Chapter 1

Theory and Practice of Option Modelling

1.1 The Role of Models in Derivatives Pricing

1.1.1 What Are Models For?

The idea that the price of a financial instrument might be arrived at using a complex mathematical formula is relatively new, and can be traced back to the Black-and-Scholes (1973) formula.[1] Of course, formulae were used before then for pricing purposes, for instance in order to convert the price of a bond into its gross redemption yield. However, these early (pre Black-and-Scholes) formulae by and large provided a very transparent transformation from one set of variables to another, and did not carry along a heavy baggage of model assumptions. The Black-and-Scholes formula changed all that, and we now live in a world where it is accepted that the value of certain illiquid derivative securities can be arrived at on the basis of a model (the acceptance of this is the basis of the practice of marking-to-model).

The models that developed from the family tree that has Black-and-Scholes at its roots shared the common assumptions that the estimation of the drift (growth rate, trend) component of the dynamics of the relevant financial driver was not relevant to arrive at the price of the derivative product. This insight directly follows from the concept of payoff replication, and is discussed in detail in this book in Chapter 2.

In order to implement these models practitioners paid more and more attention to, and began to collect, direct empirical market data at a very 'atomistic' (often transactional) level. This was done for several reasons: for instance, for assessing the reasonableness of a model's assumptions, or for seeking guidance in the development of new models, or for estimating the inputs of existing models. The very availability of this wealth of information, however, suggested new opportunities. Perhaps, embedded in these data, there could be information about the market microstructure that could provide information not only about the 'volatility' of a price series, but also about its short-term direction.

[1]Parts of this chapter have been adapted from Rebonato (2004) and from Rebonato (2003b)

Again, the practice was not strictly new, since the idea of predicting future price movements from their past history ('chartism' in a generalized sense) pre-dated Black-and-Scholes probably by decades. Yet these earlier approaches (which, incidentally, never won academic respectability), were typically based on, at most, daily observations, and purported to make predictions over time-scales of weeks and months. The new, transactional-level data, on the other hand, were made up of millions of observations, sometimes collected (as in the case of FX trades) minutes or seconds apart. The availability of these data made possible the calibration of *predictive* models, which try to anticipate stock price movements over time-scales sometimes as short as a few minutes.

This was just the type of data and models that many of the new, and, in the early 2000s, immensely popular, hedge funds required in order to try to 'get an edge' over an ever-growing competition (in 2001 one new hedge fund was being launched every week in continental Europe alone). These hedge funds and the proprietary trading desks of internationally active banks therefore become the users and developers of a second breed of models, which differed from the members of the Black-and-Scholes family because they were explicitly trying to have a predictive directional power. Unlike the early rather crude chartist approaches, these new models employed very complex and sophisticated mathematical techniques, and, if they were not being routinely published in academic journals, it had more to do with the secretive nature of the associated research than with any lack of intellectual rigour.

Two types of model had therefore developed and coexisted by the end of the 1990s: models as predictors (the 'hedge-fund models') and models as payoff-replication engines (the 'derivatives models'). With some caveats, the distinction was clear, valid and unambiguous. It is the derivatives models that are the subject of this book.

Having said that, recent developments in the derivatives industry have increasingly blurred this once-clear-cut distinction. Products have appeared whose payoff depends to first order on, say, the correlation between different equity indices (e.g. basket options), or on the correlation between FX rates and/or between different-currency yield curves (e.g. power-reverse-dual swaps), or on possibly discontinuous moves in credit spreads and default correlations (e.g. tranched credit derivatives). The Black-and-Scholes-inspired replication paradigm remains the prevalent approach when trying to price these new-breed models. Yet their value depends to first order on quantities poorly hedgeable and no easier to predict than directional market trends. The underlying model might well assume that these input quantities are deterministic (as is normally the case for correlations), but this does not take away the fact that they are difficult to estimate, that they render payoff replication very complex, if not impossible, and that their real-world realizations influence to a very large extent the variability of the option-plus-hedge portfolio. The result of this state of affairs is that, once the best hedging portfolio has been put in place, there remains an unavoidable variance of return at expiry from the complex option and its hedges.

Sure enough, *perfect* replication cannot be expected even for the simplest options: markets are not frictionless, trading cannot be continuous, bid–offer spreads do exist, etc. Yet the robustness of the Black-and-Scholes formula (discussed at several points in this book) ensures that the terminal variability of the overall total portfolio is relatively limited. The difference, however, between the early, relatively simple option payoffs and the new, more complex products, while in theory only a matter of degrees, is in practice large enough to question the validity of prices obtained on the basis of the replication approach (i.e. assuming that one can effectively hedge all the sources of uncertainty).

So, in making the price for a complex derivative product, the trader will often have to take a directional market view on the realization of quantities such as correlations between FX rates and yield curves, default frequencies or correlations among forward rates in different currencies and/or equity indices, or sub-sectors thereof. As a consequence, the distinction between predictive models (that explicitly require the ability to predict future market quantities), and models as payoff replication machines (that are supposed to work whatever the future realizations of the market quantities will be) has recently become progressively blurred. This topic is revisited in the final section of this chapter, where I argue that one could make a case for re-thinking current derivatives pricing philosophy, which still implicitly heavily relies on the existence of a replication strategy.

1.1.2 The Fundamental Approach

In option pricing there are at least two prevalent approaches (which I call in what follows the 'fundamental' and 'instrumental') to dealing with models. The general philosophy that underlies the first can be described as follows. We begin by observing certain market prices for plain-vanilla options. We assume that these prices are correct, in the sense that they embody in the best and most complete possible way all the relevant information available about the stochastic process that drives the underlying (and possibly, other variables, such as the stochastic volatility; for simplicity I will confine the discussion to the underlying, which I will also call 'the stock'). We begin by positing that this true process is of a particular form (say, a jump–diffusion). We calculate what the prices should be if indeed our guess was correct. If the call prices derived using the model are not correct, we conclude that we have not discovered the true process for the underlying. If they are better than the prices produced by another model (say, a pure diffusion) we say that we have reason to believe that the new process (the jump–diffusion) could be a more accurate description of the real process for the underlying than the old one (the pure diffusion).

Alternatively, if the model has a large number of 'free parameters' and we believe that the underlying process is correctly specified, we use all of the parameters describing the dynamics of the underlying to recover the market prices of the options. This is what is implicitly done, for instance, with some implementations of the local volatility models (see Chapters 11 and 12).

Or again, if two models give a fit of similar quality to market prices of plain-vanilla options, the conclusion is often drawn that the model that implies the process for the underlying more similar to what is statistically observed in reality is the 'better' one. The relevance of this distinction is that, despite the similarity of the two models in reproducing the *plain-vanilla* prices, better prices for *complex* options (i.e. prices in better agreement with the market practice) would be obtained if the superior process were used.

The fundamental approach sounds very sensible. It is, however, underpinned by one very strong assumption: the trader who chooses and calibrates models this way is subscribing to the view that the market-created option prices must be fully consistent with the true, but a priori unknown, process for the underlying. The market, in other words, must be a perfect information-processing machine, which absorbs all the relevant information about the unknown process followed by the 'stock', and produces prices consistent with each other (no arbitrage) and with this information set (informational efficiency).

This implicit assumption is very widespread: take, for instance, the practice of recovering all the observable option prices using a local-volatility model, discussed in Chapter 11. Even if we knew the true process *of the underlying* to be exactly a diffusion with state-dependent (local) volatility, it would only make sense to determine the shape of the local volatility from the traded option prices if we also believed that these had been correctly created in the first place on the basis of this model. We will see, however (see Chapter 11), that the local-volatility modelling approach will recover by construction any exogenous set of market prices. Therefore, in carrying out the calibration we are implicitly making two assumptions:

1. that we know, from our knowledge of financial markets (as opposed to just from the market prices of options) that the true process *for the underlying* is indeed a local-volatility diffusion; and
2. that the market has fully incorporated this information in the price-making *of plain-vanilla options*.

In other words, by following this procedure we do not allow the possibility that the true process was a local-volatility diffusion, but that the market failed to incorporate this information in the prices of plain-vanilla options.

In reality option prices are not exogenous natural phenomena, nor are they made by omniscient demi-gods with supernatural knowledge of the 'true' processes for the stochastic state variables. Option prices are made by traders who, individually, might have little or no idea about the true stochastic process for the underlying; who might be using the popular 'model of the month'; who, for a variety of institutional constraints, might be afraid to use a model at odds with the current market practice (see Section 1.2.2); or who might be prevented from doing so by the limit structures in place at their trading houses.

