Regression: Daily Changes in Yields of the Treasury 1.625s of 11/15/2050 on Daily Changes in Yields of the JNJ 2.450s of 09/01/2060. Observations Are from January 29, 2021, to May 14, 2021.**

	Regression	Reverse Regression
$\hat{\beta}$	0.842	0.921
Standard Error	2.00	2.09

[4] See Equation (A6.15) in Appendix A6.1, which expresses the variance of the P&L as a quadratic in the DV01s of each position.

11.5 PRINCIPAL COMPONENT ANALYSIS

As mentioned in the introduction to this chapter, regression hedging tends to be *ad hoc*, because the relevant bonds and estimation periods are chosen separately for each application. *Principal component analysis* is useful, by contrast, in providing a single, empirical description of the behavior of the term structure that can be applied across a portfolio of fixed income instruments.

To illustrate PCA, this section uses daily data on fixed versus three month US Dollar (USD) LIBOR swap rates from June 1, 2020, to July 16, 2021.[5] The data set consists of 13 times series, one for each of the terms from one to 10 years, as well as three with terms of 15, 20, and 30 years. These data can be summarized by the variances or standard deviations of changes in each rate and with their pairwise covariances or correlations. Another way to describe the data, however, is with 13 interest rate factors or components, where each factor represents a particular pattern of changes across the 13 rates. One factor, for example, might represent a simultaneous change of 0.2 basis points in the one-year rate, 0.6 basis points in the two-year rate, 1.2 basis points in the three-year rate, etc., up to 3.7 basis points in the 20-year rate, and 3.8 basis points in the 30-year rate. PCA is a way to construct 13 factors, or *principal components* (PCs), such that they have the following properties:

1. The sum of the variances of the PCs equals the sum of the variances of the individual rates. In this sense, the PCs capture the volatility of the set of interest rates.

[5] LIBOR swaps are being phased out at the time of this writing, but a sufficiently long time series of liquid SOFR swap rates is not yet available for the analysis of this section.

2. The PCs are uncorrelated with each other. While changes in rates of one term are highly correlated with changes in rates of another term, the PCs are constructed so that they are uncorrelated.
3. Subject to (1) and (2), each PC is chosen to have the maximum possible variance given all earlier PCs. Therefore, the first PC explains the largest fraction of the sum of the variances of the rates; the second PC explains the next largest fraction; and so forth.

PCs of rates are particularly useful because of an empirical regularity: the sum of the variances of the first three PCs is usually an overwhelming fraction of the sum of the variances across all rates. Therefore, the variances and covariances of all rates are not necessary to describe how the term structure fluctuates: the structure and volatilities of only three PCs suffice. In the simplified context of three rates, Appendix A6.2 describes the construction of PCs in more detail. The text continues with a discussion of computed PCs for USD LIBOR swaps from the data set described earlier.

Figure 11.4 graphs the first three principal components, while Table 11.5 provides similar information in tabular form. Columns (2) to (4) correspond to the three PC curves in the figure, which can be interpreted as follows. A one standard deviation increase in the "level" PC, given both in Column (2) and the solid line in the figure, is a simultaneous increase in the one-year rate of 0.23 basis points; in the

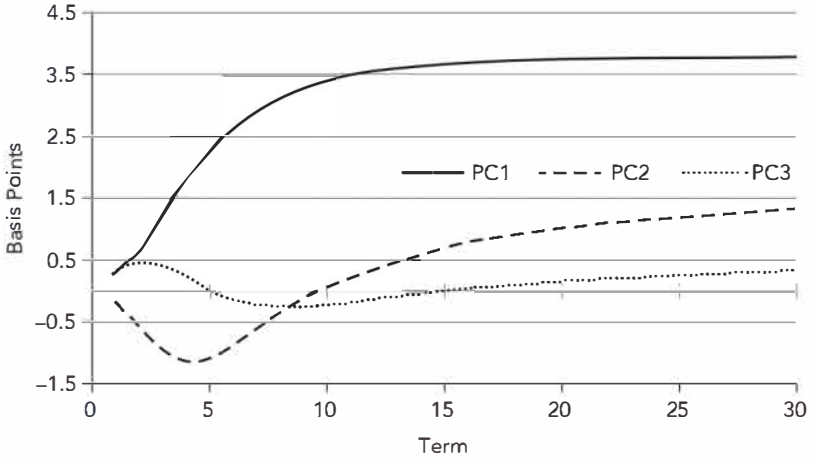


Figure 11.4 The First Three Principal Components of USD LIBOR Swap Rates, Estimated from June 1, 2020, to July 16, 2021.

Table 11.5 Principal Component Analysis of USD LIBOR Swap Rates from June 1, 2020, to July 16, 2021. Columns (2)-(6) Are in Basis Points; Columns (7)-(10) Are in Percent.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCs					% of PC Variance			
Term	Level	Slope	Short Rate	PC Vol	Total Vol	Level	Slope	Short Rate	(5)/(6)
1	0.23	−0.16	0.29	0.41	0.55	32.7	15.0	52.3	74.54
2	0.59	−0.51	0.47	0.91	0.93	42.3	31.5	26.2	97.47
3	1.18	−0.99	0.42	1.59	1.60	54.5	38.5	7.0	99.63
4	1.77	−1.16	0.23	2.13	2.13	69.1	29.7	1.2	99.72
5	2.28	−1.12	0.02	2.54	2.54	80.6	19.4	0.0	99.78
6	2.64	−0.89	−0.13	2.79	2.79	89.7	10.1	0.2	99.91
7	2.94	−0.60	−0.20	3.01	3.01	95.6	4.0	0.4	99.97
8	3.14	−0.36	−0.24	3.17	3.17	98.2	1.3	0.6	99.96
9	3.31	−0.13	−0.25	3.32	3.32	99.3	0.2	0.6	99.92
10	3.44	0.07	−0.23	3.44	3.45	99.5	0.0	0.4	99.87
15	3.65	0.71	0.00	3.72	3.72	96.3	3.7	0.0	99.98
20	3.73	1.02	0.17	3.87	3.87	92.8	7.0	0.2	99.98
30	3.77	1.35	0.36	4.02	4.02	87.9	11.3	0.8	99.93
Total	9.99	2.92	0.96	10.45	10.47	91.3	7.8	0.8	99.84

two-year rate of 0.59 basis points; in the 10-year rate of 3.44 basis points; and in the 30-year rate of 3.77 basis points. On a particular day, the change in the term structure might be best explained as a 1.5 standard deviation move in the first PC, that is, as adding 1.5 times each element of the first PC to corresponding swap rates. On another day, the change in the term structure might best be described as a −0.75 standard deviation move in the first component, that is, as subtracting 0.75 times each element of the first PC from current rates. In any case, the first PC has been traditionally called the "level" component because it has typically represented an approximately parallel shift over much of its range. In the empirical results presented here, however, the component is not particularly level for terms from one to seven years.

A one standard deviation increase in the "slope" PC, given both in Column (3) of the table and the dashed line in the figure, is a simultaneous fall in the one-year rate of 0.16 basis points; a fall in the two-year rate of 0.51 basis points; an increase in the 10-year rate of 0.07 basis points; and an increase in the 30-year rate of 1.35 basis points. This PC is said to represent a "slope" change in rates because shorter-term rates fall while longer-term rates increase, or *vice versa*.

Lastly, a one standard deviation increase in the "short-rate" PC, given both in Column (4) of the table and the dotted line in the figure, is a simultaneous small increase of very short-term rates; a small decrease in intermediate-term rates, and a small increase in long-term rates. Because of its shape across terms, this PC is often named "curvature" as well, but, in light of the full discussion in this section, this PC is particularly useful for adding explanatory power to variations in shorter-term rates.

To reiterate the sense in which the PCs describe changes in the term structure, on a given day, changes across terms might be described – picking numbers at random – as the combination of: a +1.5 standard deviation change in the first PC; a −0.4 standard deviation change in the second PC; and a −1.8 standard deviation change in the third PC. The term structure at the end of that day, therefore, would approximately equal the term structure at the end of the previous day plus the contributions from the multiples of each of the three PCs. In this way, as explained shortly, these three PCs can indeed explain an overwhelmingly large proportion of realized term structure volatility.

The small values of the PCs at very short-term rates reflect the low volatility of these rates. In the current financial environment, with the Federal Reserve promising to keep short-term rates low for an extended period of time, current economic shocks are not envisioned as impacting short-term rates until some time in the future. As a result, economic volatility is not reflected in very short-term rates but seeps gradually into intermediate- and longer-term rates as expectations of reactions to future Federal Reserve policy actions.[6]

Column (5) of Table 11.5 gives the combined standard deviation or volatility from the three principal components for each rate, while Column (6) gives the total, empirical volatility of each rate over the sample period. For the five-year rate, for example, recalling that PCs are, by construction, uncorrelated, the volatility from the three PCs is,

$$\sqrt{2.28^2 + (-1.12)^2 + 0.02^2} = 2.54 \qquad (11.26)$$

The total volatility of the five-year rate in the sample is also, to two decimal places, 2.54, but Column (10) – using more decimal places than shown in Columns (5) and (6) – reports that the ratio of five-year PC volatility to total volatility is 99.78%. Hence, the empirical variation of the five-year rate is almost completely explained by the first three PCs. Considering Column (10) as a whole, three PCs explain over 99% of the variation of rates of all terms greater than three years, 97.47% of the variation in the two-year rate, and 74.54% of the variation in the one-year rate. Hence, although there are 13 rates in the data series, three factors alone – three fixed combinations of changes in rates across terms – go a very long way in explaining the variation in all 13 rates. This is possible, intuitively, because changes in rates across terms are highly correlated; that is, nowhere near 13 factors are actually necessary to explain the variation in 13 rates. The performance of the three factors is less impressive, however, for rates of the shortest term.

Columns (7) through (9) of Table 11.5 give the variance explained by each of the first three PCs as a fraction of the total variance explained by those three PCs. For the two-year rate, for example, those fractions are calculated as follows,

$$\frac{0.5916^2}{0.9091^2} = 42.3\% \qquad (11.27)$$

$$\frac{(-0.5101)^2}{0.9091^2} = 31.5\% \qquad (11.28)$$

$$\frac{0.4650^2}{0.9091^2} = 26.2\% \qquad (11.29)$$

Note that, to avoid confusion, the values in these equations are reported to greater accuracy than those in the table.

[6] For many years before the financial crisis of 2007–2009, the first PC was hump-shaped, increasing to a peak at about five years or so, and then declining gradually over longer terms. This shape was interpreted as the Federal Reserve pegging very short-term rates but responding in the relatively near term to economic volatility. Also, because current economic volatility affects views on longer-term rates less and less with term, volatility eventually begins to decline with term. Current Federal Reserve policy, however, as discussed in the text, seems to have changed dramatically this empirical feature of rates markets.

Looking at Columns (7) to (9) across terms, the level PC dominates the other two as the main contributor to variations in the term structure. The short-rate PC, and then the slope PC, are significant contributors at the short end, however, as is the slope PC for the longest rates. These finding have significant implications for risk management. A trader of eight- to 10-year swaps, or perhaps even seven- to 15-year swaps, can defend using the one-factor metrics and hedging approaches and one-variable regression hedging described in this chapter: according to Table 11.5, in this range of terms, term structure variation can be well described by a one-factor model, like the level PC. On the other hand, traders in the three- to six-year sector or the long end might very well require two factors, while traders in the very short end may not be comfortable without a three-factor framework.

The last row of Table 11.5 computes the various statistics just discussed across the whole term structure of rates. More specifically, Columns (2) to (6) give the square root of the sum of variances across terms, and Columns (7) to (9) give the respective ratios for these totals. While the sum of variances is not a particularly interesting economic quantity – it does not represent the variance of any particularly interesting portfolio – the last row of the table does summarize two overall results of the PCA. First, 99.84% of the volatility of the 13 rates in the study is explained by the first three PCs. Second, the level PC explains over 90% of that variance, the slope PC about 8%, and the short-rate PC less than 1%.

The overall lessons from PCA are often similar across global markets. Estimated over the same time period as the USD PCs in Figure 11.4, Figures 11.5 and 11.6 graph the first three PCs of British Pound Sterling (GBP) LIBOR and Euribor swap rates, respectively. The GBP PCs are extremely similar to their USD equivalents, with respect to both shape and magnitude. The biggest difference seems to be that the slope PC in USD is more volatile. The shapes of the EUR PCs are qualitatively similar to those in USD and GBP, but volatility in EUR is significantly lower. The EUR level PC, for example, flattens in the long end at a bit above 2.5 basis points per day, whereas the USD and GBP level PCs flatten at over 3.5. This lower volatility might be explained by the current aggressiveness of the European Central Bank, relative to the Federal Reserve and the Bank of England, to keep short-term rates low over an extended period of time.

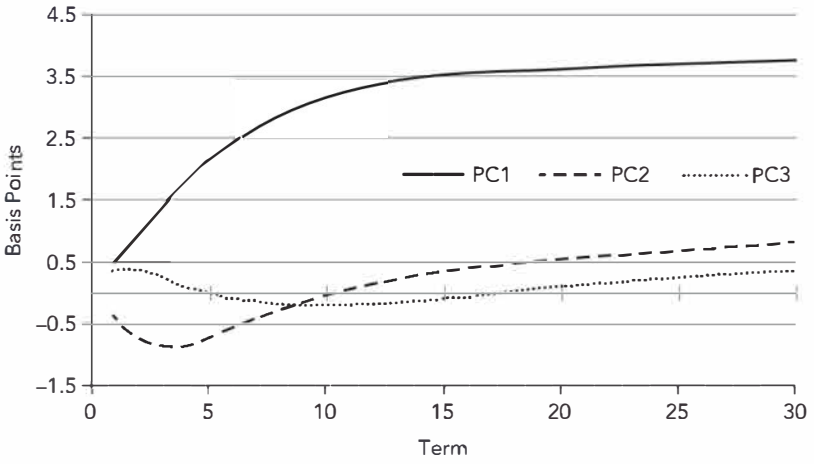


Figure 11.5 **The First Three Principal Components of GBP LIBOR Swap Rates, Estimated from June 1, 2020, to July 16, 2021.**

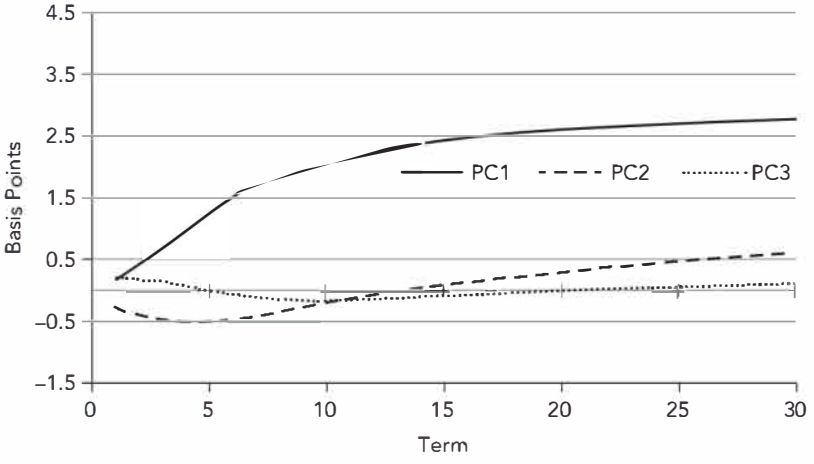


Figure 11.6 **The First Three Principal Components of Euribor Swap Rates, Estimated from June 1, 2020, to July 16, 2021.**

Hedges based on PCA are constructed like multi-factor approaches. Using the current term structure, calculate the current price of the portfolio being hedged; shift the current term structure by each PC, in turn, to get new term structures and new portfolio prices; with these new prices, calculate a portfolio '01 with respect to each PC; and find a portfolio of hedging securities that neutralizes these '01s.

This section illustrates hedging with PCA in a context described earlier, namely, hedging a relative value 10s-20s-30s butterfly, but this time in swaps. Specifically, a trader believes that USD 20-year swap rates are too high relative to 10- and 30-year swap rates, and plans, therefore, to receive fixed in 20s and pay in 10s

Table 11.6 **USD LIBOR Par Swap Rates and DV01s, as of July 16, 2021, and PC Elements from Table 11.5. Rates Are in Percent, and PC Elements Are in Basis Points.**

Term	Rate	DV01	Level	Slope	Short Rate
10	1.303	0.09347	3.44	0.07	−0.23
15	1.501	0.13387	3.65	0.71	0.00
20	1.596	0.17064	3.73	1.02	0.17
30	1.646	0.23600	3.77	1.35	0.36

and 30s.[7] Par swap rates and DV01s of par swaps are given in Table 11.6. Also, corresponding to rates of the listed terms, the table gives the elements of each of the three PCs from Table 11.5. By definition, a one standard deviation shift in each PC changes these swap rates by these PC elements. The inclusion of the 15-year swap rate is discussed later.

Assume that the trader plans to receive fixed on 100 notional amount of the 20-year swap and on F^{10} and F^{30} notional amount of 10- and 30-year swaps, respectively. Paying fixed is reflected, in this notation, with negative notional amounts. In any case, from the data in Table 11.6, the exposure of this overall relative value portfolio is hedged against a one standard deviation shift of the level and slope PCs, respectively, if the following equations obtain,

$$-F^{10}\frac{0.09347}{100}\times 3.44 - F^{30}\frac{0.23600}{100}\times 3.77 - 100\frac{0.17064}{100}\times 3.73 = 0 \quad (11.30)$$

$$-F^{10}\frac{0.09347}{100}\times 0.07 - F^{30}\frac{0.23600}{100}\times 1.35 - 100\frac{0.17064}{100}\times 1.02 = 0 \quad (11.31)$$

The first two terms of Equation (11.30) give the change in the value of the hedge position under a one standard deviation shift of the first PC, that is, a shift of +3.44 basis points in the 10-year and +3.77 basis points in the 30-year swap rate. The third term gives the change in the value of the position being hedged under the same PC shift, which is +3.73 basis points in the 20-year swap rate. Note that the negative signs indicate that receiving fixed (i.e., positive notional amounts) when rates increase results in position losses. The equation as a whole, therefore, sets the total position gain or loss under a one standard deviation shift of the first PC equal to zero. Equation (11.31) can be interpreted similarly, but under a one standard deviation shift of the second PC. Note, of course, that if these two equations hold for a one standard deviation shift, they hold for any size shift: to see this, simply multiply both sides of each equation by the intended number of standard deviations.

Solving Equations (11.30) and (11.31) reveals that $F^{10} = -49.56$ and $F^{30} = -53.60$. Or, in terms of risk weights relative to the DV01 of the 20-year swap,

$$\frac{49.56\times\frac{0.09347}{100}}{0.17064} = 27.1\% \quad (11.32)$$

$$\frac{53.60\times\frac{0.23600}{100}}{0.17064} = 74.1\% \quad (11.33)$$

Intuitively, most of the risk of the 20-year swap – 74% – is hedged with 30-year swaps, because the exposures of 20-year swaps to both the level and slope PCs more closely resemble those of 30-year swaps than of 10-year swaps. Note also that the sum of the risk weights is 101.2%, so that the DV01 of the hedge position is greater than the DV01 of the position being hedged. Only under the assumption of parallel shifts do the risk weights always sum to 100%. In the present case, more DV01 risk is needed in the hedge because the hedge includes a significant amount of 10-year swaps, which are much less sensitive to the level and slope shifts than the 20-year swaps.

In this illustration, the trader chooses to hedge with 10- and 30-year swaps. But, with only two hedging securities, the risks of only two PCs can be hedged. What is the risk of the hedged position just derived to the next most important PC, that is, the short-rate or curvature PC? Following the same logic as in Equations (11.30) and (11.31), the exposure of the hedged position to the third PC (adding a significant digit to avoid confusion) is,

$$-(-49.6)\frac{0.09347}{100}\times(-0.228) - (-53.6)\frac{0.23600}{100}\times 0.360 - 100\frac{0.17064}{100}\times(0.166) = 0.007 \quad (11.34)$$

which is less than one cent per 100 face amount. The trader might very well decide, therefore, that it is not worth the transaction costs of trading an additional swap to hedge out this residual risk from the third PC. Also, because this is a relative value trade, the trader wants to pay fixed only in swaps with rates that are believed to be too low. In any case, if hedging out the residual risk is desired, a 15-year swap can be added to the hedging portfolio; an equation for exposure to the third PC can be added to Equations (11.30) and (11.31); and, using the data from Table 11.6, the notional amounts for the 10-, 15-, and 30-year swaps can be determined. This is left as an exercise for the reader.

[7] The reader can think of this trade as "buying" 20-year bonds and "selling" 10- and 30-year bonds.

12

Arbitrage Pricing with Term Structure Models

Learning Objectives

After completing this reading, you should be able to:

- Calculate the expected discounted value of a zero-coupon security using a binomial tree.
- Construct and apply an arbitrage argument to price a call option on a zero-coupon security using replicating portfolios.
- Define risk-neutral pricing and apply it to option pricing.
- Explain the difference between true and risk-neutral probabilities and apply this difference to interest rate drift.
- Explain how the principles of arbitrage pricing of derivatives on fixed-income securities can be extended over multiple periods.
- Define option-adjusted spread (OAS) and apply it to security pricing.
- Describe the rationale behind the use of recombining trees in option pricing.
- Calculate the value of a constant-maturity Treasury swap, given an interest rate tree and the risk-neutral probabilities.
- Evaluate the advantages and disadvantages of reducing the size of the time steps on the pricing of derivatives on fixed-income securities.
- Evaluate the appropriateness of the Black-Scholes-Merton model when valuing derivatives on fixed-income securities.

Excerpt is Chapter 7 of Fixed Income Securities: Tools for Today's Markets, *Fourth Edition, by Bruce Tuckman and Angel Serrat.*

Principal components analysis reveals that the term structure of interest rates is determined by relatively few factors or random processes. Therefore, assumptions about how these few factors evolve over time, combined with arbitrage arguments, can deliver strong predictions about the prices and interest rate sensitivities of bonds and other *interest rate contingent claims* (i.e., securities with cash flows that depend on interest rates, like bond options). Formulating assumptions about the evolution of interest rate factors, pricing fixed income securities, and determining hedge ratios comprise the art and science of *term structure models*.

This chapter uses a very simple setting to show how assumptions about the evolution of the short-term rate over time allows for the arbitrage pricing of bonds of all maturities and of interest rate contingent claims. *Option-adjusted spread (OAS)* is introduced both as a metric of a security's mispricing relative to a model and as the spread that can be earned – if the model is correct – by trading that security. Chapter 13 shows how the shape of the term structure is determined by: expectations about future short-term rates, the risk premium required by investors to bear interest rate risk, and convexity, whose effect is a result of interest rate volatility. Chapter 16 then illustrates the art of modeling the evolution of short-term rates by presenting two term structure models: the classic Vasicek model and the two-factor Gauss+ model, which has proven popular in industry for both relative value and macro-style trading.

12.1 RATE AND PRICE TREES

Assume that the six-month and one-year spot rates are 2% and 2.15%, respectively. Taking these market rates as given is equivalent to taking the prices of a six-month bond and a one-year-bond as given. Securities with assumed prices are called *underlying securities* to distinguish them from the contingent claims priced by arbitrage arguments.

Next, assume that six months from now the six-month rate is either 2.50% or 1.50% with equal probability. This very strong assumption is depicted at the top of Figure 12.1 by means of a *binomial tree*, where "binomial" means that only two future values are possible. The columns in the tree represent dates. The six-month rate is 2% today, which is called date 0. Six months from now, on date 1, there are two possible outcomes or *states of the world*. The 2.50% state is called the *upstate* while the 1.50% state is called the *downstate*.

Given the current term structure of spot rates (i.e., the current six-month and one-year rates), trees can be computed for the prices of six-month and one-year zero coupon bonds. The price

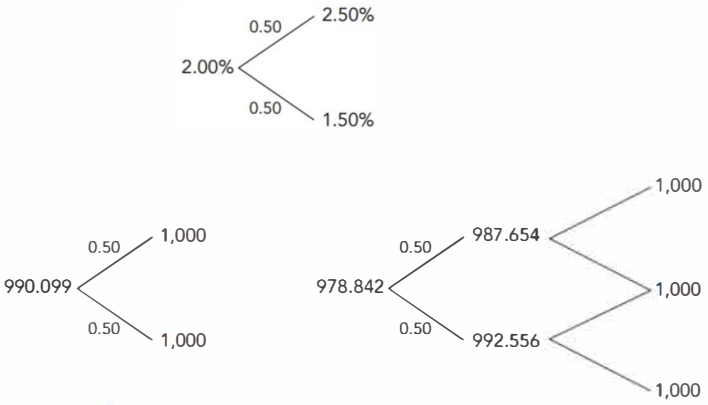


Figure 12.1 **Pricing six-month and one-year zero coupon bonds with a binomial rate tree.**

tree for $1,000 face value of the six-month zero, depicted at the bottom left of Figure 12.1, shows that the date-0 price is $1,000/ (1 + 0.02/2) = $990.099. (For readability, currency symbols are not included in price trees.) Note that, in a tree for the value of a particular security, the maturity of the security falls over time. On date 0 of the tree just discussed, the security is a six-month zero, while on date 1 the security is a maturing zero.

The price tree for $1,000 face value of a one-year zero is depicted at the bottom right of Figure 12.1. The three prices on date 2 are all $1,000, which is the face value of the one-year zero. The two prices on date 1 are found by discounting this certain $1,000 at the then-prevailing six-month rate. Hence, the date 1 upstate price is $1,000/ (1 + 0.025/2), or $987.654, and the date 1 downstate price is $1,000/ (1 + 0.015/2), or $992.556. Finally, the date 0 price is computed using the given, date 0, one-year rate of 2.15%: $1,000/$(1 + 0.0215/2)^2$, or 978.842.

The probabilities of moving up or down the tree may be used to compute average or expected values. As of date 0, the expected value of the one-year zero price on date 1 is,

$$0.5 \times \$987.654 + 0.5 \times \$992.556 = \$990.105 \quad (12.1)$$

Discounting this expected value to date 0, at the date 0, six-month rate gives an *expected discounted value* of,

$$\frac{0.5 \times \$987.654 + 0.5 \times \$992.556}{1 + \frac{.02}{2}} = \$980.302 \quad (12.2)$$

Note that the one-year zero's expected discounted value of $980.302 is not equal to its market price of $978.842. These two numbers need not be equal, because investors do not price securities by expected discounted value. Over the next six months, the one-year zero is a risky security, worth $987.654 half of the

time and \$992.556 the other half of the time, for an average or expected value of \$990.105. If investors do not like this price uncertainty, they would prefer a security worth \$990.105 on date 1 with certainty. More specifically, a security worth \$990.105 with certainty after six months would sell for \$990.105/(1 + .02/2), or \$980.302, as of date 0. By contrast, investors penalize the risky one-year zero coupon bond with an average price of \$990.105 in six months by pricing it today at \$978.842. Chapters 13 and 16 elaborate further on investor *risk aversion*.

12.2 ARBITRAGE PRICING OF DERIVATIVES

This section prices an interest rate contingent claim or derivative, in particular, a call option that expires in six months to purchase \$1,000 face value of a then six-month zero at \$990. Figure 12.2 starts the price tree for this call option based on the rates and prices in Figure 12.1. If on date 1 the six-month rate is 2.50%, and a six-month zero sells for \$987.654, the right to buy that zero at \$990 is worthless. On the other hand, if the six-month rate is 1.50%, and the price of a six-month zero is \$992.556, then the right to buy the zero at \$990 is worth \$992.556 − \$990, or \$2.556. This description of the option's terminal payoffs emphasizes the contingent claim nature of the option: its value depends on interest rates through the value of an underlying bond.

A security is priced by arbitrage by finding and pricing its replicating portfolio. In that context, because all bond cash flows are fixed or constant, the construction of the replicating portfolio is relatively simple. The present context is more difficult, because cash flows do depend on the level of rates, and the replicating portfolio must replicate the contingent claim for any possible interest rate scenario.

To price the call option of this section by arbitrage, construct a portfolio on date 0 of underlying securities, namely six-month and one-year zero coupon bonds, such that the portfolio is worth \$0 in the upstate on date 1 and \$2.556 in the downstate. Let $F^{.5}$ and F^{1} be the face values of six-month and one-year zeros in this replicating portfolio, respectively, and recall that the possible values of these bonds on date 1 are shown in Figure 12.1. These face amounts, therefore, must satisfy the following two equations,

$$F^{.5} + .987654F^{1} = \$0 \tag{12.3}$$

$$F^{.5} + .992556F^{1} = \$2.558 \tag{12.4}$$

Equation (12.3) may be interpreted as follows. In the upstate, the value of the replicating portfolio's now maturing six-month zero is its face value. The value of the once one-year zeros, now six-month zeros, is .987654 per dollar face value. Hence, the left-hand side of Equation (12.3) denotes the value of the replicating portfolio in the upstate. This value must equal \$0, the value of the option in the upstate. Similarly, Equation (12.4) sets the value of the replicating portfolio in the downstate equal to the value of the option in the downstate.

Solving Equations (12.3) and (12.4) gives $F^{.5} = -\$515.0000$ and $F^{1} = \$521.4375$. In words, the option can be replicated by buying \$521.4375 face value of one-year zeros and shorting \$515.0000 face amount of six-month zeros on date 0. Therefore, by the law of one price, the price of the option equals the price of the replicating portfolio, which, using the bond prices given earlier, is equal to,

$$\begin{aligned} .990099F^{.5} + .978842F^{1} &= -.990099 \times \$515.0000 \\ &\quad + .978842 \times \$521.4375 = \$0.504 \end{aligned} \tag{12.5}$$

Recall that pricing based on the law of one price is enforced by arbitrage. If the price of the option were less than \$0.504, arbitrageurs could buy the option, short the replicating portfolio, keep the difference, and have no future liabilities. Similarly, if the price of the option were greater than \$0.504, arbitrageurs could short the option, buy the replicating portfolio, keep the difference, and, once again, have no future liabilities. Thus, ruling out profits from riskless arbitrage implies an option price of \$0.504.

It is important to emphasize that the option cannot be priced by expected discounted value, which gives an option price of,

$$\frac{.5 \times \$0 + .5 \times \$2.555831}{1 + \frac{.02}{2}} = \$1.2653 \tag{12.6}$$

The true option price is lower, because investors dislike the risk of the call option and, as a result, will not pay as much as its expected discounted value. Put another way, the risk penalty implicit in the call option price is inherited from the risk penalty of the one-year zero, that is, from the property that the price of the one-year zero is less than its expected discounted value. Once again, the pricing of risk is discussed in Chapters 13 and 16. While this section illustrates arbitrage pricing with a call

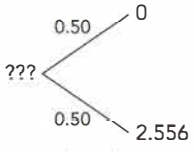


Figure 12.2 **Pricing a 990 six-month call option on a six-month zero coupon bond.**

option, it should be clear that the framework can be used to price any security with cash flows that ultimately depend on the six-month rate. For example, because the price of a five-year bond over time depends on the evolution of the six-month rate, an option on that five-year bond can be priced in this framework as well.

A remarkable feature of arbitrage pricing is that the probabilities of up and down-moves never enter into the calculation of the arbitrage price. See Equations (12.3) through (12.5). The explanation for this somewhat surprising result follows from the principles of arbitrage. Arbitrage pricing requires that the value of the replicating portfolio be the same as the value of the option in both the up- and the down-states. Therefore, the composition of the replicating portfolio is the same whether the probability of the upstate is 20%, 50%, or 80%. But if the composition of the portfolio does not depend directly on the probabilities, and if the prices of the securities in the portfolio are given, then the price of the replicating portfolio and the price of the option cannot depend directly on the probabilities either.

Despite the fact that the option price does not depend directly on the probabilities, these probabilities must have some impact on the option price. After all, as it becomes more and more likely that rates will rise to 2.50% and that bond prices will be low, the value of options to purchase bonds must fall. The resolution of this apparent paradox is that the option price depends indirectly on the probabilities through the price of the one-year zero. Were the probability of an up move to increase suddenly, the current value of a one-year zero would decline. And since the replicating portfolio is long one-year zeros, the value of the option would decline as well. In summary, a derivative like an option depends on the probabilities only through current bond prices. Given bond prices. Given bond prices, however, probabilities are not needed to derive prices determined by arbitrage.

12.3 RISK-NEUTRAL PRICING

Risk-neutral pricing is a technique that modifies an assumed interest rate process, like the one assumed at the start of this chapter, so that any contingent claim can be priced without having to construct and price its replicating portfolio. Because the technique requires that the original interest rate process be modified only once, and because this modification requires no more effort than pricing a single contingent claim by arbitrage, risk-neutral pricing is an extremely efficient way to price many contingent claims under the same assumed rate process.

In the example of this chapter, the price of a one-year zero does not equal its expected discounted value: its price is $978.842, computed from the given one-year spot rate of 2.15%, while its expected discounted value is $980.302, as derived in Equation (12.2). The probabilities of 0.5 for the up-and down-states are the assumed *true* or *real-world* probabilities. But there are other probabilities, called *risk-neutral* probabilities, which do cause the expected discounted value to equal the market price. To find these probabilities, let the risk-neutral probabilities in the up- and down-states be p and $(1 - p)$, respectively. Then, solve the following equation,

$$\frac{\$987.654p + \$992.556(1 - p)}{\left(+\frac{.02}{2}\right)} = \$978.842 \qquad (12.7)$$

to find that p = .8009. Hence, under the risk-neutral probabilities of .8009 and .1991, the expected discounted value does equal the market price.

