CHAPTER 1

What This Book Is About

> It is not my intention to detain the reader by expatiating on the variety, or the importance of the subject, which I have undertaken to treat; since the merit of the choice would serve to render the weakness of the execution still more apparent, and still less excusable. But […] it will perhaps be expected that I should explain, in a few words, the nature and limits of my general plan.
>
> Edward Gibbon, *The Decline and Fall of the Roman Empire*[1]

1.1 MY GOAL IN WRITING THIS BOOK

In this book I intend to look at yield-curve modelling from a 'structural' perspective.[2] I use the adjective *structural* in a very specific sense, to refer to those models which are created with the goal of *explaining* (as opposed to *describing*) the yield curve. What does 'explaining' mean? In the context of this book, I mean accounting for the observed yields by combining the expectations investors form about future rates (and, more generally, the economy) and the compensation they require to bear the risk inherent with holding default-free bonds. (As we shall see later, there is a third 'building block', ie, convexity.)

This provides one level of explanation, but one could go deeper. So, for instance, the degree of compensation investors require in order to bear 'interest-rate risk' could be derived ('explained') in more fundamental terms from the strategy undertaken by a rational, risk-averse investor who is faced with a set of investment opportunities and wants to maximize her utility from consumption

[1] From the Prologue.

[2] A note on terminology. In the term-structure literature the adjective 'structural' is often applied to those models that are based on a specification of the economy – a specification that may go all the way down to preferences, utility maximization and equilibrium. I use the term 'equilibrium models' to refer to these descriptions. We shall only dip our toes in these topics in Chapter 15. For those readers who already understand the meaning of the expression, structural models in this book are those that straddle the $\mathbb{P}$- (real-world) and $\mathbb{Q}$- (risk-neutral) measures. If this does not make much sense at the moment, all will be revealed.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

in a multiperiod economy. I will sketch with a broad brush the main lines of this fundamental derivation, but will not pursue this line of argument in great detail. The compensation exacted by investors for bearing market risk (the 'market price of risk') will instead be empirically related (say, via regressions) either to combinations of past and present bond prices and yields, or to past history and present values of macroeconomic variables.

Another way to look at what I try to do in this book is to say that I *describe* the market price of risk in order to *explain* the yield curve. If one took a more 'fundamental' approach, one could try to *explain* the market price of risk as well, but would still have to *describe* something more basic, say, the utility function. Sooner or later, all scientific treatments hit against this hard descriptive core; even theoretical physics is not immune to the curse, or blessing, of having to describe. See, in this respect, the Section 7 of this chapter.

In keeping with the quote that opens this chapter, I will not dwell on why yield curve modelling is important – after all, if the reader were not convinced of this, she probably would not be reading these words. Still, one may well ask, 'Why write a book on *structural* yield-curve modelling?' The answer is that since the mid-2000s there have been exciting developments in the theoretical and empirical understanding of the yield curve dynamics and of risk premia. The 'old' picture with which many of us grew up is now recognized to be in important respects qualitatively wrong. To go from the old to the new class of models requires a rather substantial piece of surgery, not a face-lift, but it is well worth the effort.

Unfortunately, the existing literature on these exciting new topics is somewhat specialized and uses elegant but, to the uninitiated, rather opaque and forbidding-sounding concepts (such as the state-price deflator or the stochastic discount factor). Gone is the simplicity with which even a relative newcomer could pick up Vasicek's paper and, with a good afternoon's work, understand what it was about.

It is therefore my intention to 'translate' and introduce these exciting new developments using the simplest mathematical tools that allow me to handle correctly (but not rigorously) the material at hand. In doing so, I will always trade off a pound of mathematical rigour for an ounce of intuition.

I will also try to explain the vocabulary of the 'new language', and rederive in the simplest possible way the old (and probably familiar) no-arbitrage results using the modern tools. This will both deepen the reader's understanding and enable her to read the current literature.

In addition to expectations and risk premia, there is a third important determinant to the shape of the yield curve, namely 'convexity'. In Part V explain in detail what convexity is, and why it is, in some sense, unique. (In a nutshell, to extract the risk premium you just have to be patient and will be 'fairly' rewarded for your patience; to earn convexity, you have to work very, very hard.) For the moment, the important point is that in the treatment I present in this book these three building blocks (expectations, risk premia, and convexity), together with

the principle of no-arbitrage, explain all that there is to know about the yield curve.[3]

1.2 WHAT MY ACCOUNT LEAVES OUT

Is it true that, once we account for expectations, risk premia and convexity, there is really nothing else to the dynamics of credit-risk-free yield curves, at least at the level of description that we have chosen? Of course it isn't. To understand what is left out some historical perspective may help.

The current modelling approach places the Expectation Hypothesis at its centre. This does not mean that 'only expectations matter', but that the only (or the main) deviations from expectations come from risk premia (and the neglected relation, convexity). As Fontaine and Garcia (2015) state '[w]hat distinguishes modern literature is the emphasis on interest rate risk as the leading (or sole) determinant of the risk premium.'[4] As a result 'sources of risk premium other than interest rate risk found a refuge in undergraduate textbooks while the academic agenda leapt forward, developing an array of sophisticated yet tractable no-arbitrage models.'[5]

So what is left behind by the expectations–risk premia–convexity triad?

To begin with, I devote little attention to liquidity, which can become very important, especially in periods of market distress.[6] However, in most market conditions the securities I deal with in this book – US Treasury bonds, German Bunds, UK gilts – are among the most liquid instruments available to investors. Liquidity, one can therefore argue, should be relatively unimportant in a reasonable hierarchy of important factors.[7] If the reader is interested in liquidity-specific issues (such as the pricing of on-the-run versus off-the-run Treasury bonds), the approach of Fontaine and Garcia (2008) discussed in some detail

[3] As noted earlier, I will mention briefly the links between my building blocks and more fundamental macreconomic and monetary-economics concepts (see Chapters 3 and 15), but I will do so simply to give the reader a qualitative understanding of the form a more fundamental approach to yield curve modelling would take.

[4] p. 463. [5] ibid., pp. 463–464.

[6] In Chapter 18 I present a general pricing methodology that will allow the reader to build her own affine model, DIY-style. Using this toolkit, there is nothing to stop the reader from introducing a factor called 'liquidity', equip it with the necessary parameter paraphernalia (reversion speed, reversion level, volatility, etc) and plug it in the multipurpose affine framework that I develop in Chapter 18. By construction, her 'fits' will be at least as good, and probably better, than before she introduced the 'liquidity' factor. However, it is not easy to find a 'principled' way to assign the correct explanatory contribution to this factor: are we really modelling liquidity, or have we just over-parametrized our model?