A strong believer in market efficiency would counter that the errors of the individual uninformed traders do not matter, in that they will either cancel each other out (if uncorrelated), or will be eliminated (arbitraged away) by a superior and more unfettered trader with the best knowledge about the true process. I discuss the implications for option pricing of this strong form of market efficiency in the next sections, but this position must be squared with several empirical observations such as, for instance, the fact that steep equity smiles suddenly appeared after the equity market crash of 1987 – did traders not know before the event that the true process had a jump component? Another 'puzzling' fact for the believer of the fundamental view is that a close-to-perfect fit to the S&P500 smile in 2001 can be obtained with a cubic polynomial (see Ait-Sahalia (2002)): do we find it easier to believe that the market 'knows' about the true process for the index, prices options accordingly and when the prices obtained by using this procedure are converted into implied volatilities they magically lie on a cubic line? Or is it not simpler to speculate that traders quote prices of plain-vanilla options with a mixture of model use, trading views about the future volatility and cubic interpolation across strikes?

The matter cannot be settled with a couple of examples. For the moment I simply stress that the common and, prima facie, very sensible ('fundamental') approach to choosing and calibrating a model that I have just described can only be justified if one assumes that the observed prices of options reflect in an informationally efficient way everything

that can be known about the true process. I will return to this topic in Section 1.4, which deals with the topic of calibration.

1.1.3 The Instrumental Approach

An alternative ('instrumental') way to look at the choice of process for the underlying is to regard a given process specification as a tool not so much for driving the underlying, but for creating present and future prices *of option*s. In this approach it is therefore natural to compare how the prices of *options* (not of the underlying) move in reality and in the model. Typically this comparison will not be made in dollar terms, but using the 'implied volatility' language. Since, however, there is a one-to-one correspondence between implied volatilities and option prices (for a given value of the underlying) we can indifferently use either language.

So, traders, who might have no idea about the true process for the underlying, might none the less form ideas about how option prices (or, more likely, the associated implied volatilities) behave over time and in different market conditions: do term structure of volatilities remain roughly similar, or do they change shape in totally unpredictable ways? Do smiles for same-expiry options suddenly appear and then permanently flatten out as calendar time goes by, or do they approximately retain their relative steepness? Do swaption matrices always retain the same shape, or do they display a few fundamental 'modes' of deformation, among which they oscillate? How will the smile surface migrate as the underlying moves? Is it 'sticky' or 'floating'?

These traders, who observe regularities in the implied volatilities (i.e. in the prices of options) rather than in the underlying, tend to have relatively little interest in determining the true process for the 'stock price', and will instead prefer a process capable of producing the desired features in the implied volatilities. The process for the underlying now becomes more instrumental than fundamental: it is simply seen as a tool to obtain something else (i.e. the correct dynamics for the implied volatilities). Indeed, some of the most recent pricing approaches have tried to dispense with the specification of the process for the underlying altogether, and have directly prescribed the dynamics of the implied volatility surface. (See, for example, Schoenbucher (2000) and Samuel (2002).)[2]

Could the process for the underlying be chosen with total disregard of the true process for the underlying, as long as it reproduces the correct behaviour for the implied volatilities? This is unfortunately not the case, because, ultimately, the option trader will want, at the very least, to delta hedge her positions, and the success of the associated trading strategy will depend on the correct specification for the dynamics of the underlying. However, there is a considerable degree of 'robustness' in the hedging process, at least as long as certain important but rather broad features of the stock price process are captured correctly. See the discussions in Sections 4.6, 13.6 and 14.11. Furthermore, if the option prices and the process for the underlying were seriously incompatible, there would lie an obvious arbitrage somewhere, and the trader who totally disregarded the plausibility of the stock dynamics would theoretically expose herself to the risk of becoming a money machine for her fellow traders. In reality, however, real arbitrages are rather complex to put in place in practice, and the threat of being at the receiving end of an arbitrage strategy is often more theoretical than real. I discuss this topic further in Chapter 17 (see, in particular, Section 17.3, 'The Trader's Dream').

[2]The difficulties and dangers in doing so are discussed in Chapter 17.

Therefore the second (instrumental) approach is more popular, and, explicitly or implicitly, more widely followed in the market. It is also the conceptual framework that I prefer, and that I will predominantly follow in this book. I will try to justify this choice in this chapter, and, as concrete situations and examples arise, in the course of the book. It is important to keep in mind, however, that neither approach is without conceptual blemishes: the first requires an extreme faith in market efficiency that I consider unwarranted; for the second to work one has to rely on a rather difficult-to-quantify 'robustness' of the hedging process *vis-à-vis* trading restrictions and mis-specifications of the process for the underlying.

The discussion of what I mean by robustness constitutes one of the recurrent themes of this book. At this stage I can give a brief account as follows. I will show (Sections 4.5 and 4.6) that, as long as a certain quantity (the quadratic variation) is deterministic and known, it makes relatively little difference for the success of the hedging programme how the exact 'partitioning' of this quantity actually occurs during the life of the option. I will then go on to argue (Sections 13.6 and 14.11) that, if the quadratic variation is instead either unknown or stochastic, the success of the hedging strategy will rely to a large extent on finding a portfolio whose dependence on the (imperfectly known) realization of the quadratic variation is similar to that of the complex product that has to be hedged. This observation brings us naturally to the topic of vega hedging.

1.1.4 A Conundrum (or, 'What is Vega Hedging For?')

Suppose that a trader has a pricing model that describes the empirical statistical properties of the process of the underlying extremely well, but that recovers the prices of traded options poorly. (By the way, according to the Efficient Market Hypothesis this could not happen, but I postpone the discussion of this point until later sections.) The trader is also aware of another model, which implies a much less realistic process for the underlying, but which reproduces the present and expected future implied volatility surfaces very well. What should the trader do? Which model should she use?

The answer, I believe, is: 'It depends'.

If the trader is a plain-vanilla option trader, she should make use of her superior knowledge, and trade and set up dynamic delta-hedging strategies based on this knowledge of the process for the underlying. She might find it difficult to do so, because her true model will prescribe different amounts of stock to be delta-neutral than the model adopted by the market consensus. Therefore her positions will probably not appear delta-neutral to the risk management function of her institution (unless she can exercise an unhealthy influence on her middle office), and she might run against VaR or other limits. Also, since her back office will presumably use, for something as relatively liquid as plain-vanilla options, mark-to-market (rather than mark-to-model), she will be able to recognize little or no profit as soon as she puts on the advantageous trade, and will have to rely on the difference between the true and wrong option values to 'trickle in' during the life of the option, as her trading strategy unfolds. (See the discussion in Section 1.2.4.) Despite these constraints, however, a plain-vanilla trader will forfeit her competitive advantage if she slavishly follows the market in every price.

If, on the other hand, she is a complex-derivatives trader, vega hedging will be for her at least as important as delta hedging (see the discussion in Section 1.3.2 as to why delta hedging becomes relatively less important for complex-derivative traders – in a

nutshell, this is because of the high correlation between the errors in the delta of the complex product, and the errors in the delta of a well-chosen vega-hedging portfolio). It is therefore crucial for her that the future vega re-hedging costs predicted by the model should be as similar as possible to the actual costs encountered during the life of the option. These costs, in turn, will be linked to the future prices of plain-vanilla calls (the future smile surface) that the 'wrong' model recovers well by definition.

Why doesn't the complex-derivatives trader, who knows the true process for the stock price, dispense with vega hedging altogether, and simply engage in a correct delta-hedging strategy? She could only do this if the true process were such as to allow for market completeness by trading just in the underlying, i.e. if any payoff could be exactly reproduced by trading dynamically in the underlying. But this is a very special, and most unlikely, case. In general, market incompleteness is the rule, not the exception, and a trader, even armed with the knowledge of the true process, cannot hope that the future vega of the complex product *will only be a function of the future realization of the stock price*.

A final observation: successful option products, whether plain-vanilla or complex, thrive if there is a strong customer demand for them (customer, in this sense, means counterparties from outside the community of professional traders). Therefore option traders do not routinely make money by pitting their intellects against each other in a (zero-sum) war-game of pricing models. It is for them much more reliable and profitable to deal with the non-trading community, by providing the end users with the financial payoff they want (e.g. interest-rate protection, principal-protected products, yield 'enhancement', cheaper funding costs), and by exacting a compensation for the technological, intellectual and risk-management costs involved in providing this service. Given this trading reality, there are greater benefits in 'being on the market smile', even when it is felt that a more realistic model would not recover these market prices, than in standing alone at odds with the market.

1.2 The Efficient Market Hypothesis and Why It Matters for Option Pricing

I mentioned in the previous section that the Efficient Market Hypothesis (EMH) has a direct bearing on option pricing. In this section I discuss why this is the case. In order to do so it is important to clarify what is meant by market efficiency, and what conditions must be met for it to prevail. In particular, I will stress that rationality of each market player is *not* required for the EMH to hold true, and therefore criticisms of its validity must take a different, and subtler, route.

1.2.1 The Three Forms of the EMH

The EMH can be formulated in forms of wider and wider applicability (see, for example, the treatment in Shleifer (2000), on which this section draws extensively). The most radical form requires that all economic agents are fully informed and perfectly rational. If they can all observe the same history (of prices, economic variables, political events, etc.), then they will all arrive at the same statistical conclusions about the real world, and will form prices by discounting the expected future cash flows from a security at a discount rate dependent on the undiversifiable uncertainty of the security and on their risk aversion.

