2024

FRM®

EXAM PART II

Market Risk Measurement and Management

GARP®

FRM® | Financial Risk Manager

GARP®
FRM® | Financial Risk Manager

2024

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.

"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

ScoutAutomatedPrintCode

00033038-00000005 / A103001321011
EEB/SK

Pearson

ISBN 10: 0-13-829218-3
ISBN 13: 978-0-13-829218-8

Contents

PREFACE

I want to thank you on behalf of GARP's Board of Trustees and our professional certification program staff for your support of the Financial Risk Manager (FRM®) program.

It's gratifying to see that in the 26 years since the first FRM examination, the FRM program has become the global standard for educating and credentialing financial risk management professionals. Its worldwide effects in furthering the understanding and acceptance of financial risk management have been highly positive and, in many ways, transformative.

COVID is thankfully in the rearview mirror. We now can be much more flexible in expanding—and in certain instances re-focusing and updating—the FRM program to address the many new challenges encountered by financial institutions globally.

Our FRM program advisory committee, consisting of senior risk professionals from around the world, that meets regularly to debate and settle the FRM program's subject coverage, has found no shortage of subjects for inclusion in the FRM curriculum.

One of the advisory committee's more-material challenges is to understand and assess where the global financial services industry is headed, and then identify issues and subjects most important for risk management professionals.

The FRM advisory committee also recommends how the FRM program covers subject matter. Its objective is to ensure that candidates who complete the FRM program successfully can be confident that their skills have been assessed objectively, and that they possess the requisite knowledge to succeed as a risk management professional anywhere in the world.

The FRM program's coverage is dynamic. The advisory committee reacts to and tries to anticipate market changes, global economic trends, technological advances, and regulatory adjustments; and assesses how these will affect the necessary knowledge and skill sets of a risk management professional.

The biggest change to the program's coverage for 2024 revolves around credit risk measurement and management. About two-thirds of the subject readings in *Credit Risk Measurement and Management* were updated for 2024.

Notably in 2023, GARP expanded the FRM program's coverage of operational resilience, an issue of rapidly growing importance around the world. Materials deal with structural vulnerabilities and areas of the financial system that may be under stress. The transmission of shocks to the financial system, and the assessment, modeling, and measurement of potential points of failure are other important covered concepts.

Also notable in 2023, GARP added two chapters on machine learning (ML) in the FRM Part I *Quantitative Analysis* book. These chapters not only introduce the ML methods risk managers need to understand, but also address key issues associated with artificial intelligence (AI) and ML, including transparency, interpretability, and explainability; data considerations; and risks that arise from the use of AI/ML, including the potential for bias, discrimination, and unethical behavior.

Throughout the FRM curriculum, GARP aims, wherever possible, to present lessons learned from noteworthy current events to contextualize program content and give FRM candidates critical insight.

As you will see from reviewing the program's coverage and readings, it keeps up with a world that is becoming more interconnected and complex by the day.

GARP is committed to offering a program that is dynamic, sophisticated, and responsive to the needs of financial institutions and risk professionals around the world.

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
Senior Technical Advisor, Operational Risk & Resilience,
Prudential Regulation Authority, Bank of England

Members

Richard Apostolik
President and CEO, GARP

Richard Brandt
MD, Operational Risk Management, Citigroup

Julian Chen, FRM
SVP, FRM Program Manager, GARP

Chris Donohue, PhD
MD, GARP Benchmarking Initiative, GARP

Donald Edgar, FRM
MD, Risk & Quantitative Analysis, BlackRock

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

Aparna Gupta
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
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
SVP, Global Head of Certifications and Educational Programs, GARP

Attilio Meucci, PhD, CFA
Founder, ARPM

Victor Ng, PhD
Chairman, Audit and Risk Committee
Former MD, Head of Risk Architecture, Goldman Sachs

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

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

Til Schuermann, PhD
Partner, Oliver Wyman

Evan Sekeris, PhD
Head of Non-Financial Risk, MUFG

Sverrir Þorvaldsson, PhD, 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.

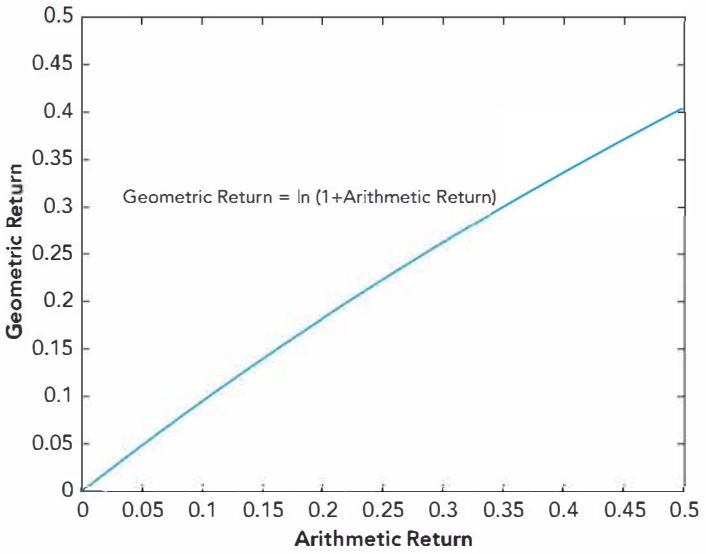


Figure 1.1 Geometric and arithmetic returns.

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:

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

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.

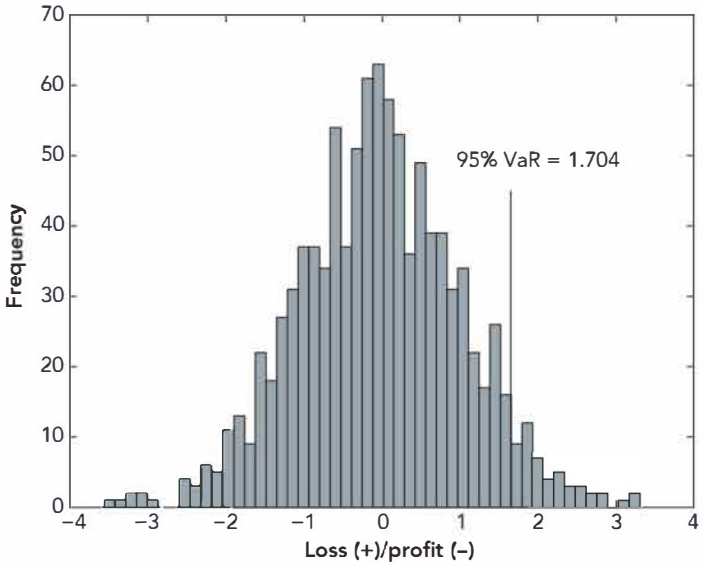


Figure 1.2 **Historical simulation VaR.**

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

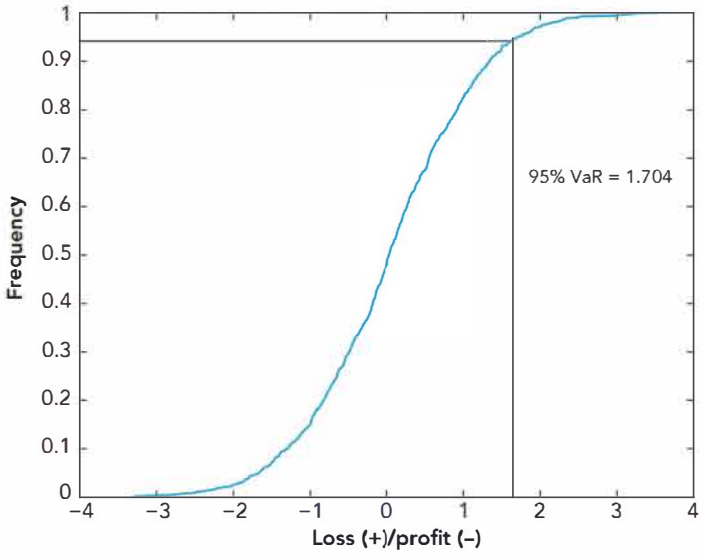


Figure 1.3 **Historical simulation via an empirical cumulative frequency function.**

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:

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

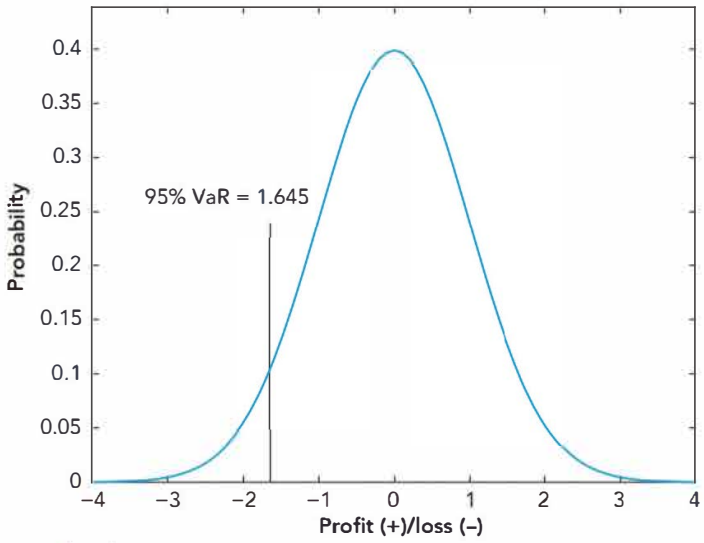


Figure 1.4 VaR with standard normally distributed profit/loss data.

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

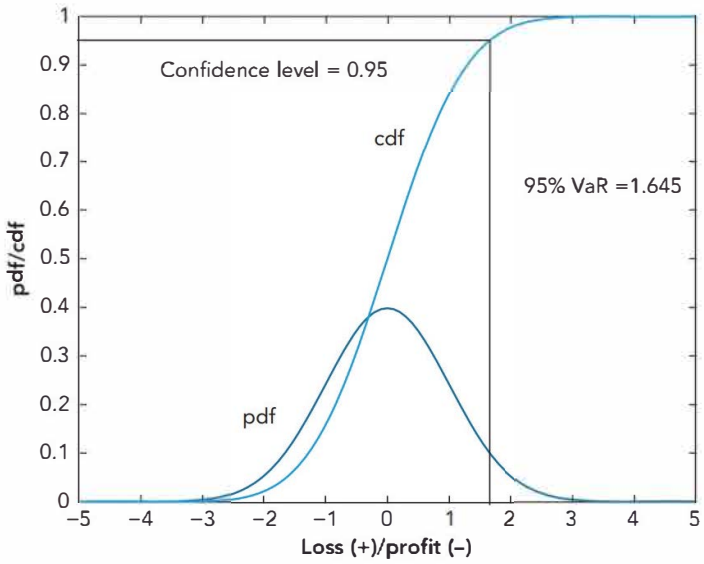


Figure 1.5 VaR with normally distributed loss/profit data.

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.

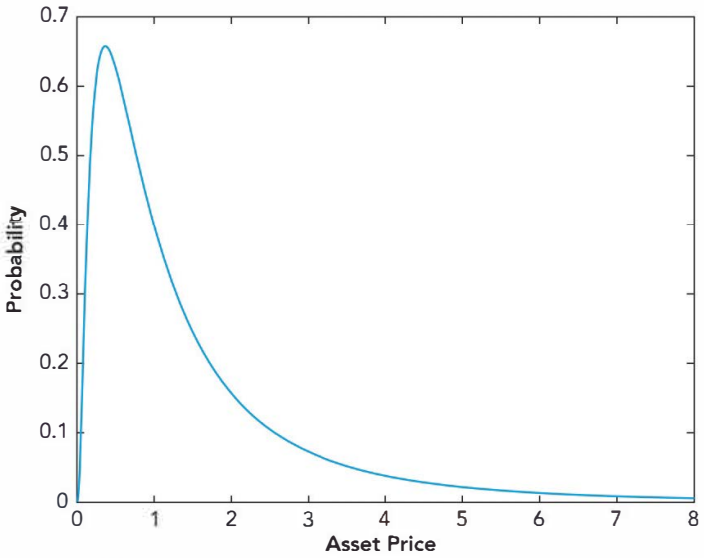


Figure 1.6 **A lognormally distributed asset price.**

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

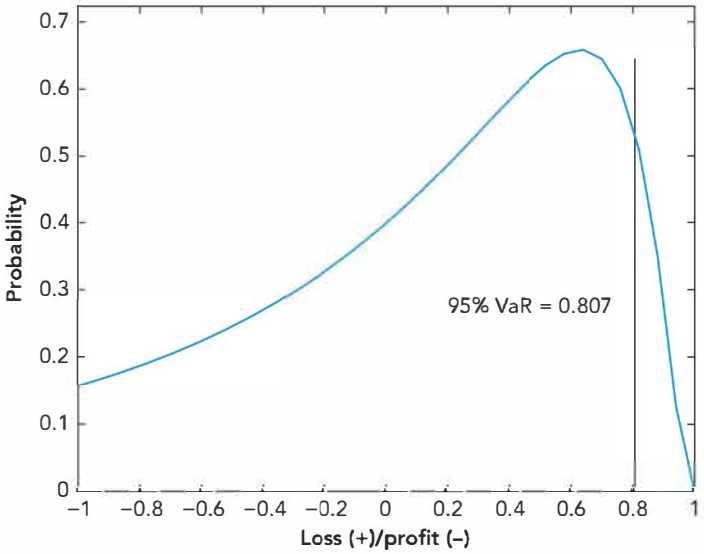


Figure 1.7 **Lognormal VaR.**

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.

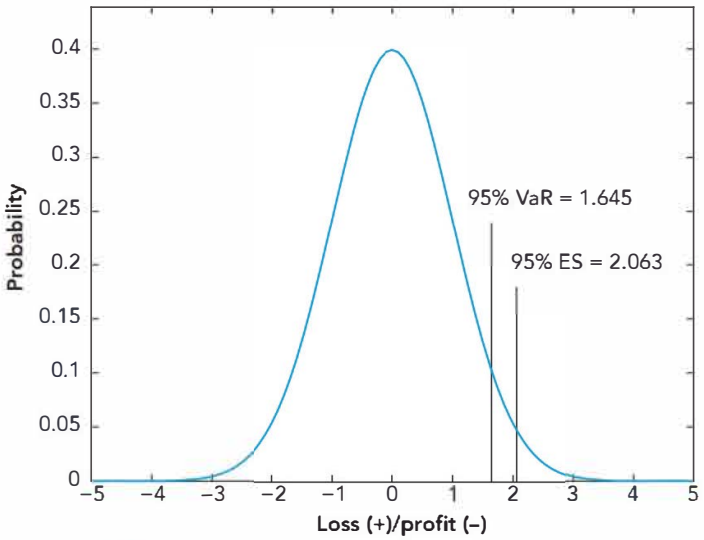


Figure 1.8 Normal VaR and ES.

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.

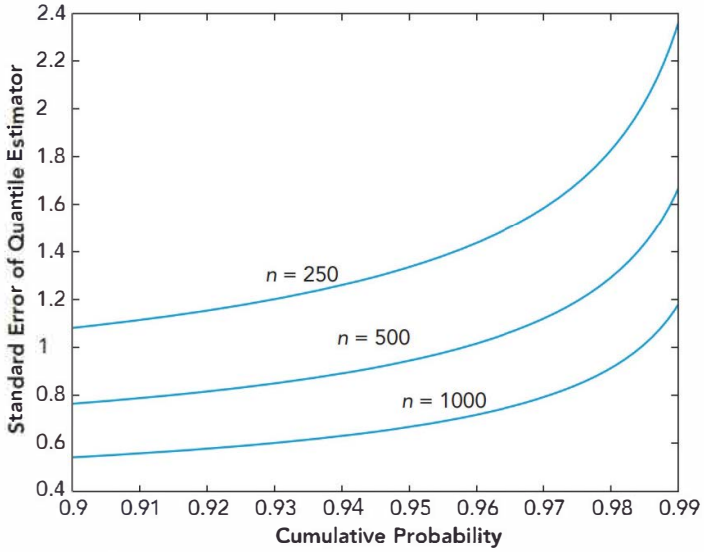


Figure 1.9 **Standard errors of quantile estimators.**

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

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

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

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

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.

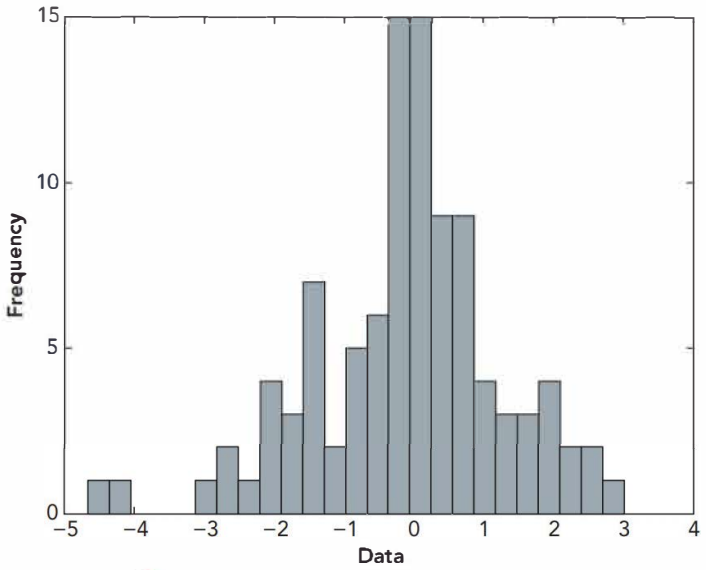


Figure 1.10 **A histogram.**

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.

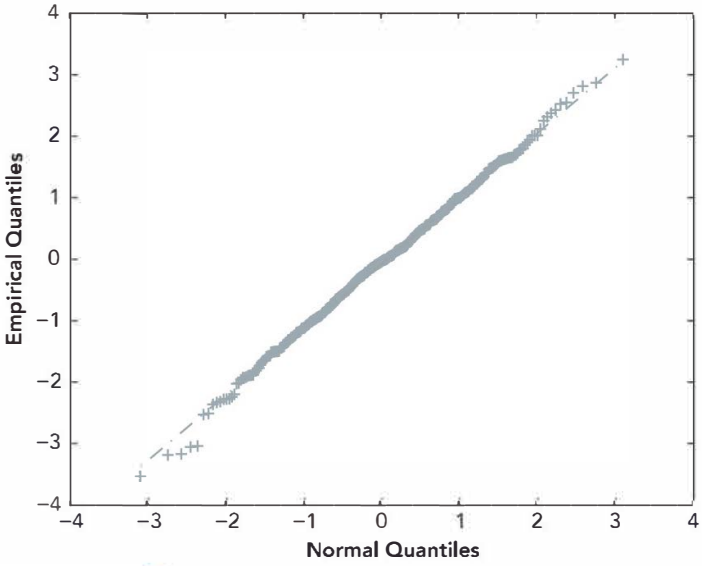


Figure 1.11 QQ plot: normal sample against normal reference distribution.

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

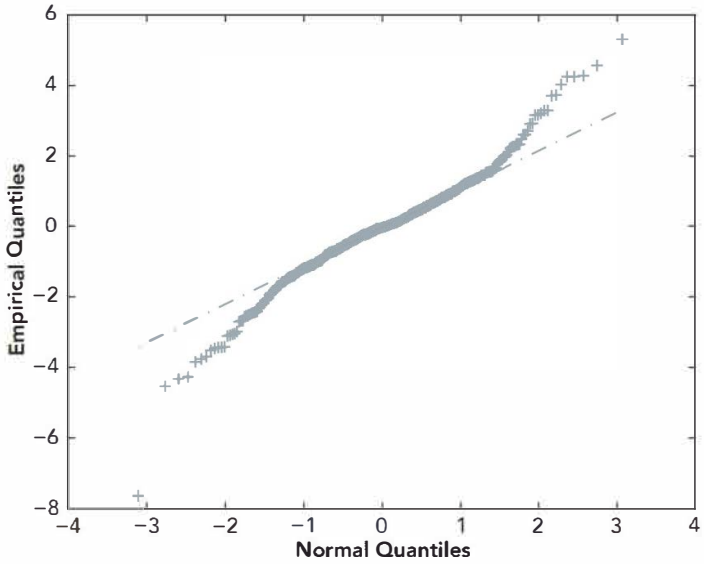


Figure 1.12 QQ plot: *t* sample against normal reference distribution.

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.

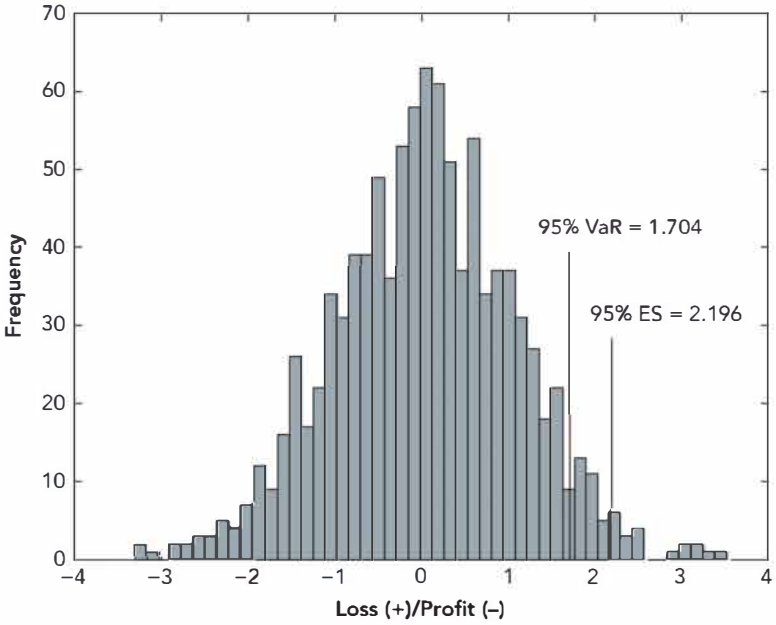


Figure 2.1 Basic historical simulation VaR and ES.

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.

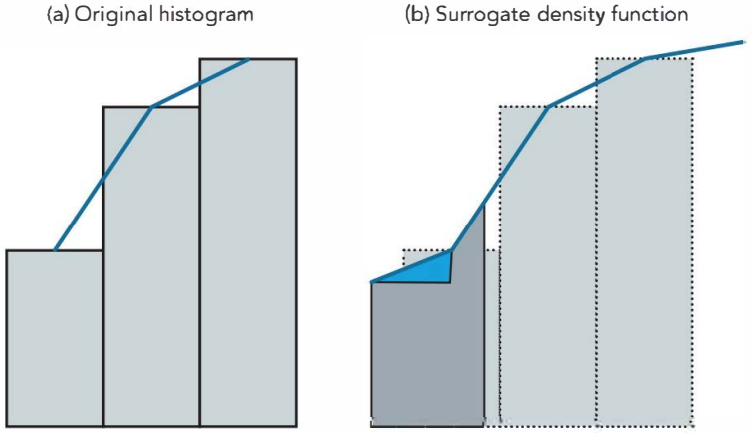


Figure 2.2 **Histograms and surrogate density functions.**

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.

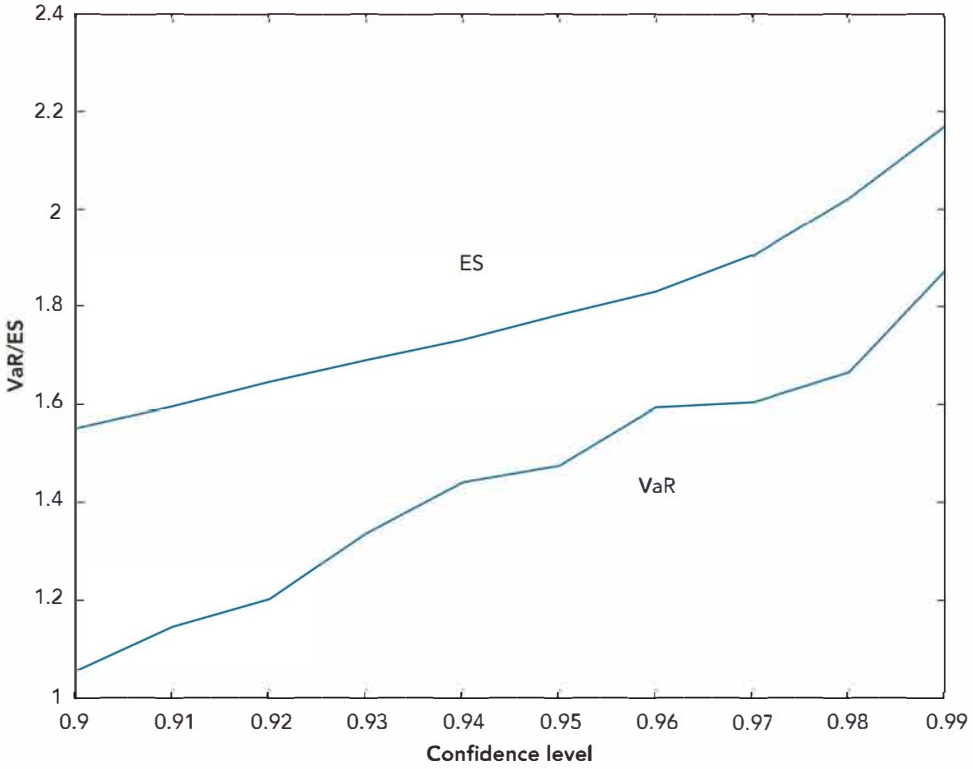


Figure 2.3 **Plots of HS VaR and ES against confidence level.**

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.

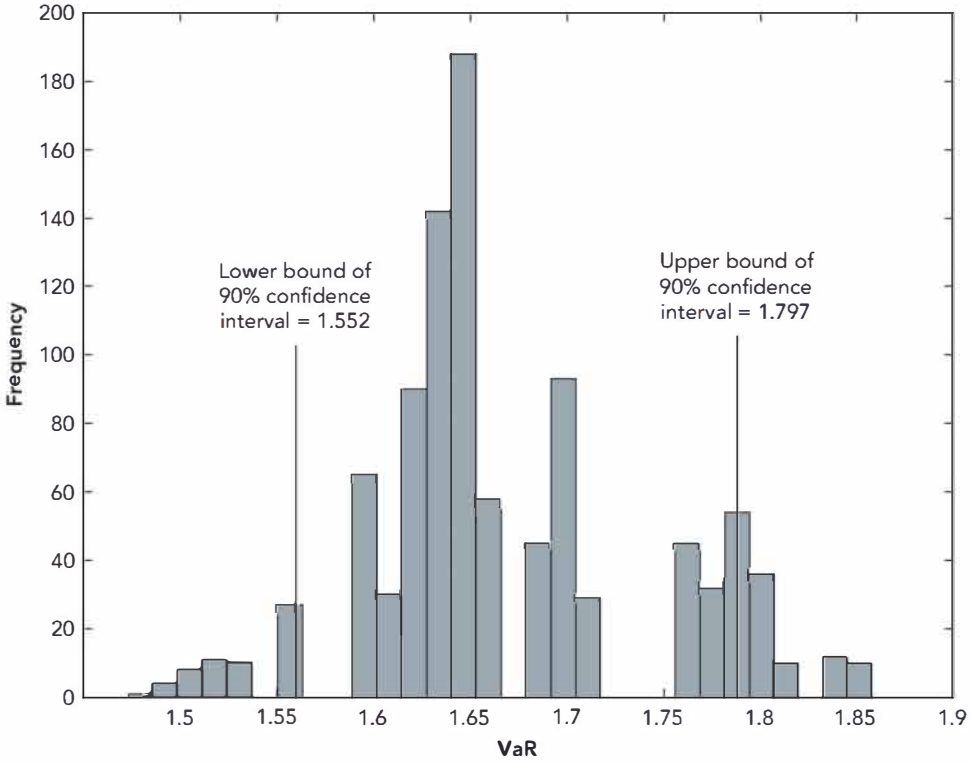


Figure 2.4 **Bootstrapped VaR.**

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).

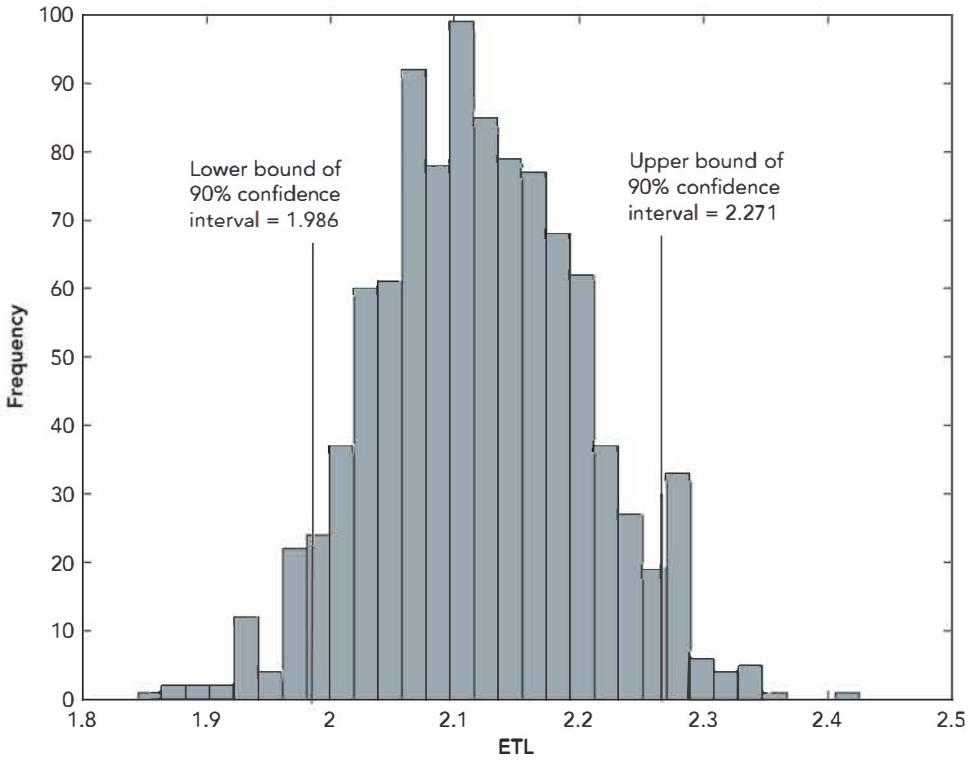


Figure 2.5 Bootstrapped ES.

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 $\lambda w(1)$; $w(3)$ could be $\lambda^2 w(1)$; 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 $\overline{\mathbf{R}}$ reflecting a current correlation matrix $\overline{\mathbf{C}}$. 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 $\overline{\mathbf{A}}$ respectively, given by a Choleski decomposition (which also implies that they are easy to obtain). We can now write **R** and $\overline{\mathbf{R}}$ as matrix products of the relevant Choleski matrices and an uncorrelated noise process ε:

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

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

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 $\overline{\mathbf{R}}$ that we are seeking:

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

The returns adjusted in this way will then have the currently prevailing correlation matrix **C** and, more generally, the currently prevailing covariance matrix $\overline{\Sigma}$.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.

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

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 $\overline{\mathbf{A}}$ 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

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

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

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

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

[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 θ, 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:

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

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 $\hat{\theta}$. 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 $\hat{\theta}$. 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 $\hat{\boldsymbol{\theta}}$ into the bias equation:

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

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 $\hat{\boldsymbol{\theta}}$, $\hat{s}_B$, 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 $\hat{\boldsymbol{\theta}}^B = (1/B)\sum_{i=1}^{B}\hat{\theta}^B(i)$. $\hat{s}_B$ 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 $\hat{\boldsymbol{\theta}}^B$, 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.

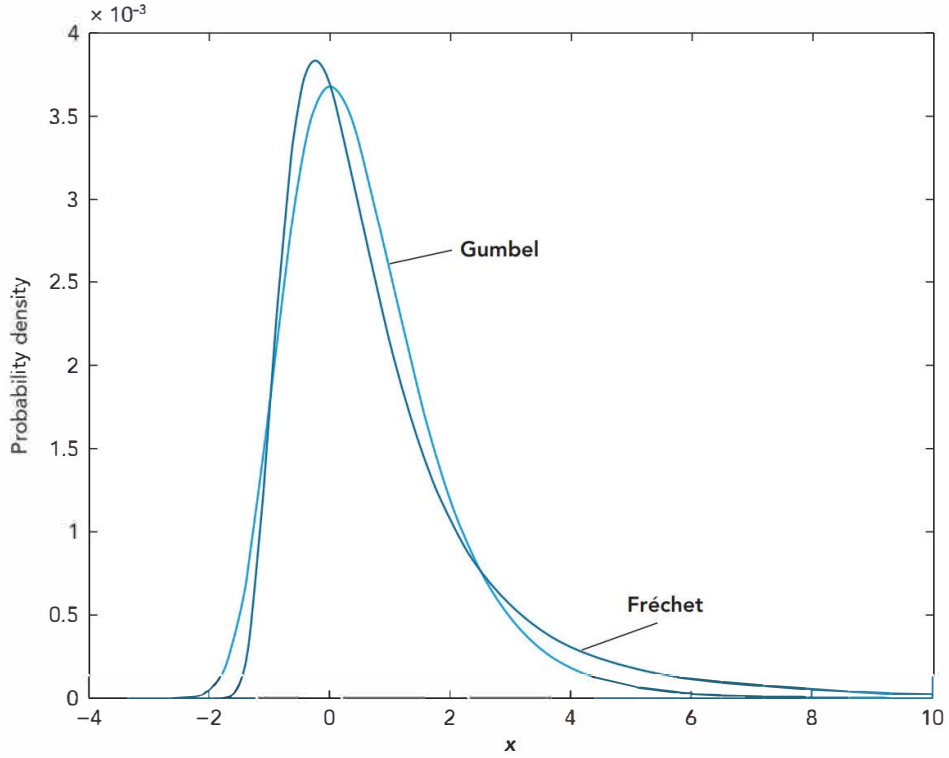


Figure 3.1 Standardised Gumbel and Fréchet probability density functions.

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:

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

which in turn implies:

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

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).

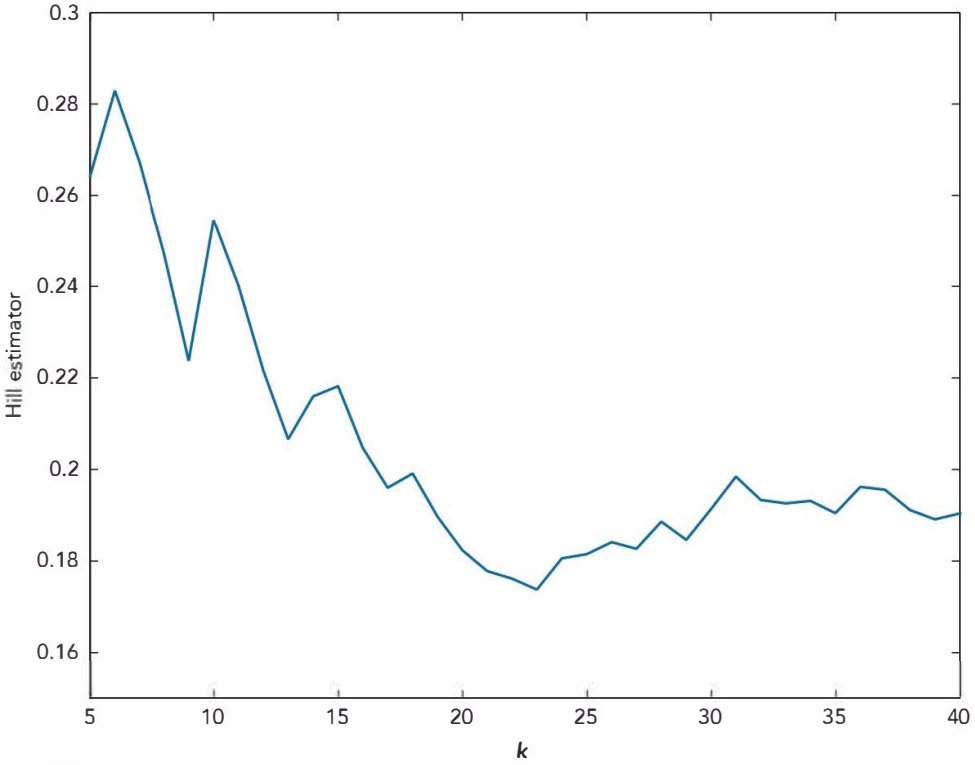


Figure 3.2 Hill plot.

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

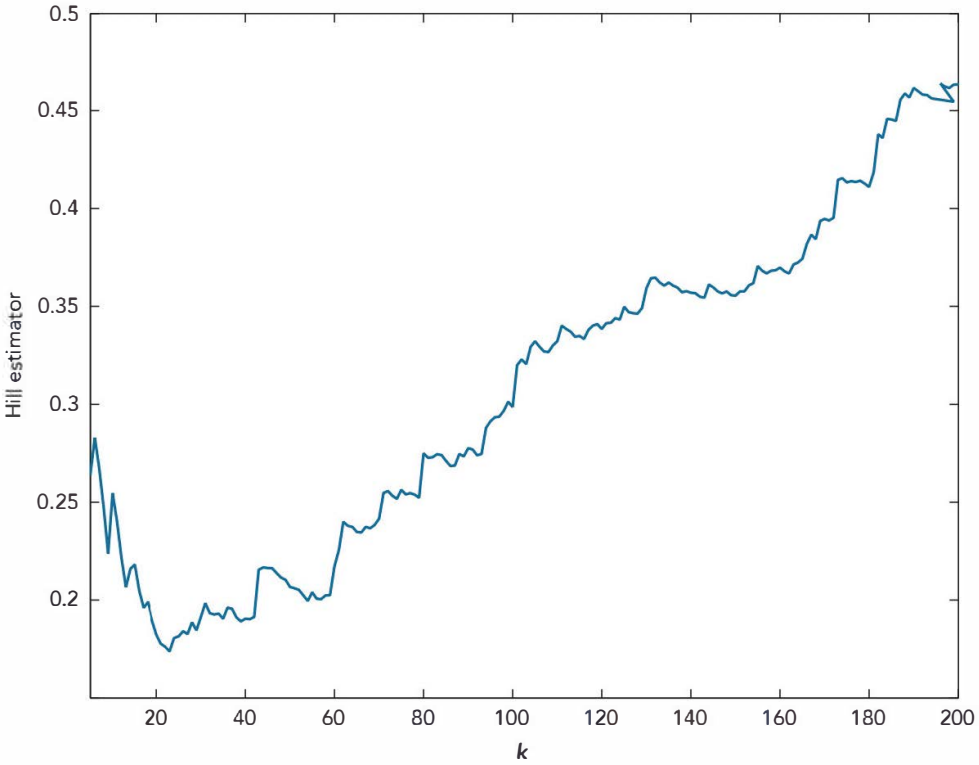


Figure 3.3 Hill horror plot.

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

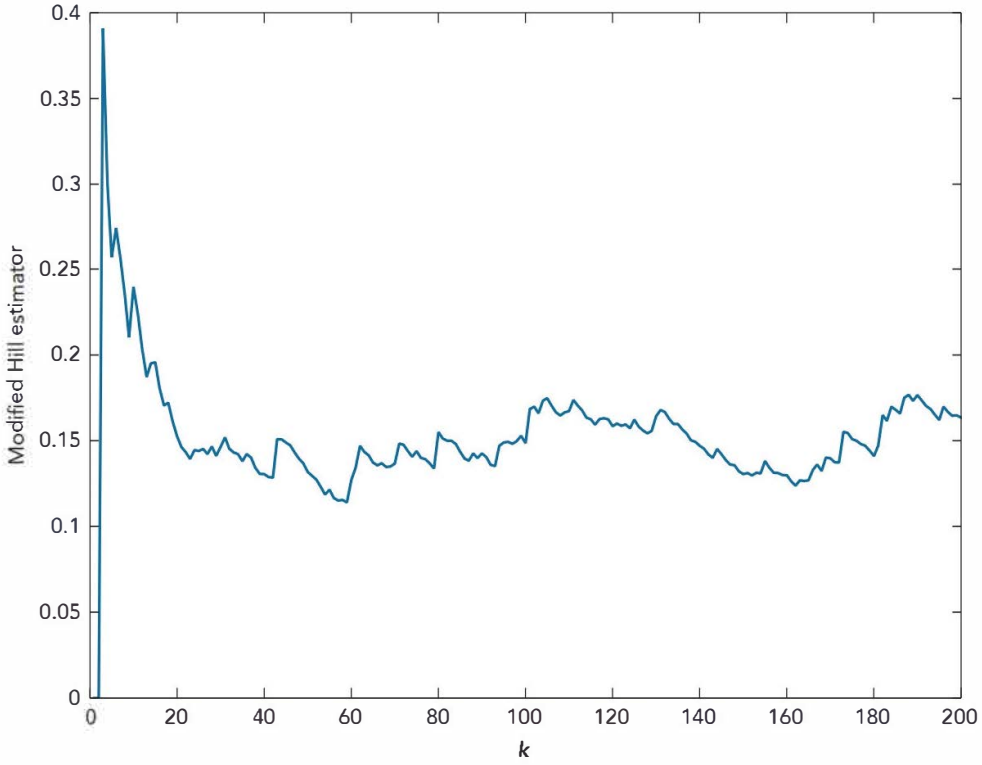


Figure 3.4 **'Hill happy plot.'**

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:

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

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.

Backtesting VaR

4

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.
- Verify a model based on exceptions or failure rates.
- 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.

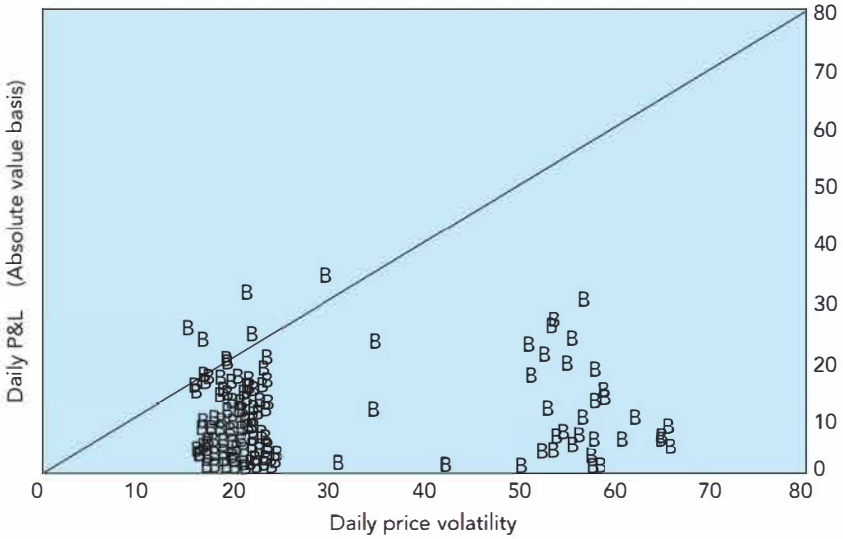


Figure 4.1 Model evaluation: Bankers trust.

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

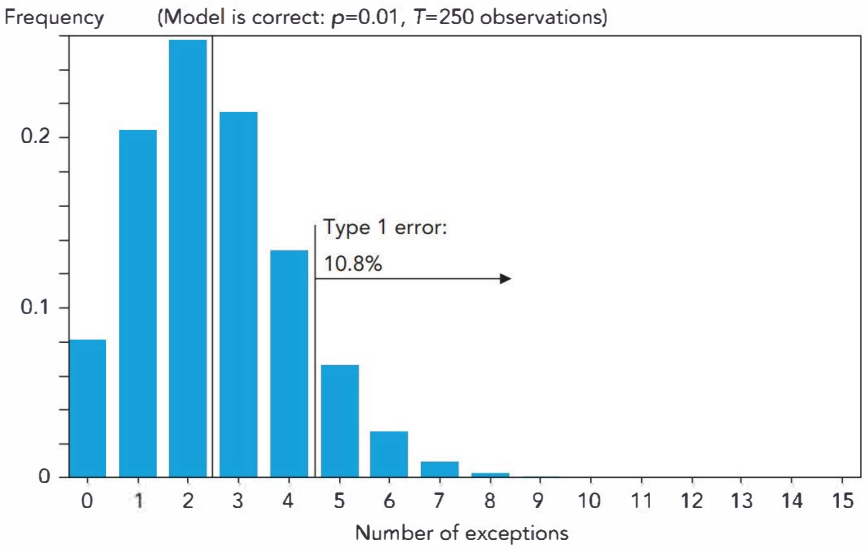


Figure 4.2 Distribution of exceptions when model is correct.

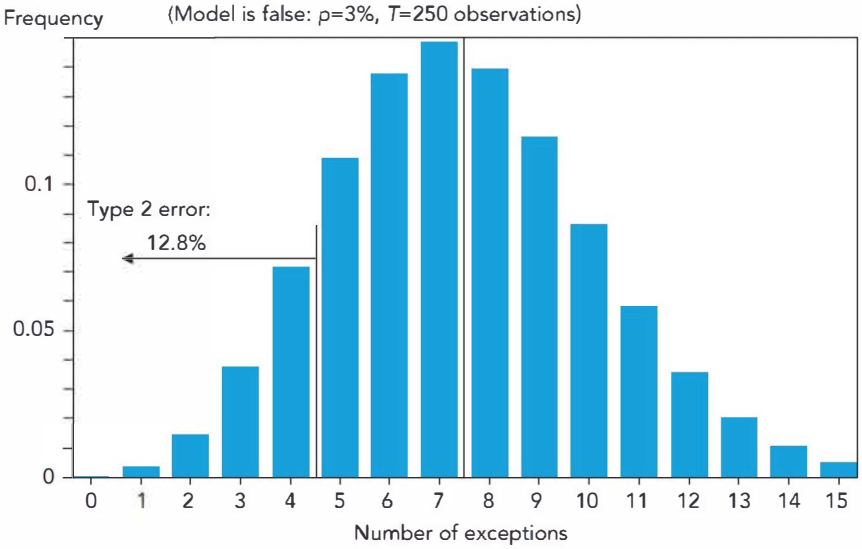


Figure 4.3 Distribution of exceptions when model is incorrect.

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

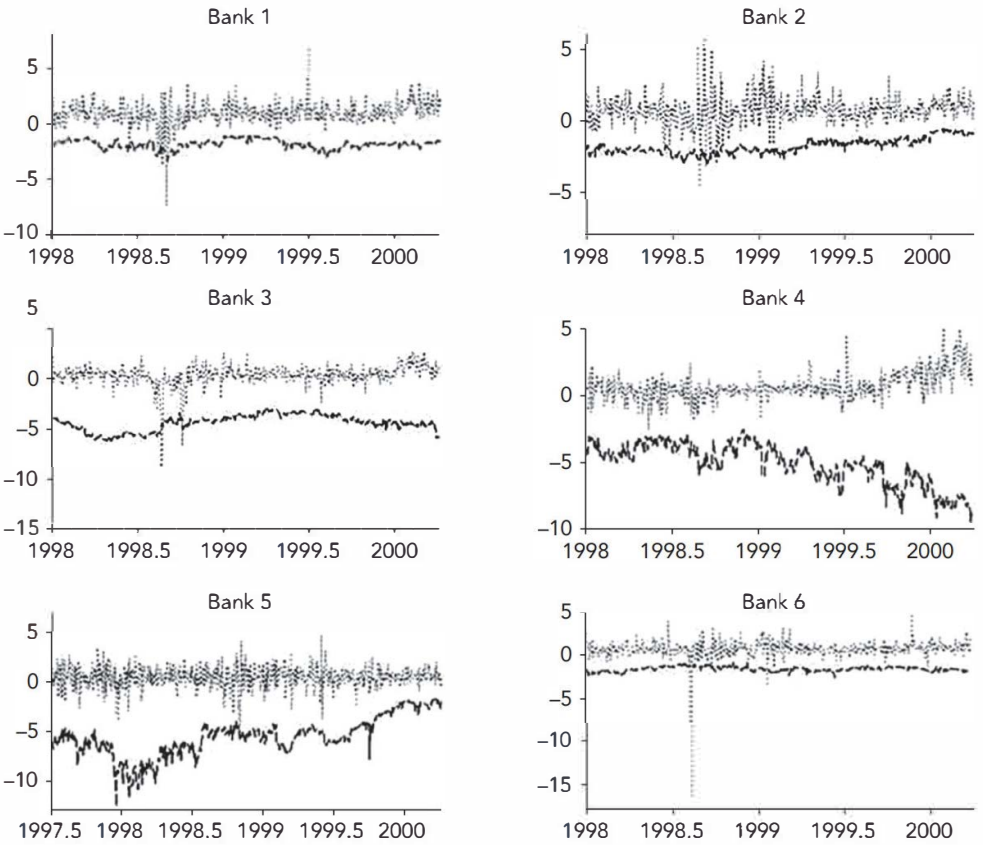


Figure 4.4 Bank VaR and trading profits.

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.

VaR Mapping

5

Learning Objectives

After completing this reading, you should be able to:

- Explain the principles underlying VaR mapping and describe the mapping process.
- Explain and demonstrate how the mapping process captures general and specific risks.
- 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 computed 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.

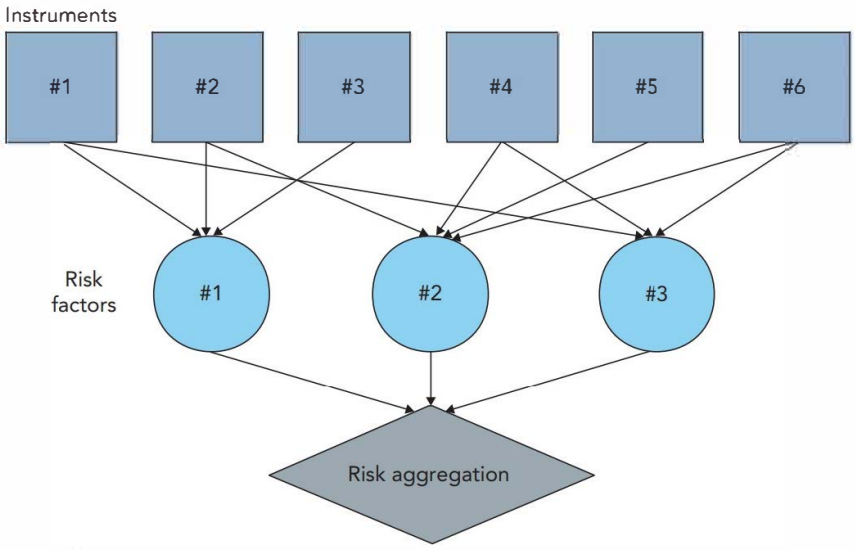


Figure 5.1 Mapping instruments on risk factors.