[7] Of the models that we explore in Part VII, two deal with liquidity. One is the D'Amico, Kim and Wei (2010) approach, which deals with nominal and *real* rates, explicitly models liquidity – and the authors make the point that the inclusion of this factor is important in order to have a correct estimation of the model parameters and a convincing description of inflation expectations. Dollar-denominated inflation-linked bonds were, especially in the early years after their introduction, far less liquid that their nominal Treasury counterparts, and a strong case can therefore be made for an explicit modelling of liquidity.

in Chapter 32, is very useful.[8] When it comes to government bonds, however, it must be kept in mind that a bond-specific maturity factor presents a serious challenge for traditional (and frictionless) no-arbitrage models, which are built on the assumption that all bonds are created exactly equal, once their return and risk characteristics are properly taken into account.[9]

The other main possible missing ingredient from the description presented in this book is market segmentation – the idea that classes of investors, such as pension funds, might have preferred 'habitats' (maturity ranges) where they 'like to' invest. According to proponents of segmentation, by so doing, these investors create an imbalance of supply and demand that arbitrageurs either do not manage to eliminate, or do eliminate, but by taking risk, for which compensation – and hence risk premium – is exacted.[10] According to researchers such as Vayanos and Vila (2009), the compensation for the risky activities of pseudo-arbitrageurs then leaves a detectable signature in the shape of a risk-premiuma contribution to various yields. Readers interested in the topic of segmentation are referred to Vayanos and Vila (2009) for a theoretical treatment along these lines, and Chen et al. (2014) for an empirical discussion of the maturity preference exhibited by insurance firms.

These topics, and other sources of imperfections such as the zero bound of rates, are well treated in Fontaine and Garcia (2015) – the title of their chapter ('Recent Advances in Old Fixed Income Topics: Liquidity, Learning, and the Lower Bound') gives a good flavour of what the reader can find in their work. As mentioned, we look at liquidity in Chapter 32, and we deal with the zero bound in Chapter 19. We do not deal with market segmentation, and only cursorily with learning-related issues; see, however, the opening sections of Chapter 28.

1.3 AFFINE MODELS

Let's therefore assume that we are happy with our identification of the three building blocks (expectations, risk premia and convexity) and of the glue

I also deal with liquidity in Chapter 32, which is devoted to the Diebold and Rudebusch approach. The treatment is based on the insight by Fontaine and Garcia (2008), and can be applied to other liquidity-unaware models as well.

[8] 'On-the-run' bonds are freshly-minted, newly-issued Treasury bonds. They enjoy special liquidity, and therefore yield several basis points less (are more expensive) than earlier-issued ('off-the-run') Treasury bonds of similar maturity. This on-the-run/off-the-run spread can become significantly larger in periods of market distress, when liquidity becomes very sought after.

[9] As Fontaine and Garcia (2008) write, 'a structural specification of the liquidity premium raises important challenges. The on-the, run-premium is a real arbitrage opportunity unless we explicitly consider the cost of shorting the more expensive bond, or, alternatively, the benefits accruing to a bondholder from a lower repo rate. These features are absent from the current crop of term-structure model' (pp. 9–10).

[10] As Fontaine and Garcia (2015) point out, liquidity and segmentation need not be looked at as totally different sources of friction or inefficiency because '[t]he clientele demand for new and old bonds is similar in spirit to the view that investors have "preferred habitats"' and '[t]he clientele demand may be scattered across bond maturities, but it can also be scattered across the illiquidity spectrum' (p. 472).

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

(noarbitrage) that holds them together. What we need next is a way to combine these ingredients in a coherent and logically consistent manner. This is what a model does, and this is why a large part of this book is devoted to discussing models of the yield curve. *Which* models, though?

Because of their unsurpassed intuitional appeal and their analytical tractability, I deal mainly with a popular class of structural models – the affine class.[11] In order to give a transparent understanding of how these models weave together these three building blocks to determine the shape of the yield curve, I will start my discussion from the simplest incarnation of affine models – the Vasicek (1977) model.[12]

The Vasicek model is unparalleled for the intuitive understanding it affords, and it is for this reason that I introduce it, perhaps unwisely, very early in the book – even, that is, before dealing with the theoretical underpinnings of term-structure modelling. Quite simply, I want the reader to have a vivid, if, at this point, probably imprecise, picture of what we will be talking about more precisely and more abstractly in the later parts of the book, when more complex, and more opaque, models come to the fore.

In general, I strongly encourage the reader who feels her intuition beginning to fail her when looking at the more complex models to adopt ruthlessly the strategy of *reductio ad Vasicek*, ie, to ask herself, 'What is the equivalent of this concept/formalism/result in the Vasicek model?' She is encouraged to do so, not because the Vasicek model is perfect, but because it lays bare with great clarity the mechanics and intuition behind more complex affine models.

For all the virtues of the Vasicek model, recent empirical evidence suggests that the explanation of risk premia Vasicek-family models afford is *qualitatively* wrong. Since the risk premium constitutes the explanatory bridge between expectations and observed prices, and since the Vasicek approach is the progenitor of all the more recent affine models, this does not seem to bode well for affine structural approaches to term-structure modelling.

Luckily, the same empirical evidence also suggests how the first-generation, Vasicek-like, affine models can be modified and enriched. I therefore present in Part VI of this book what we now know about term premia, and in Part VII how these empirical findings can be incorporated in the new-generation affine models.

[11] See, for instance, Dai and Singleton (2000) for a systematic calssification of affine models, and Duffee (2002) for a discussion of *essentially* affine models – loosely speaking, models which remain affine both in the real-world and in the pricing measures. Good reviews of affine models can be found in Bloder (2001), who also deals with Kalman filter estimation methods, and Piazzesi (2010). Extensions to stochastic affine-volatility models are found in Longstaff and Schwartz (1992) and Balduzzi et al. (1996).

[12] I must make very clear from the start that I will deal in this book with *Gaussian* affine models, which are far simpler than the square-root models of the Cox–Ingersoll–Ross (1985a, b) family. Admittedly, Gaussian affine models do allow for negative rates, but recent experience suggests that this should be considered more of a virtue than a blemish. (At the time of this writing, Germany just issued short-dated government bonds with a negative yield.)

Speaking of affine models means that we require a special type of relationship between yields and the state variables. But how should we choose these variables? As we shall discuss towards the end of the book, from a very abstract point of view, and as long as some quantities are exactly recovered by the different models, the choice of variables makes very little difference. In practice, however, this choice informs the statistical estimation techniques used in the calibration, the degree of 'structure' on the dynamics of the state variables (via the condition of no-arbitrage), the parsimony of the model and the user's ability to understand and interpret the model. Section 1.5 of this introductory chapter makes these statements more precise. First, however, we want to look a bit more carefully at the various types of yield curve models, so that the reader can clearly see what we are going to deal with and what we will not touch upon. Probably, the reader should not throw away her book receipt before reading the next section.