Risk-neutral probabilities can also be described in terms of the *drift* in interest rates. Under the true probabilities, there is a 50% chance that the six-month rate rises from 2% to 2.50%, and a 50% chance that it falls from 2% to 1.50%. Hence, the expected change in the six-month rate, or the drift of the six-month rate, is zero. Under the risk-neutral probabilities, there is an 80.09% chance of a 50-basis point increase and a 19.91% chance of a 50-basis point decrease, for an expected change of 30.09 basis points. Hence, the drift of the six-month rate under these probabilities is 30.09 basis points.

As pointed out in the previous section, the expected discounted value of the option payoff is $1.2653, while the arbitrage price is $0.504. But if expected discounted value were computed using the risk-neutral probabilities, the resulting option value would equal its arbitrage price,

$$\frac{.8009 \times \$0 + .1991 \times \$2.555831}{\left(1+\frac{.02}{2}\right)} = \$0.504 \qquad (12.8)$$

The fact that the arbitrage price of the option equals its expected discounted value under the risk-neutral probabilities is not a coincidence. In general, to value contingent claims by risk-neutral pricing, proceed as follows. First, find the risk-neutral probabilities that equate the prices of the underlying securities to their expected discounted values. (In the simple example here, the only risky, underlying security is the one-year zero.) Second, price the contingent claim by expected discounted value under these risk-neutral probabilities. The remainder of this section describes intuitively why risk-neutral pricing works. Since the argument is a bit complex, it is broken up into four steps:

1. Given trees for the underlying securities, the price of a security that is priced by arbitrage does not depend on investors' risk preferences. The reasoning is as follows.

   A security is priced by arbitrage if its cash flows can be replicated by some portfolio of underlying securities. Under the assumed process for interest rates in this chapter, the bond option is priced by arbitrage. By contrast, it is unlikely that a specific common stock can be priced by arbitrage, because no portfolio of underlying securities can mimic the idiosyncratic fluctuations of a single common stock's market value.

   If a security is priced by arbitrage, and if everyone agrees on the price evolution of the underlying securities, then everyone agrees on the replicating portfolio. In the option example, both an extremely risk-averse, retired investor and a professional gambler agree that a portfolio of $521.4375 face of one-year zeros and −$515.0000 face of six-month zeros replicates the option. And because they agree both on the composition of the replicating portfolio and on the prices of the underlying securities, they must also agree on the price of the option.

2. Imagine an economy that has the same current bond prices and possible future values of the six-month rate as the true economy. The imaginary economy is different, however, in that its investors are risk neutral. Unlike investors in the true economy, then, investors in the imaginary economy do not penalize securities for risk: they price securities by expected discounted value. In particular, under the probabilities in the imaginary economy, the expected discounted value of the one-year zero equals its market price. But, by Equation (12.7), the expected discounted value of the one-year zero does equal its market price under the risk-neutral probabilities of .8009 and .1991. Hence, these risk-neutral probabilities are the probabilities in the imaginary economy.

3. The price of the option in the imaginary economy, like any other security in that economy, is computed by expected discounted value. Since the probability of the upstate in that economy is .8009, the price of the option in that economy is given by Equation (12.8) and is $0.504.

4. Step 1 implies that, given the prices of the six-month and one-year zeros, as well as possible values of the six-month rate, the price of an option does not depend on investor risk preferences. Therefore, because the real and imaginary economies have the same bond prices and the same possible values for the six-month rate, the option price must be the same in both economies. In particular, the option price in the real economy must also equal $0.504. More generally, the price of a derivative in the real economy may be computed by expected discounted value under the risk-neutral probabilities.

12.4 ARBITRAGE PRICING IN A MULTI-PERIOD SETTING

Maintaining the binomial assumption, Figure 12.3 extends the tree from the previous section for another six months. This tree is called a *recombining tree*, because an up-move followed by a down-move, to the up–down state, lands in the same place as a down-move followed by an up-move, to the down–up state. Trees for which this is not the case are said to be *nonrecombining*. While nonrecombining trees might represent economically reasonable dynamics, they tend to be avoided as difficult or even impossible to implement. After six months there are two possible states, after one year there are four, and after N semiannual periods there are 2^N possibilities. A tree with enough semiannual steps to price 10-year securities has, in its rightmost column alone, over 500,000 nodes, and to price 20-year securities, over 500 billion. Furthermore, as discussed later in the chapter, it is often desirable to reduce the time interval between dates substantially. In short, even with modern computers, trees that grow this quickly are computationally unwieldy. This does not mean that the effects that seem to give rise to nonrecombining trees – like volatilities that change across states – cannot be modeled. It does mean, however, that such effects have to be implemented in more efficient ways.

Returning to the recombining format, as trees grow it becomes convenient to develop a notation with which to refer to particular nodes. One convention is as follows. The dates, represented by columns of the tree, are numbered from left to right starting with 0. The states, represented by rows of the tree, are numbered from bottom to top, also starting from 0. For example, in Figure 12.3, the six-month rate on date 2, state 0 is 1%. The six-month rate on state 1 of date 1 is 2.50%.

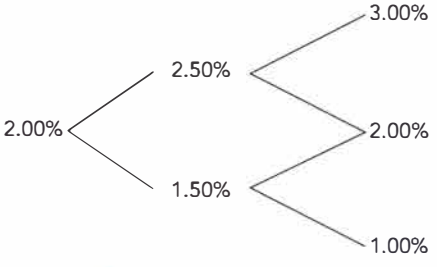


Figure 12.3 A recombining binomial rate tree.

Continuing where the option example left off, having derived the risk-neutral tree for pricing a one-year zero, the goal is to extend the tree 0 price a 1.5-year zero assuming that the 1.5-year spot rate is 2.25%. Ignoring the probabilities for a moment, several nodes of the 1.5-year zero price tree can be written down immediately, as shown in Figure 12.4. On date 3, the zero with an original term of 1.5 years matures and is worth its face value of $1,000. On date 2, the value of the then six-month zero equals its face value discounted for six months at the then-prevailing spot rates of 3%, 2%, and 1%, in states 2, 1, and 0, respectively,

$$\frac{\$1{,}000}{1+\frac{.03}{2}} = \$985.22 \qquad (12.9)$$

$$\frac{\$1{,}000}{1+\frac{.02}{2}} = \$990.10 \qquad (12.10)$$

$$\frac{\$1{,}000}{1+\frac{.01}{2}} = \$995.02 \qquad (12.11)$$

Finally, on date 0, the 1.5-year zero equals its face value discounted at the given, 1.5-year spot rate,

$$\frac{\$1.000}{\left(1+\frac{.0225}{2}\right)^3} = \$966.9954 \qquad (12.12)$$

The prices of the zero on date 1 in states 1 and 0 are denoted in Figure 12.4 by $P_{1,1}$ and $P_{1,0}$, respectively. These one-year zero prices are not known at this point.

The previous section showed that the risk-neutral probability of an up-move on date 0 is 0.8009. Letting q be the risk-neutral probability of an up-move on date 1, and, for the purposes of this section, making the simplifying assumption that the probability of moving up from state 0 is the same as the probability of moving up from state 1, the resulting tree is shown in Figure 12.5.

By definition, the expected discounted value under risk-neutral probabilities recovers market prices. With respect to the 1.5-year zero price on date 0, this requires that,

$$\frac{.8009P_{1,1} + .1991P_{1,0}}{1+\frac{.02}{2}} = \$966.995 \qquad (12.13)$$

And with respect to the prices of a then one-year zero on date 1,

$$P_{1,1} = \frac{\$985.222q + \$990.009(1-q)}{1+\frac{.025}{2}} \qquad (12.14)$$

$$P_{1,0} = \frac{\$990.099q + \$995.025(1-q)}{1+\frac{.015}{2}} \qquad (12.15)$$

Substituting Equations (12.14) and (12.15) into Equation (12.13) results in a linear equation in the one unknown, q, which can be solved to find that $q = 0.6520$. Therefore, the risk-neutral interest rate process is summarized by the tree in Figure 12.6. Furthermore, any contingent claim that depends on the six-month rate in six months and in one year may be priced by computing its discounted expected value along this tree. An example is given in the next section.

The difference between the true and risk-neutral probabilities may once again be described in terms of drift. From dates 1 to 2, the drift under the true probabilities is zero. Under the risk-neutral probabilities, the drift is computed from a 65.20% chance of a 50-basis-point increase in the six-month rate and

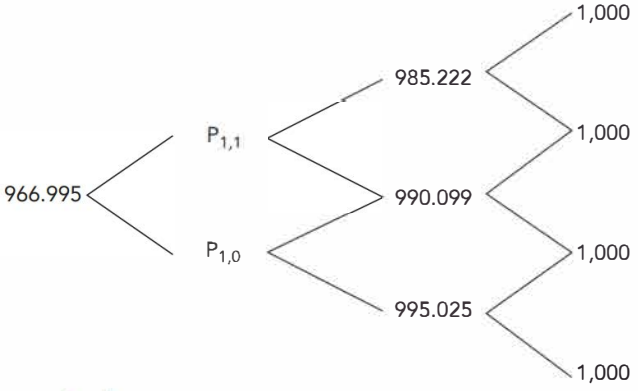


Figure 12.4 Price tree for a 1.5-year zero coupon bond.

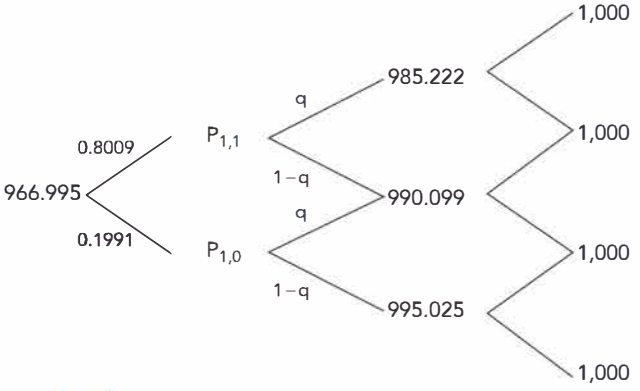


Figure 12.5 Price tree for a 1.5-year zero coupon bond, with probabilities.

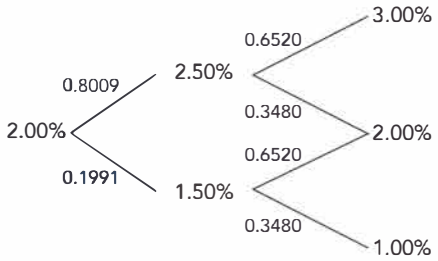


Figure 12.6 **Risk-neutral process for the six-month rate.**

a 34.80% chance of a 50-basis-point decline in the rate. These numbers give a drift or expected change of 15.20 basis points.

Substituting $q = 0.6520$ back into Equations (12.14) and (12.15) completes the tree for the price of the 1.5-year zero, which is shown in Figure 12.7. It follows immediately from this tree that the one-year spot rate in six months may be either 2.5754% or 1.5754%, because,

$$\$974.735 = \frac{\$1{,}000}{\left(1 + \frac{2.5754\%}{2}\right)^2} \tag{12.16}$$

$$\$984.430 = \frac{\$1{,}000}{\left(1 + \frac{1.5754\%}{2}\right)^2} \tag{12.17}$$

The fact that the possible values of the one-year spot rate can be extracted from the tree is at first surprising. The starting point of the example is the date 0 values of the 0.5-, 1-, and 1.5-year spot rates, along with assumptions about the evolution of the six-month rate over the next years. But because this information, in combination with arbitrage or risk-neutral arguments, is sufficient to determine the price tree of the 1.5-year zero, it is also sufficient to determine the possible values of the one-year spot rate in six months. Put another way, having specified initial spot rates

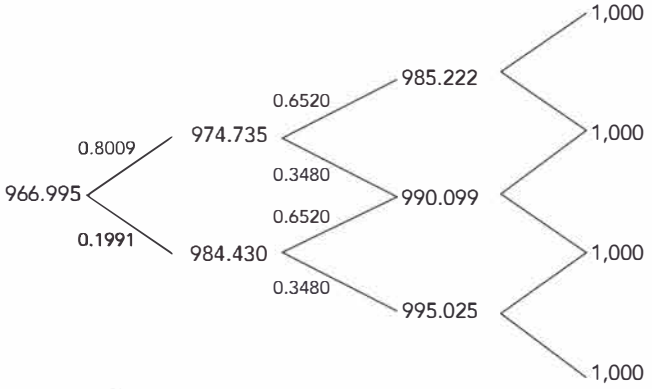


Figure 12.7 **Final price tree for a 1.5-year zero coupon bond, with probabilities.**

and the evolution of the six-month rate, a modeler may not make any further assumptions about the behavior of the one-year rate.

The six-month rate process completely determines the one-year rate process here, because the model presented has only one factor. Writing down a tree for the evolution of the six-month rate alone implicitly assumes that prices of all fixed income securities can be determined by the evolution of that rate.

This section concludes with two additional observations about multi-period settings. First, extending the tree to any number of dates requires assumptions about the future possible values of the short-term rate and the calculation of risk-neutral probabilities that recover a given set of bond prices. Second, the composition of replicating portfolios depends on date and state. For example, the replicating portfolio of a derivative as of date 0 is usually different from its replicating portfolio on date 1, state 0, and different again from its replicating portfolio on date 1, state 1. From a trading perspective, this means that replicating portfolios must be adjusted as time passes and as interest rates change. These adjustments are known as *dynamic replication*, in contrast to the *static replication* strategies of earlier chapters, like replicating a coupon bond with an unchanging portfolio of two other coupon bonds of the same maturity.

12.5 PRICING A CONSTANT-MATURITY TREASURY SWAP

Equipped with the tree in Figure 12.7, this section prices a $1,000,000 stylized *constant-maturity Treasury (CMT)* swap struck at 2%. This swap pays,

$$\$1{,}000{,}000\frac{y_{CMT} - 2\%}{2} \tag{12.18}$$

every six months until it matures, where y_{CMT} is the semiannually compounded yield – of a predetermined maturity – on the payment date. This example prices a one-year CMT swap on the six-month yield, though, in practice, CMT swaps trade most commonly on the yields of the most liquid bonds, for example, on two-, five- and 10-year Treasury yields.

Because six-month semiannually compounded yields equal six-month spot rates, rates from the tree of the previous section can be substituted into Equation (12.18) to calculate the payoffs of the CMT swap. On date 1, the state 1 and state 0 payoffs are, respectively,

$$\$1{,}000{,}000\frac{2.50\% - 2\%}{2} = \$2{,}500 \tag{12.19}$$

$$\$1{,}000{,}000\frac{1.50\% - 2\%}{2} = -\$2{,}500 \tag{12.20}$$

Similarly on date 2, the state 2, 1, and 0 payoffs are, respectively,

$$\$1{,}000{,}000\frac{3\% - 2\%}{2} = \$5{,}000 \quad (12.21)$$

$$\$1{,}000{,}000\frac{2\% - 2\%}{2} = \$0 \quad (12.22)$$

$$\$1{,}000{,}000\frac{1\% - 2\%}{2} = -\$5{,}000 \quad (12.23)$$

The possible values of the CMT swap at maturity, on date 2, are given by Equations (12.21) through (12.23). The possible values on date 1 are given by the expected discounted value of the date 2 payoffs under the risk-neutral probabilities plus the date 1 payoffs given by (12.19) and (12.20). The resulting date 1 values in states 1 and 0 are, respectively,

$$\frac{.6520 \times \$5000 + .3480 \times \$0}{1 + \frac{.0250}{2}} + \$2{,}500 = \$5{,}719.52 \quad (12.24)$$

$$\frac{.6520 \times 0 + .3480 \times (-\$5{,}000)}{1 + \frac{.0150}{2}} - \$2{,}500 = -\$4{,}227.29 \quad (12.25)$$

Finally, the value of the swap on date 0 is the expected discounted value, under the risk-neutral probabilities, of the date-1 payoffs, given by Equations (12.24) and (12.25),

$$\frac{.8009 \times \$5{,}719.52 + .1991 \times (-\$4{,}227.29)}{1 + \frac{.0200}{2}} = \$3{,}702.11 \quad (12.26)$$

The tree in Figure 12.8 summarizes the value of the stylized CMT swap ever dates and states. A value of $3,702.11 for the CMT swap might seem surprising at first. After all, the cash flows of the CMT swap are zero at a rate of 2%, and 2% is, under the true probabilities, the average rate on each date. The explanation, of course, is that the risk-neutral probabilities, not the true probabilities, determine the arbitrage price of the swap. The expected discounted value of the swap under the true probabilities can be computed by following the steps leading to Equations (12.24) through (12.26) but using the probability 0.5 for all up- and down-moves. The result of these calculations does give a value close to zero, namely, −$6.07.

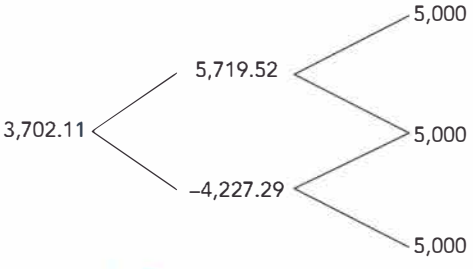


Figure 12.8 Price tree for a stylized CMT swap.

12.6 OPTION-ADJUSTED SPREAD

Option-adjusted spread is a widely used measure of the relative value of a security, that is, of its market price relative to its model value. OAS is defined as the spread such that the market price of a security is recovered when that spread is added to discount rates in the model. To illustrate, say that the market price of the CMT swap in the previous section is $3,699.18, $2.92 less than the model price. In that case, the OAS of the CMT swap turns out to be 10 basis points. To see this, add 10 basis points to the discounting rates of 2.5% and 1.5% in Equations (12.24) and (12.25), respectively, to get new swap values of,

$$\frac{.6520 \times \$5000 + .3480 \times \$0}{1 + \frac{.0260}{2}} + \$2{,}500 = \$5{,}717.93 \quad (12.27)$$

$$\frac{.6520 \times 0 + .3480 \times (-\$5{,}000)}{1 + \frac{.0160}{2}} - \$2{,}500 = -\$4{,}226.43 \quad (12.28)$$

Note that, when calculating value with an OAS spread, rates are only shifted for the purpose of discounting. Rates are not shifted for the purposes of computing cash flows. In the CMT swap example, cash flows are still computed using Equations (12.19) through (12.23).

Completing the valuation with an OAS of 10 basis points, use the results of (12.27) and (12.28) and a discount rate of 2% plus the OAS spread of 10 basis points, or 2.10%, to obtain an initial CMT swap value of,

$$\frac{.8009 \times \$5{,}717.93 + .1991 \times (-\$4{,}226.43)}{1 + \frac{.0210}{2}} = \$3{,}699.18 \quad (12.29)$$

Hence, as claimed, discounting at the risk-neutral rates plus an OAS of 10 basis points in the model recovers the given market price of $3,699.18. If a security's OAS is positive, its market price is less than its model price, which means that the security trades cheap. If the OAS is negative, the security trades rich.

Another perspective on the relative value implications of an OAS spread is the fact that the expected return of a security with an OAS, under the risk-neutral process, is the short-term rate plus the OAS per period. Very simply, discounting a security's expected value by a particular rate per period is

equivalent to that security's earning that rate per period. In the example of the CMT swap, the expected return of the fairly priced swap under the risk-neutral process over the six months from date 0 to date 1 is,

$$\frac{.8009 \times \$5{,}719.52 - .1991 \times \$4{,}227.29 - \$3{,}702.11}{\$3{,}702.11} = 1.00\% \tag{12.30}$$

which is six months' worth of the initial rate of 2%. On the other hand, with an OAS of 10 basis points, the expected return of the cheap swap is,

$$\frac{.8009 \times \$5{,}717.93 - .1991 \times \$4{,}226.43 - \$3{,}699.18}{\$3{,}699.18} = 1.05\% \tag{12.31}$$

which is six months' worth of the initial rate of 2% plus the OAS of 10 basis points, or half of 2.10%.

12.7 PROFIT AND LOSS ATTRIBUTION WITH AN OAS

This section gives a mathematical description of attribution in the context of term structure models and of securities that trade with an OAS. While the notation of this chapter is quite formal, the presentation remains intuitive.

By the definition of a one-factor model, and by the definition of OAS, the market price of a security at time t and a factor, r, which is often a rate, can be written as $P_t(r, OAS)$. Using a first-order Taylor approximation, the change in the price of the security is,

$$dP = \frac{\partial P}{\partial r}\,dr + \frac{\partial P}{\partial t}\,dt + \frac{\partial P}{\partial OAS}\,dOAS \tag{12.32}$$

where $\partial P/\partial r$ gives the change in the price of the security for a change in r, holding t and OAS constant; $\partial P/\partial t$ gives the change in price for a change in t holding r and OAS constant; and the same for $\partial P/\partial OAS$. In words, Equation (12.32) breaks down the total change in price to components of change due to changes in r, t, and OAS.

Dividing both sides of Equation (12.32) by price and taking expectations,

$$E\left[\frac{dP}{P}\right] = \frac{1}{P}\frac{\partial P}{\partial r}E[dr] + \frac{1}{P}\frac{\partial P}{\partial t}\,dt \tag{12.33}$$

Note that dP/P is the change in price divided by price, or the percentage change in price. Because the OAS calculation assumes that OAS is constant over the life of the security, moving from (12.32) to (12.33) assumes that the expected change in the OAS is zero.

As mentioned in the previous section, if expectations are taken with respect to the risk-neutral process, then, for any security priced according to the model,

$$E\left[\frac{dP}{P}\right] = r_0 dt \tag{12.34}$$

But Equation (12.34) does not apply to securities that are not priced according to the model, that is, to securities with an OAS not equal to zero. For these securities, by definition, the cash flows are discounted not at the short-term rate, but at the short-term rate plus the OAS. Equivalently, as argued in the previous section, the expected return under the risk-neutral probabilities is not the short-term rate, but the short-term rate plus the OAS. Hence, the more general form of Equation (12.34) is,

$$E\left[\frac{dP}{P}\right] = (r_0 + OAS)dt \tag{12.35}$$

Combining these pieces, substitute for $(1/P)\partial P/\partial t$ from (12.33) and then for $E[dP/P]$ from (12.35) into Equation (12.32) and rearrange terms, which breaks down the return of a security into its component parts,

$$\frac{dP}{P} = (r_0 + OAS)dt + \frac{1}{P}\frac{\partial P}{\partial r}(dr - E[dr]) + \frac{1}{P}\frac{\partial P}{\partial OAS}\,dOAS \tag{12.36}$$

Finally, multiplying through by P,

$$dP = (r_0 + OAS)\,P dt + \frac{\partial P}{\partial r}(dr - E[dr]) + \frac{\partial P}{\partial OAS}\,dOAS \tag{12.37}$$

In words, the return of a security or its P&L may be divided into a component due to the passage of time; a component due to changes in the factor; and a component due to the change in the OAS. The terms on the right-hand side of (12.37) represent, in order, carry-roll-down, gains or losses from rate changes, and gains or losses from spread change.[1] For models with predictive power, the OAS converges or trends to zero; that is, the security price converges or trends toward its fair value according to the model.

The decompositions of Equations (12.36) and (12.37) highlight the usefulness of OAS as a measure of value. If a model

[1] For expositional simplicity no explicit coupon or other direct cash flows have been included in this discussion.

is correct, a long position in a cheap security earns superior returns in two ways. First, it earns the OAS over time intervals in which the security does not converge to its fair value. Second, it earns its sensitivity to OAS times any convergence of that OAS to zero.

The decompositions also provide a framework for relative value trading. When a cheap or rich security is identified, a relative value trader buys or sells the security and hedges out all interest rate or factor risk. Mathematically, $\partial P/\partial r = 0$. In that case, the expected return or P&L depends only on the short-term rate, the OAS of the securities traded, and any OAS convergence. If the trader finances the trade at the short-term rate, that is, borrows P at rate r_0 to purchase the security, then the expected return is simply equal to the OAS plus any convergence return. If the hedge itself costs or generates funds, then the P&L also includes a return on those funds at the short-term rate. If the hedging securities are not fairly priced relative to the model, but have an OAS of their own, then the P&L also includes an OAS on the hedge. Finally, Chapter 13 explains that bearing interest rate or factor risk may earn a risk premium, in which case there would be an additional term in Equations (12.34) through (12.37) that depends on the amount of factor risk borne. But, in the relative value context, where factor risk is hedged away, any risk premium terms cancel out, and the P&L of the trade is as described in this paragraph.

12.8 REDUCING THE TIME STEP

This chapter has so far assumed that the time elapsed between dates of the tree is six months. The methodology outlined, however, adapts easily to any time step of Δt years. For monthly time steps, for example, $\Delta t = 1/12$ or .0833, and one-month rather than six-month interest rates appear on the tree. Furthermore, discounting is done over the appropriate time interval. If the rate of term Δt is r, then discounting means dividing by $1 + r\Delta t$. In the case of monthly time steps, discounting with a one-month rate of 2% means dividing by 1 + 0.02/12.

In practice there are two reasons to choose time steps smaller than six months. First, a security or portfolio of securities rarely makes all of its payments in even six-month intervals from the starting date. Reducing the time step to a month, a week, or even a day can ensure that all cash flows are sufficiently close in time to some date in the tree. Second, assuming that the six-month rate can take on only two values in six months, three values in one year, and so on, produces a tree that is too coarse for many practical pricing problems. Reducing the step size can fill the tree with enough rates to price contingent claims with sufficient accuracy.

While smaller time steps generate more realistic interest rate distributions, they require that more attention be paid to numerical issues, and they may make computations too slow for their intended uses. The choice of step size ultimately depends, therefore, on the problem at hand. When pricing a 30-year callable bond, for example, a model with weekly or even monthly time steps may provide a realistic enough interest rate distribution to generate reliable prices. By contrast, pricing a one-month bond option with any precision would require a much smaller time step. While the trees in this chapter assume that the step size is the same throughout the tree, this need not be the case. Sophisticated implementations of trees allow step size to vary across dates in order to achieve a balance between realism and computational concerns.

12.9 FIXED INCOME VERSUS EQUITY DERIVATIVES

The famous Black-Scholes-Merton (BSM) pricing analysis of stock options can be summarized as follows. Under the assumptions that the stock price evolves according to a particular random process and that the short-term interest rate is constant, it is possible to form a portfolio of the underlying stock and short-term bonds that replicates the option. Therefore, by arbitrage arguments, the price of the option equals the price of the replicating portfolio.

Consider applying this logic to price an option on a five-year bond. The starting point might be an assumption about how the price of the five-year bond evolves over time, but the task is considerably more complicated than for the price of a stock. First, the price of a bond converges to its face value at maturity, while the stock price has no similar constraint. Second, because of the maturity constraint, the volatility of a bond's price must decline as the bond approaches maturity. Hence, the simple assumption that the volatility of a stock is constant is not as appropriate for bonds. Third, because stock volatility is very large relative to short-term rate volatility, it is often acceptable to assume that the short-term interest rate is constant. By contrast, it seems inconsistent to assume that bond prices – which depend on interest rates – follow some random process, while assuming that the short-term interest rate is constant.

These objections led researchers to make assumptions about the random evolution of the interest rate rather the bond price. In that way, bond prices naturally approach par, price volatilities naturally approach zero, and no interest rate is assumed to be constant. But this approach raises another set of questions. Which interest rate is assumed to evolve in a particular way?

Assumptions about the five-year spot rate, for example, are not sufficient for two reasons. First, five-year coupon bond prices depend on shorter-term spot rates as well. Second, an option on a particular five-year bond soon depends on the prices of a bond that is no longer a five-year bond, but a bond with slightly less time to maturity. Therefore, assumptions usually have to be made about the evolution of the entire term structure of interest rates to price bond options and other derivatives. This chapter shows that, in a one-factor model, assumptions about the evolution of the short-term rate are sufficient to model the evolution of the entire term structure.

In short, there are several arguments to move beyond BSM in the fixed income context. Nevertheless, for simplicity, practitioners do use versions of BSM to price and hedge several classes of fixed income derivatives.

13

Expectations, Risk Premium, Convexity, and the Shape of the Term Structure

Learning Objectives

After completing this reading, you should be able to:

- Explain the role of interest rate expectations in determining the shape of the term structure.
- Apply a risk-neutral interest rate tree to assess the effect of volatility on the shape of the term structure.
- Estimate the convexity effect using Jensen's inequality.
- Identify the components into which the return on a bond can be decomposed, and calculate the expected return on a bond for a risk-averse investor.

Excerpt is Chapter 8 of Fixed Income Securities: Tools for Today's Markets, *Fourth Edition, by Bruce Tuckman and Angel Serrat.*

Chapter 11 shows how bonds and other interest rate contingent claims can be priced given the evolution of the short-term rate. This chapter shows how the shape of the term structure of interest rates is determined by assumptions about the evolution of the short-term rate and by assumptions about the risk premium demanded by investors for bearing interest rate risk. The first few sections of the chapter present concepts by way of example, in the simple, binomial tree framework of the previous chapter. The last section of the chapter presents the same ideas in more general setting, though at the cost of some higher-level mathematics. Chapter 16 concludes the presentation of term structure models with a detailed description of two well-known models.

13.1 EXPECTATIONS

Consider a simple framework with annual periods. Assume for the moment that the current one-year rate is 8%, and that investors know with certainty that the one-year rate in one year will be 7% and in two years will be 6%. Then, the prices of one-, two-, and three-year zero coupon bonds with a unit face value, $P(1)$, $P(2)$, and $P(3)$, are priced such that,

$$P(1) = \frac{1}{1.08}$$
$$P(2) = \frac{1}{1.08 \times 1.07}$$
$$P(3) = \frac{1}{1.08 \times 1.07 \times 1.06} \qquad (13.1)$$

But by the definition of forward rates, Equations (13.1) say that the first three forward rates are 8%, 7%, and 6%. Hence, with investor certainty as to future interest rates, that is, without any volatility around those expectations, the term structure of interest rates – here expressed in terms of forward rates – is completely determined by expectations. Consequently, depending on expectations, the term structure can take on any shape: flat, upward-sloping, downward-sloping, or some combination of these.

In practice, expectations cannot sensibly take on any arbitrary pattern. The financial community can have very specific views about short-term rates over short horizons, derived, for example, from anticipation of policy changes on central bank meeting dates and from the supply and demand conditions for funds (e.g., tax payment dates, the bond issuance calendar, quarterly balance sheet management). Over longer horizons, however, expectations are not as granular. Analysis of money market conditions is unlikely to reveal, for example, that the expected one-year rate in 29 years is very different from the expected one-year rate in 30 years. On the other hand, macroeconomic analysis might argue that the long-run expectation of the short-term rate is 4%: 1 % due to the long-run real-rate of interest and 3% due to long-run inflation.

13.2 VOLATILITY AND CONVEXITY

While investors have expectations about future short-term rates, they recognize the limits of their analyses, that is, realized rates are assumed to fluctuate randomly around expectations. Continuing with the framework of the previous chapter, consider the binomial tree for the one-year rate in the top part of Figure 13.1. The step size is one year, and the probabilities of all transitions are 50% (not shown). The level of rates and their volatility is exaggerated in this tree to illustrate the concepts of this chapter. Note that the expected value of the short-term rate in one year is 9%, as is the expected short-term rate in two years,

$$50\% \times 13\% + 50\% \times 5\% = 9\% \qquad (13.2)$$
$$50\%[50\% \times 17\% + 50\% \times 9\%] + 50\%[50\% \times 9\% + 50\% \times 1\%] = 9\% \qquad (13.3)$$

Note also that the volatility of the change in rate at any transition is 4%, or 400 basis points. For example, with the mean of the first transition calculated in Equation (13.2) to be 9%, the volatility of that transition is,

$$\sqrt{50\%[13\% - 9\%]^2 + 50\%[5\% - 9\%]^2} = 4\% \qquad (13.4)$$

The price of a one-year zero in this model is simply 1/1.09, or 0.917431. Assuming, for the moment, that investors are risk neutral, the price trees of the two- and three-year zeros can be calculated by expected discounted value, as explained in the previous chapter. These trees are shown in the Table 13.1 collects the prices of the three zero coupon bonds, along with the associated forward rates. The striking feature of the table is that the term structure of forward rates is downward sloping, despite the fact that interest rate expectations are flat at 9%. This result

Table 13.1 **Prices of Zero Coupon Bonds and Their Associated Forward Rates from the Rate Tree in Figure 13.1. Rates Are in Percent.**

Term	Price	Forward Rate
1	0.917431	9.0000
2	0.842815	8.8532
3	0.776366	8.5590

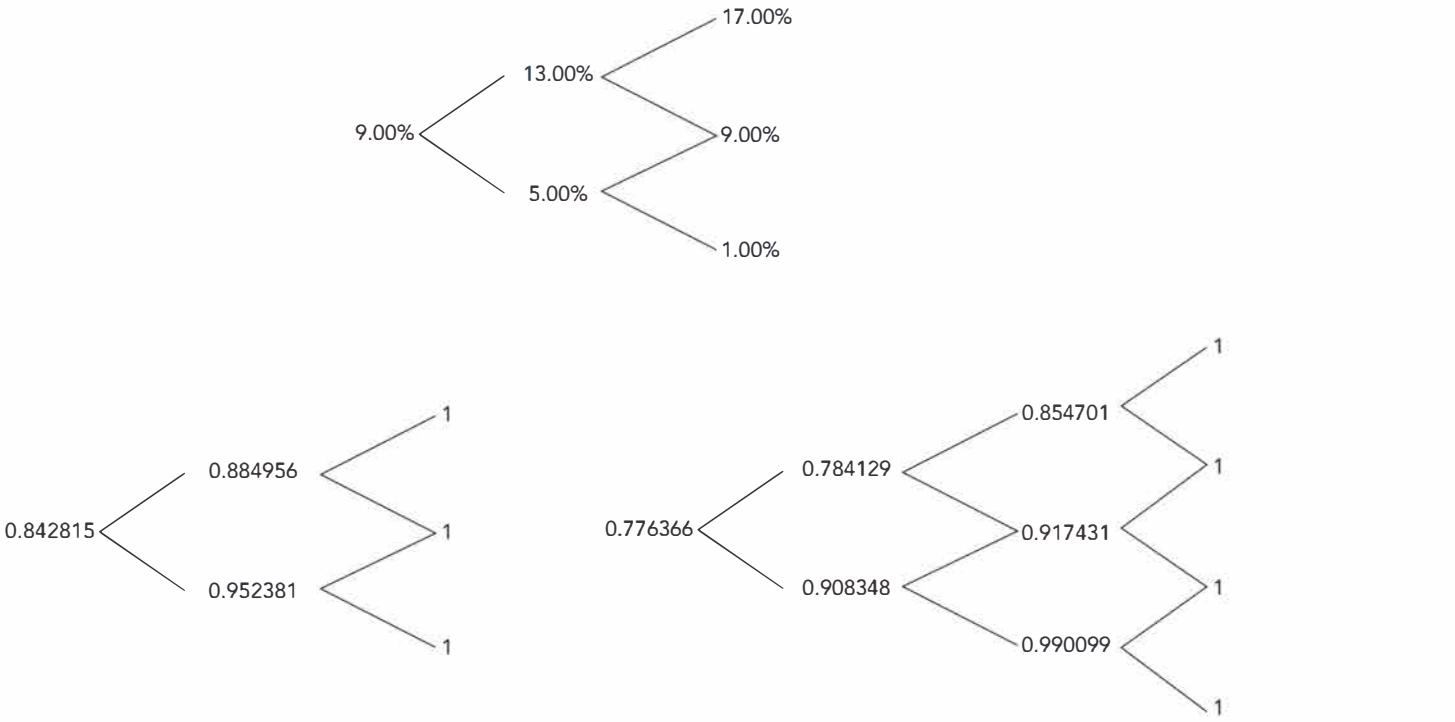


Figure 13.1 **Binomial rate tree and price trees for two- and three-year zero coupon bonds. Steps are annual, and the probabilities of all transitions are 50%.**

can be explained by the interaction of interest rate volatility with the convexity of bond prices.