In this sense the value of the security is said to embed all the information available in the market, its value to be linked to 'fundamentals', and markets to be informationally efficient. All securities are fairly priced, excess returns over the riskless rate simply reflect 'fair' compensation for excess risk, and five-dollar banknotes cannot be found lying on the pavement.

A weaker form of market efficiency (but one that arrives at the same conclusions) does not require all economic agents to be rational, but allows for a set of investors who price securities on sentiment or with imperfect information. Will this affect the market price? Actually, one can show that as long as the actions of the uninformed, irrational investors are random ('uncorrelated'), their actions will cancel out and the market will clear at the same prices that would obtain if all the agents were perfectly rational.

Surely, however, the zero-correlation assumption is far too strong to be swallowed by anybody who has witnessed the recent dotcom mania. The very essence of bubbles, after all, is that the actions of uninformed, or sentiment-driven, investors are just the opposite of uncorrelated. If this is the case, then supply and demand, rather than fundamentals, will determine the price of a security. Is this the end of the efficient market hypothesis? Not quite. Let there be irrational and co-ordinated investors. As long as there also exist rational, well-informed agents who can value securities on the basis of fundamentals *and freely trade accordingly*, price anomalies will not persist. These pseudo-arbitrageurs will in fact buy the 'irrationally cheap' securities and sell the 'sentimentally expensive' ones, and by so doing will drive the price back to the fundamentals. Whether it is due to irrationality and sentiment or to any other cause, in this framework excess demand automatically creates extra supply, and vice versa. Therefore, as long as these pseudo-arbitrageurs can freely take their positions, supply and demand will not affect equilibrium prices, these will again be based on the suitably discounted expectation of their future cash flows (the 'fundamentals') and the EMH still rules.

It is important to stress that the EMH is not only intellectually pleasing, but has also been extensively tested and has emerged, by and large, vindicated. The 'by-and-large' qualifier, however, is crucial for my argument. In the multi-trillion market of all the traded securities, a theory that accounts for the prices of 99.9% of observed instruments can at the same time be splendidly successful, and yet leave up for grabs on the pavement enough five-dollar notes to make a meaningful difference to the year-end accounts and the share prices of many a financial institution. This is more likely to be so if the instruments in question are particularly opaque. The possibility that the pseudo-arbitrageurs might not always be able to bring prices in line with fundamentals should therefore be given serious consideration.

1.2.2 Pseudo-Arbitrageurs in Crisis

What can prevent pseudo-arbitrageurs from carrying out their task of bringing prices in line with fundamentals?

To begin with, these pseudo-arbitrageurs (hedge funds, relative-value traders, etc.) often take positions not with their own money, but as agents of investors or shareholders (Shleifer (2000)). If the product is complex, and thus so is the model necessary to arrive at its price, the ultimate owners of the funds at risk might lack the knowledge, expertise or inclination to assess the fair value, and will have to rely on their agent's judgement. This trust, however, will not be extended for too long a period of time, and certainly not

for many years. Therefore, the time-span over which securities are to revert to their fundamental value must be relatively short (and almost certainly will not extend beyond the next bonus date). If the supply-and-demand dynamics were such that the mis-priced instrument might move even more violently out of line with fundamentals, the position of the pseudo-arbitrageur will swing into the red, and the 'trust-me-I-am-a-pseudo-arbitrageur' line might rapidly lose its appeal with the investors and shareholders.

Another source of danger for relative-value traders is the existence of institutional and regulatory constraints that might force the liquidation of positions before they can be shown to be 'right': the EMH does not know about the existence of stop-loss limits, VaR limits, concentration limits, etc.

Poor liquidity, often compounded with the ability of the market to guess the position of a large relative-value player, also contributes to the difficulties of pseudo-arbitrageurs. Consider for instance the case of a pseudo-arbitrageur who, on the basis of a perfectly sound model, concluded that traded equity implied volatilities are implausibly high, and entered large short-volatility trades to exploit this anomaly. If the market became aware of these positions, and if, perhaps because of the institutional constraints mentioned above, the pseudo-arbitrageur had to try to unwind these short positions before they had come in the money, the latter could experience a very painful short squeeze.

Finally, very high information costs might act as a barrier to entry, or limit the number of pseudo-arbitrageurs. Reliable models require teams of quants to devise them, scores of programmers to implement them, powerful computers to run them and expensive data sources to validate them.[3] The perceived market inefficiency must therefore be sufficiently large not only to allow risk-adjusted exceptional profits after bid–offer spreads, but also to justify the initial investment.

In short, because of all of the above the life of the pseudo-arbitrageur can be, if not nasty, brutish and short, at least unpleasant, difficult and fraught with danger. As a result, even in the presence of a severe imbalance of supply or demand, relative-value traders might be more reluctant to step in and bring prices in line with fundamentals than the EMH assumes.

1.2.3 Model Risk for Traders and Risk Managers

I have mentioned the impact of risk management on trading practice and on price formation. It is useful to explore this angle further.[4]

Within the EMH framework the goals of traders and risk managers are aligned: a superior model will bring the trader a competitive advantage, and will be recognized as such by the market with very little time lag. From this point of view, an accurate mark-to-market is purely a reflection of the best information available, and 'true' (fundamental) value, market price and model price all coincide. It therefore makes perfect sense to have a single research centre, devoted to the study and implementation of the best model, which will serve the needs of the front-office trader, of the risk manager and of the product controller just as well. Looked at from this angle, model risk is simply the risk that our current model might not be good enough, and can be reduced by creating better and

[3] When I was heading an interest-rate-derivatives trading desk I was half puzzled and half embarrassed when I discovered that the harnessed power of the farm of parallel super-mini-computers in my small trading group ranked immediately after Los Alamos National Laboratory in computing power.

[4] The following two sections have been adapted from Rebonato (2003b).

better models that will track the monotonic, if not linear, improvement of the markets' informational efficiency.

If we believe, however, that pseudo-arbitrageurs might in practice be seriously hindered in bringing all prices in line with fundamentals the interplay between true value, market value and model value can be very different. Across both sides of the EMH divide there is little doubt that, when it comes to trading-book instruments, what should be recorded for books-and-records purposes should be the best guess for the price that a given product would fetch in the market. For the EMH sceptic, however, there is no guarantee that the 'best' available model (i.e. the model that most closely prices the instrument in line with fundamentals) should produce this price. Furthermore, there is no guarantee that the market, instead of swiftly recognizing the error of its ways, might not stray even more seriously away from fundamentals.

Ultimately, whenever a trader enters a position, he must believe that, in some sense, the market is 'wrong'. For the risk manager concerned about marking an option position to model appropriately, on the other hand, the market must be right by definition. For the EMH believer the market can only be wrong for a short period of time (if at all), so there is no real disconnect between the risk manager's price, and the trader's best price. For the EMH sceptic, on the other hand, there is an irreconcilable tension between front-office pricing and the risk-management model price.

1.2.4 The Parable of the Two Volatility Traders

To illustrate further the origin of this tension, let us analyse a stylized but instructive example. Two plain-vanilla option traders (one working for Efficient Bank, and the second for Sceptical Bank) have carefully analysed the volatility behaviour of a certain stock, and both concluded that its level should be centred around 20%. The stock is not particularly liquid, and the brokers' quotes, obtained with irregular frequency, do not deviate sufficiently from this estimate to warrant entering a trade. One day, however, without anything noticeable having happened in the market, an implied volatility quote of 10% appears. Two huge five-dollar notes are now lying on the floor, and both traders swiftly pick them up. Both traders intend to crystallize the value of the mis-priced option by engaging in gamma trading, i.e. by buying the 'cheap' option they will both end up long gamma and will dynamically hold a delta-neutralizing amount of stock, as dictated by their model calibrated to 20% volatility. (See Section 4.3 for a discussion of gamma trading.)

Life is easy for the Efficient Bank trader. She will go to her risk manager, convince him that the model used to estimate volatility is correct and argue that the informationally efficient market will soon adjust the implied volatility quote back to 20%. This has important implications. First of all, the profit from the trade can be booked immediately: the front office and risk management share the same state-of-the-art model, and both concur that the price for the option in line with fundamentals should be obtained with a 20% volatility. Furthermore, should another five-dollar bill appear on the pavement, the trader at Efficient Bank will have every reason and incentive to pick it up again.