1.4 A SIMPLE TAXONOMY

There are many different types of term-structure models. They are different in part because they have been created with different purposes in mind and in part because they look at the same problem from different angles. A reasonable taxonomy may look as follows.

1. *Statistical models* aim to *describe* how the yield curve moves. Their main workhorses here are the Vector Auto-Regressive (VAR) models, which are often employed to forecast interest rates and to estimate the risk premium as the difference between the forward and the forecasted rates. This task sounds easy, but, as I discuss later in the book, the quasi-unit-root nature of the level of rates (and many more statistical pitfalls) makes estimations based purely on time-series analysis arduous, and the associated 'error bars' embarrassingly large. See, eg, the discussion in Cochrane and Piazzesi (2008).[13]

   In the attempt to improve on this state of affairs, no-arbitrage structural models, which add *cross-sectional* information to the time-series data, come to the fore. *In this book we shall take a cursory and instrumental look at statistical models, mainly to glean statistical information about one important ingredient of our structural models, ie, the market price of risk.*

   The important thing to stress is that statistical models fit observed market yield curves well and have good predictive power but lack a strong theoretical foundation, because, by themselves, they cannot guarantee absence of arbitrage among the predicted yields. Their strengths and weaknesses are therefore complementary to those of the no-arbitrage models discussed in the text that follows: these are theoretically sound, but sometimes poor at fitting the market yield

[13] See, in particular, the discussion of their Panel 1 on p. 2 of their paper.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

covariance structure and the observed yield curves, and worse at predicting their evolution. See, in this respect, the discussion in Diebold and Rudebusch (2013)[14] and Section 1 in Chapter 32.

One of the underlying themes developed in this book is the attempt to marry the predictive and fitting virtues of statistical models with the theoretical solidity of the no-arbitrage models. Chapters 32, 33 and 34 should be read in this light.

2. *Structural no-arbitrage models* (of which the Vasicek (1977) and Cox–Ingersoll–Ross (1985a, b) are the earliest and best-known textbook examples) make assumptions about how a handful of important driving factors behave; they ensure that the no-arbitrage condition is satisfied; and they derive how the three components that drive the yield curve (expectations, risk premia and convexity) should affect the shape of the yield curve. The no-arbitrage conditions ensure that the derived prices of bonds do not offer free lunches. As I explain in footnote 1, I speak of structural no-arbitrage models when they straddle the physical (real-world, $\mathbb{P}$) and risk-neutral ($\mathbb{Q}$) measures – as opposed to restricted no-arbitrage models that are formulated only in the $\mathbb{Q}$ measure.

   The distinction is important for at least two reasons. First, if we want to understand how bond prices are formed based on expectations and risk aversion, we cannot look at just one measure: market prices are compatible with an infinity of different combinations of expectations and market prices of risk.

   The second reason is subtler. It is well known that if we only look at the risk-neutral ($\mathbb{Q}$) measure three factors (as we shall see, the first three principal components) explain the movements in prices extremely well. However, if we also want to explain excess returns (risk premia) we may have to use more variables (perhaps up to five, according to Cochrane and Piazzesi (2005, 2008), Adrian, Crump and Moench (2013) and Hellerstein (2011)).[15] The message here is that variables virtually irrelevant in one measure may become important when the two measures are linked. More about this later. *Structural no-arbitrage models constitute the class of models this book is about*.

3. *'Snapshot' models* (such as the Nelson–Siegel (1987) model, or the many splines models of which Fisher, Nychka and Zervos's (1995) is probably the best known) are cross-sectional devices to *interpolate* prices or yields of bonds that we cannot observe, given a set of prices or yields that we *can* observe.[16] They also produce as a by-product the model yields of the bonds we *do* observe. If supplemented with

[14] p. 76.

[15] See in this respect the discussion on p. 140 of Cochrane and Piazzesi (2005) and on p. 3 of Hellerstein (2011).

[16] For two early, but still valid, evaluations of yield-curve estimation models, see Bliss (1997) and Anderson et al. (1996).

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

the ubiquitous but somewhat ad hoc assumption that the residuals (the differences between the model and the market prices) are mean reverting, these models give practitioners suggestions about whether a given observed bond yield (hence, price) is 'out of line' with a reasonable smooth interpolation of where it should lie.[17] Liquidity corrections such as those discussed in Fontaine and Garcia (2008) can be very important in these 'cheap/dear' analyses.

Apart from the smoothness-based assessment of the relative cheapness or dearness of different bonds, snapshot models are extremely important for structural affine models because they assume the existence of a continuum of discount bonds. So the output of snapshot models (a snapshot discount function) is the input to structural models.

In general, there is no deep meaning to the parameters of fitted snapshot models. However, some recent developments have given a time-series, dynamic interpretation to their parameters, and married them with Principal Component Analysis. (See, eg, (Diebold and Rudebusch, 2013).) So, these latest developments combine features of structural, statistical and snapshot models. We shall revisit this approach later in the chapter.

4. *Derivatives models* (eg, the Heat–Jarrow–Morton (1992), the Brace–Gatarek–Musiela (1997), the Hull and White (1990), the Black–Derman–Toy (1990), …) are based on *relative* pricing and on the enforcement of no-arbitrage. Because of this, they strongly rely on first-order cancellation of errors (between the derivative they are designed to price and the hedging instruments used to build the riskless or minimum-variance portfolio; see the discussion in Nawalha and Rebonato (2011)). Therefore they do not strive to provide a particularly realistic description of the underlying economic reality. After the first generation (Vasicek (1977), Cox et al. (1985a,b), derivatives models squarely set up camp in the risk-neutral $\mathbb{Q}$ measure, and affect a disdainful lack of interest for risk premia. I do not deal with this class of models in this book.

1.5 THE CHOICE OF VARIABLES*

1.5.1 Latent versus Observable Variables

As mentioned previously, an important theme that recurs throughout the book is that the choice of the type of state variable is a very important, and often

[17] Snapshot models are also important because all structural models use as their building blocks discount bonds, which are not traded in the market but which make mathematical analysis (immensely) easier. The output of snapshot models (the discount curve) is therefore the input to structural models.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

neglected, aspect of term-structure modelling. In this section I aim to give a first explanation of why this is the case. This is only part of the story, as the plot will thicken in Chapters 27 and 29. A health warning: this section requires an understanding of modelling issues and of mathematical formalism that is introduced in the body of the book. Consequently, it may be rather opaque at the moment and, as the saying goes, can be skipped on a first reading without loss of continuity.[18] These readers can then come back to this section after reading Chapters 27 and 29.