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 $x =$ USD 100/(1 + 5.625/200) = 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

$$c = c(S,K,\tau,r,r^*,\sigma) = Se^{-r^*\tau}N(d_1) - Ke^{-r\tau}N(d_2) \quad (5.16)$$

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,

$$\begin{aligned} dc &= \frac{\partial c}{\partial S}dS + \frac{1}{2}\frac{\partial^2 c}{\partial S^2}dS^2 + \frac{\partial c}{\partial r^*}dr^* + \frac{\partial c}{\partial r}dr + \frac{\partial c}{\partial \sigma}d\sigma + \frac{\partial c}{\partial t}dt \quad (5.17)\\ &= \Delta dS + \frac{1}{2}\Gamma dS^2 + \rho^* dr^* + \rho dr + \Lambda d\sigma + \bullet dt \end{aligned}$$

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

$$\Delta = e^{-r^*\tau}N(d_1) \quad (5.18)$$

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,

$$\begin{aligned} dc &= \Delta dS + \rho^* dr^* + \rho dr\\ &= [e^{-r^*\tau}N(d_1)]dS + [-Se^{-r^*\tau}\tau N(d_1)]dr^* + [Ke^{-r\tau}\tau N(d_2)]dr\\ &= [Se^{-r^*\tau}N(d_1)]\frac{dS}{S} + [Se^{-r^*\tau}N(d_1)]\frac{dP^*}{P^*} - [Ke^{-r\tau}N(d_2)]\frac{dP}{P}\\ &= x_1\frac{dS}{S} + x_2\frac{dP^*}{P^*} + x_3\frac{dP}{P} \quad (5.19) \end{aligned}$$

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).

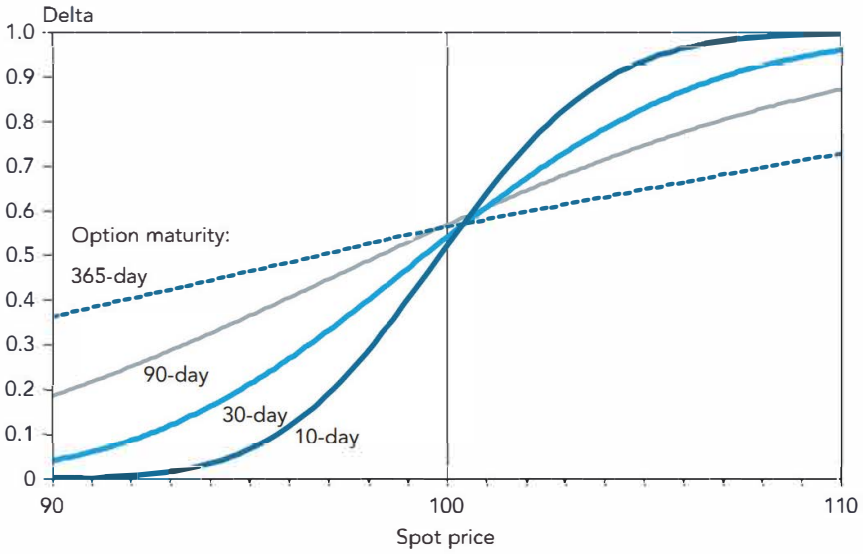


Figure 5.2 **Delta as a function of the risk factor.**

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 Messages from the Academic Literature on Risk Management for the Trading Book

Learning Objectives

After completing this reading, you should be able to:

- Explain the following lessons on VaR implementation: time horizon over which VaR is estimated, the recognition of time-varying volatility in VaR risk factors, and VaR backtesting.
- Describe exogenous and endogenous liquidity risk and explain how they might be integrated into VaR models.
- Compare VaR, expected shortfall, and other relevant risk measures.
- Compare unified and compartmentalized risk measurement.
- Compare the results of research on top-down and bottom-up risk aggregation methods.
- Describe the relationship between leverage, market value of asset, and VaR within an active balance sheet management framework.

Excerpt is reprinted by permission from the Basel Committee on Banking Supervision.

6.1 INTRODUCTION

This report summarises the findings of a working group (the "group") that surveyed the academic literature that is relevant to a fundamental review of the regulatory framework of the trading book. This joint working group embraced members of the Trading Book Group and of the Research Task Force of the Basel Committee on Banking Supervision. This report summarises its main findings. It reflects the views of individual contributing authors, and should not be construed as representing specific recommendations or guidance by the Basel Committee for national supervisors or financial institutions.

The report builds on and extends previous work by the Research Task Force on the interaction of market and credit risk (see Basel Committee on Banking Supervision (2009a)). The literature review was complemented by feedback from academic experts at a workshop hosted by the Deutsche Bundesbank in April 2010 and reflects the state of the literature at this point in time.

The key findings of the group are presented in the executive summary. The structure of the remaining report is as follows:

We address fundamental issues of a sometimes highly technical nature in current VaR-based approaches to risk measurement. More specifically, we give an overview of implementation issues including questions on the necessity of including time-variation in volatility, the appropriate time horizon over which risk is measured and backtesting of VaR. Capturing market liquidity in a VaR framework is the key question addressed in the second section. Then, we look at the pros and cons of VaR as a metric for risk and consider alternative metrics put forward in the literature. Important aspects for the future evolution of stress tests are addressed next.

The last two sections include management aspects, such as inter-risk aggregation and the borderline between the banking and trading books (which is discussed only briefly). They also expand the scope of this review by including macro-prudential aspects, such as systemic risk and pro-cyclicality. This section is concerned with an integrated versus a compartmentalised approach to risk measurement, which has become particularly important since the recent financial crisis revealed that a focus on market risk alone may provide distorted results for a trading book. This topic draws heavily on the findings of the former working group of the Research Task Force on the interaction of market and credit risk (see Basel Committee on Banking Supervision (2009a)). The last section looks at the relations between and among risk measurement, systemic risk, and potential pro-cyclical effects of risk measurement.

6.2 SELECTED LESSONS ON VaR IMPLEMENTATION

Overview

In this section we review the academic and industry literature on VaR implementation issues, as it pertains to regulatory capital calculation. The three categories of implementation issues reviewed are: (1) time horizon over which VaR is estimated; (2) the recognition of time-varying volatility in VaR risk factors; and (3) VaR backtesting. With respect to (1), we find that the appropriate VaR horizon varies across positions and depends on the position's nature and liquidity. For regulatory capital purposes, the horizon should be long, and yet the common square-root of time scaling approach for short horizon VaR (e.g., one-day VaR) may generate biased long horizon VaR (e.g., ten-day VaR) estimates. Regarding (2), we find that while many trading book risk factors exhibit time-varying volatility, there are some concerns that regulatory VaR may suffer from instability and pro-cyclicality if VaR models incorporate time-varying volatility. We also sketch several approaches to incorporate time-varying volatility in VaR. As for (3), we survey the literature on VaR backtesting and discuss several regulatory issues including whether VaR should be backtested using actual or hypothetical P&L, and whether the banks' common practice of backtesting one-day VaR provides sufficient support for their ten-day, regulatory VaR.

It is worthwhile to note that some issues related to time horizons and time-varying volatility, and to a lesser extent backtesting, also pertain to risk measures other than VaR, such as Expected Shortfall (ES). A discussion of these alternative risk measures is contained in this chapter.

Time Horizon for Regulatory VaR

One of the fundamental issues in using VaR for regulatory capital is the horizon over which VaR is calculated. The 1998 Market Risk Amendment (MRA) sets this horizon to be ten days, and it allows ten-day VaR to be estimated using square-root of time scaling of one-day VaR. This approach raises three questions: (1) Is ten days an appropriate horizon? (2) Does VaR estimation based on time scaling of daily VaRs produce accurate risk measures? (3) What role do intra-horizon risks (i.e., P&L fluctuations within ten days) play, and should such risks be taken into account in the capital framework?

Is Ten Days an Appropriate Horizon?

There seems to be consensus among academics and the industry that the appropriate horizon for VaR should depend on the characteristics of the position. In the academic literature,

Christoffersen and Diebold (2000) and Christoffersen, Diebold and Schuermann (1998) both assert that the relevant horizon will likely depend on where the portfolio lies in the firm (e.g., trading desk vs. CFO) and asset class (e.g., equity vs. fixed income), and the appropriate horizon should be assessed on an application-by-application basis. From this perspective, it appears that an across-the-board application of ten-day VaR horizon is not optimal. Indeed, one of the motivations for the Incremental Risk Charge (IRC) is to capture certain risks of credit related products at a longer horizon than ten days.

Although the literature suggests that it may be preferable to allow the risk horizon to vary across positions, Finger (2009), for instance, points out that there is no conceptual or statistical framework to justify the aggregation of a ten-day VaR and a one-year IRC. Danielsson (2002) adds that, if the purpose of VaR is to protect against losses during a liquidity crisis, the ten-day horizon at 99% refers to an event that happens roughly 25 times a decade, while a liquidity crisis is "unlikely to happen even once a decade. Hence the probability and problem are mismatched." In addition, even for the same financial product, the appropriate horizon may not be constant, because trade execution strategies depends on time-varying parameters, like transaction costs, expected price volatility, and risk aversion (Almgren and Chriss (2001), Engle and Ferstenberg (2006), Huberman and Stanzl (2005)). In addition, variation in risk aversion over the business cycle can be especially important in shortening the optimal trading horizon, potentially generating larger losses than those observable under more favourable conditions.

Danielsson (2002) questions the suitability of a ten-day horizon if VaR is to protect against a liquidity crisis, because a ten-day horizon implies a higher frequency of liquidity crisis than is observable in the data. Other authors have similarly suggested that the appropriate VaR horizon should depend on the economic purpose of VaR.[1] Smithson and Minton (1996), for instance, claim that nearly all risk managers believe a one-day horizon is valid for trading purposes but disagree on the appropriate horizon for long-term solvency or capital. Finger (2009) notes that there is "a tension between the regulatory risk horizon and the horizon at which banks manage their trading portfolios," and that the Market Risk Amendment (MRA) rules represent a compromise between regulatory and trading horizons through the introduction of the sixty-day moving average and backtesting multiplier mechanisms.

[1] For example, if VaR is expected to reduce the probability of bankruptcy, the horizon would line up with the time a bank needs to raise additional capital. If the focus is on losses while a position is being offloaded, the appropriate horizon would be more strictly related to asset characteristics.

The computation of VaR over longer horizons introduces the issue of how to account for time variation in the composition of the portfolios, especially for institutions that make markets for actively traded assets like currencies (Diebold, Hickman, Inoue and Schuermann (1998)). A common solution is to sidestep the problem of changes to portfolio composition by calculating VaR at short horizons and scaling up the results to the desired time period using the square-root of time. While simple to implement, this choice may compromise the accuracy of VaR because, as discussed in the next section, tail risk is likely to be underestimated (Bakshi and Panayotov (2010)). A second way to tackle the problem is to focus directly on calculating the portfolio VaR over the relevant horizon of interest (Hallerbach (2003)). These approaches may have limited value if the composition of the portfolio changes rapidly. Furthermore, data limitations make it challenging to study the P&L of newly traded assets. A third solution is to extend VaR models by incorporating a prediction of future trading activity, as noted by Diebold et. al. (1998): "To understand the risk over a longer horizon, we need not only robust statistical models for the underlying market price volatility, but also robust behavioural models for changes in trading positions."

Christoffersen and Diebold (2000) aptly characterised the issue of the optimal VaR horizon as "an obvious question with no obvious answer." Voices from the industry have suggested that a horizon longer than ten days may be necessary for regulatory capital purposes. It was also suggested that combining the liquidity horizon of individual positions with a constant level of risk may be an appropriate avenue.

Is Square-Root of Time Scaling a Good Idea?

Under a set of restrictive assumptions[2] on risk factors, long horizon VaR can be calculated as short horizon VaR scaled by the square root of time, if the object of interest is unconditional VaR (Kaufman (2004), McNeil, Frey and Embrechts (2005) and Danielsson and Zigrand (2006)). Unfortunately, the assumptions that justify square root of time scaling are rarely verified for financial risk factors, especially at high frequencies. Furthermore, risk management and capital computation are more often interested in assessing potential losses *conditional* on current information, and scaling today's VaR by the square root of time ignores time variation in the distribution of losses. We have not found any evidence in support of square-root of time scaling for conditional VaRs.

The accuracy of square-root of time scaling depends on the statistical properties of the data generating process of the risk factors. Diebold et. al. (1998) show that, if risk factors follow a

[2] Specifically, the risk factors have to be normally distributed with zero mean, and be independently and identically distributed ("IID") across time.

GARCH(1,1) process, scaling by the square-root of time *over-estimates* long horizon volatility and consequently VaR is over-estimated. Similar conclusions are drawn by Provizionatou, Markose and Menkens (2005). In contrast to the results that assume that risk factors exhibit time-varying volatility, Danielsson and Zigrand (2006) find that, when the underlying risk factor follows a jump diffusion process, scaling by the square root of time systematically *under-estimates* risk and the downward bias tends to increase with the time horizon. While these results argue against square-root of time scaling, it is important to acknowledge that we were not able to find immediate alternatives to square-root of time scaling in the literature. Therefore, the practical usefulness of square-root of time scaling should be recognised.[3]

Is Intra-Horizon Risk Important?

Bakshi and Panayotov (2010) discuss intra-horizon VaR (VaR-I), a risk measure that combines VaR over the regulatory horizon with P&L fluctuations over the short term, with a particular focus on models that incorporate jumps in the price process. The rationale behind intra-horizon VaR is that the maximum cumulative loss, as distinct from the end-of-period P&L, exerts a distinct effect on the capital of a financial institution. Bakshi and Panayotov (2010) suggest that VaR-I "can be important when traders operate under mark-to-market constraints and, hence, sudden losses may trigger margin calls and otherwise adversely affect the trading positions." Daily VaR does carry information on high frequency P&L but, as noted by Kritzman and Rich (2002), "Knowledge of the VaR on a daily basis does not reveal the extent to which losses may accumulate over time." Bakshi and Panayotov (2010) find that taking intra-horizon risk into account generates risk measures consistently higher than standard VaR, up to multiples of VaR, and the divergence is larger for derivative exposures.

Time-Varying Volatility in VaR

It is a stylised fact that certain asset classes, such as equities and interest rates, exhibit time-varying volatility. Accounting for time-varying volatility in VaR models has been one of the most actively studied VaR implementation issues. This section explores this topic, focusing on large and complex trading portfolios.

[3] A concept related to square-root of time scaling is the scaling of VaR to higher confidence levels. Although we were unable to find literature on this topic, we recognize that this is an important issue particularly in situations when there are inadequate data points for one to accurately estimate risks deep into the tail. Some banks use certain reference densities (e.g., Student's t with six degrees of freedom) to conduct such scaling.

Is It Necessary to Incorporate Time-Varying Volatilities and Correlations?

The industry seems to think so since many firms advocate the use of fast reacting measures of risk such as exponential time-weighted measures of volatility. The reason given is that such VaR models provide early warnings of changing market conditions and may perform better in backtesting. The academic literature has also observed that time-varying volatility in financial risk factors is important to VaR, dating back to the 1996 RiskMetrics Technical document (J.P. Morgan (1996)). Pritsker (2006) showed theoretically that using historical simulation VaR without incorporating time-varying volatility can dangerously under-estimate risk, when the true underlying risk factors exhibit time-varying volatility.

In contrast, some have argued that, depending on the purpose of VaR, capturing time-varying volatility in VaR may not be necessary, or may even be inappropriate. Christoffersen and Diebold (2000) observe that volatility forecastability decays quickly with time horizon for most equity, fixed income and foreign exchange assets. The implication is that capturing time-varying volatility may not be as important when the VaR horizon is long, compared to when the VaR horizon is relatively short. There are also concerns about pro-cyclicality and instability implications associated with regulatory VaRs that capture time-varying volatility. Dunn (2009), for instance, states that there is a "contradiction between the requirement for a risk sensitive metric to capture variations in volatility and correlation, and the regulatory requirement for a stable and forward looking basis for computing capital, that is not pro-cyclical." In reference to modelling time-varying volatility in VaR, it wrote, "Some firms mentioned a trade-off in this issue, and that for some purposes such as capital allocation, risk measures with more stable properties that reflected longer historical norms were desirable."

In summary, incorporating time-varying volatility in VaR appears to be necessary given that it is prevalent in many financial risk factors. Furthermore, many financial instruments are now priced with models with stochastic volatility features. It is logical that VaR models are constructed to account for these statistical properties. However, using VaR with time-varying volatility for regulatory capital raises the concerns of volatile and potentially pro-cyclical regulatory standards.

Methods to Incorporate Time-Varying Volatility in VaR for Large, Complex Portfolios

Beginning with J.P. Morgan (1996), the Exponentially Weighted Moving Average (EWMA) approach has been regarded as one of the industry standards for incorporating time-varying volatility in VaR. EWMA is a constrained version of an IGARCH (1,1)

model, and in the case of RiskMetrics the parameter in IGARCH was set to 0.97. An alternative and simpler approach is to weight historical data according to the weights introduced by Boudoukh, Richardson and Whitelaw (1998), where an observation from i days ago receives a weight of

$$w(i) = \frac{\theta^i(1-\theta)}{1-\theta^n}$$

Here n is the total number of days in the historical window, and θ is a number between zero and one which controls the rate of memory decay. An even simpler approach is to compute VaR with historical simulation using a short and frequently updated time series. Dunn (2009) has suggested that this method captures time-varying volatility quite well. Using simulations, Pritsker (2006) has shown that the approach of Boudoukh et. al. (1998) is not sensitive enough to pick up volatility changes. He advocated the use of Filtered Historical Simulation (FHS), first introduced by Barone-Adesi, Giannopoulos and Vosper (1999). Broadly speaking, FHS is based on the idea that risk factors should first be filtered through a GARCH model. The volatility is then updated using the model, and adhered to the filtered risk factors to constructed VaR.

Naturally, considerations should be given to how the above method can be applied to portfolios with large numbers of positions or risk factors. Barone-Adesi et. al. (1999) outlined a position-by-position FHS approach. They recommended filtering each risk factor separately, and building volatility forecasts for each factor. Analogously, EWMA and the weights introduced by Boudoukh et. al. (1998) can be applied the same way. However, weighting or filtering risk factors separately implicitly assumes that the correlation structure across risk factors does not change over time. Pritsker (2006) has pointed out that time-varying correlation is an important source of risk. Indeed, the recent crisis has highlighted the fact that correlations among many risk factors change significantly over time. One would need to be careful in handling time-varying volatilities as well as correlations.

Multivariate GARCH models such as the BEKK model of Engle and Kroner (1995), or the DCC model of Engle (2002) can be used to estimate time-varying volatilities as well as correlations. However, such multivariate GARCH models are difficult to estimate when there are a large number of risk factors. Some recent advances in the literature allow one to estimate a multivariate GARCH-type model when there are a large number of risk factors. For instance, Engle, Shephard and Sheppard (2007) proposed to average likelihoods before estimating the GARCH model with maximum likelihood. Engle and Kelly (2009) imposes a restriction on the correlation structure that helps facilitate estimation in large dimensions, but still allow correlations to change over time. Finally, Aramonte, Rodriguez and Wu (2010) estimates VaR for large portfolios comprising stocks and bonds by first reducing the dimension of risk factors using dynamic factor models, and then estimating a time-varying volatility model. The resulting VaR estimates are shown to out-perform historical simulation and FHS based on filtering risk factors one-by-one.

All in all, incorporating time-varying volatility in VaR measures is not straight forward when there are many risk factors. Time-varying correlations should be taken into account. Rather than using more involved methods, the industry appears to be taking less burdensome alternatives, such as using simple weighting of observations, or shortening the data window used to estimate VaR. These approaches compromise on accuracy, but are computationally attractive for large and complex portfolios. The recent academic literature offers promise that some of the sophisticated empirical methodologies may soon become practical for large complex portfolios.

Backtesting VaR Models

As with any type of modelling, a VaR model must be validated. In particular, backtesting has been the industry standard for validating VaR models. This section reviews some backtesting methodologies suggested by the literature, and some issues pertaining to the application of such methodologies.

Backtesting Approaches

Banks typically draw inference on the performance of VaR models using backtesting exceptions (sometimes also known as backtesting "breaches" or "violations"). For regulatory capital, the MRA imposes a multiplier on VaR depending on the number of backtesting exceptions the bank experiences.

While the MRA does not require banks to statistically test whether VaR has the correct number of exceptions, formal statistical inference is always desirable and many alternatives have been proposed in the literature. Kupiec (1995) introduced the unconditional coverage likelihood ratio tests as inference tools for whether the VaR model generated the correct number of exceptions. This methodology is simple to implement, but has two drawbacks. First, as pointed out by Kupiec (1995 and 2005), when the number of trading days used in VaR evaluation is limited (e.g., one year or approximately 250 trading days), or when the confidence level is high (e.g., 99% as in regulatory VaR), such tests have low power. This is not surprising, since one would expect only a small number of backtesting exceptions in most cases. Building a statistic out of a handful of exceptions, then, may induce high variance in the test statistic itself and the result may be sensitive to an incremental exception. Second, given that this test only counts exceptions, its power may be improved

by considering other aspects of the data such as the grouping of exceptions in time.

Christoffersen (1998) has proposed a conditional backtesting exception test that accounts for the timing as well as the number of exceptions. The test is based on the fact that when the VaR model has conditionally the correct number of exceptions, then indicator variables representing the exceptions are IID[4] Bernoulli random variables. This test, however, may still be exposed to the low power problem. To this end, Berkowitz, Christoffersen and Pelletier (2010) provided a suite of conditional tests that have good power. These tests are based on the intuition of Christoffersen (1998) (i.e., correct conditional exceptions results in IID and Bernoulli exception indicators) but derive inferences from autocorrelation, spectral, and hazard rate tests.

Aside from backtesting based on the number of exceptions, a natural measure of VaR performance is the magnitude of the exceptions. Lopez (1999), for instance, formalised this idea by introducing a quadratic loss function where loss is the difference between actual P&L and VaR, when an exception occurs. Some papers, including Pritsker (2006) and Shang (2009), also consider the use of Mean-Squared-Error (MSE) as a measure of VaR performance in backtesting. Typically, one would measure the MSE between the 'true VaR' and the VaR estimate based on the model. Clearly, this method is not directly applicable to observed portfolio P&Ls, since the true VaR is never known. Nonetheless, it can be a useful validation method prior to putting a VaR model into production: one can define data generating processes mimicking those imposed by front office pricing models, simulate position P&L enough times to construct a P&L distribution, and find the 'true VaR' based on this simulated distribution. Then, the VaR model can be applied to the generated data, and the difference between 'true VaR' and estimated VaR can be analysed.

Backtesting Issues

An important and yet ambiguous issue for backtesting is which P&L series to compare to VaR. Broadly speaking, the estimated VaR can be compared to either actual P&L (i.e., the actual portfolio P&L at the VaR horizon), or hypothetical P&L (i.e., P&L constructed based on the portfolio for which VaR was estimated). To complicate matters further, actual P&L may sometimes contain commissions and fees, which are not directly related to trading and trading risk. Franke, Härdle and Hafner (2008) and Berry (2009) described the relative merits of actual and hypothetical backtesting: actual backtesting has little value if the portfolio has changed drastically since VaR was estimated, but is simple to implement; hypothetical backtesting would make an 'apples-to-apples' comparison, but it comes with significant implementation burden given that hypothetical portfolios need to be constructed.

Another issue is the appropriate backtesting horizon. Banks typically backtest one-day ahead VaR and use it as a validation of the regulatory VaR, which is ten-day. The problem here is clear: a good one-day VaR (as validated by backtesting) does not necessarily imply a good ten-day VaR, and vice versa. Ten-day backtesting may not be ideal either, given the potentially large portfolio shifts that may take place within ten days. In that case, actual P&L backtesting in particular may not be very informative. While we were unable to find literature on this particular issue, it remains an important policy question.

Conclusions

We have reviewed the literature on a number of VaR implementation issues, including the appropriate time horizon, time-variation in the volatility of risk factors, and backtesting. We find that the optimal way of addressing these points is idiosyncratic to the problem under consideration. For instance, when estimating long horizon VaR by scaling the short horizon counterpart by the square root of time, one may overestimate VaR if the underlying P&L process exhibits time-varying volatility, but underestimate VaR if the process has jumps.

Incorporating time-varying volatility in VaR measures appears to be important to make models more realistic although it is not straight forward when there are many risk factors. The recent academic literature offers promise in this direction. While many trading book risk factors have time-varying volatility, models that incorporate this feature may, however, generate pro-cyclical VaR and also be unstable, not least because of estimation issues.

In addition, the choice of whether to evaluate a VaR model on the basis of hypothetical or actual backtesting may be affected by the characteristics of the portfolio. Indeed, actual backtesting is less informative when the composition of the portfolio has recently changed. On the other hand, while hypothetical backtesting provides a more consistent comparison, it may impose substantial computational burdens because it requires reconstructing the history of the portfolio on the basis of its current composition.

6.3 INCORPORATING LIQUIDITY

Overview

Discussing the challenging issue of how to incorporate market liquidity into a VaR model requires first of all a distinction between *exogenous* and *endogenous* liquidity. This distinction

[4] IID: independently and identically distributed.

is made from the point of view of the bank, rather than in general equilibrium terms (Bangia, Diebold, Schuermann and Stroughair (1999a) and Bervas (2006)), More specifically, *exogenous* liquidity refers to the transaction cost for trades of average size, while *endogenous* liquidity is related to the cost of unwinding portfolios large enough that the bid-ask spread cannot be taken as given, but is affected by the trades themselves.

Bangia et. al. (1999a) give a graphical representation of *exogenous* and *endogenous* liquidity that is reproduced in Figure 6.1. Below a certain size, transactions may be traded at the bid/ask price quoted in the market (*exogenous* liquidity), and above this size, the transaction will be done at a price below the initial bid or above the initial ask, depending on the sign of the trade (*endogenous* liquidity).

The exogenous component of liquidity risk corresponds to the average transaction costs set by the market for standard transaction sizes. The endogenous component corresponds to the impact on prices of the liquidation of a position in a relatively tight market, or more generally when all market participants react in the same way, and therefore applies to orders that are large enough to move market prices (Bangia et. al. (1999a), Bervas (2006)). Exogenous liquidity risk, corresponding to the normal variation of bid/ask spreads across instruments can be, from a theoretical point of view, easily integrated into a VaR framework. Endogenous risk, corresponding to the impact on market prices of the liquidation of a position, or of collective portfolio adjustments, is more difficult to include in a VaR computation. Its impact, however, may be very significant, especially for many complex derivatives held in trading books of large institutions.

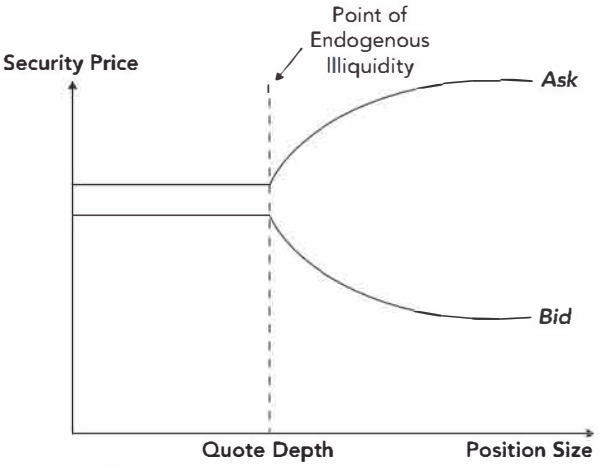


Figure 6.1 **Effect of position size on liquidation value.**

Source: Bangia et. al. (1999a).

One way to incorporate liquidity risk into VaR measures is to include new VaR risk factors that can be used to model liquidity risks. This approach is feasible only when the parameters can be deduced from market data. Liquidity reserves taken by banks on their trading portfolio according to accounting standards correspond, more or less, to reserves for exogenous liquidity. In order to integrate this risk in the VaR computation, Bangia et. al. (1999a) propose to integrate the variability of the bid/offer spread for average size transactions as a risk factor.

To take into account endogenous liquidity in the value-at-risk is more difficult, as it is not even really taken into account in the valuation of trading portfolios, but its impact on both valuation and VaR should be significant. Academic literature on the subject—portfolio valuation and VaR computation—is quite rich, but very little application has been made in particular because endogenous liquidity reserves could be considered as not compliant to accounting standards.

In the following section, we first describe how, following existing literature, exogenous liquidity might be integrated into VaR measures. We then review several aspects of endogenous liquidity risk, and detail how this risk could be integrated in portfolio valuation and VaR computation. At last, we discuss on the choice of the VaR horizon when taking into account liquidity risk.

Exogenous Liquidity

For the trading portfolio, following IAS rules, only exogenous liquidity risk will be taken into account in the valuation of cash assets and derivatives. Bangia et. al. (1999a) propose adding the bid/offer spread to characterise exogenous liquidity as a risk factor.

Their method poses that the relative spread, S = (Ask-Bid)/Mid-price, has sample mean and variance $\hat{\mu}$ and $\hat{\sigma}^2$. If the 99% quantile of the normalised distribution of S is $\hat{q}_{0.99}$, then the Cost of Liquidity is defined as

$$CoL_t = P_t\left(\frac{\hat{\mu} + \hat{q}_{0.99}\hat{\sigma}}{2}\right)$$

where P_t is today's value of the position. CoL_t is added to VaR to form a liquidity-adjusted VaR.

Endogenous Liquidity: Motivation

Adverse market conditions can generate a flight to liquid and high-quality assets, which reduces the ability to unwind positions in thinly-traded, low-quality assets. The effect can be compounded when the inventory of market makers becomes

imbalanced, thus reducing their willingness to further accommodate sell trades, and when risk management standards for traders become tighter, reducing the probability of finding a counterparty.

Margin requirements are also a source of variation in the response of assets' prices and liquidity to fundamental shocks, because higher margins increase the probability of binding funding constraints. While the choice of margin requirements is endogenous to a security's liquidity, assets with identical payoffs can have different prices depending on margin requirements and the opportunity cost of capital.

The trading activities associated with hedging may also have an impact on the dynamics of the underlying assets. For example, delta hedging an option position entails buying the asset when its price goes up, and selling it when the price goes down: if the size of these adjustments is not negligible with respect to the volumes traded on the underlying, this strategy will increase upward and downward price movements.

Such effects will be particularly important when:

- the underlying asset is not very liquid,
- the size of the positions of the investors hedging an option is important with respect to the market,
- large numbers of small investors follow the same hedging strategy,
- the market for the underlying of the derivative is subject to asymmetric information, which magnifies the sensitivity of prices to clusters of similar trades (Gennotte and Leland (1990)).

In particular, on some specific markets driven by exotic options (e.g., Power Reverse Dual Callable, some CPPI[5] strategies, etc.), even if a bank's trading book positions are small with respect to the market, this bank may be exposed to losses due to endogenous liquidity. When many other banks have the same kind of positions and none has an opposite position,[6] all these banks will have to adjust their hedging portfolio in the same way at the same time, and will then influence the market dynamics, and thus its small position may then be exposed to a significant liquidity cost.

The implications of derivative hedging have been extensively studied and derivative hedging has been identified as a potential explanation for the relation between implied volatilities and strike prices that can be observed on option markets (the so-called volatility smile). This literature includes the work of Platen and Schweizer (1998), Sircar and Papanicolaou (1998), Schönbucher and Wilmott (2000) and Subramanian (2008).

Endogenous Liquidity and Market Risk for Trading Portfolios

Several authors have studied the implications of endogenous liquidity risk for portfolio valuation and on value-at-risk measures (Jarrow and Protter (2005), Rogers and Singh (2005)). In general, these authors define an optimal liquidation strategy in a finite (or infinite time horizon) model and deduce from this strategy the market value of the portfolio which is equal to the expectation of its liquidation price. The associated VaR measure, defined as a confidence interval around this expected price, implicitly incorporates market and liquidity risks.

Some studies suggest that endogenous liquidity costs should be added to position returns before carrying out VaR calculations. To that end, Bervas (2006) suggests to incorporate Kyle's Lambda or Amihud's (2002) illiquidity ratio in returns. Both measures are based on the relationship between returns and volume. Wu (2009) applies the illiquidity cost of Amihud (2002) to stock returns and calculates the sum as "liquidity-adjusted returns." VaR is then estimated by applying a GARCH type model to the adjusted returns. Francois-Heude and Van Wynendaele (2001) suggest an approach that modifies the model of Bangia et. al. (1999a) by using average weighted bid-ask spreads, with weights based on volume. Berkowitz (2000b) proposes to incorporate price impact of an immediate liquidation via the concept of elasticity of demand. Jarrow and Subramanian (2001) modify the mean and variance that appears in the standard parametric VaR formula to incorporate means and variances of liquidation time and liquidity discount. Botha (2008) extended Jarrow and Subramanian (2001) to the two assets portfolio level. Other notable papers include Le Saout (2002) and Hisata and Yamai (2000). Finally, Acerbi and Scandolo (2008) explore the impact of market and funding liquidity on portfolio prices and risk measure. The authors revisit the coherent measures of risk criteria introduced by Artzner et. al. (1999). They explain how these criteria should be interpreted; in particular they study liquidity models that lead to solutions for the valuation of portfolios constituted of analytically tractable assets.

The liquidity risk adjustments proposed in the academic literature, for the most part, have not been applied to the trading books of banks. One reason for this may be that the suggested valuation methods are not necessarily compliant with actual

[5] CPPI: Constant Proportion Portfolio Insurance.

[6] The clients of these banks may be investors who do not dynamically hedge their positions.

accounting standards.[7] Another reason academic proposals have been slow to be adopted may be the difficulty of estimating model liquidity parameters, especially for OTC products. Indeed, the necessary data are not always available, and some of these parameters may be subjective. But recent discussions in academic circles regarding OTC transaction reporting could contribute to solve this problem.

A study of the impact of endogenous liquidity on the valuation of exotic derivatives, similar to the contributions of exogenous liquidity, would be especially welcome. When significant market movements materialise, traders will adjust their hedging strategies which may have an impact on the market dynamics if the volumes they have to trade are significant. Such effect has been suggested as a possible explanation for the significant trading losses that some banks have experienced during the last financial crisis.

Some authors have integrated liquidity risk with market and credit risk. For example, in order to evaluate a portfolio, Zheng (2006) studies optimal liquidation strategies, taking into account market and liquidity risk, together with the probability of default of an issuer or of a counterparty. Stange and Kaserer (2008) suggest calculating liquidity-adjusted VaR conditional on the market value of a position by incorporating bid-ask spread liquidity adjustments in returns; Qi and Ng (2009) discuss intraday liquidity risk and its impact on VaR.

Adjusting the VaR Time Horizon to Account for Liquidity Risk

The recent financial crisis has provided examples where a change in market liquidity conditions alters the *liquidity horizon*, i.e., the time required to unwind a position without unduly affecting the underlying instrument prices (including in a stressed market). This finding was already addressed in previous work of the Research Task Force (see Basel Committee on Banking Supervision (2009a)) and it is consistent with the literature.

Lawrence and Robinson (1997), for example, suggest that the application of a unique horizon to all positions by ignoring their size and level of liquidity is undesirable. They suggest determining the temporal horizon by the size of the position and the liquidity of the market. Haberle and Persson (2000) propose a method based on the fraction of daily volume that can be liquidated without significant impact on the market price, which can be interpreted as holding the horizon fixed and determining how much can be liquidated during that horizon. The method of Jarrow and Subramanian (2001) is also relevant in this context as it requires an estimate of the average liquidation time.

Previous work of the Research Task Force suggests an *interdependence* between risk assessment and liquidity horizon: On the one hand the exposures of banks to market risk and credit risk may vary with a risk horizon that is set dependent on market liquidity. If liquidity decreases, for example, the risk horizon lengthens and the exposure to credit risk typically increases. On the other hand, liquidity conditions are also affected by perceptions of market and credit risk. A higher estimate of credit risk for example, may adversely affect the willingness to trade and thereby market liquidity (see Basel Committee on Banking Supervision (2009a)).

Liquidation horizons vary over the business cycle, increasing during times of market stress. Besides transaction costs or the size of the position relative to the market, a trade execution strategy also depends on factors like expected price volatility and risk aversion (Huberman and Stanzl (2005)). If, for instance, risk aversion increases during a crisis, an investor may choose to trade more rapidly than during normal times, thus generating higher losses than those observable under favourable economic conditions.

Conclusions

Both exogenous and endogenous liquidity risks are important; endogenous liquidity risk is particularly relevant for exotic/complex trading positions. While exogenous liquidity is partially incorporated in the valuation of trading portfolios, endogenous liquidity is typically not, even though its impact may be substantial. Although endogenous liquidity risk is especially relevant under stress conditions, portfolios may be subject to significant endogenous liquidity costs under *all* market conditions, depending on their size or on the positions of other market participants.

The academic literature suggests as a first step to adjust valuation methods in order to take endogenous liquidity risk into account. Then a VaR integrating liquidity risk could be computed. Notwithstanding academic findings on this topic, in practice, the ability to model exogenous and endogenous liquidity may be constrained by limited data availability, especially for OTC instruments.

[7] For example, IAS 39 specify in AG72: "The appropriate quoted market price for an asset held or liability to be issued is usually the current bid price. . . . The fair value of a portfolio of financial instruments is the product of the number of units of the instrument and its quoted market price," and in AG75: "The objective of using a valuation technique is to establish what the transaction price would have been on the measurement date in an arm's length exchange motivated by normal business considerations."

6.4 RISK MEASURES

Overview

This section compares selected risk measures that appear to be relevant for risk management purposes either today or in the future. The alternative measures considered include VaR, expected shortfall and spectral measures of risk. The key features used to decide among alternative risk measurement approaches include ease of calculation, numerical stability, the possibility to calculate risk contributions of individual assets to portfolio risk, backtesting possibilities, incentives created for risk managers, and, linked to the latter, the relation between risk measures and regulators' objectives. Although few financial institutions currently make use of VaR alternatives, those that do are often considered as technologically leading in the industry.

In the literature, risk measures are usually defined as functions of *random variables* (portfolio losses or returns in most cases). This seems to be a trivial aspect but is actually a substantial restriction because it binds the analysis to one point of time; while this time horizon can be varied, a *joint* analysis of a portfolio's losses at several times, which may be important for asset/liabilities management, is excluded. Risk measures being a function of random loss variables also means that these variables are not an attribute of risk measures; the probability distributions of the variables are specified in a preceding step, and the analysis of risk measures is not an analysis of whether the random variables are correctly specified.

In our discussion of alternative measures we focus on VaR because of its high relevance to the industry today and on Expected Shortfall and Spectral Measures because of their advantages and hence a potentially growing importance in the future. Other risk measures, such as variance or upper-tail moments are briefly sketched for completeness.

VaR

Concept of VaR and Its Problems

VaR has become a standard measure used in financial risk management due to its conceptual simplicity, computational facility, and ready applicability. Given some random loss L and a confidence level α, $VaR_\alpha(L)$ is defined as the quantile of L at the probability α. The quantile is not necessarily unique if there are regions where the loss distribution function F_L does not grow. For these cases, McNeil et. al. (2005) define the VaR as the smallest, i.e., most optimistic quantile:

$$VaR_\alpha(L) = \inf\{l : F_L(l) \geq \alpha\}$$

Despite its prevalence in risk management and regulation, VaR has several conceptual problems. Artzner, Delbaen, Eber and Heath (1999) point out that VaR measures only quantiles of losses, and thus disregards any loss beyond the VaR level. As a consequence, a risk manager who strictly relies on VaR as the only risk measure may be tempted to avoid losses within the confidence level while increasing losses beyond the VaR level. This incentive sharply contrasts with the interests of regulators since losses beyond the VaR level are associated with cases where regulators or deposit insurers have to step in and bear some of the bank's losses. Hence, VaR provides the risk manager with incentives to neglect the severity of those losses that regulators are most interested in.

Neglecting the severity of losses in the tail of the distribution also has a positive flipside: it makes back-testing easier or possible in the first place simply because empirical quantiles are per se robust to extreme outliers, unlike typical estimators of the expected shortfall, e.g., (see below).

VaR is criticised for not being a *coherent* risk measure, which means that VaR lacks an axiomatic foundation as proposed by Artzner et. al. (1999). They set out the following four consistency rules. A risk measure R is called *coherent* if it satisfies the following axioms.

- Subadditivity (diversification) $R(L_1 + L_2) \leq R(L_1) + R(L_2)$
- Positive homogeneity (scaling) $R(\lambda L) = \lambda R(L)$, for every $\lambda > 0$
- Monotonicity $R(L_1) < R(L_2)$ if $L_1 < L_2$
- Transition property $R(L + a) < R(L) - a$

VaR is not coherent because it may violate the subadditivity criterion. For why subadditivity indeed makes sense we quote from McNeil et. al. (2005):

- "Subadditivity reflects the idea that risk can be reduced by diversification, . . . the use of non-subadditive risk measures in a Markowitz-type portfolio optimisation problem may lead to optimal portfolios that are very concentrated and that would be deemed quite risky by normal economic standards.
- If a regulator uses a non-subadditive risk measure in determining the regulatory capital for a financial institution, that institution has an incentive to legally break up into various subsidiaries in order to reduce its regulatory capital requirements . . . .
- Subadditivity makes decentralisation of risk-management systems possible. Consider as an example two trading desks with positions leading to losses L_1 and L_2. Imagine that a risk manager wants to ensure that $R(L)$, the risk of the overall loss $L = L_1 + L_2$, does not exceed some number M. If he uses a subadditive risk measure R, he may simply choose bounds M_1 and M_2 such that $M_1 + M_2 \leq M$ and impose on each of the desks the constraint that $R(L_i) \leq M_i$; subadditivity then ensures automatically that $R(L) \leq M_1 + M_2 \leq M$."

Remark 1: Related to non-coherency of VaR, Basak and Shapiro (2001) create an example where VaR-based risk management may possibly be problematic. They analyse optimal, dynamic portfolio and wealth/consumption policies of utility maximising investors who must also manage market-risk exposure using VaR. They find that VaR risk managers often optimally choose a larger exposure to risky assets than non-VaR risk managers and consequently incur larger losses when losses occur.

Remark 2: At first glance, subadditivity and positive homogeneity may not appear as meaningful concepts when risk measures are applied to counterparty credit risk (CCR) or other types of credit risk. For example, assume there is CCR involved with some position in the trading book. Doubling the position can *more than* double the CCR simply because not only the exposure doubles but also because the position becoming extremely profitable can *make* the counterparty go bankrupt. This appears to contradict the postulate of positive homogeneity which claims $R(2L) = 2R(L)$. However, it is not that positive homogeneity is wrong for CCR but rather that this idea reflects a misunderstanding of risk measures as functions of positions. Generally, the risk of the doubled position will not be $2L$ but rather a random variable with a wider probability distribution. Similar effects are possible for subadditivity; the issue is related to the Unified versus Compartmentalised Risk Measurement section on whether a compartmentalised measurement of risk is appropriate. For instance, there may exist two positions which cause individual risks L_1 and L_2, respectively, if held alone, but the risk of holding both positions may be more severe than $L_1 + L_2$ for similar reasons as in the above example. Knowing this, one might question the subadditivity property as such because it requires $R(L_1 + L_2) \leq R(L_1) + R(L_2)$. However, not subadditivity is to blame but a potential misunderstanding of $L_1 + L_2$ as the risk of holding both positions together. These considerations imply two lessons:

- It may be problematic to assume that a vector of assets linearly maps into the associated vector of random losses.
- If a "risk measure" is defined as a composed mapping *from positions* (via the loss variable) to numbers, this mapping is generally not coherent. Assuming coherence can lead to an underestimation of risk.

Is VaR Failing Subadditivity Relevant in Practice?

The favourite textbook example of VaR violating subadditivity is constructed with the help of two large losses the probability of which is lower than the confidence level of the VaR. When measured separately, each loss can have zero VaR but when aggregated, the probability that either of the losses occurs may exceed the confidence level so that the VaR of the aggregated loss is positive.

As the textbook example relies on jumps in the loss distribution one might conjecture that VaR works properly if loss distributions are smooth or if discrete losses are superimposed by sufficiently large smooth ones. Whether this intuition is correct ultimately depends on the situation and particularly on the tail thickness of the loss distributions:

- McNeil et. al. (2005) present an example of a *continuous* two-dimensional loss distribution in which VaR violates subadditivity. While this is alarming in that it does not build on the abovementioned textbook ingredients, the example is still rather artificial.
- If the joint distribution of risk factors is elliptical (multivariate normal, e.g.), VaR is subadditive; see McNeil et. al. (2005), Theorem 6.8.
- Gourier, Farkas, and Abbate (2009) give an example where the sum of some fat-tailed, continuously distributed, and independent (!) random variables has a larger VaR than the sum of individual VaRs. The example is rather exotic as one of the random variables has infinite mean. While VaR fails in that case, it must be conceded that there is no coherent *and* practicable alternative at all because any coherent risk measure must be infinite then.[8]
- Danielsson, Jorgensen, Samorodnitsky, Sarma, and de Vries (2005) prove that VaR is subadditive for a sufficiently high confidence level if the total loss has finite mean. Note, however, that this is not an "all clear" signal but an asymptotic result only. Generally it may happen that subadditivity is only achieved for impracticably high confidence levels.
- Degen, Embrechts, and Lambrigger (2007) restrict their analysis to a parametric class of distributions but gain valuable insight into the interplay of tail thickness, confidence level, and subadditivity. For example, they find the 99%-VaR to be superadditive even for very moderate tail indices above 6, which means that moments of order 6 and lower may exist.[9] These are realistic cases in market risk. The dependence structure between the individual losses generally aggravates the problem but has surprisingly low impact in the cases considered.

---

[8] Gourier et. al. (2009) refer to Delbaen (2002) who shows in Theorem 13 that, given a continuous distribution and some continuity of the risk measure, any coherent risk measure larger or equal than the α-VaR cannot fall short of the α-expected shortfall; the latter is already infinite in Gourier's example so that no useful coherent measure of that risk can exist.

[9] The higher the tail index, the thinner is the tail. Degen et. al. (2007) consider random variables the distribution tail of which is as thick as that of a transform $\exp(gZ + 0{,}5hZ^2)$ of a standard normal Z; e.g., $g = 2.3$ and $h = 0.25$ make the VaR super-additive; the tail index is 4 in this example.

To sum up, while the literature provides us with conditions that assure VaR is subadditive (and thus coherent), these conditions are generally *not fulfilled* in the market risk context; for example, Balaban, Ouenniche and Politou (2005) estimate tail indices between 1 and 2 for UK stock index returns over holding periods between 1 and 10 days, meaning that these tails are substantially heavier than necessary for assuring the subadditivity of VaR in general.

Expected Shortfall

Expected shortfall (ES) is the most well-known risk measure following VaR. It is conceptually intuitive and has firm theoretical backgrounds; see, e.g., Dunn (2009), Artzner et. al. (1999), Acerbi and Tasche (2002), Sy (2006), and Yamai and Yoshiba (2005). Therefore, it is now preferred to VaR by an increasing number of risk managers in the industry.

ES corrects three shortcomings of VaR. First, ES does account for the severity of losses beyond the confidence threshold. This property is especially important for regulators, who are, as discussed above, concerned about exactly these losses. Second, it is always subadditive and coherent. Third, it mitigates the impact that the particular choice of a single confidence level may have on risk management decisions, while there is seldom an objective reason for this choice.

To define ES, let L be a random loss with distribution function F_L and $\alpha \in (0,1)$ a confidence level (close to 1). Recall that the α-VaR is defined as the α-quantile of F_L. The ES at level α is defined by

$$ES_\alpha \equiv \frac{1}{1-\alpha}\int_\alpha^1 VaR_u(L)du \qquad (6.1)$$

and can thus be understood as an average of all VaRs from level α up to 1. ES is a coherent risk measure—and so subadditive. It is continuous in α and thus avoids cliff effects that may appear when the distribution has discrete components.

If the loss distribution is continuous, there is an even more intuitive representation:

$$ES_\alpha = E(L|L \geq VaR_\alpha), \qquad (6.2)$$

i.e., ES is then the expected loss conditional on this loss belonging to the $100(1-\alpha)$ percent worst losses. This measure has several other names like tail conditional expectation (TCE) or conditional VaR (CVaR). It is the key to simulations-based calculations of ES but care has to be taken as it does not always coincide with ES, and it is also not necessarily subadditive. The technical problem arises if the distribution function jumps from a value below the VaR confidence level to a value above it. Then, a correction term must be introduced into (2) to reconcile it with the correct ES from (1); see Acerbi and Tasche (2002).

The calculation of ES and the marginal contributions of assets to portfolio ES is more challenging than the corresponding calculations for VaR, especially for high confidence levels, because a formula for the α-quantiles of the loss distribution is often missing. Simulations need to be done in most cases. Since the introduction of expected shortfall, substantial progress has been made on computational issues, mainly through the application of importance sampling techniques (Kalkbrener, Lotter, and Overbeck (2004), Egloff, Leippold and Jöhri (2005), or Kalkbrener, Kennedy, and Popp (2007)). Research suggests that computational techniques have advanced to a point that expected shortfall is a viable risk management option for financial institutions.

Remark 3: In Remark 1, it is noted that utility optimisation using a VaR constraint can lead to perverse investment decisions. Risk measures which control the first moment of a random variable (such as ES) have been proposed to overcome this problem. However, recently Wylie, Zhang, and Siu (2010) showed that in the context of hedging both ES and VaR can give rise to discontinuous hedging behaviour that can lead investors to take extremely high-risk positions even when apparently minimising the risk measures.

Backtesting ES

Intuitively, backtesting ES is more complicated and/or less powerful than backtesting VaR because the robust statistic given by the number of VaR violations, as the most common VaR backtest statistic, must be replaced by something that accounts for the *magnitude* of VaR exceedances so that ES backtests by nature have to cope with the size of outliers.