The coincidence of the front-office and middle-office models has yet another consequence. The trader works on a 'volatility-arbitrage' desk, and her managers are happy for her to take a view on volatility, but not to take a substantial position in the underlying. They have therefore granted her very tight delta limits. This, however, creates no problem, because her strategy is to be delta-neutral at every point in time and to enjoy

the fact that she has bought (at 10%) 'cheap convexity' (see Chapter 4). Crucially, in order to crystallize the model profits from trade, she will engage in a dynamic hedging strategy based on the superior model (calibrated with a 20% volatility), not on the temporarily erroneous market model. Since middle office again shares the same model, the risk manager calculates the delta of the position exactly in the same way as the trader, and therefore sees the whole portfolio perfectly within the desk's delta limits (actually, fully delta neutral).

Life is much harder for the trader at Sceptical Bank. She also works on a volatility-arbitrage desk with tight delta limits, and her middle-office function also recognizes that the model she uses is sound and plausible and concurs that the market must be going through a phase of summer madness. The similarities, however, virtually end here. Her risk-management function does not believe that a superior model must be endorsed by the market with effectively no delay, and therefore is not prepared to recognize the model value implied by the 10% trade as an immediate profit. A model provision will be set aside. Since the trader will not be able to book (all) the model profit upfront, she will have to rely on the profit dripping into the position over the life of the option as a result of trading the gamma. This process will be relatively slow, the more so the longer the maturity of the option. During this period the trader is exposed to the risk that another 'rogue' volatility quote, say at 5%, might even create a negative mark-to-market for her position. Her reaction to a second five-dollar bill will therefore be considerably different from that of her colleague at Efficient Bank. Furthermore, in order to carry out her gamma-trading programme she would like to buy and sell delta amounts of stock based on her best estimate of the 'true' volatility (20%). Middle office, however, who have observed the 10% trade, uses the model calibrated with the lower volatility to calculate the delta exposure of the trade, and therefore does not regard her position as delta-neutral at all. She utilizes more VaR than her colleague, might soon hit against her delta limit, and, if her trading performance is measured on the basis of VaR utilization, she will be deemed to be doing, on a risk-adjusted basis, more poorly than her colleague.

This parable could be expanded further, but the central message is clear: different views about market efficiency can generate very different behaviours and incentives for otherwise identical traders. In the real world, however, financial institutions are organized much more along the lines of Sceptical Bank than of Efficient Bank, and this creates a strong disincentive for a trader to stray too much from the path of the commonly accepted pricing model.

The strength of this disincentive should not be underestimated. The story about Efficient Bank and Sceptical Bank might have been contrived and over-stylized, but a real-life example can bring home the same point with greater force and clarity. As a trader enters a complex derivative transaction for which no transparent market prices are available, the product control function of her institution faces the problem of how to ascribe a value to the trade. Commercial data providers exist to fulfil this need. One such major provider active in the United Kingdom and in the United States acts as follows: the prices of non-visible trades are collected *from the product control functions* of several participating institutions; the outliers are excluded and an average price is created from the remaining quotes; information is then fed back to the contributing banks about the average price and about how many standard deviations away from the consensus their original quote was. If the quotes submitted by an institution are consistently away from the consensus, *the institution is expelled from the contributing group, and will no longer see the consensus*

prices. When this happens, the product control function of that bank will no longer be able to discover the price of the opaque derivative, but simply knows that the bank's pricing model (even if it might be, perhaps, a 'better' one) is away from the industry consensus. A model reserve will have to be applied that will typically negate the extra value ascribed by the trader's model to the exotic product. Furthermore, since there is prima facie evidence that where the trader would like to mark her trades is away from the market consensus, a limit on the maximum notional size of the 'offending' positions is likely to be imposed.

Therefore, while today's models are indubitably more effective (at least in certain respects) than the early ones, I do not believe that the 'linear evolution' paradigm, possibly applicable to some areas of physics,[5] and according to which later models are 'better' in the sense of being closer to the 'phenomenon', is necessarily suited to describing the evolution of derivatives models. This real-life example shows that the disincentives against straying away from market consensus (ultimately, the withdrawal of market information) can be even more powerful in practice than in the parable of the two traders. Models can, and do, evolve, but in a less unfettered manner than traditional 'linear' accounts (and the EMH) assume.

Model inertia is therefore certainly a very significant feature to take into account when analysing models. There are however other aspects of market practice that have a profound influence on how models are developed, tested and used. Given their importance, some of these are discussed below.

1.3 Market Practice

1.3.1 Different Users of Derivatives Models

To understand derivatives models it is essential to grasp how they are put to use. In general, a pricing model can be of interest to plain-vanilla-option traders, to relative-value traders and to complex-derivatives traders. Relative-value and plain-vanilla traders are interested in models because of their ability to predict how option prices should move relative to the underlying, and relative to each other, given a certain move in the underlying. For both these classes of user, models should therefore have not just a *descriptive*, but also a *prescriptive* dimension.

The situation is different for complex-derivative traders, who do not have access to readily visible market prices for the structured products they trade in, and therefore require the models to 'create' these price, given the observable market inputs for the underlying and the plain-vanilla implied volatilities. Since complex traders will, in general, vega hedge their positions, exact recovery of the plain-vanilla hedging instruments – the descriptive aspect of a model – becomes paramount. The recovery of present and future option prices is linked to the current and future vega hedging and to model calibration. These very important practices are discussed in the next section.

[5] I realize that, with this statement, I am walking into a philosophical thicket, and that some philosophers of science would deny even models in fundamental physics any claim of being 'true' in an absolute sense. I stay clear of this controversy, and the more mundane points I make about the evolution of derivatives models remain valid irrespective of whether an absolute or a 'social' view of scientific progress is more valid.

1.3.2 In-Model and Out-of-Model Hedging

Possibly no aspect of derivatives trading has a deeper-reaching impact on pricing than the joint practices of out-of-model hedging and model recalibration. In-model hedging refers to the practice of hedging a complex option by taking positions in 'delta' amounts of traded instruments to neutralize the uncertainty from the *stochastic* drivers of the process for the underlying. In a Black-and-Scholes world, neutralizing the movements in an option price by buying a delta amount of stock is a classical example of in-model hedging.

Out-of-model hedging is the taking of positions to neutralize the sensitivity of a complex product to variations in input quantities *that the model assumes deterministic* (e.g. volatility). In a Black-and-Scholes world, vega hedging is a prime example of out-of-model hedging.

Needless to say, out-of-model hedging is on conceptually rather shaky ground: if the volatility is deterministic and perfectly known, as many models used to arrive at the price assume it to be, there would be no need to undertake vega hedging. Furthermore, calculating the vega statistics means estimating the dependence on changes in volatility of a price that has been arrived at assuming the self-same volatility to be both deterministic and perfectly known. Despite these logical problems, the adoption of out-of-model hedging in general, and of vega hedging in particular, is universal in the complex-derivatives trading community. The trader who engages in this logically dubious vega hedging *at inception of a trade* can at least console herself as follows. If her model has been correctly calibrated to the current market prices of the vega-hedging instruments, she will be adding to the original delta-neutral portfolio another self-financing, delta-neutral and fairly-valued portfolio of options. By so doing she will simply have exchanged part of her wealth from cash into stock and fairly-priced options, and this can have no impact on the value of the complex trade (because her model was correctly calibrated to the current market prices of the hedging instruments). But what about *future* vega re-hedging transactions? If, conditional on a future level for the underlying, these vega trades will be carried out in the real world at the same future prices that the model ascribes to them today, once again the trader has notionally just added lots of self-financing, delta-neutral and fairly-valued portfolios of forward-starting options. The economic effect of this is zero. This is no longer true, however, if the model predicts today future re-hedging costs different from what will be encountered in reality. If systematic, this difference between the real and theoretical level of future re-hedging transactions can make the whole strategy non-self-financing, and cause money to bleed in or (most likely) out of the trader's account.

Similarly important, universal and difficult to justify theoretically is the practice of re-calibrating a model to the current market plain-vanilla prices throughout the life of the complex trade. Let us look at this practice in some detail. Because of the need to vega hedge at the start of the life of a complex transaction, a trader will begin by calibrating her model in such a way as to recover the current (day-0) prices of all the options needed for hedging. Once the model has been calibrated in this manner, the price of a complex derivative will be calculated, and the trader will begin the dynamic hedging strategy to be carried out until the option expiry. Let us now move a few days into the trade. On day 2, the same model calibration used on day 0 will not in general produce spot (i.e. day-2) plain-vanilla option prices in line with the market. Therefore if the future re-hedging transactions were carried out with the model's parameter as per day-0's calibration, their model prices would not coincide with the market prices. To avoid this, the trader will re-calibrate the model on the basis of these new benchmark option prices, and re-calculate

the price of the complex instrument on the basis of the new calibration, *assumed again to be valid until the option expiry*. As an unrepentant sinner, therefore, every morning the trader who re-calibrates a model admits that yesterday's calibration (and price) had been wrong, yet makes a price today (with the new parameters) that rests on the assumption that the new calibration will be valid until the product's expiry.