Every model comes equipped with a number of parameters (*constant*[19] quantities that describe the plumbing of the model – say, the volatility of the short rate or the speed with which a variable returns to its reversion level), and a, usually much smaller, number of state variables, ie, quantities that, according to the model, should vary *stochastically* during the life of a bond.

What is a parameter and what is a state variable is a modelling choice, not a fact of nature: for instance, in one affine model (say, the Vasicek) the volatility of the short rate may play the role of a parameter; in another (say, the Longstaff and Schwartz (1992) model) it may become a state variables; ditto for the reversion level. So, the choice of what to treat as a fixed building block and what to model as a stochastic variable reflects a messy trade-off between richness of the description, ability to estimate the model parameters,[20] analytical tractability, parsimony of the model and the aesthetic sensitivity of the modeller.

The difficult choices faced by the model developer are not limited to the state-variable/parameter dichotomy. She will also have to choose the nature of the state variables. Two main routes are open here: the latent-variable and the specified-variable approaches.

With the first approach the modeller will start from some latent (unspecified) variables and impose that these variables (whatever their meaning) should follow a particular process; she will impose a simple link (typically an affine transformation) between the latent variables and some observable variables; and she will then estimate indirectly the statistical properties (the parameters) of the latent variables, usually by econometric analysis (say, using Kalman filter techniques) of the time series of the observable quantities. The D'Amico, Kim and Wei (2010) model, that we study in Chapter 31, is a prime and popular example in this mould.

The beauty of the latent-variable approach is that we do not make any assumptions (which could, of course, be wrong) about what 'really' drives the yield curve. The drawback likewise is that do not make any assumptions (right

[18] Sections marked with an asterisk can be skipped on a first reading.

[19] Parameters may have a deterministic time dependence, but in this case the 'meta-parameters' of the deterministic function of time become the constant quantities.

[20] Whenever a parameter is 'promoted' to a state variable, it receives as a dowry its own set of process parameters: so, for instance, the moment we allow the volatility to become stochastic, we are immediately faced with the problem of assigning the volatility of volatility, its drift and the correlation between the shocks to the volatility and the shocks to the yield curve.

or wrong as they may be) about what 'really' drives the yield curve. This has several unpleasant consequences.

To begin with, it is difficult to restrict, on the basis of our understanding of the meaning of the variables we use, the number of the admissible values for the model parameters. The price to pay for enforcing a Newtonian *hypotheses-non-fingo* attitude to the choice of variables is the risk of overparametrization: once p observable variables and their q lags are added to m latent variables one has to deal with $(p+m)(pq+m)$ parameters, which can quickly add up to $O\left(10^1\right)$ if not $O\left(10^2\right)$ 'degrees of freedom'.[21] As Johnny von Neumann pointed out, '[w]ith four parameters I can fit an elephant, and with five I can make it wiggle his trunk'.[22] As Mayer, Khairy and Howard (2010) prove, this is no empty boast, and Figure 1.1 shows how they indeed achieved the feat with four (complex) parameters.

The second problem with latent-variable approaches is that they do not lend themselves to easy 'sanity checks'. For instance, if we estimate a reversion level of the short rate, after making the appropriate adjustments for risk aversion, we can assess if this roughly squares with past experience and future expectations. Or, from the estimated reversion-speed coefficient for the target rate, we can impute, again after adjusting for risk, a half-life for the short rate, and assess whether this is reasonable.[23] But how are we to make these semi-quantitative sanity checks for the reversion level or reversion speed of a variable whose meaning we do not specify?

One can rebut: surely, a latent-variable model provides a mapping (a 'dictionary') capable of translating latent variables into observables. If this is the case, does it really make much of a difference whether we work with latent or observable variables, as long as we can go to our model dictionary and translate from one set of variables to the next? Can't we do our sanity checks after looking up the variable translation in our dictionary?

Indeed, it would make little difference if there were a unique correspondence between sets of acceptable values for the observable variables and combinations of latent variables. However, more often than not, very different combinations of latent variables can give rise to observable quantities in very similar ranges, as shown in the lower half of Figure 1.2. If we find ourselves in the case depicted in the lower half of the figure, which of the latent-variable bubbles (all converging into the same region of acceptability for the observable variables)

[21] The Ang and Piazzes: ((2003)) model, with its 2 observable variables, their associated 12 lags, and the 3 latent factors requires 135 parameters for the pricing kernel to be defined. Similarly, the Gaussian QMLE affine model discussed in Diebold and Rudebusch (2013) comes equipped with 139 parameters eager to be estimated: in their words, this is 'a challenging if not absurd situation' (pp. 37–38).

[22] Quoted in Mayer et al. (2010).

[23] Some lazy people say 'After adjusting for risk (ie, after moving from one measure to the other) 'anything can happen to the drift, and the risk-neutral drift could become anything '. This is emphatically not true, and this is where a 'principled' structural approach to term-structure modelling makes a difference. In reality, any transformation from the real-world to term-structure the risk-neutral measure implies a price of risk function that can and should be interrogated for plausibility, and for consistency with the empirical information. (Much) more about this later.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

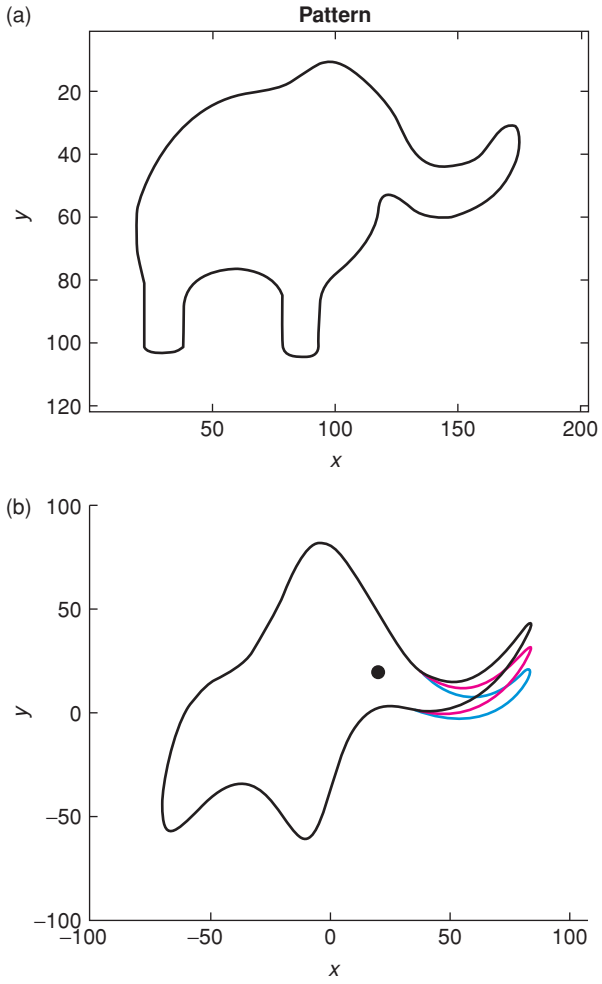

Figure 1.1 (a) As shown, it is indeed possible to draw an elephant with four (complex parameters). (b) With a fifth, Mayer et al. (2010) can both locate its right eye, and make it wiggle its trunk.