Whether specialised ES backtests are good or not, one simple option is always available: during an ES calculation, the VaR at the same α can be generated as a by-product with low additional effort. One can backtest this VaR with traditional methods; if the VaR is rejected, the corresponding ES calculation can hardly be correct. Of course, VaR backtest acceptance does not guarantee the correctness of the ES calculation, and this would be true even if the VaR backtest were always right.

Some backtests verify if the VaR correctly adjusts for changes in risk dynamics ("conditional coverage"; see Berkowitz and O'Brien (2002)). According to Pritsker (2006), they exploit the fact that exceedances of a correctly calculated VaR "should not help forecast future exceedances. Therefore, the autocorrelation function of the VaR exceedances should be equal to 0 at all lags." It is hard to decide whether ES or VaR is verified with these backtests because VaR exceedances are the very constituents of ES.

Because backtests that are strictly focused on some historical estimator of the risk measure, like the number of VaR violations, often have low power, several authors propose to backtest the whole distribution (or at least the tail), for instance by transforming loss realisations with the forecasted loss distribution: If the latter is correct, the transformed sample must be equally distributed on [0,1]. This hypothesis can be tested (Berkowitz (2001)). While not all backtests of this kind could be used in regulation,[10] Kerkhof and Melenberg (2004) follow this approach to develop test statistics directly applicable to VaR and ES. The test statistic for the ES involves, besides the forecasted ES and VaR, also the calculation of the ES of the *squared* loss, which would be a tolerable extra effort in practice.

Kerkhof and Melenberg (2004) show that their backtest statistics for ES perform better than those for VaR. They also derive regulatory multiplication factors for their backtests and conclude that "the resulting regulatory capital scheme using expected shortfall compares favourably to the current Basel Accord backtesting scheme." It is important to notice that, according to Kerkhof and Melenberg (2004), a comparison of an α-ES with an α-VaR is not "fair" in the context of economic or regulatory capital. Since $ES_\alpha \geq VaR_\alpha$ for the same confidence level α, they lower the confidence level α' for the ES such that $ES(\alpha') \approx VaR(\alpha)$. The intuition is that a regulator would require roughly the same amount of capital for a fixed portfolio, irrespective of the risk measure in use.

This aspect is important not only in the context of backtesting but also when estimation errors for ES and VaR are compared. Yamai and Yoshiba (2005) find ES estimates of fat (generalised Pareto distributed) tailed losses to be much more volatile than their VaR counterparts but they compare ES and VaR at the same confidence level. A comparison in the spirit of Kerkhof and Melenberg (2004) seems not to have been conducted so far but could easily be done.

Wong (2008) suggests another backtest statistic for ES that accounts for the small samples of VaR exceedances. The statistic is derived for normally distributed losses and turns out to perform very well under these assumptions. The test is also powerful in detecting non-normal VaR exceedances. For the case that a bank models non-normal losses when calculating the ES, Wong suggests to derive adapted saddle-point approximations for the estimator's distribution or to use the sample transform as used by Berkowitz (2001) and Kerkhof and Melenberg (2004).

[10] Some of these tests require that the bank fully specifies the loss distribution in the tail, not just the risk measure (ES or VaR). While this should not be a problem for bank internal purposes, a fully specified tail distribution would entail a fairly complex interface between bank and supervisor.

These results are promising, but in the context of banking regulation it must be taken into account that Wong's backtest would require that banks provide more information than they currently do for regulatory backtests. At present, past returns are compared with reported VaRs. With Wong's backtest, the bank would also have to report its estimates of tail thickness, which is potentially involved with weird incentives. For instance, banks might, keeping minimum capital constant, be tempted to rely on certain tail distributions under which Wong's backtest has particular low power so that it is difficult to provide firm evidence of wrong risk reporting. Whether such concerns are substantial is left to future research.

Spectral Risk Measures

Spectral risk measures (SRM) are a promising generalisation of ES (Acerbi (2002)). While the α-ES assigns equal weight to all β-VaRs with $\beta \geq \alpha$ but zero to all others, an SRM allows these weights to be chosen more freely. This is implemented by a weight function $w:[0,1] \to [0, \infty)$ that integrates to 1. An SRM is formally defined as

$$SRM = \int_0^1 w(u) VaR_u(L) du$$

Expected shortfall is a special case of spectral measure, where $w(u) = (1 - \alpha)^{-1} 1_{\{\alpha \leq u \leq 1\}}$. The definition of SRM is restricted to functions w that increase over [0,1], which ensures that the risk measure is coherent. This restriction also implies that larger losses are taken more seriously than smaller losses and thus the function w establishes a relationship to risk aversion. The intuition is that a financial institution is not very risk averse for small losses, which can be absorbed by income, but becomes increasingly risk averse to larger losses. As there may be a level of loss where employing additional capital to absorb yet higher loss is no longer desirable, such losses should be given the highest weights from a regulator's angle because often the public would have to bear such losses. Intuitively, a weight function that increases can also be thought of as marginal costs that rise while losses become increasingly rare, i.e., large.

Another advantage of SRM over ES (and VaR, a fortiori) is that they are not bound to a single confidence level. Rather, one can choose w to grow continuously with losses and thereby make the risk measure react to changes in the loss distribution more smoothly than the ES, and avoid the risk that an atom in the distribution being slightly above or below the confidence level has large effects.

If the underlying risk model is simulation-based, the additional effort to calculate an SRM as opposed to the ES seems negligible; the simulated VaR realisation are just differently weighed (Acerbi (2002)).

In spite of their theoretical advantages, SRMs other than ES are still seldom used in practice.[11] However, insurers use the closely related concept of *distortion measures* (see the next section). Prominent examples such as the measure based on the Wang transformation (see next page) are also SRMs.

Remark 4: Leaving aside that w must be increasing to meet the definition of SRM, VaR is a limiting case of spectral risk measures: for instance, the sequence of SRMs based on the weight functions $w_n(u) \equiv 0.5n1_{\{\alpha-n^{-1}\le u<\alpha+n^{-1}\}}$ converges to the α-VaR.

Other Risk Measures

There also are a number of other risk measures which are briefly introduced in this subsection.

Distortion risk measures: These measures are used in actuarial risk measurement. The definition is very general; both spectral risk measures (including ES) and the VaR are nested. To define distortion risk measures, let D be any distribution function on [0,1] that is right-continuous and increasing with $D(0) = 0$ and $D(1) = 1$. This D is called the *distortion function*. A *distortion risk measure* of loss L is defined as

$$DM(L) = \int_0^1 VaR_u(L)dD(u)$$

Each spectral risk measure is clearly a distortion risk measure; to see this, recall that the weight function w integrates to 1 and observe that the SRM and the distortion measure defined by the antiderivative $D(u) \equiv \int_0^u w(s)ds$ are identical.

Distortion risk measures are not necessarily coherent; the definition allows for distortion functions with a non-monotonous derivative (this is just the weight function of the corresponding SRM), whereas Acerbi (2002) has shown that the monotonicity of w is also necessary for the risk measure to be coherent.[12]

The VaR has a representation as a distortion risk measure by $D_{VaR}(u) = 1_{\{u\ge\alpha\}}$.

The Wang transform (Wang (2001)) $D_\theta^{Wang}(u) = \Phi(\Phi^{-1}(u) + \log\theta)$, where Φ denotes the Gaussian distribution function and $\theta < 1$, is an interesting distortion function. The corresponding risk measure is also a spectral risk measure because the first derivative of D_θ^{Wang} is strictly increasing. Hence the Wang transform indeed implements risk aversion over the whole range of losses but particularly in the tail. It has been applied to the pricing of catastrophe insurance contracts and exotic option pricing where Black-Scholes assumptions cannot be applied.

Variance: The variance is historically the most important risk measure and widely used in practice. It has many desirable properties but at least two drawbacks from a regulatory perspective. McNeil et. al. (2005) state "if we want to work with variance, we have to assume that the second moment of the loss distribution exists. . . . [V]ariance is a good measure of risk only for distributions which are (approximately) symmetric. . . . However, in many areas of risk management, we deal with highly skewed distributions."

The **mean deviation**, defined as $MD(L) \equiv \mathbf{E}|L - \mathbf{E}L|$, can do without second moments but suffers from the same problems with skewed distributions as the variance. It is less accessible to analytical treatment than the variance and therefore rarely used as a risk measure.

Upper partial moments (see McNeil et. al. (2005)): Given a loss distribution F_L, an exponent $k \ge 0$ and a reference point q, which could be some VaR, the upper partial moment $UPM(k,q)$ is defined as

$$UPM(k,q) = \int_q^\infty (l - q)^k \, dF_L(l)$$

Hence, for $k > 1$ an UPM measures losses beyond the threshold q with increasing weight. It is therefore related to spectral risk measures in spirit but not equivalent in analytic terms. The higher k is, the more conservative is the UPM. For $k = 1$ and continuous loss distributions, there is a close relationship with expected shortfall:

$$UPM(1, VaR_\alpha) = (1 - \alpha)(ES_\alpha - VaR_\alpha)$$

Left-tail measure: In a similar vein of mean deviation and lower (upper) partial moment, Wu and Xiao (2002) propose a *left-tail measure*, defined as the conditional standard deviation of VaR exceedances, i.e.,

$$LTM \equiv \sqrt{E\left\{[L - E(L|L \ge VaR_\alpha)]^2 \mid L \ge VaR_\alpha\right\}}$$

Wu and Xiao (2002) show that the left-tail measure is useful particularly for the measurement of non-normal tail risks. This risk measure has several undesirable features such as a lack of coherency and a heavy burden of calculation.

Conclusions

While VaR has been criticised for its lack of coherence, until recently it was unclear whether this flaw is relevant for real

[11] At least one reputable risk consulting company reports it is currently implementing an SRM-based risk management system for some of its clients.

[12] Wang (2001) claims all smooth distortion measures are coherent. This is wrong as subadditivity is missing in general. Wang (2001) means to build on Wang, Young and Panjer (1997) which, however, state that a distortion measure is subadditive *if it is convex* (in our notation). The latter is correct and conforms to Acerbi (2002).

asset portfolios, particularly for risks in the trading book. Degen et. al. (2007) have shown that the lack of coherence can be an important problem for trading book risk measurement. A risk measurement based on VaR is thus not necessarily conservative.

The ES avoids the major flaws of VaR but its fundamental difference from VaR—that it accounts for the magnitude of losses beyond a threshold—is an equally important advantage. By this, it aligns the interests of bank managers and owners to those of the public much better than VaR.

Much of the criticism of ES that has been brought forward in defence of VaR could be refuted. Advanced simulation techniques have helped to make ES calculations stable enough, and ES and VaR backtests have similar power, if compared on the basis that both risk measures have roughly the same value.

Spectral risk measures are a promising generalisation of expected shortfall. The main advantages are improved smoothness and the intuitive link to risk aversion. If the underlying risk model is simulations-based, the additional calculation effort as opposed to ES seems negligible.

6.5 STRESS TESTING PRACTICES FOR MARKET RISK

Overview

VaR limitations have been highlighted by the recent financial turmoil. Financial industry and regulators now regard stress tests as no less important than VaR methods for assessing a bank's risk exposure. A new emphasis on stress testing exercises derives also from the amended Basel II framework which requires banks to compute a valid stressed VaR number.

A stress test can be defined as a risk management tool used to evaluate the potential impact on portfolio values of unlikely, although plausible, events or movements in a set of financial variables (Lopez (2005)). They are designed to explore the tails of the distribution of losses beyond the threshold (typically 99%) used in value-at-risk (VaR) analysis.

However, stress testing exercises often are designed and implemented on an ad hoc compartmentalised basis, and the results of stress tests are not integrated with the results of traditional market risk (or VaR) models. The absence of an integrated framework creates problems for risk managers, who have to choose which set of risk exposures are more reliable. There is also the related problem that traditional stress testing exercises typically remain silent on the likelihood of stress-test scenarios.

A survey of stress testing practices conducted by the Basel Committee in 2005 showed that most stress tests are designed around a series of scenarios based either on historical events, hypothetical events, or some combination of the two. Such methods have been criticised by Berkowitz (2000a). Without using a risk model the probability of each scenario is unknown, making its importance difficult to evaluate. There is also the possibility that many extreme yet plausible scenarios are not even considered.

Berkowitz proposed the integration of stress testing into formal risk modelling by assigning probabilities to stress-test scenarios. The resulting risk estimates incorporate both traditional market risk estimates and the outcomes of stress tests, as well as the probabilities of each. Therefore, they provide an integrated set of risk indicators and estimates to work with.

Incorporating Stress Testing into Market-Risk Modelling

Traditional stress testing exercises can be classified into three main types, which differ in how the scenarios are constructed:

1. historical scenarios;
2. predefined or set-piece scenarios where the impact on P/L of adverse changes in a series of given risk factors is simulated;
3. mechanical-search stress tests, based on automated routines to cover prospective changes in risk factors, then the P/L is evaluated under each set of risk-factor changes, and the worst-case results are reported.

All these approaches depend critically on the choice of scenarios. A related problem is that the results of stress tests are difficult to interpret because they give no idea of the probabilities of the events concerned (Berkowitz (2000a)). These criticisms can be addressed by integrating stress testing into the market risk modelling process and assigning probabilities to the scenarios used in stress testing. Once scenarios are put in probabilistic form, a unified and coherent risk measurement system is obtained rather than two incompatible ones and backtesting procedures can be applied to impose some (albeit limited) check on scenarios. Inevitably, the choice of scenarios will remain subjective, but even there, the need to assign probabilities to scenarios will impose some discipline on risk management.

Several authors have developed an integrated approach to stress testing including Kupiec (1998) who examines cross-market effects resulting from a market shock and

Aragones et. al. (2001) who incorporated hypothetical stress events into an Extreme Value Theory (EVT) framework.

Alexander and Sheedy (2008) analysed the problem of determining the most suitable risk model in which to conduct a stress test. Obviously if the model is mis-specified, their approach is vulnerable to a considerable degree of model risk. Hence a significant part of their research is supported through backtests, which are designed to reduce the model risk in risk models that are used for stress testing. They conduct backtests for eight risk models, including both conditional and unconditional models and four possible return distributions. Their backtesting experiment suggests that unconditional historical simulation, currently the most popular VaR methodology in the industry according to Perignon and Smith (2006), is likely to be mis-specified and is therefore unsuited for stress testing purposes.

Breuer et. al. (2009) define an operational definition to three requirements which the Basel Committee specifies for stress tests: plausibility and severity of stress scenarios as well as suggestiveness of risk-reducing actions. The basic idea of their approach is to define a suitable region of plausibility in terms of the risk-factor distribution and search systematically for the scenario with the worst portfolio loss over this region. One key innovation of their approach compared with the existing literature is the solution of two open problems. They suggest a measure of plausibility that is not dependent to the problem of dimensional dependence of maximum loss and they derive a way to consistently deal with situations where some but not all risk factors are stressed. They show that setting the non-stressed risk factors to their conditional expected value given the value of the stressed risk factors, the procedure first suggested by Kupiec (1998), maximises plausibility among the various approaches used in the literature. Furthermore, Breuer et. al. (2010b) propose a new method for analyzing multi-period stress scenarios for portfolio credit risk more systematically than in the current practice of macro stress testing. This method quantifies the plausibility of scenarios by considering the distance of the stress scenario from an average scenario. For a given level of plausibility their method searches systematically for the most adverse scenario for the given portfolio.

Finally, as a general point, it must be underlined that for the purposes of calculating the P&L impact of stress shock-factors it is generally assumed that the shock occurs instantaneously, i.e., that traders have no opportunity to re-hedge or adjust their positions, and it is ignored the impact of declining tenors for, for example, futures and options contracts. Apart from simplifying the calculations, such an assumption could be unreasonable in some cases given the practical experience of the actions of traders during historical events, and it may generate inconsistent results by amplifying the magnitude of the losses. Such issues have not yet been addressed in the literature.

Stressed VaR

The pressing technical issue now facing financial institutions that intend to comply with the amended Basel II framework is to understand how to calculate a valid stressed VaR number. After the revisions of July 2009, banks have to calculate a VaR using the risk engine it normally uses but "with model inputs calibrated to historical data from a continuous 12-month period of significant financial stress relevant to the bank's portfolio" (Basel Committee on Banking Supervision (2009b)).

An over-simplistic interpretation of this specification might be to increase the assumed volatilities of the securities in a portfolio. This would have the effect of lengthening the tails of the Gaussian (normal) loss distributions that underlie all standard VaR calculations.

However, in order to calculate stressed VaR accurately it is also necessary to stress the correlation matrix used in all VaR methodologies. It is a repeated observation that during times of extreme volatility, such as occurs during every market crash, correlations are dramatically perturbed relative to their 'normal' historical values. In general, most correlations tend to increase during market crises, asymptotically approaching 1.0 during periods of complete meltdown, such as occurred in 1987, 1998 and 2008.

One possibility is to adopt the conditional stress test approach of Kupiec (1998). In this approach, the risk factor distributions are conditional on an extreme value realisation of one or more of the risk factors. Conditional on a large move of at least one factor, the conditional factor covariance matrix exhibits much higher correlations among the remaining factors. In this approach, the apparent shift in the correlation structure is a consequence of conditioning the distribution on a large factor shock. The unconditional correlations remain unchanged. Analysing a large number of stress test results for currency portfolios over the Asian currency crisis period, Kupiec shows that the conditional stress test process performs extremely well as very few stress test violations are recorded during this crisis period.

An alternative approach to conditional correlation is to stress the unconditional correlation matrix of the risk factors. Unfortunately, this approach is not as straightforward as the conditional correlation approach or stretching the tails of the loss distributions. The VaR calculation engine requires a correlation

matrix that satisfies the mathematical property of positive definiteness, which is a way of saying that all of the correlations are internally consistent with each other. Noisy or erroneous historical price data can result in matrices that are not positive definite. Perturbing the correlation matrix, which is necessary for a true stressed VaR calculation, may result in correlation matrices that also violate the internal consistency requirement. If the matrix is not positive definite the VaR calculus will fail, so methods have to be devised to modify the stressed matrix until it becomes positive definite. Kupiec (1998) discusses some practical methods that can be used to address this problem.

Besides these technical issues one may also more fundamentally consider concepts that are not covered by the current regulatory definition of stressed VaR. A more sophisticated approach might include not only linear transforms of multivariate normal risk factors but also employing 'fat-tailed' distributions to model the extreme loss events more accurately. Examples of those 'extreme value theory' distributions are the Gumbel, Generalised Pareto, Weibull, Fréchet, and the Tukey g&h distributions.

However, one should keep in mind that the stressed VaR is from a theoretical perspective an imperfect solution—its purpose is to reflect that current market conditions may not lead to an accurate assessment of the risk in a more stressful environment. Extreme value theory distributions may already incorporate extreme market conditions and could in principle make a stressed VaR redundant. In general, these distributions are flexible enough to obtain very good fits but serious robustness issues arise instead, as regulators and risk managers had to learn in the context of operational risk, for instance.

Conclusions

More recent research advocates the integration of stress testing into the risk modelling framework. This would overcome drawbacks of reconciling stand-alone stress test results with standard VaR model output.

Progress has also been achieved in theoretical research on the selection of stress scenarios. In one approach, for example, the "optimal" scenario is defined by the maximum loss event in a certain region of plausibility of the risk factor distribution.

The regulatory "stressed VaR" approach is still too recent to have been analyzed in the academic literature. Certain methods that could be meaningful in this context can be identified in the earlier literature on stress testing. Employing fat-tailed distributions for the risk factors and replacing the standard correlation matrix with a stressed one are two examples.

6.6 UNIFIED VERSUS COMPARTMENTALISED RISK MEASUREMENT

Overview

In this section, we survey the academic literature on the implications of modelling the aggregate risks present across a bank's trading and banking books using either a compartmentalised approach—namely, the sum of risks measured separately—or a unified approach that considers the interaction between these risks explicitly. Finally, we survey the recent literature on the systemic implications of the current regulatory capital requirements that aggregate capital requirements across risk types.

In many financial institutions, aggregate economic capital needs are calculated using a two step procedure. First, capital is calculated for individual risk types, most prominently for credit, market and operational risk. In a second step, the stand-alone economic capital requirements are added up to obtain the overall capital requirement for the bank.

The Basel framework for regulatory capital uses a similar idea. As discussed by Cuenot, Masschelein, Pritsker, Schuermann and Siddique (2006), the Basel framework is based on a "building block" approach such that a bank's regulatory capital requirement is the sum of the capital requirements for each of the defined risk categories (i.e., market, credit and operational risk), which are calculated separately within the formulas and rules that make up Pillar 1. Capital requirements for other risk categories are determined by the supervisory process that fits within Pillar 2; see Figure 6.2 which is reproduced from Cuenot

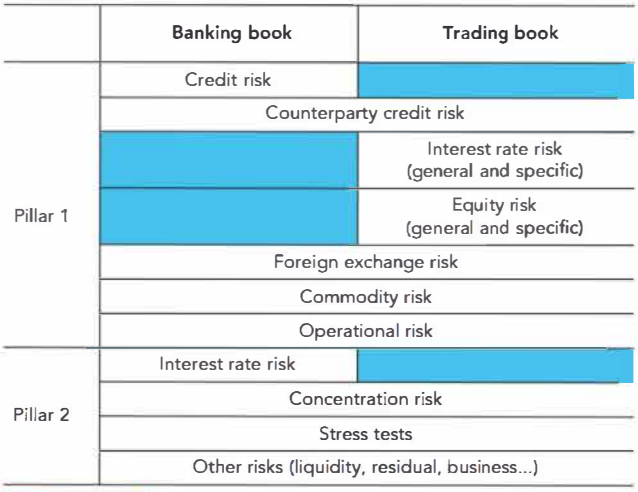


Figure 6.2 Overview of risk categories relevant for banking book and trading book in Pillar 1 and Pillar 2.

Source: Cuenot et. al. (2006).

et. al. (2006). This approach is therefore often referred to as a non-integrated approach to risk measurement. An integrated approach would, by contrast, calculate capital for all the risks borne by a bank simultaneously in one single step and accounting for possible correlations and interactions, as opposed to adding up compartmentalised risk calculations.

Pressure to reconsider the regulatory compartmentalised approach came mainly from the financial industry, where it has been frequently argued that a procedure that simply adds up economic capital estimates across portfolios ignores diversification benefits. These alleged benefits have been estimated to be between 10 and 30% for banks (see Brockmann and Kalkbrener (2010)).

Capital diversification arguments and estimates of potential capital savings are partially supported in the academic literature. More recently this view and the estimates have been fundamentally challenged by the Basel Committee (Basel Committee on Banking Supervision (2009)) and by Breuer et. al. (2010a). These papers have pointed out that nonlinear interaction between risk categories may even lead to compounding effects. This fact questions whether the compartmentalised approach will in general give a conservative and prudent upper bound for economic capital.

Is this a merely academic debate or does it have practical implications for reform considerations related to the trading book? In this section, we survey the main arguments and give a brief review of the main papers and their findings. We then discuss policy implications that might be relevant for a discussion related to potential future reform related to the trading book.

Aggregation of Risk: Diversification versus Compounding Effects

Diversification is a term from portfolio theory referring to the mix of a variety of investments within a portfolio. Since different investments will develop differently in the future with value losses in some investment offset by value gains in another investment, the overall portfolio risk is reduced through the spreading of risk. In a similar way, the assets of a bank can be thought of as an overall portfolio that can be divided into subportfolios. If risk analysis is done by looking at risk measures at the level of the subportfolios and the risk measures are added up, the intuition of diversification suggests that we should arrive at a conservative risk measure for the bank as a whole.

So, what is wrong with this straightforward intuition about diversification between market, credit and other risk categories? The flaw in the intuition lies in the fact that it is usually not possible to divide the overall portfolio of a bank into subportfolios purely consisting of market, credit and operational risk; these risk categories are too intertwined in a modern financial institution to possibly separate in a meaningful way. In short, we cannot construct a subportfolio of risk factors. It is therefore incorrect to think of the banking book as a subportfolio of the overall bank portfolio for which only credit risk is relevant. It is also incorrect to view the trading book as another subportfolio related solely to market risk.

A simple way to summarise this argument is to consider a portfolio of loans. The interest rate risk related to such a portfolio is usually counted as a market risk, and this risk affects the bank's refinancing costs and the revaluation of these loans. If the interest rate risk is borne by the creditors in some way, this market risk suddenly may transform into a credit risk for the bank. So, do assets with a value that fluctuates with interest rates belong in a subportfolio for market risk or in a subportfolio of credit risk? They clearly belong to both, because each loan has a market risk component as well as a credit risk component simultaneously. Trading book positions with counterparty risks or positions related to carry trades fall into the same category.

Breuer et. al. (2010a) consider portfolios of foreign currency loans, which are loans denominated in a foreign currency extended to domestic creditors with income in domestic currency. The credit risk in these portfolios is always a function of the market risk (i.e., exchange rate movements), and the risk of each position in a foreign currency loan portfolio has simultaneously a credit and a market risk component. Adding up capital and hoping for an upper bound amounts to ignoring possible "malign risk interactions" as they are called in Breuer et. al. (2010a). This issue has been known in the market risk literature for a long time as "wrong way risk." Wrong way risk is the risk arising from the problem that the value of a trading position is inversely correlated with the default risk of some counterparty.

From these examples, we see that a formation of subportfolios along the lines of risk factors—and for that matter across banking and trading books—is usually not possible. Breuer et. al. (2010a) indeed show that the ability to form subportfolios along the lines of risk categories is a sufficient condition for diversification effects to occur. Since we can in general not form such subportfolios, we must anticipate the possibility that there can be risk compounding effects between the banking and the trading book. In short, while the intuition of diversification is inviting, it does not apply to the interaction of banking and trading books since there are in general no pure subportfolios of market, credit or operational risks.

This insight is important because it demonstrates that "diversification effects" that are derived from papers using a so-called

"top-down" approach are often assuming what they want to derive. By construction, the assumption of splitting up the bank portfolio into subportfolios according to market, credit and operational risk assumes that this can indeed be done. If such a split were possible, it follows from the results in Breuer et. al. (2010a) that diversification effects must occur necessarily.

To estimate the quantitative dimension of the problem, we therefore must focus on papers working with a "bottom-up" approach. We also need to examine the results of papers based on the "top-down" approach that assumes risk separability at the beginning of the analysis. In this section, we survey several key papers that use either of these risk aggregation methods. As part of this literature survey, we provide a summary of recent papers that estimate the range and magnitude of these differences between compartmentalised and unified risk measures. Our proposed measure is a simple ratio of these two measures, as used in other papers, such as Breuer et. al. (2010a). In that paper, the authors adopt the term "inter-risk diversification index" for the ratio; see also the related measure in Alessandri and Drehmann (2010). Ratio values greater than one indicate risk compounding, and values less than one indicate risk diversification. In the summary tables later in this chapter, we list the various papers, the portfolio analysed, the risk measures used, the horizon over which the risks are measured, and these risk ratios.

Papers Using the "Bottom-Up" Approach

As mentioned above, a common assumption of most current risk measurement models is that market and credit risks are separable and can be addressed independently. Yet, as noted as early as Jarrow and Turnbull (2000), economic theory clearly does not support this simplifying assumption.

While the reasons behind this common assumption are mostly operational in nature, some studies have used numerical simulation techniques to generate results. For example, Barnhill and Maxwell (2002) examine the economic value of a portfolio of risky fixed income securities, which they define as a function of changes in the risk-free interest rate, bond spreads, exchange rates, and the credit quality of the bond issuers. They develop a numerical simulation methodology for assessing the VaR of such a portfolio when all of these risks are correlated. Barnhill et. al. (2000) use this methodology to examine capital ratios for a representative South African bank. However, in these studies, the authors do not examine the differing values of their chosen risk measures using a unified risk measurement approach versus a compartmentalised approach that sums the independent risk measures.

The study by Jobst, Mitra and Zenios (2006) provides some analysis along these lines. The authors construct a simulation model, based on Jobst and Zenios (2001), in which the risk underlying the future value of a bond portfolio is decomposed into:

- the risk of a borrower's rating change (including default);
- the risk that credit spreads will change; and
- the risk that risk-free interest rates will change.

Note that the first item is more narrowly defined to represent the portfolio's credit risk, while the last item is more narrowly defined to represent the portfolio's market risk. However, the middle item is sensitive to both risks and challenges the notion that market and credit risk can be readily separated in this analysis. The authors use portfolios of US corporate bonds and one-year VaR and CVaR risk measures at the 95%, 99% and 99.9% confidence levels for their analysis.

In their analysis, the authors generate risk measures under three sets of assumptions. To concentrate on the pure credit risk contributions to portfolio losses, they simulate only rating migration and default events as well as recovery rates, while assuming that future interest rates and credit spreads are deterministic. The authors then allow future credit spreads to be stochastically determined, and finally, they allow future interest rates to be stochastically determined. Note that the latter case provides an integrated or unified risk measurement, according to our definition for this survey.[13]

The authors' results are quite strong regarding the magnitude of the risk measures across risk types and credit ratings. For AAA-rated bonds, the authors find that the unified risk measures at all three tail percentiles are on the order of ten times the pure credit risk measures, since highly-rated bonds are unlikely to default. As the credit quality of the portfolio declines, the ratio between the unified risk measures and the risk measures for pure credit risk drops to just above one for C-rated bonds.

Table 6.1 presents a short summary of several papers for which we can directly examine the ratio of unified to compartmentalised risk measures for bottom-up models. As mentioned earlier, the recent work of Breuer et. al. (2008, 2010a) provides a leading example of how market and credit risk cannot be readily separated in a portfolio, a fact that complicates risk measurement and works to undermine the simple assumptions underling additive risk measures.

[13] Note, however, that the authors do not conduct an analysis of a market risk scenario (i.e., deterministic ratings and stochastic credit spreads and interest rates). Thus, we cannot examine their ratio of unified to compartmentalised risk measures as discussed above.

In Breuer et. al. (2010a), the authors present analysis of hypothetical loan portfolios for which the impact of market and credit risk fluctuations are not linearly separable. They argue that changes in aggregate portfolio value caused by market and credit risk fluctuations in isolation should sum up to the integrated change incorporating all risk interactions very rarely. The magnitude and direction of the discrepancy between these two types of risk assessments can vary broadly. For example, the authors examine a portfolio of foreign currency loans for which exchange rate fluctuations (i.e., market risk) affect the size of loan payments and hence the ability of the borrowers to repay the loan (i.e., credit risk). For their empirically calibrated example, they use expected shortfall at various tail percentiles as their risk measure and examine portfolios of BBB+ and B+ rated loans. Their analysis shows that changes in market and credit risks can cause compounding losses such that the sum of value changes from the individual risk factors are smaller than the value change due to accounting for integrated risk factors.

In particular, their reported inter-risk diversification index for expected shortfall increased sharply as the tail quantile decreased, which suggests that the sum of the two separate risk measures becomes much less useful as an approximation of the total integrated risk in the portfolio as we go further into the tail. These index values also increase for all but the most extreme tail percentiles as the original loan rating is lowered. The authors argue that this example presents evidence of a "malign interaction of market and credit risk which cannot be captured by providing separately for market risk and credit risk capital." The authors show a similar qualitative outcome for domestic currency loans (i.e., loans for which default probability are simply a function of interest rates), although the index values are much lower.

In Breuer et. al. (2008), the authors use a similar analytical framework to examine variable rate loans in which the interaction between market and credit risk can be analysed. In particular, they model the dependence of credit risk factors—such as the loans' default probabilities (PD), exposure at default (EAD), and loss-given-default (LGD)—on the interest rate environment. A key risk of variable rate loans is the danger of increased defaults triggered by adverse rate moves. For these loans, market and credit risk factors cannot be readily separated, and their individual risk measures cannot be readily aggregated back to a unified risk measure. They conduct a simulation study based on portfolios of 100 loans of equal size by borrowers rated B+ or BBB+ over a one-year horizon using the expected shortfall measure at various tail percentiles. They find that the ratio of unified expected shortfall to the sum of the separate expected shortfalls is slightly greater than one, suggesting that risk compounding effects can occur. Furthermore, these compounding effects are more pronounced for lower-rated loans and higher loan-to-value ratios.

In contrast to this work, the paper by Grundke (2005) lays out a bottom-up model that assumes the separability of interest rate risk (i.e., market risk) and credit spread risk (i.e., credit risk). The author examines a calibrated multi-factor credit risk model that accommodates various asset value correlations, correlations between credit spreads and other model factors, and distributional assumptions for innovations. The author examines hypothetical loan portfolios of varying credit quality over a three-year horizon, both with and without the joint modelling of interest rates and credit spreads. To assess the joint impact of interest rate and credit risk, the author uses forward market interest rates instead of separate interest rate and credit spread processes. Interestingly, the reported VaR measures at various tail percentiles lead to ratios of unified VaR measures to summed VaR measures that range widely from near zero to one, which seems to be due mainly to the separability of the interest rate risk (i.e., market risk) and credit spread risk (i.e., credit risk) in the model.

Kupiec (2007) proposes a single-factor, migration-style credit risk model that accounts for market risk. This modelling approach generates a portfolio loss distribution that accounts for the non-diversifiable elements of the interactions between market and credit risks. The integrated exposure distribution of the model is used to examine capital allocations at various thresholds. These integrated capital allocations are compared to the separated assessments. The results show that capital allocations derived from a unified risk measure importantly alter the estimates of the minimum capital needed to achieve a given target solvency margin. The capital amount could be larger or smaller than capital allocations estimated from compartmentalised risk measures. Regarding specifically the Basel II AIRB approach, the author argues that the results show that no further diversification benefit is needed for banking book positions since no market risk capital is required. Thus, Basel II AIRB capital requirements fall significantly short of the capital required by a unified risk measure.

Numerically speaking, the risk measure used in this study is the amount of capital that the unified and the compartmentalised capital approaches generate as the appropriate value to assure funding costs of a certain magnitude calibrated to historical funding rates for specific credit ratings. The hypothetical portfolios of interest are corporate loans with various rating categories represented in proportion to historical data. The author examines a wide variety of alternative separated approaches with which to calculate economic capital measures, ranging from three different alternative credit risk models to several methods for measuring market risk. Correspondingly, the range of inter-risk diversification index values is quite wide for the AAA- and BBB-rated portfolios, ranging from about 0.60

to almost 4.00. In summary, the author's capital calculations show that capital allocations derived from a unified market and credit risk measure can be larger or smaller than capital allocations that are estimated from aggregated compartmentalised risk measures.

The studies discussed above examine the different risk implications of a unified risk measurement approach relative to a compartmentalised approach for specific portfolios. In contrast, Drehmann et. al. (2010) examine a hypothetical bank calibrated to be representative of the UK banking system as a whole. Within their analytical framework, they do not explicitly assume that market and credit risk are separable. The authors decompose the total risk in their bank scenario analysis into:

- the impact of credit risk from non-interest rate factors,
- the impact of interest rate risk (excluding the effect of changes in interest rates on credit risk), and
- the impact of the interaction of credit risk and interest rate risk.

The latter is calculated as the difference between the total impact of the scenario shock and the sum of the first two components.

Their simulations confirm that interest rate risk and credit risk must be assessed jointly for the whole portfolio to gauge overall risk correctly. In particular, the authors find in their simulations that if banks gauged credit risk by solely monitoring their write-offs, aggregate risk would be underestimated in the short term since a rate increase would also lower its net interest income and profits. Correspondingly, the bank's aggregate risk would be overestimated in the long run as net interest income and profits recover while write-offs continue to rise.

Their main variable of interest is net profits over twelve quarters after their macroeconomic stress scenario hits their representative bank, although they also report separate measures of write-offs and net interest income. They report that the interaction between interest rate and credit risk accounts for about 60% of the decline in capital adequacy for their calibrated bank. While the decline in capital adequacy does not perfectly match our other risk measures, we can still think of the diversification index here as the ratio of the capital decline for the unified risk framework relative to the capital decline that would come from separate identification of market and credit risks. Given their reported numbers, that ratio here is $100\%/(100\% - 60\%) = 2.5$, which suggests a very clear contribution of this interaction to risk management concerns.

Following up on the work of Drehmann et. al. (2010), Alessandri and Drehmann (2010) develop an integrated economic capital model that jointly accounts for credit and interest rate risk in the banking book; i.e., where all exposures are held to maturity. Note that they explicitly examine repricing mismatches (and thus market and credit risks) that typically arise between a bank's assets and liabilities.

For a hypothetical, average UK bank with exposures to only the UK and US, they find that the difference between aggregated and unified economic capital levels is often significant but depends on various bank features, such as the granularity of assets, the funding structure or bank pricing behaviour. They derive capital for the banking book over a one year horizon. For credit and interest rate risk, they define unexpected losses and thus economic capital as the difference between VaR at the specified 99% confidence level and expected losses. Note that their measures of economic capital for just credit risk and just interest rate risk do not fully disentangle these risks as the credit risk measure incorporates the effects of higher interest rates on default probabilities and the latter the effect of higher credit risk on income. The key point is that the framework represents a plausible description of how current capital models for the banking book capture these risks.

The authors examine the ratio of unified economic capital to the sum of the component measures at three VaR quantiles. For the 95th percentile of portfolio losses, unified capital measure is near zero, and thus the ratio is nearly zero as well. For the 99th percentile, the ratio is quite small at 0.03, but the ratio rises quickly to just over 50% for the 99.9th percentile. Note, however, that this result still suggests that the compartmentalised approach is more conservative than the unified approach. The authors examine certain modifications of their assumptions—such as infinitely fine-grained portfolios to increase the correlation of portfolio credit risk with the macroeconomic factors, banking funding scenarios from all short-term debt that is frequently repriced to all long-term debt that is repriced only on a yearly basis—and find some interesting difference with the base case scenario. However, the lower integrated capital charge holds.

On balance, these authors conclude that the bank's capital is mismeasured if risk interdependencies are ignored. In particular, the addition of economic capital for interest rate and credit risk derived separately provides an upper bound relative to the integrated capital level. Two key factors determine this outcome. First, the credit risk in this bank is largely idiosyncratic and thus less dependent on the macroeconomic environment; and second, bank assets that are frequently repriced lead to a reduction in bank risk. Given that these conditions may be viewed as special cases, the authors recommend that "As a consequence, risk managers and regulators should work on the presumption that interactions between risk types may be such that the overall level of capital is higher than the sum of capital derived from risks independently."

Papers Using the "Top-Down" Approach

An alternative method for determining total firm risk, primarily for enterprise-wide risk management, is to aggregate risks calculated for different business lines or different risk types using so-called "top-down" approaches. An important difference is that top-down approaches always reference an institution as a whole, whereas bottom-up approaches can range from the portfolio level up to an institutional level. With respect to market and credit risk, the top-down approach explicitly assumes that the risks are separable and can be aggregated in some way. As outlined by Cuenot et. al. (2006), firms may compute their market and credit risk capital separately and aggregate the two risk types by imposing some form of correlation between them. The top-down approach thus does not require a common scenario across risk types, but because the correct form of aggregation is not known, the approach "loses the advantages of logical coherence." In addition, as suggested by Breuer et. al. (2008, 2010a), the assumption of separable risk will generally prevent the ability to gauge the degree of risk compounding that might be present and instead typically provide support for risk diversification.

The literature is unclear on whether the combination of financial business lines within one organisation leads to an increase or decrease in risk. The literature as surveyed by Saunders and Walters (1994) and Stiroh (2004) suggests mixed results. However, as surveyed by Kuritzkes, Schuermann and Weiner (2003), several studies, including their own, suggest that reductions in economic capital arise from the combination of banking and insurance firms. The papers surveyed in Table 6.2 and below find this result as well for various risk combinations at the firm level.

For example, Dimakos and Aas (2004) decompose the joint risk distribution for a Norwegian bank with an insurance subsidiary into a set of conditional probabilities and impose sufficient conditional independence that only pair-wise dependence remains; the total risk is then just the sum of the conditional marginals (plus the unconditional credit risk, which serves as their anchor). Their simulations indicate that total risk measured using near tails (95%–99%) is about 10%–12% less than the sum of the individual risks. In terms of our proposed ratio, the value ranges from 0.88 to 0.90. Using the far tail (99.97%), they find that total risk is often overestimated by more than 20% using the additive method. In terms of our proposed ratio of unified risk measure to the sum of the compartmentalised risk measures, its value would be 0.80.

Similarly, Kuritzkes et. al. (2003) examine the unified risk profile of a "typical banking-insurance conglomerate" using the simplifying assumption of joint normality across the risk types, which allows for a closed-form solution. They use a broad set of parameters to arrive at a range of risk aggregation and diversification results for a financial conglomerate. Based on survey data for Dutch banks on the correlations between losses within specific risk categories, their calculations of economic capital at the 99.9% level is lower for the unified, firm-level calculation than for the sum of the risk-specific, compartmentalised calculations. The ratio of these two quantities ranges from 0.72 through 0.85 based on correlation assumptions across market, credit and operational risk.

Rosenberg and Schuermann (2006) conduct a more detailed, top-down analysis of a representative large, internationally active bank that uses copulas to construct the joint distribution of losses. The copula technique combines the marginal loss distributions for different business lines or risk types into a joint distribution for all risk types and takes account of the interactions across risk types based on assumptions. Using a copula, parametric or nonparametric marginals with different tail shapes can be combined into a joint risk distribution that can span a range of dependence types beyond correlation, such as tail dependence. The aggregation of market, credit and operational risk requires knowledge of the marginal distributions of the risk components as well as their relative weights. Rosenberg and Schuermann assign inter-risk correlations and specify a copula, such as the Student-t copula, which captures tail dependence as a function of the degrees of freedom. They impose correlations of 50% for market and credit risk, and 20% for the other two correlations with operational risk; all based on triangulation with existing studies and surveys.[14]

Rosenberg and Schuermann find several interesting results, such as that changing the inter-risk correlation between market and credit risk has a relatively small impact on total risk compared to changes in the correlation of operational risk with the other risk types. The authors examine the sensitivity of their risk estimates to business mix, dependence structure, risk measure, and estimation method. Overall, they find that "assumptions about operational exposures and correlations are much more important for accurate risk estimates than assumptions about relative market and credit exposures or correlations." Comparing their VaR measures for the 0.1% tail to the sum of the three different VaR measures for the three risk types, they find diversification benefits in all cases. For our benchmark measure of the ratio between the unified risk measure and the compartmentalised risk measure, their results suggest values ranging from 0.42 to 0.89. They found similar results when the expected shortfall (ES) measure was used.

[14] Note that different correlation values could lead to risk compounding, but it is not clear what those values might be and what values would be implied by the bottom-up exercises discussed here.

Note that the authors state that the sum of the separate risk measures is always the most conservative and overestimates risk, "since it fixes the correlation matrix at unity, when in fact the empirical correlations are much lower." While the statement of imposing unit correlation is mathematically correct, it is based on the assumption that the risk categories can be linearly separated. If that assumption were not correct, as suggested by papers cited above, the linear correlations could actually be greater than one and lead to risk compounding.

Finally, Kuritzkes and Schuermann (2007) examine the distribution of earnings volatility for US bank holding companies with at least USD 1 billion in assets over the period from 1986.Q2 to 2005.Q1; specially, they examine the 99.9% tail of this distribution. Using a decomposition methodology based on the definition of net income, the authors find that market risk accounts for just 5% of total risk at the 99.9% level, while operational risk accounts for 12% of total risk. Using their risk measure of the lower tail of the earnings distribution, as measured by the return on risk-weighted assets, their calculations suggest that the ratio of the integrated risk measure to the sum of the disaggregated risk measures ranges from 0.53 through 0.63.

Conclusions

Academic studies have generally found that at a high level of aggregation, such as at the holding company level, the ratio of the risk measures for the unified approach to that of the separated approach is often less than one, i.e., risk diversification is prevalent and ignored by the separated approach. However, this approach often assumes that diversification is present. At a lower level of aggregation, such as at the portfolio level, this ratio is also often found to be less than one, but important examples arise in which risk compounding (i.e., a ratio greater than one) is found. These results suggest, at a minimum, that the assumption of risk diversification cannot be applied without questioning, especially for portfolios subject to both market and credit risk, regardless of where they reside on the balance sheet.

Recent literature on the systemic implications of the current regulatory capital requirements that aggregate capital requirements across risk types suggests that this compartmentalised approach can—at least in general—be argued to contribute to the amplification of systemic risk, which is counter to its intentions.

In terms of policy implications, the academic literature suggests that if we are able to divide risk types easily across the trading book and the banking book (as is assumed in the top-down studies), diversification benefits appear to be certain, and aggregation of capital requirements across the books is conservative. However, recent studies have shown that if this risk separation cannot be done completely, simple aggregation of compartmentalised measures may not be conservative and, in fact, may understate the total risk. Such an outcome would clearly be undesirable as the necessary amount of capital could be underestimated by a significant margin.

These conclusions seem to directly question whether separate capital requirements for the trading and banking books provide a reasonable path to setting the appropriate level of capital for the entire firm. If we retained the different capital treatments, attempts could be made to fully detail each type of risk within each book, and the subsequent aggregation might then be considered conservative. However, performing such an analysis within the current and traditional separation between a trading and a banking book would require important changes in operational procedures. An alternative approach might be to develop a system of book keeping and risk allocation that does not artificially assign positions into different books when its risk characteristics are interrelated.

6.7 RISK MANAGEMENT AND VALUE-AT-RISK IN A SYSTEMIC CONTEXT

Overview

In this section, we survey the research literature on the systemic consequences of individual risk management systems and regulatory capital charges that rely on them. At the time when the Basel Committee implemented the MRA in 1996, risk management and banking regulation still was a subject that had received relatively little attention in the academic literature. Perhaps the most important change brought to the Basel framework by the MRA was the ability for banks to use their own quantitative risk models for determining the capital requirements for market risk.

Both conceptually and procedurally, this amendment was a significant departure from the previous regulatory approaches to determine bank capital. The conceptual innovation was that the notion of risk on which the new regulation relied was much closer to the notions of risk that were in use in the financial, economic and statistical research literature. Procedurally the amendment amounted to an official recognition that financial institutions themselves are in the best positions to assess their risk exposures. The new regulatory approach seemed to suggest that using and relying on this knowledge might be the best way to cope with methodological problems of risk assessment in a rapidly changing economic environment.

At the time of the amendment and in the years after, the academic literature on risk management and regulation largely

accepted the conceptual reasoning behind the amendment and confined itself mostly to developing the technology of quantitative risk management itself. The discussion in the economics community remained sparse and largely sceptical.

Hellwig (1995, 1996) raised several important issues related to this new regulatory approach that did not take hold very much in the regulatory community but sound very modern in the current debate about the recent financial crises: Hellwig discussed incentive problems. Banks may find it desirable to bias their model development towards the goal of minimising capital. With hindsight, we know that the practice of determining capital based on VaR models helped large and international active banks to reduce greatly the amount of capital to be held against any given asset during the pre-crisis boom years. He also pointed out the difficulties related to using statistical techniques which work under the assumption of a stationary world in a non-stationary environment like financial markets. He also criticised the separation between market and credit risk while he acknowledged that quantitative models of integrated risk measurement are subject to the general problems outlined above.

During the discussion of the new Basel II framework, in May 2001, a group of academics at the Financial Markets Group (FMG) of the London School of Economics wrote a paper that raised a concern with respect to the use of value-at-risk that is more fundamental.[15] In the report's executive summary, there is a conclusion that calls into question the conceptual construction of the 1996 amendment: "The proposed regulations fail to consider the fact that risk is endogenous. Value-at-risk can destabilise and induce crashes when they would not otherwise occur."

In the current practice of risk management and regulation, these conclusions so far have only partly lead to a serious reconsideration of the framework initiated and extended more than a decade ago. In the current regulatory discussion, the general view seems to be that the conclusions from the financial crisis call for suitable expansions and amendments to the prevailing framework. In the meantime, the conclusions derived in the FMG paper have received more substantive underpinnings from academic research, both empirically and theoretically. The papers of Adrian and Shin (2008), the book of Shin (2008a) and joint work by Danielsson, Shin and Zigrand (2009) suggest that the use of value-at-risk models in regulation intended to function as a "fire extinguisher," function in practice rather like a "fire accelerant."[16] Rather than suggesting improving the VaR-based capital regulations by various refinements and amendments to the concepts in place, this literature suggests to abandon this approach and remove a VaR-based capital requirement from the regulatory framework. It should not be ignored, however, that some of the new regulatory initiatives will likely dampen procyclical effects in the future. The *stressed VaR* introduced by the July 2009 revisions of the Market Risk Framework is a case in point: its calculation is based on estimates from bad historical periods of the economy and so acts rather "through the cycle." Admittedly, the stressed VaR is only one addend of total trading book capital.

Although the literature for this section generally refers to VaR as the risk measure at issue, it is important to bear in mind that the term VaR should be interpreted here in a wide sense since the results generally do not depend on this specific risk measure.

In the following we give a brief outline of the main arguments and explain the boom and bust amplification mechanism identified in this literature. We then go through some of the policy conclusions suggested by this analysis.

Intermediation, Leverage and Value-at-Risk: Empirical Evidence

Adrian and Shin (2010) empirically investigated the relationship between leverage and balance sheet size of the five major US investment banks shortly before the financial crises. All these institutions meanwhile left the broker-dealer sector, either because they were taken over or went bankrupt or were converted to bank holding companies. A major reason why these institutions are particularly interesting is because they all show a very clear picture of how financial intermediation works in a capital markets-based financial system with active balance sheet management through risk management systems.

When an intermediary actively manages its balance sheet, leverage becomes procyclical because risk models and economic capital require balance sheet adjustments as a response to changes in financial market prices and measured risks. This relationship follows from simple balance sheet mechanics. The following example is taken from Shin (2008a, pp. 24 ff.) Assume a balance sheet is given with 100 in assets and a liability side which consists of 90 in debt claims and 10 in equity shares. Leverage is defined as the ratio of total assets to equity, 10 in our example. If we assume more generally that the market value of assets is A and make the simplifying assumption that the value of debt stays roughly constant at 90 for small changes in A, we see that total leverage is given by:

$$L = \frac{A}{A - 90}$$

Leverage is thus related inversely to the market value of total assets. When net worth increases, because A is rising, leverage

[15] See Danielsson et. al. (2001).