A short detour is required at this point to present and explain *Jensen's inequality* as applied to bond pricing. For a random variable, like the one-year rate, r,

$$E\left[\frac{1}{1+r}\right] > \frac{1}{E[1+r]} = \frac{1}{1+E[r]} \qquad (13.5)$$

In words, the expected price of a bond is greater than the price of a bond at the expected interest rate.

This inequality is easily explained by Figure 13.2. In the figure, the rate can take on two values, r^d and r^u, with equal probability, resulting in an expected value just between them, $E[r]$. Each possible value of r has an associated price, and the expected value of price, $E[1/(1+r)]$, is graphically depicted as the vertical-axis coordinate of the dotted line connecting the points $\{r^d, 1/(1+r^d)\}$ and $\{r^u, 1/(1+r^u)\}$. Because of the curvature or convexity of the price-rate curve, however, this expected price exceeds the price at a rate of $E[r]$, which is $1/(1+E[r])$. And this is exactly the relationship described in Equation (13.5).

Returning to the role of volatility and convexity, let f denote the one-year rate, one year forward, and consider the date-0 price of a two-year zero coupon bond, as expressed in Equation

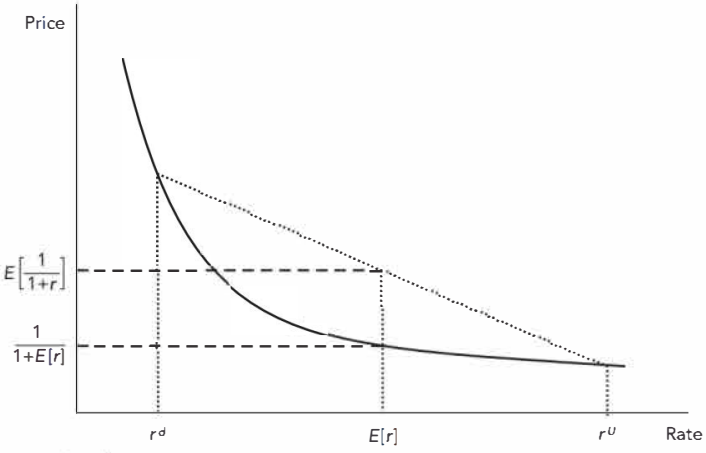


Figure 13.2 **An illustration of jensen's inequality as applied to bond pricing.**

(13.6). By definition, the price of the two-year zero equals its unit face amount discounted by 9% over the first year and by f over the second year. By the logic of pricing along the tree, this price also equals the discounted expected value of the date-1 price of the bond. Multiplying both sides of (13.6) by 1.09 and invoking Jensen's inequality in Equation (13.5), gives (13.7). And from this equation, (13.8) follows directly: the one-year rate, one year forward, is less than the expected one-year rate in one year,

$$0.8428 \equiv \frac{1}{(1.09)(1+f)} = \frac{1}{1.09}\left[50\% \times \frac{1}{1.13} + 50\% \times \frac{1}{1.05}\right] \tag{13.6}$$

$$\frac{1}{1+f} > \frac{1}{50\% \times 1.13 + 50\% \times 1.05} = \frac{1}{1.09} \tag{13.7}$$

$$f < 9\% \tag{13.8}$$

13.3 AN ANALYTICAL DECOMPOSITION OF FORWARD RATES

This section derives a general decomposition of forward rates in terms of expectations, convexity, and risk premium. The level of mathematics here is higher than used in most of the book, but the discussion still aims at intuition.

Assume that all bond prices are determined by the instantaneous rate, r, which takes on the value of r_t at time t. Let $P_t(r_t, T)$ be the price of a T-year zero coupon bond at time t. By *Ito's lemma*, a discussion of which is beyond the scope of this book,

$$dP = \frac{\partial P}{\partial r}dr + \frac{\partial P}{\partial t}dt + \frac{1}{2}\frac{\partial^2 P}{\partial r^2}\sigma^2\,dt \tag{13.9}$$

where dP, dr, and dt are the changes in price, rate, and time over the next instant, respectively, and σ is the volatility of changes in r. The two first-order partial derivatives in Equation (13.9) denote the instantaneous change in the bond price for a unit change in the rate (with time unchanged) and for a unit change in time (with rate unchanged), respectively. Finally, the second order partial derivative in the equation gives the instantaneous change in $\partial P/\partial r$ (with time unchanged). Dividing both sides of (13.9) by price,

$$\frac{dP}{P} = \frac{1}{P}\frac{\partial P}{\partial r}dr + \frac{1}{P}\frac{\partial P}{\partial t}dt + \frac{1}{2}\frac{1}{P}\frac{\partial^2 P}{\partial r^2}\sigma^2 dt \tag{13.10}$$

Equation (13.10) breaks down the instantaneous return on the zero coupon bond into three components, but this decomposition can be written more intuitively by invoking several ideas from earlier chapters.

First, in terms of instantaneous compounded forward rates, $f(t)$, the price of a T-year zero coupon bond is (from Section A2.1),

$$P = e^{-\int_0^T f(s)ds} \tag{13.11}$$

Then, differentiating both sides of (13.11) with respect to t, recognizing that increasing t decreases T one-for-one,

$$\frac{\partial P}{\partial t} = -\frac{\partial P}{\partial T} = f(T)P \tag{13.12}$$

Second, by the definitions of duration, D, and convexity, C,

$$D \equiv -\frac{1}{P}\frac{\partial P}{\partial r} \tag{13.13}$$

$$C = \frac{1}{P}\frac{\partial^2 P}{\partial r^2} \tag{13.14}$$

Now, substituting Equations (13.12) through (13.14) into the return decomposition (13.10),

$$\frac{\partial P}{P} = f(T)dt - Ddr + \frac{1}{2}C\sigma^2\,dt \tag{13.15}$$

Equation (13.15) gives the return decomposition in terms of the following three components. The first is the return due to the passage of time, which, in this case, is the forward rate, $f(T)$.[1] The second and third components are returns due to changes in the rate. The second term says that increases in rate reduce bond return in proportion to duration. The third term says that the volatility of rates – movement of rates either up or down – increases return in proportion to convexity. To appreciate this term, recall that, across portfolios with the same duration, more convex portfolios increase more in value as rates change (at a fixed moment in time), whether rates rise or fall.

To draw conclusions about expected returns, take the expectation of both sides of (13.15),

$$E\left[\frac{\partial P}{P}\right] = f(T)dt - DE[dr] + \frac{1}{2}C\sigma^2\,dt \tag{13.16}$$

The intuition of this decomposition is the same as for Equation (13.15), but with the duration component depending not on the change in rate but on the expected change in rate.

The next step in the analysis introduces the concept of a risk premium. Risk-neutral investors, who do not require a risk premium, demand that each bond offer an expected return equal to the short-term rate of interest. Mathematically,

$$E\left[\frac{dP}{P}\right] = r_0 dt \tag{13.17}$$

Risk averse investors, however, demand higher expected returns for bonds with greater interest rate risk. The interest rate risk of a bond over the next instant may be measured by its duration with respect to the interest rate factor, and that risk-averse

[1] Note that the result here is related to the result that the T-year zero coupon bond earns the forward rate corresponding to its term, $f(T)$, under the assumption of an unchanged term structure and, implicitly, no change in the short-term rate and no interest rate volatility.

investors demand a risk premium proportional to duration. This risk premium may depend on time and on the level of rates, but not on the characteristics of any individual bond. The discussion proceeds here, however, as if the risk premium were constant and denoted by λ. In that case, the expected return equation for risk-averse investors is,

$$E\left[\frac{dP}{P}\right] = r_0 dt + \lambda D dt \qquad (13.18)$$

Say, for example, that the short-term rate is 1%, that the duration of a bond is five, and that the risk premium is 10 basis points per year of duration risk. Then, according to Equation (13.18), the bond's expected return is 1% + 5 × 0.1% = 1.5% per year.

Another useful way to think of the risk premium is in terms of the *Sharpe ratio* (SR) of a security, defined as its expected excess return (over the short-term rate) divided by the standard deviation of its return. Because the random part of a bond's return comes from its duration times the change in rate, as in Equation (13.15), the standard deviation of the return equals the duration times the standard deviation of rates. Therefore, the SR of a bond may be written as,

$$SR = \frac{E[dP/P] - r_0 dt}{\sigma D dt} = \frac{\lambda}{\sigma} \qquad (13.19)$$

where the second equality follows from Equation (13.18). For example, if the risk premium is 10 basis points per year, and if the standard deviation of rates is 100 basis points per year, then the Sharpe ratio of bond investments is 10%.

The decomposition of returns can now be combined with the economics of the risk premium to draw conclusions about the shape of the term structure of forward rates. Equating the expressions for expected returns in the right-hand sides of Equations (13.16) and (13.18),

$$f(T) = \left\{ r_0 + E\left[\frac{dr}{dt}\right] D \right\} + \lambda D - \frac{1}{2} C\sigma^2 \qquad (13.20)$$

Equation (13.20) mathematically describes the determinants of forward rates. The three terms represent the impacts of expectations, risk premium, and convexity, respectively. The first term says that the forward rate is composed of the instantaneous interest rate plus the expected change in that rate times the duration of the zero coupon bond corresponding to the term of the forward rate. In other words, the higher the instantaneous rate, the higher the forward rate; the more rates are expected to increase, the higher the forward rate; and the greater the corresponding duration, the greater the effect of expected rate changes on the forward rate.

The second term on the right-hand side of (13.20) says that the forward rate increases with the risk premium in proportion to the corresponding duration. In other words, the greater the corresponding interest rate risk and the greater the risk premium, the greater the forward rate.

A drift in the short-term rate of a certain number of basis points has the same effect on bond pricing as a risk premium of that number of basis points per year of duration risk. Equation (13.20) formalizes this statement. Increasing the risk premium or increasing the expected short-term rate by the same amount are indistinguishable from the observation of forward rates. This means that the term structure of interest rates cannot, on its own, be used to separate expectations of rate changes from risk premium. From a modeling perspective, this means that only the risk-neutral process is relevant for pricing. Dividing the risk-neutral drift into expectations and risk premium might be very useful for economic perspectives and for macro-style trading, but this division is not observable from a cross section of bond prices alone.

The first two terms of Equation (13.20) can also be cast in terms of theories of the term structure of interest rates. (Put aside the convexity term for the moment.) Under the *pure expectations hypothesis*, the risk premium, λ, is zero, and the term structure of forward rates is determined by expectations, $E[dr/dt]$. In this view of the world, the most natural "no-change" scenario is that short-term rates evolve as expected and that forward rates are realized. At the opposite extreme, under the *pure risk premium hypothesis*, the market has no expectations about rates, that is, $E[dr/dt] = 0$, and the term structure of forward rates is determined by the risk premium. In this view of the world, the most natural "no-change" scenario is that short-term rates stay the same, as expected, which is an unchanged term structure. The reality, of course, can be between the two extremes, such the the term structure is determined by a mix of expectations and risk premium.

To conclude the discussion of Equation (13.20), the third term shows that the forward rate is reduced because of volatility and the convexity of the zero corresponding to the term of the forward rate by 0.5 $C\sigma^2$. Using this to reinterpret Equation (13.16), the indirect reduction in return through the forward rate, because of convexity, is exactly offset by the direct increase in return, because of convexity. Put another way, the expected return condition of Equation (13.18) ensures that there is no net advantage of convexity. The significance of this reasoning for investment and hedging decisions is introduced in the context of establishing long- and short-convexity positions.

14

The Art of Term Structure Models: Drift

Learning Objectives

After completing this reading, you should be able to:

- Construct and describe the effectiveness of a short-term interest rate tree assuming normally distributed rates, both with and without drift.
- Calculate the short-term rate change and standard deviation of the rate change using a model with normally distributed rates and no drift.
- Describe methods for addressing the possibility of negative short-term rates in term structure models.
- Construct a short-term rate tree under the Ho-Lee Model with time-dependent drift.
- Describe uses and benefits of the arbitrage-free models and assess the issue of fitting models to market prices.
- Describe the process of constructing a simple and recombining tree for a short-term rate under the Vasicek Model with mean reversion.
- Calculate the Vasicek Model rate change, standard deviation of the rate change, expected rate in T years, and half-life.
- Describe the effectiveness of the Vasicek Model.

Excerpt is Chapter 9 of Fixed Income Securities: Tools for Today's Markets, *Third Edition, by Bruce Tuckman and Angel Serrat.*

Assumptions about the true and risk-neutral short-term rate processes determine the term structure of interest rates and the prices of fixed income derivatives. The goal of this chapter is to describe the most common building blocks of short-term rate models. Selecting and rearranging these building blocks to create suitable models for the purpose at hand is the art of term structure modeling.

This chapter begins with an extremely simple model with no drift and normally distributed rates. The next sections add and discuss the implications of alternate specifications of the drift: a constant drift, a time-deterministic shift, and a mean-reverting drift.

14.1 MODEL 1: NORMALLY DISTRIBUTED RATES AND NO DRIFT

The particularly simple model of this section will be called Model 1. The continuously compounded, instantaneous rate r_t is assumed to evolve according to the following equation:

$$dr = \sigma dw \qquad (14.1)$$

The quantity dr denotes the change in the rate over a small time interval, dt, measured in years; σ denotes the annual *basis-point volatility* of rate changes; and dw denotes a normally distributed random variable with a mean of zero and a standard deviation of $\sqrt{dt}$.[1]

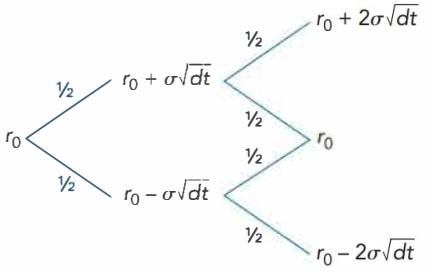
Say, for example, that the current value of the short-term rate is 6.18%, that volatility equals 113 basis points per year, and that the time interval under consideration is one month or 1/12 years. Mathematically, $r_0 = 6.18\%$; $\sigma = 1.13\%$; and $dt = 1/12$. A month passes and the random variable dw, with its zero mean and its standard deviation of $\sqrt{\frac{1}{12}}$ or .2887, happens to take on a value of .15. With these values, the change in the short-term rate given by (14.1) is

$$dr = 1.13\% \times .15 = .17\% \qquad (14.2)$$

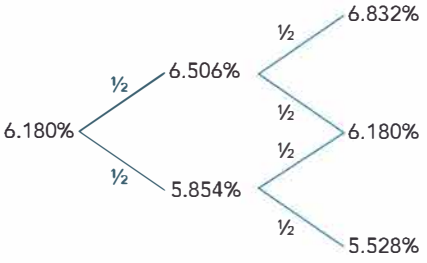
or 17 basis points. Since the short-term rate started at 6.18%, the short-term rate after a month is 6.35%.

Since the expected value of dw is zero, (14.1) says that the expected change in the rate, or the drift, is zero. Since the standard deviation of dw is $\sqrt{dt}$, the standard deviation of the change in the rate is $\sigma\sqrt{dt}$. For the sake of brevity, the standard deviation of the change in the rate will be referred to as simply the standard deviation of the rate. Continuing with the numerical example, the process (14.1) says that the drift is zero and that the standard deviation of the rate is $\sigma\sqrt{dt}$, which is $1.13\% \times \sqrt{\frac{1}{12}} = .326\%$ or 32.6 basis points per month.

A rate tree may be used to approximate the process (14.1). A tree over dates 0 to 2 takes the following form:



In the case of the numerical example, substituting the sample values into the tree gives the following:



To understand why these trees are representations of the process (14.1), consider the transition from date 0 to date 1. The change in the interest rate in the up-state is $\sigma\sqrt{dt}$ and the change in the down-state is $-\sigma\sqrt{dt}$. Therefore, with the probabilities given in the tree, the expected change in the rate, often denoted $E[dr]$, is

$$E[dr] = .5 \times \sigma\sqrt{dt} + .5 \times -\sigma\sqrt{dt} = 0 \qquad (14.3)$$

The variance of the rate, often denoted $V[dr]$, from date 0 to date 1 is computed as follows:

$$\begin{aligned} V[dr] &= E[dr^2] - \{E[dr]\}^2 \\ &= .5 \times (\sigma\sqrt{dt})^2 + .5 \times (-\sigma\sqrt{dt})^2 - 0 \\ &= \sigma^2 dt \end{aligned} \qquad (14.4)$$

Note that the first line of (14.4) follows from the definition of variance. Since the variance is $\sigma^2 dt$, the standard deviation, which is the square root of the variance, is $\sigma\sqrt{dt}$.

Equations (14.3) and (14.4) show that the drift and volatility implied by the tree match the drift and volatility of the interest rate process (14.1). The process and the tree are not identical because the random variable in the process, having a normal distribution, can take on any value while a single step in the tree leads to only two possible values. In the example, when dw takes on a value of .15, the short rate changes from 6.18%

[1] It is beyond the mathematical scope of the text to explain why the random variable dw is denoted as a change. But the text uses this notation since it is the convention of the field.

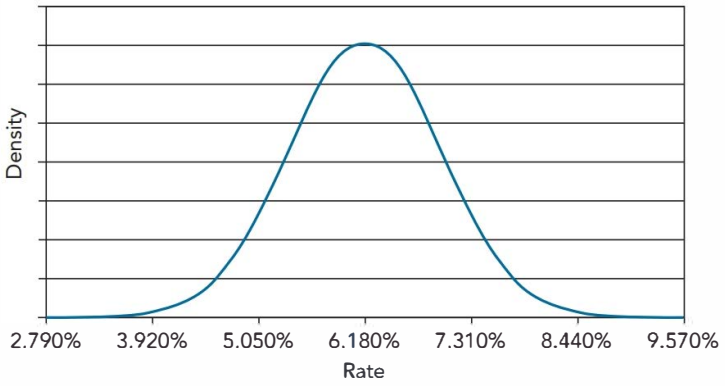


Figure 14.1 **Distribution of short rates after one year, Model 1.**

to 6.35%. In the tree, however, the only two possible rates are 6.506% and 5.854%. Nevertheless, after a sufficient number of time steps the branches of the tree used to approximate the process (14.1) will be numerous enough to approximate a normal distribution. Figure 14.1 shows the distribution of short rates after one year, or the *terminal distribution* after one year, in Model 1 with $r_0 = 6.18\%$ and $\sigma = 1.13\%$. The tick marks on the horizontal axis are one standard deviation apart from one another.

Models in which the terminal distribution of interest rates has a normal distribution, like Model 1, are called *normal* or *Gaussian* models. One problem with these models is that the short-term rate can become negative. A negative short-term rate does not make much economic sense because people would never lend money at a negative rate when they can hold cash and earn a zero rate instead.[2] The distribution in Figure 14.1, drawn to encompass three standard deviations above and below the mean, shows that over a horizon of one year the interest rate process will almost certainly not exhibit negative interest rates. The probability that the short-term rate in the process (14.1) becomes negative, however, increases with the horizon. Over 10 years, for example, the standard deviation of the terminal distribution in the numerical example is $1.13\% \times \sqrt{10}$ or 3.573%. Starting with a short-term rate of 6.18%, a random negative shock of only $6.18\%/3.573\%$ or 1.73 standard deviations would push rates below zero.

The extent to which the possibility of negative rates makes a model unusable depends on the application. For securities whose value depends mostly on the average path of the interest rate, like coupon bonds, the possibility of negative rates typically does not rule out an otherwise desirable model. For securities that are asymmetrically sensitive to the probability of low interest rates, however, using a normal model could be dangerous. Consider the extreme example of a 10-year call option to buy a long-term coupon bond at a yield of 0%. The model of this section would assign that option much too high a value because the model assigns too much probability to negative rates.

The challenge of negative rates for term structure models is much more acute, of course, when the current level of rates is low, as it is at the time of this writing. Changing the distribution of interest rates is one solution. To take but one of many examples, lognormally distributed rates, as will be seen in Chapter 15, cannot become negative. As will become clear later in that chapter, however, building a model around a probability distribution that rules out negative rates or makes them less likely may result in volatilities that are unacceptable for the purpose at hand.

Another popular method of ruling out negative rates is to construct rate trees with whatever distribution is desired, as done in this section, and then simply set all negative rates to zero.[3] In this methodology, rates in the original tree are called the shadow rates of interest while the rates in the adjusted tree could be called the observed rates of interest. When the observed rate hits zero, it can remain there for a while until the shadow rate crosses back to a positive rate. The economic justification for this framework is that the observed interest rate should be constrained to be positive only because investors have the alternative of investing in cash. But the shadow rate, the result of aggregate savings, investment, and consumption decisions, may very well be negative. Of course, the probability distribution of the observed interest rate is not the same as that of the originally postulated shadow rate. The change, however, is localized around zero and negative rates. By contrast, changing the form of the probability distribution changes dynamics across the entire range of rates.

Returning now to Model 1, techniques may be used to price fixed coupon bonds. Figure 14.2 graphs the semiannually compounded par, spot, and forward rate curves for the numerical example along with data from U.S. dollar swap par rates. The initial value of the short-term rate in the example, 6.18%, is set so that the model and market 10-year, semiannually compounded par rates are equal at 6.086%. All of the other data points shown are quite different from their model values. The desirability of fitting market data exactly is discussed in its own section, but Figure 14.2 clearly demonstrates that the simple model of this section does not have enough flexibility to capture the simplest of term structure shapes.

[2] Actually, the interest rate can be slightly negative if a security or bank account were safer or more convenient than holding cash.

[3] Fischer Black, "Interest Rates as Options," *Journal of Finance*, Vol. 50, 1995, pp. 1371–1376.

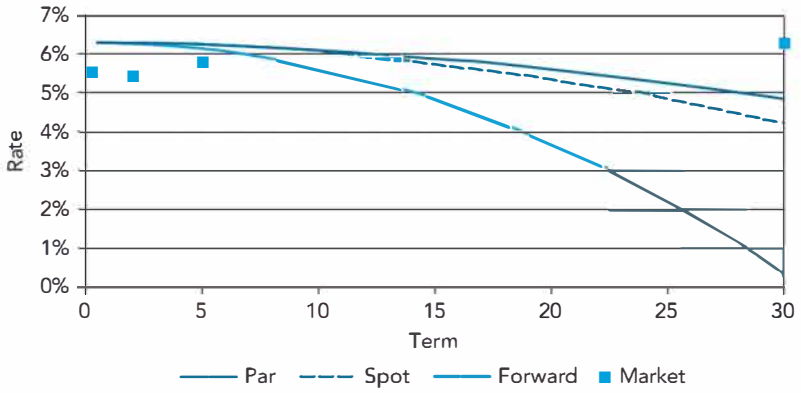


Figure 14.2 Rate curves from Model 1 and selected market swap rates, February 16, 2001.

Table 14.1 Convexity Effects on par Rates in a Parameterization of Model 1

Term (years)	Convexity (bps)
2	−0.8
5	−5.1
10	−18.8
30	−135.3

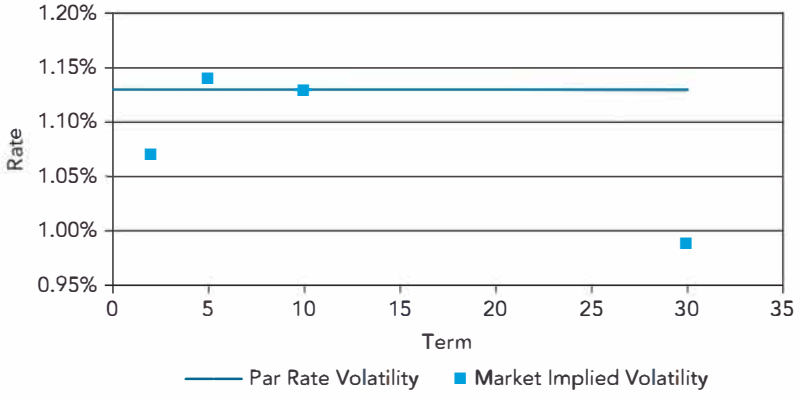


Figure 14.3 Par rate volatility from Model 1 and selected implied volatilities, February 16, 2001.

The model term structure is downward-sloping. As the model has no drift, rates decline with term solely because of convexity. Table 14.1 shows the magnitudes of the convexity effect on par rates of selected terms.[4] The numbers are realistic in the sense that a volatility of 113 basis points a year is reasonable. In fact, the volatility of the 10-year swap rate on the data date, as implied by options markets, was 113 basis points. The convexity numbers are not necessarily realistic, however, because, as this chapter will demonstrate, the magnitude of the convexity effect depends on the model and Model 1 is almost certainly not the best model of interest rate behavior.

The term structure of volatility in Model 1 is constant at 113 basis points per year. In other words, the standard deviation of changes in the par rate of any maturity is 113 basis points per year. As shown in Figure 14.3, this implication fails to capture the implied volatility structure in the market. The volatility data in Figure 14.3 show that the term structure of volatility is humped, i.e., that volatility initially rises with term but eventually declines. As this shape is a feature of fixed income markets, it will be revisited again in this chapter and in Chapter 15.

The last aspect of this model to be analyzed is its factor structure. The model's only factor is the short-term rate. If this rate increases by 10 semiannually compounded basis points, how would the term structure change? In this simple model, the answer is that all rates would increase by 10 basis points. (See the closed-form solution for spot rates in Model 1 in the Appendix in Chapter 15). Therefore, Model 1 is a model of parallel shifts.

14.2 MODEL 2: DRIFT AND RISK PREMIUM

The term structures implied by Model 1 always look like Figure 14.2: relatively flat for early terms and then downward sloping. Term structure tends to slope upward and that this behavior might be explained by the existence of a risk premium. The model of this section, to be called Model 2, adds a drift to Model 1, interpreted as a risk premium, in order to obtain a richer model in an economically coherent way.

The dynamics of the risk-neutral process in Model 2 are written as

$$dr = \lambda dt + \sigma dw \tag{14.5}$$

The process (14.5) differs from that of Model 1 by adding a drift to the short-term rate equal to λdt. For this section, consider the values $r_0 = 5.138\%$, $\lambda = .229\%$, and $\sigma = 1.10\%$. If the

[4] The convexity effect is the difference between the par yield in the model with its assumed volatility and the par yield in the same structural model but with a volatility of zero.

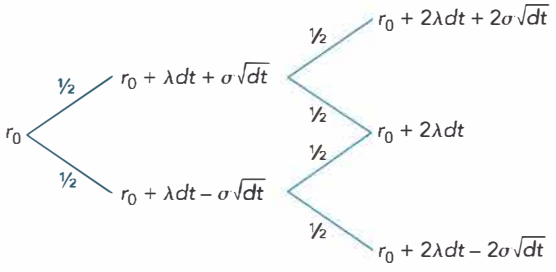
realization of the random variable dw is again .15 over a month, then the change in rate is

$$dr = .229\% \times \frac{1}{12} + 1.10\% \times .15 = .1841\% \qquad (14.6)$$

Starting from 5.138%, the new rate is 5.322%.

The drift of the rate is .229% × 1/12 or 1.9 basis points per month, and the standard deviation is $1.10\% \times \sqrt{\frac{1}{12}}$ or 31.75 basis points per month. The drift in the risk-neutral process is a combination of the true expected change in the interest rate and of a risk premium. A drift of 1.9 basis points per month may arise because the market expects the short-term rate to increase by 1.9 basis points a month, because the short-term rate is expected to increase by one basis point with a risk premium of .9 basis points, or because the short-term rate is expected to fall by .1 basis points with a risk premium of two basis points.

The tree approximating this model is



It is easy to verify that the drift and standard deviation of the tree match those of the process (14.5).

The terminal distribution of the numerical example of this process after one year is normal with a mean of 5.138% + 1 × .229% or 5.367% and a standard deviation of 110 basis points. After 10 years, the terminal distribution is normal with a mean of 5.138% + 10 × .229% or 7.428% and a standard deviation of $1.10\% \times \sqrt{10}$ or 347.9 basis points. Note that the constant drift, by raising the mean of the terminal distribution, makes it less likely that the risk-neutral process will exhibit negative rates.

Figure 14.4 shows the rate curves in this example along with par swap rate data. The values of r_0 and λ are calibrated to match the 2- and 10-year par swap rates, while the value of σ is chosen to be the average implied volatility of the 2- and 10-year par rates. The results are satisfying in that the resulting curve can match the data much more closely than did the curve of Model 1 shown in Figure 14.2. Slightly unsatisfying is the relatively high value of λ required. Interpreted as a risk premium alone, a value of .229% with a volatility of 110 basis points implies a relatively high Sharpe ratio of about .21. On the other hand, interpreting λ as a combination of true drift and risk premium is difficult in the long end where, it is difficult to make a case for rising expected rates. These interpretive difficulties arise because Model 2 is still not flexible enough to explain the shape of the term structure in an economically meaningful way. In fact, the use of r_0 and λ to match the 2- and 10-year rates in this relatively inflexible model may explain why the model curve overshoots the 30-year par rate by about 25 basis points.

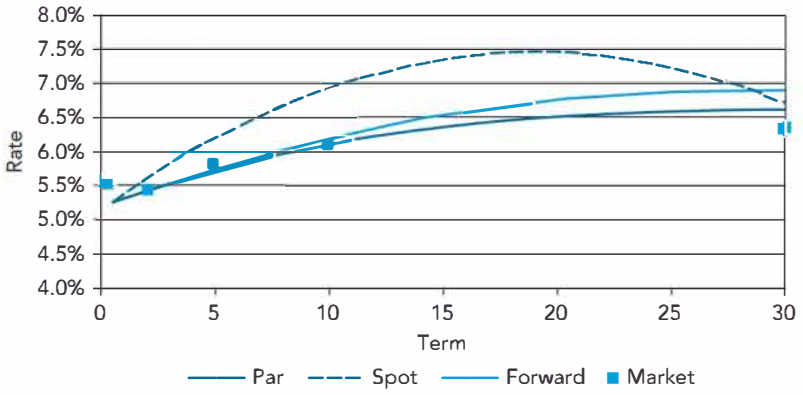


Figure 14.4 **Rate curves from Model 2 and selected market swap rates, February 16, 2001.**

Moving from Model 1 with zero drift to Model 2 with a constant drift does not qualitatively change the term structure of volatility, the magnitude of convexity effects, or the parallel-shift nature of the model.

Models 1 and 2 would be called equilibrium models because no effort has been made to match the initial term structure closely. The next section presents a generalization of Model 2 that is in the class of arbitrage-free models.

14.3 THE HO-LEE MODEL: TIME-DEPENDENT DRIFT

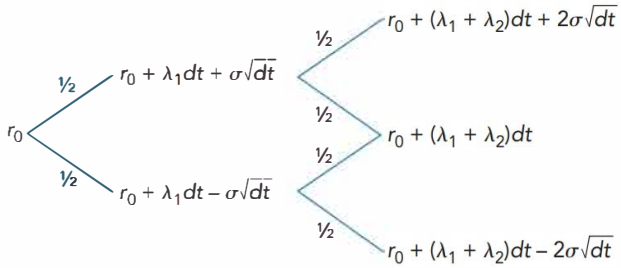
The dynamics of the risk-neutral process in the Ho-Lee model are written as

$$dr = \lambda_t dt + \sigma dw \qquad (14.7)$$

In contrast to Model 2, the drift here depends on time. In other words, the drift of the process may change from date to date. It might be an annualized drift of −20 basis points over the first month, of 20 basis points over the second month, and so on. A drift that varies with time is called a *time-dependent* drift. Just

as with a constant drift, the time-dependent drift over each time period represents some combination of the risk premium and of expected changes in the short-term rate.