are we going to choose? In essence, the problem is the following: we have some observable variables (such as, say, market yields or yield volatilities) that we use to calibrate our model. We are interested in some not directly observable quantities (such as, say, expectations or risk premia). In the upper half of Figure 1.2 we then have two sets of values for latent variables. One set maps to 'good' values for the observable quantities to which we 'fit' our model and the other to 'bad' values. The two sets make different predictions for the quantities we are actually interested in gaining information about (say, the risk premia), but it is not difficult to choose which set of values for the latent variables we should choose: the upper one, that maps into the 'good' region for the fitting observables.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

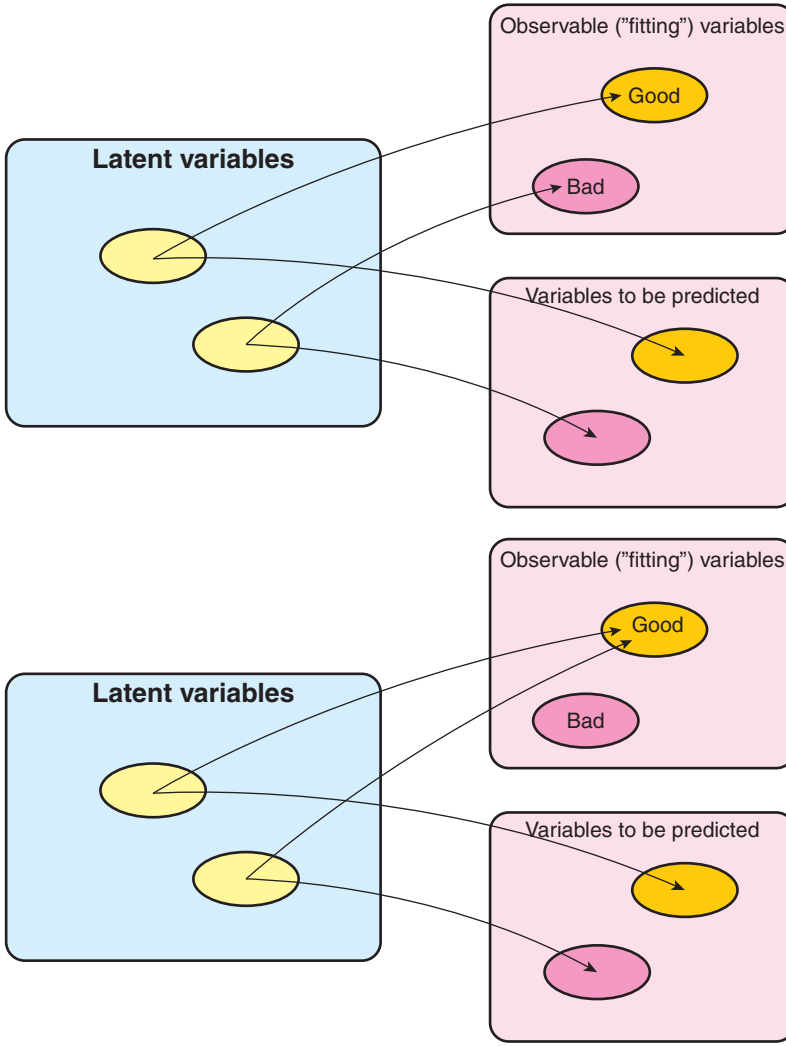

Figure 1.2 A 'helpful' (upper panel) and 'awkward' (lower panel) mapping from latent variables to observables and predicted quantities. See the text for a discussion.

The situation is trickier in the lower half of Figure 1.2: now two very different combinations (sets) of values for the latent variables map into the same region of acceptability for the observable quantities. However, they give rise to very different predictions for the future values of the variables-which, after all, is what we are interested in obtaining. Which set should we choose?

So much for (some of) the problem one faces using latent variables. Not everything is problem-free, however, when we use identifiable state variables. We have two distinct sets of problem here: the spanning problem and the constraint problem. Let's look at both in turn.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

1.5.2 The Spanning Problem*

The first requisite of a successful set of state variables is that they should span the changes in the yield curve. But there is an additional requirement: if we are interested in a structural description, we must also include variables that describe the risk premia. There is no a priori reason why the two sets of variables should coincide – ie, that the variables that 'span' the yield curve variations should also account for the market price of risk.

What does this mean in practice? Every term-structure model establishes a mapping between the state variables and the observable yields. Suppose that, as is the case for affine models, this link is linear. Also suppose that, departing from the latent-variable approach, we have specified the exact nature of our state variables. For instance, if we took a macrofinancial approach, we may have chosen as state variables a set of macroeconomic quantities; or perhaps we may require our state variables to be some of the yield-curve principal components.[24] Now, roughly speaking, a good spanning is achieved when, given the mapping afforded by the model, the model-produced variability in the observable yields 'looks like' the yield variability observed in reality.

Whether this turns out to be the case strongly depends on the variables we have chosen. And, with specified-variable models, just throwing in more and more variables is no guarantee that we are going to fare any better. To explain why this may be the case, let's consider a contrived example. Suppose that a modeller decided that what drives the yield curve is real economic activity – and nothing else. On the face of it, the claim sounds somewhat audacious, but not absurd. Suppose also that the same modeller established that changes in real economic activity are associated with[25] changes in the slope, but not the level, of the yield curve. (Again, I am not saying that any of this is true: I am just explaining what 'spanning' means.) Armed with this insight, the modeller decides to use real economic activity as *the* driver of the yield curve. Then it is clear that changes in real economic activity, given the mapping allowed by the model, will generate a change only in the slope, but not in the level, of the yield curve. But this is not what we observe in reality, where changes in the level usually dominate slope changes. Here, the chosen state variable did not adequately span the observed changes in what we want to describe.

The example was obviously far from realistic, but the problem of ensuring that state variables chosen a priori do span changes in the yield curve is a real one. The model we propose in Chapter 35 suggests one possible solution to this problem.

There is another aspect to the spanning problem. If we are interested purely in pricing – and, therefore, not in a structural description that weaves together expectations and risk premia – the only spanning requirement is that the chosen variables should be able to account for the observed variability in the observed *market* yields. Recall, however, that market yields are made up of an

[24] We explain what principal components are in Chapter 6.