[16] See Hellwig (2009).

goes down, when net worth decreases, because A is falling, leverage increases.

Consider now what happens if an intermediary actively manages its balance sheet to maintain a constant leverage of 10. If asset prices rise by 1%, the bank can take on an additional amount of 9 in debt, its assets have grown to 110, its equity is 11, and the debt is 99. If asset values shrink by 1%, leverage rises. The bank can adjust its leverage by selling securities worth 9 and pay down a value 9 of debt to bring the balance sheet back to the targeted leverage ratio.

This kind of behaviour leads to a destabilising feedback loop, because it induces an *increase* in asset purchases as asset prices are rising and a *sale of* assets when prices are falling. Whereas the textbook market mechanism is self stabilising because the reaction to a price increase is a reduction in quantity demanded and an expansion in quantity supplied, and to a price decreases an expansion in quantity demanded and a contraction in quantity supplied, active balance sheet management reverses this self stabilising mechanism into a destabilising positive feedback loop.

Adrian and Shin (2010) document this positive relationship between total assets and leverage for all of the (former) big Wall Street investment banks. Furthermore, they produce econometric evidence that the balance sheet adjustments brought about by active risk management of financial institutions indeed has an impact on risk premiums and aggregate volatility in financial markets.

What Has All This to Do with VaR-Based Regulation?

Why would a bank target a constant leverage and what is the role of value-at-risk in all of this? The book of Shin (2008a) and the papers by Shin (2008b) and Adrian and Shin (2008) as well as by Danielsson, Shin and Zigrand (2009) explore this role in more detail.

If we consider the future value of bank assets A as a random variable, the value-at-risk (VaR) at a confidence level c is defined by

$$\Pr(A < A_0 - VaR) \leq 1 - c$$

The VaR is equal to the equity capital the firm must hold to be solvent with probability c. The economic capital is tied to the overall value-at-risk.

If a bank adjusts its balance sheet to target a ratio of value-at-risk to economic capital then bank capital to meet VaR is

$$K = \lambda \times VaR,$$

where λ is the proportion of capital to be held per total value-at-risk. This proportion may vary with time. Leverage is thus

$$L = \frac{A}{K} = \frac{1}{\lambda} \times \frac{A}{VaR}$$

Since VaR per value of assets is countercyclical, it directly follows that leverage is procyclical as the data in Adrian and Shin (2008) indeed show.[17]

The systemic consequences of this built-in risk limiting technology at the level of individual institutions works in the aggregate as an amplifier of financial boom and bust cycles. The mechanism by which the systemic amplification works is risk perception and the pricing of risk, even if all network effects and complex interconnectedness patterns in the financial system are absent.[18]

Consider intermediaries who run a VaR-based risk management system and start with a balance sheet consisting of risk-free debt and equity. Now an asset boom takes place, leading to an expansion in the values of securities. Since debt was risk-free to begin with, without any balance sheet adjustment, this leads to a pure expansion in equity. The VaR constraint is relaxed through the asset boom and creates new balance sheet capacity to take on more risky securities or increase its debt. The boom gets amplified by the portfolio decisions of the leveraged banking system.

Put differently, in a system of investors driven by a VaR constraint, investors' demand follows and amplifies the most recent price changes in the financial market. Price increases and balance sheet effects become intertwined through the active VaR-driven risk management of financial institutions.

Of course, the described mechanism also works on the way down. A negative shock drives down market values, tightening the VaR constraints of leveraged investors. These investors have to sell assets to reduce leverage to the new VaR constraint. By hardwiring VaR-driven capital management in banking regulation, a positive feedback loop with potent destabilising force both in booms and busts has been built into the financial system.

The mechanisms described in this section have been theoretically analysed in Shin (2008a), Danielsson, Shin, Zigrand (2009) theoretically and with explicit reference to value-at-risk. They are also central in the work of Geanakoplos (2009), although there the connection with VaR is not made explicit.

---

[17] This formal derivation of the procyclicality of VaR is taken directly from Shin (2008a).

[18] For this point, see also Geanakoplos (2009), who has shown in a theoretical model how risk-free debt may nevertheless give rise to fluctuations in leverage and risk pricing and thus create systemic spillover effects.

Conclusions

A literature stream on the systemic consequences of individual risk management systems as the basis of regulatory capital charges has found that the mechanical link between measured risks derived from risk models and historical data and regulatory capital charges can work as a systemic amplifier of boom and bust cycles.

The central mechanism that leads to this feedback loop works through the pricing of risk. In good times, when measured risks look benign, a financial institution that targets a regulatory capital requirement as a function of a model-based risk measure has slack capacity in its balance sheet that it can either use to buy additional risky assets or to increase its debt. This means that we have a mechanism where institutions are buying more risky assets when the price of these assets is rising and where they are buying less of these assets when prices are falling. The stabilising properties of the market mechanism are turned on their head. By this mechanic link of measured risk to regulatory capital a powerful amplifier of booms and busts is created at the system level counteracting the intention of the regulation to make the system as a whole safer.

It is important to recognise that while the current system may implement a set of rules that limit the risk taken at the level of individual institutions, the system may also enable institutions to take on more risk when times are good and thereby lay the foundations for a subsequent crisis. The very actions that are intended to make the system safer may have the potential to generate systemic risk in the system.

These results question a regulatory approach that accepts industry risk models as an input to determine regulatory capital charges. This critique applies in particular to the use of VaR to determine regulatory capital for the trading book but it questions also an overall trend in recent regulation.

The amplifying mechanism identified in this section will be at work no matter how sophisticated VaR becomes, whether it is replaced by more sophisticated risk measures, like expected shortfall, or whether it goes beyond the naïve categorisation of risk classes (market, credit and operational) towards a more integrated risk measurement. These changes generally do not address the problems raised by the papers reviewed in this section. One exception is the *stressed VaR* introduced in July 2009. This new component of trading book capital acts more "through the cycle" than the "normal" VaR. Still some argue that what may be needed is a less mechanical approach to capital adequacy that takes into account a system-wide perspective on endogenous risk. The academic literature has identified many potential shortcomings in the currently regulatory approach for bank capital but it has yet to develop an alternative approach that simultaneously satisfies all the (sometimes conflicting) regulatory policy objectives.

References

Acerbi, C (2002): "Spectral measures of risk: a coherent representation of subjective risk aversion," *Journal of Banking and Finance*, vol 26, no 7, pp 1505–1518.

Acerbi, C and G Scandolo (2008): "Liquidity risk theory and coherent measures of risk," *Quantitative Finance*, vol 8, no 7, pp 681–692.

Acerbi, C and D Tasche (2002): "On the coherence of expected shortfall," *Journal of Banking and Finance*, vol 26, pp 1487–1503.

Adrian, T and H S Shin (2010): "Liquidity and leverage," *Journal of Financial Intermediation*, vol 19, no 3, pp 418–437.

——— (2008), "Financial intermediary leverage and value at risk," Federal Reserve Bank of New York, staff report no 338.

Alessandri, P and M Drehmann (2010): "An economic capital model integrating credit and interest rate risk in the banking book," *Journal of Banking and Finance*, vol 34, pp 730–742.

Alexander, C and E Sheedy (2008): "Developing a stress testing framework based on market risk models," *Journal of Banking and Finance*, vol 32, no 10, pp 2220–2236.

Almgren, R and N Chriss (2001): "Optimal execution of portfolio transactions," *Journal of Risk*, vol 3, pp 5–39.

Amihud, Y (2002): "Illiquidity and stock returns: cross-section and time-series effects," *Journal of Financial Markets*, pp 31–56.

Aragones, J, C Blanco and K Dowd (2001): "Incorporating stress tests into market risk modeling," *Derivatives Quarterly*, pp 44–49.

Aramonte, S, M Rodriguez and J Wu (2010): "Portfolio value-at-risk: a dynamic factor approach," Federal Reserve Board.

Artzner, P F, J Delbaen, J Eber and D Heath (1999): "Coherent measures of risk," *Mathematical Finance*, 203–228.

Bakshi, G and G Panayotov (2010): "First-passage probability, jump models, and intra-horizon risk," *Journal of Financial Economics*, vol 95, pp 20–40.

Balaban, E, J Ouenniche and D Politou (2005), "A note on return distribution of UK stock indices," *Applied Economics Letters*, vol 12, pp 573–576.

Bangia, A, F X Diebold, T Schuermann and J D Stroughair (1999a): "Modeling liquidity risk, with implication for traditional

market risk measurement and management," Wharton working paper.

——— (1999b): "Liquidity on the outside," *Risk*, December, pp 68–73.

Barnhill, T and W Maxwell (2002): "Modeling correlated interest rate, exchange rate, and credit risk in fixed income portfolios," *Journal of Banking and Finance*, vol 26, pp 347–374.

Barnhill, T M, P Papapanagiotou and L Schumacher (2000): "Measuring integrated market and credit risks in bank portfolios: an application to a set of hypothetical banks operating in South Africa," IMF Working Paper #2000–212.

Barone-Adesi, G, F Bourgoin and K Giannopoulos (1998): "Don't look back," *Risk*, November, pp 100–103.

Basak, S and A Shapiro (2001): "Value-at-risk-based risk management: optimal policies and asset prices," *The Review of Financial Studies*, pp 371–405.

Basel Committee on Banking Supervision (2009a): *Findings on the interaction of market and credit risk*, Working Paper no 16, Basel.

——— (2009b): *Revisions to the Basel II market risk framework*, http://www.bis.org/publ/bcbs158.pdf, July.

Berkowitz, J (2000a): "A coherent framework for stress-testing," *Journal of Risk*, vol 2, pp 1–11.

——— (2000b): "Incorporating liquidity risk into value-at-risk models," Working Paper, University of Houston, September.

——— (2001): "Testing density forecasts, with applications to risk management," *Journal of Business and Economic Statistics*, vol 19, no 4, pp 465–474.

Berkowitz, J and J O'Brien (2002): "How accurate are value-at-risk models at commercial banks?," *Journal of Finance*, vol 57, pp 1093–1111.

Berkowitz, J, P F Christoffersen and D Pelletier (2010): "Evaluating value-at-risk models with desk-level data," *Management Science*.

Berry, R P (2009): "Back testing value-at-risk," *Investment Analytics and Consulting*, September.

Bervas, A (2006): "Market liquidity and its incorporation into risk management," Banque de France Financial Stability Review.

Botha, A (2008): "Portfolio liquidity-adjusted value at risk," SAJEMS NS, pp 203–216.

Boudoukh, J, M Richardson and R Whitelaw (1998): "The best of both worlds," *Risk*, May, pp 64–67.

Breuer, T, M Jandačka, K Rheinberger and M Summer (2008): "Compounding effects between market and credit risk: the case of variable rate loans," in A Resti (ed), *The Second Pillar in Basel II and the Challenge of Economic Capital*, London: Risk Books.

——— (2009): "How to find plausible, severe, and useful stress scenarios," *International Journal of Central Banking*, September.

——— (2010a): "Does adding up of economic capital for market and credit risk amount to a conservative risk estimate?," *Journal of Banking and Finance*, vol 34, pp 703–712.

Breuer, T, M Jandačka, J Mencia and M Summer (2010b): "A systematic approach to multi-period stress testing of portfolio credit risk," Bank of Spain Working Paper, June.

Brockmann, M and M Kalkbrener (2010): "On the aggregation of risk," *Journal of Risk*, vol 12, no 3.

Campbell, S D (2005): "A review of backtesting and backtesting procedures," FEDS Working Paper Series.

Christoffersen, P (1998): "Evaluating interval forecasts," *International Economic Review*, pp 841–862.

Christoffersen, P and F Diebold (2000): "How relevant is volatility forecasting for financial risk management?," *The Review of Economics and Statistics*, vol 82, no 1, pp 12–22.

Christoffersen, P, F Diebold and T Schuermann (1998): "Horizon problems and extreme events in financial risk management," *FRBNY Economic Policy Review*, October 1998, pp 109–118.

Crouhy, M, D Galai and R Mark (2003): *Risk management*, McGraw-Hill.

Cuenot, S, N Masschelein, M Pritsker, T Schuermann and A Siddique (2006): "Interaction of market and credit risk: framework and literature review," manuscript, Basel Committee Research Task Force Working Group.

Danielsson, J (2002): "The emperor has no clothes: limits to risk modelling," *Journal of Banking and Finance*, vol 26, pp 1273–1296.

Danielsson, J, P Embrechts, C Goodhart, C Keating, F Muennich, O Renault and H Shin (2001): *An academic response to Basel II*.

Danielsson, J, B N Jorgensen, G Samorodnitsky, M Sarma and C G de Vries (2005): "Subadditivity re-examined: the case for value-at-risk," FMG Discussion Papers, London School of Economics.

Danielsson, J and J Zigrand (2006): "On time-scaling of risk and the square-root-of-time rule," *Journal of Banking and Finance*, vol 30, pp 2701–2713.

Danielsson, J, H Shin and J Zigrand (2009): "Risk appetite and endogenous risk," mimeo, http://www.princeton.edu/~hsshin/www/riskappetite.pdf.

Degen, M, P Embrechts and D Lambrigger (2007): "The quantitative modeling of operational risk: between g-and-h and EVT," ASTIN Bulletin, vol 37, no 2, pp 265–291.

Delbaen, F (2002), "Coherent risk measures," lecture notes for University of Pisa lectures, draft.

Diebold, F, A Hickman, A Inoue and T Schuermann (1998), "Scale models," *Risk*, no 11, pp 104–107.

Dimakos, X and K Aas (2004): "Integrated risk modeling," *Statistical Modelling*, vol 4, no 4, pp 265–277.

Drehmann, M, S Sorensen and M Stringa (2010): "The integrated impact of credit and interest rate risk on banks: a dynamic framework and stress testing application," *Journal of Banking and Finance*, vol 34, pp 713–742.

Dunn, G (2009), "Modelling market risk," UK FSA.

Egloff, D, M Leippold and S Jöhri (2005): "Optimal importance sampling for credit portfolios with stochastic approximation," working paper, http://ssrn.com/abstract=1002631.

Engle, R F (2002): "Dynamic conditional correlation—a simple class of multivariate GARCH models," *Journal of Business and Economic Statistics*, pp 339–350.

Engle, R F and R Ferstenberg (2006): "Execution risk," NBER working paper 12165.

Engle, R F and B Kelly (2009): "Dynamic equicorrelation," Stern School of Business.

Engle, R F and K F Kroner (1995): "Multivariate simultaneous generalized ARCH," *Econometric Theory*, pp 122–150.

Engle, R F, N Shephard and K Sheppard (2007): "Fitting and testing vast dimensional time-varying covariance models," New York University.

Finger, C (2009), "IRC comments," *RiskMetrics Group Research Monthly*, February.

Francois-Heude, A and P Van Wynendaele (2001): "Integrating liquidity risk in a parametric intraday VaR framework," Université de Perpignan, France, Facultés Universitaires Catholiques de Mons, Belgium.

Franke, J, W K Härdle and C M Hafner (2008): "Value at risk and backtesting," in *Statistics of Financial Markets*, pp 321–332, Berlin, Heidelberg: Springer.

Geanakoplos, J (2009): "The leverage cycle," Cowles Foundation Discussion Paper no 1715R.

Gennotte, G and H Leland (1990): "Market liquidity, hedging, and crashes," *American Economic Review*, vol 80, no 5, pp 999–1021.

Gourier, E, W Farkas and D Abbate (2009): "Operational risk quantification using extreme value theory and copulas: from theory to practice," *Journal of Operational Risk*, vol 4, no 3, pp 3–26.

Grundke, P (2005): "Risk measurement with integrated market and credit portfolio models," *Journal of Risk*, vol 7, pp 63–94.

Häberle, R and P Persson (2000): "Incorporating market liqudity constraints in VaR," *Bankers Markets & Investors*, vol 44, 01/01/2000.

Hallerbach, W G (2003): "Decomposing portfolio value-at-risk: a general analysis," *Journal of Risk*, vol 5, no 2, pp 1–18.

Hellwig, M (1995), "Systemic aspects of risk management in banking and finance," *Swiss Journal of Economics and Statistics*, vol 131, no 4/2, pp 723–737.

——— (1996), "Capital adequacy rules as instruments for the regulation of banks," *Swiss Journal of Economics and Statistics*, vol 132, no 4/2, pp 609–612.

——— (2009): "Brandbeschleuniger im Finanzsystem," *Max Planck Research*, vol 2, pp 10–15.

Hisata, Y and Y Yamai (2000): "Research toward the practical application of liquidity risk evaluation methods," *Monetary and Economic Studies*, pp 83–128.

Huberman, G and W Stanzl (2005), "Optimal liquidity trading," *Review of Finance*, vol 9, no 2, pp 165–200.

International Accounting Standard 39, *Financial instruments: recognition and measurement*, last version 31 December 2008.

Jarrow, R and S Turnbull (2000): "The intersection of market and credit risk," *Journal of Banking and Finance*, vol 24, no 1–2, pp 271–299.

Jarrow, R and A Subramanian (2001): "The liquidity discount," *Mathematical Finance*, pp 447–474.

Jarrow, R and Ph Protter (2005): "Liquidity risk and risk measure computation," working paper.

Jobst, N J, G Mitra and S A Zenios (2006): "Integrating market and credit risk: a simulation and optimisation perspective," *Journal of Banking and Finance*, vol 30, pp 717–742.

Jobst, N J S A Zenios (2001): "The tail that wags the dog: integrating credit risk in asset portfolios," *Journal of Risk Finance*, vol 3, pp 31–43.

J.P. Morgan (1996): *RiskMetrics Technical Document*, http://www.riskmetrics.com/system/files/private/td4e.pdf.

Kalkbrener, M, H Lotter and L Overbeck (2004): "Sensible and efficient capital allocation for credit portfolios," *Risk*, January, pp 19–24.

Kalkbrener, M, A Kennedy and M Popp (2007): "Efficient calculation of expected shortfall contributions in large credit portfolios," *Journal of Computational Finance*, vol 11, no 2, pp 45–77.

Kaufman, R (2004), *Long-term risk management*, PhD thesis, ETH Zurich.

Kerkhof, J and B Melenberg (2004): "Backtesting for risk-based regulatory capital," *Journal of Banking and Finance*, vol 28, no 8, pp 1845–1865.

Kritzman, M and D Rich (2002): "The mismeasurement of risk," *Financial Analysts Journal*, vol 58, no 3, pp 91–99.

Kupiec, P (1998), "Stress testing in a value-at-risk framework," *Journal of Derivatives*, pp 7–24.

——— (1995), "Techniques for verifying the accuracy of risk measurement models," *Journal of Derivatives*, pp 73–84.

——— (2007), "An integrated structural model for portfolio market and credit risk," manuscript, Federal Deposit Insurance Corporation.

Kuritzkes, A and T Schuermann (2007): "What we know, don't know and can't know about bank risk: a view from the trenches," forthcoming in F X Diebold, N Doherty and R J Herring (eds), *The Known, The Unknown and The Unknowable in Financial Risk Management*, Princeton University Press.

Kuritzkes, A, T Schuermann and S M Weiner (2003): "Risk measurement, risk management and capital adequacy in financial conglomerates," Wharton Financial Institutions Center Working Paper #2003-02.

Lawrence, C and G Robinson (1997): "Liquidity, dynamic hedging and VaR," in: *Risk management for financial institutions*, Risk Publications, London, pp 63–72.

Le Saout, E (2002): "Intégration du risque de liquidité dans les modèles de valeur en risque," *Bankers Markets & Investors*, vol 61, November–December, pp 15–25.

Lopez, J (1999): "Methods for evaluating value-at-risk estimates," *Federal Reserve Bank of San Francisco Review*, 2, pp 3–17.

——— (2005), "Stress tests: useful complements to financial risk models," *FRBSF Economic Letter*, pp 119–124.

McNeil, A, R Frey and P Embrechts (2005): *Quantitative risk management*, Princeton.

Perignon, C and D Smith (2006): "The level and quality of value-at-risk disclosure by commercial banks," Simon Fraser University.

Platen, E and M Schweizer (1998): "On feedback effects from hedging derivatives," *Mathematical Finance*, vol 8, pp 67–84.

Pritsker, M (2006), "The hidden dangers of historical simulation," *Journal of Banking and Finance*, vol 30, no 2, pp 561–582.

Provizionatou, V, S Markose and O Menkens (2005): "Empirical scaling rules for value-at-risk," University of Essex.

Qi, J and W L Ng (2009): "Liquidity adjusted intraday value at risk," *Proceedings of the World Congress of Engineering*.

Rogers, L C G and S Singh (2005): "Option pricing in an illiquid market," Technical Report, University of Cambridge.

Rosenberg, J and T Schuermann (2006): "A general approach to integrated risk management with skewed, fat-tailed risks," *Journal of Financial Economics*, vol 79, pp 569–614.

Saunders, A and I Walter (1994): *Universal banking in the United States: what could we gain? What could we lose?*, Oxford University Press, New York.

Schönbucher, P J and P Wilmott (2000): "The feedback effects of hedging in illiquid markets," *SIAM Journal on Applied Mathematics*, vol 61, pp 232–272.

Shang, D (2009): "ARCH-based value-at-risk with heavy-tailed errors," London School of Economics.

Shin, H S (2008a): "Risk and liquidity," Clarendon Lectures, Oxford University Press, forthcoming.

——— (2008b), "Risk and liquidity in a system context," *Journal of Financial Intermediation*, vol 17, no 3, pp 315–329.

Sircar, K R and G Papanicolaou (1998): "Generalized Black-Scholes models accounting for increased market volatility from hedging strategies," *Applied Mathematical Finance*, vol 5, no 1, pp 45–82.

Smithson, C and L Minton (1996): "Value-at-risk," *Risk*, September, pp 38–39.

Stange, S and C Kaserer (2008): "Why and how to integrate liquidity risk into a VaR framework," CEFS Working Paper.

Stiroh, K J (2004): "Diversification in banking: Is noninterest income the answer?," *Journal of Money, Credit and Banking*, vol 36, no 5, pp 853–82.

Subramanian, A (2008): "Optimal liquidation by a large investor," *SIAM Journal of Applied Mathematics*, vol 68, no 4, pp 1168–1201.

Sy, W (2006): "On the coherence of VaR risk measures for Lévy Stable distributions," Australian Prudential Regulation Authority.

Wang, S (2001): "A risk measure that goes beyond coherence," University of Waterloo, Institute of Insurance and Pension.

Wang, S, V Young and H Panjer (1997): "Axiomatic characterization of insurance prices," *Insurance: Mathematics and Economics*, vol 21, pp 173–183.

Wong, W K (2008): "Backtesting trading risk of commercial banks using expected shortfall," *Journal of Banking and Finance*, vol 32, no 7, pp 1404–1415.

Wu, G and Z Xiao (2002): "An analysis of risk measures," *Journal of Risk*, vol 4, no 4, pp 53–75.

Wu, L (2009): "Incorporating liquidity risk in value-at-risk based on liquidity adjusted returns," Southwestern University of Economics and Finance.

Wylie, J, Q Zhang and T Siu (2010): "Can expected shortfall and value-at-risk be used to statistically hedge options?," *Quantitative Finance*, vol 10, no 6, pp 575–583.

Yamai, Y and Y Yoshiba (2005): "Value-at-risk versus expected shortfall: a practical perspective," *Journal of Banking and Finance*, vol 29, pp 997–1015.

Zheng, H (2006): "Interaction of credit and liquidity risks: modelling and valuation," *Journal of Banking and Finance*, vol 30, pp 391–407.

ANNEX

Table 6.1 Summary of "Bottom-Up" Risk Aggregation Papers in the Survey

Research Paper	Portfolio Analysed	Horizon	Risk Measure Used	Ratio of Unified Risk Measure to Sum of Compartmentalised Risk Measures
Breuer, Jandačka, Rheinberger and Summer (2010a)	Hypothetical portfolios of foreign-exchange denominated loans of rating: BBB+ B+	One year	Expected shortfall at: the 1% level the 0.1% level the 1% level the 0.1% level	 1.94 8.22 3.54 7.59
Breuer, Jandačka, Rheinberger and Summer (2008)	Hypothetical portfolios of variable rate loans of rating: BBB+ B+	One year	Expected shortfall at: the 1% level the 0.1% level the 1% level the 0.1% level	 1.11 1.16 1.06 1.10
Grundke (2005)	Hypothetical portfolios of loans with various credit ratings, asset value correlations, distributional assumptions, and correlations between the risk-free rate, credit spreads and firm asset returns	Three years	VaR at: the 1% level the 0.1% level	 0.07–0.97 0.09–1.00
Kupiec (2007)	Hypothetical portfolio of corporate loans with various rating categories calibrated to historical data	Six months	Portfolio losses at funding cost levels consistent with: AAA rating BBB rating	 0.60–3.65 0.61–3.81
Drehmann, Sorensen and Stringa (2010)	Hypothetical UK bank	Three years	Decline in capital over the horizon	2.5
Alessandri and Drehmann (2008)	Hypothetical UK bank	One year	Value-at-risk at: the 1% level the 0.1% level	 0.03 0.50

Table 6.2 Summary of "Top-Down" Risk Aggregation Papers in the Survey

Research Paper	Portfolio Analysed	Horizon	Risk Measure Used	Ratio of Unified Risk Measure to Sum of Compartmentalised Risk Measures
Dimakos and Aas (2004)	Norwegian financial conglomerate	—	Total risk exposure at: the 1% level the 0.1% level	 0.90 0.80
Rosenberg and Schuermann (2008)	Hypothetical, internationally-active financial conglomerate	One year	Value-at-risk based on a normal copula at the 0.1% level. (Note: similar results using expected shortfall.)	0.42–0.89 based on different correlation assumptions between market, credit and operational risk.
Kuritzkes, Schuermann and Weiner (2003)	Representative Dutch bank	One year	Economic capital	0.72–0.85 based on different correlation assumptions between market, credit and operational risk.
Kuritzkes and Schuermann (2007)	US banking system from 1986.Q2 through 2005.Q1	—	Tail quantile of the earnings distribution at: the 1% level the 0.1% level	 0.63 0.63

7

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.
- Relate correlation risk 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.

7.1 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.

7.2 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 p.
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

7.3 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

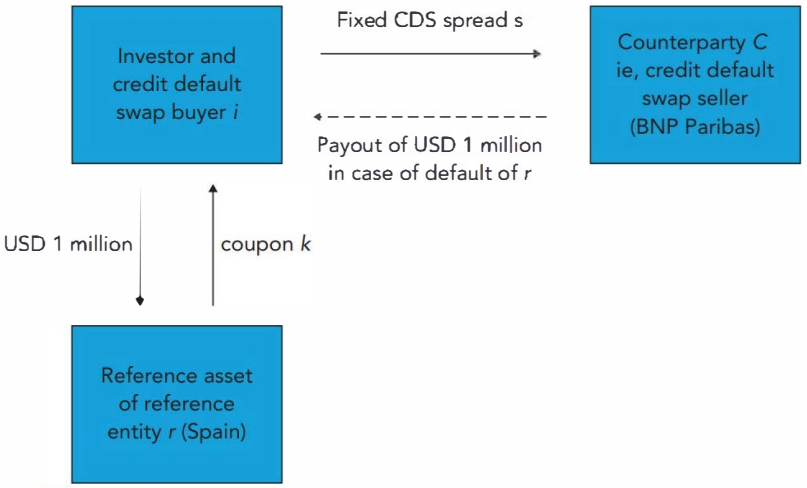


Figure 7.1 **An investor hedging their Spanish bond exposure with a CDS.**

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 7.1).

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 8, 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 7.1.

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 7.2.

From Figure 7.2, 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 7.2 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 7.1) in the CDS minus the yield of a riskless bond. See Meissner 2005, for details.

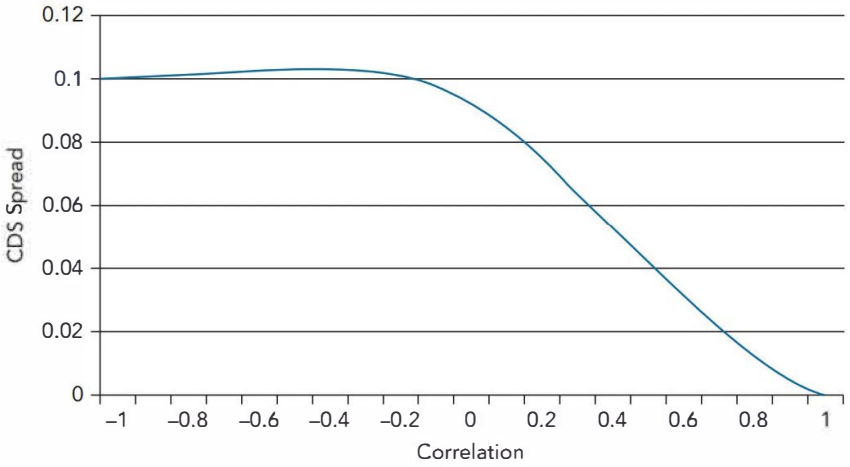


Figure 7.2 **CDS spread s of a hedged[2] bond purchase (as displayed in Figure 7.1) with respect to the default correlation between the reference entity *r* and the counterparty c.**

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 7.2 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.

7.4 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 7.1.

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:

$$\mu_P = w_X\mu_X + w_Y\mu_Y \tag{7.1}$$

where $w_X + w_Y = 1$

The standard deviation of returns, called *volatility*, is derived for asset *X* with equation:

$$\sigma_X = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \mu_X)^2} \tag{7.2}$$

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 7.2 can be computed with = stdev in Excel and std in MATLAB. From our example in Table 7.1, we find that $\sigma_X = 44.51\%$ and $\sigma_Y = 47.58\%$.

Table 7.1 **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.

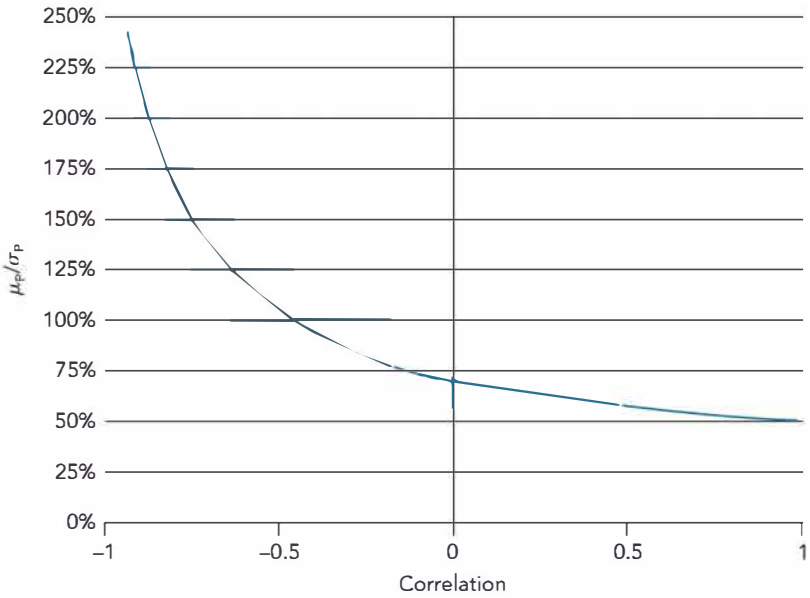


Figure 7.3 **The negative relationship of the portfolio return/portfolio risk ratio μ_P/σ_P with respect to the correlation ρ of the assets in the portfolio (input data are from Table 7.1).**

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:

$$COV_{XY} = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \mu_X)(y_t - \mu_Y) \tag{7.3}$$

For our example in Table 7.1 we derive $COV_{XY} = -0.1567$. Equation (7.3) 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:

$$\rho_{XY} = \frac{COV_{XY}}{\sigma_X \sigma_Y} \tag{7.4}$$

For our example in Table 1, $\rho_{XY} = -0.7403$, showing that the returns of assets X and Y are highly negatively correlated. Equation (7.4) is "correl" in Excel and "corrcoef" in MATLAB. For the derivation of the numerical examples of equations (7.2) to (7.4) and more information on the covariances see the appendix of Chapter 1 and www.dersoft.com/matrixprimer.xlsx, sheet "Covariance Matrix".

We can calculate the standard deviation for our two-asset portfolio P as:

$$\sigma_P = \sqrt{w_X^2\sigma_X^2 + w_Y^2\sigma_Y^2 + 2w_X w_Y COV_{XY}} \tag{7.5}$$

With equal weights, ie, $w_X = w_Y = 0.5$, the example in Table 7.1 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 (7.1) and (7.5). In Figure 7.3 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 7.3 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 (7.1) to (7.5) are derived within the framework of the Pearson correlation approach.

"Only by great risks can great results be achieved"

Xerxes

7.5 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.

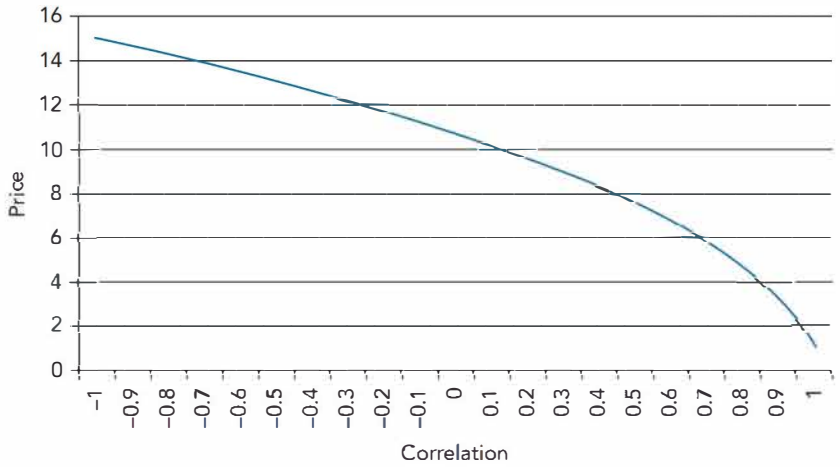


Figure 7.4 **Exchange option price with respect to correlation of the assets in the portfolio.**

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 (7.5):

$$\sigma_E = \sqrt{\sigma_A^2 + \sigma_B^2 - 2COV_{AB}} \qquad (7.5a)$$

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 7.4.

From Figure 7.4 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 7.2 and 7.3, the correlation approach underlying Figure 7.4 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 7.4.

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 7.5.

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 7.5 is the correlation between the assets that actually occur during the time of the swap. It is calculated as:

$$\rho_{realised} = \frac{2}{n^2 - n}\sum_{i>j}\rho_{i,j} \tag{7.6}$$

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:

$$N(\rho_{realised} - \rho_{fixed}) \tag{7.7}$$

where N is the notional amount. Let's look at an example of a correlation swap.

Example 7.1

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 7.2.

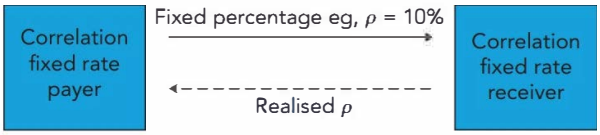


Figure 7.5 A correlation swap with a fixed 10% correlation rate.

The average correlation between the three assets is derived by equation (7.6). We only apply the correlations in the shaded area from Table 7.2, 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 (7.7), 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 7.8, as well as in Chapter 8, 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 8, there is a positive relationship between correlation and volatility.

Table 7.2 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:

$$VaR_P = \sigma_P \alpha \sqrt{x} \tag{7.8}$$

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:

$$\sigma_P = \sqrt{\beta_h C \beta_v} \tag{7.9}$$

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 7.2

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 (7.9) 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:

$$\text{Cov}_{11} = \rho_{11}\sigma_1\sigma_1 = 1 \times 0.02 \times 0.02 = 0.0004^{8} \quad (7.10)$$

$$\text{Cov}_{12} = \rho_{12}\sigma_1\sigma_2 = 0.7 \times 0.02 \times 0.01 = 0.00014$$

$$\text{Cov}_{21} = \rho_{21}\sigma_2\sigma_1 = 0.7 \times 0.01 \times 0.02 = 0.00014$$

$$\text{Cov}_{22} = \rho_{22}\sigma_2\sigma_2 = 1 \times 0.01 \times 0.01 = 0.0001$$

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 (7.9). 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 (7.8) from Excel as =normsinv(0.99) = 2.3264, or MATLAB as norminv(0.99) = 2.3264.

Following equation (7.8), 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 (7.3). In Equation (7.3) we derived the covariance "from scratch", inputting the return values and means. In Equation (7.10) 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 7.6 shows the impact.

As expected, we observe from Figure 7.6 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.

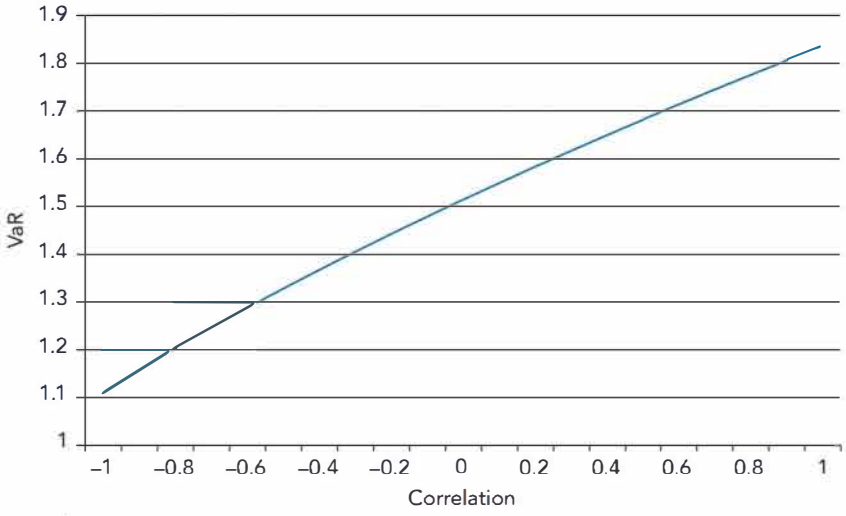


Figure 7.6 **VaR of the two-asset portfolio of Example 7.2 with respect to correlation ρ.**

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.

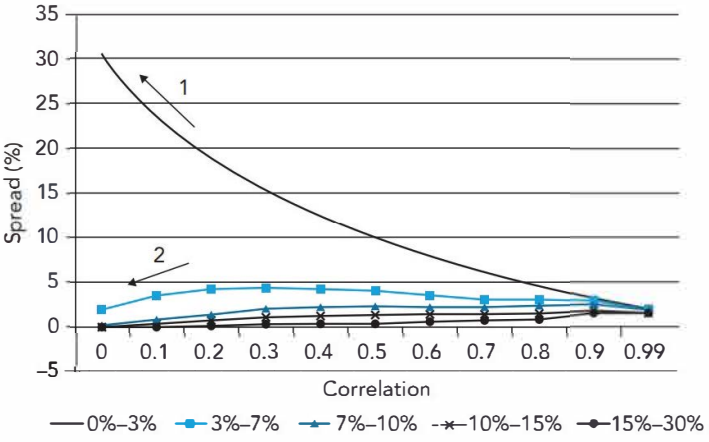


Figure 7.7 **CDO tranche spreads with respect to correlation between the assets in the CDO.**

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 7.7 shows the dilemma. Hedge funds had invested in the equity tranche[10] (0% to 3% in Figure 7.7) 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 7.7). However, as we can see from Figure 7.7, 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.

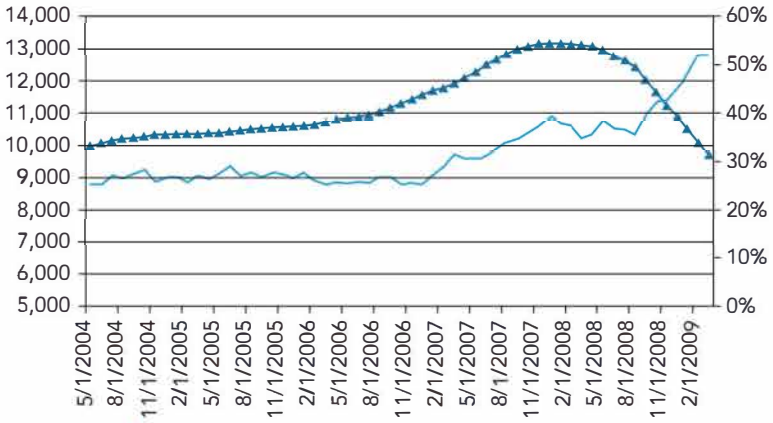


Figure 7.8 Relationship between the Dow (graph with triangles, numerical values on left axis) and correlation between the stocks in the Dow (numerical values on right axis) before and during the systemic 2007–09 global financial crisis; one-year moving average of monthly correlations.

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 7.8. 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 7.7. 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 7.1. 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 7.1. 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.

7.6 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 7.3 shows the default correlation from 1981 to 2001 of 6,907 companies, of which 674 defaulted.

The default correlations in Table 7.3 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 7.3 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 7.4.

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 7.4 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 7.4. 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 7.4 with the famous (now infamous) copula model, which will be discussed in

[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 7.3 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 7.4 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.

Chapter 7. This will allow us to 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

7.7 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 7.8 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 7.8 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 7.8 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 8, 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.

7.8 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 7.3.

Naturally, concentration and correlation risk are closely related. Let's verify this in an example.

Example 7.3

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 7.9.

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 7.10.

The exact joint default probability depends on the correlation model and correlation parameter values, which will be discussed in Chapters 4 to 8. 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 7.9.

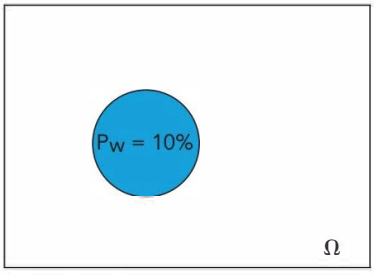


Figure 7.9 **Probability space for the default probability of a single loan to W.**

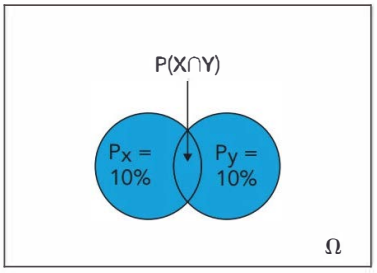


Figure 7.10 **Probability space for loans to companies X and Y.**

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 7.11.

Hence, from Figures 7.9 to 7.11 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 (7.11) gives the joint probability of default for the two binomial events:

$$P(X \cap Y) = \rho_{XY}\sqrt{P_X(1 - P_X)\,P_Y(1 - P_Y)} + P_X P_Y \quad (7.11)$$

where ρ_{XY} is the correlation coefficient and

$$\sqrt{P_X(1 - P_X)} \quad (7.12)$$

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 (7.12), 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 (7.11) 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 7.9 with $P_X = P_Y = 10\%$ as the probability event.

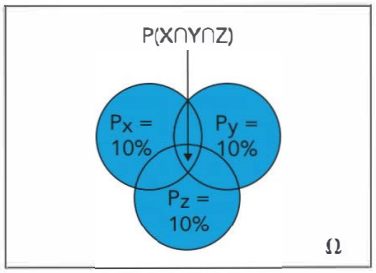


Figure 7.11 **Probability space for loans to companies X, Y and Z.**

If we now decrease the correlation coefficient, we can see from equation (7.11) 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 (7.11) 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.

7.9 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 AI 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. See Chapter 3 for a more detailed discussion on the Pearson correlation model.

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

7.1 What two types of financial correlations exist?

7.2 What is "wrong-way correlation risk" or for short "wrong-way risk"?

7.3 Correlations can be non-monotonous. What does this mean?

7.4 Correlations are critical in many areas in finance. Name five.

7.5 High diversification is related with low correlation. Why is this considered one of the few "free lunches" in finance?

7.6 Create a numerical example and show why a lower correlation results in a higher return/risk ratio.

7.7 What is "correlation trading"?

7.8 What is "pairs trading"?

7.9 Name three correlation options, in which a lower correlation results in a higher option price.

7.10 Name one correlation option where a lower correlation results in a lower option price.

7.11 Create a numerical example of a two-asset portfolio and show that lower correlation coefficient leads to a lower VaR number.

7.12 Why do correlations typically increase in a systemic market crash?

7.13 In 2005, a correlation crisis with respect to CDOs occurred that led to the default of several hedge funds. What happened?

7.14 In the global financial crisis 2007–09, many investors in the presumably safe super-senior tranches got hurt. What exactly happened?

7.15 What is the main objective of the Basel III accord?

7.16 The Basel accords have no legal authority. So why do most developed countries implement them?

7.17 How is correlation risk related to market risk and credit risk?

7.18 How is correlation risk related to systemic risk and concentration risk?

7.19 How can we measure the joint probability of occurrence of a binomial event as default or no-default?

7.20 Can it be that two binomial events are negatively correlated but they have a positive probability of joint default?

7.21 What is value-at-risk (VaR) and credit value-at-risk (CVaR)? How are they related?

7.22 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?

7.23 What do the terms "measure of association" and "measure of dependence" refer to in statistics?

8 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.

8.1 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 8.1 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 8.2 shows the volatility of the averaged monthly correlations. For the calculation of volatility, see Chapter 7.

From Figures 8.1 and 8.2, we observe the somewhat erratic behaviour of Dow correlation levels and volatility. However, Table 8.1 reveals some expected results:

From Table 8.1, 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 8.1. In addition, we had already displayed in Chapter 7, Figure 7.8, that correlation

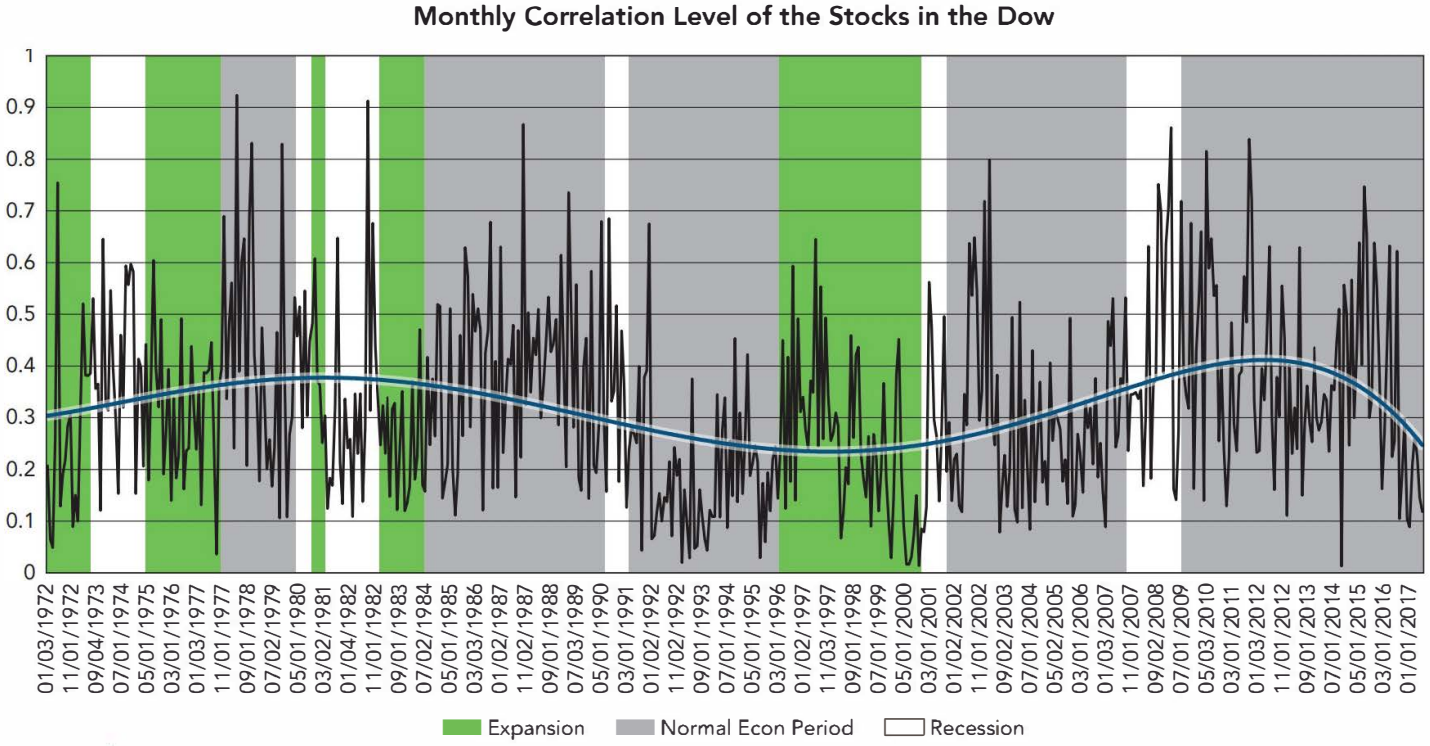


Figure 8.1 Average correlation of monthly 30 × 30 Dow stock return bins. The light grey background displays an expansionary economic period, the medium gray background a normal economic period and the white background represents a recession. The horizontal line shows the polynomial trendline of order 4.