The flexibility of the Ho-Lee model is easily seen from its corresponding tree:



The free parameters λ_1 and λ_2 may be used to match the prices of securities with fixed cash flows. The procedure may be described as follows. With $dt = 1/12$, set r_0 equal to the one-month rate. Then find λ_1 such that the model produces a two-month spot rate equal to that in the market. Then find λ_2 such that the model produces a three-month spot rate equal to that in the market. Continue in this fashion until the tree ends.

The rate curves resulting from this model match all the rates that are input into the model. Just as adding a constant drift to Model 1 to obtain Model 2 does not affect the shape of the term structure of volatility nor the parallel-shift characteristic of the model, adding a time-dependent drift does not change these features either.

14.4 DESIRABILITY OF FITTING TO THE TERM STRUCTURE

The desirability of matching market prices is the central issue in deciding between arbitrage-free and equilibrium models. Not surprisingly, the choice depends on the purpose of building the model in the first place.

One important use of arbitrage-free models is for quoting the prices of securities that are not actively traded based on the prices of more liquid securities. A customer might ask a swap desk to quote a rate on a swap to a particular date, say three years and four months away, while liquid market prices might be observed only for three- and four-year swaps, or sometimes only for two- and five-year swaps. In this situation, the swap desk may price the odd-maturity swap using an arbitrage-free model essentially as a means of interpolating between observed market prices.

Interpolating by means of arbitrage-free models may very well be superior to other curve-fitting methods, from linear interpolation to more sophisticated approaches. The potential superiority of arbitrage-free models arises from their being based on economic and financial reasoning. In an arbitrage-free model, the expectations and risk premium built into neighboring swap rates and the convexity implied by the model's volatility assumptions are used to compute, for example, the three-year and four-month swap rate. In a purely mathematical curve fitting technique, by contrast, the chosen functional form heavily determines the intermediate swap rate. Selecting linear or quadratic interpolation, for example, results in intermediate swap rates with no obvious economic or financial justification. This potential superiority of arbitrage-free models depends crucially on the validity of the assumptions built into the models. A poor volatility assumption, for example, resulting in a poor estimate of the effect of convexity, might make an arbitrage-free model perform worse than a less financially sophisticated technique.

Another important use of arbitrage-free models is to value and hedge derivative securities for the purpose of making markets or for proprietary trading. For these purposes, many practitioners wish to assume that some set of underlying securities is priced fairly. For example, when trading an option on a 10-year bond, many practitioners assume that the 10-year bond is itself priced fairly. (An analysis of the fairness of the bond can always be done separately.) Since arbitrage-free models match the prices of many traded securities by construction, these models are ideal for the purpose of pricing derivatives given the prices of underlying securities.

That a model matches market prices does not necessarily imply that it provides fair values and accurate hedges for derivative securities. The argument for fitting models to market prices is that a good deal of information about the future behavior of interest rates is incorporated into market prices, and, therefore, a model fitted to those prices captures that interest rate behavior. While this is a perfectly reasonable argument, two warnings are appropriate. First, a mediocre or bad model cannot be rescued by calibrating it to match market prices. If, for example, the parallel shift assumption is not a good enough description of reality for the application at hand, adding a time-dependent drift to a parallel shift model so as to match a set of market prices will not make the model any more suitable for that application. Second, the argument for fitting to market prices assumes that those market prices are fair in the context of the model. In many situations, however, particular securities, particular classes of securities, or particular maturity ranges of securities have been distorted due to supply and demand imbalances,

taxes, liquidity differences, and other factors unrelated to interest rate models. In these cases, fitting to market prices will make a model worse by attributing these outside factors to the interest rate process. If, for example, a large bank liquidates its portfolio of bonds or swaps with approximately seven years to maturity and, in the process, depresses prices and raises rates around that maturity, it would be incorrect to assume that expectations of rates seven years in the future have risen. Being careful with the word *fair*, the seven-year securities in this example are fair in the sense that liquidity considerations at a particular time require their prices to be relatively low. The seven-year securities are not fair, however, with respect to the expected evolution of interest rates and the market risk premium. For this reason, in fact, investors and traders might buy these relatively cheap bonds or swaps and hold them past the liquidity event in the hope of selling at a profit.

Another way to express the problem of fitting the drift to the term structure is to recognize that the drift of a risk-neutral process arises only from expectations and risk premium. A model that assumes one drift from years 15 to 16 and another drift from years 16 to 17 implicitly assumes one of two things. First, the expectation today of the one-year rate in 15 years differs from the expectation today of the one-year rate in 16 years. Second, the risk premium in 15 years differs in a particular way from the risk premium in 16 years. Since neither of these assumptions is particularly plausible, a fitted drift that changes dramatically from one year to the next is likely to be erroneously attributing non-interest rate effects to the interest rate process.

If the purpose of a model is to value bonds or swaps relative to one another, then taking a large number of bond or swap prices as given is clearly inappropriate: arbitrage-free models, by construction, conclude that all of these bond or swap prices are fair relative to one another. Investors wanting to choose among securities, market makers looking to pick up value by strategically selecting hedging securities, or traders looking to profit from temporary mispricings must, therefore, rely on equilibrium models.

Having starkly contrasted arbitrage-free and equilibrium models, it should be noted that, in practice, there need not be a clear line between the two approaches. A model might posit a deterministic drift for a few years to reflect relatively short-term interest rate forecasts and posit a constant drift from then on. Another model might take the prices of 2-, 5-, 10- and 30-year bond or swap rates as given, thus assuming that the most liquid securities are fair while allowing the model to value other securities. The proper blending of the arbitrage-free and equilibrium approaches is an important part of the art of term structure modeling.

14.5 THE VASICEK MODEL: MEAN REVERSION

Assuming that the economy tends toward some equilibrium based on such fundamental factors as the productivity of capital, long-term monetary policy, and so on, short-term rates will be characterized by *mean reversion*. When the short-term rate is above its long-run equilibrium value, the drift is negative, driving the rate down toward this long-run value. When the rate is below its equilibrium value, the drift is positive, driving the rate up toward this value. In addition to being a reasonable assumption about short rates,[5] mean reversion enables a model to capture several features of term structure behavior in an economically intuitive way.

The risk-neutral dynamics of the Vasicek model[6] are written as

$$dr = k(\theta - r)dt + \sigma dw \tag{14.8}$$

The constant θ denotes the long-run value or central tendency of the short-term rate in the risk-neutral process and the positive constant k denotes the speed of mean reversion. Note that in this specification, the greater the difference between r and θ, the greater the expected change in the short-term rate toward θ.

Because the process (14.8) is the risk-neutral process, the drift combines both interest rate expectations and risk premium. Furthermore, market prices do not depend on how the risk-neutral drift is divided between the two. Nevertheless, in order to understand whether or not the parameters of a model make sense, it is useful to make assumptions sufficient to separate the drift and the risk premium. Assuming, for example, that the true interest rate process exhibits mean reversion to a long-term value r_∞ and, as assumed previously, that the risk premium

[5] While reasonable, mean reversion is a strong assumption. Long time series of interest rates from relatively stable markets might display mean reversion because there happened to be no catastrophe over the time period, that is, precisely because a long time series exists. Hyperinflation, for example, is not consistent with mean reversion and results in the destruction of a currency and its associated interest rates. When mean reversion ends, the time series ends. In short, the most severe critics of mean reversion would say that interest rates mean revert until they don't.

[6] O. Vasicek, "An Equilibrium Characterization of the Term Structure, *Journal of Financial Economics*, 5, 1977, pp. 177–188. It is appropriate to add that this paper started the literature on short-term rate models. The particular dynamics of the model described in this section, which is commonly known as the Vasicek model, is a very small part of the contribution of that paper.

enters into the risk-neutral process as a constant drift, the Vasicek model takes the following form:

$$\begin{aligned} dr &= k(r_\infty - r)dt + \lambda dt + \sigma dw \\ &= k\left(\left[r_\infty + \frac{\lambda}{k}\right] - r\right)dt + \sigma dw \end{aligned} \qquad (14.9)$$

The process in (14.8) is identical to that in (14.9) so long as

$$\theta \equiv r_\infty + \frac{\lambda}{k} \qquad (14.10)$$

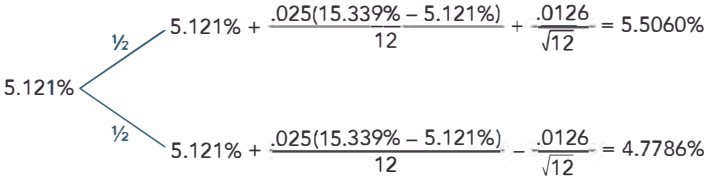
Note that very many combinations of r_∞ and λ give the same θ and, through the risk-neutral process (14.8), the same market prices.

For the purposes of this section, let $k = .025$, $\sigma = 126$ basis points per year, $r_\infty = 6.179\%$, and $\lambda = .229\%$. According to (14.10), then, $\theta = 15.339\%$. With these parameters, the process (14.8) says that over the next month the expected change in the short rate is

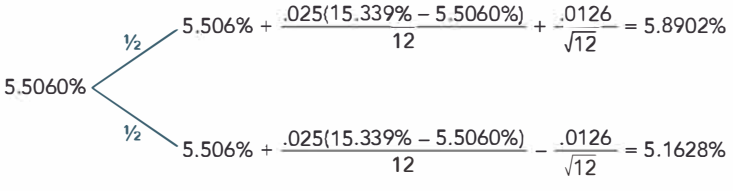
$$.025 \times (15.339\% - 5.121\%)\frac{1}{12} = .0213\% \qquad (14.11)$$

or 2.13 basis points. The volatility over the next month is $126 \times \sqrt{\frac{1}{12}}$ or 36.4 basis points.

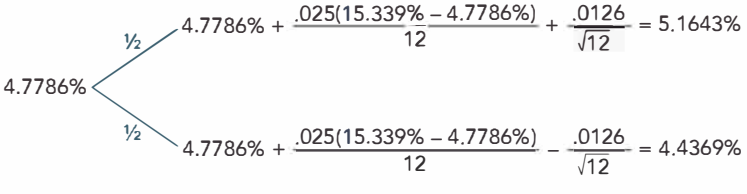
Representing this process with a tree is not quite so straightforward as the simpler processes described previously because the most obvious representation leads to a nonrecombining tree. Over the first time step,



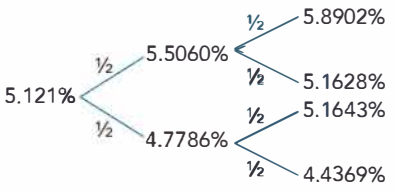
To extend the tree from date 1 to date 2, start from the up state of 5.5060%. The tree branching from there is



while the tree branching from the date 1 down-state of 4.7786% is



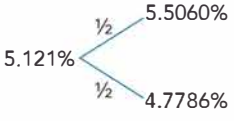
To summarize, the most straightforward tree representation of (14.8) takes the following form:



This tree does not recombine since the drift increases with the difference between the short rate and θ. Since 4.7786% is further from θ than 5.5060%, the drift from 4.7786% is greater than the drift from 5.5060%. In this model, the volatility component of an up move followed by a down move does perfectly cancel the volatility component of a down move followed by an up move. But since the drift from 4.7786% is greater, the move up from 4.7786% produces a larger short-term rate than a move down from 5.5060%.

There are many ways to represent the Vasicek model with a recombining tree. One method is presented here, but it is beyond the scope of this book to discuss the numerical efficiency of the various possibilities.

The first time step of the tree may be taken as shown previously:



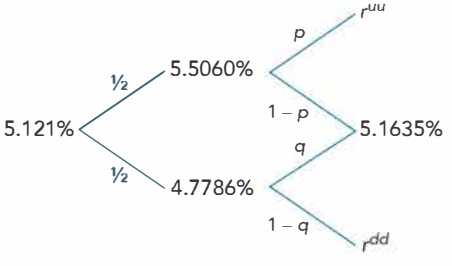
Next, fix the center node of the tree on date 2. Since the expected perturbation due to volatility over each time step is zero, the drift alone determines the expected value of the process after each time step. After the first time step, the expected value is

$$5.121\% + .025\,(15.339\% - 5.121\%)\frac{1}{2} = 5.1423\% \qquad (14.12)$$

After the second time step, the expected value is

$$5.1423\% + .025\,(15.339\% - 5.1423\%)\frac{1}{12} = 5.1635\% \qquad (14.13)$$

Take this value as the center node on date 2 of the recombining tree:



The parts of the tree to be solved for, namely, the missing probabilities and interest rate values, are given variable names.

According to the process (14.8) and the parameter values set in this section, the expected rate and standard deviation of the rate from 5.5060% are, respectively,

$$5.5060\% + .025\,(15.339\% - 5.5060\%)\,\frac{1}{12} = 5.5265\% \qquad (14.14)$$

and

$$1.26\%\sqrt{\frac{1}{12}} = .3637\% \qquad (14.15)$$

For the recombining tree to match this expectation and standard deviation, it must be the case that

$$p \times r^{uu} + (1 - p) \times 5.1635\% = 5.5265\% \qquad (14.16)$$

and, by the definition of standard deviation,

$$\sqrt{p(r^{uu} - 5.5265\%)^2 + (1 - p)(5.6135\% - 5.5265\%)^2} = .3637\% \qquad (14.17)$$

Solving Equations (14.16 and (14.17), $r^{uu} = 5.8909\%$ and $p = .4990$.

The same procedure may be followed to compute r^{dd} and q. The expected rate from 4.7786% is

$$4.7786\% + .025\,(15.339\% - 4.7786\%)\,\frac{1}{12} = 4.8006\%. \qquad (14.18)$$

and the standard deviation is again 36.37 basis points. Starting from 4.7786%, then, it must be the case that

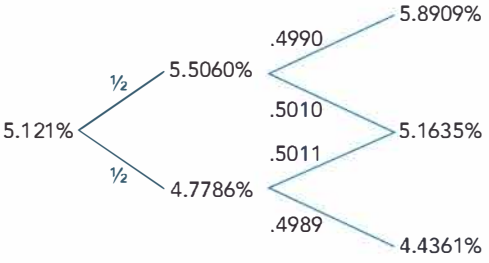
$$q \times 5.1635\% + (1 - q) \times r^{dd} = 4.8006\% \qquad (14.19)$$

and

$$\sqrt{q(5.1635\% - 4.8006\%)^2 + (1 + q)(r^{dd} - 4.8006\%)^2} = .3637\% \qquad (14.20)$$

Solving Equations (14.19) and (14.20), $r^{dd} = 4.4361\%$ and $q = .5011$.

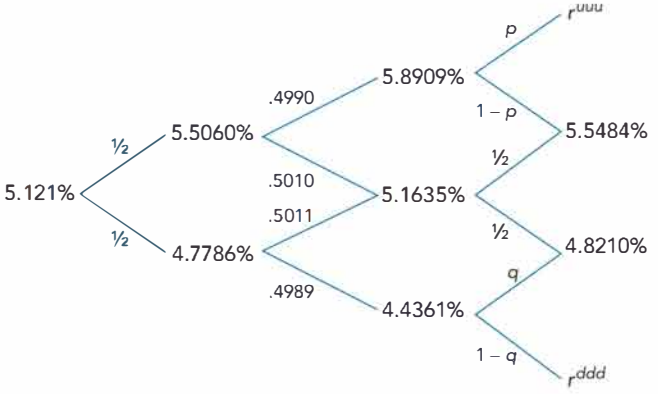
Putting the results from the up- and downstates together, a recombining tree approximating the process (14.8) with the parameters of this section is



To extend the tree to the next date, begin again at the center. From the center node of date 2, the expected rate of the process is

$$5.1635\% + .025 \times (15.339\% - 5.1635\%)\,\frac{1}{12} = 5.1847\% \qquad (14.21)$$

As in constructing the tree for date 1, adding and subtracting the standard deviation of .3637% to the average value 5.1847% (obtaining 5.5484% and 4.8210%) and using probabilities of 50% for up and down movements satisfy the requirements of the process at the center of the tree:



The unknown parameters can be solved for in the same manner as described in building the tree on date 2.

The text now turns to the effects of mean reversion on the term structure. Figure 14.5 illustrates the impact of mean reversion on the terminal, risk-neutral distributions of the short rate at different horizons. The expectation or mean of the short-term rate as a function of horizon gradually rises from its current value of 5.121% toward its limiting value of $\theta = 15.339\%$. Because the mean-reverting parameter $k = .025$ is relatively small, the horizon expectation rises very slowly toward 15.339%. While

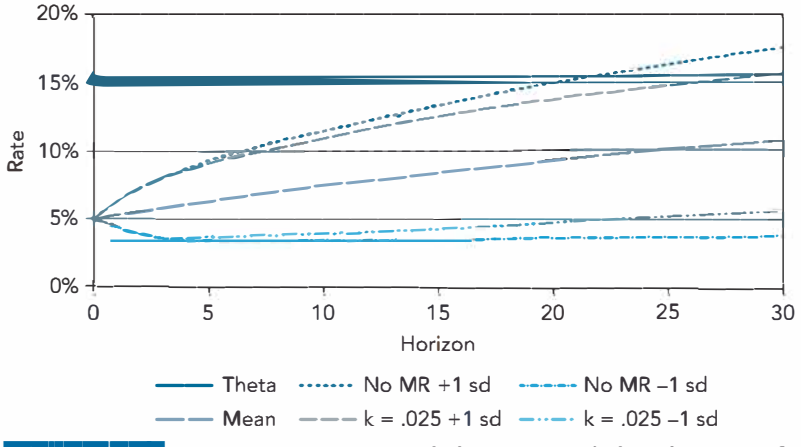


Figure 14.5 **Mean reversion and the terminal distribution of short rates.**

mathematically beyond the scope of this book, it can be shown that the distance between the current value of a factor and its goal decays exponentially at the mean-reverting rate. Since the interest rate is currently 15.339% − 5.121% or 10.218% away from its goal, the distance between the expected rate at a 10-year horizon and the goal is

$$10.2180\% \times e^{-.025\times 10} = 7.9578\% \qquad (14.22)$$

Therefore, the expectation of the rate in 10 years is 15.3390% − 7.9578% or 7.3812%.

For completeness, the expectation of the rate in the Vasicek model after T years is

$$r_0 e^{-kT} + \theta(1 - e^{-kT}) \qquad (14.23)$$

In words, the expectation is a weighted average of the current short rate and its long-run value, where the weight on the current short rate decays exponentially at a speed determined by the mean-reverting parameter.

The mean-reverting parameter is not a particularly intuitive way of describing how long it takes a factor to revert to its long-term goal. A more intuitive quantity is the factor's *half-life*, defined as the time it takes the factor to progress half the distance toward its goal. In the example of this section, the half-life of the interest rate, τ, is given by the following equation:

$$(15.339\% - 5.121\%)e^{-.025\tau} = \frac{1}{2}(15.339\% - 5.121\%) \qquad (14.24)$$

Solving,

$$e^{-.025\tau} = \frac{1}{2}$$

$$\tau = \frac{\ln(2)}{.025}$$

$$\tau = 27.73 \qquad (14.25)$$

where $\ln(\cdot)$ is the natural logarithm function. In words, the interest rate factor takes 27.73 years to cover half the distance between its starting value and its goal. This can be seen visually in Figure 14.5 where the expected rate 30 years from now is about halfway between its current value and θ. Larger mean-reverting parameters produce shorter half lives.

Figure 14.5 also shows one-standard deviation intervals around expectations both for the mean-reverting process of this section and for a process with the same expectation and the same σ but without mean reversion ("No MR"). The standard deviation of the terminal distribution of the short rate after T years in the Vasicek model is

$$\sqrt{\frac{\sigma^2}{2k}(1 - e^{-2kT})} \qquad (14.26)$$

In the numerical example, with a mean-reverting parameter of .025 and a volatility of 126 basis points, the short rate in 10 years is normally distributed with an expected value of 7.3812%, derived earlier, and a standard deviation of

$$\sqrt{\frac{.0126^2}{2 \times .025}(1 - e^{-2\times.025\times 10})} \qquad (14.27)$$

or 353 basis points. Using the same expected value and σ but no mean reversion the standard deviation is $\sigma\sqrt{T} = 1.26\%\sqrt{10}$ or 398 basis points. Pulling the interest rate toward a long-term goal dampens volatility relative to processes without mean reversion, particularly at long horizons.

To avoid confusion in terminology, note that the mean-reverting model in this section sets volatility equal to 125 basis points "per year." Because of mean reversion, however, this does not mean that the standard deviation of the terminal distribution after T years increases with the square root of time. Without mean reversion, this is the case, as mentioned in the previous paragraph. With mean reversion, the standard deviation increases with horizon more slowly than that, producing a standard deviation of only 353 basis points after 10 years.

Figure 14.6 graphs the rate curves in this parameterization of the Vasicek model. The values of r_0 and θ were calibrated to match the 2- and 10-year par rates in the market. As a result, Figure 14.6 qualitatively resembles Figure 14.4. The mean reversion parameter might have been used to make the model fit the observed term structure more closely, but, as discussed in the next paragraph, this parameter was used to produce a particular term structure of volatility. In conclusion, Figure 14.6 shows that the model as calibrated in this section is probably not flexible enough to produce the range of term structures observed in practice.

A model with mean reversion and a model without mean reversion result in dramatically different term structures of volatility.

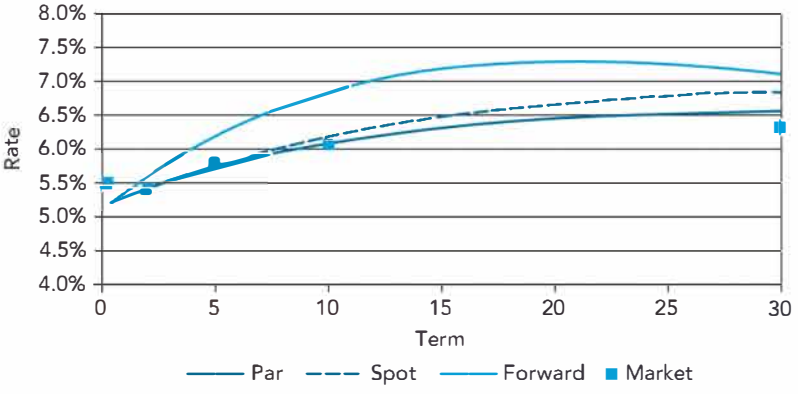


Figure 14.6 **Rate curves from the Vasicek model and selected market swap rates, February 16, 2001.**

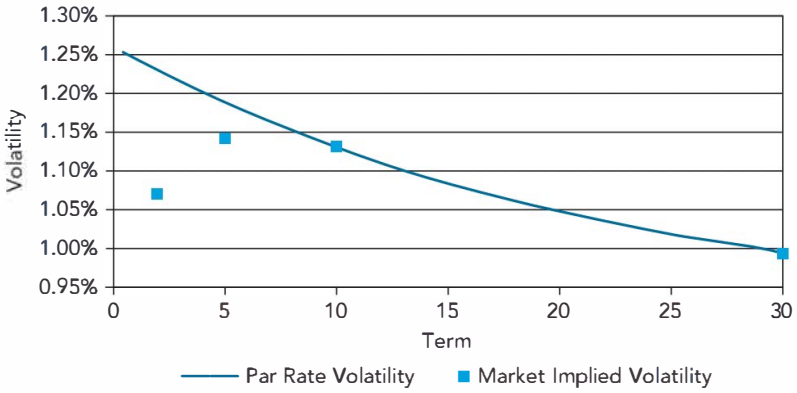


Figure 14.7 **Par rate volatility from the Vasicek model and selected implied volatilities, February 16, 2001.**

Table 14.2 **Convexity Effects on par Rates in a Parameterization of the Vasicek Model**

Term (years)	Convexity (bps)
2	−1.0
5	−5.8
10	−19.1
30	−74.7

Figure 14.7 shows that the volatilities of par rates decline with term in the Vasicek model. In this example the mean reversion and volatility parameters are chosen to fit the implied 10- and 30-year volatilities. As a result, the model matches the market at those two terms but overstates the volatility for shorter terms. While Figure 14.7 certainly shows an improvement relative to the flat term structure of volatility shown in Figure 14.3, mean reversion in this model generates a term structure of volatility that slopes downward everywhere.

Since mean reversion lowers the volatility of longer-term par rates, it must also lower the impact of convexity on these rates. Table 14.2 reports the convexity effect at several terms. Recall that the convexity effects listed in Table 14.1 are generated from a model with no mean reversion and a volatility of 113 basis points per year. Since this section sets volatility equal to 126 basis points per year and since mean reversion is relatively slow, the convexity effects for terms up to 10 years are slightly larger in Table 14.2 than in Table 14.1. But by a term of 30 years the dampening effect of mean reversion on volatility manifests itself, and the convexity effect in the Vasicek model of about 75 basis points is substantially below the 135 basis point in the model without mean reversion.

Figure 14.8 shows the shape of the interest rate factor in a mean-reverting model, that is, how the spot rate curve is affected by a 10-basis point increase in the short-term rate. By definition, short-term rates rise by about 10 basis points but longer term rates are impacted less. The 30-year spot rate, for example, rises by only 7 basis points. Hence a model with mean reversion is not a parallel shift model.

The implications of mean reversion for the term structure of volatility and factor shape may be better understood by reinterpreting the assumption that short rates tend toward a long-term goal. Assuming that short rates move as a result of some news or shock to the economic system, mean reversion implies that the effect of this shock eventually dissipates. After all, regardless of the shock, the short rate is assumed to arrive ultimately at the same long-term goal.

Economic news is said to be *long-lived* if it changes the market's view of the economy many years in the future. For example, news of a technological innovation that raises productivity would be a relatively long-lived shock to the system. Economic news is said to be *short-lived* if it changes the market's view of the economy in the near but not far future. An example of this kind of shock might be news that retail sales were lower than expected due to excessively cold weather over the holiday season. In this interpretation, mean reversion measures the length of economic news in a term structure model. A very low mean reversion parameter, i.e., a very long half-life, implies that news is long-lived and that it will affect the short rate for many years to come. On the other hand, a very high mean reversion parameter, i.e., a very short half-life, implies that news is short-lived and that it affects the short rate for a relatively short period of time. In reality, of course, some news is short-lived while other news is long-lived, a feature captured by the multi-factor Gauss + model.

Interpreting mean reversion as the length of economic news explains the factor structure and the downward-sloping term

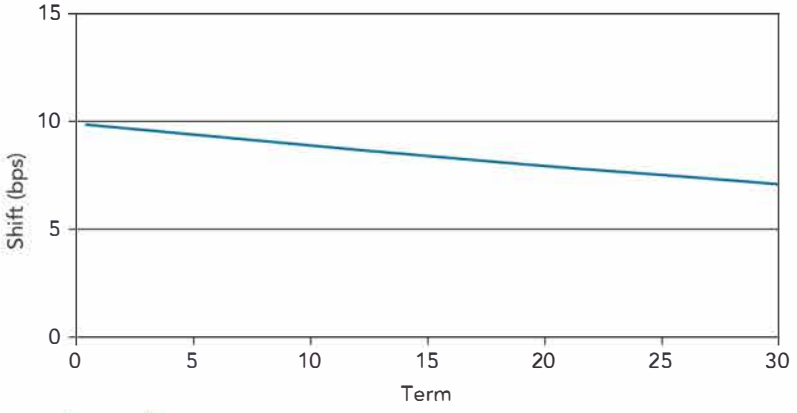


Figure 14.8 **Sensitivity of spot rates in the Vasicek model to a 10-basis-point change in the factor.**

structure of volatility in the Vasicek model. Rates of every term are combinations of current economic conditions, as measured by the short-term rate, and of long-term economic conditions, as measured by the long-term value of the short rate (i.e., θ). In a model with no mean reversion, rates are determined exclusively by current economic conditions. Shocks to the short-term rate affect all rates equally, giving rise to parallel shifts and a flat term structure of volatility. In a model with mean reversion, short-term rates are determined mostly by current economic conditions while longer-term rates are determined mostly by long-term economic conditions. As a result, shocks to the short rate affect short-term rates more than longer-term rates and give rise to a downward-sloping term structure of volatility and a downward-sloping factor structure.

15

The Art of Term Structure Models: Volatility and Distribution

Learning Objectives

After completing this reading, you should be able to:

- Describe the short-term rate process under a model with time-dependent volatility.
- Calculate the short-term rate change and determine the behavior of the standard deviation of the rate change using a model with time-dependent volatility.
- Assess the efficacy of time-dependent volatility models.
- Describe the short-term rate process under the Cox-Ingersoll-Ross (CIR) and lognormal models.
- Calculate the short-term rate change and describe the basis point volatility using the CIR and lognormal models.
- Describe lognormal models with deterministic drift and mean reversion.

Excerpt is Chapter 10 of Fixed Income Securities: Tools for Today's Markets, *Third Edition, by Bruce Tuckman and Angel Serrat.*

This chapter continues the presentation of the elements of term structure modeling, focusing on the volatility of interest rates and on models in which rates are not normally distributed.

15.1 TIME-DEPENDENT VOLATILITY: MODEL 3

Just as a time-dependent drift may be used to fit many bond or swap rates, a time-dependent volatility function may be used to fit many option prices. A particularly simple model with a time-dependent volatility function might be written as follows:

$$dr = \lambda(t)dt + \sigma(t)dw \qquad (15.1)$$

Unlike the models presented in Chapter 14, the volatility of the short rate in Equation (15.1) depends on time. If, for example, the function $\sigma(t)$ were such that $\sigma(1) = 1.26\%$ and $\sigma(2) = 1.20\%$, then the volatility of the short rate in one year is 126 basis points per year while the volatility of the short rate in two years is 120 basis points per year.

To illustrate the features of time-dependent volatility, consider the following special case of (15.1) that will be called Model 3:

$$dr = \lambda(t)dt + \sigma e^{-\alpha t}dw \qquad (15.2)$$

In (15.2), the volatility of the short rate starts at the constant σ and then exponentially declines to zero. Volatility could have easily been designed to decline to another constant instead of zero, but Model 3 serves its pedagogical purpose well enough.

Setting $\sigma = 126$ basis points and $\alpha = .025$, Figure 15.1 graphs the standard deviation of the terminal distribution of the short rate at various horizons.[1] Note that the standard deviation rises rapidly with horizon at first but then rises more slowly. The particular shape of the curve depends, of course, on the volatility function chosen for (15.2), but very many shapes are possible with the more general volatility specification in (15.1).

Deterministic volatility functions are popular, particularly among market makers in interest rate options. Consider the example of caplets. At expiration, a caplet pays the difference between the short rate and a strike, if positive, on some notional amount. Furthermore, the value of a caplet depends on the distribution of the short rate at the caplet's expiration. Therefore, the flexibility of the deterministic functions $\lambda(t)$ and $\sigma(t)$ may be used to match the market prices of caplets expiring on many different dates.

[1] This result is presented without derivation.

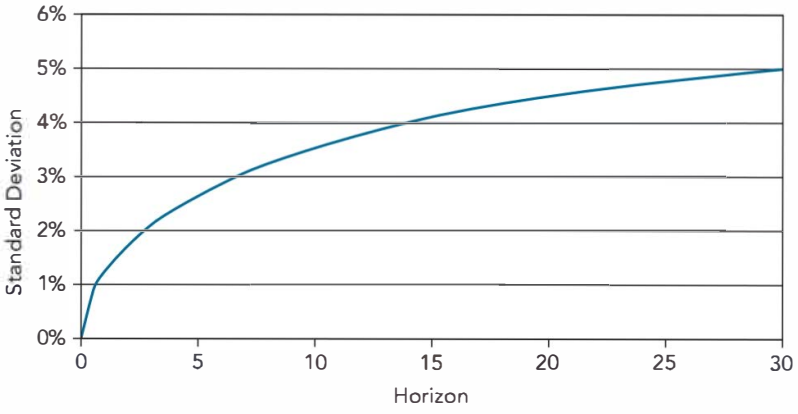


Figure 15.1 Standard deviation of terminal distributions of short rates in Model 3.

The behavior of standard deviation as a function of horizon in Figure 15.1 resembles the impact of mean reversion on horizon standard deviation in Figure 13.5. In fact, setting the initial volatility and decay rate in Model 3 equal to the volatility and mean reversion rate of the numerical example of the Vasicek model, the standard deviations of the terminal distributions from the two models turn out to be identical. Furthermore, if the time-dependent drift in Model 3 matches the average path of rates in the numerical example of the Vasicek model, then the two models produce exactly the same terminal distributions.

While these parameterizations of the two models give equivalent terminal distributions, the models remain very different in other ways. As is the case for any model without mean reversion, Model 3 is a parallel shift model. Also, the term structure of volatility in Model 3 is flat. Since the volatility in Model 3 changes over time, the term structure of volatility is flat at levels that change over time, but it is still always flat.

The arguments for and against using time-dependent volatility resemble those for and against using a time-dependent drift. If the purpose of the model is to quote fixed income options prices that are not easily observable, then a model with time-dependent volatility provides a means of interpolating from known to unknown option prices. If, however, the purpose of the model is to value and hedge fixed income securities, including options, then a model with mean reversion might be preferred for two reasons.

First, while mean reversion is based on the economic intuitions outlined earlier, time-dependent volatility relies on the difficult argument that the market has a forecast of short-term volatility in the distant future. A modification of the model that addresses this objection, by the way, is to assume that volatility depends on time in the near future and then settles at a constant.

Second, the downward-sloping factor structure and term structure of volatility in mean-reverting models capture the

behavior of interest rate movements better than parallel shifts and a flat term structure of volatility. It may very well be that the Vasicek model does not capture the behavior of interest rates sufficiently well to be used for a particular valuation or hedging purpose. But in that case it is unlikely that a parallel shift model calibrated to match caplet prices will be better suited for that purpose.