These two practices are closely linked. In a friction-less market, if a model did not have to be re-calibrated during its life, future vega transactions would have no economic impact. If a model only needed to be calibrated once and for all, it would always imply the same future prices for plain-vanilla options in the same future attainable states of the world. Therefore, contingent on a particular realization of the stock price or rate, these trades would be transacted at the future conditional prices for the plain-vanilla hedging instruments implicit in the day-0 calibration. Exchanging in the future an amount of money equal to the fair value (according to the model) of the future plain-vanilla options required for re-hedging has no economic effect today, and, as a consequence, would not affect today's model price of the complex derivative. This is no longer true, however, if, in order to recover the future spot plain-vanilla prices, the model has to be re-calibrated day after day.

Clearly, no model will be able to predict exactly what the future re-hedging costs will be (even if the Black-and-Scholes approach assumes this to be possible). It is however important to use a model that, as much as possible, 'knows' about possible future re-hedging costs and assigns to them the correct (risk-adjusted) probabilities. Since the true cost of an option is linked (in a complete market, is equal) to the cost of the replicating portfolio, the trader will therefore have to keep two possible sources of cost in mind: the hedging costs incurred at inception, and those encountered during the life of the option. As for the initial costs, these come from in-model hedging (e.g. the cost of the delta amount of stock), and from out-of-model hedging. I have argued that the latter have, in theory and in the absence of bid–offer spreads, no economic effect today, since the trader who has correctly calibrated her model is simply buying at fair value a series of fairly priced options. Indeed, neglecting again bid–offer spreads, after booking these out-of-model hedging initial trades, the P&L account of the trader will display no change. For a small number of products (such as, for instance, European digital options) an initial portfolio is all that is needed to hedge exactly the 'complex' trade until expiry. See, for instance, the discussion in Chapter 17. For this type of trade recovery of today's plain-vanilla prices is all that matters, and the trader does not have to worry whether the future conditional option prices predicted by the model will be in line with reality or not. These pseudo-complex trades, however, are few, and, by and large, uninteresting variations on the plain-vanilla theme. For all bona fide complex trades some degree of in- and out-of-model vega *re*-hedging will always have to be undertaken. The more the future vega trades will be important, the more the trader will be sensitive to the correct prediction by the model of the future conditional plain-vanilla option prices, and the less exact recovery of today's prices becomes the only relevant criterion in assessing the quality of a model.

Choosing good inputs to a model therefore means recovering today's prices in such a way that tomorrow's volatilities and correlations, as predicted by the model, will produce future plain-vanilla option prices as similar as possible to what will be encountered in the market. Given the joint practices of vega re-hedging and model re-calibration, the 'best' calibration methodology is therefore the one that will require as little future re-estimation of the model parameters as possible.

Looking at the problem in this light, one of the most important questions is: How should the model inputs that will give rise to the more stable calibration and to the smallest re-hedging 'surprises' be estimated?[6] Answering this fundamental question requires choosing the source of information (statistical analysis or market 'implication') that can best serve the trading practice. This is therefore the topic of the next section.

1.4 The Calibration Debate

In principle, to calibrate a model in order to price complex derivatives, one could follow two distinct routes: one could prescribe the whole real-world dynamics for the driving factor(s) (e.g. the stock price or the short rate) and for the associated risk premia. Given these inputs, the equilibrium prices for *all* assets (the underlying *and* the derivatives) can be obtained. This approach is called 'absolute pricing'. Alternatively, one could assign the volatility and correlation functions (the covariance structure, for short) of the stochastic state variables. On the basis of this much more limited information, the prices of options *given the price of the underlying* can be obtained. This approach is called 'relative pricing'. Readers familiar with the Black-and-Scholes approach might find that the first (absolute) approach 'goes against the grain', since it fails to take advantage of the greatest strength of relative pricing, i.e. the irrelevance of the difficult-to-estimate real-world drift. Yet, in the interest-rate arena, estimation of the real-world dynamics of the driving factor (typically, the short rate) and a separate specification of the associated risk premium was the approach of choice for the first term-structure models. I discuss the evolution of these models in detail in Rebonato (2004), where I explain why this practice was abandoned in favour of working directly in the risk-neutral measure.

While both routes are in principle possible, these days for practical pricing purposes the relative-pricing route is almost universally adopted for derivatives. Therefore, the specification of a relative-pricing, arbitrage-free model in a complete-market setting has in current trading practice become tantamount to assigning the covariance structure among the state variables (or just the volatility, if only one stochastic variable describes the financial universe).

When this is combined with the market practices of out-of-model hedging and model re-calibration discussed in the previous section, it produces some important consequences, which have a direct bearing on calibration. This is because any choice of volatilities and correlations will determine the model-implied future conditional prices of the plain-vanilla options required to carry out the future re-hedging trades. As a consequence, the universal practices of re-calibrating the model and of re-balancing the vega hedges during the life of the complex trade require that the model should recover to a satisfactory degree the future conditional prices of the hedging instruments. The fundamental calibration question therefore becomes: 'What sources of information can most reliably provide an estimate of the covariance structure capable of producing these desirable future prices?' The answer to this question is as, if not more, important than choosing the 'best' model.

[6] Strictly speaking, this statement does not tell the full story. If markets were complete, then it would be possible to set up at inception strategies that would produce gains or losses that exactly offset the 'surprises' coming from the real-world realizations. This possibility is discussed in Section 1.4, where I argue that the complete-market hypothesis is a conceptually useful but often unrealistic idealization.

The question that we have posed is important and it is worthwhile pausing and sketching the logical itinerary followed to formulate it. The starting point is the predominance in current complex-derivatives pricing of the relative-pricing approach. In this approach the real-world drifts and risk premia become irrelevant (this is obvious for the Black-and-Scholes approach, but it is also true for interest-rate modelling). Therefore only the specification of volatility and correlation is required for no-arbitrage pricing (again, obviously for 'stock-like' problems, but, given the Heath-Jarrow-and-Morton (1987, 1989) insight discussed in Chapter 19, also true in a more complex form for interest rates). At the same time, volatilities and correlations determine the present and future smile surfaces. In turn, the future smile surfaces determine future re-hedging costs (this is because of the joint practices of model re-calibration and of vega re-hedging). Therefore the choice of volatilities and correlations will both give concrete form to the no-arbitrage pricing condition and 'predict' the future conditional re-hedging costs. I have also argued that the first desideratum of a model used for the relative pricing of complex products is that it should 'know' not only about the present, but also about the future (re)-hedging costs. Putting all the pieces together it therefore follows that the calibration procedure (i.e. choosing the volatilities and correlations that enter the model) should be carried out in such a way as to ensure that the chosen model will produce the future properties for the hedging options that we desire. This, incidentally, is the reason why this book is called *Volatility and Correlation*. In option pricing no decision is more important than how to choose these two quantities. Yet there is no universal agreement among practitioners or academics as to how volatilities and correlations should be estimated. The next section explains why this is the case.

1.4.1 Historical vs Implied Calibration

The estimation of volatilities and correlations can be arrived at either using historical estimation or via the implied route. With the historical approach the model inputs are determined on the basis of a statistical analysis of the time series of the relevant market quantities. With the implied approach, the input quantities are determined so as to recover the observed market prices of plain-vanilla options. When, for a given set of market prices, there is only one input function to determine, the 'implied' solution is unambiguous. This is the classic case of the Black-and-Scholes implied volatility. But, once this volatility has been estimated from the price of an option, can one use this implied quantity to price something else (e.g. a different-strike option, or an option with a different payoff)? And, what should one do when different combinations of input quantities can give rise to the same set of observable market prices? This is the case, for instance, of market swaption prices, which can be recovered with a variety of possible combinations of correlation and instantaneous volatility functions, as discussed in Chapters 19 and 20. If we are pricing a forward-rate-dependent complex product, should we use the forward-rate correlation 'implied' by the swaption prices, or the one estimated on the basis of statistical analysis? Should we rely on the FX/interest-rate correlation implied by the price of a quanto swap to price a power reverse dual swap, or should we use the available historical information? What use should we make of the information about the implied equity correlation that one can extract from the price of an equity index basket option?

In the context of derivatives pricing, both academics and practitioners have tended to embrace the implied route with far more enthusiasm than the statistical approach. It is common to find in the literature statements such as the following by Alexander (2003):

> ...correlation forecasts [are] difficult to obtain, particularly for short maturity forward rates. The forecasts have a great deal of uncertainty. This is one of the many reasons why we should seek to use market data rather than historical data for calibrating correlations....

Furthermore, the ability of a model to recover simultaneously as many 'market-implied' features as possible (e.g. implied instantaneous volatilities *and* correlations from the prices of caplets *and* swaptions), has generally been regarded as highly desirable. See, for instance, Schoenmakers and Coffey (2000), Brace and Womersley (2000), De Jong *et al.* (1999) and Marris and Lane (2002).[7]

What are the reasons for this preference? Under what assumptions can it be justified? Much as the implied route might appear 'natural' to a trading community trained in the footsteps of the Black-and-Scholes approach, it should not be taken as self-evident. This is the topic explored in the next section.