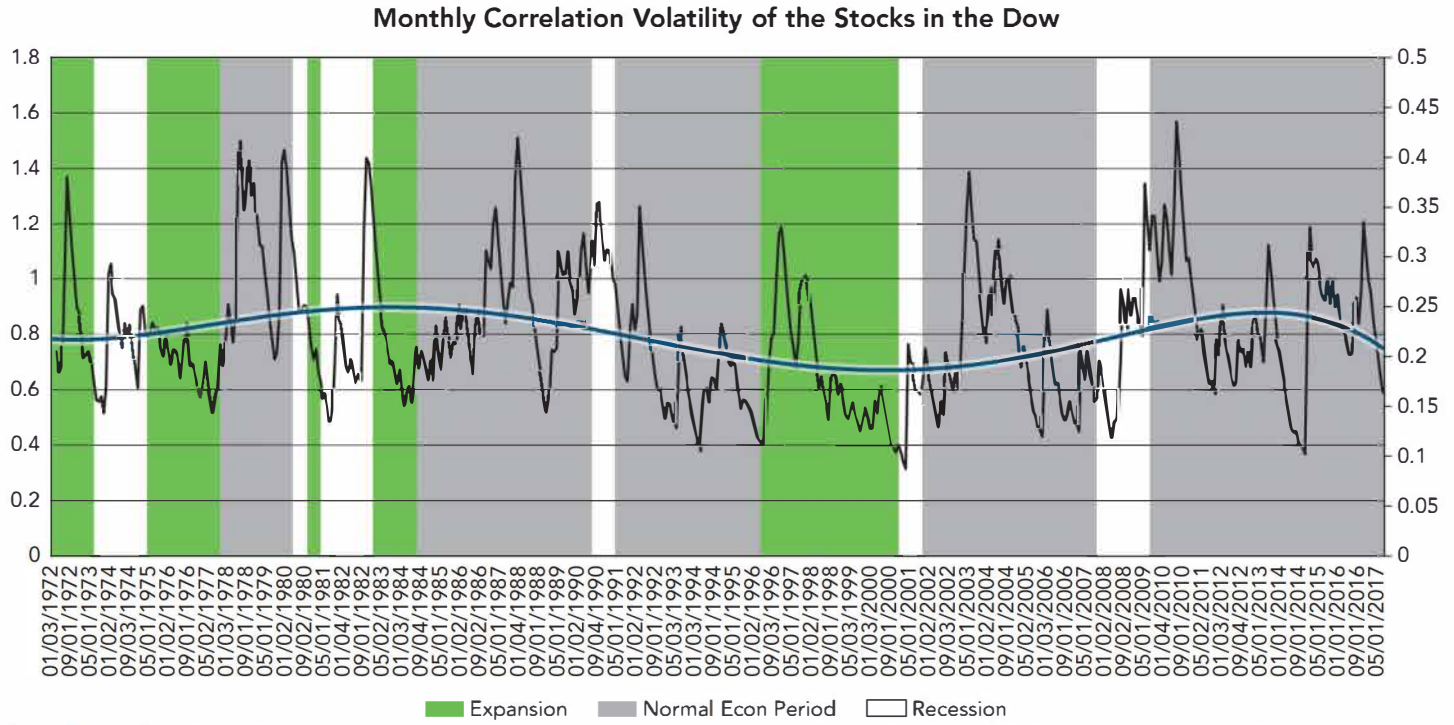


Figure 8.2 Correlation volatility of the average correlation of monthly 30 × 30 Dow stock return bins with respect to the state of the economy. The horizontal line shows the polynomial trendline of order 4.

Table 8.1 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%

levels increased sharply in the great recession from 2007 to 2009. In a recession, macroeconomic factors seem to dominate idiosyncratic factors, leading to a downturn across multiple stocks.

A further expected result in Table 8.1 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 8.1 displays the higher correlation risk in bad economic times, which traders and risk managers should consider in their trading and risk management.

From Table 8.1, we observe a generally positive relationship between correlation level and correlation volatility. This is verified in more detail in Figure 8.3.

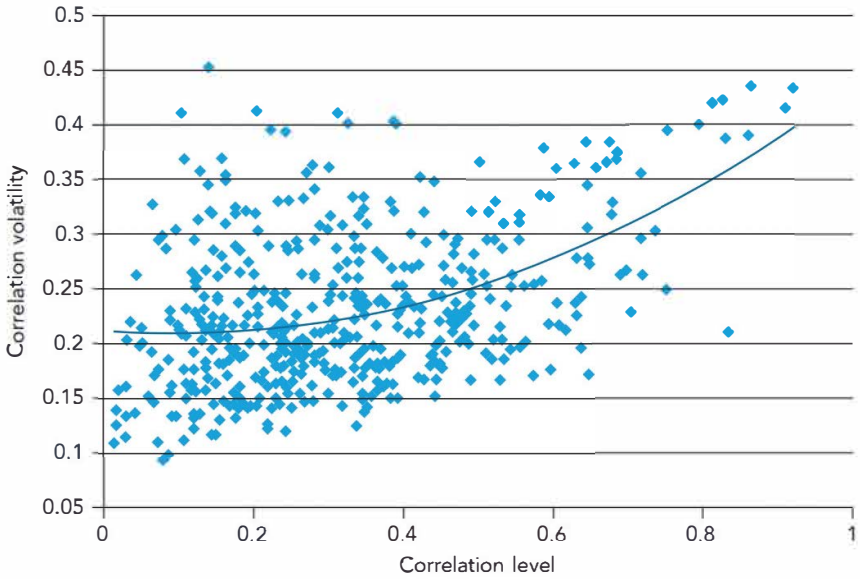


Figure 8.3 Positive relationship between correlation level and correlation volatility with a polynomial trendline of order 2 (data from 1972 to 2017).

8.2 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

$$\frac{\partial(S_t - S_{t-1})}{\partial S_{t-1}} < 0 \tag{8.1}$$

where

S_t: Price at time t

S_{t-1}: Price at the previous point in time $t - 1$

∂: Partial derivative coefficient

Equation (8.1) 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:

$$S_t - S_{t-1} = a(\mu_s - S_{t-1})\Delta t + \sigma_s \varepsilon \sqrt{\Delta t} \tag{8.2}$$

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 (8.2), $\sigma_s \varepsilon \sqrt{\Delta t}$.

For ease of explanation, let's assume $\Delta t = 1$. Then, from equation (8.2), 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 (8.2) without stochasticity reduces to

$$S_t - S_{t-1} = a(\mu_s - S_{t-1}) \tag{8.3}$$

or

$$S_t - S_{t-1} = a\mu_s - aS_{t-1} \tag{8.4}$$

To find the mean reversion rate "a", we can run a standard regression analysis of the form

$$Y = \alpha + \beta X$$

Following equation (8.4) we are regressing $S_t - S_{t-1}$ with respect to S_{t-1}:

$$\underbrace{S_t - S_{t-1}}_{\gamma} = \underbrace{a\mu_s}_{\alpha} - \underbrace{aS_{t-1}}_{\beta X} \tag{8.5}$$

Importantly, from equation (8.5), we observe that the regression coefficient β is equal to the negative mean-reversion parameter "a".

We now run a regression of equation (8.5) 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 8.4.

The regression function in Figure 8.4 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%

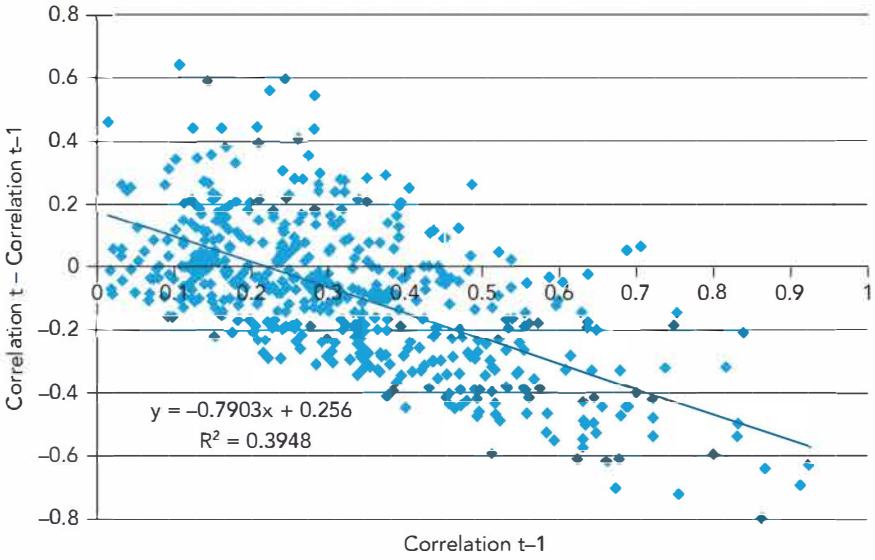


Figure 8.4 **Regression function (2.5) for 534 monthly average Dow stock return correlations from 1972 to 2017.**

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 8.1. 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 8.1: 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 (8.3) or (8.4)?

Solving equation (8.3) 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 (8.3) with the mean reversion rate of 79.03%, we expect the correlation in March 2017 to be 31.07%.[1]

8.3 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

$$AC(\rho_t, \rho_{t-1}) = \frac{COV(\rho_t, \rho_{t-1})}{\sigma(\rho_t)\sigma(\rho_{t-1})} \tag{8.6}$$

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 (8.6) 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 (8.6), 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 (8.2).

Figure 8.5 shows the autocorrelation with respect to different time lags.

From Figure 8.5, 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.

8.4 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 8.6.

From Figure 8.6, 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.

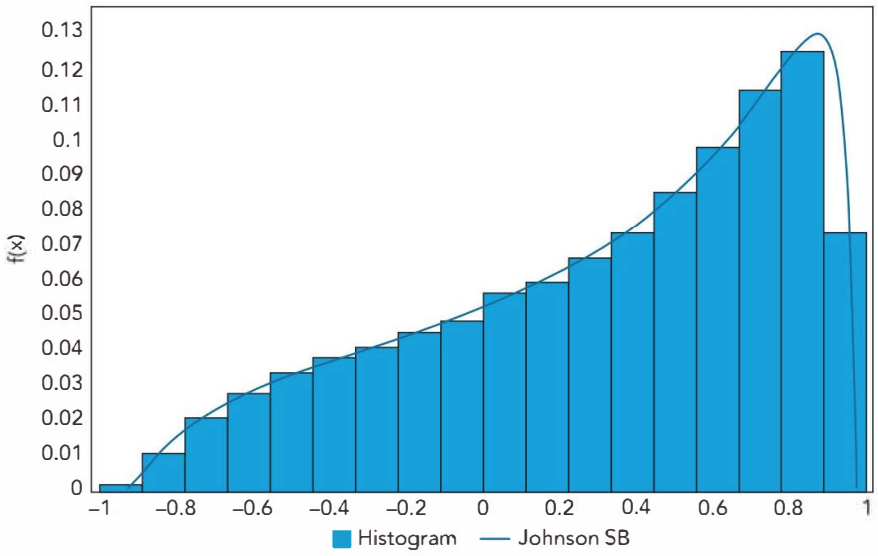


Figure 8.6 **Histogram of 464,580 correlations between the Dow 30 stocks from 1972 to 2017; the continuous line shows the Johnson SB distribution, which provided the best fit.**

We tested 61 distributions for fitting the histogram in Figure 8.6, 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.

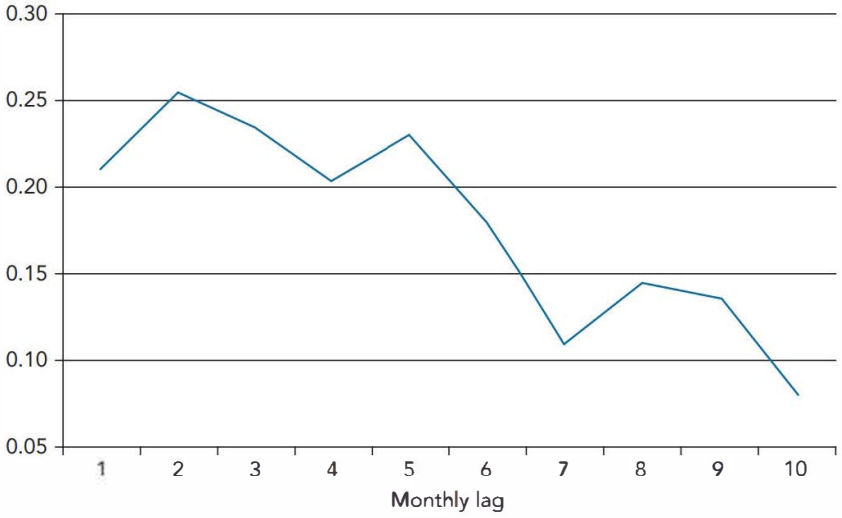


Figure 8.5 **Autocorrelation of monthly average 30 × 30 Dow stock correlations from 1972 to 2017; the time period of the lags is months.**

8.5 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 8.2 displays the relationship of a change in the correlation volatility preceding the start of a recession.

From Table 8.2, 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 8.2, 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 8.1), which often precedes a recession. However, the relationship between a decline in volatility and the severity of the

Table 8.2 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.

8.6 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 8.4) 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

8.1 In which state of the economy are equity correlations the highest?

8.2 In which state of the economy is equity correlation volatility high?

8.3 What follows from Questions 1 and 2 for risk management?

8.4 What is mean reversion?

8.5 How can we quantify mean reversion?

8.6 What is autocorrelation? Name two approaches for how to quantify autocorrelation.

8.7 For equity correlations, we see the typical decrease of autocorrelation with respect to time lags. What does that mean?

8.8 How are mean reversion and autocorrelation related?

8.9 What is the distribution of equity correlations?

8.10 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?

9 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.*

9.1 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.

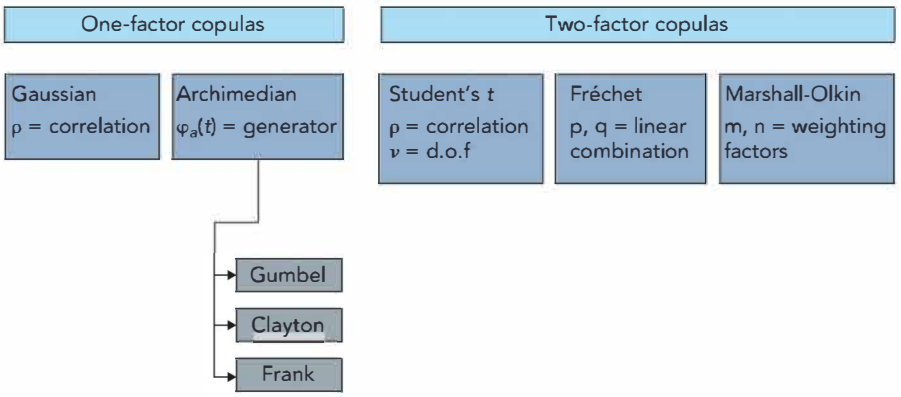


Figure 9.1 Popular copula functions in finance.

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:

$$C : [0, 1]^n \rightarrow [0, 1] \tag{9.1}$$

More explicitly, let $G_i(u_i)$ be a univariate, uniform distribution with $u_i = u_1, \ldots, u_n$, and $i \in \mathbf{N}$. Then there exists a copula function C such that

$$C[G_1(u_1), \ldots, G_n(u_n)] = F_n[F_1^{-1}(G_1(u_1)), \ldots, F_n^{-1}(G_n(u_n)); \rho_F] \tag{9.2}$$

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 (9.2) 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 (9.2), 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 9.1 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

$$C_G[G_1(u_1), \ldots, G_n(u_n)] = M_n[N^{-1}(G_1(u_1)), \ldots, N^{-1}(G_n(u_n)); \rho_M] \tag{9.3}$$

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 (9.3) 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},

$$C_{GD}[Q_i(t), \ldots, Q_n(t)] = M_n[N^{-1}(Q_1(t)), \ldots N^{-1}(Q_n(t)); \rho_M] \tag{9.4}$$

[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 (9.4) 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 (9.4) 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 (9.4) are abscise (x-axis) values of the standard normal distribution. For a numerical example see example 9.1 and Figure 9.2 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 9.1 Let's assume we have two companies, B and Caa, with their estimated default probabilities for years 1 to 10 as displayed in Table 9.1.

Default probabilities for investment-grade companies typically increase in time, since uncertainty increases with time. However, in Table 9.1 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 (9.4). First, we map the cumulative default probabilities $Q(t)$, which are in columns 3 and 5 in Table 9.1, 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 9.2. In the lower graph of Figure 9.2, 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 9.1 (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 9.1 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 9.2, 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 (9.4).

Since we have only $n = 2$ companies B and Caa in our example, equation (9.4) reduces to

$$M_2[N^{-1}(Q_B(t)), N^{-1}(Q_{Caa}(t)); \rho] \tag{9.5}$$

From equation (9.5) 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.

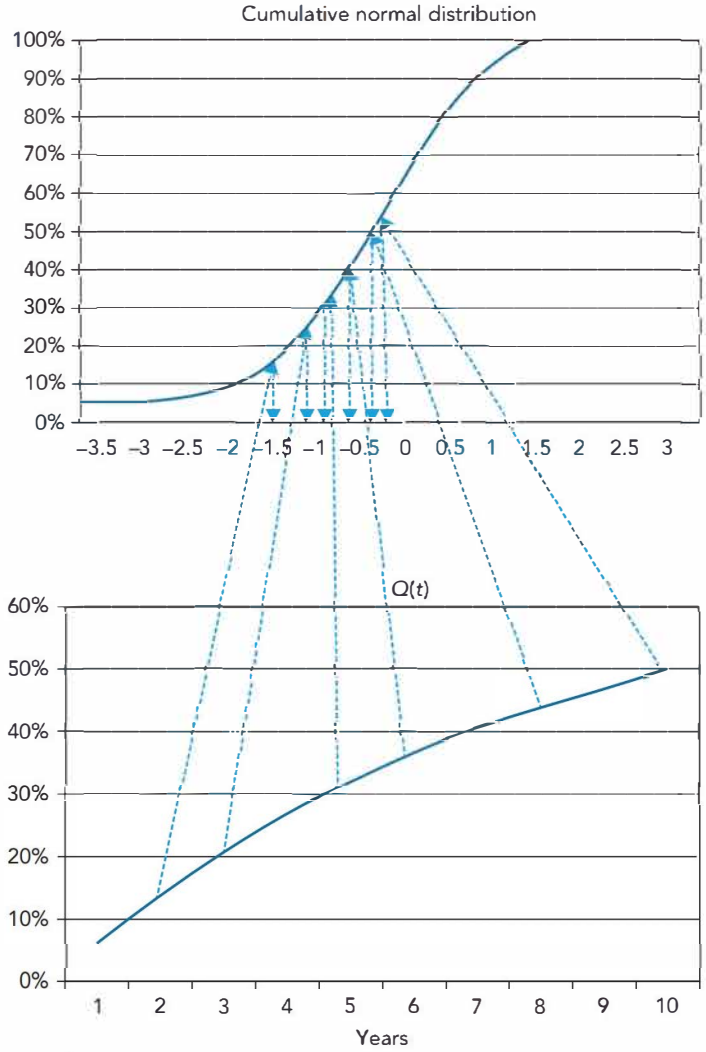


Figure 9.2 **Graphical representation of the copula mapping $N^{-1}(Q(t))$.**

Table 9.1 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 9.2 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 9.3.

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

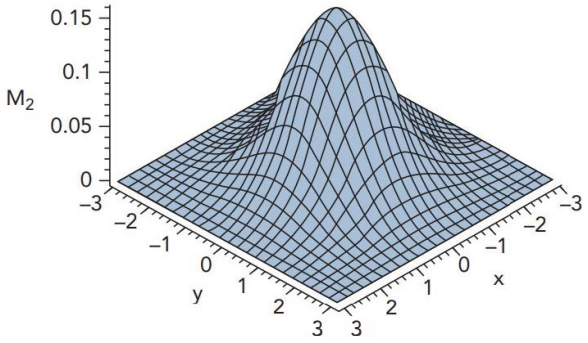


Figure 9.3 Bivariate (non-cumulative) normal distribution.

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

$$Q(t_b \leq 1 \cap t_{Caa} \leq 1)$$
$$\equiv M(x_B \leq -1.5133 \cap x_{Caa} \leq -0.7118, \rho = 0.4) = 3.44\% \quad (9.6)$$

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 9.2.

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

$$Q(t_B \leq 3 \cap t_{Caa} \leq 5)$$
$$\equiv M(x_B \leq -0.8054 \cap x_{Caa} \leq 0.2557, \rho = 0.4) = 16.93\% \quad (9.7)$$

Equations (9.6) and (9.7) 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 (9.6) 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 (9.5) 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,

$$M_n(\cdot) = Q_i(\tau_i) \text{ or} \quad (9.8)$$
$$\tau_i = Q_i^{-1}(M_n(\cdot)) \quad (9.9)$$

There is no closed-form solution for equation (9.8) or (9.9). To find the solution, we first take the sample $M_n(\cdot)$ and use equation (9.8) 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 9.4, 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.

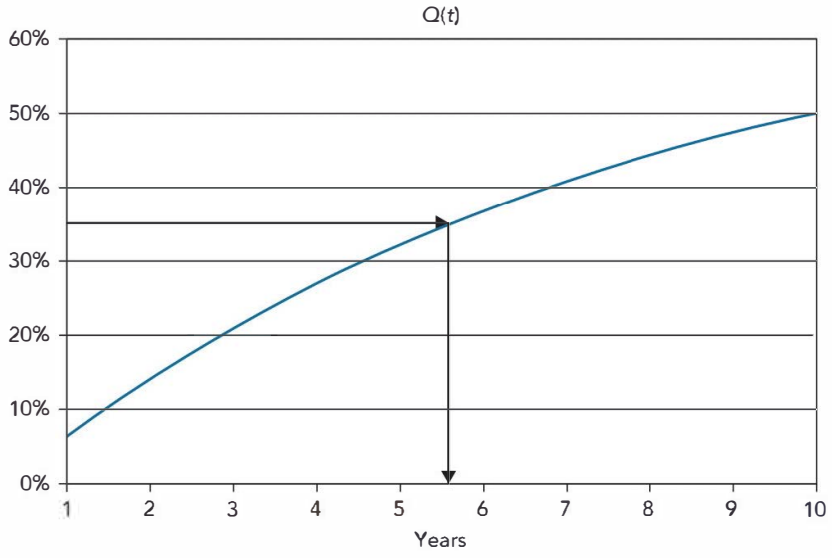


Figure 9.4 **Finding the default time τ of 5.5 years from equation (9.8) for a random sample of the n-variate normal distribution $M_n(\cdot)$ of 35%.**

10

Empirical Approaches to Risk Metrics and Hedging

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.
- 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, *Third Edition, by Bruce Tuckman and Angel Serrat.*

Central to the *DV*01-style metrics and hedges and the multifactor metrics and hedges are implicit assumptions about how rates of different term structures change relative to one another. In this chapter, the necessary assumptions are derived directly from data on rate changes.

The chapter begins with single-variable hedging based on regression analysis. In the example of the section, a trader tries to hedge the interest rate risk of U.S. nominal versus real rates. This example shows that empirical models do not always describe the data very precisely and that this imprecision expresses itself in the volatility of the profit and loss of trades that depend on the empirical analysis.

The chapter continues with two-factor hedging based on multiple regression. The example for this section is that of an EUR swap market maker who hedges a customer trade of 20-year swaps with 10- and 30-year swaps. The quality of this hedge is shown to be quite a bit better than that of nominal *versus* real rates. Before concluding the discussion of regression techniques, the chapter comments on level *versus* change regressions.

The final section of the chapter introduces principal component analysis, which is an empirical description of how rates move together across the curve. In addition to its use as a hedging tool, the analysis provides an intuitive description of the empirical behavior of the term structure. The data illustrations for this section are taken from USD, EUR, GBP, and JPY swap markets. Considerable effort has been made to present this material at as low a level of mathematics as possible.

A theme across the illustrations of the chapter is that empirical relationships are far from static and that hedges estimated over one period of time may not work very well over subsequent periods.

10.1 SINGLE-VARIABLE REGRESSION-BASED HEDGING

This section considers the construction of a relative value trade in which a trader sells a U.S. Treasury bond and buys a U.S. Treasury TIPS (Treasury Inflation Protected Securities). As mentioned in the Overview, TIPS make real or inflation-adjusted payments by regularly indexing their principal amount outstanding for inflation. Investors in TIPS, therefore, require a relatively low real rate of return. By contrast, investors in U.S. Treasury bonds—called nominal bonds when distinguishing them from TIPS—require a real rate of return plus compensation for expected inflation plus, perhaps, an inflation risk premium. Thus the spread between rates of nominal bonds and TIPS reflects market views about inflation. In the relative value trade of this section, a trader bets that this inflation-induced spread will increase.

The trader plans to short USD 100 million of the (nominal) $3\frac{5}{8}$s of August 15, 2019, and, against that, to buy some amount of the TIPS $1\frac{7}{8}$s of July 15, 2019. Table 10.1 shows representative yields and *DV*01s of the two bonds. The TIPS sells at a relatively low yield, or high price, because its cash flows are protected from inflation while the *DV*01 of the TIPS is relatively high because its yield is low. In any case, what face amount of the TIPS should be bought so that the trade is hedged against the level of interest rates, i.e., to both rates moving up or down together, and exposed only to the spread between nominal and real rates?

One choice is to make the trade *DV*01-neutral, i.e., to buy F^R face amount of TIPS such that

$$F^R \times \frac{.081}{100} = 100\text{mm} \times \frac{.067}{100}$$

$$F^R = 100\text{mm} \times \frac{.067}{.081} = \text{USD } 82.7\text{mm} \qquad (10.1)$$

This hedge ensures that if the yield on the TIPS and the nominal bond both increase or decrease by the same number of basis points, the trade will neither make nor lose money. But the trader has doubts about this choice because changes in yields on TIPS and nominal bonds may very well not be one-for-one. To investigate, the trader collects data on daily changes in yield of these two bonds from August 17, 2009, to July 2, 2010, which are then graphed in Figure 10.1, along with a regression line, to be discussed shortly. It is immediately apparent from the graph that, for example, a five basis-point change in the yield of the TIPS does not imply, with very high confidence, a unique change in the nominal yield, nor even an average change of five basis points. In fact, while the daily change in the real yield was about five basis points several times over the study period, the change in the nominal yield over those particular days ranged from 2.2 to 8.4 basis points. This lack of a one-to-one yield relationship calls the *DV*01 hedge into question. For context, by the way, it should be noted that graphing the changes in the yield of one nominal Treasury against changes in the yield of another, of similar maturity, would result in data points much more tightly surrounding the regression line.

Table 10.1 **Yields and *DV*01s of a TIPS and a Nominal U.S. Treasury as of May 28, 2010**

Bond	Yield (%)	*DV*01
TIPS $1\frac{7}{8}$s of 7/15/19	1.237	.081
$3\frac{5}{8}$s of 8/15/19	3.275	.067

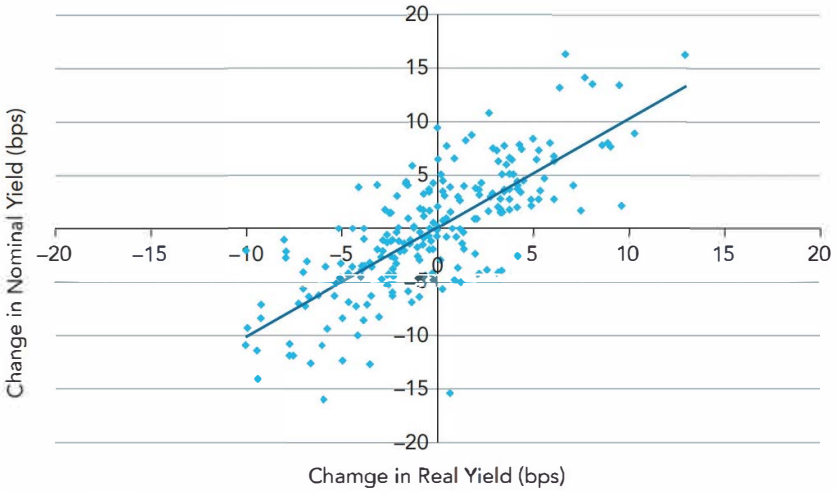


Figure 10.1 **Regression of changes in the yield of the Treasury 3⅝s of August 15, 2019, on changes in the yield of the TIPS 1.875s of July 15, 2019, from August 17, 2009, to July 2, 2010.**

With respect to improving on the *DV*01 hedge, there is not much the trader can do about the dispersion of the change in the nominal yield for a given change in the real yield. That is part of the risk of the trade and will be discussed later. But the trader can estimate the average change in the nominal yield for a given change in the real yield and adjust the *DV*01 hedge accordingly. For example, were it to turn out—as it will—that the nominal yield in the data changes by 1.0189 basis points per basis-point change in the real yield, the trader could adjust the hedge such that

$$F^R \times \frac{.081}{100} = 100\text{mm} \times \frac{.067}{100} \times 1.0189$$

$$F^R = \text{USD } 100\text{mm} \times \frac{.067}{.081} \times 1.0189 = \text{USD } 84.3\text{mm} \quad (10.2)$$

Relative to the *DV*01 hedge of USD 82.7 million in (10.1), the hedge in (10.2) increases the amount of TIPS to compensate for the empirical fact that, on average, the nominal yield changes by more than one basis point for every basis-point change in the real yield.

The next subsection introduces *regression analysis*, which is used both to estimate the coefficient 1.0189, used in Equation (10.2), and to assess the properties of the resulting hedge.

Least-Squares Regression Analysis

Let Δy_t^N and Δy_t^R be the changes in the yields of the nominal and real bonds, respectively, and assume that

$$\Delta y_t^N = \alpha + \beta \Delta y_t^R + \varepsilon_t \quad (10.3)$$

According to Equation (10.3), changes in the real yield, the *independent variable*, are used to predict changes in the nominal yield, the *dependent variable*. The intercept, α, and the slope, β, need to be estimated from the data. The error term ε_t is the deviation of the nominal yield change on a particular day from the change predicted by the model. *Least-squares estimation* of (10.3), to be discussed presently, requires that the model be a true description of the dynamics in question and that the errors have the same probability distribution, are independent of each other, and are uncorrelated with the independent variable.[1]

As an example of the relationship between the nominal and real yields in (10.3), say that the parameters estimated with the data, denoted $\hat{\alpha}$ and $\hat{\beta}$, are 0 and 1.02 respectively. Then, if Δy_t^R is 5 basis points on a particular day, the predicted change in the nominal yield, written $\Delta \hat{y}_t^N$, is

$$\Delta \hat{y}_t^N = \hat{\alpha} + \hat{\beta} \Delta y_t^R$$
$$= 0 + 1.02 \times 5 = 5.1 \quad (10.4)$$

Furthermore, should it turn out that the nominal yield changes by 5.5 basis points on that day, then the realized error that day, written $\hat{\varepsilon}_t$, following Equation (10.3), is defined as

$$\hat{\varepsilon}_t = \Delta y_t^N - \hat{\alpha} - \hat{\beta} \Delta y_t^R$$
$$= \Delta y_t^N - \Delta \hat{y}_t^N \quad (10.5)$$

In this example,

$$\hat{\varepsilon}_t = 5.5 - 5.1 = .4 \quad (10.6)$$

Least-squares estimation of α and β finds the estimates $\hat{\alpha}$ and $\hat{\beta}$ that minimize the sum of the squares of the realized error terms over the observation period,

$$\sum_t \hat{\varepsilon}_t^2 = \sum_t (\Delta y_t^N - \hat{\alpha} - \hat{\beta} \Delta y_t^R)^2 \quad (10.7)$$

[1] Since the nominal rate is the real rate plus the inflation rate, the error term in Equation (10.3) contains the change in the inflation rate. Therefore, the assumption that the independent variable be uncorrelated with the error term requires here that the real rate be uncorrelated with the inflation rate. This is a tolerable, though far from ideal, assumption: the inflation rate can have effects on the real economy and, consequently, on the real rate.

If the regression were specified such that the real rate were the dependent variable and the nominal rate the independent variable, the requirement that the error and the dependent variable be uncorrelated would certainly not be met. In that case, the error term contains the inflation rate and there is no credible argument that the nominal rate is even approximately uncorrelated with the inflation rate. Consequently, a more advanced estimation procedure would be required, like that of *instrumental variables*.

where the equality follows from (10.5). The 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 values are penalized substantially more than smaller errors.

Least-squares estimation is available through many statistical packages and spreadsheet add-ins. A typical summary of the regression output from estimating Equation (10.3) using the data in Figure 10.1 is given in Table 10.2. The $\hat{\beta}$ reported in the table is 1.0189, which says that, over the sample period, the nominal yield increases by 1.0189 basis points per basis-point increase in real yields. The constant term of the regression, $\hat{\alpha}$, is not very different from zero, which is typically the case in regressions of changes in a yield on changes in a comparable yield. The economic interpretation of this regularity is that a yield does not usually trend up or down while a comparable yield is not changing.

Table 10.2 reports standard errors of $\hat{\alpha}$ and $\hat{\beta}$ of .2529 and .0525, respectively. Under the assumptions of least squares and the availability of sufficient data, the parameters $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed with means equal to the true model values, α and β respectively, and with standard deviations that can be estimated as the standard errors given in the table. Therefore, relying on the properties of the normal distribution, the confidence interval $.0503 \pm 2 \times .2529$ or (−.4555, .5561) has a 95% chance of falling around the true value α. And since this confidence interval does include the value zero, one cannot reject the statistical hypothesis that $\alpha = 0$. Similarly, the 95% confidence interval with respect to β is $1.0189 \pm 2 \times .0595$, or (.8999, 1.1379). So, while regression hedging makes heavy use of the point estimate $\hat{\beta} = 1.0189$, the true value of β may very well be somewhat higher or lower.

Substituting the estimated coefficients from Table 10.2 into the predicted regression equation in the first line of (10.4),

$$\Delta\hat{y}_t^N = \hat{\alpha} + \hat{\beta}\Delta y_t^R$$
$$\Delta\hat{y}_t^N = .0503 + 1.0189 \times \Delta y_t^R \qquad (10.8)$$

This relationship is known as the *fitted regression line* and is the straight line through the data that appears in Figure 10.1.

Table 10.2 reports two other useful statistics, the R-squared and the standard error of the regression. The R-squared in this case is 56.3%, which means that 56.3% of the variance of changes in the nominal yield can be explained by the model. In a one-variable regression, the R-squared is just the square of the correlation of the two changes, so the correlation between changes in the nominal and real yields is the square root of 56.3% or about 7.5%. This is a relatively low number compared with typical correlations between changes in two nominal yields, echoing the comment made in reference to the relatively wide dispersion of the points around the regression line in Figure 10.1.

Table 10.2 Regression Analysis of Changes in the Yield of the 3⅝s of August 15, 2019, on the Changes in Yield of the TIPS 1⅞s of July 15, 2019, from August 17, 2009, to July 2, 2010

No. of Observations	**229**	
R-Squared	**56.3%**	
Standard Error	**3.82**	
Regression Coefficients	**Value**	**Std. Error**
Constant ($\hat{\alpha}$)	0.0503	.2529
Change in Real Yield ($\hat{\beta}$)	1.0189	.0595

The second useful statistic reported in Table 10.2 is the standard error of the regression, denoted here by $\hat{\sigma}$ and given as 3.82 basis points. Algebraically, $\hat{\sigma}$ is essentially the standard deviation of the realized error terms $\hat{\varepsilon}_t$,[2] defined in Equation (10.5). Graphically, each $\hat{\varepsilon}_t$ is the vertical line from a data point directly down or up to the regression line and $\hat{\sigma}$ is essentially the standard deviation of these distances. Either way, $\hat{\sigma}$ measures how well the model fits the data in the same units as the dependent variable, which, in this case, are basis points.

The Regression Hedge

The use of the regression coefficient in the hedging example of this section was discussed in the development of Equation (10.2). More formally, denoting the face amounts of the real and nominal bonds by F^R and F^N and their *DV*01s by $DV01^R$ and $DV01^N$, the regression-based hedge, characterized earlier as the *DV*01 hedge adjusted for the average change of nominal yields relative to real yields, can be written as follows:

$$F^R = -F^N \times \frac{DV01^N}{DV01^R} \times \hat{\beta} \qquad (10.9)$$

It turns out, however, that this regression hedge has an even stronger justification. The profit and loss (P&L) of the hedged position over a day is

$$-F^R \times \frac{DV01^R}{100}\Delta y_T^R - F^N \times \frac{DV01^N}{100}\Delta y_t^N \qquad (10.10)$$

[2] If the number of observations is n, the standard error of the regression is actually defined as the square root of $\frac{\sum_t \hat{\varepsilon}_t^2}{(n-2)}$. The average of the $\hat{\varepsilon}_t$ in a regression with a constant is zero by construction, so the standard error of the regression differs from the standard deviation of the errors only because of the division by $n - 2$ instead of $n - 1$.

Appendix A in this chapter shows that the hedge of Equation (10.9) minimizes the variance of the P&L in (10.10) over the data set shown in Figure 10.1 and used to estimate the regression parameters of Table 10.2.

In the example of this section, $F^N = -\text{USD } 100\text{mm}$, $\hat{\beta} = 1.0189$, $DV01^N = .067$, and $DV01^R = .081$, so, from (10.9), as derived before, $F^R = \text{USD } 84.279\text{mm}$. Because the estimated β happens to be close to one, the regression hedge of about USD 84.3 million is not very different from the *DV*01 hedge of USD 82.7 million calculated earlier. In fact, some practitioners would describe this hedge in terms of the *DV*01 hedge. Rearranging the terms of (10.9),

$$\frac{-F^R \times DV01^R}{F^N \times DV01^N} = \hat{\beta} = 101.89\% \qquad (10.11)$$

In words, the risk of the (TIPS) hedging portfolio, measured by *DV*01, is 101.89% of the risk of the underlying (nominal) position, measured by *DV*01. Alternatively, the *risk weight* of the hedge portfolio is 101.89%. This terminology does connect the hedge to the common *DV*01 benchmark but is somewhat misleading because the whole point of the regression-based hedge is that the risks of the two securities cannot properly be measured by the *DV*01 alone. It should also be noted at this point that the regression-based and *DV*01 hedges are certainly not always this close in magnitude, even in other cases of hedging TIPS *versus* nominals, as will be illustrated in the next subsection.

An advantage of the regression framework for hedging is that it automatically provides an estimate of the volatility of the hedged portfolio. To see this, substitute F^R from (10.9) into the P&L expression (10.10) and rearrange terms to get the following expression for the P&L of the hedged position:

$$-F^N \times \frac{DV01^N}{100}(\Delta y_t^N - \hat{\beta}\Delta y_t^R) \qquad (10.12)$$

From the definition of $\hat{\varepsilon}_t$ in (10.5), the term in parentheses equals $\hat{\varepsilon}_t + \hat{\alpha}$. But since $\hat{\alpha}$ is typically not very important, the standard error of the regression $\hat{\sigma}$ can be used to approximate the standard deviation of $\Delta y_t^N - \hat{\beta}\Delta y_t^R$. Hence, the standard deviation of the P&L in (10.12) is approximately

$$F^N \times \frac{DV01^N}{100} \times \hat{\sigma} \qquad (10.13)$$

In the present example, recalling that the standard error of the regression can be found in Table 10.2, the daily volatility of the P&L of the hedged portfolio is approximately

$$\text{USD } 100\text{mm} \times \frac{.067}{100} \times 3.82 = \text{USD } 255,940 \qquad (10.14)$$

The trader would have to compare this volatility with an expected gain to decide whether or not the risk-return profile of the trade is attractive.

The Stability of Regression Coefficients over Time

An important difficulty in using regression-based hedging in practice is that the hedger can never be sure that the hedge coefficient, β, is constant over time. Put another way, the errors around the regression line might be random outcomes around a stable relationship, as described by Equation (10.3), or they might be manifestations of a changing relationship. In the former situation a hedger can safely continue to use a previously estimated $\hat{\beta}$ for hedging while, in the latter situation, the hedger should re-estimate the hedge coefficient with more recent data, if available, or with data from a past, more relevant time period. But how can the hedger know which situation prevails?

A useful start for thinking about the stability of an estimated regression coefficient is to estimate that coefficient over different periods of time and then observe if the result is stable or not. To this end, with the same data as before, Figure 10.2 graphs $\hat{\beta}$ for regressions over rolling 30-day windows. This means that the full data set of changes from August 18, 2009, to July 2, 2010, is used in 30-day increments, as follows: the first $\hat{\beta}$ comes from a regression of changes from August 18, 2009, to September 28, 2009; the second $\hat{\beta}$ from that regression from August 19, 2009, to September 29, 2009, etc.; and the last $\hat{\beta}$ from May 24, 2010, to July 2, 2010. The estimates of β in the figure certainly do vary over time, but the range of .75 to 1.29 is not extremely surprising given the previously computed 95% confidence interval with respect to β of (.8999, 1.1379). More troublesome, perhaps, is the fact that the most recent values of $\hat{\beta}$ have been trending up, which may indicate a change in regime in which even higher values of β characterize the relationship between nominal and real rates.

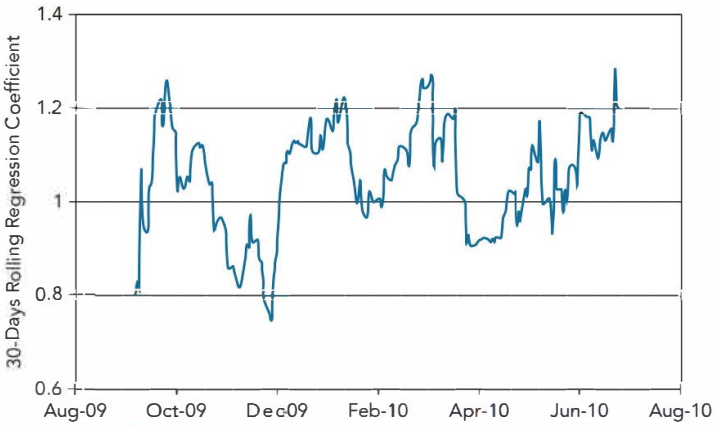


Figure 10.2 **Rolling 30-day regression coefficient for the change in yield of the Treasury 3⅝s of August 15, 2019, on the change in yield of the TIPS 1⅞s of July 15, 2019.**

Table 10.3 Regression Analysis of Changes in the Yield of the 6½s of February 15, 2010, on the Changes in Yield of the TIPS 4¼s of January 15, 2010, from February 15, 2000, to February 15, 2002

No. of Observations	519	
R-Squared	43.0%	
Standard Error	4.70	
Regression Coefficients	**Value**	**Std. Error**
Constant ($\hat{\alpha}$)	−.0267	.2067
Change in Real Yield ($\hat{\beta}$)	1.5618	.0790

For a bit more perspective before closing this subsection, the period February 15, 2000, to February 15, 2002, when rates were substantially higher, was characterized by significantly higher levels of $\hat{\beta}$ and higher levels of uncertainty with respect to the regression relationship. The two bonds used in this analysis are the TIPS 4¼s of January 15, 2010, and the Treasury 6½s of February 15, 2010. Summary statistics for the regression of changes in yields of the nominal 6½s on the real 4¼s are given in Table 10.3.

Compared with Table 10.2, the estimated β here is 50% larger and the precision of this regression, measured by the R-squared or the standard error of the regression, is substantially worse. The contrast across periods again emphasizes the potential pitfalls of relying on estimated relationships persisting over time. This does not imply, of course, that blindly assuming a β of one, as in *DV*01 hedging, is a generally superior approach.

10.2 TWO-VARIABLE REGRESSION-BASED HEDGING

To illustrate regression hedging with two independent variables, this section considers the case of a market maker in EUR interest rate swaps. An algebraic introduction is followed by an empirical analysis.

The market maker in question has bought or received fixed in relatively illiquid 20-year swaps from a customer and needs to hedge the resulting interest rate exposure. Immediately paying fixed or selling 20-year swaps would sacrifice too much if not all of the spread paid by the customer, so the market maker chooses instead to sell a combination of 10- and 30-year swaps. Furthermore, the market maker is willing to rely on a two-variable regression model to describe the relationship between changes in 20-year swap rates and changes in 10- and 30-year swap rates:

$$\Delta y_t^{20} = \alpha + \beta^{10}\Delta y_t^{10} + \beta^{30}\Delta y_t^{30} + \varepsilon_t \qquad (10.15)$$

Equation (10.15) can be estimated by least squares, analogously to the single-variable case, by minimizing

$$\sum_t (\Delta y_t^{20} - \hat{\alpha} - \hat{\beta}^{10}\Delta y_t^{10} - \hat{\beta}^{30}\Delta y_t^{30})^2 \qquad (10.16)$$

with respect to the parameters $\hat{\alpha}$, $\hat{\beta}^{10}$ and $\hat{\beta}^{30}$. The estimation of these parameters then provides a predicted change for the 20-year swap rate:

$$\Delta \hat{y}_t^{20} = \hat{\alpha} + \hat{\beta}^{10}\Delta y_t^{10} + \hat{\beta}^{30}\Delta y_t^{30} \qquad (10.17)$$

To derive the notional face amount of the 10- and 30-year swaps, F^{10} and F^{30}, respectively, required to hedge F^{20} face amount of the 20-year swaps, generalize the reasoning given in the single-variable case as follows. Write the P&L of the hedged position as

$$-F^{20}\frac{DV01^{20}}{100}\Delta y_t^{20} - F^{10}\frac{DV01^{10}}{100}\Delta y_t^{10} - F^{30}\frac{DV01^{30}}{100}\Delta y_t^{30} \qquad (10.18)$$

Then substitute the predicted change in the 20-year rate from (10.17) into (10.18), retaining only the terms depending on Δy_t^{10} and Δy_t^{30}, to obtain

$$\left[-F^{20}\frac{DV01^{20}}{100}\hat{\beta}^{10} - F^{10}\frac{DV01^{10}}{100}\right]\Delta y_t^{10} + \left[-F^{20}\frac{DV01^{20}}{100}\hat{\beta}^{30} - F^{30}\frac{DV01^{30}}{100}\right]\Delta y_t^{30} \qquad (10.19)$$

Finally, choose F^{10} and F^{30} to set the terms in brackets equal to zero, i.e., to eliminate the dependence of the predicted P&L on changes in the 10- and 30-year rates. This leads to two equations with the following solutions:

$$F^{10} = -F^{20}\frac{DV01^{20}}{DV01^{10}}\hat{\beta}^{10} \qquad (10.20)$$

$$F^{30} = -F^{20}\frac{DV01^{20}}{DV01^{30}}\hat{\beta}^{30} \qquad (10.21)$$

As in the single-variable case, this 10s-30s hedge of the 20-year can be expressed in terms of risk weights. More specifically, the *DV*01 risk in the 10-year part of the hedge and the *DV*01 risk in the 30-year part of the hedge can both be expressed as a fraction of the *DV*01 risk of the 20-year. Mathematically, these risk weights can be found by rearranging (10.20) and (10.21):

$$\frac{-F^{10}\times DV01^{10}}{F^{20}\times DV01^{20}} = \hat{\beta}^{10} \qquad (10.22)$$

$$\frac{-F^{30}\times DV01^{30}}{F^{20}\times DV01^{20}} = \hat{\beta}^{30} \qquad (10.23)$$

Proceeding now to the empirical analysis, the market maker, as of July 2006, performs an initial regression analysis using data on changes in the 10-, 20-, and 30-year EUR swap rates from July 2,

2001, to July 3, 2006. Summary statistics for the regression of changes in the 20-year EUR swap rate on changes in the 10- and 30-year EUR swap rates are given in Table 10.4. The statistical quality of these results, characteristic of all regressions of like rates, are far superior to those of the nominal against real yields of the previous section: the R-squared or percent variance explained by the regression is 99.8%; the standard error of the regression is only .14 basis points; and the 95% confidence intervals with respect to the two coefficients are extremely narrow, i.e., (.2153, .2289) for the 10-year and (.7691, .7839) for the 30-year. Lastly, in a result similar to those of the regressions of the previous section, the constant is insignificantly different from zero.

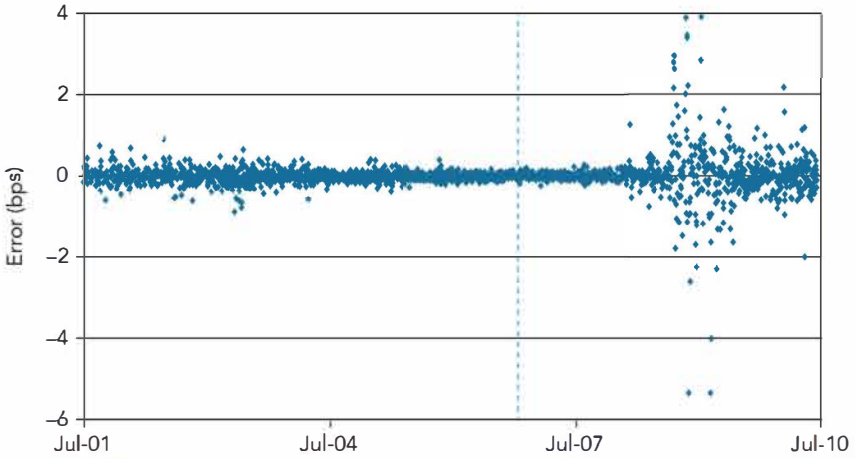


Figure 10.3 **In- and out-of-sample errors for a regression of changes of 20-year and 10- and 30-year EUR swap rates with estimation period July 2, 2001, to July 3, 2006.**

Applying the risk-weight interpretation of the regression coefficients given in Equations (10.22) and (10.23), the results in Table 10.4 say that 22.21% of the *DV*01 of the 20-year swap should be hedged with a 10-year swap and 77.65% with a 30-year swap. The sum of these weights, 99.86%, happens to be very close to one, meaning that the *DV*01 of the regression hedge very nearly matches the *DV*01 of the 20-year swap, although this certainly need not be the case: minimizing the variance of the P&L of a hedged position, when rates are not assumed to move in parallel, need not result in a *DV*01-neutral portfolio.

Tight as the in-sample regression relationship seems to be, the real test of the hedge is whether it works out-of-sample.[3] To this end, Figure 10.3 tracks the errors of the hedge over time. All of these errors are computed as the realized change in the 20-year yield minus the predicted change for that yield based on the estimated regression in Table 10.4:

$$\hat{\varepsilon}_t = \Delta y_t^{20} - (-.0014 + .2221\Delta y_t^{10} + .7765\Delta y_t^{30}) \quad (10.24)$$

The errors to the left of the vertical dotted line are in-sample in that the same Δy_t^{20} used to compute $\hat{\varepsilon}_t$ in (10.24) were also used to compute the coefficient estimates −.0014, .2221, and .7765. In other words, it is not that surprising that the $\hat{\varepsilon}_t$ to the left of the dotted line are small because the regression coefficients were estimated to minimize the sum of squares of these errors. By contrast, the errors to the right of the dotted line are out-of-sample: these $\hat{\varepsilon}_t$ are computed from realizations of Δy_t^{20} after July 3, 2006, but using the regression coefficients estimated over the period from July 2, 2001, to July 3, 2006. It is, therefore, the size and behavior of these out-of-sample errors that provide evidence as to the stability of the estimated coefficients over time.