15.2 THE COX-INGERSOLL-ROSS AND LOGNORMAL MODELS: VOLATILITY AS A FUNCTION OF THE SHORT RATE

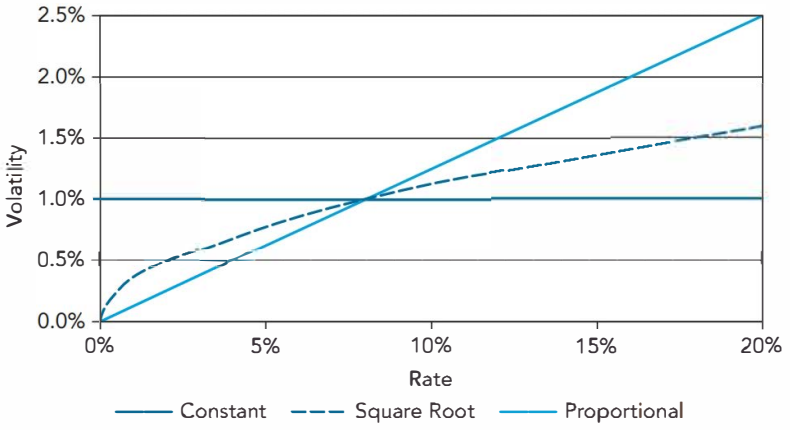


Figure 15.2 Three volatility specifications.

The models presented so far assume that the basis-point volatility of the short rate is independent of the level of the short rate. This is almost certainly not true at extreme levels of the short rate. Periods of high inflation and high short-term interest rates are inherently unstable and, as a result, the basis-point volatility of the short rate tends to be high. Also, when the short-term rate is very low, its basis-point volatility is limited by the fact that interest rates cannot decline much below zero.

Economic arguments of this sort have led to specifying the basis-point volatility of the short rate as an increasing function of the short rate. The risk-neutral dynamics of the Cox-Ingersoll-Ross (CIR) model are

$$dr = k(\theta - r)dt + \sigma\sqrt{r}dw \tag{15.3}$$

Since the first term on the right-hand side of (15.3) is not a random variable and since the standard deviation of dw equals $\sqrt{dt}$ by definition, the annualized standard deviation of dr (i.e., the basis-point volatility) is proportional to the square root of the rate. Put another way, in the CIR model the parameter σ is constant, but basis-point volatility is not: annualized basis-point volatility equals $\sigma\sqrt{r}$ and increases with the level of the short rate.

Another popular specification is that the basis-point volatility is proportional to rate. In this case the parameter σ is often called *yield volatility*. Two examples of this volatility specification are the Courtadon model,

$$dr = k(\theta - r)dt + \sigma r dw \tag{15.4}$$

and the simplest *lognormal model*, to be called Model 4, a variation of which will be discussed in the next section:

$$dr = ardt + \sigma r dw \tag{15.5}$$

In these two specifications, yield volatility is constant but basis-point volatility equals σr and increases with the level of the rate.

Figure 15.2 graphs the basis-point volatility as a function of rate for the cases of the constant, square root, and proportional specifications. For comparison purposes, σ is set in all three cases such that basis-point volatility equals 100 at a short rate of 8%. Mathematically,

$$\sigma^{bp} = .01 \tag{15.6}$$

$$\sigma^{CIR} \times \sqrt{8\%} = 1\% \Rightarrow \sigma^{CIR} = .0354 \tag{15.7}$$

$$\sigma^{y} \times 8\% = 1\% \Rightarrow \sigma^{y} = 12.5\% \tag{15.8}$$

Note that the units of these volatility measures are somewhat different. Basis-point volatility is in the units of an interest rate (e.g., 100 basis points), while yield volatility is expressed as a percentage of the short rate (e.g., 12.5%).

As shown in Figure 15.2, the CIR and proportional volatility specifications have basis-point volatility increasing with rate but at different speeds. Both models have the basis-point volatility equal to zero at a rate of zero.

The property that basis-point volatility equals zero when the short rate is zero, combined with the condition that the drift is positive when the rate is zero, guarantees that the short rate cannot become negative. In some respects this is an improvement over models with constant basis-point volatility that allow interest rates to become negative. It should be noted again, however, that choosing a model depends on the purpose at hand. Consider a trader who believes the following. One, the assumption of constant volatility is best in the current economic environment. Two, the possibility of negative rates has a small impact on the pricing of the securities under consideration. And three, the computational simplicity of constant volatility models has great value. This trader might very well opt for a model that allows some probability of negative rates.

Figure 15.3 graphs terminal distributions of the short rate after 10 years under the CIR, normal, and lognormal volatility specifications. In order to emphasize the difference in the

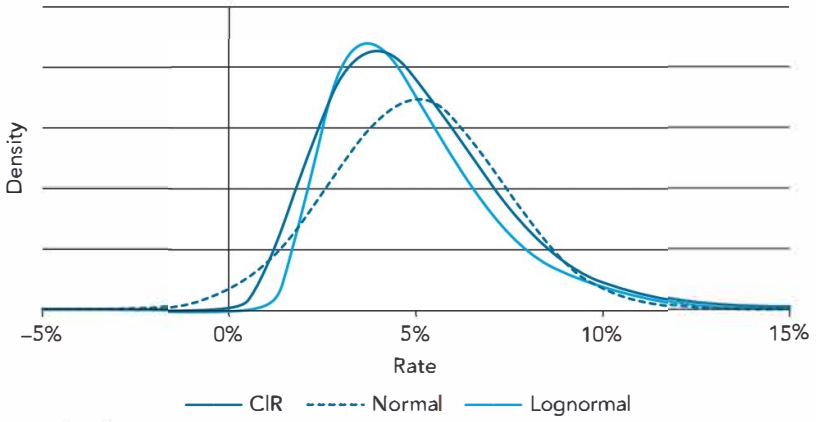


Figure 15.3 **Terminal distributions of the short rate after ten years in CIR, normal, and lognormal models.**

shape of the three distributions, the parameters have been chosen so that all of the distributions have an expected value of 5% and a standard deviation of 2.32%. The figure illustrates the advantage of the CIR and lognormal models with respect to not allowing negative rates. The figure also indicates that out-of-the-money option prices could differ significantly under the three models. Even if, as in this case, the mean and volatility of the three distributions are the same, the probability of outcomes away from the means are different enough to generate significantly different options prices. More generally, the shape of the distribution used in an interest rate model is an important determinant of that model's performance.

15.3 TREE FOR THE ORIGINAL SALOMON BROTHERS MODEL

This section shows how to construct a binomial tree to approximate the dynamics for a lognormal model with a deterministic drift, a model attributed here to researchers at Salomon Brothers in the '80s. The dynamics of the model are as follows:

$$dr = \tilde{a}(t)rdt + \sigma rdw \qquad (15.9)$$

By Ito's Lemma, which is beyond the mathematical scope of this book,

$$d[\ln(r)] = \frac{dr}{r} - \frac{1}{2}\sigma^2 dt \qquad (15.10)$$

Substituting (15.9) into (15.10),

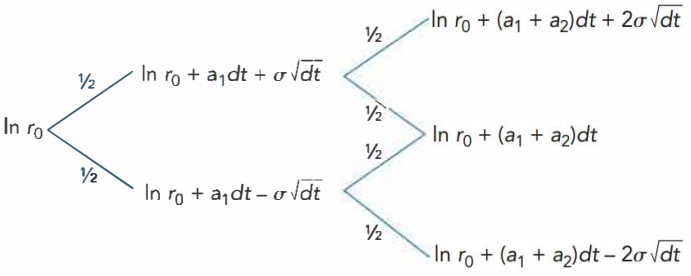
$$d[\ln(r)] = \left[\tilde{a}(t) - \frac{1}{2}\sigma^2\right]dt + \sigma dw \qquad (15.11)$$

Redefining the notation of the time-dependent drift so that $a(t) = \tilde{a}(t) - \frac{1}{2}\sigma^2$, Equation (15.11) becomes

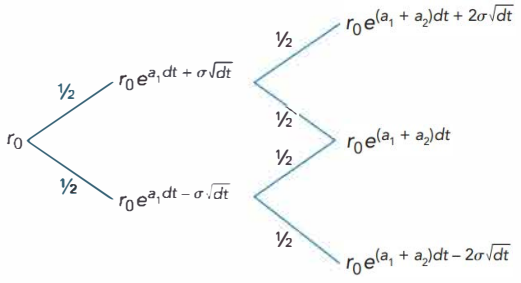
$$d[\ln(r)] = a(t)dt + \sigma dw \qquad (15.12)$$

Equation (15.12) says that the natural logarithm of the short rate is normally distributed. Furthermore, by definition, a random variable has a lognormal distribution if its natural logarithm has a normal distribution. Therefore, (15.12) implies that the short rate has a lognormal distribution.

Equation (15.12) may be described as the Ho-Lee model based on the natural logarithm of the short rate instead of on the short rate itself. Adapting the tree for the Ho-Lee model accordingly, the tree for the first three dates is



To express this tree in rate, as opposed to the natural logarithm of the rate, exponentiate each node:



This tree shows that the perturbations to the short rate in a lognormal model are multiplicative as opposed to the additive perturbations in normal models. This observation, in turn, reveals why the short rate in this model cannot become negative. Since e^x is positive for any value of x, so long as r_0 is positive every node of the lognormal tree results in a positive rate.

The tree also reveals why volatility in a lognormal model is expressed as a percentage of the rate. Recall the mathematical fact that, for small values of x, $e^x \approx 1 + x$. Setting $a_1 = 0$ and $dt = 1$, for example, the top node of date 1 may be approximated as

$$r_0 e^{\sigma} \approx r_0(1 + \sigma) \qquad (15.13)$$

Volatility is clearly a percentage of the rate in equation (15.13). If, for example, $\sigma = 12.5\%$, then the short rate in the up-state is 12.5% above the initial short rate.

As in the Ho-Lee model, the constants that determine the drift (i.e., a_1 and a_2) may be used to match market bond prices.

15.4 THE BLACK-KARASINSKI MODEL: A LOGNORMAL MODEL WITH MEAN REVERSION

The final model to be presented in this chapter is a lognormal model with mean reversion called the Black-Karasinski model. The model allows volatility, mean reversion, and the central tendency of the short rate to depend on time, firmly placing the model in the arbitrage-free class. A user may, of course, use or remove as much time dependence as desired.

The dynamics of the model are written as

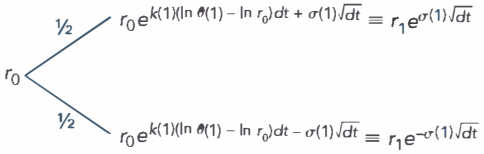
$$dr = k(t)(ln\tilde{\theta}(t) - \ln r)dt + \sigma(t)rdw \tag{15.14}$$

or, equivalently,[2] as

$$d[\ln r] = k(t)(\ln \theta(t) - \ln r)dt + \sigma(t)dw \tag{15.15}$$

In words, Equation (15.15) says that the natural logarithm of the short rate is normally distributed. It reverts to $\ln \theta(t)$ at a speed of $k(t)$ with a volatility of $\sigma(t)$. Viewed another way, the natural logarithm of the short rate follows a time-dependent version of the Vasicek model.

As in the previous section, the corresponding tree may be written in terms of the rate or the natural logarithm of the rate. Choosing the former, the process over the first date is



The variable r_1 is introduced for readability. The natural logarithms of the rates in the up and down-states are

$$\ln r_1 + \sigma(1)\sqrt{dt} \tag{15.16}$$

and

$$\ln r_1 - \sigma(1)\sqrt{dt} \tag{15.17}$$

[2] This derivation is similar to that of moving from Equation (15.9) to Equation (15.12).

respectively. It follows that the step down from the up-state requires a rate of

$$r_1 e^{\sigma(1)\sqrt{dt}} e^{k(2)[\ln\theta(2) - \{\ln r_1 + \sigma(1)\sqrt{dt}\}]dt - \sigma(2)\sqrt{dt}} \tag{15.18}$$

while the step up from the down-state requires a rate of

$$r_1 e^{-\sigma(1)\sqrt{dt}} e^{k(2)[\ln\theta(2) - \{\ln r_1 - \sigma(1)\sqrt{dt}\}]dt + \sigma(2)\sqrt{dt}} \tag{15.19}$$

A little algebra shows that the tree recombines only if

$$k(2) = \frac{\sigma(1) - \sigma(2)}{\sigma(1)dt} \tag{15.20}$$

Imposing the restriction (15.20) would require that the mean reversion speed be completely determined by the time-dependent volatility function. But these elements of a term structure model serve two distinct purposes. As demonstrated in this chapter, mean reversion controls the term structure of volatility while time-dependent volatility controls the future volatility of the short-term rate (and the prices of options that expire at different times). To create a model flexible enough to control mean reversion and time-dependent volatility separately, the model has to construct a recombining tree without imposing (15.20). To do so it allows the length of the time step, dt, to change over time.

Rewriting Equations (15.18) and (15.19) with the time steps labeled dt_1 and dt_2 gives the following values for the up-down and down-up rates:

$$r_1 e^{\sigma(1)\sqrt{dt_1}} e^{k(2)[\ln\theta(2) - \{\ln r_1 + \sigma(1)\sqrt{dt_1}\}]dt_2 - \sigma(2)\sqrt{dt_2}} \tag{15.21}$$

$$r_1 e^{-\sigma(1)\sqrt{dt_1}} e^{k(2)[\ln\theta(2) - \{\ln r_1 - \sigma(1)\sqrt{dt_1}\}]dt_2 + \sigma(2)\sqrt{dt_2}} \tag{15.22}$$

A little algebra now shows that the tree recombines if

$$k(2) = \frac{1}{dt_2}\left[1 - \frac{\sigma(2)\sqrt{dt_2}}{\sigma(1)\sqrt{dt_1}}\right] \tag{15.23}$$

The length of the first time step can be set arbitrarily. The length of the second time step is set to satisfy (15.23), allowing the user freedom in choosing the mean reversion and volatility functions independently.

15.5 APPENDIX

Closed-Form Solutions for Spot Rates

This appendix lists formulas for spot rates, without derivation, in various models mentioned in the text. These can be useful for some applications and also to gain intuition about applying term structure models. The spot rates of term T, $\hat{r}(T)$, are continuously compounded rates.

Model 1

$$\hat{r}(T) = r_0 - \frac{\sigma^2 T^2}{6} \tag{15.24}$$

Model 2

$$\hat{r}(T) = r_0 + \frac{\lambda T}{2} - \frac{\sigma^2 T^2}{6} \tag{15.25}$$

Vasicek

$$\hat{r}(T) = \theta + \frac{1 - e^{-kT}}{kT}(r_0 - \theta) - \frac{\sigma^2}{2k^2}\left(1 + \frac{1 - e^{-2kT}}{2kT} - 2\frac{1 - e^{-kT}}{kT}\right) \tag{15.26}$$

Model 3 with $\lambda(t) = \lambda$

$$\hat{r}(T) = r_0 + \frac{\lambda T}{2} - \sigma^2 \frac{2\alpha^2 T^2 - 2\alpha T + 1 - e^{-2\alpha T}}{8\alpha^3 T} \tag{15.27}$$

Cox-Ingersoll-Ross

Let $P(T)$ be the price of a zero-coupon bond maturing at time T (from which the spot rate can be easily calculated). Then,

$$P(T) = A(T)e^{-B(T)r_0} \tag{15.28}$$

where

$$A(T) = \left[\frac{2he^{(k+h)T/2}}{2h + (k + h)(e^{ht} - 1)}\right]^{2k\theta/\sigma^2} \tag{15.29}$$

$$B(T) = \frac{2(e^{ht} - 1)}{2h + (k + h)(e^{hT} - 1)} \tag{15.30}$$

$$h = \sqrt{k^2 + 2\sigma^2} \tag{15.31}$$

16

The Vasicek and Gauss+ Models

Learning Objectives

After completing this reading, you should be able to:

- Describe the structure of the Gauss+ model and discuss the implications of this structure for the model's ability to replicate empirically observed interest rate dynamics.
- Compare and contrast the dynamics, features, and applications of the Vasicek model and the Gauss+ model.
- Calculate changes in the short-term, medium-term, and long-term interest rate factors under the Gauss+ model.
- Explain how the parameters of the Gauss+ model can be estimated from empirical data.

Excerpt is Chapter 9 of Fixed Income Securities: Tools for Today's Markets, *Fourth Edition, by Bruce Tuckman and Angel Serrat.*

This chapter, the last on term structure models, presents the well-known Vasicek and Gauss+ models. The Vasicek model started the literature on short-term rate models;[1] remains an extremely good starting point for learning about these models; and can still be used in some applied contexts. The Gauss+ model has proven very popular for proprietary trading, for both relative value and macro-style trading. The presentation of this model here is directed toward determined readers who would like to implement a term structure model for their own trading purposes.

16.1 THE VASICEK MODEL

The Vasicek model assumes *mean reversion* to set the expected path of the short-term rate. When below its long-term value, the short-term rate is expected to increase; when above its long-term value, the short-term rate is expected to decrease. Mathematically, the risk-neutral dynamics for the short-term rate, r, are given by,

$$dr = k(\theta - r)dt + \sigma dw \tag{16.1}$$

In words, the instantaneous change in the short-term rate, dr, is determined by a trend or drift plus a random fluctuation or shock. The drift is equal to the parameter of mean reversion, k, times the distance between the long-run value of the short-term rate, θ, and its current value. Because all variables are expressed in annualized terms, the dt factor adjusts for the actual passage of time. Say, for example, that $r = 2\%$; $k = 0.0165$, and $\theta = 11\%$. Then the drift of the short-term rate in Equation (16.1) is 0.0165(11% – 2%) or 0.1485% or 14.85 basis points per year. The drift over a month, therefore, with $dt = 1/12$, would be 14.85/12 or about 1.2 basis points per month. The shock around the drift in Equation (16.1) is σdw, where dw is a normally distributed random variable with mean equal to zero and standard deviation equal to $\sqrt{dt}$. The shock, therefore, is normally distributed with mean zero and standard deviation $\sigma\sqrt{dt}$. For example, if σ is 0.95%, or 95 basis points per year, then the volatility of the shock over a month is $95 \times \sqrt{1/12} = 27.4$ basis points.

Fixed income security prices may incorporate a risk premium that is indistinguishable from a drift in the evolution of the short-term rate. Along these lines, Equation (16.1) can be viewed as containing a drift due to a risk premium. Assume for the purposes of this section that the risk premium is a known constant of λ basis points per year, and that the long-run value of the short-term rate under the true or real-world probabilities is r_∞. In that case, the true process of the short-term rate with the addition of a drift due to the risk premium is,

$$\begin{aligned} dr &= k(r_\infty - r)dt + \lambda dt + \sigma dw \\ &= k\left(\left[r_\infty + \frac{\lambda}{k}\right] - r\right)dt + \sigma dw \end{aligned} \tag{16.2}$$

$$\theta \equiv r_\infty + \frac{\lambda}{k} \tag{16.3}$$

Equation (16.3) neatly emphasizes the inability to distinguish expectations from risk premium by observing security prices: an infinite number of combinations of r_∞ and λ give the same θ and, therefore, the same risk-neutral price process in Equation (16.1).

One reason that the Vasicek model is useful, both for learning about term structure models and for some simple pricing and hedging applications, is that most rates and prices from the model can be expressed through simple formulae. For the most complex securities, numerical methods, like binomial trees, are needed. The text continues by presenting analytic solutions of the model, of which some of the most useful are,

$$E[r_t] = r_0 e^{-kt} + \theta(1 - e^{-kt}) \tag{16.4}$$

$$V[r_t] = \sigma^2 \frac{1 - e^{-2kt}}{2k} \tag{16.5}$$

$$f(t) = \theta + e^{-kt}(r_0 - \theta) - \frac{\sigma^2}{2k^2}(1 + e^{-2kt} - 2e^{-kt}) \tag{16.6}$$

$$\hat{r}(t) = \theta + \frac{1 - e^{-kt}}{kt}(r_0 - \theta) - \frac{\sigma^2}{2k^2}\left(1 + \frac{1 - e^{-2kt}}{2kt} - 2\frac{1 - e^{-kt}}{kt}\right) \tag{16.7}$$

where $E[r_t]$ gives today's expectation of the short-term rate at time t, $V[r_t]$ gives the variance of the short-term rate at time t, $f(t)$ is the continuously compounded forward rate of term t; and $\hat{r}(t)$ is the continuously compounded spot rate of term t.

Figures 16.1 through 16.3 illustrate these formulae with the parameter values given earlier. The expected short-term rate, according to Equation (16.4) and the solid line in Figure 16.1, moves gradually from r_0 today ($t = 0$) to θ in the very distant future ($t = \infty$). The mean reversion parameter governing the speed of that adjustment, $k = 0.0165$, is sometimes quoted instead as a *half-life*. From Equation (16.4), a shock to r_0 decays according to the factor e^{-kt}. And half of such a shock decays away after time h, such that,

$$\begin{aligned} e^{-kh} &= \frac{1}{2} \\ h &= \ln(2)/k \end{aligned} \tag{16.8}$$

[1] Vasicek, O. (1977), "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics* 5.

For the relatively small mean reverting parameter, $k = 0.0165$, h is over 42 years, which means that any shock to r affects rate expectations over a very long period of time. Equivalently, the expected rate takes a very long time to revert from the current rate to θ.

The standard deviation of the short-term rate around its expectations is given by the square root of (16.5) and shown by the dashed lines in Figure 16.1. With a volatility parameter of 95 basis points per year and a very slow mean reversion, the standard deviation around expectations is quite wide. The figure does show, however, that mean reversion narrows this standard deviation. Without mean reversion, the standard deviation of the short-term rate is greater, at $\sigma\sqrt{t}$. Put another way, the pull of the short-term rate to the constant value θ reduces the standard deviation of the short-term rate as of any future date.

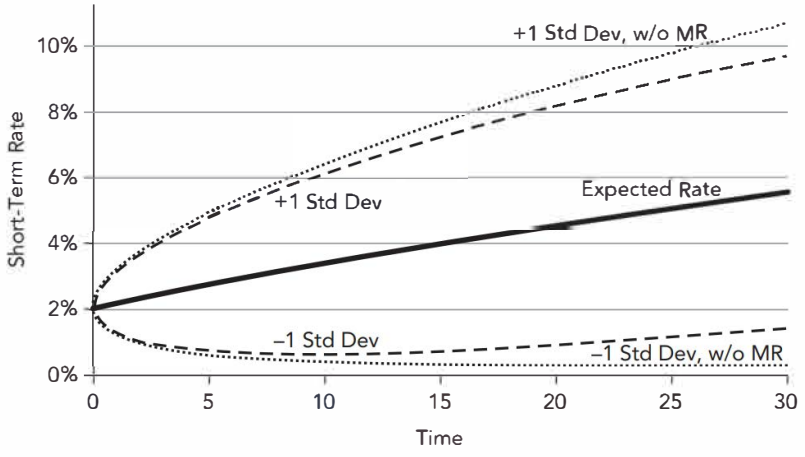


Figure 16.1 **Expectations of the continuously compounded short-term rate in the Vasicek model, with one-standard-deviation bands. Light dotted lines give bands without any mean reversion. Model parameters are $r_0 = 2\%$, $\theta = 11\%$, $k = 0.0165$, and $\sigma = 0.95\%$.**

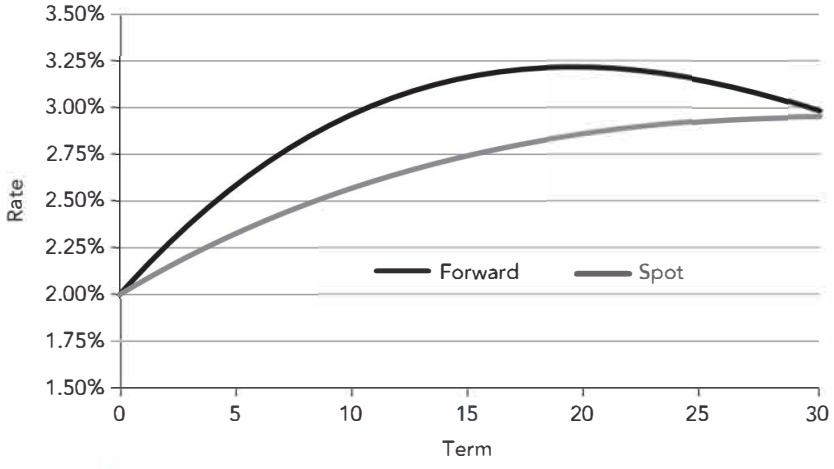


Figure 16.2 **Continuously compounded forward and spot rates in the Vasicek model. Model parameters are $r_0 = 2\%$, $\theta = 11\%$, $k = 0.0165$, and $\sigma = 0.95\%$.**

Equations (16.6) and (16.7), illustrated in Figure 16.2, give the continuously compounded forward and spot rates in the model. The shape of the forward curve is discussed presently. Recall, however, so long as the forward curve is above the spot curve, spot rates are increasing.

Figure 16.3 shows the term structure of forward rate volatility in the model, that is, the instantaneous volatility of forward rates of different terms. Because the only volatility in the model is the volatility of changes in the short-term rate, the volatility of $f(t)$ – from Equation (16.6) – is just σe^{-kt}. The mean reversion feature of the model, therefore, captures the empirical regularity that, for longer terms, the term structure of volatility is downward sloping. Empirical volatilities of short-term rates are much lower than indicated in this figure, however, because the central bank pegs short-term rates. The Gauss+ model, discussed next, has the flexibility to capture both a low short-term rate volatility and an ultimately declining term structure of volatilities. In any case, note from Equation (16.6) that the sensitivity of each forward rate to changes in the short-term rate is e^{-kt}. Hence, these sensitivities across terms have the same shape as in Figure 16.3.

Figure 16.4, the last presented on the Vasicek model, decomposes the forward rate curve into expectations, risk premium, and convexity using the values $\lambda = 0.125\%$, which – given $\theta = 11\%$ and Equation (16.3) – means that $r_\infty = 3.424\%$. In this decomposition, expectations are mildly increasing over the coming years. Forward rates out to 10 year or so increase much more rapidly than expectations, however, due to the risk premium of 12.5 basis points per year. For longer terms, however, the (negative) convexity term grows rapidly, not only moderating the impacts of expectations and convexity but also actually causing forward rates to decline with term.

The Vasicek model has some limited uses for practitioners. It is a relatively simple model, which is a great advantage. Furthermore, a single factor can explain a large fraction of term structure variability, particular across longer maturities. The parameters r_0, k, and θ can be jointly calibrated to approximate both the shape of the term structure and the shape of rate sensitivities to the factor (i.e., Figure 16.3). The parameter σ can be used to approximate an implied option volatility at one point of the term structure. With these considerations in mind, the

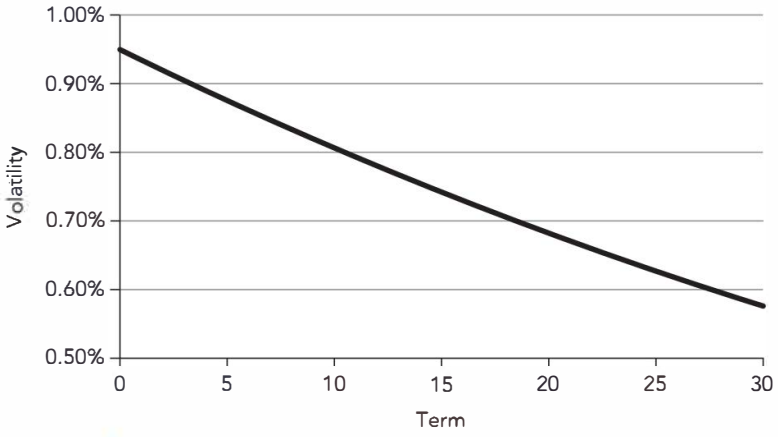


Figure 16.3 **Term structure of forward rate volatilities in the Vasicek model. Parameters are r_0 = 2%, θ = 11%, k = 0.0165, and σ = 0.95%.**

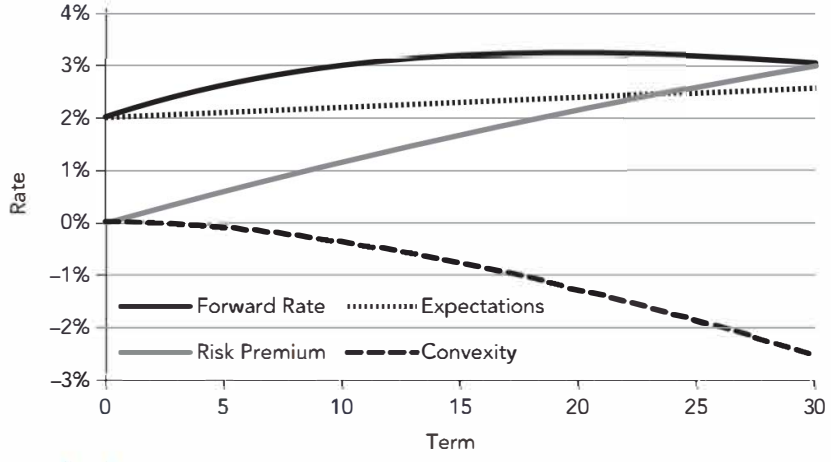


Figure 16.4 **Decomposition of the forward rates in the Vasicek model into expectations, risk premium, and convexity. Model parameters are r_0 = 2%, λ = 0.125%, r_∞ = 3.424% k = 0.0165, and σ = 0.95%.**

model can reasonably be used, for example, to price, compare values, and hedge long-term bonds that are first callable after some intermediate number of years. The model is flexible enough to match the prices of noncallable bonds from 10 to 30 years, and also to match the most relevant volatility, namely, the volatility of 10-year rates. Bond sensitivities to changes in interest rates, defined in the model as changes in r, can be computed by shifting r_0, recomputing prices, and computing DV01s or durations. To the extent that bonds of one maturity are hedged with bonds of another maturity, the effectiveness of the resulting hedges depend on the reliability of the shape in Figure 16.3.

The model might also be used for trading and hedging relatively long-term bonds. The value of r_0 might be set each day so that the model 10-year rate matches the market 10-year rate. Deviations of calculated prices of longer-term bonds from model predictions might then be taken as signals of relative value, with hedge ratios calculated as described in the previous paragraph. The success of relative value trading along these lines depends on the extent to which the model captures the equilibrium or steady state of the shape of the term structure. Also, as before, hedging effectiveness depends on the reliability of the term structure of sensitivities to the factor.

The Vasicek model is not very widely used, however, because it is too simple for most applications. First, the model is not flexible enough to approximate the wide variety of observed market term structures. Or, put another way, calibrating the model to match one point on the term structure relies too heavily on the model to approximate all other points on the term structure. Second, while one factor does explain a lot of term structure variability, a second factor can capture significantly more variability. More than one factor is needed to hedge intermediate- and shorter-term bonds. Third, the Vasicek model cannot capture the empirical regularity, mentioned already, that the term structure of volatilities and, therefore, the term structure of factor sensitivities, typically rises quickly with term and then flattens or declines. This limitation means that the model cannot simultaneously handle options or other volatility-sensitive products that are spread out across the term structure, nor can it reliably hedge bonds most sensitive to one segment of the term structure with bonds most sensitive to another segment. With these caveats, the text turns to a much more flexible term structure model.

16.2 THE GAUSS+ MODEL

The Gauss+ model is well-known among practitioners as a tool for proprietary trading and hedging. The assumptions of the model are intuitively appealing, and they reasonably balance the goals of tractability and of capturing the empirical complexity of term structure dynamics. The goal of the text is to introduce the model, both in theory – through its equations and solution – and in application – through a full estimation of its parameters using recent data. Introducing a detailed estimation procedure here is noteworthy, because methodologies across the industry vary widely.

The dynamics of the cascade *form* of the model are given in Equations (16.9) through (16.12). The factors r, m, and

l denote the short-term rate of interest, a medium-term factor, and a long-term factor, respectively. The parameters μ and ρ are discussed presently. The mean reversion parameters of the factors are, α_r, α_m, and α_l, respectively, and the volatility parameters for the medium- and long-term factors are σ_m and σ_l, respectively. The two random variables in the model are dW^1 and dW^2. The subscript t denotes time-t observations of the factors, of changes in the factors, and of the random variables. Finally, then, the equations are,

$$dr_t = -\alpha_r(r_t - m_t)dt \tag{16.9}$$

$$dm_t = -\alpha_m(m_t - l_t)dt + \sigma_m(\rho dW_t^1 + \sqrt{1 - \rho^2}dW_t^2) \tag{16.10}$$

$$dl_t = -\alpha_l(l_t - \mu)dt + \sigma_l dW_t^1 \tag{16.11}$$

$$E[dW_t^1 dW_t^2] = 0 \tag{16.12}$$

Given the structure of the model, it turns out that the medium- and long-term factors can be thought of as rates. The short-term rate mean reverts to the medium-term factor, which is meant to reflect business cycles and monetary policy factors. The medium-term factor reverts to the long-term factor, which is meant to reflect long-term expectations of inflation and the real interest rate, which ultimately depend on long-term trends in demographics, production technology, and so forth. And the long-term factor reverts to a constant, μ, which, as in the Vasicek model, can be thought of as including both a long-term expectation of the short-term rate and a risk premium. The mean reversion parameters are expected to be consistent with these economic interpretations; that is, the short-term rate reverts quickly to the medium-term factor; the medium-term factor reverts more slowly to the long-term factor; and the long-term factor reverts slowest of all to its target.