[25] And *fully* reflected in – see the end of the section about this important point.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

expectation and a risk-premium component.[26] If we are interested in modelling risk premia as well, it is not obvious that 'just' knowing about the present and past prices (yields) can give us the optimal information set for the prediction of excess returns. It may very well be – and indeed, Ludvigson and Ng (2009) show that it *is* the case – that macroeconomic variables have explanatory power over and above what is embedded in the prices. This raises an important question, that we discuss towards the end of the book: does the traditional affine modelling setting allow a simple 'augmentation' of the yield-curve–based state variables to include the *full* information provided by the macro quantities? Can we do this in theory? Can we do so in practice? Understanding this point would require too long a detour at this stage, but it is important to keep this *caveat* in mind for later discussion.

1.5.3 The Constraint Problem*

As mentioned previously, there is a second problem when one uses non-latent (specified) variables – the 'constraint' problem. To understand the nature of this problem, a good place to start is with the work by Dai and Singleton (2000). In an important paper published at the beginning of the new millennium, they produced a very general classification of affine models and they obtained the following result. Start from N state variables (factors), x_t, and impose that they should follow a diffusive process of the form[27]

$$dx_t = a\left(x_t\right)dt + \sqrt{b(x_t)}dz_t. \tag{1.1}$$

As for the "drift' and 'variance' coefficients', $a\left(x_t\right)$ and $\underline{b}(x_t)$, they can depend on the state variables, $\overrightarrow{x}_t$, but only in a linear (or, rather, affine[28]) fashion

$$\begin{aligned} \text{drift} &= a_0 + a_1 x_t, \\ \text{variance} &= b_0 + b_1 x_t. \end{aligned} \tag{1.2}$$

Now, if the short rate, r_t, can be written as a linear combination of these N factors plus a constant,

$$r_t = c_0 + c_1^{\mathrm{T}} x_t, \tag{1.3}$$

then Dai and Singleton (2000) show that the bond prices, P_t^T, can always be written as exponentially affine functions of the factors, ie, as a function of the form

$$P_t^T = e^{A_t^T + \left(\overrightarrow{B}_t^T\right)^{\mathrm{T}} \overrightarrow{x}_t}. \tag{1.4}$$

[26] And, of course, a convexity contribution as well.

[27] We have not introduced our notation yet, and the reader may at this stage be unfamiliar with the matrix formulation of a mean-reverting process – or may not have seen a mean-reverting process at all. We ask the reader to 'go with the flow' for the moment and promise that all we be explained in due course. In particular, we define and explain the matrix notation in detail in Chapter 17.

[28] We discuss the difference between linear and affine function in Chapter 18.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

Note that, apart from the short-rate requirement that $r_t = c_0 + c_1^{\mathrm{T}} x_t$, the factors can be totally general. However, we know[29] that the time-t yield for maturity T is defined as

$$y_t^T = -\frac{1}{T-t} \log P_t^T. \tag{1.5}$$

By taking the logarithm of the expression for the bond price (Equation (1.4)), we see that in affine models yields always have this very simple form:

$$y_t^T = u_t + \overrightarrow{g}^T \overrightarrow{x} \tag{1.6}$$

for some row vector $\overrightarrow{g}^T$.

Now, in specified-variable approaches, the modeller assigns a priori the link between the state variables and the yields. Let this assigned relationship be of the form

$$y_t^T = \phi_t + \overrightarrow{\Phi}^T \overrightarrow{x}. \tag{1.7}$$

For instance, Duffie and Kan (1996) simply identify the factors with the yields themselves ($\phi_t = 0$ and $\Phi_t = I$). More interestingly, as we have seen, macro-financial models link the observable yields (or linear functions thereof) to macroeconomic observables via some structural models. Or, again, we may use as state variables some special combinations of yields, such as their principal components. See, eg, the approaches described in Chapters 33 and 34.

In general, modifying the terminology in Saroka (2014), let's call *specified-variable models*[30] all models in which the loadings ϕ_t and Φ_t are assigned *a priori* by the modeller on the basis of her knowledge of (or prejudices about) how the world works.

Now, as we saw earlier, working with non-latent factors has obvious important advantages. However, even leaving to one side the spanning problem alluded to above, Equations (1.6) and (1.7) immediately suggest that working with prespecified factors must bring about strong issues of internal consistency: indeed, once absence of arbitrage is imposed, any exogenous, a priori specification of the loadings ϕ_t and Φ_t must imply a relationship between yields-as-by-products of the model (Equation (1.6)) and yields-as-specified-by-the-modeller (1.7). This means that, for consistency, we must have

$$\phi_t = u_t \tag{1.8}$$

$$\Phi_t = g_t^T. \tag{1.9}$$

As we shall see, this can place severe restrictions on the admissible stochastic behaviour of the state variables.[31]

Let's give a concrete example. Suppose that we choose to work with principal components as state variables – a natural enough choice (which we pursue in Chapter 33) given how much we know about the principal components obtained

[29] See Chapter 2.

[30] Saroka (2014) calls them *observable affine-factor models*.

[31] This topic is discussed in detail in Joslin, Singleton and Zhu (2011).

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

from the covariance matrix of yield curve changes. (Why work with specified variables unless we know a lot about them, after all?) As principal components are a particular linear combination of yields, our principal component–based model would certainly fall in the specified-variable model category.

Given this choice of principal components as state variables, let's now evolve them to some future time, τ. These future values of the state variables determine the future bond prices via Equation (1.4). And, as we have seen, in an affine setting future bond prices require future yields to be a linear (affine, really) function of the state variables – see again Equation (1.6).

But remember that we have required our state variables to be principal components. We have not defined principal components yet, but at this stage we can just say that they are some special (and fixed!) linear combinations of yields – such as the linear combinations in Equation (1.7). But if this is the case, the econometrically determined coefficients that link at all times yields and principal components must be linked to the model coefficients that give yields as linear functions of log prices! Achieving this internal consistency, as we shall see, is not a trivial task.

This is another way to look at the same problem. In an affine framework, we may then *like* to require that each principal component should display a nice and simple mean-reverting behaviour.[32] For instance, we may want to impose that, as each principal component moves away from its own reversion level, it is attracted back towards it by a 'spring' of appropriate strength, and that the strength of this mean-reversion is unaffected by what the other principal components are doing.[33]

We may well *like* to impose this simple behaviour, but the internal consistency constraints mentioned above and discussed in detail in Chapter 33 (see also Rebonato, Saroka and Putyatin (2017) and Saroka (2014)) tell us that it is not in our gift to do so.