From inspection of Figure 10.3 the out-of-sample errors are indeed small, for the most part, until August and September 2008, a peak in the financial crisis of 2007–2009. After then the daily errors ran as high as about four basis points and as low as about −5.3 basis points. And while the accuracy of the relationship seems to have recovered somewhat to the far right-end of the graph, by the summer of 2009, the errors there are not nearly so well behaved as at the start of the out-of-sample period.

It is obvious and easy to say that the market maker, during the turbulence of a financial crisis, should have replaced the regression of Table 10.4 and the resulting hedging rule. But replace these with what? What does the market maker do at that time, before there exist sufficient post-crisis data points? And what does the

Table 10.4 **Regression Analysis of Changes in the Yield of the 20-year EUR Swap Rate on Changes in the 10- and 30-Year EUR Swap Rates from July 2, 2001, to July 3, 2006**

No. of Observations	**1281**	
R-Squared	**99.8%**	
Standard Error	**.14**	
Regression Coefficients	**Value**	**Std. Error**
Constant ($\hat{\alpha}$)	−.0014	.0040
Change in 10-Year Swap Rate ($\hat{\beta}^{10}$)	.2221	.0034
Change in 30-Year Swap Rate ($\hat{\beta}^{30}$)	.7765	.0037

[3] The phrase *in-sample* refers to behavior within the period of estimation, in this case July 2, 2001, to July 3, 2006. The phrase *out-of-sample* refers to behavior outside the period of estimation, usually after but possibly before that period as well.

market maker do after the worst of the crisis: estimate a regression from data during the crisis or revert to some earlier, more stable period? These are the kinds of issues that make regression hedging an art rather than a science. In any case, it should again be emphasized that avoiding these issues by blindly resorting to a one-security *DV*01 hedge, or a two-security *DV*01 hedge with arbitrarily assigned risk weights, like 50%-50%, is even less satisfying.

10.3 LEVEL VERSUS CHANGE REGRESSIONS

When estimating regression-based hedges, some practitioners regress changes in yields on changes in yields, as in the previous sections, while others prefer to regress yields on yields. Mathematically, in the single-variable case, the level-on-level regression with dependent variable y and independent variable x is

$$y_t = \alpha + \beta x_t + \varepsilon_t \tag{10.25}$$

while the change-on-change regression is[4]

$$y_t - y_{t-1} = \Delta y_t = \beta \Delta x_t + \Delta \varepsilon_t \tag{10.26}$$

By theory that is beyond the scope of this book, if the error terms ε_t are independently and identically distributed random variables with mean zero and are uncorrelated with the independent variable, then so are the $\Delta\varepsilon_t$, and least squares on either (10.25) or (10.26) will result in coefficient estimators that are *unbiased*,[5] *consistent*,[6] and *efficient*, i.e., of *minimum variance*, in the class of linear estimators. If the error terms of either specification are not independent of each other, however, then the least-squares coefficients of that specification are not necessarily efficient, but retain their unbiasedness and consistency.

To illustrate the economics behind the assumption that error terms are independent of each other, say that $\alpha = 0$, that $\beta = 1$, that y is the yield on a coupon bond, and that x is the yield on another, near-maturity coupon bond. Say further that the yield on the x-bond was 5% yesterday and 5% again today while the yield on the y-bond was 1% yesterday. Because the yield on the x-bond is 5% today, the level Equation (10.25) predicts that the yield on the y-bond will be 5% today, despite its being 1% yesterday. But if the market yield was so far off yesterday's prediction, with a realized error of −4%, then it is more likely that the error today will be not far from −4% and that the yield of the y-bond yield will be closer to 1% than the 5% predicted by (10.25). Put another way, the errors in (10.25) are not likely to be independent of each other, as assumed, but rather persistent, or correlated over time.

The change regression (10.26) assumes the opposite extreme with respect to the errors, i.e., that they are completely persistent. Continuing with the example of the previous paragraph, with the yield on the y-bond at 1% yesterday and the yield on the x-bond unchanged from yesterday, the change regression predicts that y-bond will remain at 1%. But, as reasoned above, it is more likely that the y-bond yield will move some of the way back from 1% to 5%. Hence, the error terms in (10.26) are also unlikely to be independent of each other.

The first lesson to be drawn from this discussion is that because the error terms in both (10.26) and (10.25) are likely to be correlated over time, i.e., *serially correlated*, their estimated coefficients are not efficient. But, with nothing to gainsay the validity of the other assumptions concerning the error terms, the estimated coefficients of both the level and change specifications are still unbiased and consistent.

The second lesson to be drawn from the discussion of this section is that there is a more sensible way to model the relationship between two bond yields than either (10.26) or (10.25). In particular, model the behavior that the y-bond's yield will, on average, move somewhat closer from 1% to 5%. Mathematically, assume (10.25) with the error dynamics

$$\varepsilon_t = \rho \varepsilon_{t-1} + \upsilon_t \tag{10.27}$$

for some constant $\rho < 1$. Assumption (10.27) says that today's error consists of some portion of yesterday's error plus a new random fluctuation. In terms of the numerical example, if $\rho = 75\%$, then yesterday's error of −4% would generate an average error today of 75% × − 4% or −3% and, therefore, an expected y-bond yield of 5% − 3% or 2%. In this way the error structure (10.27) has the yield of the y-bond converging to its predicted value of 5% given the yield of the x-bond at 5%. While beyond the scope of this book, the procedure for estimating (10.25) with the error structure (10.27) is presented in many statistical texts.

10.4 PRINCIPAL COMPONENTS ANALYSIS

Overview

Regression analysis tries to explain the changes in the yield of one bond relative to changes in the yields of a small number of other bonds. It is often useful, however, to have a single,

[4] It is usual to include a constant term in the change-on-change regression, but for the purposes of this section, to maintain consistency across the two specifications, this constant term is omitted.

[5] An unbiased estimator of a parameter is such that its expectation equals the true value of that parameter.

[6] A consistent estimator of a parameter, with enough data, becomes arbitrarily close to the true value of the parameter.

empirical description of the behavior of the term structure that can be applied across all bonds. *Principal Components* (PCs) provide such an empirical description.

To fix ideas, consider the set of swap rates from 1 to 30 years at annual maturities. One way to describe the time series fluctuations of these rates is through the variances of the rates and their pairwise covariances or correlations. Another way to describe the data, however, is to create 30 interest rate factors or components, where each factor describes a change in each of the 30 rates. So, for example, one factor might be a simultaneous change of 5 basis points in the 1-year rate, 4.9 basis points in the 2-year rate, 4.8 basis points in the 3-year rate, etc. Principal Components Analysis (PCA) sets up these 30 such factors with 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 this set of interest rates.
2. The PCs are uncorrelated with each other. While changes in the individual rates are, of course, highly correlated with each other, the PCs are constructed so that they are uncorrelated.
3. Subject to these two properties or constraints, each PC is chosen to have the maximum possible variance given all earlier PCs. In other words, the first PC explains the largest fraction of the sum of the variances of the rates; the second PC explains the next largest fraction, etc.

PCs of rates are particularly useful because of an empirical regularity: the sum of the variances of the first three PCs is usually quite close to the sum of variances of all the rates. Hence, rather than describing movements in the term structure by describing the variance of each rate and all pairs of correlation, one can simply describe the structure and volatility of each of only three PCs.

The next subsections illustrate PCs and their uses in the context of USD and then global swap markets. For interested readers, Appendix B in this chapter describes the construction of PCs with slightly more mathematical detail, using the simpler context of three interest rates and three PCs. Fully general and more mathematical descriptions are available in numerous other books and articles.

PCAs for USD Swap Rates

Figure 10.4 graphs the first three principal components from daily data on USD swap rates while Table 10.5 provides a selection of the same information in tabular form. Thirty different data series are used, one series for each annual maturity from one to 30 years, and the observation period spans from October 2001 to October 2008. (Data from more recent dates will be presented and discussed later in this section.)

Columns (2) to (4) in Table 10.5 correspond to the three PC curves in Figure 10.4. These components can be interpreted as follows. A one standard-deviation increase in the "Level" PC, given in column (2), is a simultaneous 3.80 basis-point increase in the one-year swap rate, a 5.86 basis-point increase in the 2-year, etc., and a 5.38 basis-point increase in the 30-year. This PC is said to represent a "level" change in rates because rates of all maturities move up or down together by, very roughly, the same amount. A one standard-deviation increase in the "Slope" PC, given in column (3), is a simultaneous 2.74 basis-point *drop* in the 1-year rate, a 3.09 basis-point drop in the 2-year rate, etc., and a 6.74 basis-point *increase* in the 30-year rate. This PC is said to represent a "slope" change in rates because short-term rates fall while longer-term rates increase, or *vice versa*. Finally, a one standard-deviation increase in the "Short Rate" PC, given in column (4), is made up of simultaneous increases in short-term rates (e.g., one- and two-year terms), small decreases in intermediate-term rates (e.g., 5- and 10-year terms), and small increases in long-term rates (e.g., 20- and 30-year terms). While this PC is often called a "curvature" change, because intermediate-term rates move in the opposite direction from short- and long-term rates, the short-term rates moves dominate. Hence, the third PC is interpreted here as an additional factor to describe movements in short-term rates.

One feature of the shape of the level PC warrants additional discussion. Short-term rates might be expected to be more volatile than longer-term rates because changes in short-term rates are determined by current economic conditions, which are relatively volatile, while longer-term rates are determined mostly by expectations of future economic conditions, which

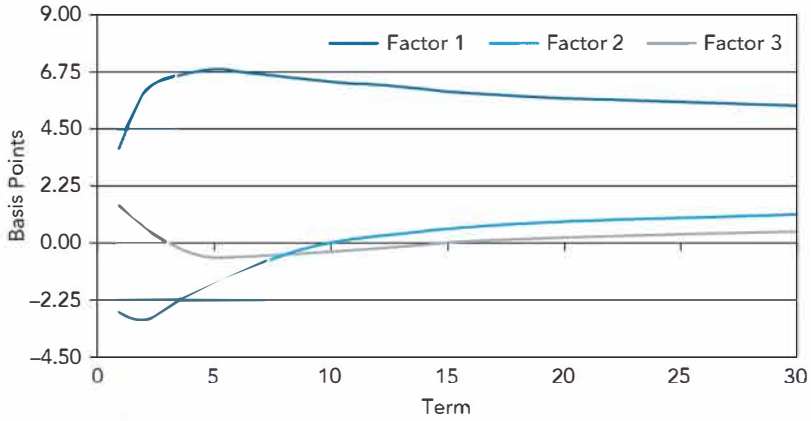


Figure 10.4 **The first three principal components from USD swap rates from October 2001 to October 2008.**

Table 10.5 Selected Results of Principal Components for the USD Swap Curve from October 1, 2001, to October 2, 2008. Units are Basis Points or Percentages

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	PCs					% of PC Variances			
Term	Level	Slope	Short Rate	PC Vol	Total Vol	Level	Slope	Short Rate	PC Vol/ Total Vol (%)
1	3.80	−2.74	1.48	4.91	4.96	59.8	31.0	9.1	99.05
2	5.86	−3.09	0.59	6.65	6.67	77.7	21.5	0.8	99.74
5	6.85	−1.53	−0.57	7.04	7.06	94.7	4.7	0.7	99.85
10	6.35	0.06	−0.34	6.36	6.37	99.7	0.0	0.3	99.83
20	5.69	0.82	0.14	5.75	5.75	97.9	2.0	0.1	99.95
30	5.38	1.09	0.39	5.51	5.52	95.6	3.9	0.5	99.79
Total	32.47	6.74	2.28	33.25	33.29	95.4	4.1	0.5	99.87

are relatively less volatile. But since the Board of Governors of the Federal Reserve System, like many other central banks, anchors the very short-term rate at some desired level, the volatility of very short-term rates is significantly dampened. The level factor, which, as will be discussed shortly, explains the vast majority of term structure movements, and reflects this behavior on the part of central banks: very short-term rates move relatively little. Then, at longer maturities, the original effect prevails and longer-term rates move less than intermediate and shorter-term rates.

Column (5) of Table 10.5 gives the combined standard deviation or volatility of the three principal components for a given rate, and column (6) gives the total or empirical volatility of that rate. For the one-year rate, for example, recalling that the principal components are uncorrelated, the combined volatility, in basis points, from the three components is

$$\sqrt{3.80^2 + (-2.74)^2 + 1.48^2} = 4.91 \qquad (10.28)$$

The total or empirical volatility of the one-year rate, however, computed directly from the time series data, is 4.96 basis points. Column (10) of the table gives the ratio of columns (5) and (6), which, for the 1-year rate is 4.9099/4.9572 or 99.05%. (For readability, many of the entries of Table 10.5 are rounded although calculations are carried out to higher precision.)

Columns (7) through (9) of Table 10.5 give the ratios of the variance of each PC component to the total PC variance. For the 1-year rate, these ratios are $3.80^2/4.91^2 = 59.9\%$; $(-2.74)^2/4.91^2 = 31.1\%$; and $1.48^2/4.91^2 = 9.1\%$.

Finally, the last row of the table gives statistics on the square root of the sum of the variances across rates of different maturities. The sum of the variances is not a particularly interesting economic quantity—it does not, for example, represent the variance of any interesting portfolio—but, as mentioned in the overview of PCA, this sum is used to ensure that the PCs capture all of the volatility of the underlying interest rate series.

Having explained the calculations of Figure 10.4 and Table 10.5, the text can turn to interpretation. First and foremost, column (10) of Table 10.5 shows that, for rates of all maturities, the three principal components explain over 99% of rate volatility. And, across all rates, the three PCs explain 99.87% of the sum of the variability of these rates. While these findings represent relatively recent data on U.S. swap rates, similarly high explanatory powers characterize the first three components of other kinds of rates, like U.S. government bond yields and rates in fixed income markets in other countries. These results provide a great deal of comfort to hedgers: while in theory many factors (and, therefore, securities) might be required to hedge the interest rate risk of a particular portfolio, in practice, three factors cover the vast majority of the risk.

Columns (7) through (9) of Table 10.5 show that the level component is far and away the most important in explaining the volatility of the term structure. The construction of principal components, described in the overview, does ensure that the first component is the most important component, but the extreme dominance of this component is a feature of the data. This finding is useful for thinking about the costs and

benefits of adding a second or third factor to a one-factor hedging framework. Interestingly too, the dominance of the first factor is significantly muted in the very short end of the curve. This implies that hedging one short-term bond with another will not be so effective as hedging one longer-term bond with another. Or, put another way, relatively more factors or hedging securities are needed to hedge portfolios that are concentrated at the short end of the curve. This makes intuitive sense in the context of the extensive information market participants have about near-term events and their effects on rates relative to the information they have on events further into the future.

Hedging with PCA and an Application to Butterfly Weights

A PCA-based hedge for a portfolio would proceed along the lines of the multi-factor approaches. Start with the current price of the portfolio under the current term structure. Then, shift each principal component in turn to obtain new term structures and new portfolio prices. Next, calculate an '01 with respect to each principal component using the difference between the respective shifted price and the original price. Finally, using these portfolio '01s and analogously constructed '01s for a chosen set of hedging securities, find the portfolio of hedging securities that neutralizes the risk of the portfolio to the movement of each PC.

PCA is particularly useful for constructing empirically-based hedges for large portfolios; it is impractical to perform and assess individual regressions for every security in a large portfolio. For illustration purposes, however, this subsection will illustrate how PCA is used, in practice, to hedge a *butterfly* trade. Most typically, butterfly trades use three securities and either buy the security of intermediate maturity and short the *wings* or short the intermediate security and buy the wings.

To take a relatively common butterfly, consider a trader who believes that the 5-year swap rate is too high relative to the 2- and 10-year swap rates and is, therefore, planning to receive in the 5-year and pay in the 2- and 10-year. As of May 28, 2010, the par swap rates and *DV*01s of the swaps of relevant terms are listed in Table 10.6. (The 30-year data will be used shortly.) To calculate the PCA hedge ratios, assume that the trader will receive on 100 notional amount of 5-year swaps and will trade F^2 and F^{10} notional amount of 2- and 10-year swaps. Using the data from Tables 10.5 and 10.6, the equation that neutralizes the overall portfolio's exposure to the level PC is

$$-F^2\frac{.0197}{100}\times 5.86 - F^{10}\frac{.0842}{100}\times 6.35 - 100\times\frac{.0468}{100}\times 6.85 = 0 \tag{10.29}$$

Table 10.6 Par Swap Rates and *DV*01s as of May 28, 2010

Term	Rate	*DV*01
2	1.235%	.0197
5	2.427%	.0468
10	3.388%	.0842
30	4.032%	.1731

Similarly, the equation that neutralizes the overall exposure to the slope PC is

$$-F^2\frac{.0197}{100}\times(-3.09) - F^{10}\frac{.0842}{100}\times .06 - 100 \times\frac{.0468}{100}\times(-1.53) = 0 \tag{10.30}$$

Solving, $F^2 = -120.26$ and $F^{10} = -34.06$ or, in terms of risk weights relative to the *DV*01 of the five-year swap,

$$\frac{120.26\times\frac{.0197}{100}}{.0468} = 50.6\% \tag{10.31}$$

$$\frac{34.06\times\frac{.0842}{100}}{.0468} = 61.3\% \tag{10.32}$$

In words, the *DV*01 of the five-year swap is hedged 50.6% by the two-year swap and 61.3% by the 10-year swap. Note that the sum of the risk weights is not 100%: the hedge neutralizes exposures to the level and slope PCs, not exposures to parallel shifts. To the extent that the term structure changes as assumed, i.e., as some combination of the first two PCs, then the hedge will work exactly. On the other hand, to the extent that the actual change deviates from a combination of these two PCs, the hedge will not, *ex post*, have fully hedged interest rate risk.

Hedging the interest rate risk of the five-year swap with two other swaps in not uncommon, a practice supported by the large fraction of rate variance explained by the first two PCs. A trader might also decide, however, to hedge the third PC as well. A hedge against the first three PCs, found by generalizing the two-security hedge just discussed, gives rise to risk weights of 28.1%, 139.1%, and −67.4% in the 2-, 10-, and 30-year swaps, respectively, i.e., pay in the 2- and 10-year, but receive in the 30-year.

Is hedging the third PC worthwhile? The answer depends on the trader's risk preferences, but the following analysis is useful.

Say that the trader hedges the first two components alone and then the third component experiences a one standard-deviation decrease. The P&L of the trade, per 100 face amount of the 5-year swap, would be

$$\left[-120.26 \times \frac{.0197}{100} \times .59 + 100 \times \frac{.0468}{100} \times (-.57) - 34.06 \times \frac{.0842}{100} \times (-.34)\right] = -.031 \qquad (10.33)$$

or, for a two standard-deviation move, a loss of a bit more than 6 cents per 100 face amount of the 5-year swap. As these two standard deviations of short rate risk equates to not even 1.5 basis points of convergence of the 5-year swap, a trader might very well not bother with this third leg of the hedge.

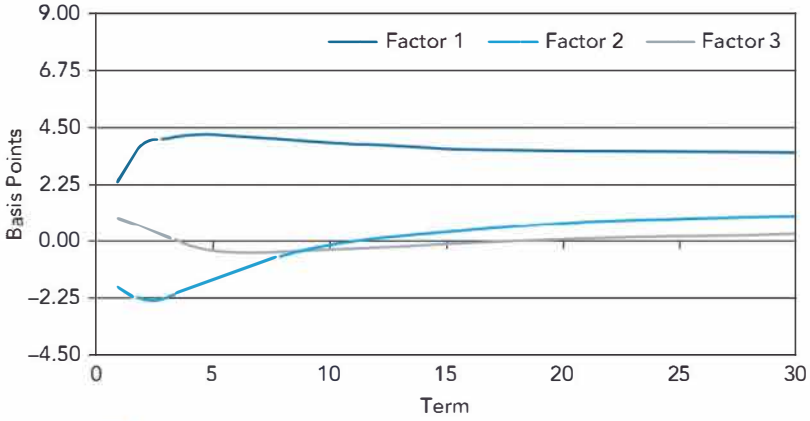


Figure 10.5 The first three principal components from EUR swap rates from October 2001 to October 2008.

Principal Component Analysis of EUR, GBP, and JPY Swap Rates

Figures 10.5 to 10.7 show the first three PCs for the EUR, GBP, and JPY swap rate curves over the same sample period as the USD PCs in Figure 10.4. The striking fact about these graphs is that the shape of the PCs are very much the same across USD, EUR, and GBP. The only significant difference is in magnitudes, with the USD level component entailing larger-sized moves than the level components of EUR and GBP. The PCs of the JPY curve are certainly similar to those of these other countries, but the level component in JPY does not have the same hump: in JPY the first PC does not peak at the five-year maturity point as do the other curves, but increases monotonically with maturity before ultimately leveling off. The significance of this difference in shape will be discussed in the next subsection.

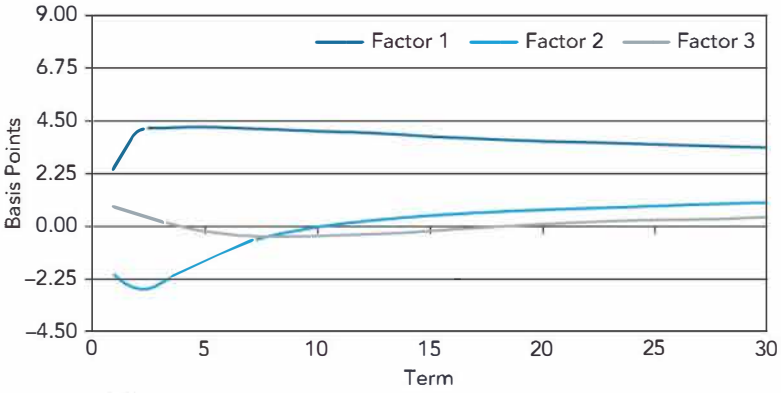


Figure 10.6 The first three principal components from GBP swap rates from October 2001 to October 2008.

The Shape of PCs over Time

As with any empirically based hedging methodology, a decision has to be made about the relevant time period over which to estimate parameters. This is an issue for regression-based methods, as discussed in this chapter, and it is no less an issue for PCA. As will be discussed in this subsection, the qualitative shapes of PCs have, until very recently, remained remarkably stable. This does not imply, however, that differences in PCs estimated over different time periods can be ignored in the sense that they have no important effects on the quality of hedges.

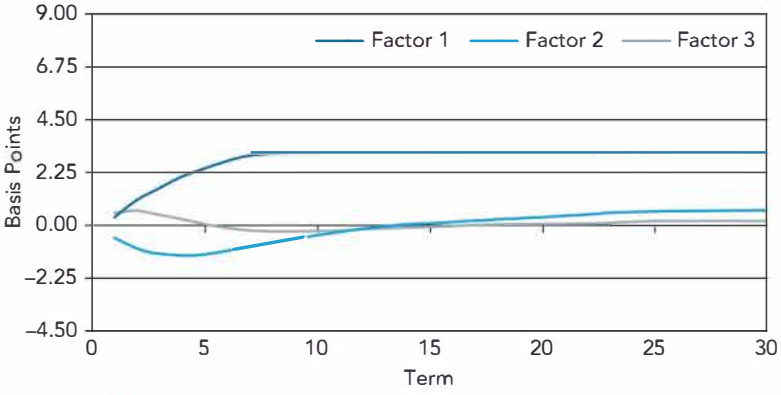


Figure 10.7 The first three principal components from JPY swap rates from October 2001 to October 2008.

But having made this point, the text focuses on the relatively recent changes in the shapes of PCs around the world.

Figure 10.4 showed the first three USD PCs computed over the period 2001 to 2008, but, for quite some time, the qualitative shapes of these PCs was pretty much the same.[7] The volatility of rates has changed over time, and with it the magnitude or height of the PC curves, but the qualitative shapes have not changed much. Most recently, however, there has been a qualitative change to the shape of the first PC in USD, EUR, and GBP. In fact, these shapes have become more like the past shape of the first PC in JPY!

Figures 10.8 and 10.9 contrast the level PC over the historical period October 2001 to October 2008 with that of the post-crisis period, October 2008 to October 2010. Figure 10.8 makes the comparison for USD and EUR while Figure 10.9 does the same for GBP and JPY. The historical maximum of the level PC at a term of about five years in USD, EUR, and GBP has been pushed out dramatically to 10 years and beyond. In fact, these shapes now more closely resemble the level PC of JPY over the earlier estimation period. One explanation for this is the increasing certainty that central banks will maintain easy monetary conditions and low rates for an extended period of time. This dampens the volatility of short- and intermediate-term rates relative to that of longer-term rates, lowers the absolute volatility of short-term rates, and increases the volatility of long-term rates, reflecting the uncertainty of the ultimate results of central bank policy. Meanwhile, the level PC for JPY in the most recent period has become even more pronouncedly upward-sloping, consistent with an even longer period of central-bank control over the short-term rate.

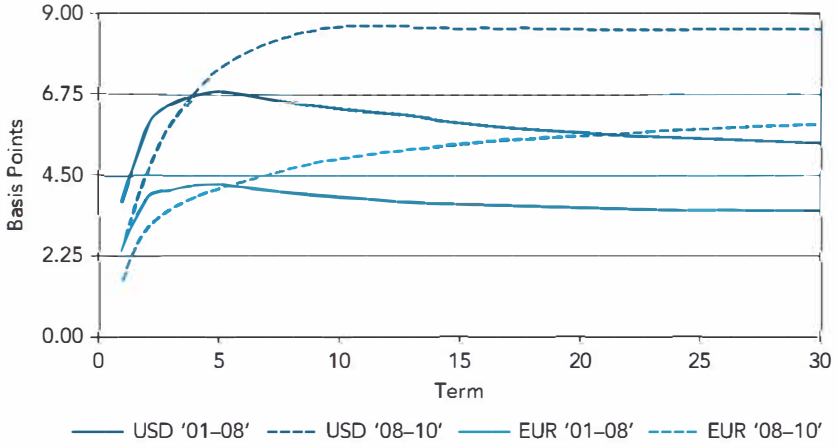


Figure 10.8 The first principal component in USD and EUR swap rates estimated from October 2001 to October 2008 and from October 2008 to October 2010.

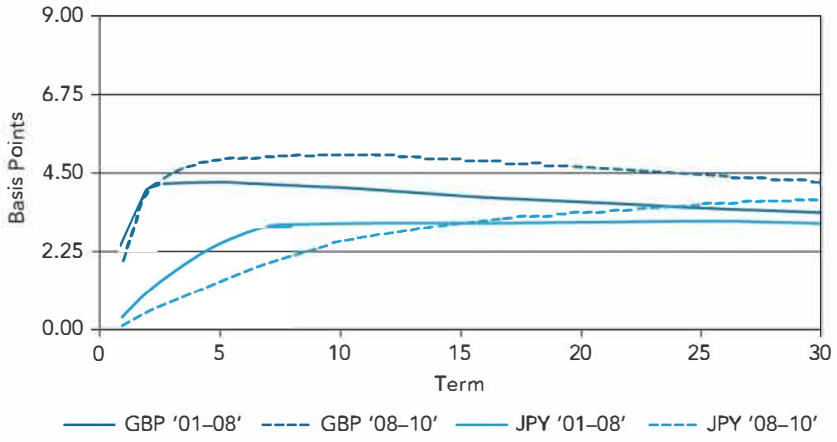


Figure 10.9 The first principal component in GBP and JPY swap rates estimated from October 2001 to October 2008 and from October 2008 to October 2010.

APPENDIX A

The Least-Squares Hedge Minimizes the Variance of the P&L of the Hedged Position

The P&L of the hedged position, given in (10.10) and repeated here, is

$$-F^R \times \frac{DV01^R}{100}\Delta y_t^R - F^N \times \frac{DV01^N}{100}\Delta y_t^N \qquad (10.34)$$

[7] See, for example, Figure 2 of Bulent Baygun, Janet Showers, and George Cherpelis, Salomon Smith Barney, "Principles of Principal Components," January 31, 2000. The shapes of the three PCs in that graph, covering the period from January 1989 to February 1998, are qualitatively extremely similar to those of Figure 10.4 in this chapter.

Let $V(\cdot)$ and $Cov(\cdot,\cdot)$ denote the variance and covariance functions. The variance of the P&L expression in (10.34) is

$$\left(F^R \times \frac{DV01^R}{100}\right)^2 V(\Delta y_t^R) + \left(F^N \times \frac{DV01^N}{100}\right)^2 V(\Delta y_t^N) + 2\left(F^R \times \frac{DV01^R}{100}\right)\left(F^N \times \frac{DV01^N}{100}\right) Cov(\Delta y_t^R, \Delta y_t^N) \quad (10.35)$$

To find the face amount F^R that minimizes this variance, differentiate (10.35) with respect to F^R and set the result to zero:

$$2F^R\left(\frac{DV01^R}{100}\right)^2 V(\Delta y_t^R) + 2F^N \frac{DV01^R}{100}\frac{DV01^N}{100} Cov(\Delta y_t^R, \Delta y_t^N) = 0 \quad (10.36)$$

Then, rearranging terms,

$$F^N \times DV01^N \times \frac{Cov(\Delta y_t^R, \Delta y_t^N)}{V(\Delta y_t^R)} = -F^R \times DV01^R \quad (10.37)$$

But, by the properties of least squares, not derived in this text,

$$\hat{\beta} \equiv \frac{Cov(\Delta y_t^R, \Delta y_t^N)}{V(\Delta y_t^R)} \quad (10.38)$$

Therefore, substituting (10.38) into (10.37) gives the regression hedging rule (10.9) of the text.

APPENDIX B

Constructing Principal Components from Three Rates

The goal of this appendix is to demonstrate the construction and properties of PCs with a minimum of mathematics. To this end, consider three swap rates, the 10-year, 20-year, and 30-year. Over some sample period, the volatilities of these rates, in basis points per day, are 4.25, 4.20, and 4.15. Furthermore, the correlations among these rates are given in the correlation matrix of Table 10.7.

Table 10.7 Correlation Matrix for Swap Rate Example

Term	10-Year	20-Year	30-Year
10-Year	1.00	0.95	0.90
20-Year	0.95	1.00	0.99
30-Year	0.90	0.99	1.00

The combination of data on volatilities and correlations are usefully combined into a *variance-covariance* matrix, denoted by V, where the element in the i^{th} row and j^{th} column gives the covariance of the rate of term i with the rate of term j, or, the correlation of i and j times the standard deviation of i times the standard deviation of j. For example, the covariance of the 20-year swap rate with the 30-year swap rate is $.99 \times 4.20 \times 4.15$, or 17.26. The variance-covariance matrix for the example of this appendix is

$$\mathbf{V} = \begin{pmatrix} 18.06 & 16.96 & 15.87 \\ 16.96 & 17.64 & 17.26 \\ 15.87 & 17.26 & 17.22 \end{pmatrix} \quad (10.39)$$

One use of a variance-covariance matrix is to write succinctly the variance of a particular portfolio of the relevant securities. Consider a portfolio with a total $DV01$ of .50 in the 10-year swap, −1.0 in the 20-year swap, and .60 in the 30-year swap. Without matrix notation, then, the dollar variance of the portfolio, denoted by σ^2 would be given by

$$\begin{aligned}\sigma^2 &= .5^2 4.25^2 + (-1)^2 4.20^2 + .6^2 4.15^2 \\ &\quad + 2 \times .5 \times (-1) \times .95 \times 4.25 \times 4.20 \\ &\quad + 2 \times .5 \times .6 \times .90 \times 4.25 \times 4.15 \\ &\quad + 2 \times (-1) \times .6 \times .99 \times 4.20 \times 4.15 \\ &= .464^2 \end{aligned} \quad (10.40)$$

With matrix notation, letting the transpose of the vector w be $w' = (.5, -1, .6)$, the dollar variance of the portfolio is given more compactly by

$$\mathbf{w'Vw} = (.5, -1, .6)\begin{pmatrix} 18.06 & 16.96 & 15.87 \\ 16.96 & 17.64 & 17.26 \\ 15.87 & 17.26 & 17.22 \end{pmatrix}\begin{pmatrix} .5 \\ -1 \\ .6 \end{pmatrix} \quad (10.41)$$

Finally, note that the sum of the variances of the rates is $4.25^2 + 4.20^2 + 4.15^2 = 52.925$, or, for a measure of total volatility, take the square root of that sum to get 7.27 basis points.

Returning now to principal components, the idea is to create three factors that capture the same information as the variance-covariance matrix. The procedure is as follows. Denote the first principal component by the vector $\mathbf{a} = (a_1, a_2, a_3)'$. Then find the elements of this vector by maximizing $\mathbf{a'Va}$ such that $\mathbf{a'a} = \mathbf{1}$. As mentioned in the PCA overview, this maximization ensures that, among the three PCs to be found, the first PC explains the largest fraction of the variance. The constraint, $\mathbf{a'a} = \mathbf{1}$, along with a similar constraint placed on the other PCs, will ensure that the total variance of the PCs equals the total variance of the underlying data. Performing this maximization, which can be done with the solver in Excel, $\mathbf{a} = (.5758, .5866, .5696)$. Note that the variance of this first component is $\mathbf{a'Va} = 51.041$ which is 51.041/52.925 or 96.44% of the total variance of the rates.

The second principal component, denoted by the vector $\mathbf{b} = (b_1, b_2, b_3)$ is found by maximizing $\mathbf{b}'\mathbf{V}\mathbf{b}$ such that $\mathbf{b}'\mathbf{b} = \mathbf{1}$ and $\mathbf{b}'\mathbf{a} = \mathbf{0}$. The maximization and the first constraint are analogous to those for finding the first principal component. The second constraint requires that the PC **b** is uncorrelated with the first PC, **a**. Solving, gives $\mathbf{b} = (-.7815, .1902, .5941)$. Note that $\mathbf{b}'\mathbf{V}\mathbf{b} = 1.867$ which explains 1.867/52.925 or 3.53% of the total variance of the rates.

Finally, the third PC, denoted by $\mathbf{c} = (c_1, c_2, c_3)$ is found by solving the three equations, $\mathbf{c}'\mathbf{c} = \mathbf{1}$; $\mathbf{c}'\mathbf{a} = \mathbf{0}$; and $\mathbf{c}'\mathbf{b} = \mathbf{0}$. The solution is $\mathbf{c} = (.2402, 2.7872, .5680)$.

As will be clear in a moment, it turns out to be more intuitive to work with a different scaling of the PCs, namely, by multiplying each by its volatility. In the example, this means multiplying the first PC by $\sqrt{51.041}$ or 7.14; the second PC by $\sqrt{1.867}$ or 1.37; and the third by $\sqrt{.017}$ or .13. This gives the PCs, to be denoted $\tilde{\mathbf{a}}$, $\tilde{\mathbf{b}}$, $\tilde{\mathbf{c}}$, as recorded in Table 10.8.

Under this scaling the PCs have a very intuitive interpretation: a one standard-deviation increase of the first PC or factor is a 4.114 basis-point increase in the 10-year rate, a 4.191 basis-point increase in the 20-year rate, and a 4.069 basis-point increase in the 30-year rate. Similarly, a one standard-deviation increase of the second PC is a 1.068 basis-point drop in the 10-year rate, a .260 basis-point increase in the 20-year rate, and a .812 basis-point increase in the 30-year rate. Finally, a one standard-deviation increase of the third PC constitutes changes of .032, –.103, and .075 basis points in each of the rates, respectively.

To appreciate the scaling of the PCs in Table 10.8, note the following implications:

- By construction, the PCs are uncorrelated. Hence, the volatility of the 10-year rate can be recovered from Table 10.8 as

$$\sqrt{4.114^2 + (-1.068)^2 + .032^2} = 4.25 \qquad (10.42)$$

And the volatilities of the 20- and 30-year rates can be recovered equivalently.

- The variance of each PC is the sum of squares of its elements, or, its volatility is the square root of that sum of squares. For the three PCs,

$$\sqrt{4.114^2 + 4.191^2 + 4.069^2} = 7.14 \qquad (10.43)$$

$$\sqrt{(-1.068)^2 + .260^2 + .812^2} = 1.37 \qquad (10.44)$$

$$\sqrt{.032^2 + (-.103)^2 + .075^2} = .13 \qquad (10.45)$$

Table 10.8 Transformed PCs for the Swap Rate Example

Term	1st PC	2nd PC	3rd PC
10-Year	4.114	−1.068	.032
20-Year	4.191	.260	−.103
30-Year	4.069	.812	.075

- The square root of the sum of the variances of the PCs is the square root of the sum of the variances of the rates, which quantity was given above as 7.27 basis points:

$$\sqrt{7.14^2 + 1.37^2 + .13^2} = \sqrt{52.925} = 7.27 \qquad (10.46)$$

- The volatility of any portfolio can be found by computing its volatility with respect to each of the PCs and then taking the square root of the sum of the resulting variances. Returning to the portfolio with *DV*01 weights of $\mathbf{w}' = (.5, -1, .6)$, its volatility with respect to each of the PCs can be computed as in Equations (10.47) through (10.49). Then, adding the sum of these squares and taking the square root, gives a portfolio volatility of .464, as computed earlier from the variances and covariances.

$$\sqrt{(\mathbf{w}'\tilde{\mathbf{a}})^2} = \sqrt{(.5 \times 4.114 - 1 \times 4.191 + .6 \times 4.069)^2} = .3074 \qquad (10.47)$$

$$\sqrt{(\mathbf{w}'\tilde{\mathbf{b}})^2} = \sqrt{(.5 \times (-1.068 - 1 \times .260 + .6 \times .812)^2} = .3068 \qquad (10.48)$$

$$\sqrt{(\mathbf{w}'\tilde{\mathbf{c}})^2} = \sqrt{(.5 \times .032 - 1 \times (-.103) + .6 \times .075)^2} = .1640 \qquad (10.49)$$

In summary, the PCs in Table 10.8 contain the same information as the variances and covariances, but have the interpretation of one standard-deviation changes in the level, slope, and short rate factors. Of course, the power of the methodology is evident not in a simple example like this, but when, as in the text, changes in 30 rates can be adequately expressed with changes in three factors.

11

The Science of 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.
- Distinguish 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, *Third Edition, by Bruce Tuckman and Angel Serrat.*

This chapter uses a very simple setting to show how to price interest rate contingent claims relative to a set of underlying securities by arbitrage arguments. Unlike the arbitrage pricing of securities with fixed cash flows, the techniques of this chapter require strong assumptions about how interest rates evolve in the future. This chapter also introduces *option-adjusted spread* (OAS) as the most popular measure of deviations of market prices from those predicted by models.

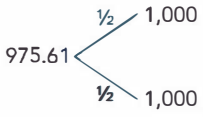
11.1 RATE AND PRICE TREES

Assume that the six-month and one-year spot rates are 5% and 5.15% 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 will be either 4.50% or 5.50% with equal probability. This very strong assumption is depicted by means of a *binomial tree*, where "binomial" means that only two future values are possible:

Note that the columns in the tree represent dates. The six-month rate is 5% today, which will be called date 0. On the next date six months from now, which will be called date 1, there are two possible outcomes or states *of the world*. The 5.50% state will be called the *up-state* while the 4.50% state will be called the *down-state*.

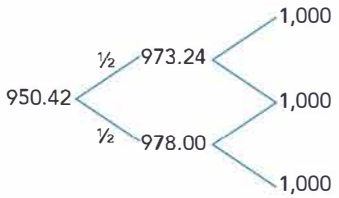
Given the current term structure of spot rates (i.e., the current six-month and one-year rates), trees for the prices of six-month and one-year zero-coupon bonds may be computed. The price tree for USD 1,000 face value of the six-month zero is



since USD 1,000/(1 + .05/2) = USD 975.61. (For easy 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 with the date. On date 0 of the preceding tree the security is a six-month zero, while on date 1 the security is a maturing zero.

The price tree for USD 1,000 face value of a one-year zero is the following:



The three date 2 prices of USD 1,000 are, of course, the maturity values of the one-year zero. The two date 1 prices come from discounting this certain USD 1,000 at the then-prevailing six-month rate. Hence, the date 1 up-state price is USD 1,000/(1 + .05/2) or USD 973.2360, and the date 1 down-state price is USD 1,000/(1 + .045/2) or USD 977.9951. Finally, the date 0 price is computed using the given date 0 one-year rate of 5.15%: USD 1,000/(1 + .0515/2)^2 or 950.423.

The probabilities of moving up or down the tree may be used to compute the average or expected values. As of date 0, the expected value of the one-year zero's price on date 1 is

$$\frac{1}{2}\,\text{USD } 973.24 + \frac{1}{2}\,\text{USD } 978.00 = \text{USD } 975.62 \qquad (11.1)$$

Discounting this expected value to date 0 at the date 0, six-month rate gives an *expected discounted value*[1] of

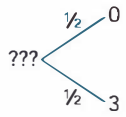
$$\frac{\frac{1}{2}\,\text{USD } 973.24 + \frac{1}{2}\,\text{USD } 978.00}{\left(1 + \frac{0.05}{2}\right)} = \text{USD } 951.82 \qquad (11.2)$$

Note that the one-year zero's expected discounted value of USD 951.82 does not equal its given market price of USD 950.42. 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 USD 973.24 half of the time and USD 978 the other half of the time for an average or expected value of USD 975.62. If investors do not like this price uncertainty, they would prefer a security worth USD 975.62 on date 1 with certainty. More specifically, a security worth USD 975.62 with certainty after six months would sell for USD 975.62/(1 + .05/2) or USD 951.82 as of date 0. By contrast, investors penalize the risky one-year zero-coupon bond with an average price of USD 975.62 after six months by pricing it at USD 950.42. The next chapter elaborates further on investor *risk aversion* and how large an impact it might be expected to have on bond prices.

[1] Over one period, discounting the expected value and taking the expectation of discounted values are the same. But, over many periods the two are different and, with the approach taken by the short rate models, taking the expectation of discounted values is correct—hence the choice of the term "expected discounted value."

11.2 ARBITRAGE PRICING OF DERIVATIVES

The text now turns to the pricing of a derivative security. What is the price of a call option, maturing in six months, to purchase USD 1,000 face value of a then six-month zero at USD 975? Begin with the price tree for this call option:



If on date 1 the six-month rate is 5.50% and a six-month zero sells for USD 973.23, the right to buy that zero at USD 975 is worthless. On the other hand, if the six-month rate turns out to be 4.50% and the price of a six-month zero is USD 978, then the right to buy the zero at USD 975 is worth USD 978 – USD 975 or USD 3. This description of the option's terminal payoffs emphasizes the derivative nature of the option: its value depends on the value of an underlying security.

A security is priced by arbitrage by finding and pricing its replicating portfolio. When, as in that context, cash flows do not depend on the levels of rates, the construction of the replicating portfolio is relatively simple. The derivative context is more difficult because cash flows do depend on the levels of rates, and the replicating portfolio must replicate the derivative security for any possible interest rate scenario.

To price the option by arbitrage, construct a portfolio on date 0 of underlying securities, namely six-month and one-year zero-coupon bonds, that will be worth USD 0 in the up-state on date 1 and USD 3 in the down-state. To solve this problem, let $F^{.5}$ and F^{1} be the face values of six-month and one-year zeros in the replicating portfolio, respectively. Then, these values must satisfy the following two equations:

$$F^{.5} + .97324F^{1} = \text{USD } 0 \qquad (11.3)$$

$$F^{.5} + .97800F^{1} = \text{USD } 3 \qquad (11.4)$$

Equation (11.3) may be interpreted as follows. In the up-state, 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 .97324 per dollar face value. Hence, the left-hand side of Equation (11.3) denotes the value of the replicating portfolio in the up-state. This value must equal USD 0, the value of the option in the up-state. Similarly, Equation (11.4) requires that the value of the replicating portfolio in the down-state equal the value of the option in the down-state.

Solving Equations (11.3) and (11.4), $F^{.5} = -\text{USD } 613.3866$ and $F^{1} = +\text{USD } 630.2521$. In words, on date 0 the option can be replicated by buying about USD 630.25 face value of one-year zeros and simultaneously shorting about USD 613.39 face amount of six-month zeros. Since this is the case, the law of one price requires that the price of the option equal the price of the replicating portfolio. But this portfolio's price is known and is equal to

$$.97561F^{.5} + .95042F^{1} = -.97561 \times \text{USD } 613.3866 + .95042 \times \text{USD } 630.2521 = \text{USD } .58 \qquad (11.5)$$

Therefore, the price of the option must be USD .58.

Recall that pricing based on the law of one price is enforced by arbitrage. If the price of the option were less than USD .58, 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 USD.58, 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 USD.58.

It is important to emphasize that the option cannot be priced by expected discounted value. Under that method, the option price would appear to be

$$\frac{.5 \times \text{USD } 0 + .5 \times \text{USD } 3}{1 + \frac{.05}{2}} = \text{USD } 1.46 \qquad (11.6)$$

The true option price is less than this value 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 magnitude of this effect is discussed in the next chapter.

This section illustrates arbitrage pricing with a call option, but it should be clear that arbitrage can be used to price any security with cash flows that depend on the six-month rate. Consider, for example, a security that, in six months, requires a payment of USD 200 in the up-state but generates a payment of USD 1,000 in the down-state. Proceeding as in the option example, find the portfolio of six-month and one-year zeros that replicates these two terminal payoffs, price this replicating portfolio as of date 0, and conclude that the price of the hypothetical security equals the price of the replicating portfolio.

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 (11.3) to (11.5). The explanation for this somewhat surprising observation follows from the principles of arbitrage. Arbitrage pricing requires that the value of the replicating portfolio matches 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 up-state 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 hence 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 5.50% 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, however, probabilities are not needed to derive arbitrage-free prices.

11.3 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. Since the original interest rate process has to be modified only once, and since 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. The price of the one-year zero is USD 950.42, computed from the given one-year spot rate of 5.15%. At the same time, the expected discounted value of the one-year zero is USD 951.82, as derived in Equation (11.2) and reproduced here:

$$\frac{\frac{1}{2}\,\text{USD } 973.24 + \frac{1}{2}\,\text{USD } 978.00}{\left(1 + \frac{.05}{2}\right)} = \text{USD } 951.82 \qquad (11.7)$$

The probabilities of ½ for the up and down-states are the assumed true or real-world probabilities. But there are other probabilities, called *risk-neutral* probabilities, that 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:

$$\frac{\text{USD } 973.24p + \text{USD } 978.00(1 - p)}{1 + \frac{.05}{2}} = \text{USD } 950.42 \quad (11.8)$$

The solution is $p = .8024$. In words, under the risk-neutral probabilities of .8024 and .1976 the expected discounted value equals the market price.

In later chapters the difference between true and risk-neutral probabilities is 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 5% to 5.50% and a 50% chance that it falls from 5% to 4.50%. 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 80.24% chance of a 50-basis point increase in the six-month rate and a 19.76% chance of a 50-basis point decline for an expected change of 30.24 basis points. Hence the drift of the six-month rate under these probabilities is 30.24 basis points.

As pointed out in the previous section, the expected discounted value of the option payoff is USD 1.46, while the arbitrage price is USD.58. But what if expected discounted value is computed using the risk-neutral probabilities? The resulting option value would be:

$$\frac{.8024 \times \text{USD } 0 + .1976 \times \text{USD } 3}{1 + \frac{.05}{2}} = \text{USD.58} \qquad (11.9)$$

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 price of the underlying securities with their expected discounted values. (In the simple example of this chapter 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 will describe intuitively why risk-neutral pricing works. Since the argument is a bit complex, it is broken up into four steps.

Step 1: Given trees for the underlying securities, the price of a security that is priced by arbitrage does not depend on investors' risk preferences. This assertion can be supported as follows.

A security is priced by arbitrage if one can construct a portfolio that replicates its cash flows. Under the assumed process for interest rates in this chapter, for example, the sample 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 in a single common stock's market value.

If a security is priced by arbitrage and everyone agrees on the price evolution of the underlying securities, then everyone will agree on the replicating portfolio. In the option example, both an extremely risk-averse, retired investor and a professional gambler would agree that a portfolio of USD 630.25 face of one-year zeros and −USD 613.39 face of six-month zeros replicates the option. And since they agree on the composition of the replicating portfolio and on the prices of the underlying securities, they must also agree on the price of the derivative.

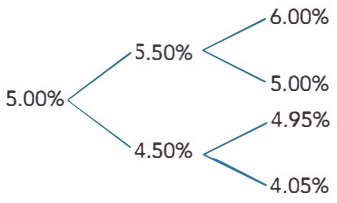
Step 2: Imagine an economy identical to the true economy with respect to current bond prices and the possible value of the six-month rate over time but different in that the investors in the imaginary economy are risk neutral. Unlike investors in the true economy, investors in the imaginary economy do not penalize securities for risk and, therefore, price securities by expected discounted value. It follows that, under the probabilities in the imaginary economy, the expected discounted value of the one-year zero equals its market price. But these probabilities satisfy Equation (11.8), namely the risk-neutral probabilities of .8024 and .1976.

Step 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 up-state in that economy is .8024, the price of the option in that economy is given by Equation (11.9) and is, therefore, USD .58.

Step 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. It follows that since 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 equal USD .58, the option price in the imaginary economy. More generally, the price of a derivative in the real economy may be computed by expected discounted value under the risk-neutral probabilities.

11.4 ARBITRAGE PRICING IN A MULTI-PERIOD SETTING

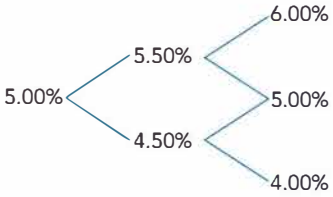
Maintaining the binomial assumption, the tree of the previous section might be extended for another six months as follows:



When, as in this tree, an up move followed by a down move does not give the same rate as a down move followed by an up move, the tree is said to be *nonrecombining*. From an economic perspective, there is nothing wrong with this kind of tree. To justify this particular tree, for example, one might argue that when short rates are 5% or higher they tend to change in increments of 50 basis points. But when rates fall below 5%, the size of the change starts to decrease. In particular, at a rate of 4.50%, the short rate may change by only 45 basis points. A volatility process that depends on the level of rates exhibits *state-dependent* volatility.