1.4.2 The Logical Underpinning of the Implied Approach

I will show in Chapter 3 that, in a classic Black-and-Scholes world, once the volatilities have been determined for two options with expiries T_1 and T_2 on the same underlying, for any other option whose payoff depends on the realization of the same volatility from time T_1 to time T_2, one can safely assume that this *future* volatility will be exactly as implied by the two spot volatilities estimated *today*. The precise value of this future volatility prevailing from time T_1 to time T_2 will be obtained in Chapter 3, but, for the purpose of the present discussion the relevant point is that, if unrestricted trading in two options with expiries T_1 and T_2 is possible, the trader does not have to rely on this market-implied guess to be correct, or even plausible. As long as she can trade in the two spot options freely, she can 'lock-in' this future value of the volatility. This practice is very similar to the ability to lock-in a future borrowing rate by trading in two pure discount bonds, and is reflected in the fact that the equilibrium value of a forward rate does not reflect an expectation of future rates, but is obtained from arbitrage considerations. See Rebonato (2002) for a careful discussion of this point.

The justification for this practice is to be found in the completeness of the relevant markets, which in turn stems from the assumed deterministic nature of the volatility function in the Black-and-Scholes world: even if the future volatility were to turn out to be different from what is implied by today's spot prices, by trading in the options available today we will be able to make enough money in each possible future state of the world to compensate us for any discrepancy between the implied and the actually realized value.

In the case of the same-currency correlation function mentioned above, I discuss in Sections 19.5 and 19.6 (see also Rebonato (2002)) that, if caplets, swaptions *and serial options* were liquidly traded, one would indeed find oneself in a situation of market

[7] Incidentally, this practice might have been partly motivated or encouraged by the fact that the LIBOR market model has just enough degrees of freedom to fit exactly, if one so wanted, all the caplet and European swaption prices.

completeness,[8] and one would be able to put in place strategies able to 'lock-in' any value of the correlation function implied by these joint prices, no matter how econometrically implausible. This is conceptually equivalent to the statement that, if two discount bonds, P_1 and P_2, maturing in one and two years' time trade in the market at prices $\exp[-y_1 T_1]$ and $\exp[-y_2 T_2]$, I will be able to synthetically borrow money for one year in one year's time at the rate $\exp[-(y_1 - 2y_2)] - 1$. In the interest-rate case, however, serial options are *not* liquidly traded, the market is *not* complete, and the future correlation *cannot* be locked-in with certainty by means of any strategy initiated at time 0. More generally, market incompleteness is the rule and not the exception, and arises when the volatility is stochastic,[9] when jumps are present, and, in general, when the payoffs from the available hedging instruments do not span all the possible future states of the world.

Does this mean that the implied approach is in general useless, and should be abandoned? Not necessarily. While market incompleteness might prevent the trader from locking-in the quantities of interest, yet, under certain conditions, the market-implied estimates might still convey useful information, namely the best collective guess produced by the market as to their value.

This statement must be treated with great care. In general, a market-implied quantity (say, the implied jump frequency in a jump–diffusion model) cannot be directly related to the corresponding real-world quantity, because of risk aversion: assume, for instance, that a set of market players are afraid of equity market crashes and cannot perfectly hedge against them. A different set of market players might provide 'insurance' to them by selling put options that would be priced assuming more frequent and more severe down jumps than observed in reality. This is discussed in detail in Smile Tale 1 in Section 6.4.

Yet, even if investors are risk averse, the market-implied quantity might still clear at the econometric level. This might happen if there is no net imbalance of supply and demand for an option from individually risk-averse investors. As I discuss in Section 6.6, in this case if a trade does take place it will do so at a level that does not incorporate risk aversion and simply reflects the unadjusted views 'of the market'.

Unfortunately, this state of 'natural' balance of demand and supply is in reality rather rare. Another condition can, however, apply, and has been alluded to before: the existence of pseudo-arbitrageurs who do not have a preferred trading habitat and who can take positions judging each trade on the basis of an appropriately discounted expectation of its future payoffs. If pseudo-arbitrageurs exist and can carry out their trades in as large a size as desired, excess demand creates its own supply, and vice versa. If this is the case, again supply and demand ultimately have no direct effect on price formation. In this setting prices do contain useful and direct (i.e. unpolluted-by-risk-aversion) information about the real world.

So far we have reached the conclusion that implying the values of financial quantities from market prices can be justified either if markets are complete, or if they are informationally efficient. Since we can safely rule out the first possibility, we should look carefully at the second.

[8] For market completeness to hold, one must also assume, of course, that the deterministic-volatility and deterministic-correlation assumptions hold exactly, that these quantities are perfectly known, and that the process for the forward rates has no jumps.

[9] It is often claimed that the market incompleteness arising from stochastic volatility can be easily exorcised by trading in another plain-vanilla option. While theoretically correct, I discuss the limitations of this solution in Chapter 13.

1.4.3 Are Derivatives Markets Informationally Efficient?

A large body of literature has appeared in the last 15 years or so, which challenges one of the pillars of the classical financial asset pricing, namely the EMH. The name generically applied to these rather disparate studies is that of 'behavioural finance'. In short, two joint claims are implicitly made (although not always explicitly articulated) by the proponents of this school, namely that (i) at least some investors arrive at decisions that are not informationally efficient and (ii) mechanisms that would allow better-informed traders to exploit and eliminate the results of these 'irrationalities' are not always effective. Since the prime mechanism to enforce efficiency is the ability to carry out (pseudo-)arbitrage, an important line of critique of the EMH has been developed (see, for example, Shleifer and Vishny (1997)) which shows that pseudo-arbitrage can in reality be very risky, and that, therefore, the pricing results of irrational decisions made on the basis of psychological features such as, say, over-confidence might persist over long periods of time.

In order to account for the *origin* of the pricing inefficiencies, the original emphasis was put on the psychological features of investors, such as, for instance, over-confidence, anchoring, framing effects, etc. (see, for example, Shefrin (2000) and Shiller (2000)), whence the name *behavioural* finance. Simplifying greatly, behavioural finance questions the assumption that market participants process new information as Bayesian agents, and claims that they maximize their utility function not over total wealth, but over gains and losses from a given reference point. The argument, based on the difficulty and riskiness of pseudo-arbitrage, can, however, still be applied if the price of an asset (and, in our case, of a derivative) is disconnected from 'fundamentals' for any (i.e. not necessarily for psychological) reasons: agency relationships, for instance, can give rise to discrepancies between the observed prices and those predicted by the EMH even if (and, actually, especially when) all the players are fully rational. This is important, because in the derivatives area (largely the arena of professional and sophisticated traders), it is more likely that institutional set-ups, rather than psychological biases, might be at the root of possible price deviations from fundamentals.

The relevance of these possible informational inefficiencies for derivatives pricing can be seen as follows. First, according to the EMH, prices are arrived at by discounting future expected payoffs using an appropriate discount factor.[10] The second step in the argument is that new information (a change in 'fundamentals') can lead an informed trader to reassess the current price for a derivative (a new expectation is produced by the new information), but supply and demand pressures *per se* cannot: if the 'fundamentals' have not changed, a demand-driven increase in the price of a substitutable security will immediately entice pseudo-arbitrageurs to short the 'irrationally expensive' security, and bring it back in line with fundamentals. The more two securities (or bundles of securities) are similar, the less undiversifiable risk will remain, and the more pseudo-arbitrageurs will be enticed to enter 'correcting' trades.

So, answering the question, 'To what extent should one make use of market-implied quantities as input to a model?' means addressing the two joint questions: 'Are there reasons to believe that a systematic imbalance of supply or demand might be present in

[10] 'Appropriate' in this context means, on the one hand, that it takes the riskiness of the cash flows into account, but, on the other, that it is only affected by non-diversifiable risk. So, if a security (an option) can be replicated by another (the hedging portfolio), then no idiosyncratic risk will be left and the appropriate discount factor is derived from riskless bonds.

the interest-rate plain-vanilla market?' and, if so, 'Are there reasons to believe that the activity of pseudo-arbitrageurs might entail substantial risks?'

Possible Mechanisms to Create a Supply/Demand Imbalance

The dynamics of the supply of and demand for interest-rate derivatives products are very complex, especially in US$, where the mortgage-backed securities market creates a large demand for a variety of derivatives products. If, to begin with, one focuses attention on the non-USD market, in broad terms some relatively simple patterns can be identified. On the one hand there are investors looking for 'yield enhancement' and issuers in search of 'advantageous' funding rates; on the other hand there are floating-rate borrowers who want to reduce their risk by purchasing interest-rate protection. In order to obtain the advantageous yields or funding rates, investors or issuers, respectively, tend to sell the right to call or put a bond, i.e. swaption-type optionality, which is typically 'sold-on' to investment houses. See Rebonato (1998a) for a more detailed description of these structures. Professional traders will therefore find themselves systematically *long* swaption optionality.