In sum, if the reader is still with us, the message of this section is that the problem with specified-variable models is general: by specifying the variables, we add constraints to their dynamics over and above the no-arbitrage restrictions. These constraints come from imposing Equations (1.6) and (1.7). Taken together, these two sets of equations will dictate part of the dynamics of the state variables. So, we *can* choose principal components (or any other set of specified variables) as state variables, *or* we can choose (latent) state variables and assign a nice and simple mean-reverting behaviour (in the $\mathbb{Q}$-measure). What we *cannot* do is choose principal components (or any other set of specified variables) as state variables *and* impose that they should follow a 'nice and simple' mean-reverting behaviour (in the $\mathbb{Q}$-measure).

I must stress here that there is nothing special about principal components in this impossibility result. Once we have made a priori the modelling choice

[32] We should add 'in the $\mathbb{Q}$ measure'. Please bear with us and go with the flow for the moment.

[33] The situation we are describing here corresponds to a diagonal reversion-speed matrix. For readers not familiar with mean-reverting (Ornstein–Uhlenbeck) processes, we give a first intuitive presentation in Chapter 8, and a more thorough treatment in Chapters 15 to 17 and 33.

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

of Equation (1.7), ie, once, on the basis of our domain knowledge, we choose to assign the link between the yields and the state variables, we lose part of our ability to assign the precise nature of their dynamics.

All of this may sound rather abstract at this stage, but it will, hopefully, become clearer when we deal with the various models.

1.6 WHY DO WE NEED NO-ARBITRAGE MODELS AFTER ALL?

Before getting started in earnest we have to answer one more important question.

I said that a substantial part of this book is devoted to term-structure models. What's so good about models, and about *no-arbitrage* models in particular? Suppose that our main interest is in predicting future rates, or in decomposing yields into the three components of expectations, term premia and convexity. Why can't we just rely on statistical regularities, as uncovered by careful econometric analysis, to extract this information?

To be even more concrete, suppose that we look at the strategy of investing in an n-year maturity bond, funding it with a 1-year-maturity bond, selling the $n-1$-maturity bond after one year, and repaying our loan. (This is what excess return studies essentially investigate.) We want to know in what configurations of the yield curve (level, slope, curvature), or in what states of the economy, we expect this strategy to be profitable. Why are statistical models not enough to answer this question? If no arbitrage opportunities are indeed found in the market, surely the observed market prices should reflect this. To the extent that econometric estimation reveals these regularities, the no-arbitrage conditions should automatically be present in the estimated models.[34] Granted, by themselves statistical models may not explain a lot, but, as far as detecting empirical regularities, surely they should be unsurpassed. Or are they?

One could take an even more provocative stance. Unless one determines the market price of risk from an equilibrium asset model and from the utility function of the representative investor – a feat that, for the purpose of predicting excess returns, very few modellers are brave enough to attempt – one has to take a rather uninspiring two-step approach: first one must estimate risk premia by extracting empirical information about excess returns and/or the real-world behaviour of rates (say, their reversion levels, their reversion speeds, if any, etc). Then one has to pour this empirical content into the often-funny-shaped vessel of the model. (I describe the model as a funny-shaped vessel because one can carry out this exercise in translation of information only in the rather restricted way that any given model allows.)

So, a cynic may say, the model purely regurgitates the empirical econometric information it has been fed, and can only do so imperfectly: it is a halting and

[34] Diebold and Rudebusch (2013, Section 1.5.3, p. 16 and passim) ask the question: 'Is the Imposition of No-Arbitrage Useful?': 'if reality is arbitrage-free, and if a model provides a very good description of reality, then imposition of no-arbitrage would presumably have little effect.'

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

stuttering rendition of lines it has 'learnt by heart', not arrived at itself. (The lines are 'learnt by heart' because we almost certainly eschewed the utility-function-based approach.)

There is more than a grain of truth in these objections. Still, I think that there is value in a disciplined and skeptical use of models. Of course, a model affords an *understanding* of how reality works that no purely-data-driven statistical analysis can afford. But I believe that a model can tell us something useful even if we are interested only in prediction. As I think that my case for the importance of models can be better made after reviewing the actual model performance than in the abstract, I will defer the 'argument for the defence' to the last chapter of the book. This 'existential' question, however, should be kept vividly in mind by the critical reader.[35]

1.7 STAMP COLLECTING AND SHALLOW VERSUS DEEP EXPLANATIONS

I said in the opening paragraphs of this chapter that my goal is to present a *structural* approach to yield-curve modelling. Having said that, it is good to keep in mind that the boundaries between descriptive and structural approaches are a bit arbitrary, and that what we may proudly call 'structural' in this book could be regarded as 'descriptive' by a hard-core financial economist. To understand this point, let's consider again for a moment one important component of yield-curve modelling, the explanation of excess returns.

We can start from the observation that term premia seem to be time varying. If we stop here, we are clearly just describing. After carrying out some clever regressions, however, we may find that, say, the slope of the yield curve 'explains' a large portion of these time-varying excess returns. We may feel better, because we are now able to explain the time variation of excess returns in terms of something else: term premia are high when the curve is steep, and low when it is flat or inverted. But why is it so?

Perhaps we can relate the slope of the yield curve to the business cycle. This sounds encouraging. But this new explanation just moves the goal posts: why should the slope of the yield curve be linked to the business cycle?

[35] I stress that what I provide in Chapter 31 is *my* explanation of why models are useful. For a different discussion of the need for no-arbitrage ('cross-sectional') restrictions, see Piazzesi (2010), p. 695 and passim. Ang and Piazzesi (2003) also discuss from an econometric perspective how imposing no-arbitrage helps the out-of-sample prediction of yields. In particular, Piazzesi (2010, pp. 694–695) mentions among the five reasons for using no-arbitrage models the problem of the 'missing bond yields', ie, yields for 'odd' maturities whose value can be recovered using a no-arbitrage model from a small set of reference yields. As Piazzesi (2010) points out, this can be important for markets with sparse reference points (such as energy markets), but this is unlikely to be a major consideration for the government bond markets of most G7 economies. Piazzesi (2010, p. 695) also mentions the advantage of having consistency between the time series and the cross-sectional yield equations, and the ability to split an observed yield into its expectation and term-premium components. Regarding the last split, it should be pointed out that the same split can also be achieved by a statistical analysis of excess returns.

Perhaps we can come up with a compelling story about state-dependent (in this case, business-cycle–dependent) degrees of risk aversion. Or perhaps we can look at how other asset classes behave (co-vary) during different phases of the business cycle, and argue that in some parts of the business cycle bonds act as effective diversifiers.[36] Or whatever. Also this, however, does not provide the 'final' answer to the series of new questions that every new explanation opens up.