While the medium- and long-term factors trend as described in the previous paragraph, they also fluctuate around these trends. With respect to the evolution of the long-term factor in Equation (16.11), the fluctuation over a short time dt is $\sigma_l dW_t^1$. Because dW_t^1 is normally distributed with mean zero and standard deviation $\sqrt{dt}$, the instantaneous fluctuation of the long-term factor around its trend has mean zero and volatility $\sigma_l\sqrt{dt}$. The random terms in Equation (16.10) look complicated, but they simply ensure that the instantaneous fluctuation of the medium-term factor around its trend has a volatility of $\sigma_m\sqrt{dt}$ and a correlation of ρ with the fluctuation of the long-term factor around its trend. To see this, note that dW_t^2 also has mean zero and standard deviation $\sqrt{dt}$, and, from Equation (16.12), zero correlation with dW_t^1. It then follows from Equation (16.10) that the standard deviation of dm_t is,

$$\sqrt{\sigma_m^2(\rho^2 dt + [1 - \rho^2]dt)} = \sigma_m\sqrt{dt} \tag{16.13}$$

that the covariance of dm and dl is,

$$Cov[\sigma_m(\rho dW^1 + \sqrt{1 - \rho^2}dW^2), \sigma_l dW^1] = \rho\sigma_m\sigma_l dt \tag{16.14}$$

and, therefore, that the correlation of dm and dl is,

$$\frac{\rho\sigma_m\sigma_l dt}{\sigma_m\sqrt{dt} \times \sigma_l\sqrt{dt}} = \rho \tag{16.15}$$

The evolution of the short-term rate in the model, Equation (16.9), is meant to reflect how central banks conduct rate policy. The Fed, for example, keeps the short-term policy rate pegged or fixed at a target, but moves that target over time in a manner deemed appropriate for the state of the business cycle and monetary conditions. Mathematically, in Equation (16.9), the short-term rate is fixed over the very short time interval, dt, in the sense that there is no random variable shocking the dynamics of r. The rate, r, is pushed gradually, however, toward the medium-term factor, m, which in turn reverts to the long factor, l.

The medium- and long-term factors move expectations of the short-term rate as of future dates. Because these expectations move smoothly over time, the medium- and long-term factors are assumed to move in a continuous fashion. The short-term rate, by contrast, changes in the real world by discrete amounts, on a set of fixed dates, according to central bank policy decisions. The model approximates the future outcomes of this process, however, by a continuous process starting at today's short-term rate.

The lack of a volatility term in Equation (16.9) is an important feature of the Gauss+ model. As pointed out in the discussion of the Vasicek model, mean-reverting factors generate a downward-sloping term structure of volatility. Largely because of central banks, however, empirical and implied term structures of volatility tend to have a hump; that is, volatility is low for very short-term rates before increasing to a peak at intermediate-term or longer rates. In the Gauss+ model, the lack of a random shock in the dynamics of r keeps short-term rate volatility low. In this way, the Gauss+ model can match empirically observed hump-shaped term structures of volatility.

In passing, the Gauss+ model gets its name from the lack of a volatility term in Equation (16.9). The "Gauss" part of the name indicates that interest rates have a normal or Gaussian distribution. But while most one-, two-, or three-factor normal models have a corresponding number of sources of risk, the Gauss+ model, strictly speaking, has three factors, but only two sources of risk. The model has three state *variable*s in the sense that describing the state of the world in the model requires knowing the three factors, r, m, and l. There are only two sources of risk, however, namely, dW^1 and dW^2. The "+" in the name, therefore, indicates the somewhat unusual presence of a factor or state variable that is not also a source of risk.

As a final comment on the structure of the model, Equations (16.9) through (16.12) are the risk-neutral dynamics of the model. Additional assumptions allow for the identification of an implicit risk premium. An application of the model developed next shows how this may be done, assuming that only the long-term factor earns a risk premium.[2]

A simplified overview of how the parameters of the Gauss+ model might be estimated from bond or swap data is presented here, using daily data from January 2014 to January 2022 on the fed funds target rate and on zero coupon bond prices of various maturities, which are derived from the prices of US Treasury bonds. The time series of zero coupon bond prices are published by the Federal Reserve Bank of New York and are publicly available.[3]

Consistent with the interpretation of the model, the short-term rate, r, is taken each day as equal to the fed funds target rate on that day. While a general collateral repo rate is theoretically more consistent with a term structure of Treasury interest rates, repo rates exhibit occasional idiosyncratic jumps that complicate the estimation without significant offsetting advantages.

Once estimated, the model factors are chosen to "fit" or match, each day, the one-year rates, two and 10 years forward, and the fed funds target rate. Furthermore, as discussed presently, with these two- and 10-year forward rates fair by construction, the model becomes a tool for trading value in other parts of the curve relative to these fitted points. For ease of exposition, by the way, all forward rates mentioned from this point denote one-year rates some number of years forward.

Although the model ultimately fits the two- and 10-year forward rates, note that the factors m and l are in no way those forward rates themselves, in the way that r equals the fed funds rate. The relationships of all forward rates to m and l depend on the estimated model parameters and, in particular, on the mean reversion parameters. More specifically, the larger α_r, the faster changes to the factors m and l make their way into the term structure of rates. The larger α_m, the faster m converges to l, and the less m affects longer-term yields. And the smaller α_l, the slower l converges to μ, and the more similar or parallel is the effect of l on all longer-term yields.

16.3 A PRACTICAL ESTIMATION METHOD

The estimation method presented here proceeds in stages, with each stage estimating a subset of the model's parameters.[4]

Figure 16.5 shows the coefficients of regressing the changes in the zero yields of various terms on changes in the two-year zero yield – the dark gray bars – and on changes in the 10-year zero yield – the light gray bars. For example, the regression of the three-year yield gives a coefficient of 0.91 on the two-year yield and 0.22 on the 10-year yield. By contrast, the regression of the 15-year yield gives a coefficient of −0.20 on the two-year yield and 1.10 on the 10-year yield. In any case, as explained before, all these regression coefficients implicitly describe the mean reversion parameters of the model. This stage of the estimation, therefore, chooses the parameters α_r, α_m, and α_l so that the model captures the empirically observed regression coefficients as closely as possible. This staging is possible because the model regression coefficients depend only on the mean reversion parameters, not on the volatility parameters. In any case, the mottled dark and light gray bars in Figure 16.5 show the success of this stage of the estimation. There are a set of mean reversion parameters, given hereafter, such that implied regression coefficients from the model very closely match the empirically observed regression coefficients. And, while not reported here, these matches are within statistical confidence intervals.

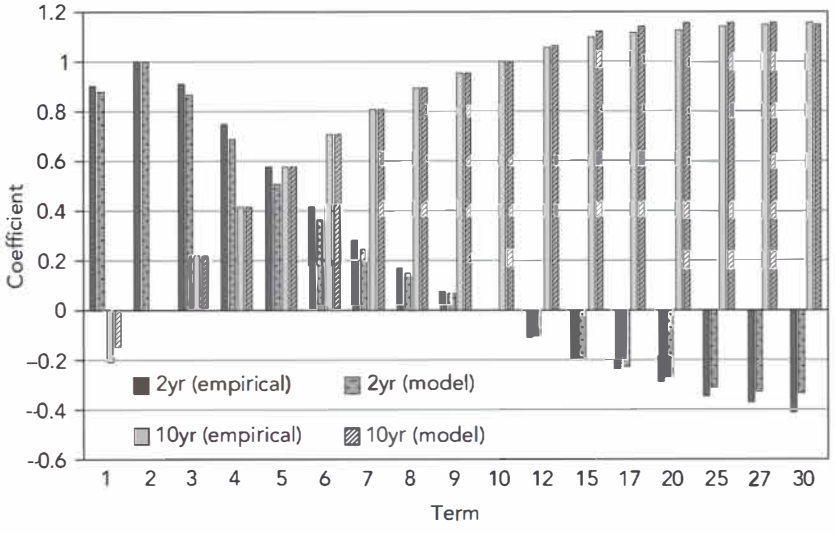


Figure 16.5 **Coefficients of regressing zero coupon bond yields of various terms on two and 10-year zero coupon bond yields, from empirical analysis and as implied by the estimated Gauss+ model.**

[2] See, for example, Cochrane, J., and Piazzesi, M. (2009), "Decomposing the Yield Curve," *AFA 2010 Atlanta Meetings Paper*, January 26.

[3] See Gürkaynak, R., Sack, B., and Wright, J. (2006), "The US Treasury Yield Curve: 1961 to the Present." Data are available at https://www.federalreserve.gov/data/nominal-yield-curve.html.

[4] This approach is significantly easier to implement than the maximum likelihood methods that are standard in the term structure literature.

The next stage of the estimation is to find the volatility and correlation parameters, σ_m, σ_l, and ρ, so that the term structure of volatilities in the model matches the term structure of volatilities in the data as closely as possible. The result of this optimization is shown in Figure 16.6. Once again, the model is flexible enough to do an excellent job of matching empirical properties of the term structure of interest rates.

The last remaining parameter to be estimated is μ, the value to which the short-term rate reverts, over the very long-term. The estimation procedure suggested here finds the μ that minimizes the sum of the squared errors of observed yields relative to model yields across the whole data sample.

Following the estimation procedure described, Table 16.1 reports the resulting Gauss+ parameter values. The mean reversion parameters are in the order expected, with the central bank reaction fastest, the speed of the medium-term factor's reversion to the long-term factor next, and the speed of the long-term factor's reversion to μ slowest. Alternatively, the time for each process to converge halfway to its target, given in the half-life column, is about eight months for *r*, 13 months for *m*, and 42 years for *l*. Because of the mean reverting nature of the factors, the volatility parameters of 109 and 96 basis points for the medium- and long-term factors, respectively, translate into the lower zero yield volatilities shown in Figure 16.6. The parameter μ, as the very long-run target for the short-term rate, might seem high at over 10%, but the long-term factor reverts very slowly to this target, and that target includes a risk premium, which is discussed further next.

The term-structure properties of the estimated model are well described by Figure 16.7, which graphs the change in forward rates for a change in each of the factors as a function of term. For example, the seven-year forward, changes by 0.9 basis points for every basis point change in the long-term factor. Taken as a whole, the figure shows that short-term factor affects only the very short end of the curve. The medium-term factor, which drives the two- to three-year part of the curve, can be thought of as capturing monetary policy in the sense of encapsulating where the market believes the short-term rate will be in two to three years. The long-term factor has its biggest impact in six to eight years and, beyond 10 years, is the sole factor driving forward rate changes and volatilities.

Table 16.1 **Estimated Parameters of the Gauss+ Model from US Treasury Zero Coupon Yields, January 2014 to January 2022. Half-Life Is in Years.**

Parameter	Estimate	Half-Life
α_r	1.0547	0.66
α_m	0.6358	1.09
α_l	0.0165	42.01
σ_m	109.2 bps	
σ_l	96.4 bps	
ρ	0.212	
μ	10.555%	

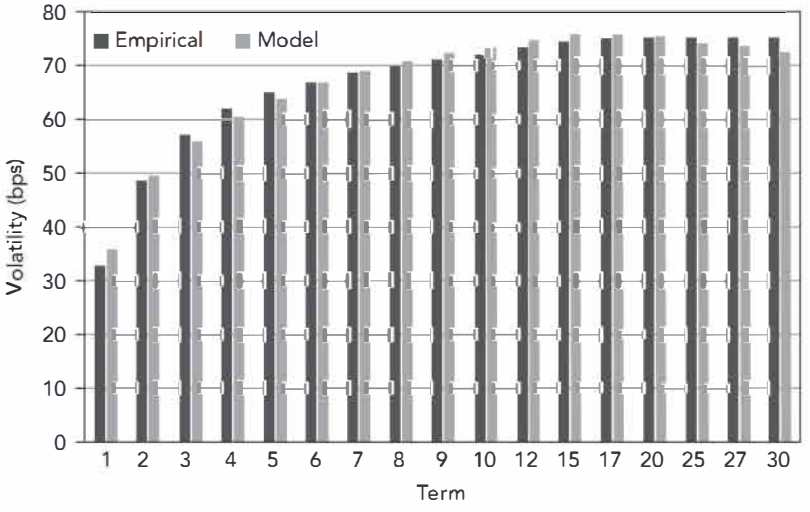


Figure 16.6 **Yield volatility in annual basis points, from empirical analysis and as implied by the estimated Gauss+ model.**

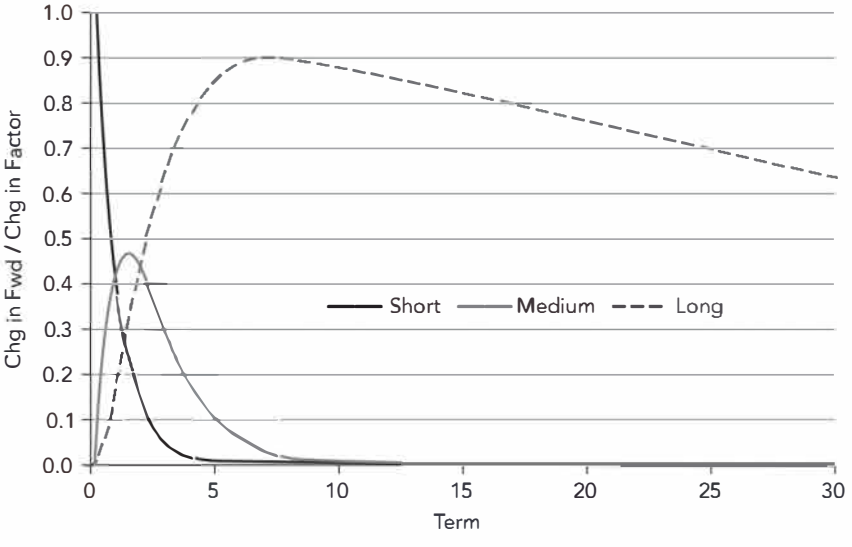


Figure 16.7 **Changes in forward rates relative to changes in the short-rate, the medium-term factor, and the long-term**

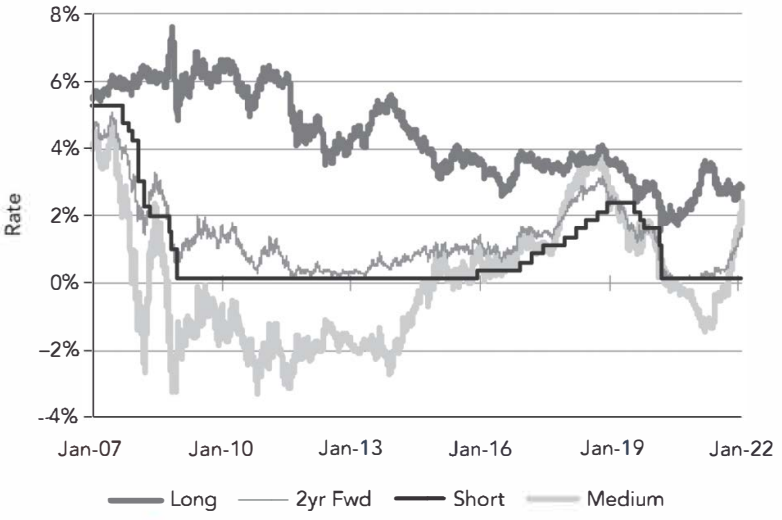


Figure 16.8 **The two-year forward rate and Gauss+ factors extracted from daily market data.**

The time series properties of the model can be described by graphing its factors over time. As mentioned already, the short-term rate is set each day to the fed funds target rate, and the medium- and long-term factors are set so as to match the model and market two- and 10-year forward rates. Figure 16.8 graphs these empirically recovered market factors from January 2007 to January 2022.[5] The two-year forward rate is included in the graph to focus the interpretation of the medium-term factor. When rates are high, this factor closely tracks the two-year forward rate, confirming that the medium-term factor loosely corresponds to where the market expects the short-term rate to be in two years. When rates are low, however, near the zero lower bound, the medium-term factor can fall into deeply negative territory. In this sense, the medium-term factor is a "forward shadow rate" that reflects future expectations of the short-term rate and that can be traded explicitly in the Gauss+ model. This interpretation differs from models in which the "shadow rate" is what the short-term rate would be now if it were not bounded above zero.[6]

[5] The model is estimated using data from January 2014, but the resulting parameters are used to extract model factors back through 2007. Also, instead of the long-term factor itself, the figure graphs the long-term factor shifted forward 10 years. This allows the series to be more easily interpreted as an approximation for expectations of the short-term rate in 10 years.

[6] See, for example, Wu, J., and Xia, F. (2016), "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound," *Journal of Money Credit and Banking* 48, March-April; Bauer, M., and Rudebusch, G. (2016), "Monetary Policy Expectations at the Lower Bound," *Journal of Money Credit and Banking* 48, October; and Kim, D., and Singleton, K. (2012), "Term Structure Models and the Zero Bound: An Investigation of Japanese Yields," *Journal of Econometrics* 170, September.

In any case, the nature of the medium-rate factor as a leading indicator of the short-term rate can be seen by comparing those two series in Figure 16.8. Particularly striking is the steep increase in the medium-term factor starting in early 2014, a couple of years before the Federal Reserve began raising rates.

16.4 RELATIVE VALUE AND MACRO-STYLE TRADING WITH THE GAUSS+ MODEL

In the context of a term structure model, a relative value trade is one that is not exposed to or hedged against changes in the factors. Ideally, a trader would find an individual security that is cheap relative to the model and, from various analyses, is expected to revert soon to being fair to the model. The trader would then buy that security; hedge some or all of its factor exposure by selling fair or rich securities in the same part of the curve; and earn the resulting profits. In practice, however, individual securities are often persistently cheap or rich to the model, so that convergence is not expected in the short run, and securities neighboring in term are often all fair or rich together. Most of the time, therefore, relative value opportunities arise in which a trader receives forward rates in one part of the curve that are high relative to the model but expected to converge promptly to the model; pays forward rates in another part of the curve that are too low, but expected to converge promptly; and structures trades to minimize factor exposures. In this synopsis, the speed at which any detected mispricing is likely to correct itself is an important trading consideration.

As an example of using the Gauss+ model along these lines, consider the following framework. Compute the time series of the nine-year forward rate minus the model-predicted equivalent rate, where each observation can be called a fitting error. Then construct a signal from this times series as the difference between its five-day and 40-day moving averages. If the demeaned fitting error at a given time is positive, so that the nine-year forward is too high relative to the model, and if the signal is negative, so that fitting errors have started to fall, that is, they have started to converge to the model, then receiving at that forward rate might be considered an attractive trade.

The problem with receiving or paying the nine-year forward in isolation, however, is that it has significant exposure to the medium- and long-term factors, which is not consistent with the spirit of relative value trading. One solution is to pool together and size several attractive relative value trades so that their exposure to the factors is minimal. In addition, traders can

diversify across mean reverting trades. It turns out, for example, that nine and five-year fitting errors tend to be positively correlated. This fact makes it attractive to pay in one rate and receive in the other. Figure 16.9 shows, in fact, that the difference between the nine- and five-year signals is strongly mean reverting, which is one of the most important properties of relative value trades.

Unlike relative value trading, which finds value in the absence of factor exposure, macro trading takes direct or indirect views on the factors. Simple examples include positions based on predictions of changes in rates or in the slope of the term structure that differ from what is priced in the market. A more complex example, which has attracted more interest over time, is trying to trade the long-run level of the short-term rate. As explained earlier, long-term forward rates are a combination of expectations, risk premium, and convexity. Within the structure of the Gauss+ model, with the help of a strategic assumption, long-term forward rates can be separated into these three components. A macro trader can then decide that long-term forwards are too low or too high and position accordingly. The details of this decomposition of forward rates are complex. The text continues with an intuitive approach.

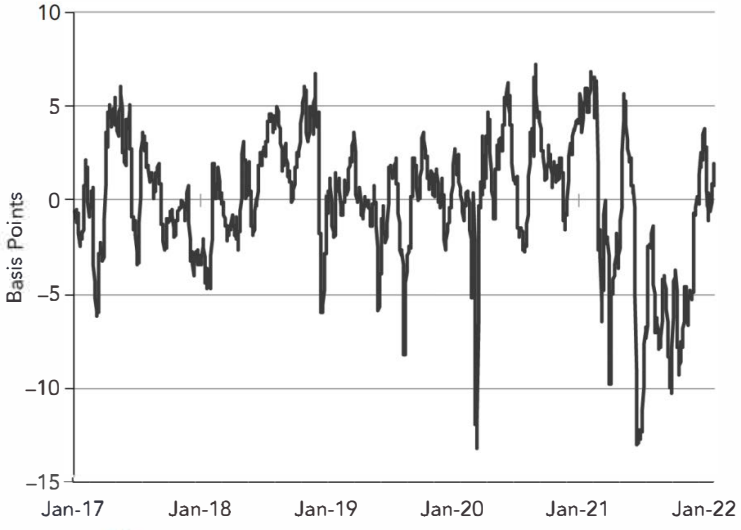


Figure 16.9 **Difference between the nine- and five-year signals. Each signal is based on a difference of moving averages of deviations of market from Gauss+ model rates.**

In the context of any term structure model, it is relatively straightforward to determine the effect of convexity on forward rates. It is much more difficult, however, to separate expectations from risk premium. A number of approaches appear in the academic literature, but these are not without various shortcomings.[7] The method proposed here relies on one key assumption: expectations of future short-term rates do not change past some point in the future. Anyone with a view of the short-term rate in 15 years, for example, has the same view of the short-term rate in 20 or in 30 years. Consequently, any difference in forward rates beyond some term is attributable not to rate expectations, but solely to risk premium and convexity. In this way, expectations and risk premium can be separated and calculated from observable rates.

Using the assumption just described and the parameters of the Gauss+ model estimated previously, Figure 16.10 graphs the risk premium on the 10-year forward rate over time, measured along the left axis. A value of 60 basis points on any day, for example, means that, on that day, 60 basis points of the 10-year forward rate is attributable to risk premium. The lighter line is the level of the 10-year forward rate, measured along the right axis. The risk premium often moves in the opposite direction of rates. As rates decline, the term

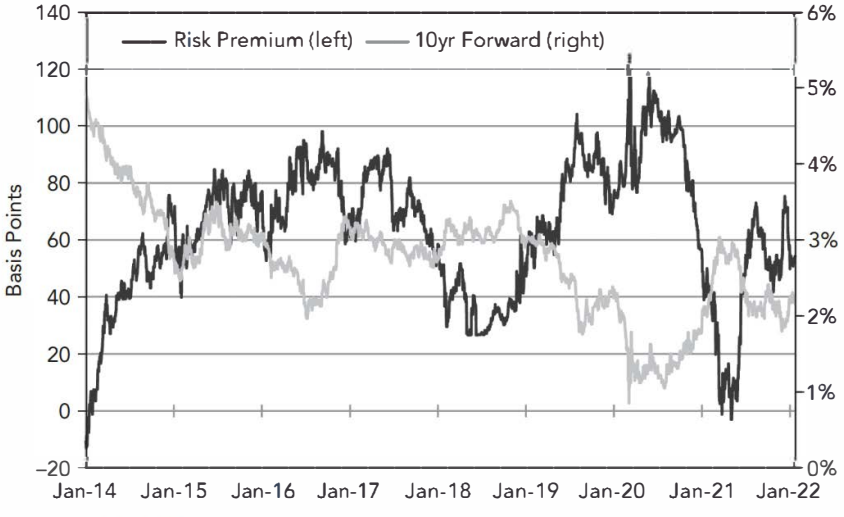


Figure 16.10 **Estimated risk premium on the 10-year forward rate.**

[7] See, for example, Adrian, T., Crump, R., and Moench, E. (2013), "Pricing the Term Structure with Linear Regressions," *Journal of Financial Economics* 110(1), October; Ang, A., and Piazzesi, M. (2003), "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables," *Journal of Monetary Economics* 50(4); Cieslak, A. (2018), "Short-Rate Expectations and Unexpected Returns in Treasury Bonds," *Review of Financial Studies* 31(9); and Kim, D., and Orphanides, A. (2012), "Term Structure Estimation with Survey Data on Interest Rate Forecasts," *Journal of Financial and Quantitative Analysis* 47(1).

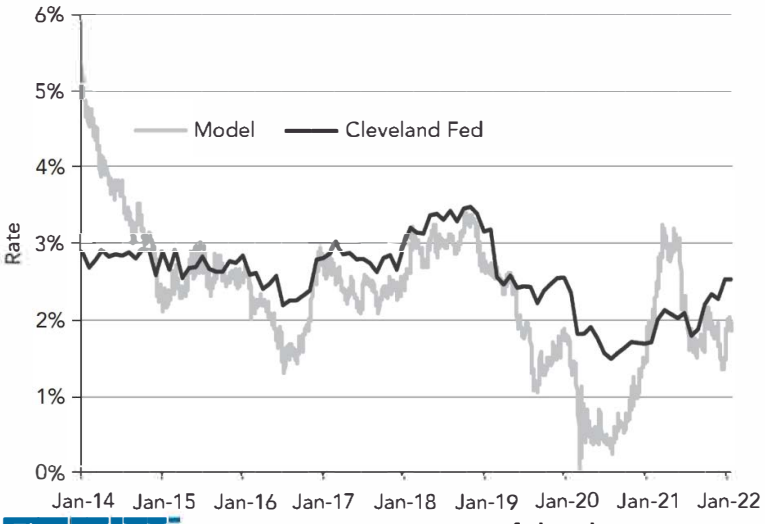


Figure 16.11 **Long-run expectations of the short-term rate, as implied by Gauss+ fitted to market rates and by fundamental analysis at the Federal Reserve Bank of Cleveland.**

structure typically steepens, which the model interprets as an increase in risk premium. This observed behavior is consistent with the low-inflation regime over the last several decades, in which government bond prices increase as other risky assets fall in value. In that environment, as rates fall, and have less room to fall even further, bonds are less able to hedge the declining value of other assets and, as a result, are worth less themselves.

The flip-side of identifying the risk premium, of course, is identifying the long-run expectation of the short-term rate in the estimated Gauss+ model, more specifically, the 10-year forward rate minus the term-appropriate risk premium plus the term-appropriate convexity. The time series of this expectation is shown in Figure 16.11, along with a different estimate, formed from real rate forecasts and inflation estimates at the Federal Reserve Bank of Cleveland. While the model and outside series track each other quite well over time, there are trading opportunities to use the difference between the two series as a measure of value. Put another way, the difference between the Gauss+ market-implied view – the long-run rate priced in the market – and the exogenous, economist-generated, fundamental view – what one thinks the long-run rate should be – can be used as a basis for taking outright long or short positions in bonds.

17

Volatility Smiles and Volatility Surfaces

Learning Objectives

After completing this reading, you should be able to:

- Describe a volatility smile and volatility skew.
- Explain the implications of put-call parity on the implied volatility of call and put options.
- Compare the shape of the volatility smile (or skew) to the shape of the implied distribution of the underlying asset price and to the pricing of options on the underlying asset.
- Describe characteristics of foreign exchange rate distributions and their implications on option prices and implied volatility.
- Describe the volatility smile for equity options and foreign currency options and provide possible explanations for its shape.
- Describe alternative ways of characterizing the volatility smile.
- Describe volatility term structures and volatility surfaces and how they may be used to price options.
- Explain the impact of the volatility smile on the calculation of an option's Greek letter risk measures.
- Explain the impact of a single asset price jump on a volatility smile.

Excerpt is Chapter 20 of Options, Futures, and Other Derivatives, *Eleventh Edition, by John C. Hull.*

How close are the market prices of options to those predicted by the Black–Scholes–Merton model? Do traders really use the Black–Scholes–Merton model when determining a price for an option? Are the probability distributions of asset prices really log-normal? This chapter answers these questions. It explains that traders do use the Black–Scholes–Merton model—but not in exactly the way that Black, Scholes, and Merton originally intended. This is because they allow the volatility used to price an option to depend on its strike price and time to maturity.

A plot of the implied volatility of an option with a certain life as a function of its strike price is known as a *volatility smile*. A three-dimensional plot of the implied volatility as a function of both strike price and time to maturity is known as a *volatility surface*. This chapter describes the volatility smiles and volatility surfaces that traders use in equity and foreign currency markets. It explains the relationship between a volatility smile and the risk-neutral probability distribution being assumed for the future asset price. It also discusses how traders use volatility surfaces as option-pricing tools.

17.1 IMPLIED VOLATILITIES OF CALLS AND PUTS

This section shows that the implied volatility of a European call option is the same as that of a European put option when they have the same strike price and time to maturity. This is a particularly convenient result. It shows that when talking about a volatility smile or volatility surface we do not have to worry about whether the options are calls or puts.

Put–call parity provides a relationship between the prices of European call and put options when they have the same strike price and time to maturity. With a dividend yield on the underlying asset of q, the relationship is

$$p + S_0 e^{-qT} = c + Ke^{-rT} \tag{17.1}$$

As usual, c and p are the European call and put price. They have the same strike price, K, and time to maturity, T. The variable S_0 is the price of the underlying asset today, and r is the risk-free interest rate for maturity T.

A key feature of the put–call parity relationship is that it is based on a relatively simple no-arbitrage argument. It does not require any assumption about the probability distribution of the asset price in the future. It is true both when the asset price distribution is lognormal and when it is not lognormal.

Suppose that, for a particular value of the volatility, p_{BS} and c_{BS} are the values of European put and call options calculated using the Black–Scholes–Merton model. Suppose further that p_{mkt} and c_{mkt} are the market values of these options. Because put–call parity holds for the Black–Scholes–Merton model, we must have

$$p_{BS} + S_0 e^{-qT} = c_{BS} + Ke^{-rT}$$

In the absence of arbitrage opportunities, put–call parity also holds for the market prices, so that

$$p_{mkt} + S_0 e^{-qT} = c_{mkt} + Ke^{-rT}$$

Subtracting these two equations, we get

$$p_{BS} - p_{mkt} = c_{BS} - c_{mkt} \tag{17.2}$$

This shows that the dollar pricing error when the Black–Scholes–Merton model is used to price a European put option should be exactly the same as the dollar pricing error when it is used to price a European call option with the same strike price and time to maturity.

Suppose that the implied volatility of the put option is 22%. This means that $p_{BS} = p_{mkt}$ when a volatility of 22% is used in the Black–Scholes–Merton model. From equation (17.2), it follows that $c_{BS} = c_{mkt}$ when this volatility is used. The implied volatility of the call is, therefore, also 22%. This argument shows that the implied volatility of a European call option is always the same as the implied volatility of a European put option when the two have the same strike price and maturity date. To put this another way, for a given strike price and maturity, the correct volatility to use in conjunction with the Black–Scholes–Merton model to price a European call should always be the same as that used to price a European put. This means that the volatility smile (i.e., the relationship between implied volatility and strike price for a particular maturity) is the same for European calls and European puts. More generally, it means that the volatility surface (i.e., the implied volatility as a function of strike price and time to maturity) is the same for European calls and European puts. These results are also true to a good approximation for American options.

Example 17.1

The value of a foreign currency is \$0.60.The risk-free interest rate is 5% per annum in the United States and 10% per annum in the foreign country. The market price of a European call option on the foreign currency with a maturity of 1 year and a strike price of \$0.59 is 0.0236. DerivaGem shows that the implied volatility of the call is 14.5%. For there to be no arbitrage, the put–call parity relationship in equation (17.1) must apply with q equal to the foreign risk-free rate. The price p of a European put

option with a strike price of $0.59 and maturity of 1 year therefore satisfies

$$p + 0.60e^{-0.10\times 1} = 0.0236 + 0.59e^{-0.05\times 1}$$

so that $p = 0.0419$. DerivaGem shows that, when the put has this price, its implied volatility is also 14.5%. This is what we expect from the analysis just given.

17.2 VOLATILITY SMILE FOR FOREIGN CURRENCY OPTIONS

The volatility smile used by traders to price foreign currency options tends to have the general form shown in Figure 17.1. The implied volatility is relatively low for at-the-money options. It becomes progressively higher as an option moves either into the money or out of the money.

In the appendix at the end of this chapter, we show how to determine the risk-neutral probability distribution for an asset price at a future time from the volatility smile given by options maturing at that time. We refer to this as the *implied distribution*. The volatility smile in Figure 17.1 corresponds to the implied distribution shown by the solid line in Figure 17.2. A lognormal distribution with the same mean and standard deviation as the implied distribution is shown by the dashed line in Figure 17.2. It can be seen that the implied distribution has heavier tails than the lognormal distribution.[1]

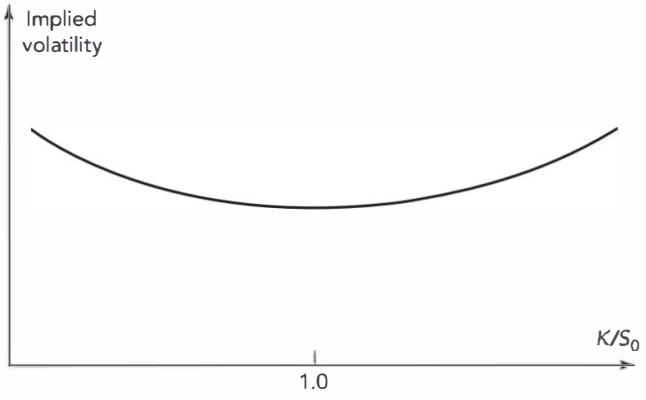


Figure 17.1 Volatility smile for foreign currency options (K = strike price, S_0 = current exchange rate).

[1] This is known as *kurtosis*. Note that, in addition to having a heavier tail, the implied distribution is more "peaked." Both small and large movements in the exchange rate are more likely than with the lognormal distribution. Intermediate movements are less likely.