At the same time, floating-rate corporate borrowers will seek to limit their exposure to rising rates by purchasing caps from the same trading desks. In the non-USD markets, the latter will therefore find themselves systematically *long* swaption optionality and *short* caplet optionality.

In the USD market the picture is made much more complex by the presence of the Government-sponsored mortgage Agencies, who retain in their investment portfolios very large amounts of mortgage collateral. These mortgages expose the Agencies to pre-payment risk, and make them aggressive bidders of swaption volatility with expiries and underlying maturities dictated by the pre-payment speeds. The demand for swaption volatility is so high that the Agencies choose, or – some commentators claim – are forced, to complement their purchases of over-the-counter European and Bermudan swaptions with a funding programme largely based on callable debt (callable Agency debentures). It is important to point out that the demand for swaption optionality by the Agencies is localized in particular expiries and maturities, which do not necessarily coincide with the expiries and maturities of the swaption optionality that banks receive from investors and non-Agency issuers. As a result of this market dynamics significant supply/demand imbalances of different types of volatility persist, the complex-derivatives trader cannot act as a pure 'volatility broker' and has to warehouse and manage a substantial amount of 'volatility basis risk'.

Moving to the FX area, let us look at Figure 1.1, which shows time series of the implied volatilities for USD/JPY FX options of different expiries. Ignoring smile effects, which should anyway be rather limited for expiries of five or more years, we shall see in Chapter 3 that an increasing series of at-the-money implied volatilities requires that the future instantaneous (spot) volatility should also increase, and even more dramatically so, over time. Looking again at Figure 1.1, do we really believe that an informationally efficient market is conveying, via the implied volatility quotes, information about the expected future spot volatility? Today's implied volatility for options expiring in 10 years' time can be read from the graph to be around 15%. Yet the future 10-year average volatility embedded in today's 20- and 30-year option prices is approximately twice as large (and three times as large as today's spot volatility). Even a cursory examination

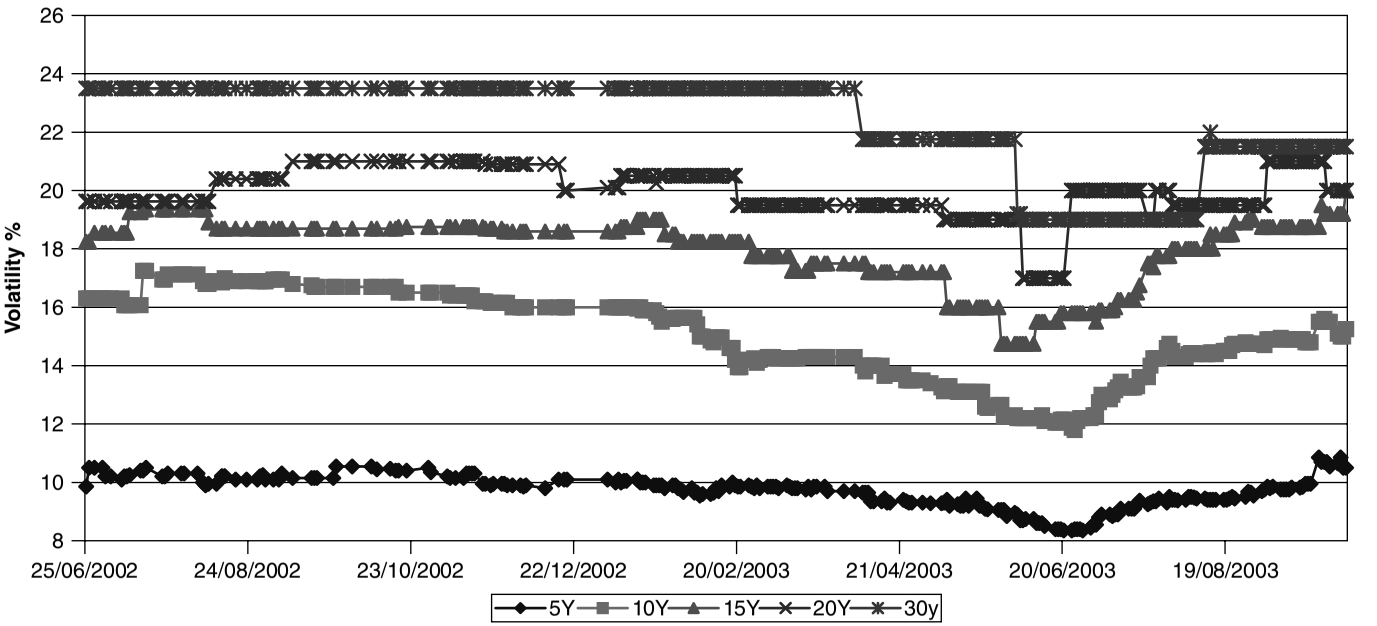

Figure 1.1 Time series of implied volatilities for USD/JPY options of expiry from 5 to 30 years.

of this figure appears to cast serious doubts over the proposition that the implied estimation of volatilities provides a reliable estimate of market expectations. Is there a more convincing explanation?

In order to understand the USD/JPY option-volatility case one should recall that in the first years of the twenty-first century very large volumes of yield-enhancing notes were marketed to JPY investors. (See Jeffrey (2003) for a story of the fortunes of this product.) Embedded in these notes were derivatives called power-reverse-dual swaps, which effectively conferred to the investor a series of very-long-dated callable FX options. The banks that provided these products therefore found themselves short of long-dated FX option volatility, and bid up the prices of the long-dated FX options required to hedge their exposure. These very-long-dated derivatives were virtually unheard of before the appearance of power reverse duals, which *de facto* created the market for this 'underlying'. No other economically motivated market flows were therefore in place to redress the demand imbalance. In my opinion, therefore, it is much simpler to explain the observed pattern in terms of a supply-and-demand effect, than of expectations of future behaviour of the spot volatility of the USD/JPY rate. But if this is correct, the implied at-the-money values do not provide any indication about future realization of the quantity of interest. In particular, unless we are dealing with a complete-market situation and the trader can 'lock-in' these values via static hedging strategies, these implied values are of no assistance in estimating the volatility levels at which future re-hedging trades will have to be carried out.

A different dynamics of supply and demand applies in the equity derivatives market. This market is largely driven by two sets of players: on the one hand, fund managers seeking to protect the value of their equity-invested portfolios; and on the other, especially in the United Kingdom and in continental Europe, retail investors receiving a deleveraged equity return in exchange for capital protection. Both sets of players are therefore net *buyers* of optionality, supplied by the derivatives desks of trading houses. Once again, the market dynamics appear to give rise to a net supply/demand imbalance.

A case can therefore be made as to why there might be systematic imbalances in the derivatives market: in the equity arena ultimately because of a desire for downside protection. In the FX case (at least in the example described above) because of the need to hedge the exposure created by large volumes of investor products. In the interest-rate area because of a more complex dynamics: issuers, investors and the Agencies create a systematic excess demand of cap volatility and excess supply or demand (depending on the exact location in the swaption matrix) of swaption volatility. Clearly, however, both caps and swaptions share the same underlyings (ultimately, forward rates) and should therefore in principle be 'cross-arbitrageable'.

Because of this state of affairs, pseudo-arbitrageurs should be enticed to take advantage of whatever move away from 'fundamentals' the supply/demand imbalance might create in the relative prices of caps and swaptions. Similarly, they should be tempted to sell the expensive put equity volatility and hedge their positions either with the underlying or with cheaper at-the-money options. Are there reasons to believe that the effectiveness of the pseudo-arbitrageurs to carry out these trades might in practice by hampered?

Possible Limitations to Pseudo-Arbitrageur Activity

What can prevent pseudo-arbitrageurs from carrying out their task of bringing prices in line with fundamentals? Without repeating the arguments presented earlier in this

chapter (see Section 1.2.2) I will recall the existence of agency relationships (hedge funds, relative-value traders, etc. often take positions not with their own money, but as agents of investors or shareholders); and the existence of institutional and regulatory constraints (stop-loss limits, VaR limits, size constraints, concentration limits, etc.) that might force the liquidation of positions before they can be shown to be 'right'. Furthermore, poor liquidity, often compounded with the ability of the market to guess the position of a large relative-value player, also contributes to the difficulties of pseudo-arbitrageurs. In this context, the role played by pseudo-arbitrageurs as ultimate providers of liquidity has been discussed by Shleifer (2000). Finally, very high information costs might act as a barrier to entry, or limit the number, of pseudo-arbitrageurs.

In short, because of all of the above, even in the presence of a significant imbalance of supply or demand, relative-value traders might be more reluctant to step in and bring prices in line with fundamentals than the EMH assumes.