The point here is that even the most 'structural' approaches sooner or later end up hitting against a hard descriptive wall. Perhaps the goal of most scientific enquiries is to make this collision occur later rather than sooner. Physicists used to say that, outside their domain, everything is just stamp collecting. This may well be true, but physicists, too, at some point, must begin to collect stamps.

This important reminder is to put in perspective what we are trying to do in this book. First, we do not want to stop at a 'shallow' explanation of bond behaviour, an explanation which is but a small step away from the most basic observations. ('Bonds are exposed to the risk of rates going up' is one such shallow explanation.) Like good physicists, we want to delay at least for a while the moment when we begin collecting stamps. At the same time, we must keep in mind that the idea of reaching the 'ultimate' explanation is futile, and that we will stop somewhere halfway between an eighteenth-century-like specimen collection of 'curios' and the Grand Unified Theory.

1.8 THE IDEAL READER AND PLAN OF THE BOOK

> "And what is the use of a book," thought Alice, "without pictures or conversations in it?"
>
> *Alice's Adventures in Wonderland*, L. Carroll

Who is the ideal reader of this work? The stock-in-trade recommendation for a writer is to imagine the 'ideal reader' peeking over her (that is, the writer's) shoulder as she writes. This is all well and good, but in this case of limited help, because the ideal reader I have in mind will change significantly from the first pages of the book to its last chapters. How so?

Of course, I hope that, towards the later chapters of the book, the reader will have become progressively more familiar and comfortable with some simple mathematical techniques that I will introduce as we go along, and with which she may have not been familiar from the start. But, more importantly, I as well hope that she will have also become subtler and more sophisticated in her thinking about term-structure modelling. This changing reader will therefore have to deploy not just a wider and wider set of mathematical tools, but also a progressively subtler and deeper mode of financial reasoning.

[36] On the important topic of the stock-bond correlation, see the good paper by Johnson et al. (2015).

https://doi.org/10.1017/9781316694169.001 Published online by Cambridge University Press

The tone of the book, and the level of mathematical sophistication will therefore change, and become somewhat more demanding, in the later parts if the book. I have always chosen, however, the path (or the shortcut) of least mathematical resistance, and, given a choice, I have invariably chosen the mathematically simplest (yet correct) way to present the topic at hand.

As a consequence, I think that an undergraduate student, an MBA student, an MSc student or a quantitatively conversant investment professional should certainly be able to follow the arguments and the derivations presented in the book. However, I believe that also a graduate student, a proficient 'quant' or an academic will find in the book a fresh perspective, and something to agree or disagree with. I hope, in sum, that this book will make the novice think, and the expert think again.[37] Above all, I have strived to provide the precise tools with which all readers can reproduce the results presented in the body of the work, and tinker with their own variations on the affine theme.

The book is therefore organized as follows.

In Part I (which contains this introduction) I lay the foundations of the book: I state what the topic is, I explain my goal and my strategy, I define my notation, I introduce some mathematical tools, and I present some topics (such as some rudiments of Monetary Economics) that will appear over and over again in the body of the work. Chapters 4 and 5 are particularly important in Part I, because they give the first introduction to the risks to which a bond is exposed and the compensations investors can exact for bearing these risks. It is here that the real-world and risk-neutral measures make their first appearance.

Part II is devoted to presenting two of the three building blocks of term-structure building, namely, expectations and convexity. Convexity will be revisited in greater detail in Part V, but, in keeping with my general strategy, I have chosen to offer early on a taste of the main course. As another amuse bouche, I present at this early stage (Chapter 8) an incomplete, but hopefully inspiring, first look at the Vasicek model.

In Part III I introduce the glue that holds together the three building blocks, namely the conditions of no-arbitrage. I do so from a variety of perspectives, with different levels of sophistication, and for different types of assets (ie, nominal and real bonds). In particular, I present and derive the no-arbitrage conditions using both the traditional (Partial Differential Equation [PDE]–based) approach and the modern language based on the stochastic discount factor and the pricing kernel.

With all of the building blocks in place, and the conditions of arbitrage thoroughly explained, I return in Part IV to the Vasicek model (Chapter 16), with a simple derivation of its salient results, and a deeper discussion of its strengths and weaknesses than what was presented in Chapter 8. By the end of Chapter 16 the reader will have understood not only how to 'solve' the Vasicek model, but also the reason why, despite it being so dearly loved by modellers who prize parsimony and simplicity, it can take us only on part of our journey, and why more complex approaches will be needed.

[37] Paraphrased from Jackson (2015, p. 9).

In this part of the book I also present a generalization to many state variables of the results obtained in Chapter 16 for the Vasicek model. Chapter 17 provides a gentle introduction to the notation and the techniques presented in Chapter 18. In this latter chapter general pricing results are presented that the reader will be able to apply to virtually all the models that she will come across – or that she may care to create herself. Chapter 19 closes Part IV with a discussion of the shadow rate – a topic of salient relevance especially, but not only, during times of ultra-low rates.

In Part V I return to the topic of convexity, and I present in this part of the book both theoretical and empirical results.

Part VI, which deals with excess returns, is very important, because it presents the bridge between the real-world and the risk-neutral description. The particular bridge we have chosen in order to cross this chasm is the empirical study of excess returns, to which we devote no fewer than seven chapters. I present in this part of the book a detailed critical discussion of the traditional (eg, Fama and Bliss, 1987) and of the modern (eg, Cochrane and Piazzesi, 2005; Cieslak and Povala, 2010a) return-predicting factors. Once these empirical results are presented, the reader will also understand another important shortcoming of simple Vasicek-like models when it comes to a structural account of yield curve modelling – in a nutshell, why the Vasicek form of the market price of risk is *qualitatively* wrong.

Now that the case for progressing beyond the simple Vasicek framework has been fully made, in Part VII I present and critically discuss a number of models that, to different extents and from different perspectives, attempt to overcome the limitations of the simple Vasicek-like models discussed in Parts I to VI. This is no encyclopaedia of Gaussian affine models (no such thing can exist); rather, it is a presentation of models organized in such a way as to answer, in turn, some of the questions raised by the analysis in Parts I to VI.

By the end of Part VII the inquisitive ideal reader will, indeed, have changed a lot, and should be ready for experimenting with her own version of a Gaussian affine model in order to deal with the specific problems she may face. This is, indeed, what several of my Oxford students have chosen to do after reading early drafts of this book.

To conclude, a few words about the figures. In the hope that even Alice may find in this book something to interest her, I have given special care to the many pictures that complement the text. These are not meant as decoration for an otherwise dull page, but, with their captions, are an integral part of the story the book tells. The reader will find her efforts well rewarded if she spends almost as much time examining the graphs and their descriptions as reading the text. However, sorry, Alice, I have not been able to put a lot of conversations in my book.