Despite the economic reasonableness of nonrecombining trees, practitioners tend to avoid them because such trees are 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. So, for example, a tree with semiannual steps large enough to price 10-year securities will, in its rightmost column alone, have over 500,000 nodes, while a tree used to price 20-year securities will in its rightmost column have over 500 billion nodes. Furthermore, as discussed later in the chapter, it is often desirable to reduce substantially the time interval between dates. In short, even with modern computers, trees that grow this quickly are computationally unwieldy. This doesn't mean, by the way, that the effects that give rise to nonrecombining trees, like state-dependent volatility, have to be abandoned. It simply means that these effects must be implemented in a more efficient way.

Trees in which the up-down and down-up-states have the same value are called *recombining trees*. An example of this type of tree that builds on the two-date tree of the previous sections is

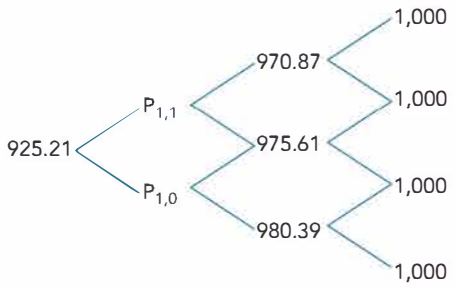


Note that there are two nodes after six months, three after one year, and so on. A tree with weekly rather than semiannual steps capable of pricing a 30-year security would have only $52 \times 30 + 1$ or 1,561 nodes in its rightmost column. Evidently, recombining trees are much more manageable than nonrecombining trees from a computational viewpoint.

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 the preceding tree the six-month rate on date 2, state 0 is 4%. The six-month rate on state 1 of date 1 is 5.50%.

Continuing where the option example left off, having derived the risk-neutral tree for the pricing of a one-year zero, the goal is to extend the tree for the pricing of a 1.5-year zero assuming

that the 1.5-year spot rate is 5.25%. Ignoring the probabilities for a moment, several nodes of the 1.5-year zero price tree can be written down immediately:



On date 3, the zero with an original term of 1.5 years matures and is worth its face value of USD 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 6%, 5%, and 4% in states 2, 1, and 0, respectively:

$$\frac{\text{USD } 1{,}000}{1 + \frac{.06}{2}} = \text{USD } 970.87 \qquad (11.10)$$

$$\frac{\text{USD } 1{,}000}{1 + \frac{.05}{2}} = \text{USD } 975.61 \qquad (11.11)$$

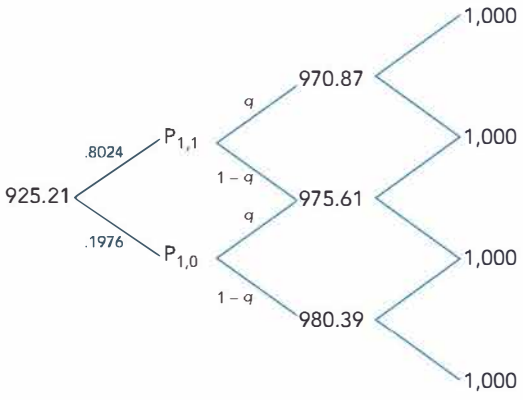
$$\frac{\text{USD } 1{,}000}{1 + \frac{.04}{2}} = \text{USD } 980.39 \qquad (11.12)$$

Finally, on date 0, the 1.5-year zero equals its face value discounted at the given 1.5-year spot rate:

$$\frac{\text{USD } 1{,}000}{\left(1 + \frac{.0525}{2}\right)^3} = \text{USD } 925.21 \qquad (11.13)$$

The prices of the zero on date 1 in states 1 and 0 are denoted $P_{1,1}$ and $P_{1,0}$ respectively. The then one-year zero prices are not known because, at this point in the development, possible values of the one-year rate in six months are not available.

The previous section showed that the risk-neutral probability of an up move on date 0 is .8024. Letting q be the risk-neutral probability of an up move on date 1,[2] the tree becomes



[2] For simplicity alone, this example assumes that the probability of moving up from state 0 equals the probability of moving up from state 1.

By definition, expected discounted value under risk-neutral probabilities must produce market prices. With respect to the 1.5-year zero price on date 0, this requires that

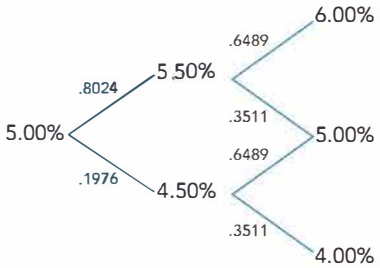
$$\frac{.8024P_{1,1} + .1976P_{1,0}}{1 + \frac{.05}{2}} = \text{USD } 925.21 \qquad (11.14)$$

With respect to the prices of a then one-year zero on date 1,

$$P_{1,1} = \frac{\text{USD } 970.87q + \text{USD } 975.61(1 - q)}{1 + \frac{.055}{2}} \qquad (11.15)$$

$$P_{1,0} = \frac{\text{USD } 975.61q + \text{USD } 980.39(1 - q)}{1 + \frac{.045}{2}} \qquad (11.16)$$

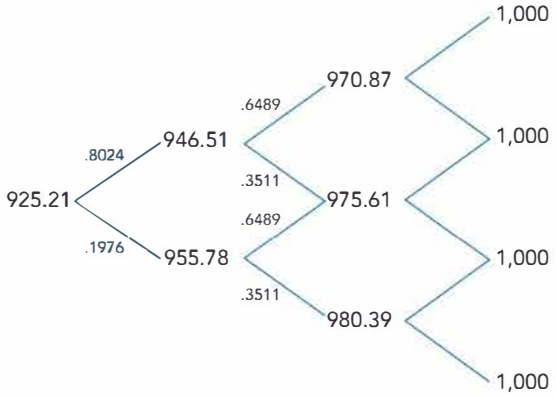
While Equations (11.14) through (11.16) may appear complicated, substituting (11.15) and (11.16) into (11.14) results in a linear equation in the one unknown, q. Solving this resulting equation reveals that $q = .6489$. Therefore, the risk-neutral interest rate process may be summarized by the following tree:



Furthermore, any derivative security 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 appears 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 64.89% chance of a 50-basis point increase in the six-month rate and a 35.11% chance of a 50-basis point decline in the rate. These numbers give a drift or expected change of 14.89 basis points.

Substituting $q = .6489$ back into Equations (11.15) and (11.16) completes the tree for the price of the 1.5-year zero:



It follows immediately from this tree that the one-year spot rate six months from now may be either 5.5736% or 4.5743% since

$$\text{USD } 946.51 = \frac{\text{USD } 1{,}000}{(1 + \frac{5.5736\%}{2})^2} \tag{11.17}$$

$$\text{USD } 955.78 = \frac{\text{USD } 1{,}000}{(1 + \frac{4.5743\%}{2})^2} \tag{11.18}$$

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 .5-, 1-, and 1.5-year spot rates as well as an assumption about the evolution of the six-month rate over the next year. But since 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 sufficient to determine the possible values of the one-year spot rate in six months. Considering this fact from another point of view, having specified initial spot rates 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 because the model presented here 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.

Just as some replicating portfolio can reproduce the cash flows of a security from date 0 to date 1, some other replicating portfolios can reproduce the cash flows of a security from date 1 to date 2. The composition of these replicating portfolios depends on the date and state. More specifically, the replicating portfolios held on date 0, on state 0 of date 1, and on state 1 of date 1 are usually different. From the trading perspective, the replicating portfolio must be adjusted as time passes and as interest rates change. This process is known as *dynamic replication*, in contrast to the *static replication strategies*. As an example of static replication, the portfolio of zero-coupon bonds that replicates a coupon bond does not change over time nor with the level of rates.

Having built a tree out to date 2 it should be clear how to extend the tree to any number of dates. Assumptions about the future possible values of the short-term rate have to be extrapolated further into the future and risk-neutral probabilities have to be calculated to recover a given set of bond prices.

11.5 EXAMPLE: PRICING A CONSTANT-MATURITY TREASURY SWAP

Equipped with the last tree of interest rates in the previous section, this section prices a particular derivative security, namely USD 1,000,000 face value of a stylized *constant-maturity Treasury (CMT) swap* struck at 5%. This swap pays

$$\text{USD } 1{,}000{,}000\frac{y_{CMT} - 5\%}{2} \tag{11.19}$$

every six months until it matures, where y_{CMT} is a semiannually compounded yield, of a predetermined maturity, on the payment date. The text prices a one-year CMT swap on the six-month yield. In practice, CMT swaps trade most commonly on the yields of the most liquid maturities, i.e., on 2-, 5- and 10-year yields.

Since six-month semiannually compounded yields equal six-month spot rates, rates from the tree of the previous section can be substituted into (11.19) to calculate the payoffs of the CMT swap. On date 1, the state 1 and state 0 payoffs are, respectively,

$$\text{USD } 1{,}000{,}000\frac{5.50\% - 5\%}{2} = \text{USD } -2{,}500 \tag{11.20}$$

$$\text{USD } 1{,}000{,}000\frac{4.50\% - 5\%}{2} = \text{USD } 2{,}500 \tag{11.21}$$

Similarly on date 2, the state 2, 1, and 0 payoffs are, respectively,

$$\text{USD } 1{,}000{,}000\frac{6\% - 5\%}{2} = \text{USD } -5{,}000 \tag{11.22}$$

$$\text{USD } 1{,}000{,}000\frac{5\% - 5\%}{2} = \text{USD } 0 \tag{11.23}$$

$$\text{USD } 1{,}000{,}000\frac{6\% - 5\%}{2} = \text{USD } -5{,}000 \tag{11.24}$$

The possible values of the CMT swap at maturity, on date 2, are given by Equations (11.22) through (11.24). 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 (11.20) and (11.21). The resulting date 1 values in states 1 and 0, respectively, are

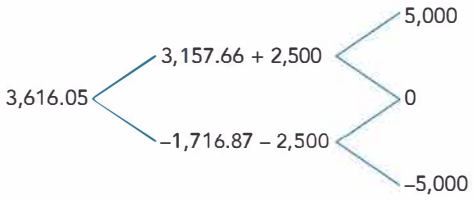
$$\frac{.6489 \times \text{USD } 5{,}000 + .3511 \times \text{USD } 0}{1 + \frac{.05}{2}} + \text{USD } 2{,}500 = \text{USD } 5{,}657.66 \tag{11.25}$$

$$\frac{.6489 \times 0 + .3511 \times (-\text{USD } 5{,}000)}{1 + \frac{.045}{2}} - \text{USD } 2{,}500 = -\text{USD } 4{,}216.87 \tag{11.26}$$

Finally, the value of the swap on date 0 is the expected discounted value of the date 1 payoffs, given by (11.25) and (11.26), under the risk-neutral probabilities:

$$\frac{.8024 \times \text{USD } 5{,}657.66 + .1976 \times (-\text{USD } 4{,}216.87)}{1 + \frac{.05}{2}} = \text{USD } 3{,}616.05 \tag{11.27}$$

The following tree summarizes the value of the stylized CMT swap over dates and states:



A value of USD 3,616.05 for the CMT swap might seem surprising at first. After all, the cash flows of the CMT swap are zero at a rate of 5%, and 5% is, under the real probabilities, the average rate on each date. The explanation, of course, is that the risk-neutral probabilities, not the real probabilities, determine the arbitrage price of the swap. The expected discounted value of the swap under the real probabilities can be computed by following the steps leading to (11.25) through (11.27) but using .5 for all up and down moves. The result of these calculations does give a value close to zero, namely −USD 5.80.

The expected cash flow of the CMT swap on both dates 1 and 2, under the real probabilities, is zero. It follows immediately that the discounted value of these expected cash flows is zero. At the same time, the expected discounted value of the CMT swap is −USD 5.80.

11.6 OPTION-ADJUSTED SPREAD

Option-adjusted spread (OAS) 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 equals its model price when discounted values are computed at risk-neutral rates plus that spread. To illustrate, say that the market price of the CMT swap in the previous section is USD 3,613.25, USD 2.80 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 5.5% and 4.5% in Equations (11.25) and (11.26), respectively, to get new swap values of

$$\frac{.6489 \times \text{USD } 5{,}000 + .3511 \times \text{USD } 0}{1 + \frac{.056}{2}} + \text{USD } 2{,}500 = \text{USD } 5{,}656.13 \tag{11.28}$$

$$\frac{.6489 \times 0 + .3511 \times (-\text{USD } 5{,}000)}{1 + \frac{.046}{2}} - \text{USD } 2{,}500 = -\text{USD } 4{,}216.03 \tag{11.29}$$

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 (11.20) through (11.24).

Completing the valuation with an OAS of 10 basis points, use the results of (11.28) and (11.29) and a discount rate of 5% plus the OAS spread of 10 basis points, or 5.10%, to obtain an initial CMT swap value of

$$\frac{.8024 \times \text{USD } 5{,}656.13 + .1976 \times (-\text{USD } 4{,}216.03)}{1 + \frac{.051}{2}} = \text{USD } 3{,}613.25 \tag{11.30}$$

Hence, as claimed, discounting at the risk-neutral rates plus an OAS of 10 basis points produces a model price equal to the given market price of USD 3,613.25.

If a security's OAS is positive, its market price is less than its model price, so 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

$$\frac{.8024 \times \text{USD } 5{,}657.66 - .1976 \times \text{USD } 4{,}216.87 - \text{USD } 3{,}616.05}{\text{USD } 3{,}616.05} = 2.5\% \tag{11.31}$$

which is six month's worth of the initial rate of 5%. On the other hand, the expected return of the cheap swap, with an OAS of 10 basis points, is

$$\frac{.8024 \times \text{USD } 5{,}656.13 - .1976 \times \text{USD } 4{,}216.03 - \text{USD } 3{,}613.25}{\text{USD } 3{,}613.25} = 2.55\% \tag{11.32}$$

which is six month's worth of the initial rate of 5% plus the OAS of 10 basis points, or half of 5.10%.

11.7 PROFIT AND LOSS ATTRIBUTION WITH AN OAS

We introduced profit and loss (P&L) attribution. This section gives a mathematical description of attribution in the context of term structure models and of securities that trade with an OAS.

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

value of x can be written as $P_t(x, OAS)$. Using a first-order Taylor approximation, the change in the price of the security is

$$dP = \frac{\partial P}{\partial x}dx + \frac{\partial P}{\partial t}dt + \frac{\partial P}{\partial OAS}dOAS \tag{11.33}$$

Dividing by the price and taking expectations,

$$E\left[\frac{dP}{P}\right] = \frac{1}{P}\frac{\partial P}{\partial x}E[dx] + \frac{1}{P}\frac{\partial P}{\partial t}dt \tag{11.34}$$

Since the OAS calculation assumes that OAS is constant over the life of the security, moving from (11.33) to (11.34) 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,[3] then, for any security priced according to the model,

$$E\left[\frac{dP}{P}\right] = rdt \tag{11.35}$$

But Equation (11.35) 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 (11.35), is

$$E\left[\frac{dP}{P}\right] = (r + OAS)dt \tag{11.36}$$

Combining these pieces, substitute (11.34) and (11.36) into (11.33) and rearrange terms to break down the return of a security into its component parts:

$$\frac{dP}{P} = (r + OAS)dt + \frac{1}{P}\frac{\partial P}{\partial x}(dx - E[dx]) + \frac{1}{P}\frac{\partial P}{\partial OAS}dOAS \tag{11.37}$$

Finally, multiplying through by P,

$$dP = (r + OAS)Pdt + \frac{\partial P}{\partial x}(dx - E[dx]) + \frac{\partial P}{\partial OAS}dOAS \tag{11.38}$$

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 (11.38) represent, in order, carry-roll-down,[4] gains or losses from rate changes, and gains or losses from spread change. For models with predictive power, the OAS converges or tends to zero, or, equivalently, the security price converges or tends toward its fair value according to the model.

The decompositions (11.37) and (11.38) highlight the usefulness of OAS as a measure of the value of a security with respect to a particular model. According to the model, 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 the extent of any convergence.

The decomposition equations also provide a framework for thinking about 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. In terms of the decompositions, $\partial P/\partial x = 0$. In that case, the expected return or P&L depends only on the short-term rate, the OAS, and any convergence. Furthermore, if the trader finances the trade at the short-term rate, i.e., borrows P at a rate r to purchase the security, the expected return is simply equal to the OAS plus any convergence return.

11.8 REDUCING THE TIME STEP

To this point this chapter has assumed that the time elapsed between dates of the tree is six months. The methodology outlined previously, however, can be easily adapted 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 must be 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 5% means dividing by $1 + .05/12$

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

[3] Taking expected values with respect to the true probabilities would add a risk premium term to the right-hand side of this equation. See Chapter 12.

[4] For expositional simplicity, no explicit coupon or other direct cash flows have been included in this discussion.

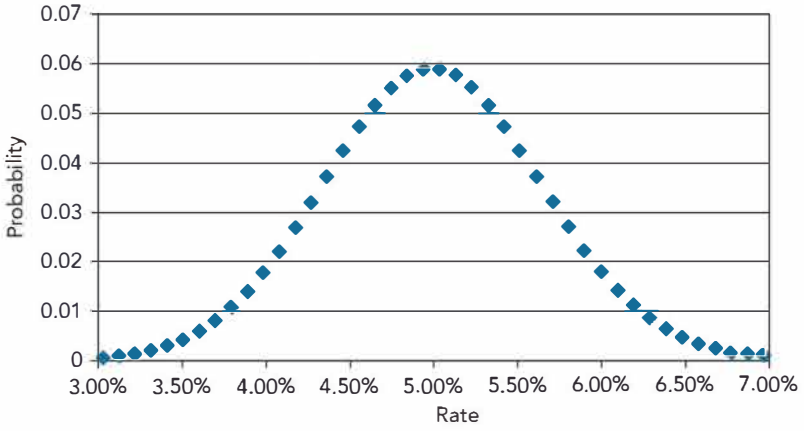


Figure 11.1 **Sample probability distribution of the six-month rate in six months with daily time steps.**

for many practical pricing problems. Reducing the step size can fill the tree with enough rates to price contingent claims with sufficient accuracy. Figure 11.1 illustrates this point by showing a relatively realistic-looking probability distribution of the six-month rate in six months from a tree with daily time steps, a drift of zero, and a horizon standard deviation of 65 basis points.

While smaller time steps generate more realistic interest rate distributions, it is not the case that smaller time steps are always desirable. First, the greater the number of computations in pricing a security, the more attention must be paid to numerical issues like round-off error. Second, since decreasing the time step increases computation time, practitioners requiring quick results cannot make the time step too small. Customers calling market makers in options on swaps, or *swaptions*, for example, expect price quotations within minutes if not sooner. Hence, the time step in a model used to price swaptions must be consistent with the market maker's required response time.

The best choice of step size ultimately depends on the problem at hand. When pricing a 30-year callable bond, for example, a model with monthly time steps may provide a realistic enough interest rate distribution to generate reliable prices. The same monthly steps, however, will certainly be inadequate to price a one-month bond option: that tree would imply only two possible rates on the option expiration date.

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.

11.9 FIXED INCOME VERSUS EQUITY DERIVATIVES

While the ideas behind pricing fixed income and equity derivatives are similar in many ways, there are important differences as well. In particular, it is worth describing why models created for the stock market cannot be adopted without modification for use in fixed income markets.

The famous Black-Scholes-Merton pricing analysis of stock options can be summarized as follows. Under the assumption 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 stocks and short-term bonds that replicates the payoffs of an option. Therefore, by arbitrage arguments, the price of the option must equal the known price of the replicating portfolio.

Say that an investor wants to price an option on a five-year bond by a direct application of this logic. The investor would have to begin by making an assumption about how the price of the five-year bond evolves over time. But this is considerably more complicated than making assumptions about how the price of a stock evolves over time. First, the price of a bond must converge to its face value at maturity while the random process describing the stock price need not be constrained in any similar way. Second, because of the maturity constraint, the volatility of a bond's price must eventually get smaller as the bond approaches maturity. The simpler assumption that the volatility of a stock is constant is not so appropriate for bonds. Third, since stock volatility is very large relative to short-term rate volatility, it may be relatively harmless to assume that the short-term rate is constant. By contrast, it can be difficult to defend the assumption that a bond price follows some random process while the short-term interest rate is constant.[5]

These objections led researchers to make assumptions about the random evolution of the interest rate rather than of the bond price. In that way bond prices would naturally approach par, price volatilities would naturally approach zero, and the interest rate would not be assumed to be constant. But this approach raises another set of questions. Which interest rate

[5] Because these three objections are less important in the case of short-term options on long-term bonds, practitioners do use stock-like models in this fixed income context. Also, it is often sufficient to assume, somewhat more satisfactorily, that the relevant discount factor is uncorrelated with the price of the underlying bond.

is assumed to evolve in a particular way? Making assumptions about the 5-year rate over time is not particularly helpful for two reasons. First, 5-year coupon bond prices depend on shorter-term rates as well. Second, pricing an option on a 5-year bond requires assumptions about the bond's future possible prices. But knowing the 5-year rate over time is insufficient because, in a very short time, the option's underlying security will no longer be a 5-year bond. Therefore, one must often make assumptions about the evolution of the entire term structure of interest rates to price bond options and other derivatives. In the one-factor case described in this chapter it has been shown that modeling the evolution of the short-term rate is sufficient, combined with arbitrage arguments, to build a model of the entire term structure. In short, despite the enormous importance of the Black-Scholes-Merton analysis, the fixed income context does demand special attention.

Having reached the conclusion at the end of the previous paragraph, there are some contexts in which practitioners invoke assumptions so that the Black-Scholes-Merton models can be applied in place of more difficult-to-implement term structure models.

12

The Evolution of Short Rates 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.
- Evaluate the impact of changes in maturity, yield, and volatility on the convexity of a security.
- Calculate the price and return of a zero-coupon bond incorporating a risk premium.

Excerpt is Chapter 8 of Fixed Income Securities: Tools for Today's Markets, *Third Edition, by Bruce Tuckman and Angel Serrat.*

This chapter presents a framework for understanding the shape of the term structure. In particular, it is shown how spot or forward rates are determined by expectations of future short-term rates, the volatility of short-term rates, and an interest rate risk premium. To conclude the chapter, this framework is applied to swap curves in the United States and Japan.

12.1 INTRODUCTION

From assumptions about the interest rate process for the short-term rate and from an initial term structure implied by market prices, Chapter 11 showed how to derive a risk-neutral process that can be used to price all fixed income securities by arbitrage. Models that follow this approach, i.e., models that take the initial term structure as given, are called *arbitrage-free* models. A different approach, however, is to start with assumptions about the interest rate process and about the risk premium demanded by the market for bearing interest rate risk and then derive the risk-neutral process. Models of this sort do not necessarily match the initial term structure and are called *equilibrium* models.[1]

This chapter describes how assumptions about the interest rate process and about the risk premium determine the level and shape of the term structure. For equilibrium models, an understanding of the relationships between the model assumptions and the shape of the term structure is important in order to make reasonable assumptions in the first place. For arbitrage-free models, an understanding of these relationships reveals the assumptions implied by the market through the observed term structure.

Many economists might find this chapter remarkably narrow. An economist asked about the shape of the term structure would undoubtedly make reference to such macroeconomic factors as the marginal productivity of capital, the propensity to save, and expected inflation. The more modest goal of this chapter is to connect the dynamics of the short-term rate of interest and the risk premium with the shape of the term structure. While this goal does fall short of answers that an economist might provide, it is more ambitious than the derivation of arbitrage restrictions on bond and derivative prices given underlying bond prices.

[1] This nomenclature is somewhat misleading. Equilibrium models, in the context of their assumptions, which do not include market prices for the initial term structure, are also arbitrage-free.

12.2 EXPECTATIONS

The word *expectations* implies uncertainty. Investors might expect the one-year rate to be 10%, but know there is a good chance it will turn out to be 8% or 12%. For the purposes of this section alone, the text assumes away uncertainty so that the statement that investors expect or forecast a rate of 10% means that investors assume that the rate will be 10%. The sections to follow reintroduce uncertainty.

To highlight the role of interest rate forecasts in determining the shape of the term structure, consider the following simple example. The one-year interest rate is currently 10% and all investors forecast that the one-year interest rate next year and the year after will also be 10%. In that case, investors will discount cash flows using forward rates of 10%. In particular, the price of one-, two- and three-year zero-coupon bonds per dollar face value (using annual compounding) will be

$$P^1 = \frac{1}{1.10} \tag{12.1}$$

$$P^2 = \frac{1}{(1.10)(1.10)} = \frac{1}{1.10^2} \tag{12.2}$$

$$P^3 = \frac{1}{(1.10)(1.10)(1.10)} = \frac{1}{1.10^3} \tag{12.3}$$

From inspection of Equations (12.1) through (12.3), the term structure of spot rates in this example is flat at 10%. Very simply, investors are willing to lock in 10% for two or three years because they assume that the one-year rate will always be 10%.

Now assume that the one-year rate is still 10%, but that all investors forecast the one-year rate next year to be 12% and the one-year rate in two years to be 14%. In that case, the one-year spot rate is still 10%. The two-year spot rate, $\hat{r}(2)$, is such that

$$P^2 = \frac{1}{(1.10)(1.12)} = \frac{1}{(1 + \hat{r}(2))^2} \tag{12.4}$$

Solving, $\hat{r}(2) = 10.995\%$. Similarly, the three-year spot rate, $\hat{r}(3)$, is such that

$$P^3 = \frac{1}{(1.10)(1.12)(1.14)} = \frac{1}{(1 + \hat{r}(3))^3} \tag{12.5}$$

Solving, $\hat{r}(3) = 11.998\%$. Hence, the evolution of the one-year rate from 10% to 12% to 14% generates an upward-sloping term structure of spot rates: 10%, 10.995%, and 11.988%. In this case, investors require rates above 10% when locking up their money for two or three years because they assume one-year rates will be higher than 10%. No investor, for example, would buy a two-year zero at a yield of 10% when it is possible to buy a one-year zero at 10% and, when it matures, buy another one-year zero at 12%.

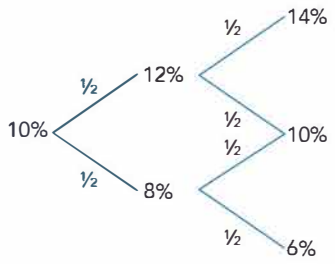
Finally, assume that the one-year rate is 10%, but that investors forecast that it will fall to 8% in one year and to 6% in two years. In that case, it is easy to show that the term structure of spot rates will be downward-sloping. In particular, $\hat{r}(1) = 10\%$, $\hat{r}(2) = 8.995\%$, and $\hat{r}(3) = 7.988\%$.

These simple examples reveal that expectations can cause the term structure to take on any of a myriad of shapes. Over short horizons, the financial community can have very specific views about future short-term rates. Over longer horizons, however, expectations cannot be so granular. It would be difficult, for example, to defend the position that the expectation for the one-year rate 29 years from now is substantially different from the expectation of the one-year rate 30 years from now. On the other hand, an argument can be made that the long-run expectation of the short-term rate is 5%: 3% due to the long-run real rate of interest and 2% due to long-run inflation. Hence, forecasts can be very useful in describing the shape and level of the term structure over short-term horizons and the level of rates at very long horizons. This conclusion has important implications for extracting expectations from observed interest rates and for choosing among term structure models.

12.3 VOLATILITY AND CONVEXITY

This section drops the assumption that investors believe that their forecasts will be realized and assumes instead that investors understand the volatility around their expectations. To isolate the implications of volatility on the shape of the term structure, this section assumes that investors are risk-neutral so that they price securities by expected discounted value. The next section drops this assumption.

Assume that the following tree gives the true process for the one-year rate:



Note that the expected interest rate on date 1 is $.5 \times 8\% + .5 \times 12\%$ or 10% and that the expected rate on date 2 is $.25 \times 14\% + .5 \times 10\% + .25 \times 6\%$ or 10%. In the previous section, with no volatility around expectations, flat expectations of 10% imply a flat term structure of spot rates. That is not the case in the presence of volatility.

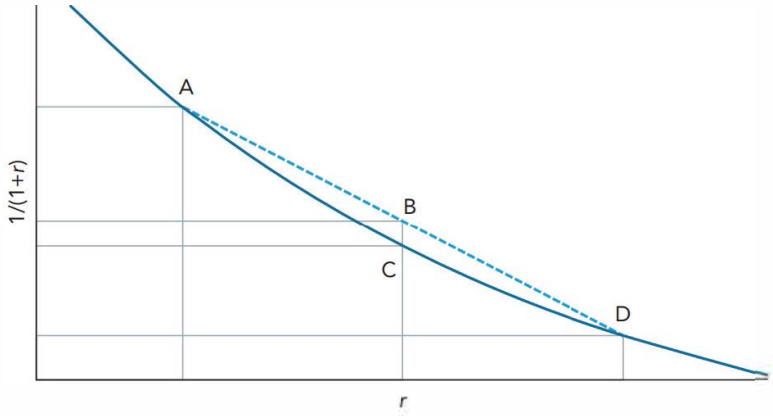
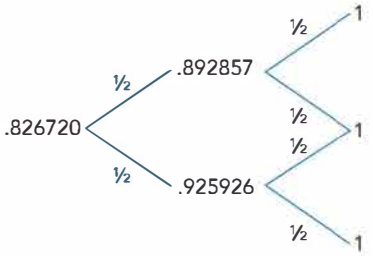


Figure 12.1 An illustration of convexity.

The price of a one-year zero is, by definition, $1/1.10$ or .909091, implying a one-year spot rate of 10%. Under the assumption of risk-neutrality, the price of a two-year zero may be calculated by discounting the terminal cash flow using the preceding interest rate tree:



Hence, the two-year spot rate is such that $.82672 = (1 + \hat{r}(2))^{-2}$, implying that $\hat{r}(2) = 9.982\%$.

Even though the one-year rate is 10% and the expected one-year rate in one year is 10%, the two-year spot rate is 9.982%. The 1.8-basis point difference between the spot rate that would obtain in the absence of uncertainty, 10%, and the spot rate in the presence of volatility, 9.982%, is the effect of convexity on that spot rate. This convexity effect arises from the mathematical fact, a special case of *Jensen's Inequality*, that

$$E\left[\frac{1}{1+r}\right] > \frac{1}{E[1+r]} = \frac{1}{1+E[r]} \qquad (12.6)$$

Figure 12.1 graphically illustrates this equation. There are two possible values of r and, consequently, of the function $1/1+r$ in the figure,[2] shown as points A and D. The height or vertical-axis

[2] The curve shown is actually a power of $1/1+r$; i.e., the price of a longer-term zero-coupon bond, so that the curvature is more visible.

coordinate of point B is the average of these two function values. Under the assumption that the two possible values of r occur with equal probability, this average can be thought of as $E[1/(1+r)]$ in (12.6). And under the same assumption, the horizontal-axis coordinates of the points B and C can be thought of as $E[r]$ so that the height of point C can be thought of as $1/(1+E[r])$ Clearly, the height of B is greater than that of C, or $E[1/(1+r)] > 1/(1+E[r])$. To summarize, Equation (12.6) is true because the pricing function of a zero-coupon bond, $1/(1+r)$, is convex rather than concave.

Returning to the example of this section, Equation (12.6) may be used to show why the one-year spot rate is less than 10%. The spot rate one year from now may be 12% or 8%. According to (12.6),

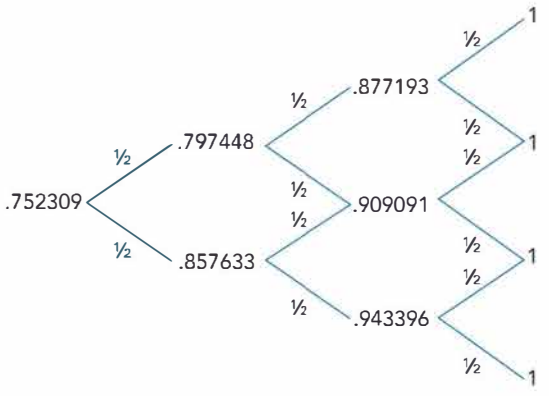
$$.5 \times \frac{1}{1.12} + .5 \times \frac{1}{1.08} > \frac{1}{.5 \times 1.12 + .5 \times 1.08} = \frac{1}{1.10} \quad (12.7)$$

Dividing both sides by 1.10,

$$\frac{1}{1.10}\left[.5 \times \frac{1}{1.12} + .5 \times \frac{1}{1.08} \right] > \frac{1}{1.10^2} \quad (12.8)$$

The left-hand side of (12.8) is the price of the two-year zero-coupon bond today. In words, then, Equation (12.8) says that the price of the two-year zero is greater than the result of discounting the terminal cash flow by 10% over the first period and by the expected rate of 10% over the second period. It follows immediately that the yield of the two-year zero, or the two-year spot rate, is less than 10%.

The tree presented at the start of this section may also be used to price a three-year zero. The resulting price tree is

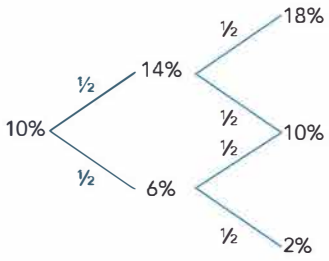


The three-year spot rate, such that $.752309 = (1 + \hat{r}(3))^{-3}$, is 9.952%. Therefore, the value of convexity in this spot rate is 10% − 9.952% or 4.8 basis points, whereas the value of convexity in the two-year spot rate was only 1.8 basis points.

It is generally true that, all else equal, the value of convexity increases with maturity. This will become evident shortly. For now, suffice it to say that the convexity of the pricing function of a zero maturing in N years, $(1 + r)^{-N}$, increases with N. In terms of Figure 12.1, the longer the maturity of the illustrated pricing function, the more convex the curve.

Securities with greater convexity perform better when yields change a lot and perform worse when yields do not change by much. The discussion in this section shows that convexity does, in fact, lower bond yields. The mathematical development in a later section ties these observations together by showing exactly how the advantages of convexity are offset by lower yields.

The previous section assumes no interest rate volatility and, consequently, yields are completely determined by forecasts. In this section, with the introduction of volatility, yield is reduced by the value of convexity. So it may be said that the value of convexity arises from volatility. Furthermore, the value of convexity increases with volatility. In the tree introduced at the start of the section, the standard deviation of rates is 200 basis points a year.[3] Now consider a tree with a standard deviation of 400 basis points:



The expected one-year rate in one year and in two years is still 10%. Spot rates and convexity values for this case may be derived along the same lines as before. Figure 12.2 graphs three term structures of spot rates: one with no volatility around the expectation of 10%; one with a volatility of 200 basis points a year (the tree of the first example); and one with a volatility of 400 basis points per year (the tree preceding this paragraph). Note that the value of convexity, measured by the distance between the rates assuming no volatility and the rates assuming volatility, increases with volatility. Figure 12.2 also illustrates that the value of convexity increases with maturity.

For very short terms and realistic levels of volatility, the value of convexity is quite small. But since simple examples must rely on short terms, convexity effects would hardly be discernible without raising volatility to unrealistic levels. Therefore, this section had to make use of unrealistically high volatility. The application at the end of this chapter uses realistic volatility levels to present typical convexity values.

[3] Chapter 13 describes the computation of the standard deviation of rates implied by an interest rate tree.

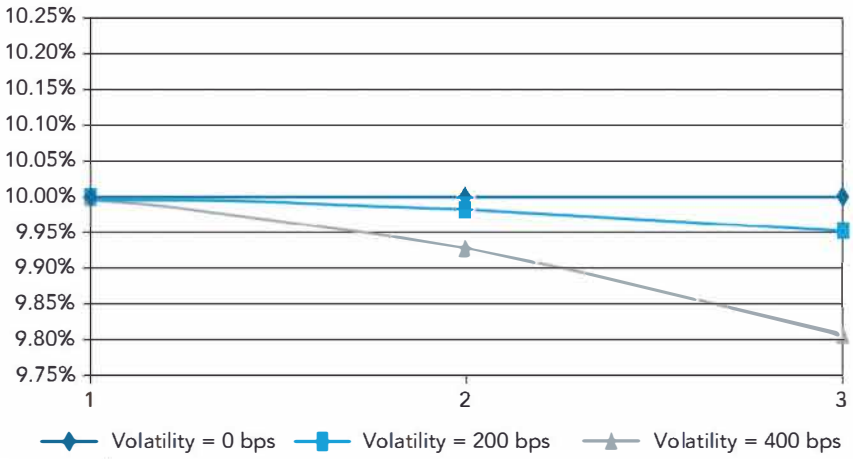


Figure 12.2 **Volatility and the shape of the term structure in three-date binomial models.**

12.4 RISK PREMIUM

To illustrate the effect of risk premium on the term structure, consider again the second interest rate tree presented in the preceding section, with a volatility of 400 basis points per year. Risk-neutral investors would price a two-year zero by the following calculation:

$$.827541 = \frac{.5\left[\frac{1}{1.14} + \frac{1}{1.06}\right]}{1.10} = \frac{.5[.877193 + .943396]}{1.10} \qquad (12.9)$$

By discounting the expected future price by 10%, Equation (12.9) implies that the expected return from owning the two-year zero over the next year is 10%. To verify this statement, calculate this expected return directly:

$$.5\frac{.877193 - .827541}{.827541} + .5\frac{.943396 - .827541}{.827541} = .5 \times 6\% + .5 \times 14\% = 10\% \qquad (12.10)$$

Would investors really invest in this two-year zero offering an expected return of 10% over the next year? The return will, in fact, be either 6% or 14%. While these two returns do average to 10%, an investor could, instead, buy a one-year zero with a certain return of 10%. Presented with this choice, any risk-averse investor would prefer an investment with a certain return of 10% to an investment with a risky return that averages 10%. In other words, investors require compensation for bearing interest rate risk.[4]

Risk-averse investors demand a return higher than 10% for the two-year zero over the next year. This return can be effected by pricing the zero-coupon bond one year from now at less than the prices of $1/1.14$ and $1/1.06$. Equivalently, future cash flows could be discounted at rates higher than the possible rates of 14% and 6%. The next section shows that adding, for example, 20 basis points to each of these rates is equivalent to assuming that investors demand an extra 20 basis points for each year of duration risk. Assuming this is indeed the fair market *risk premium*, the price of the two-year zero would be computed as follows:

$$.826035 = \frac{.5\left[\frac{1}{1.142} + \frac{1}{1.062}\right]}{1.10} \qquad (12.11)$$

The price in (12.11) is below the value obtained in (12.9) which assumes that investors are risk-neutral. Put another way, the increase in the discounting rates has increased the expected return of the two-year zero. In one year, if the interest rate is 14%, then the price of a one-year zero will be $1/1.14$ or .877193. If the rate is 6%, then the price will be $1/1.06$ or .943396. Therefore, the expected return of the two-year zero priced at .826035 is

$$\frac{.5[.877193 + .943396] - .826035}{.826035} = 10.20\% \qquad (12.12)$$

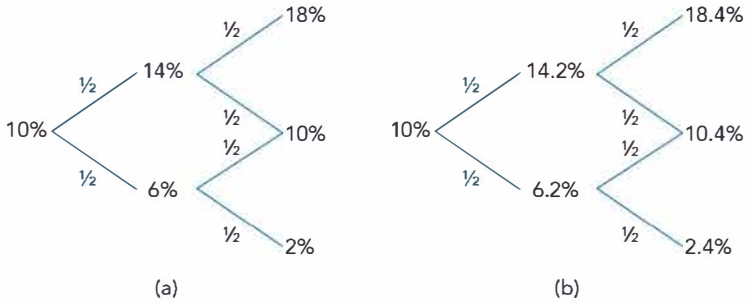
Hence, recalling that the one-year zero has a certain return of 10%, the risk-averse investors in this example demand 20 basis points in expected return to compensate them for the one year of duration risk inherent in the two-year zero.[5]

Continuing with the assumption that investors require 20 basis points for each year of duration risk, the three-year zero, with its approximately two years of duration risk,[6] needs to offer an expected return of 40 basis points. The next section shows that this return can be effected by pricing the three-year zero as if rates next year are 20 basis points above their true values and as if rates the year after next are 40 basis points above their true values. To summarize, consider trees (a) and (b) below. If tree (a) depicts the actual or true interest rate process, then pricing with tree (b) provides investors with a risk premium of 20 basis points for each year of duration risk. If this risk premium is, in fact, embedded in market prices, then, by definition, tree (b) is the risk-neutral interest rate process.

[4] This is an oversimplification. See the discussion at the end of the section.

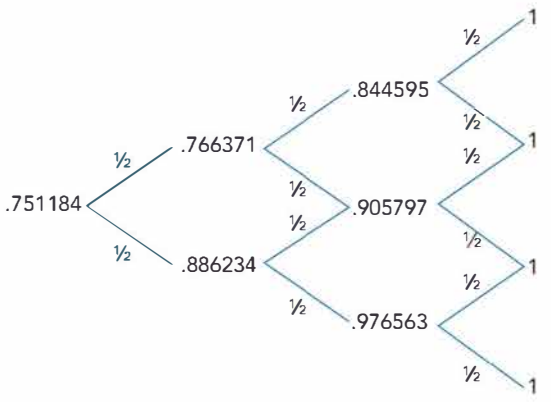
[5] The reader should keep in mind that a two-year zero has one year of interest rate risk only in this stylized example: it has been assumed that rates can move only once a year. In reality, rates can move at any time, so a two-year zero has two full years of interest rate risk.

[6] A three-year zero has two years of interest rate risk only in this stylized example. See the previous footnote.



The text now verifies that pricing the three-year zero with the risk-neutral process does offer an expected return of 10.4%, assuming that rates actually move according to the true process.

The price of the three-year zero can be computed by discounting using the risk-neutral tree:



To find the expected return of the three-year zero over the next year, proceed as follows. Two years from now the three-year zero will be a one-year zero with no interest rate risk.[7] Therefore, its price will be determined by discounting at the actual interest rate at that time: $1/1.18$ or .847458, $1/1.10$ or .909091, and $1/1.02$ or .980392. One year from now, however, the three-year zero will be a two-year zero with one year of duration risk. Therefore, its price at that time will be determined by using the risk-neutral rates of 14.20% and 6.20%. In particular, the two possible prices of the three-year zero in one year are

$$.769067 = \frac{.5(.847458 + .909091)}{1.142} \qquad (12.13)$$

and

$$.889587 = \frac{.5(.909091 + .980392)}{1.062} \qquad (12.14)$$

[7] Once again, this is an artifact of this example in which rates change only once a year.

Finally, then, the expected return of the three-year zero over the next year is

$$\frac{.5(.769067 + .889587) - .751184}{.751184} = 10.40\% \qquad (12.15)$$

To summarize, in order to compensate investors for two years of duration risk, the return on the three-year zero is 40 basis points above a one-year zero's certain return of 10%.

Continuing with the assumption of 400-basis-point volatility, Figure 12.3 graphs the term structure of spot rates for three cases: no risk premium, a risk premium of 20 basis points per year of duration risk, and a risk premium of 40 basis points. In the case of no risk premium, the term structure of spot rates is downward-sloping due to convexity. A risk premium of 20 basis points pushes up spot rates while convexity pulls them down. In the short end, the risk premium effect dominates and the term structure is mildly upward-sloping. In the long end, the convexity effect dominates and the term structure is mildly downward-sloping. The next section clarifies why risk premium tends to dominate in the short end while convexity tends to dominate in the long end. Finally, a risk premium as large as 40 basis points dominates the convexity effect and the term structure of spot rates is upward-sloping. The convexity effect is still evident, however, from the fact that the curve increases more rapidly from one to two years than from two to three years.

Just as the section on volatility uses unrealistically high levels of volatility to illustrate its effects, this section uses unrealistically high levels of the risk premium to illustrate its effects. The application at the end of this chapter focuses on reasonable magnitudes for the various effects in the context of the USD and JPY swap markets.

Before closing this section, a few remarks on the sources of an interest rate risk premium are in order. Asset pricing theory (e.g., the Capital Asset Pricing Model, or CAPM) teaches that assets whose returns are positively correlated with aggregate wealth or consumption will earn a risk premium. Consider, for example, a traded stock index. That asset will almost certainly do well if the economy is doing well and poorly if the economy is doing poorly. But investors, as a group, already have a lot of exposure to the economy. To entice them to hold a little more of the economy in the form of a traded stock index requires the payment of a risk premium; i.e., the index must offer an expected return greater than the risk-free rate of return. On the other hand, say that there exists an asset that is negatively correlated with the economy. Holdings in that asset allow investors to reduce their exposure to the economy. As a result, investors would accept an expected return on that asset below the risk-free rate of return. That asset, in other words, would have a negative risk premium.

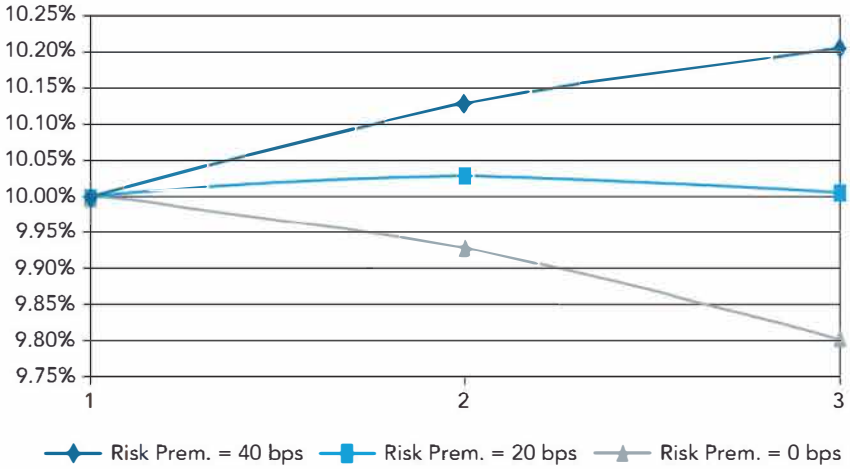


Figure 12.3 **Volatility, risk premium, and the shape of the term structure in three-date binomial models.**

This section assumes that bonds with interest rate risk earn a risk premium. In terms of asset pricing theory, this is equivalent to assuming that bond returns are positively correlated with the economy or, equivalently, that falling interest rates are associated with good times. One argument supporting this assumption is that interest rates fall when inflation and expected inflation fall and that low inflation is correlated with good times.

The concept of a risk premium in fixed income markets has probably gained favor more for its empirical usefulness than for its theoretical solidity. On average, over the past 75 years, the term structure of interest rates has sloped upward.[8] While the market may from time to time expect that interest rates will rise, it is hard to believe that the market expects interest rates to rise on average. Therefore, expectations cannot explain a term structure of interest rates that, on average, slopes upward. Convexity, of course, leads to a downward-sloping term structure. Hence, of the three effects described in this chapter, only a positive risk premium can explain a term structure that, on average, slopes upward.

An uncomfortable fact, however, is that over earlier time periods the term structure has, on average, been flat.[9] Whether this means that an interest rate risk premium is a relatively recent phenomenon that is here to stay or that the experience of persistently upward-sloping curves is only partially due to a risk premium is a question beyond the scope of this book. In short, the theoretical and empirical questions with respect to the existence of an interest rate risk premium have not been settled.

[8] See, for example, Homer, S., and Richard Sylla, *A History of Interest Rates*, 3rd Edition, Revised, Rutgers University Press, 1996, pp. 394-409.

[9] Ibid.

13

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.*

Chapters 11 and 12 show that 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.

13.1 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:

$$dr = \sigma dw \tag{13.1}$$

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]

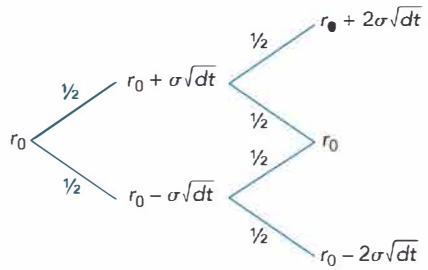
Say, for example, that the current value of the short-term rate is 6.18%, that volatility equals 113 basis points per year, and that the time interval under consideration is one month or ½ years. Mathematically, $r_0 = 6.18\%$; $\sigma = 1.13\%$; and $dt = 1/12$. A month passes and the random variable 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 (13.1) is

$$dr = 1.13\% \times .15 = .17\% \tag{13.2}$$

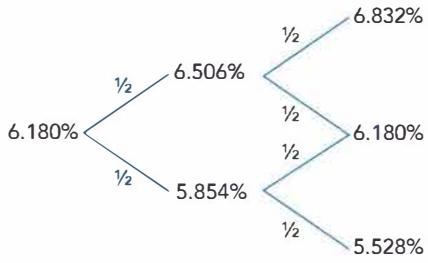
or 17 basis points. Since the short-term rate started at 6.18%, the short-term rate after a month is 6.35%.

Since the expected value of dw is zero, (13.1) 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 (13.1) 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 (13.1). 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 (13.1), 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

$$E[dr] = .5 \times \sigma\sqrt{dt} + .5 \times -\sigma\sqrt{dt} = 0 \tag{13.3}$$

The variance of the rate, often denoted $V[dr]$, from date 0 to date 1 is computed as follows:

$$\begin{aligned} V[dr] &= E[dr^2] - \{E[dr]\}^2 \\ &= .5 \times (\sigma\sqrt{dt})^2 + .5 \times (-\sigma\sqrt{dt})^2 - 0 \\ &= \sigma^2 dt \end{aligned} \tag{13.4}$$

Note that the first line of (13.4) 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 (13.3) and (13.4) show that the drift and volatility implied by the tree match the drift and volatility of the interest rate process (13.1). 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.

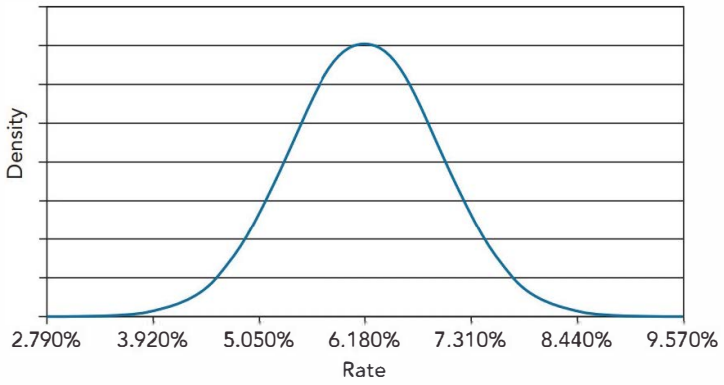


Figure 13.1 Distribution of short rates after one year, Model 1.

to 6.35%. In the tree, however, the only two possible rates are 6.506% and 5.854%. Nevertheless, as shown in Chapter 9, after a sufficient number of time steps the branches of the tree used to approximate the process (13.1) will be numerous enough to approximate a normal distribution. Figure 13.1 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 13.1, 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 (13.1) 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 14, 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, the techniques of Chapter 11 may be used to price fixed coupon bonds. Figure 13.2 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 13.2 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.