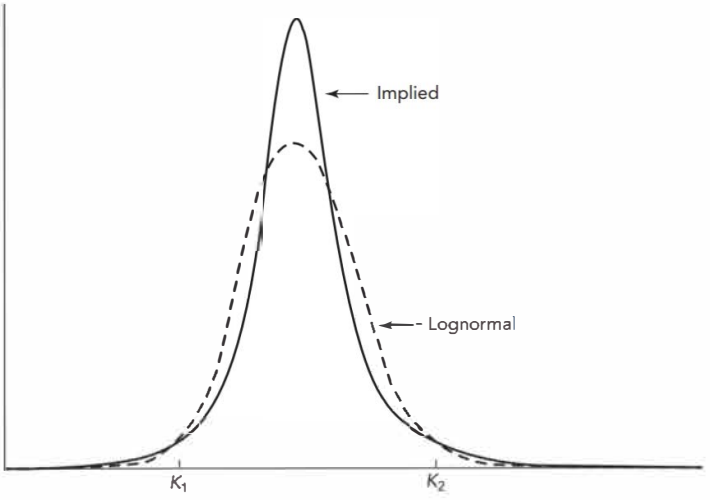


Figure 17.2 Implied and lognormal distribution for foreign currency options.

To see that Figures 17.1 and 17.2 are consistent with each other, consider first a deep-out-of-the-money call option with a high strike price of K_2 (K_2/S_0 well above 1.0). This option pays off only if the exchange rate proves to be above K_2. Figure 17.2 shows that the probability of this is higher for the implied probability distribution than for the lognormal distribution. We therefore expect the implied distribution to give a relatively high price for the option. A relatively high price leads to a relatively high implied volatility—and this is exactly what we observe in Figure 17.1 for the option. The two figures are therefore consistent with each other for high strike prices. Consider next a deep-out-of-the-money put option with a low strike price of K_1 (K_1/S_0 well below 1.0). This option pays off only if the exchange rate proves to be below K_1. Figure 17.2 shows that the probability of this is also higher for the implied probability distribution than for the lognormal distribution. We therefore expect the implied distribution to give a relatively high price, and a relatively high implied volatility, for this option as well. Again, this is exactly what we observe in Figure 17.1.

Empirical Results

We have just shown that the volatility smile used by traders for foreign currency options implies that they consider that the lognormal distribution understates the probability of extreme movements in exchange rates. To test whether they are right, Table 17.1 examines the daily movements in 10 different exchange rates over a 10-year period between 2005 and 2015. The exchange rates are those between the U.S. dollar and the following currencies: Australian dollar, British pound, Canadian dollar, Danish krone, euro, Japanese yen, Mexican peso,

Table 17.1 Percentage of Days When Daily Exchange Rate Moves are Greater than 1, 2, . . . , 6 Standard Deviations (SD = Standard Deviation of Daily Change)

	Real world	*Lognormal model*
>1 SD	23.32	31.73
>2 SD	4.67	4.55
>3 SD	1.30	0.27
>4 SD	0.49	0.01
>5 SD	0.24	0.00
>6 SD	0.13	0.00

New Zealand dollar, Swedish krona, and Swiss franc. The first step in the production of the table is to calculate the standard deviation of daily percentage change in each exchange rate. The next stage is to note how often the actual percentage change exceeded 1 standard deviation, 2 standard deviations, and so on. The final stage is to calculate how often this would have happened if the percentage changes had been normally distributed. (The lognormal model implies that percentage changes are almost exactly normally distributed over a one-day time period.)

Daily changes exceed 3 standard deviations on 1.30% of days. The lognormal model predicts that this should happen on only 0.27% of days. Daily changes exceed 4, 5, and 6 standard deviations on 0.49%, 0.24%, and 0.13% of days, respectively. The lognormal model predicts that we should hardly ever observe this happening. The table therefore provides evidence to support the existence of heavy tails (Figure 17.2) and the volatility smile used by traders (Figure 17.1). Business Snapshot 17.1 shows how you could have made money if you had done the analysis in Table 17.1 ahead of the rest of the market.

BUSINESS SNAPSHOT 17.1 MAKING MONEY FROM FOREIGN CURRENCY OPTIONS

Black, Scholes, and Merton in their option pricing model assume that the underlying asset price has a lognormal distribution at future times. This is equivalent to the assumption that asset price changes over a short period of time, such as one day, are normally distributed. Suppose that most market participants are comfortable with the Black–Scholes–Merton assumptions for exchange rates. You have just done the analysis in Table 17.1 and know that the lognormal assumption is not a good one for exchange rates. What should you do?

The answer is that you should buy deep-out-of-the-money call and put options on a variety of different currencies and wait. These options will be relatively inexpensive and more of them will close in the money than the lognormal model predicts. The present value of your payoffs will on average be much greater than the cost of the options.

In the mid-1980s, a few traders knew about the heavy tails of foreign exchange probability distributions. Everyone else thought that the lognormal assumption of Black–Scholes–Merton was reasonable. The few traders who were well informed followed the strategy we have described—and made lots of money. By the late 1980s everyone realized that foreign currency options should be priced with a volatility smile and the trading opportunity disappeared.

Reasons for the Smile in Foreign Currency Options

Why are exchange rates not lognormally distributed? Two of the conditions for an asset price to have a lognormal distribution are:

1. The volatility of the asset is constant.
2. The price of the asset changes smoothly with no jumps.

In practice, neither of these conditions is satisfied for an exchange rate. The volatility of an exchange rate is far from constant, and exchange rates frequently exhibit jumps, sometimes in response to the actions of central banks. It turns out that both a nonconstant volatility and jumps will have the effect of making extreme outcomes more likely.

The impact of jumps and nonconstant volatility depends on the option maturity. As the maturity of the option is increased, the percentage impact of a nonconstant volatility on prices becomes more pronounced, but its percentage impact on implied volatility usually becomes less pronounced. The percentage impact of jumps on both prices and the implied volatility becomes less pronounced as the maturity of the option is increased.[2] The result of all this is that the volatility smile becomes less pronounced as option maturity increases.

17.3 VOLATILITY SMILE FOR EQUITY OPTIONS

Prior to the crash of 1987, there was no marked volatility smile for equity options. Since 1987, the volatility smile used by traders to price equity options (both on individual stocks and

[2] When we look at sufficiently long-dated options, jumps tend to get "averaged out," so that the exchange rate distribution when there are jumps is almost indistinguishable from the one obtained when the exchange rate changes smoothly.

on stock indices) has tended to look like that in Figure 17.3. This is sometimes referred to as a *volatility skew*. The volatility decreases as the strike price increases. The volatility used to price a low-strike-price option (i.e., a deep-out-of-the-money put or a deep-in-the-money call) is significantly higher than that used to price a high-strike-price option (i.e., a deep-in-the-money put or a deep-out-of-the-money call).

The volatility smile for equity options corresponds to the implied probability distribution given by the solid line in Figure 17.4. A lognormal distribution with the same mean and standard deviation as the implied distribution is shown by the dotted line. It can be seen that the implied distribution has a heavier left tail and a less heavy right tail than the lognormal distribution.

To see that Figures 17.3 and 17.4 are consistent with each other, we proceed as for Figures 17.1 and 17.2 and consider options that are deep out of the money. From Figure 17.4, a deep-out-of-the-money call with a strike price of K_2 (K_2/S_0 well above 1.0) has a lower price when the implied distribution is used than when the lognormal distribution is used. This is because the option pays off only if the stock price proves to be above K_2, and the probability of this is lower for the implied probability distribution than for the lognormal distribution. Therefore, we expect the implied distribution to give a relatively low price for the option. A relatively low price leads to a relatively low implied volatility—and this is exactly what we observe in Figure 17.3 for the option. Consider next a deep-out-of-the-money put option with a strike price of K_1.

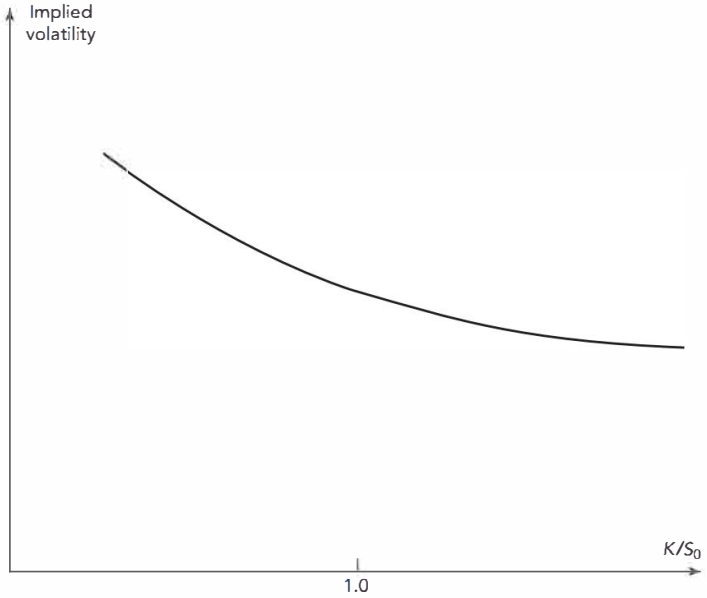


Figure 17.3 Volatility smile for equities (K = strike price, S_0 = current equity price).

This option pays off only if the stock price proves to be below K_1 (K_1/S_0 well below 1.0). Figure 17.4 shows that the probability of this is higher for the implied probability distribution than for the lognormal distribution. We therefore expect the implied distribution to give a relatively high price, and a relatively high implied volatility, for this option. Again, this is exactly what we observe in Figure 17.3.

The Reason for the Smile in Equity Options

There is a negative correlation between equity prices and volatility.[3] As prices move down (up), volatilities tend to move up (down). There are several possible reasons for this. One concerns leverage. As equity prices move down (up), leverage increases (decreases) and as a result volatility increases (decreases). Another is referred to as the *volatility feedback effect*. As volatility increases (decreases) because of external factors, investors require a higher (lower) return and as a result the stock price declines (increases). A further explanation is crashophobia (see Business Snapshot 17.2).

Whatever the reason for the negative correlation, it means that stock price declines are accompanied by increases in volatility, making even greater declines possible. Stock price increases are accompanied by decreases in volatility, making further stock price increases less likely. This explains the heavy left tail and thin right tail of the implied distribution in Figure 17.4.

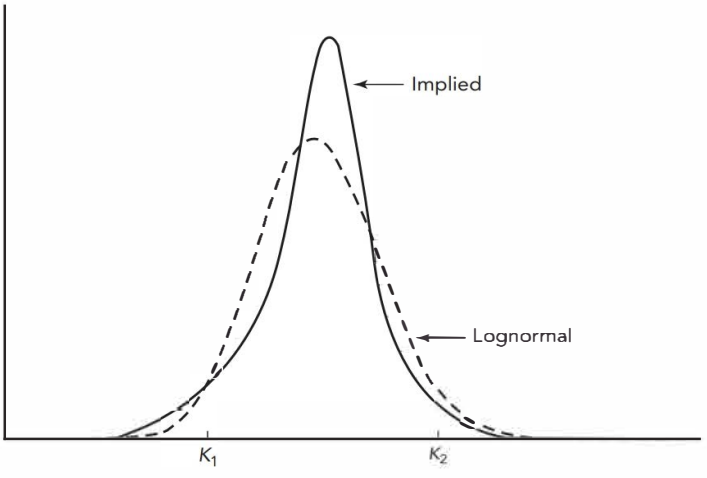


Figure 17.4 Implied distribution and lognormal distribution for equity options.

[3] For a machine learning investigation of this, see J. Cao, J. Chen, and J. Hull, "A Neural Network Approach to Understanding Implied Volatility Movements," *Quantitative Finance*, 20, 9 (2020): 1405–13.

BUSINESS SNAPSHOT 17.2 CRASHOPHOBIA

It is interesting that the pattern in Figure 17.3 for equities has existed only since the stock market crash of October 1987. Prior to October 1987, implied volatilities were much less dependent on strike price. This has led Mark Rubinstein to suggest that one reason for the equity volatility smile may be "crashophobia." Traders are concerned about the possibility of another crash similar to October 1987, and they price options accordingly.

There is some empirical support for this explanation. Declines in the S&P 500 tend to be accompanied by a steepening of the volatility skew. When the S&P increases, the skew tends to become less steep.

17.4 ALTERNATIVE WAYS OF CHARACTERIZING THE VOLATILITY SMILE

There are a number of ways of characterizing the volatility smile. Sometimes it is shown as the relationship between implied volatility and strike price K. However, this relationship depends on the price of the asset. As the price of the asset increases (decreases), the central at-the-money strike price increases (decreases) so that the curve relating the implied volatility to the strike price moves to right (left).[4] For this reason the implied volatility is often plotted as a function of the strike price divided by the current asset price, K/S_0. This is what we have done Figures 17.1 and 17.3.

A refinement of this is to calculate the volatility smile as the relationship between the implied volatility and K/F_0, where F_0 is the forward price of the asset for a contract maturing at the same time as the options that are considered. Traders also often define an "at-the-money" option as an option where $K = F_0$, not as an option where $K = S_0$. The argument for this is that F_0, not S_0, is the expected stock price on the option's maturity date in a risk-neutral world.

Yet another approach to defining the volatility smile is as the relationship between the implied volatility and the delta of the option. This approach sometimes makes it possible to apply volatility smiles to options other than European and American calls and puts. When the approach is used, an at-the-money option is then defined as a call option with a delta of 0.5 or a put option with a delta of −0.5. These are referred to as "50-delta options."

[4] Research by Derman suggests that this adjustment is sometimes "sticky" in the case of exchange-traded options. See E. Derman, "Regimes of Volatility," *Risk*, April 1999: 55–59.

17.5 THE VOLATILITY TERM STRUCTURE AND VOLATILITY SURFACES

Traders allow the implied volatility to depend on time to maturity as well as strike price. Implied volatility tends to be an increasing function of maturity when short-dated volatilities are historically low. This is because there is then an expectation that volatilities will increase. Similarly, volatility tends to be a decreasing function of maturity when short-dated volatilities are historically high. This is because there is then an expectation that volatilities will decrease.

Volatility surfaces combine volatility smiles with the volatility term structure to tabulate the volatilities appropriate for pricing an option with any strike price and any maturity. An example of a volatility surface that might be used for foreign currency options is given in Table 17.2.

One dimension of Table 17.2 is K/S_0; the other is time to maturity. The main body of the table shows implied volatilities calculated from the Black–Scholes–Merton model. At any given time, some of the entries in the table are likely to correspond to options for which reliable market data are available. The implied volatilities for these options are calculated directly from their market prices and entered into the table. The rest of the table is typically determined using interpolation. The table shows that the volatility smile becomes less pronounced as the option maturity increases. As mentioned earlier, this is what is observed for currency options. (It is also what is observed for options on most other assets.)

Table 17.2 Volatility Surface.

	K/S_0				
	0.90	**0.95**	**1.00**	**1.05**	**1.10**
1 month	14.2	13.0	12.0	13.1	14.5
3 month	14.0	13.0	12.0	13.1	14.2
6 month	14.1	13.3	12.5	13.4	14.3
1 year	14.7	14.0	13.5	14.0	14.8
2 year	15.0	14.4	14.0	14.5	15.1
5 year	14.8	14.6	14.4	14.7	15.0

When a new option has to be valued, financial engineers look up the appropriate volatility in the table. For example, when valuing a 9-month option with a K/S_0 ratio of 1.05, a financial engineer would interpolate between 13.4 and 14.0 in Table 17.2 to obtain a volatility of 13.7%. This is the volatility that would be used in the Black–Scholes–Merton formula or a binomial tree. When valuing a 1.5-year option with a K/S_0 ratio of 0.925, a two-dimensional (bilinear) interpolation would be used to give an implied volatility of 14.525%.

The shape of the volatility smile depends on the option maturity. As illustrated in Table 17.2, the smile tends to become less pronounced as the option maturity increases. Define T as the time to maturity and F_0 as the forward price of the asset for a contract maturing at the same time as the option. Some financial engineers choose to define the volatility smile as the relationship between implied volatility and

$$\frac{1}{\sqrt{T}}\ln\left(\frac{K}{F_0}\right)$$

rather than as the relationship between the implied volatility and K. The smile is then usually much less dependent on the time to maturity.

17.6 MINIMUM VARIANCE DELTA

The formulas for delta and other Greek letters assume that the implied volatility remains the same when the asset price changes. This is not what is usually expected to happen. Consider, for example, a stock or stock index option.

As explained in Section 17.3, there is a negative correlation between equity prices and volatility. The delta that takes this relationship between implied volatilities and equity prices into account is referred to as the *minimum variance delta*. It is:

$$\Delta_{MV} = \frac{\partial f_{BSM}}{\partial S} + \frac{\partial f_{BSM}}{\partial \sigma_{imp}}\frac{\partial E(\sigma_{imp})}{\partial S}$$

where f_{BSM} is the Black–Scholes–Merton price of the option, σ_{imp} is the option's implied volatility, $E(\sigma_{imp})$ denotes the expectation of σ_{imp} as a function of the equity price, S. This gives

$$\Delta_{MV} = \Delta_{BSM} + \nu_{BSM}\frac{\partial E(\sigma_{imp})}{\partial S}$$

where Δ_{BSM} and ν_{BSM} are the delta and vega calculated from the Black–Scholes–Merton (constant volatility) model. Because ν_{BSM} is positive and, as we have just explained $\partial E(\sigma_{imp})/\partial S$ is negative, the minimum variance delta is less than the Black–Scholes-Merton delta.[5]

17.7 THE ROLE OF THE MODEL

How important is the option-pricing model if traders are prepared to use a different volatility for every option? It can be argued that the Black–Scholes–Merton model is no more than a sophisticated interpolation tool used by traders for ensuring that an option is priced consistently with the market prices of other actively traded options. If traders stopped using Black–Scholes–Merton and switched to another plausible model, then the volatility surface and the shape of the smile would change, but arguably the dollar prices quoted in the market would not change appreciably. Greek letters and therefore hedging strategies do depend on the model used. An unrealistic model is liable to lead to poor hedging.

Models have most effect on the pricing of derivatives when similar derivatives do not trade actively in the market. For example, the pricing of many of the nonstandard exotic derivatives is model-dependent.

17.8 WHEN A SINGLE LARGE JUMP IS ANTICIPATED

Let us now consider an example of how an unusual volatility smile might arise in equity markets. Suppose that a stock price is currently \$50 and an important news announcement due in a few days is expected either to increase the stock price by \$8 or to reduce it by \$8. (This announcement could concern the outcome of a takeover attempt or the verdict in an important lawsuit.) The probability distribution of the stock price in, say, 1 month might then consist of a mixture of two lognormal distributions, the first corresponding to favorable news, the second to unfavorable news. The situation is illustrated in Figure 17.5. The solid line shows the mixture-of-lognormals distribution for the stock price in 1 month; the dashed line shows a lognormal distribution with the same mean and standard deviation as this distribution.

The true probability distribution is bimodal (certainly not lognormal). One easy way to investigate the general effect of a bimodal stock price distribution is to consider the extreme case

[5] For a further discussion of this, see, for example, J. C. Hull and A. White, "Optimal Delta Hedging of Options," *Journal of Banking and Finance*, 82 (September 2017), 180–190.

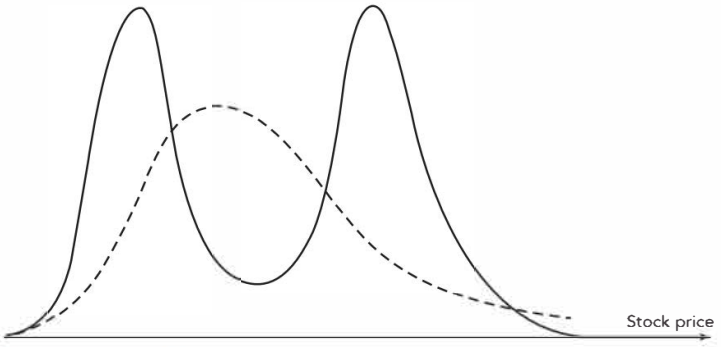


Figure 17.5 Effect of a single large jump. The solid line is the true distribution; the dashed line is the lognormal distribution.

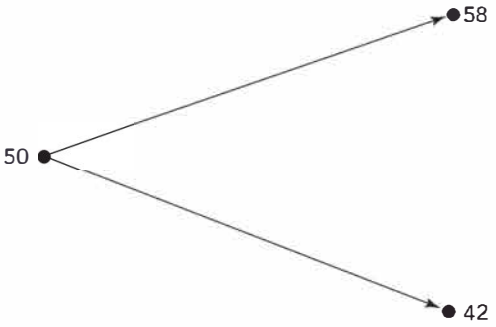


Figure 17.6 Change in stock price in 1 month.

where there are only two possible future stock prices. This is what we will now do.

Suppose that the stock price is currently \$50 and it is known that in 1 month it will be either \$42 or \$58. Suppose further that the risk-free rate is 12% per annum. The situation is illustrated in Figure 17.6. In this case, $u = 1.16$, $d = 0.84$, $a = 1.0101$, and $p = 0.5314$. The results from valuing a range of different options are shown in Table 17.3. The first column shows alternative strike prices; the second shows prices of 1-month European call options; the third shows the prices of one-month European put option prices; and the fourth shows implied volatilities. (As shown in Section 17.1, the implied volatility of a European put option is the same as that of a European call option when they have the same strike price and maturity.) Figure 17.7 displays

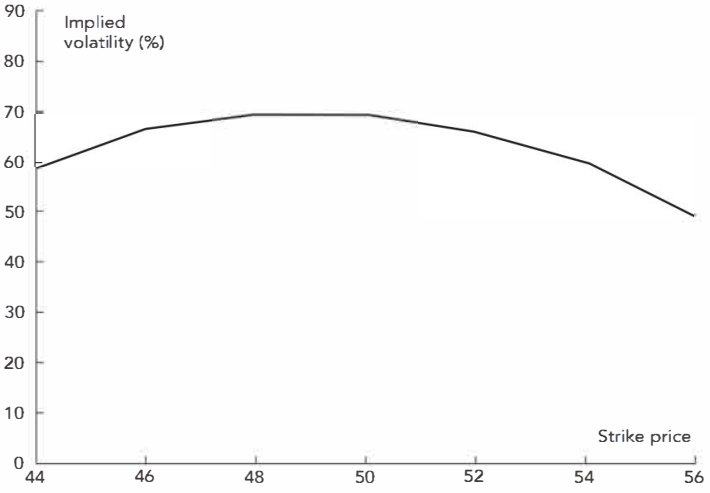


Figure 17.7 Volatility smile for situation in Table 17.3.

Table 17.3 Implied Volatilities in Situation Where It is Known That the Stock Price Will Move from \$50 to Either \$42 or \$58.

Strike price ($)	*Call price ($)*	*Put price ($)*	*Implied volatility (%)*
42	8.42	0.00	0.0
44	7.37	0.93	58.8
46	6.31	1.86	66.6
48	5.26	2.78	69.5
50	4.21	3.71	69.2
52	3.16	4.64	66.1
54	2.10	5.57	60.0
56	1.05	6.50	49.0
58	0.00	7.42	0.0

the volatility smile from Table 17.3. It is actually a "frown" (the opposite of that observed for currencies) with volatilities declining as we move out of or into the money. The volatility implied from an option with a strike price of 50 will overprice an option with a strike price of 44 or 56.

SUMMARY

The Black–Scholes–Merton model and its extensions assume that the probability distribution of the underlying asset at any given future time is lognormal. This assumption is not the one made by traders. They assume the probability distribution of an equity price has a heavier left tail and a less heavy right tail than the lognormal distribution. They also assume that the probability distribution of an exchange rate has a heavier right tail and a heavier left tail than the lognormal distribution.

Traders use volatility smiles to allow for nonlognormality. The volatility smile defines the relationship between the implied volatility of an option and its strike price. For equity options, the volatility smile tends to be downward sloping. This means that out-of-the-money puts and in-the-money calls tend to have high implied volatilities whereas out-of-the-money calls and in-the-money puts tend to have low implied volatilities. For foreign currency options, the volatility smile is U-shaped. Both out-of-the-money and in-the-money options have higher implied volatilities than at-the-money options.

Often traders also use a volatility term structure. The implied volatility of an option then depends on its life. When volatility smiles and volatility term structures are combined, they produce a volatility surface. This defines implied volatility as a function of both the strike price and the time to maturity.

FURTHER READING

Bakshi, G., C. Cao, and Z. Chen. "Empirical Performance of Alternative Option Pricing Models," *Journal of Finance*, 52, No. 5 (December 1997): 2004–49.

Bates, D. S. "Post'87 Crash Fears in the S&P Futures Market," *Journal of Econometrics*, 94 (January/February 2000): 181–238.

Daglish, T., J. C. Hull, and W. Suo. "Volatility Surfaces: Theory, Rules of Thumb, and Empirical Evidence," *Quantitative Finance*, 7, 5 (2007), 507–24.

Derman, E. "Regimes of Volatility," *Risk*, April 1999: 55–59.

Ederington, L. H., and W. Guan. "Why Are Those Options Smiling," *Journal of Derivatives*, 10, 2 (2002): 9–34.

Hull, J. C., and A. White. "Optimal Delta Hedging of Options," *Journal of Banking and Finance*, 82 (September, 2017): 180–190.

Jackwerth, J. C., and M. Rubinstein. "Recovering Probability Distributions from Option Prices," *Journal of Finance*, 51 (December 1996): 1611–31.

Melick, W. R., and C. P. Thomas. "Recovering an Asset's Implied Probability Density Function from Option Prices: An Application to Crude Oil during the Gulf Crisis," *Journal of Financial and Quantitative Analysis*, 32, 1 (March 1997): 91–115.

Reiswich, D., and U. Wystup. "FX Volatility Smile Construction," Working Paper, Frankfurt School of Finance and Management, April 2010.

Rubinstein, M. "Nonparametric Tests of Alternative Option Pricing Models Using All Reported Trades and Quotes on the 30 Most Active CBOE Option Classes from August 23, 1976, through August 31, 1978," *Journal of Finance*, 40 (June 1985): 455–80.

SHORT CONCEPT QUESTIONS

17.1 What is a volatility smile?

17.2 What is a volatility surface?

17.3 What is the difference between the volatility smiles that are typically observed for equities and currencies?

17.4 What is a minimum variance delta?

17.5 How is a volatility surface used in option pricing?

PRACTICE QUESTIONS

17.6 What volatility smile is likely to be observed when:

(a) Both tails of the stock price distribution are less heavy than those of the lognormal distribution?

(b) The right tail is heavier, and the left tail is less heavy, than that of a lognormal distribution?

17.7 What volatility smile is likely to be caused by jumps in the underlying asset price? Is the pattern likely to be more pronounced for a 2-year option than for a 3-month option?

17.8 A European call and put option have the same strike price and time to maturity. The call has an implied volatility of 30% and the put has an implied volatility of 25%. What trades would you do?

17.9 Explain carefully why a distribution with a heavier left tail and less heavy right tail than the lognormal distribution gives rise to a downward sloping volatility smile.

17.10 The market price of a European call is $3.00 and its price given by Black–Scholes–Merton model with a volatility of 30% is $3.50. The price given by this Black–Scholes–Merton model for a European put option with the same strike price and time to maturity is $1.00. What should the market price of the put option be? Explain the reasons for your answer.

17.11 Explain what is meant by "crashophobia."

17.12 A stock price is currently $20. Tomorrow, news is expected to be announced that will either increase the price by $5 or decrease the price by $5. What are the problems in using Black–Scholes–Merton to value 1-month options on the stock?

17.13 What volatility smile is likely to be observed for 6-month options when the volatility is uncertain and positively correlated with the stock price?

17.14 Explain the problems in testing a stock option pricing model empirically.

17.15 Suppose that a central bank's policy is to allow an exchange rate to fluctuate between 0.97 and 1.03. What pattern of implied volatilities for options on the exchange rate would you expect to see?

17.16 Option traders sometimes refer to deep-out-of-the-money options as being options on volatility. Why do you think they do this?

17.17 A European call option on a certain stock has a strike price of $30, a time to maturity of 1 year, and an implied volatility of 30%. A European put option on the same stock has a strike price of $30, a time to maturity of 1 year, and an implied volatility of 33%. What is the arbitrage opportunity open to a trader? Does the arbitrage work only when the lognormal assumption underlying Black–Scholes–Merton holds? Explain carefully the reasons for your answer.

17.18 Suppose that the result of a major lawsuit affecting a company is due to be announced tomorrow. The company's stock price is currently $60. If the ruling is favorable to the company, the stock price is expected to jump to $75. If it is unfavorable, the stock is expected to jump to $50. What is the risk-neutral probability of a favorable ruling?

Assume that the volatility of the company's stock will be 25% for 6 months after the ruling if the ruling is favorable and 40% if it is unfavorable. Use DerivaGem to calculate the relationship between implied volatility and strike price for 6-month European options on the company today. The company does not pay dividends. Assume that the 6-month risk-free rate is 6%. Consider call options with strike prices of $30, $40, $50, $60, $70, and $80.

17.19 An exchange rate is currently 0.8000. The volatility of the exchange rate is quoted as 12% and interest rates in the two countries are the same. Using the lognormal assumption, estimate the probability that the exchange rate in 3 months will be (a) less than 0.7000, (b) between 0.7000 and 0.7500, (c) between 0.7500 and 0.8000, (d) between 0.8000 and 0.8500, (e) between 0.8500 and 0.9000, and (f) greater than 0.9000. Based on the volatility smile usually observed in the market for exchange rates, which of these estimates would you expect to be too low and which would you expect to be too high?

17.20 A stock price is $40. A 6-month European call option on the stock with a strike price of $30 has an implied volatility of 35%. A 6-month European call option on the stock with a strike price of $50 has an implied volatility of 28%. The 6-month risk-free rate is 5% and no dividends are expected. Explain why the two implied volatilities are different. Use DerivaGem to calculate the prices of the two options. Use put–call parity to calculate the prices of 6-month European put options with strike prices of $30 and $50. Use DerivaGem to calculate the implied volatilities of these two put options.

17.21 "The Black–Scholes–Merton model is used by traders as an interpolation tool." Discuss this view.

17.22 Using Table 17.2, calculate the implied volatility a trader would use for an 8-month option with $K/S_0 = 1.04$.

17.23 A company's stock is selling for $4. The company has no outstanding debt. Analysts consider the liquidation value of the company to be at least $300,000 and there are 100,000 shares outstanding. What volatility smile would you expect to see?

17.24 Data for a number of foreign currencies are provided on the author's website: http://www2.rotman.utoronto.ca/~hull/data

Choose a currency and use the data to produce a table similar to Table 17.1.

17.25 Data for a number of stock indices are provided on the author's website: http://www2.rotman.utoronto.ca/~hull/data

Choose an index and test whether a three-standard-deviation down movement happens more often than a three-standard-deviation up movement.

17.26 Consider a European call and a European put with the same strike price and time to maturity. Show that they change in value by the same amount when the volatility increases from a level σ_1 to a new level σ_2 within a short period of time. (*Hint*: Use put–call parity.)

17.27 An exchange rate is currently 1.0 and the implied volatilities of 6-month European options with strike prices 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3 are 13%, 12%, 11%, 10%, 11%, 12%, 13%. The domestic and foreign risk-free rates are both 2.5%. Calculate the implied probability distribution using an approach similar to that used for Example 17A.1 in the appendix to this chapter. Compare it with the implied distribution where all the implied volatilities are 11.5%.

17.28 Using Table 17.2, calculate the implied volatility a trader would use for an 11-month option with $K/S_0 = 0.98$.

APPENDIX

Determining implied risk-neutral distributions from volatility smiles

The price of a European call option on an asset with strike price K and maturity T is given by

$$c = e^{-rT} \int_{S_T=K}^{\infty} (S_T - K)\, g(S_T) dS_T$$

where r is the interest rate (assumed constant), S_T is the asset price at time T, and g is the risk-neutral probability density function of S_T. Differentiating once with respect to K gives

$$\frac{\partial c}{\partial K} = -e^{-rT} \int_{S_T=K}^{\infty} g(S_T) dS_T$$

Differentiating again with respect to K gives

$$\frac{\partial^2 c}{\partial K^2} = e^{-rT} g(K)$$

This shows that the probability density function g is given by

$$g(K) = e^{rT} \frac{\partial^2 c}{\partial K^2} \qquad (17A.1)$$

This result, which is from Breeden and Litzenberger (1978), allows risk-neutral probability distributions to be estimated from volatility smiles.[6] Suppose that c_1, c_2, and c_3 are the prices of T-year European call options with strike prices of $K - \delta$, K, and $K + \delta$, respectively. Assuming δ is small, an estimate of $g(K)$, obtained by approximating the partial derivative in equation (17A.1), is

$$e^{rT} \frac{c_1 + c_3 - 2c_2}{\delta^2}$$

For another way of understanding this formula, suppose you set up a butterfly spread with strike prices $K - \delta$, K, and $K + \delta$, and maturity T. This means that you buy a call with strike price $K - \delta$, buy a call with strike price $K + \sigma$, and sell two calls with strike price K. The value of your position is $c_1 + c_3 - 2c_2$. The value of the position can also be calculated by integrating the payoff over the risk-neutral probability distribution, $g(S_T)$, and discounting at the risk-free rate. The payoff is shown in Figure 17A.1. Since δ is small, we can assume that $g(S_T) = g(K)$ in the whole of the range

[6] See D. T. Breeden and R. H. Litzenberger, "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51 (1978), 621–51.

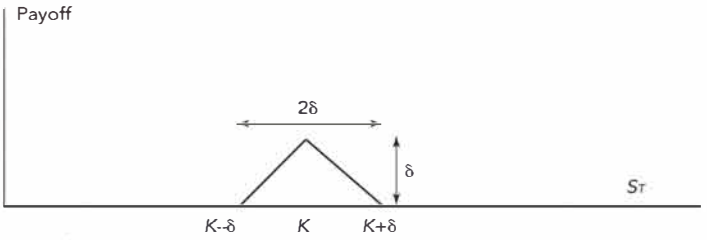


Figure 17A.1 Payoff from butterfly spread.

$K - \delta < S_T < K + \delta$, where the payoff is nonzero. The area under the "spike" in Figure 17A.1 is $0.5 \times 2\delta \times \delta = \delta^2$. The value of the payoff (when δ is small) is therefore $e^{-rT} g(K)\delta^2$. It follows that

$$e^{-rT} g(K)\delta^2 = c_1 + c_3 - 2c_2$$

which leads directly to

$$g(K) = e^{rT} \frac{c_1 + c_3 - 2c_2}{\delta^2} \qquad (17A.2)$$

Example 17A.1

Suppose that the price of a non-dividend-paying stock is \$10, the risk-free interest rate is 3%, and the implied volatilities of 3-month European options with strike prices of \$6, \$7, \$8, \$9, \$10, \$11, \$12, \$13, \$14 are 30%, 29%, 28%, 27%, 26%, 25%, 24%, 23%, 22%, respectively. One way of applying the above results is as follows. Assume that $g(S_T)$ is constant between $S_T = 6$ and $S_T = 7$, constant between $S_T = 7$ and $S_T = 8$, and so on. Define:

$$g(S_T) = g_1 \text{ for } 6 \le S_T < 7$$
$$g(S_T) = g_2 \text{ for } 7 \le S_T < 8$$
$$g(S_T) = g_3 \text{ for } 8 \le S_T < 9$$
$$g(S_T) = g_4 \text{ for } 9 \le S_T < 10$$
$$g(S_T) = g_5 \text{ for } 10 \le S_T < 11$$
$$g(S_T) = g_6 \text{ for } 11 \le S_T < 12$$
$$g(S_T) = g_7 \text{ for } 12 \le S_T < 13$$
$$g(S_T) = g_8 \text{ for } 13 \le S_T < 14$$

The value of g_1 can be calculated by interpolating to get the implied volatility for a 3-month option with a strike price of \$6.5 as 29.5%. This means that options with strike prices of \$6, \$6.5, and \$7 have implied volatilities of 30%, 29.5%, and 29%, respectively. From DerivaGem their prices are \$4.045, \$3.549, and \$3.055, respectively. Using equation (17A.2), with $K = 6.5$ and $\delta = 0.5$, gives

$$g_1 = \frac{e^{0.03\times 0.25}(4.045 + 3.055 - 2 \times 3.549)}{0.5^2} = 0.0057$$

Similar calculations show that

$$g_2 = 0.0444,\ g_3 = 0.1545,\ g_4 = 0.2781$$

$$g_5 = 0.2813,\ g_6 = 0.1659,\ g_7 = 0.0573,\ g_8 = 0.0113$$

Figure 17A.2 displays the implied distribution. (Note that the area under the probability distribution is 0.9985. The probability that $S_T < 6$ or $S_T > 14$ is therefore 0.0015.) Although not obvious from Figure 17A.2, the implied distribution does have a heavier left tail and less heavy right tail than a lognormal distribution. For the lognormal distribution based on a single volatility of 26%, the probability of a stock price between \$6 and \$7 is 0.0031 (compared with 0.0057 in Figure 17A.2) and the probability of a stock price between \$13 and \$14 is 0.0167 (compared with 0.0113 in Figure 17A.2).

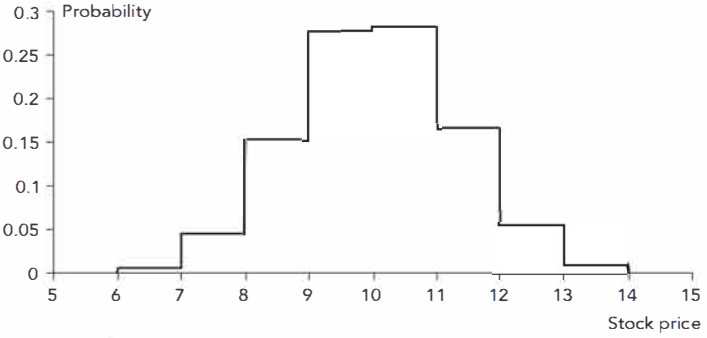


Figure 17A.2 **Implied probability distribution for Example 17A.1.**

18

Fundamental Review of the Trading Book

Learning Objectives

After completing this reading, you should be able to:

- Describe the changes to the Basel framework for calculating market risk capital under the Fundamental Review of the Trading Book (FRTB) and the motivations for these changes.
- Compare the various liquidity horizons proposed by the FRTB for different asset classes and explain how a bank can calculate its expected shortfall using the various horizons.
- Explain the FRTB revisions to Basel regulations in the following areas:
  - Classification of positions in the trading book compared to the banking book.
  - Backtesting, profit and loss attribution, credit risk, and securitizations.

Excerpt is Chapter 27 of Risk Management and Financial Institutions, *Sixth Edition, by John C. Hull.*

In May 2012, the Basel Committee on Banking Supervision issued a consultative document proposing major revisions to the way regulatory capital for market risk is calculated. This is referred to as the "Fundamental Review of the Trading Book" (FRTB).[1] The Basel Committee then followed its usual process of requesting comments from banks, revising the proposals, and carrying out Quantitative Impact Studies (QISs).[2] The final version of the rules was published by the Basel Committee in January 2019.[3] The internationally agreed implementation date is January 2023, but many jurisdictions indicated that they will be a year or two later than that.

FRTB's approach to determining capital for market risk is much more complex than the approaches previously used by regulators. The purpose of this chapter is to outline its main features.

18.1 BACKGROUND

The Basel I calculations of market risk capital were based on a value at risk (VaR) calculated for a 10-day horizon with a 99% confidence level. The VaR was "current" in the sense that calculations made on a particular day were based on the behavior of market variables during an immediately preceding period of time (typically, one to four years). Basel II.5 required banks to calculate a "stressed VaR" measure in addition to the current measure. As explained in Sections 12.1 and 26.1, this is VaR where calculations are based on the behavior of market variables during a 250-day period of stressed market conditions. To determine the stressed period, banks were required to go back through time searching for a 250-day period where the observed movements in market variables would lead to significant financial stress for the current portfolio.

FRTB changes the measure used for determining market risk capital. Instead of VaR with a 99% confidence level, it uses expected shortfall (ES) with a 97.5% confidence level. The measure is actually stressed ES with a 97.5% confidence. This means that, as in the case of stressed VaR, calculations are based on the way market variables have been observed to move during stressed market conditions.

For normal distributions, VaR with a 99% confidence and ES with a 97.5% confidence are almost exactly the same. Suppose losses have a normal distribution with a mean μ and standard deviation σ. The 99% VaR is $\mu + 2.326\sigma$ while the 97.5% expected shortfall is $\mu + 2.338\sigma$.[4] (See Problem 18.2.) For non-normal distributions, they are not equivalent. When the loss distribution has a heavier tail than a normal distribution, the 97.5% ES can be considerably greater than the 99% VaR.

Under FRTB, the 10-day time horizon used in Basel I and Basel II.5 is changed to reflect the liquidity of the market variable being considered. FRTB considers changes to market variables that would take place (in stressed market conditions) over periods of time reflecting their liquidity. The changes are referred to as *shocks*. The market variables are referred to as *risk factors*. The periods of time considered are referred to as *liquidity horizons*. Five different liquidity horizons are specified: 10 days, 20 days, 40 days, 60 days, and 120 days. The allocation of risk factors to these liquidity horizons is indicated in Table 18.1.

FRTB specifies both a standardized approach and an internal models approach for calculating market risk capital. Even when banks have been approved to use the internal models approach, they are required by regulators to calculate required capital under both approaches. This is consistent with the Basel Committee's plans to use standardized approaches to provide a floor for capital requirements. As discussed in Section 26.4, in December 2017, the Basel Committee announced a move to a situation where total required capital is at least 72.5% of that given by standardized approaches. It will achieve this by 2028 with a five-year phase-in period. These changes are a culmination of a trend by the Basel Committee since the 2008 crisis to place less reliance on internal models and to use standardized models to provide a floor for capital requirements.

A difference between FRTB and previous market risk regulatory requirements is that most calculations are carried out at the trading desk level. Furthermore, permission to use the internal models approach is granted on a desk-by-desk basis. Therefore it is possible that, at a particular point in time, a bank's foreign currency trading desk has permission to use the internal models approach while the equity trading desk does not.

In earlier chapters, we saw how the ways in which capital is calculated for the trading book and the banking book are quite different. This potentially gives rise to regulatory arbitrage where banks choose to allocate instruments to either the trading book or the banking book so as to minimize capital. In Basel II.5, the incremental risk charge made this less attractive. FRTB counteracts regulatory arbitrage by defining more clearly than previously the differences between the two books.

[1] See Bank for International Settlements, "Consultative Document: Fundamental Review of the Trading Book," May 2012.
[2] QISs are calculations carried out by banks to estimate the impact of proposed regulatory changes on capital requirements.
[3] See Bank for International Settlements, "Minimum Capital Requirements for Market Risk," January 2019.

[4] From equation (11.2), the ES for a normal distribution with mean μ and standard deviation σ is $\mu + \sigma \exp(-Y^2/2)/[\sqrt{2\pi}(1 - X)]$ where X is the confidence level and Y is the point on a normal distribution that has a probability of $1 - X$ of being exceeded. This can also be written $\mu + \sigma^2 f(VaR)/(1 - X)$ where f is the probability density function for the loss.

Table 18.1 Allocation of Risk Factors to Liquidity Horizons

Risk Factor	Horizon (days)
Interest rate (dependent on currency)	10–60
Interest rate volatility	60
Credit spread: sovereign, investment grade	20
Credit spread: sovereign, non-investment grade	40
Credit spread: corporate, investment grade	40
Credit spread: corporate, non-investment grade	60
Credit spread: other	120
Credit spread volatility	120
Equity price: large cap	10
Equity price: small cap	20
Equity price: large cap volatility	20
Equity price: small cap volatility	60
Equity: other	60
Foreign exchange rate (dependent on currency)	10–40
Foreign exchange volatility	40
Energy price	20
Precious metal price	20
Other commodities price	60
Energy price volatility	60
Precious metal volatility	60
Other commodities price volatility	120
Commodity (other)	120

18.2 STANDARDIZED APPROACH

Under the standardized approach, the capital requirement is the sum of three components: a risk charge calculated using a risk sensitivity approach, a default risk charge, and a residual risk add-on.

Consider the first component. Seven risk classes (corresponding to trading desks) are defined (general interest rate risk, foreign exchange risk, commodity risk, equity risk, and three categories of credit spread risk). Within each risk class, a delta risk charge, vega risk charge, and curvature risk charge are calculated.

The delta risk charge for a risk class is calculated using the risk weights and weighted sensitivity approach described in Section 13.6:

$$\text{Delta Risk Charge} = \sum_i \sum_j > \rho_{ij}\delta_i\delta_j W_i W_j \qquad (18.1)$$

where the summations are taken over all risk factors in the risk class. The risk weights, W_i, and the correlations between risk factors, ρ_{ij}, are determined by the Basel Committee.[5] The sensitivities (or deltas), δ_i, are determined by the bank. In the case of risk factors such as equity prices, exchange rates, or commodity prices, the deltas measure the sensitivity of the portfolio to percentage changes. For example, if a 1% increase in a commodity price would increase the value of a portfolio by $3,000, the delta would be 3,000/0.01 = 300,000. In the case of risk factors such as interest rates and credit spreads, the deltas are defined in terms of absolute changes. For example, if the effect of an interest rate increasing by one basis point (0.0001) is to reduce the value of a portfolio by $200, the delta with respect to that interest rate would be $-200/0.0001 = -2{,}000{,}000$.

Consider how the risk weights, W_i, might be set by regulators. Suppose first that all risk factors are equity prices, exchange rates, or commodity prices, so the deltas are sensitivities to percentage changes. If W_i were set equal to the daily volatility of risk factor i for all i, the risk charge in equation (18.1) would equal the standard deviation of change in the value of the portfolio per day. If W_i were set equal to the daily volatility of risk factor i in stressed market conditions (the stressed daily volatility) for all i, equation (18.1) would give the standard deviation of the daily change of the portfolio in stressed market conditions. In practice, the W_i are set equal to multiples of the stressed daily volatility to reflect the liquidity horizon and the confidence level that regulators wish to consider. Suppose that the stressed daily volatility of risk factor i is estimated as 2% and that the risk factor has a 20-day liquidity horizon. The risk weight might be set as $0.02 \times \sqrt{20} \times 2.338 = 0.209$. (Note that the 2.338 multiplier reflects the amount by which a standard deviation has to be multiplied to get ES with a 97.5% confidence when a normal distribution is assumed.)

Now suppose that the risk factors are interest rates and credit spreads so that deltas are sensitivities with respect to actual changes measured in basis points. The W_i for risk factor i is set equal to a multiple of the stressed daily standard deviation for all i. If the multiple were 1, the formula would give the standard

[5] Banks are required to test the effect of multiplying the correlations specified by the Basel Committee by 1.25, 1.00, and 0.75 and then set the capital charge equal to the greatest result obtained.

deviation of the value of the portfolio in one day. In practice the multiple is determined as just described to reflect the liquidity horizon and confidence level.

Vega risk is handled similarly to delta risk.[6] A vega risk charge is calculated for each risk class using equation (18.1). The risk factors (counted by the i and j) are now volatilities. The summation is taken over all volatilities in the risk class. The parameter δ_i is actually a vega. It is the sensitivity of the value of the portfolio to small changes in volatility i.[7] The parameter ρ_{ij} is the correlation between changes in volatility i and volatility j, and W_i is the risk weight for volatility i. The latter is determined similarly to the delta risk weights to reflect the volatility of the volatility i, its liquidity horizon, and the confidence level.

There are assumed to be no diversification benefits between risk factors in different risk classes and between the vega risks and delta risks within a risk class. The end product of the calculations we have described so far is therefore the sum of the delta risk charges across the seven risk classes plus the sum of the vega risk charges across the seven risk classes.

18.2.1 Term Structures

In the case of risk factors such as interest rates, volatilities, and credit spreads, there is usually a term structure defined by a number of points. For example, an interest rate term structure is sometimes defined by 10 points. These are the zero-coupon interest rates for maturities of 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 10 years, 15 years, 20 years, and 30 years. Each vertex of the term structure is a separate risk factor for the purposes of using equation (18.1). The delta of a portfolio with respect to a one basis point move in one of the vertices on the term structure is calculated by increasing the position of the vertex by one basis point while making no change to the other vertices. The Basel Committee defines risk weights for each vertex of the term structure and correlations between the vertices of the same term structure.

A simplification is used when correlations between points on different term structures are defined. The correlations between point A on term structure 1 and point B on term structure 2 are assumed to be the same for all A and B.

[6] This works well because most of the value of a derivative is in many cases approximately linearly dependent on volatility.
[7] Banks can choose whether it is percentage or actual changes in volatility that are considered.

18.2.2 Curvature Risk Charge

The curvature risk charge is a capital charge for a bank's gamma risk exposure under the standardized approach. Consider the exposure of a portfolio to the ith risk factor. Banks are required to test the effect of increasing and decreasing the risk factor by its risk weight, W_i. If the portfolio is linearly dependent on the risk factor, the impact of an increase of W_i in the risk factor is $W_i\delta_i$. Similarly, the impact of a decrease of W_i in the risk factor is $-\delta_i W_i$. To evaluate the impact of curvature net of the delta effect, the standardized approach therefore calculates

1. $W_i\delta_i$ minus the impact of a increase of W_i in the risk factor, and
2. $-W_i\delta_i$ minus the impact of a decrease in the risk factor of W_i.

The curvature risk charge for the risk factor is the greater of these two. If the impact of curvature net of delta is negative, it is counted as zero. The calculation is illustrated in Figure 18.1. In Figure 18.1a, the portfolio value is currently given by point O. If there were no curvature, an increase of W_i in the risk factor would lead to the portfolio value at point C, whereas a decrease of W_i in the risk factor would lead to the portfolio value at point A. Because of curvature, an increase of W_i leads to the portfolio value at point D, and a decrease of W_i leads to the portfolio value at point B. Since AB > CD, the risk charge is AB. In Figure 18.1b, the risk charge is zero because curvature actually increases the value of the position (relative to what delta would suggest) for both increases and decreases in the risk factor. (Figure 18.1a could correspond to a short position in an option; Figure 18.1b could correspond to a long position in an option.)

When there are several risk factors, each is handled similarly to Figure 18.1. When there is a term structure (e.g., for interest rates, credit spreads, and volatilities), all points are shifted by the same amount for the purpose of calculating the effect of curvature. The shift is the largest W_i for the points on the term structure. In the case of an interest rate term structure, the W_i corresponding to the three-month vertex might be the largest W_i, so this would define an upward and downward parallel shift in the term structure. The delta effect is removed for each point on the term structure by using the δ_i for that point.

The curvature risk charges for different risk factors are combined to determine a total curvature risk charge. When diversification benefits are allowed, aggregation formulas broadly similar to those used for deltas are used with correlations specified by the Basel Committee.

18.2.3 Default Risk Charge

Risks associated with counterparty credit spread changes are handled separately from risks associated with counterparty

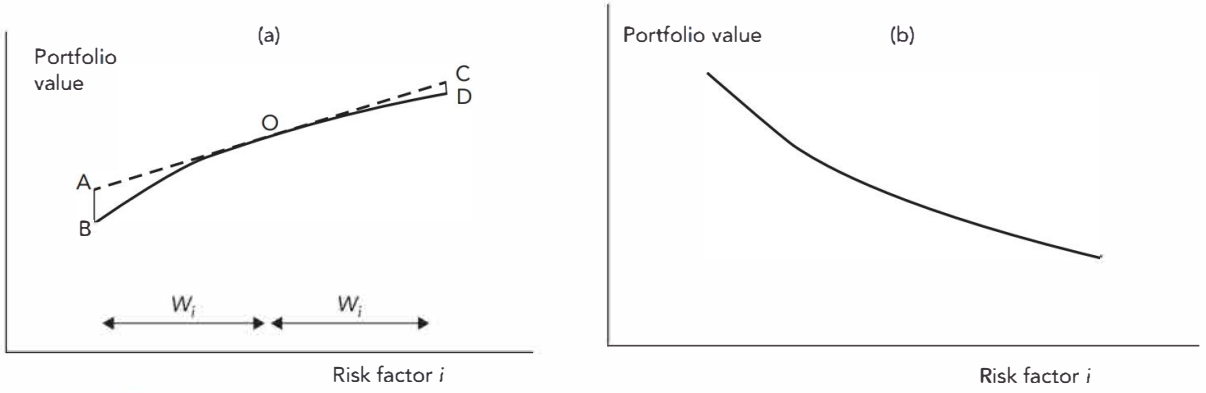


Figure 18.1 Calculation of curvature risk charge for a risk factor

In Figure 18.1a, the curvature risk charge is AB; in Figure 18.1b, it is zero.

defaults in FRTB. In the standardized approach, credit spread risks are handled using the delta/vega/curvature approach described earlier. Default risks, sometimes referred to as *jump-to-default* (JTD) risks, are handled by a separate default risk charge. This is calculated by multiplying each exposure by a loss given default (LGD) and a default risk weight. Both the LGD and the risk weight are specified by the Basel Committee. For example, the LGD for senior debt is specified as 75% and the default risk weight for a counterparty rated A is 3%. Equity positions are subject to a default risk charge with an LGD = 100%. Rules for offsetting exposures are specified.

18.2.4 Residual Risk Add-On

The residual risk add-on considers risks that cannot be handled by the delta/vega/curvature approach described earlier. It includes exotic options when they cannot be considered as linear combinations of plain vanilla options. The add-on is calculated by multiplying the notional amount of the transaction by a risk weight that is specified by the Basel Committee. In the case of exotic options the risk weight is 1%.

18.2.5 A Simplified Approach

In this section, we have described the standardized approach that the Basel Committee requires all large banks to use. In June 2017 the Basel Committee published a consultative document outlining a simplified standardized approach that it proposes for smaller banks.[8] This has been included in the final January 2019 document. The full approach is simplified in a number of ways. For example, vega and gamma risk do not have to be considered. This should make FRTB more attractive to jurisdictions such as the United States that have many small banks that tend to enter into only relatively simple transactions.

18.3 INTERNAL MODELS APPROACH

The internal models approach requires banks to estimate stressed ES with a 97.5% confidence. FRTB does not prescribe a particular method for doing this. Typically the historical simulation approach is likely to be used. Risk factors are allocated to liquidity horizons as indicated in Table 18.1. Define:

Category 1 Risk Factors: Risk factors with a time horizon of 10 days

Category 2 Risk Factors: Risk factors with a time horizon of 20 days

Category 3 Risk Factors: Risk factors with a time horizon of 40 days

Category 4 Risk Factors: Risk factors with a time horizon of 60 days

Category 5 Risk Factors: Risk factors with a time horizon of 120 days

As we shall see, all calculations are based on considering 10-day changes in the risk factors. In Basel I and Basel II.5, banks are allowed to deduce the impact of 10-day changes from the impact of one-day changes using a $\sqrt{10}$ multiplier. In FRTB, banks are required to consider changes over periods of 10 days that occurred during a stressed period in the past. Econometricians naturally prefer that non-overlapping periods be used when VaR or ES is being estimated using historical simulation, because they want observations on the losses to be

[8] See Basel Committee on Banking Supervision, "Simplified Alternative to the Standardized Approach to Market Risk Capital Requirements," June 2017.

independent. However, this is not feasible when 10-day changes are considered, because it would require a very long historical period. FRTB requires banks to base their estimates on overlapping 10-day periods. The first simulation trial assumes that the percentage changes in all risk factors over the next 10 days will be the same as their changes between Day 0 and Day 10 of the stressed period; the second simulation trial assumes that the percentage changes in all risk factors over the next 10 days will be the same as their changes between Day 1 and Day 11 of the stressed period; and so on.

Banks are first required to calculate ES when 10-day changes are made to all risk factors. (We will denote this by ES_1.) They are then required to calculate ES when 10-day changes are made to all risk factors in categories 2 and above with risk factors in category 1 being kept constant. (We will denote this by ES_2.) They are then required to calculate ES when 10-day changes are made to all risk factors in categories 3, 4, and 5 with risk factors in categories 1 and 2 being kept constant. (We will denote this by ES_3.) They are then required to calculate ES when 10-day changes are made to all risk factors in categories 4 and 5 with risk factors in categories 1, 2, and 3 being kept constant. (We will denote this by ES_4.) Finally, they are required to calculate ES_5, which is the effect of making 10-day changes only to category 5 risk factors.

The liquidity-adjusted ES is calculated as

$$\sqrt{ES_1^2 + \sum_{j=2}^{5}\left(ES_j\sqrt{\frac{LH_j - LH_{j-1}}{10}}\right)^2} \qquad (18.2)$$

where LH_j is the liquidity horizon for category j. To understand equation (18.2), suppose first that all risk factors are in category 1 or 2 so that only ES_1 and ES_2 are calculated. It is assumed that the behavior of all risk factors during a 10-day period is independent of the behavior of category 2 risk factors during a further 10-day period. An extension of the square root rule then leads to the liquidity-adjusted ES being

$$\sqrt{ES_1^2 + ES_2^2}$$

Now suppose that there are also category 3 risk factors. The expression $\sqrt{ES_1^2 + ES_2^2}$ would be correct if the category 3 risk factors had a 20-day instead of a 40-day liquidity horizon. We assume that the behavior of the category 3 risk factors over an additional 20 days is independent of the behavior of all the risk factors over the periods already considered. We also assume that the ES for the category 3 risk factors over 20 days is $\sqrt{2}$ times their ES over 10 days. This leads to a liquidity-adjusted ES of:

$$\sqrt{ES_1^2 + ES_2^2 + 2ES_3^2}$$

Continuing in this way, we obtain equation (18.2). This is referred to as the *cascade approach* to calculating ES (and can be used for VaR as well).

Calculations are carried out for each desk. If there are six desks, this means the internal models approach, as we have described it so far, requires $5 \times 6 = 30$ ES calculations. As mentioned, the use of overlapping time periods is less than ideal because changes in successive historical simulation trials are not independent. This does not bias the results, but it reduces the effective sample size, making results more noisy than they would otherwise be.

FRTB represents a movement away from basing calculations on one-day changes. Presumably the Basel Committee has decided that, in spite of the lack of independence of observations, a measure calculated from 10-day changes provides more relevant information than a measure calculated from one-day changes. This could be the case if changes on successive days are not independent, but changes in successive 10-day periods can reasonably be assumed to be independent.

The calculation of a stressed measure (VaR or ES) requires banks to search for the period in the past when market variable changes would be worst for their current port folio. (The search must go back as far as 2007.) When Basel II.5 was implemented, a problem was encountered in that banks found that historical data were not available for some of their current risk factors. It was therefore not possible to know how these risk factors would have behaved during the 250-day periods in the past that were candidates for the reference stressed period. FRTB handles this by allowing the search for stressed periods to involve a subset of risk factors, provided that at least 75% of the current risk factors are used. The expected shortfalls that are calculated are scaled up by the ratio of ES for the most recent 12 months using all risk factors to ES for the most recent 12 months using the subset of risk factors. (This potentially doubles the number of ES calculations from 30 to 60.)

Banks are required to calculate ES for the whole portfolio as well for each of six trading desks. The ES for a trading desk is referred to as a *partial expected shortfall*. It is determined by shocking the risk factors belonging to the trading desk while keeping all other risk factors fixed. The sum of the partial expected shortfalls is always greater than the ES for the whole portfolio. What we will refer to as the weighted expected shortfall (WES) is a weighted average of (a) the ES for the whole portfolio and (b) the sum of the partial expected shortfalls. Specifically:

$$WES = \lambda \times EST + (1 - \lambda) \times \sum_j ESP_j$$

where EST is the expected shortfall calculated for the total portfolio and ESP_j is the jth partial expected shortfall. The parameter λ is set by the Basel Committee to be 0.5.

Some risk factors are categorized as *non-modelable*. Specifically, if there are less than 24 observations on a risk factor in a year or more than one month between successive observations, the risk factor is classified as non-modelable. Such risk factors are handled by special rules involving stress tests.

The total capital requirement for day t is

$$\max(\text{WES}_{t-1} + \text{NMC}_{t-1}, m_c \times \text{WES}_{\text{avg}} + \text{NMC}_{\text{avg}})$$

where WES_{t-1} is the WES for day $t-1$, NMC_{t-1} is the capital charge calculated for non-modelable risk factors on day $t-1$, WES_{avg} is the average WES for the previous 60 days, and NMC_{avg} is the average capital charge calculated for the non-modelable risk factors over the previous 60 days. The parameter m_c is at minimum 1.5.

18.3.1 Back-Testing

FRTB does not back-test the stressed ES measures that are used to calculate capital under the internal models approach for two reasons. First, it is more difficult to back-test ES than VaR. Second, it is not possible to back-test a stressed measure at all. The stressed data upon which a stressed measure is based are extreme data that statistically speaking are not expected to be observed with the same frequency in the future as they were during the stressed period.

FRTB back-tests a bank's models by asking each trading desk to back-test a VaR measure calculated over a one-day horizon and the most recent 12 months of data. Both 99% and 97.5% confidence levels are to be used. If there are more than 12 exceptions for the 99% VaR or more than 30 exceptions for the 97.5% VaR, the trading desk is required to calculate capital using the standardized approach until neither of these two conditions continues to exist.

Banks may be asked by regulators to carry out other back-tests. Some of these could involve calculating the *p*-value of the profit or loss on each day. This is the probability of observing a profit that is less than the actual profit or a loss that is greater than the actual loss. If the model is working perfectly, the *p*-values obtained should be uniformly distributed.

18.3.2 Profit and Loss Attribution

Another test used by the regulators is known as *profit and loss attribution*. Banks are required to compare the actual profit or loss in a day with that predicted by their models. Two measures must be calculated. The measures are:

$$\frac{\text{Mean of } U}{\text{Standard Deviation of } V}$$

$$\frac{\text{Variance of } U}{\text{Variance of } V}$$

where U denotes the difference between the actual and model profit/loss in a day and V denotes the actual profit/loss in a day.[9] Regulators expect the first measure to be between -10% and $+10\%$ and the second measure to be less than 20%. When there are four or more situations in a 12-month period where the ratios are outside these ranges, the desk must use the standardized approach for determining capital.

18.3.3 Credit Risk

As mentioned, FRTB distinguishes two types of credit risk exposure to a company:

1. *Credit spread risk* is the risk that the company's credit spread will change, causing the mark-to-market value of the instrument to change.
2. *Jump-to-default risk* is the risk that there will be a default by the company.

Under the internal models approach, the credit spread risk is handled in a similar way to market risks. Table 18.1 shows that the liquidity horizon for credit spread varies from 20 to 120 days and the liquidity horizon for a credit spread volatility is 120 days. The jump-to-default risk is handled in the same way as default risks in the banking book. In the internal models approach, the capital charge is based on a VaR calculation with a one-year time horizon and a 99.9% confidence level.

18.3.4 Securitizations

The comprehensive risk measure (CRM) charge was introduced in Basel II.5 to cover the risks in products created by securitizations such as asset-backed securities and collateralized debt obligations (see Section 26.1). The CRM rules allow a bank (with regulatory approval) to use its own models. The Basel Committee has concluded that this is unsatisfactory because there is too much variation in the capital charges calculated by different banks for the same portfolio. It has therefore decided that under FRTB the standardized approach must be used for securitizations.

[9] The "actual" profit/loss should be the profit and loss that would occur if there had been no trading in a day. This is sometimes referred to as the *hypothetical profit and loss*.

18.4 TRADING BOOK VS. BANKING BOOK

The FRTB addresses whether instruments should be put in the trading book or the banking book. Roughly speaking, the trading book consists of instruments that the bank intends to trade. The banking book consists of instruments that are expected to be held to maturity. Instruments in the banking book are subject to credit risk capital whereas those in the trading book are subject to market risk capital. The two sorts of capital are calculated in quite different ways. This has in the past given rise to regulatory arbitrage. For example, banks have often chosen to hold credit-dependent instruments in the trading book because they are then subject to less regulatory capital than they would be if they had been placed in the banking book.

The FRTB attempts to make the distinction between the trading book and the banking book clearer and less subjective. To be in the trading book, it will no longer be sufficient for a bank to have an "intent to trade." It must be able to trade and manage the underlying risks on a trading desk. The day-to-day changes in value should affect equity and pose risks to solvency. The FRTB provides rules for determining for different types of instruments whether they should be placed in the trading book or the banking book.

An important point is that instruments are assigned to the banking book or the trading book when they are initiated and there are strict rules preventing them from being subsequently moved between the two books. Transfers from one book to another can happen only in extraordinary circumstances. (Examples given of extraordinary circumstances are the closing of trading desks and a change in accounting standards with regard to the recognition of fair value.) Any capital benefit as a result of moving items between the books will be disallowed.

SUMMARY

FRTB is a major change to the way capital is calculated for market risk. After over 20 years of using VaR with a 10-day time horizon and 99% confidence to determine market risk capital, regulators are switching to using ES with a 97.5% confidence level and varying time horizons. The time horizons, which can be as high as 120 days, are designed to incorporate liquidity considerations into the capital calculations. The change that is considered to a risk factor when capital is calculated reflects movements in the risk factor over a period of time equal to the liquidity horizon in stressed market conditions.

The Basel Committee has specified a standardized approach and an internal models approach. Even when they have been approved by their supervisors to use the internal models approach, banks must also implement the standardized approach. Regulatory capital under the standardized approach is based on formulas involving the delta, vega, and gamma exposures of the trading book. Regulatory capital under the internal models approach is based on the calculation of stressed expected shortfall. Calculations are carried out separately for each trading desk.

Further Reading

Bank for International Settlements. "Minimum Capital Requirements for Market Risk," January 2019.

Practice Questions and Problems

18.1 Outline the differences between the way market risk capital is calculated in (a) Basel I, (b) Basel II.5, and (c) the FRTB.

18.2 Use footnote 4 to verify that when losses have a normal distribution with mean μ and standard deviation σ the 97.5% expected shortfall is $\mu + 2.338\sigma$.

18.3 Explain why the use of overlapping time periods proposed by the FRTB does not give independent observations on the changes in variables.

18.4 What are the advantages of expected shortfall over value at risk?

18.5 What is the difference between the trading book and the banking book? Why are regulators concerned about specifying whether an instrument should be one or the other in the FRTB?

18.6 How are credit trades handled under the FRTB?

Further Question

18.7 Suppose that an investor owns the $10 million portfolio in Table 12.1 on July 8, 2020. Suppose that the 250 days ending April 30, 2020, constitute the stressed period for the portfolio. Calculate the 97.5% expected shortfall using the over-lapping periods method in conjunction with historical simulation and the cascade approach. Relevant data on the indices is on the author's website (see "Worksheets for Value at Risk Example"). For the purposes of this problem, assume that S&P 500 and FTSE have a 10-day liquidity horizon, CAC 40 has a 40-day liquidity horizon, and Nikkei 225 has a 20-day liquidity horizon. For each day during the stressed period, consider the change in a variable over a 10-day period ending on the day.

Index

C

D

E

F

G

H

O

P

Q

R

S

T

U

V

W

Z