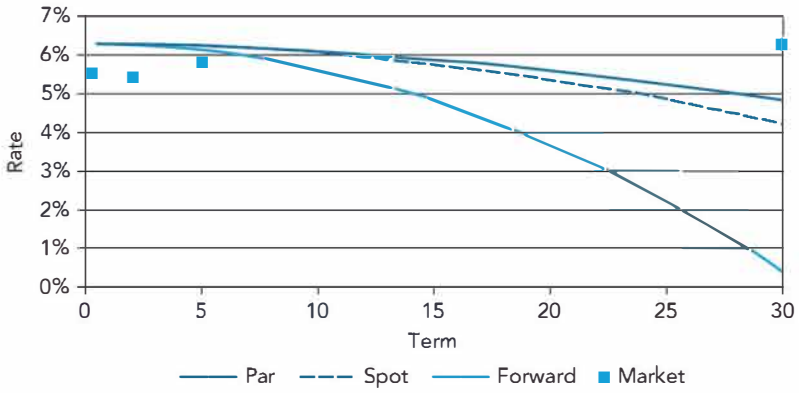


Figure 13.2 **Rate curves from Model 1 and selected market swap rates, February 16, 2001.**

Table 13.1 **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

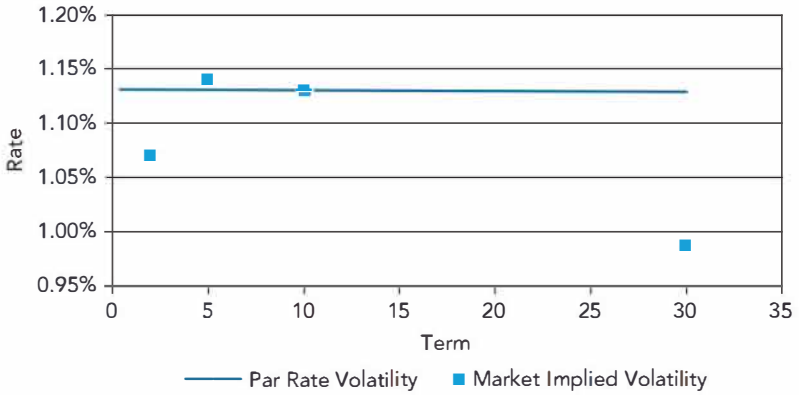


Figure 13.3 **Par rate volatility from Model 1 and selected implied volatilities, February 16, 2001.**

The model term structure is downward-sloping. As the model has no drift, rates decline with term solely because of convexity. Table 13.1 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 13.3, this implication fails to capture the implied volatility structure in the market. The volatility data in Figure 13.3 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 14.

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 14). Therefore, Model 1 is a model of parallel shifts.

13.2 MODEL 2: DRIFT AND RISK PREMIUM

The term structures implied by Model 1 always look like Figure 13.2: relatively flat for early terms and then downward sloping. Chapter 12 pointed out that the 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

$$dr = \lambda dt + \sigma dw \tag{13.5}$$

The process (13.5) 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.

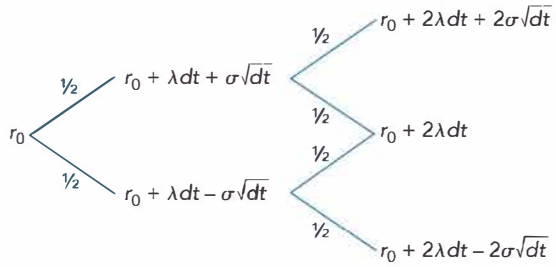
realization of the random variable dw is again .15 over a month, then the change in rate is

$$dr = .229\% \times \frac{1}{12} + 1.10\% \times .15 = .1841\% \qquad (13.6)$$

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. As discussed in Chapter 12, 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 (13.5).

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 13.4 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 13.2. 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, as argued in Chapter 12, 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.

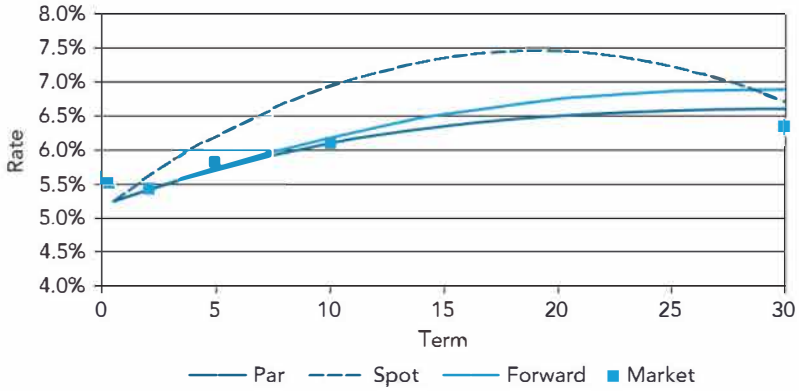


Figure 13.4 **Rate curves from Model 2 and selected market swap rates, February 16, 2001.**

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.

13.3 THE HO-LEE MODEL: TIME-DEPENDENT DRIFT

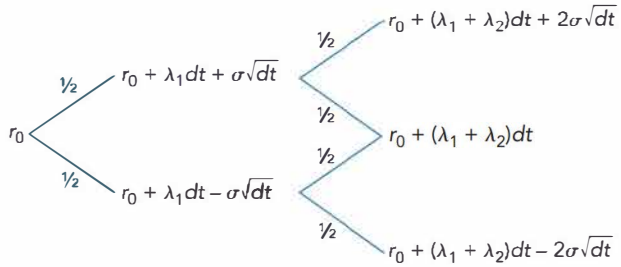
The dynamics of the risk-neutral process in the Ho-Lee model are written as

$$dr = \lambda_t dt + \sigma dw \qquad (13.7)$$

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 procedure is very much like that used to construct the trees in Chapter 11. The only difference is that Chapter 11 adjusts the probabilities to match the spot rate curve while this section adjusts the rates. As it turns out, the two procedures are equivalent so long as the step size is small enough.

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.

13.4 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.

13.5 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

$$dr = k(\theta - r)dt + \sigma dw \tag{13.8}$$

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 (13.8) 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:

$$\begin{aligned} dr &= k(r_\infty - r)dt + \lambda dt + \sigma dw \\ &= k\left(\left[r_\infty + \frac{\lambda}{k}\right] - r\right)dt + \sigma dw \end{aligned} \qquad (13.9)$$

The process in (13.8) is identical to that in (13.9) so long as

$$\theta \equiv r_\infty + \frac{\lambda}{k} \qquad (13.10)$$

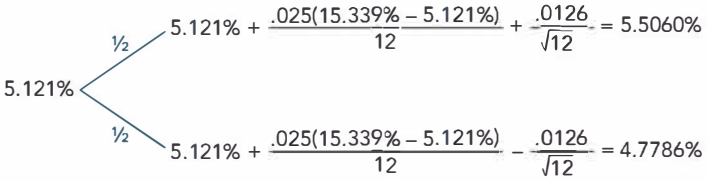
Note that very many combinations of r_∞ and λ give the same θ and, through the risk-neutral process (13.8), the same market prices.

For the purposes of this section, let $k = .025$, $\sigma = 126$ basis points per year, $r_\infty = 6.179\%$, and $\lambda = .229\%$. According to (13.10), then, $\theta = 15.339\%$. With these parameters, the process (13.8) says that over the next month the expected change in the short rate is

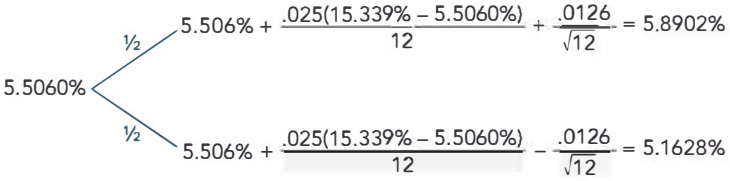
$$.025 \times (15.339\% - 5.121\%)\frac{1}{12} = .0213\% \qquad (13.11)$$

or 2.13 basis points. The volatility over the next month is $126 \times \sqrt{\frac{1}{12}}$ or 36.4 basis points.

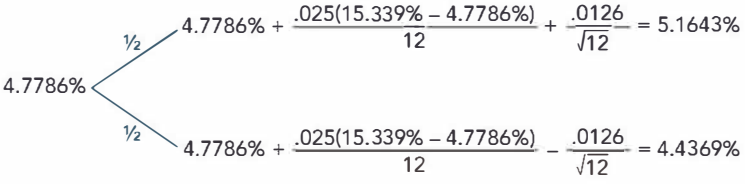
Representing this process with a tree is not quite so straightforward as the simpler processes described previously because the most obvious representation leads to a nonrecombining tree. Over the first time step,



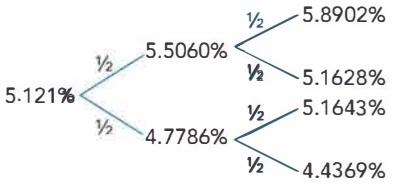
To extend the tree from date 1 to date 2, start from the up state of 5.5060%. The tree branching from there is



while the tree branching from the date 1 down-state of 4.7786% is



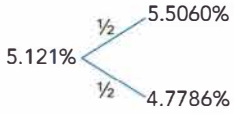
To summarize, the most straightforward tree representation of (13.8) takes the following form:



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:



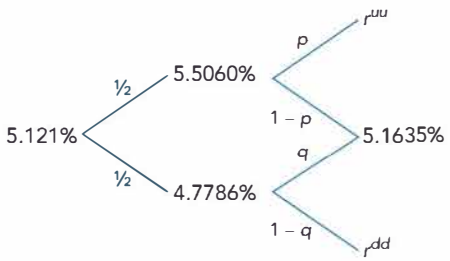
Next, fix the center node of the tree on date 2. Since the expected perturbation due to volatility over each time step is zero, the drift alone determines the expected value of the process after each time step. After the first time step, the expected value is

$$5.121\% + .025\,(15.339\% - 5.121\%)\frac{1}{2} = 5.1423\% \qquad (13.12)$$

After the second time step, the expected value is

$$5.1423\% + .025\,(15.339\% - 5.1423\%)\frac{1}{12} = 5.1635\% \qquad (13.13)$$

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 (13.8) and the parameter values set in this section, the expected rate and standard deviation of the rate from 5.5060% are, respectively,

$$5.5060\% + .025\,(15.339\% - 5.5060\%)\,\frac{1}{12} = 5.5265\% \qquad (13.14)$$

and

$$1.26\%\sqrt{\frac{1}{12}} = .3637\% \qquad (13.15)$$

For the recombining tree to match this expectation and standard deviation, it must be the case that

$$p \times r^{uu} + (1 - p) \times 5.1635\% = 5.5265\% \qquad (13.16)$$

and, by the definition of standard deviation,

$$\sqrt{p(r^{uu} - 5.5265\%)^2 + (1 - p)(5.6135\% - 5.5265\%)^2} = .3637\% \qquad (13.17)$$

Solving Equations (13.16 and (13.17), $r^{uu} = 5.8909\%$ and $p = .4990$.

The same procedure may be followed to compute r^{dd} and q. The expected rate from 4.7786% is

$$4.7786\% + .025\,(15.339\% - 4.7786\%)\,\frac{1}{12} = 4.8006\%. \qquad (13.18)$$

and the standard deviation is again 36.37 basis points. Starting from 4.7786%, then, it must be the case that

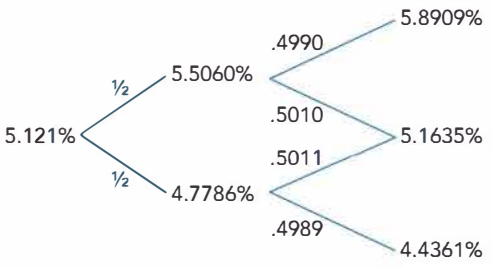
$$q \times 5.1635\% + (1 - q) \times r^{dd} = 4.8006\% \qquad (13.19)$$

and

$$\sqrt{q(5.1635\% - 4.8006\%)^2 + (1 + q)(r^{dd} - 4.8006\%)^2} = .3637\% \qquad (13.20)$$

Solving Equations (13.19) and (13.20), $r^{dd} = 4.4361\%$ and $q = .5011$.

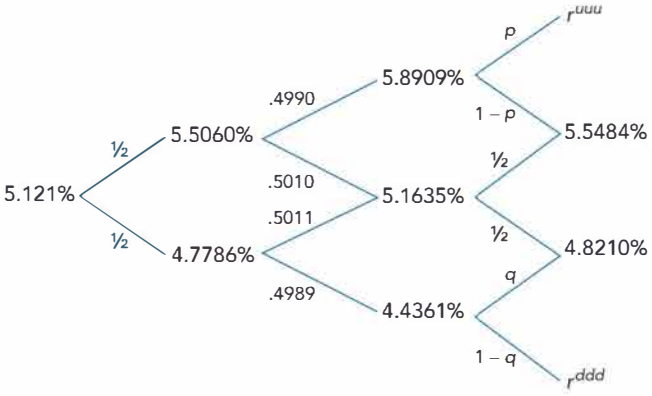
Putting the results from the up- and downstates together, a recombining tree approximating the process (13.8) with the parameters of this section is



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

$$5.1635\% + .025 \times (15.339\% - 5.1635\%)\frac{1}{12} = 5.1847\% \qquad (13.21)$$

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 13.5 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

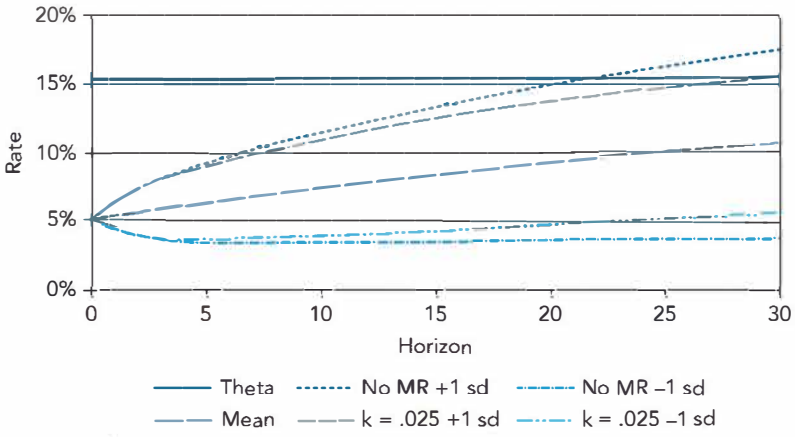


Figure 13.5 **Mean reversion and the terminal distribution of short rates.**

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

$$10.2180\% \times e^{-.025\times 10} = 7.9578\% \qquad (13.22)$$

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

$$r_0 e^{-kT} + \theta(1 - e^{-kT}) \qquad (13.23)$$

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:

$$(15.339\% - 5.121\%)e^{-.025\tau} = \frac{1}{2}(15.339\% - 5.121\%) \qquad (13.24)$$

Solving,

$$e^{-.025\tau} = \frac{1}{2}$$

$$\tau = \frac{\ln(2)}{.025}$$

$$\tau = 27.73 \qquad (13.25)$$

where ln (·) 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 13.5 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 13.5 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

$$\sqrt{\frac{\sigma^2}{2k}(1 - e^{-2kT})} \qquad (13.26)$$

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

$$\sqrt{\frac{.0126^2}{2 \times .025}(1 - e^{-2\times .025\times 10})} \qquad (13.27)$$

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 13.6 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 13.6 qualitatively resembles Figure 13.4. 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 13.6 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.

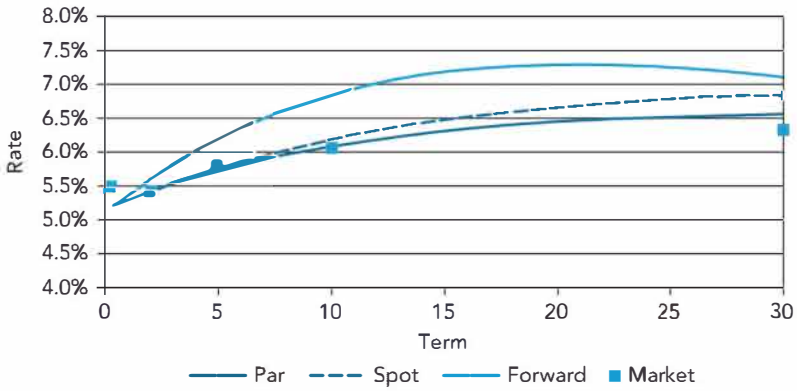


Figure 13.6 Rate curves from the Vasicek model and selected market swap rates, February 16, 2001.

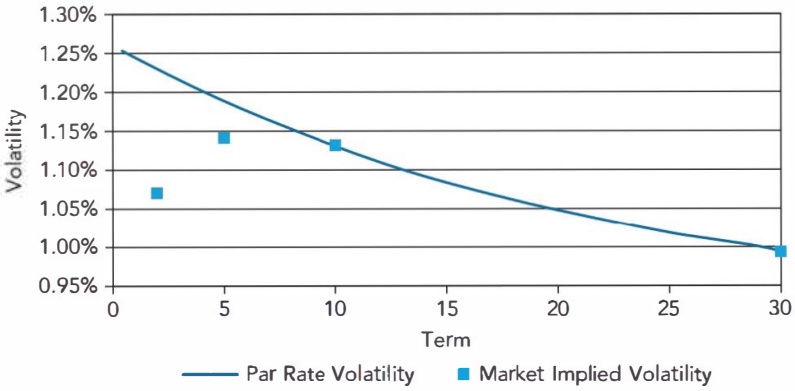


Figure 13.7 **Par rate volatility from the Vasicek model and selected implied volatilities, February 16, 2001.**

Table 13.2 **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 13.7 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 13.7 certainly shows an improvement relative to the flat term structure of volatility shown in Figure 13.3, 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 13.2 reports the convexity effect at several terms. Recall that the convexity effects listed in Table 13.1 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 13.2 than in Table 13.1. 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 13.8 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

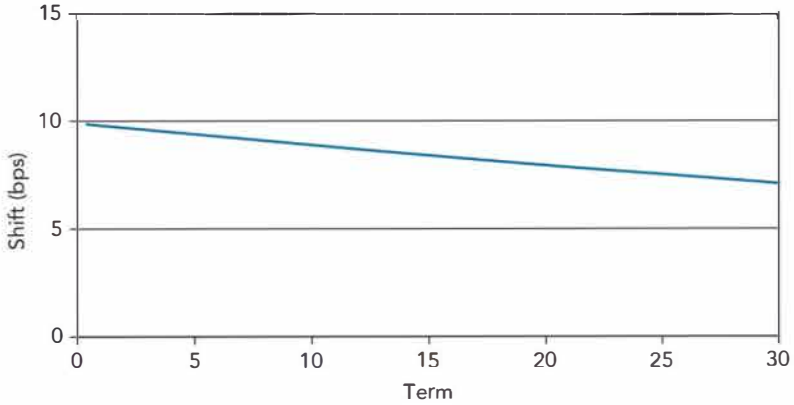


Figure 13.8 **Sensitivity of spot rates in the Vasicek model to a 10-basis-point change in the factor.**

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.

14

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.

14.1 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:

$$dr = \lambda(t)dt + \sigma(t)dw \qquad (14.1)$$

Unlike the models presented in Chapter 13, the volatility of the short rate in Equation (14.1) 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 (14.1) that will be called Model 3:

$$dr = \lambda(t)dt + \sigma e^{-\alpha t}dw \qquad (14.2)$$

In (14.2), 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 14.1 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 (14.2), but very many shapes are possible with the more general volatility specification in (14.1).

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.

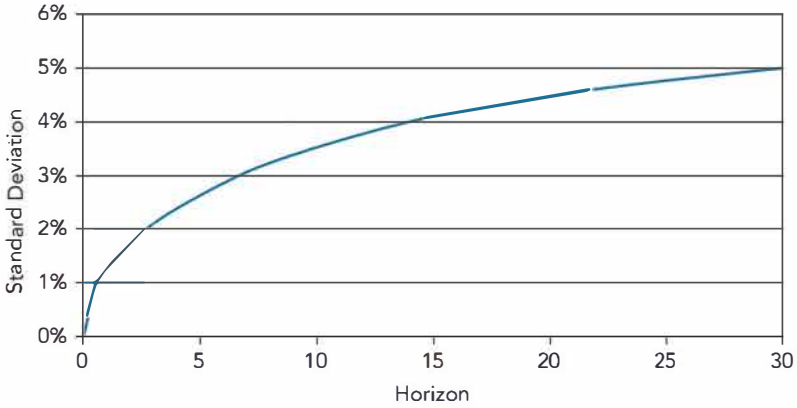


Figure 14.1 **Standard deviation of terminal distributions of short rates in Model 3.**

The behavior of standard deviation as a function of horizon in Figure 14.1 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.

14.2 THE COX-INGERSOLL-ROSS AND LOGNORMAL MODELS: VOLATILITY AS A FUNCTION OF THE SHORT RATE

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

$$dr = k(\theta - r)dt + \sigma\sqrt{r}dw \tag{14.3}$$

Since the first term on the right-hand side of (14.3) 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,

$$dr = k(\theta - r)dt + \sigma r dw \tag{14.4}$$

and the simplest *lognormal model*, to be called Model 4, a variation of which will be discussed in the next section:

$$dr = ardt + \sigma r dw \tag{14.5}$$

In these two specifications, yield volatility is constant but basis-point volatility equals σr and increases with the level of the rate.

Figure 14.2 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,

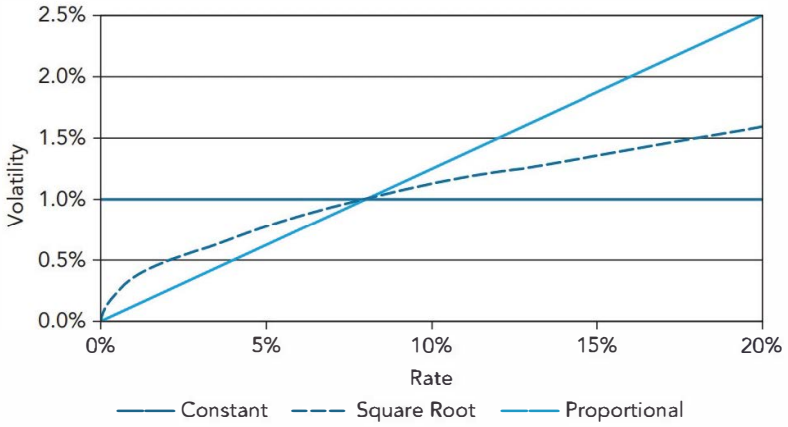


Figure 14.2 Three volatility specifications.

$$\sigma^{bp} = .01 \tag{14.6}$$

$$\sigma^{CIR} \times \sqrt{8\%} = 1\% \Rightarrow \sigma^{CIR} = .0354 \tag{14.7}$$

$$\sigma^{y} \times 8\% = 1\% \Rightarrow \sigma^{y} = 12.5\% \tag{14.8}$$

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 14.2, 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 14.3 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

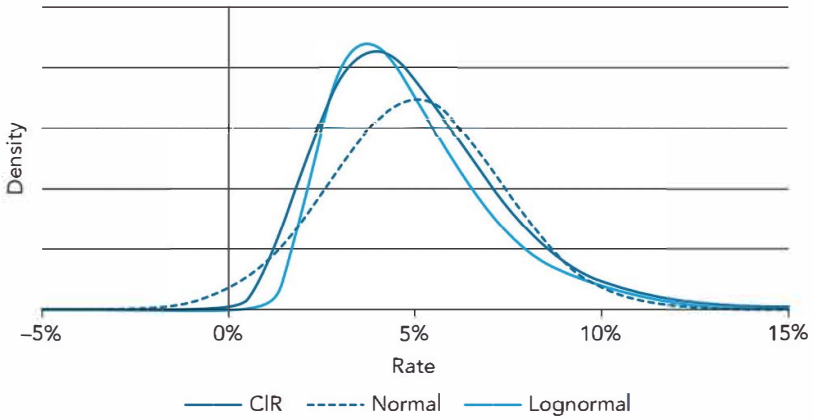


Figure 14.3 **Terminal distributions of the short rate after ten years in CIR, normal, and lognormal models.**

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.

14.3 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:

$$dr = \tilde{a}(t)rdt + \sigma rdw \qquad (14.9)$$

By Ito's Lemma, which is beyond the mathematical scope of this book,

$$d[\ln(r)] = \frac{dr}{r} - \frac{1}{2}\sigma^2 dt \qquad (14.10)$$

Substituting (14.9) into (14.10),

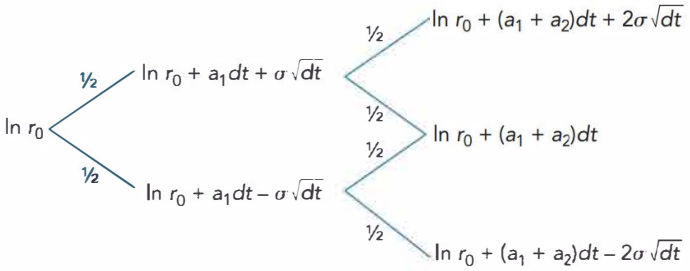
$$d[\ln(r)] = \left[\tilde{a}(t) - \frac{1}{2}\sigma^2\right]dt + \sigma dw \qquad (14.11)$$

Redefining the notation of the time-dependent drift so that $a(t) = \tilde{a}(t) - \frac{1}{2}\sigma^2$, Equation (14.11) becomes

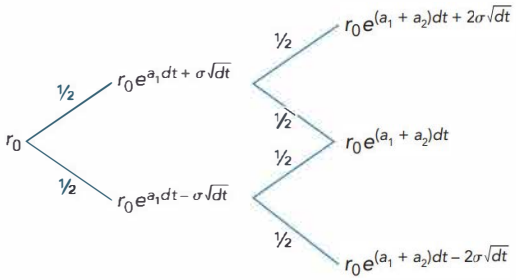
$$d[\ln(r)] = a(t)dt + \sigma dw \qquad (14.12)$$

Equation (14.12) 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, (14.12) implies that the short rate has a lognormal distribution.

Equation (14.12) 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

$$r_0 e^{\sigma} \approx r_0(1 + \sigma) \qquad (14.13)$$

Volatility is clearly a percentage of the rate in equation (14.13). 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.

14.4 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

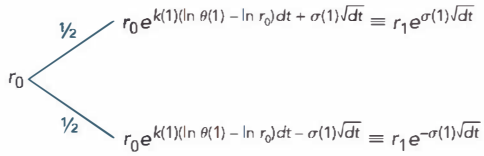
$$dr = k(t)(ln\tilde{\theta}(t) - \ln r)dt + \sigma(t)rdw \tag{14.14}$$

or, equivalently,[2] as

$$d[\ln r] = k(t)(\ln \theta(t) - \ln r)dt + \sigma(t)dw \tag{14.15}$$

In words, Equation (14.15) 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

$$\ln r_1 + \sigma(1)\sqrt{dt} \tag{14.16}$$

and

$$\ln r_1 - \sigma(1)\sqrt{dt} \tag{14.17}$$

[2] This derivation is similar to that of moving from Equation (14.9) to Equation (14.12).

respectively. It follows that the step down from the up-state requires a rate of

$$r_1 e^{\sigma(1)\sqrt{dt}} e^{k(2)[\ln\theta(2) - \{\ln r_1 + \sigma(1)\sqrt{dt}\}]dt - \sigma(2)\sqrt{dt}} \tag{14.18}$$

while the step up from the down-state requires a rate of

$$r_1 e^{-\sigma(1)\sqrt{dt}} e^{k(2)[\ln\theta(2) - \{\ln r_1 - \sigma(1)\sqrt{dt}\}]dt + \sigma(2)\sqrt{dt}} \tag{14.19}$$

A little algebra shows that the tree recombines only if

$$k(2) = \frac{\sigma(1) - \sigma(2)}{\sigma(1)dt} \tag{14.20}$$

Imposing the restriction (14.20) 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 (14.20). To do so it allows the length of the time step, dt, to change over time.

Rewriting Equations (14.18) and (14.19) with the time steps labeled dt_1 and dt_2 gives the following values for the up-down and down-up rates:

$$r_1 e^{\sigma(1)\sqrt{dt_1}} e^{k(2)[\ln\theta(2) - \{\ln r_1 + \sigma(1)\sqrt{dt_1}\}]dt_2 - \sigma(2)\sqrt{dt_2}} \tag{14.21}$$

$$r_1 e^{-\sigma(1)\sqrt{dt_1}} e^{k(2)[\ln\theta(2) - \{\ln r_1 - \sigma(1)\sqrt{dt_1}\}]dt_2 + \sigma(2)\sqrt{dt_2}} \tag{14.22}$$

A little algebra now shows that the tree recombines if

$$k(2) = \frac{1}{dt_2}\left[1 - \frac{\sigma(2)\sqrt{dt_2}}{\sigma(1)\sqrt{dt_1}}\right] \tag{14.23}$$

The length of the first time step can be set arbitrarily. The length of the second time step is set to satisfy (14.23), allowing the user freedom in choosing the mean reversion and volatility functions independently.

14.5 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

$$\hat{r}(T) = r_0 - \frac{\sigma^2 T^2}{6} \tag{14.24}$$

Model 2

$$\hat{r}(T) = r_0 + \frac{\lambda T}{2} - \frac{\sigma^2 T^2}{6} \tag{14.25}$$

Vasicek

$$\hat{r}(T) = \theta + \frac{1 - e^{-kT}}{kT}(r_0 - \theta) - \frac{\sigma^2}{2k^2}\left(1 + \frac{1 - e^{-2kT}}{2kT} - 2\frac{1 - e^{-kT}}{kT}\right) \tag{14.26}$$

Model 3 with $\lambda(t) = \lambda$

$$\hat{r}(T) = r_0 + \frac{\lambda T}{2} - \sigma^2 \frac{2\alpha^2 T^2 - 2\alpha T + 1 - e^{-2\alpha T}}{8\alpha^3 T} \tag{14.27}$$

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,

$$P(T) = A(T)e^{-B(T)r_0} \tag{14.28}$$

where

$$A(T) = \left[\frac{2he^{(k+h)T/2}}{2h + (k + h)(e^{ht} - 1)}\right]^{2k\theta/\sigma^2} \tag{14.29}$$

$$B(T) = \frac{2(e^{ht} - 1)}{2h + (k + h)(e^{hT} - 1)} \tag{14.30}$$

$$h = \sqrt{k^2 + 2\sigma^2} \tag{14.31}$$

Volatility Smiles 15

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, *Tenth 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 lognormal? 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*. This chapter describes the volatility smiles 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 option traders use volatility surfaces as pricing tools.

15.1 WHY THE VOLATILITY SMILE IS THE SAME FOR 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 means that the volatility smile for European calls with a certain maturity is the same as that for European puts with the same maturity. This is a particularly convenient result. It shows that when talking about a volatility smile we do not have to worry about whether the options are calls or puts.

As explained in earlier chapters, 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

$$p + S_0 e^{-qT} = c + Ke^{-rT} \tag{15.1}$$

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

$$p_{BS} - p_{mkt} = c_{BS} - c_{mkt} \tag{15.2}$$

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 (15.2), 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 15.1

The value of a foreign currency is USD 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 USD 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 (15.1) must apply with q equal to the foreign risk-free rate. The price p of a European put option with a strike price of USD 0.59 and maturity of 1 year therefore satisfies

$$p + 0.60e^{-0.10\times1} = 0.0236 + 0.59e^{-0.05\times1}$$

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.

15.2 FOREIGN CURRENCY OPTIONS

The volatility smile used by traders to price foreign currency options has the general form shown in Figure 15.1. 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 15.1 corresponds to the implied distribution shown by the solid line in Figure 15.2. A lognormal distribution with the same mean and standard deviation as the implied distribution is shown by the dashed line in Figure 15.2. It can be seen that the implied distribution has heavier tails than the lognormal distribution.[1]

To see that Figures 15.1 and 15.2 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 15.2 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 15.1 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 15.2 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 15.1.

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,

[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.

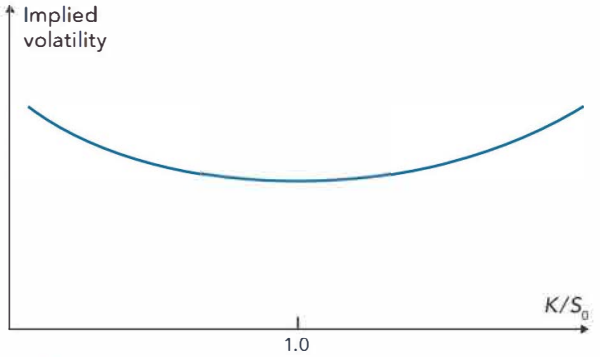


Figure 15.1 Volatility smile for foreign currency options (K = strike price, S_0 = current exchange rate).

Table 15.1 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, 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

Table 15.1 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

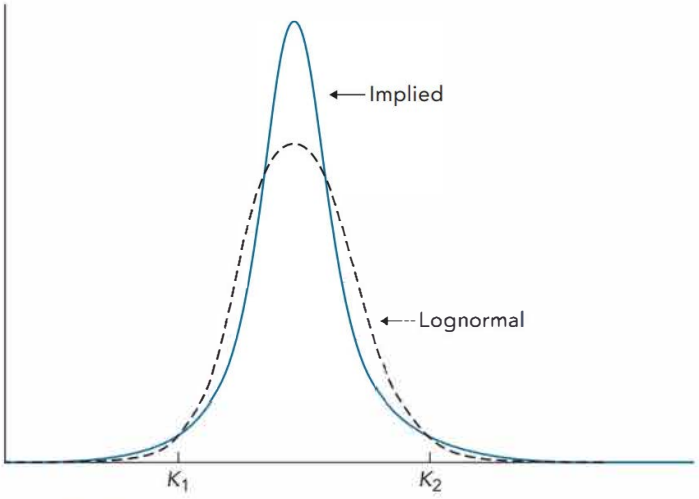


Figure 15.2 **Implied and lognormal distribution for foreign currency options.**

happening. The table therefore provides evidence to support the existence of heavy tails (Figure 15.2) and the volatility smile used by traders (Figure 15.1). Business Snapshot 15.1 shows how you could have made money if you had done the analysis in Table 15.1 ahead of the rest of the market.

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]

[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.

BUSINESS SNAPSHOT 15.1 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 15.1 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.

The result of all this is that the volatility smile becomes less pronounced as option maturity increases.

15.3 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 on stock indices) has had the general form shown in Figure 15.3. 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 15.4. A lognormal distribution with the same mean and standard

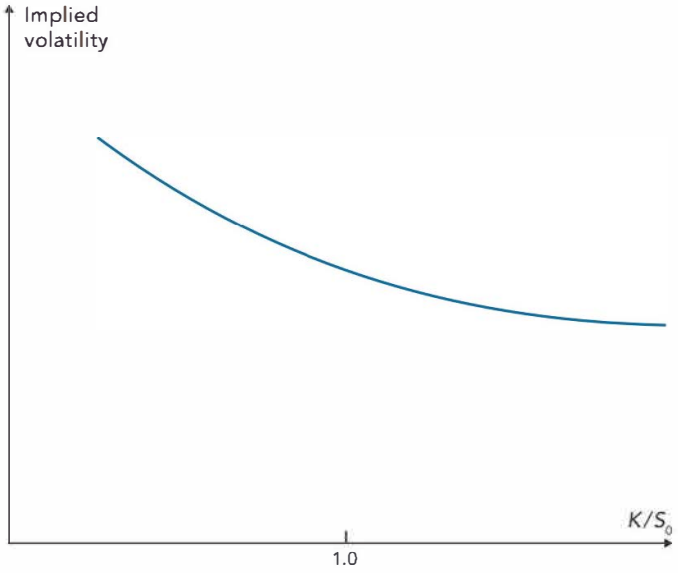


Figure 15.3 Volatility smile for equities (K = strike price, S_0 = current equity price).

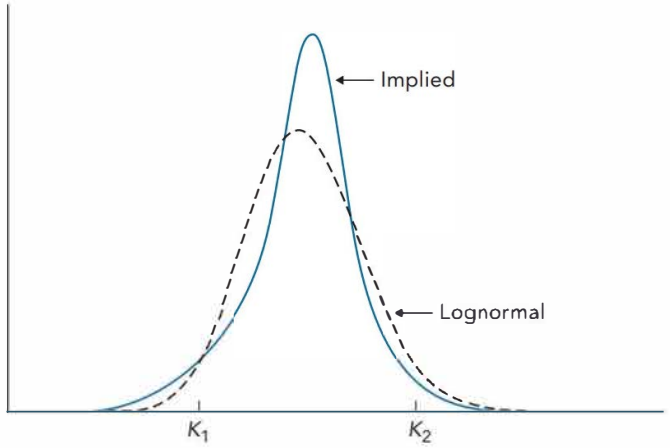


Figure 15.4 Implied distribution and lognormal distribution for equity options.

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 15.3 and 15.4 are consistent with each other, we proceed as for Figures 15.1 and 15.2 and consider options that are deep out of the money. From Figure 15.4, 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 15.3 for the option. Consider next a deep-out-of-the-money put option with a strike price of K_1. This option pays off only if the stock price proves to be below K_1 (K_1/S_0 well below 1.0). Figure 15.4 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 15.3.

The Reason for the Smile in Equity Options

There is a negative correlation between equity prices and volatility. 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 15.2).

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 15.4.

BUSINESS SNAPSHOT 15.2 Crashophobia

It is interesting that the pattern in Figure 15.3 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.

15.4 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 the right (left).[3] 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 in Figures 15.1 and 15.3.

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."

15.5 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.

[3] 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.

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 15.2.

One dimension of Table 15.2 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.)

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 15.2 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 15.2, 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

$$\frac{1}{\sqrt{T}} \ln\left(\frac{K}{F_0}\right)$$

Table 15.2 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

rather than as the relationship between the implied volatility and K. The smile is then usually much less dependent on the time to maturity.

15.6 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. The volatility smile has the shape shown in Figure 15.3. Two phenomena can be identified:

1. As the equity price increases (decreases), K/S_0 decreases (increases) and the volatility increases (decreases). In other words, the option moves up the curve in Figure 15.3 when the equity price increases and down the curve when the equity price decreases.
2. There is a negative correlation between equity prices and their volatilities. When the equity price increases, the whole curve in Figure 15.3 tends to move down; when the equity price decreases, the whole curve in Figure 15.3 tends to move up.

It turns out that the second effect dominates the first, so that implied volatilities tend to move down (up) when the equity price moves up (down). 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.[4]

[4] For a further discussion of this, see, for example, J. C. Hull and A. White, "Optimal Delta Hedging of Options," Working paper, University of Toronto, 2016.

15.7 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.

15.8 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 USD 50 and an important news announcement due in a few days is expected either to increase the stock price by USD 8 or to reduce it by USD 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 15.5. 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 where there are only two possible future stock prices. This is what we will now do.

Suppose that the stock price is currently USD 50 and that it is known that in 1 month it will be either USD 42 or USD 58. Suppose further that the risk-free rate is 12% per annum. The situation is illustrated in Figure 15.6. Options can be valued using the binomial model. 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 15.3. The first column shows

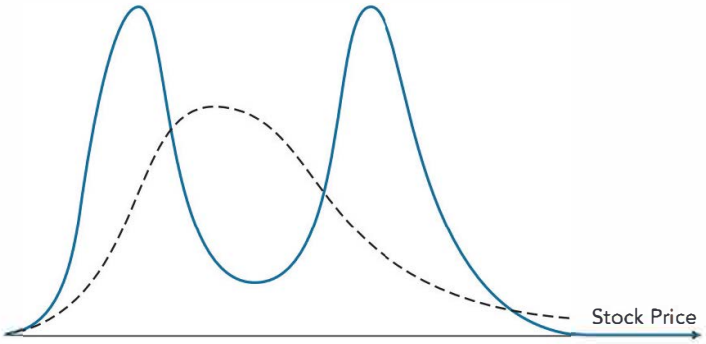


Figure 15.5 Effect of a single large jump. The solid line is the true distribution; the dashed line is the lognormal distribution.

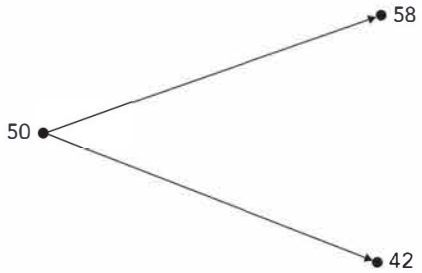


Figure 15.6 Change in stock price in 1 month.

Table 15.3 Implied Volatilities in Situation Where it is Known that the Stock Price will Move from USD 50 to Either USD 42 or USD 58

Strike Price (USD)	Call Price (USD)	Put Price (USD)	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

alternative strike prices; the second column shows prices of 1-month European call options; the third column shows the prices of one-month European put option prices; the fourth column shows implied volatilities. (The implied volatility of a

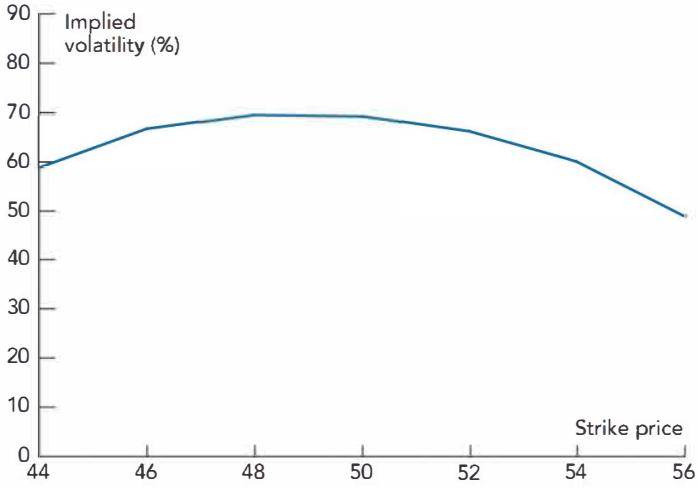


Figure 15.7 Volatility smile for situation in Table 15.3.

European put option is the same as that of a European call option when they have the same strike price and maturity.) Figure 15.7 displays the volatility smile from Table 15.3. 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.

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

$$g(K) = e^{rT}\frac{\partial^2 c}{\partial K^2} \qquad (15A.1)$$

This result, which is from Breeden and Litzenberger (1978), allows risk-neutral probability distributions to be estimated from volatility smiles.[5] 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 (15A.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 $K + \delta$, 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 15A.1. Since δ is small, we can assume that $g(S_T) = g(K)$ in the whole of the range $K - \delta < S_T < K + \delta$, where the payoff is nonzero. The area under the "spike" in Figure 15A.1 is

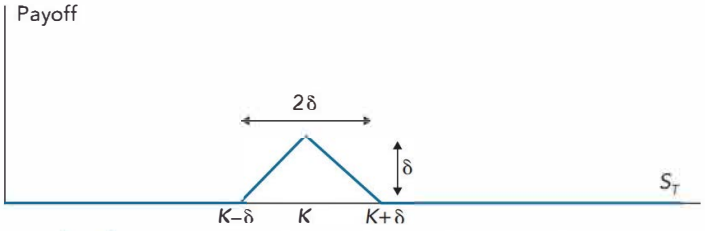


Figure 15A.1 Payoff from butterfly spread.

$0.5 \times 2\delta \times \delta = \delta^2$. 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

$$g(K) = e^{rT}\frac{c_1 + c_3 - 2c_2}{\delta^2} \qquad (15A.2)$$

Example 15A.1

Suppose that the price of a non-dividend-paying stock is USD 10, the risk-free interest rate is 3%, and the implied volatilities of 3-month European options with strike prices of USD 6, USD 7, USD 8, USD 9, USD 10, USD 11, USD 12, USD 13, USD 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:

$$g(S_T) = g_1 \quad \text{for} \quad 6 \le S_T < 7$$
$$g(S_T) = g_2 \quad \text{for} \quad 7 \le S_T < 8$$
$$g(S_T) = g_3 \quad \text{for} \quad 8 \le S_T < 9$$
$$g(S_T) = g_4 \quad \text{for} \quad 9 \le S_T < 10$$
$$g(S_T) = g_5 \quad \text{for} \quad 10 \le S_T < 11$$
$$g(S_T) = g_6 \quad \text{for} \quad 11 \le S_T < 12$$
$$g(S_T) = g_7 \quad \text{for} \quad 12 \le S_T < 13$$
$$g(S_T) = g_8 \quad \text{for} \quad 13 \le S_T < 14$$

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 USD 6.5 as 29.5%. This means that options with strike prices of USD 6, USD 6.5, and USD 7 have implied volatilities of 30%, 29.5%, and 29%, respectively. From DerivaGem their prices are USD 4.045, USD 3.549, and USD 3.055, respectively. Using equation (15A.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

$$g_2 = 0.0444, \quad g_3 = 0.1545, \quad g_4 = 0.2781$$
$$g_5 = 0.2813, \quad g_6 = 0.1659, \quad g_7 = 0.0573, \quad g_8 = 0.0113$$

[5] See D. T. Breeden and R. H. Litzenberger, "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51 (1978), 621–51.

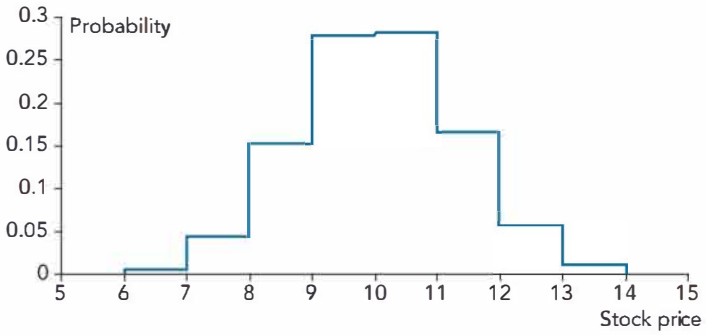


Figure 15A.2 **Implied probability distribution for Example 15A.1.**

Figure 15A.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 15A.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 USD 6 and USD 7 is 0.0031 (compared with 0.0057 in Figure 15A.2) and the probability of a stock price between USD 13 and USD 14 is 0.0167 (compared with 0.0113 in Figure 15A.2).

16

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 18 of Risk Management and Financial Institutions, *Fifth Edition, by John C. Hull.*

Note: This chapter references the December 2014 proposal for the Fundamental Review of the Trading Book. The final version was published under the title "Minimum capital requirements for market risk" (Basel Committee on Banking Supervision Publication 352, January 2016). It is freely available on the GARP website.

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 2016.[3] This requires banks to implement the new rules in 2019, but in December 2017, the implemented year was revised to 2022.

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.

16.1 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. 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] 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 16.1.

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 "Use of Standardized Approaches and SA-CCR," 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 2027 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.

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 2016.

[4] 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 $\mu + \sigma^2 f(\text{VaR})/(1 - X)$ where f is the probability density function for the loss.

Table 16.1 Allocation of Market Variables 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

16.2 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 dual 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:

$$\text{Risk Charge} = \sum_i \sum_j \rho_{ij}\delta_i\delta_j W_i W_j \qquad (16.1)$$

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 weighted 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 USD 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 USD 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 16.1 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 16.1 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 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.

[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.

Vega risk is handled similarly to delta risk. A vega risk charge is calculated for each risk class using Equation 16.1. 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[6]. 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.

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 typically defined by 10 points. These are the zero-coupon interest rates tor 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 16.1. 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.

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 16.1. In Figure 16.1a, 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 16.1b, 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 16.1a could correspond to a short position in an option; Figure 16.1b could correspond to a long position in an option.)

When there are several risk factors, each is handled similarly to Figure 16.1. 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 is often 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.

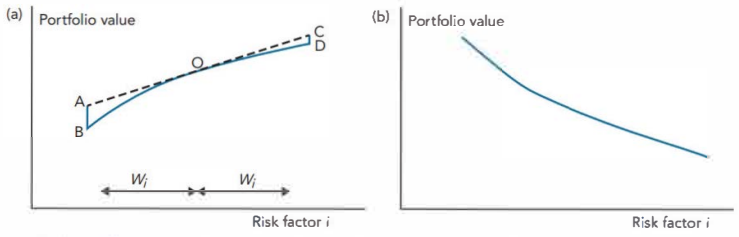


Figure 16.1 Calculation of curvature risk charge for a risk factor. In Figure 16.1a, the curvature risk charge is AB; in Figure 16.1b, it is zero.

[6] Banks can choose whether it is percentage or actual changes in volatility that are considered.

Default Risk Charge

Risks associated with counterparty credit spread changes are handled separately from risks associated with counterparty 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 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.

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%.

A Simplified Approach

In this section, we have described the standardized approach that the Basel Committee requires all large banks to use. It is worth noting that in June 2017 the Basel Committee published a consultative document outlining a simplified standardized approach that it proposes for smaller banks.[7] 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.

16.3 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 16.1. 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 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

$$\sqrt{ES_1^2 + \sum_{j=2}^{5}\left(ES_j\sqrt{\frac{LH_j - LH_{j-1}}{10}}\right)^2} \qquad (16.2)$$

[7] See Basel Committee on Banking Supervision, "Simplified Alternative to the Standardized Approach to Market Risk Capital Requirements," June 2017.

where LH_j is the liquidity horizon for category j. To understand Equation 16.2, 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 16.2. 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 portfolio. (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:

$$\text{WES} = \lambda \times \text{EST} + (1 - \lambda) \times \sum_j \text{ESP}_j$$

where EST is the expected shortfall calculated for the total portfolio and ESP_j is 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.

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.

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.[8] Regulators expect the first measure to be between and −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.

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 other market risks. Table 16.1 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.

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. 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.

16.4 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

[8] 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*.

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 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 2016.

INDEX

C

D

E

F

G

H

I

J

K

L

M

N

T

U

V

W

Y