Empirical Evidence

The literature covering empirical tests of market efficiency is far too large to survey even in a cursory manner (a recent count of papers in behavioural finance aimed at displaying failures of the EMH exceeded 2000), and, beyond the statement that markets are 'by and large' efficient, there has been a strong, and unfortunate, polarization of academic opinion. However, questions of more limited scope can be posed, such as: 'Is there any evidence that the mechanisms discussed above do hamper the activity of pseudo-arbitrageurs?', or 'Does the relative price of caplets and swaptions provide any indication of price deviations from fundamentals compatible with the supply/demand dynamics discussed in Section 1.2.1?'. Both questions are difficult to answer, the first because of the clearly secretive nature of the activities of pseudo-arbitrageurs (hedge funds and proprietary traders); the second, because showing that some prices are 'incorrect' always requires working under a joint hypothesis: what it tested is the deviation from a pricing model, given that the pricing model itself is correct. None the less, some pertinent observations *can* be made.

Starting from the question regarding the actual impact of the factors discussed above on the effectiveness of the pseudo-arbitrageurs, some indirect evidence can be obtained from reconstructions of the market events that surrounded the near-collapse of the LTCM hedge fund in 1998. (See Jorion (2000) and Das (2002) for a description of the events from a risk-management perspective.) Both Dunbar (2000) and Scholes (2000), although from different perspectives and drawing different conclusions, describe a situation where, for instance, the long-dated (5-year) equity implied volatility had reached in the autumn of 1998 levels that would imply for the next several years a realized volatility much higher than what had ever been observed in the past over similar periods. Yet traders (LTCM *in primis*) who attempted to short volatility found their positions moving farther into the red before the volatility finally declined. Many traders/arbitrageurs, including LTCM, faced with margin calls and with their request for additional 'arbitrage' capital from the (technically uninformed) investors turned down, had to cut the positions at a loss before the 'correctness' of their views could be proven. Similarly, swap spreads, which are but tenuously linked to the default risk of banks, reached during the same period levels difficult to reconcile with any plausibly risk-adjusted probability of bank default.[11]

[11] Given the role of pseudo-arbitrageurs as providers of liquidity alluded to above, Scholes (2000) argues that, in order to assess the 'fair' level of the swap spread, one should factor in the (time-varying) price for liquidity.

Finally, and most relevantly for the topic of this chapter, a reconstruction of events from market participants suggests that a major international investment house observed during the same period swaptions and caplet volatilities to move away from levels that most models could explain. In particular, for most plausible instantaneous volatility functions that recovered caplet prices, the 'implied correlation' was very different from correlations estimated statistically. The same house is widely thought to have put large (and 'correct'), swaption–caplet 'arbitrage' trades in place, only to have to unwind them at a loss as the positions temporarily moved even more strongly away from 'fundamentals'.

If the account given above is correct, and if therefore supply and demand cannot always be arbitraged away efficiently by proprietary traders, one would expect to observe some systematic effects. For instance, recall that investment houses tend to be long swaption optionality, and short caplet optionality. If this is the case, the market-implied forward-rate instantaneous volatilities estimated from swaption prices would be systematically lower than the same quantities estimated from caplet prices. Similarly, but less directly, the implied jump frequency required to recover the equity smile would have to be substantially higher than the corresponding statistically estimated quantity.

For reasons explained in Chapter 14, the equity 'evidence' is too indirect to lend itself to an unambiguous analysis, but we *can* say something about the caplet–swaption case. Rebonato (2002) displays graphs of the instantaneous volatilities of forward rates for several currencies estimated from the caplet and from the swaption markets.[12] The instantaneous volatility functions thus estimated turned out to have a very similar qualitative shape irrespective of the instruments (caplets or swaptions) used for their estimation, but to be systematically lower in all currencies when estimated from swaption data. This would be consistent with the supply-and-demand story for caplet and swaption optionality outlined above. Similarly, using different market data, Rebonato (2002) finds that the implied correlation required to price a set of co-terminal swaptions given the market prices of the caplets is much lower than what is historically observed. Again, swaptions would appear too 'cheap' relative to caplets, providing further indirect corroboration for the same 'story'.

It should be stressed that these results must be interpreted with care, because what is tested is the joint assumption that supply and demand skew the prices of a theoretically replicable set of securities *and* that the model used for the pricing (with the chosen parametric forms for the volatility and the correlation) is correct. Even with these caveats, however, the results appear to provide some corroboration for the hypothesis that supply/demand imbalances do affect the relative prices of caplets and swaptions.

1.4.4 Back to Calibration

This discussion brings us back to the calibration issue. The prevalent market practice, as evidenced by the references quoted above, seems to favour the 'implied' estimation approach. This would only make sense, however, if complex derivatives markets were

If this view is correct, there would be no market inefficiency at play. From the point of view of the calibration of a model, however, the existence of an important, and unhedgeable, risk factor (in this case liquidity) neglected in the pricing would raise similar concerns as to the appropriateness of using the market-implied estimate.

[12] A variety of correlation functions were used in the calibration to swaption prices. For econometrically plausible correlations the results displayed weak dependence on the level and details of the correlation functions. See also De Jong *et al*. (1999) on this point.

either complete or informationally efficient. The traded assets, however, certainly do not span all the possible states of the world (see, for example, the discussion in Section 19.5). And I have argued that there appear to be sufficient reasons to doubt the informational efficiency of the plain-vanilla instruments used for model calibration.

If these conditions are not met, then 'implying' values of financial quantities from option prices might indeed convey information, but probably more about the supply-and-demand dynamics than about intrinsic features of the underlying process. Therefore the generally accepted practice of fitting the free parameters of a model so as to recover the prices of as many plain-vanilla instruments as possible should be carefully questioned in each application. I have argued in this first chapter, and I will do so again at several points in this book, that a more relevant criterion for choosing these input functions should be their ability to recover in a plausible way current *and future* prices of the relevant re-hedging instruments.

1.4.5 A Practical Recommendation

I would like to offer a final recommendation to the trader who is tempted to 'imply' from market prices financial quantities (e.g. statistical properties of the process of the underlying from prices of the associated options). Whenever such a process of 'implication' is considered, the trader should ask the question: 'If the prices failed to reflect the best estimate of the quantity to be estimated, how easy would it be for a clever trader to exploit this error?' If the answer is that the pseudo-arbitrage is simple and relatively riskless, it is reasonable to expect that the market prices do convey useful information. If, instead, the 'correcting trades' are, for any reason, difficult to put in place or to exploit, then the trader should be much more cautious before relying on the 'best market estimate'.

Let me give a concrete example, related to the quote by Alexander (2000) reported above. Let us suppose that, perhaps for reasons of supply and demand, the correlation among forward rates implied by caplet and swaption prices was found to be out of line with what statistical analysis suggests plausible. How easy would it be to exploit this 'market error'? How can the trader put in place close-to-riskless pseudo-arbitrage strategies? More concretely: how would a trader deal with the gamma mismatch of a vega-neutral portfolio of swaptions and caplets? How can a trader hedge a product (the swaption) sensitive to most of the important modes of deformation of the swaption matrix (see Chapter 26) with products (caplets) which are only affected by a much more restricted set of eigenmodes? If the trader can find a satisfactory answer to these questions (I cannot), she can be reasonably confident that the market-implied correlation will convey useful information, *usable for trades other than caplet–swaption strategies*. If not, I believe that relying on the efficiency of the market as an information-processing machine is unwarranted.

1.5 Across-Markets Comparison of Pricing and Modelling Practices

In the mortgage-backed-securities (MBS) area pre-payment models are coupled with interest-rate models in order to produce the present value of the expected cash flows arising from a pool of mortgages (see, for example, Fabozzi (2001) or Hayre (2001) for a description of the market and of the prevalent pricing approaches). As a first stage in

arriving at the price for, say, a pass-through, the cash flows (including pre-payments) are discounted at the riskless (LIBOR) rate. The implicit assumption in doing so is that the hedge ratios suggested not only by the interest-rate model, but also by the pre-payment model, should provide a perfect hedge for the cash-flow uncertainty. From the modelling point of view, this is perfectly justifiable, because the vast majority of pre-payment models use only interest rates as state variables, and therefore allow for theoretically perfect hedging of the pre-payment risk by trading in the interest-rate-sensitive underlying products (swaps, caps, swaptions, indexed-principal swaps, spread locks, etc.). However, it has always been recognized in the MBS market that other variables (such as unemployment, GDP growth, etc.) strongly affect pre-payments, and that these variables are very imperfectly correlated with the interest-rate state variables. Whatever the models might claim, it has therefore always been very clear that hedging against these sources of uncertainty is in practice very difficult.

Because of this, the concept of the option-adjusted spread (OAS) has been introduced. The OAS is defined to be the spread to be added to the risk-less (LIBOR-derived) forward rates in order to obtain the discount factors to present-value the expected cash flows. A non-zero OAS therefore explicitly adjusts the price for all the undiversifiable (unhedgeable) sources of risk, for model uncertainty, for liquidity effects, etc. It is by no means a second-order effect, since, especially in periods of great pre-payment uncertainty (e.g. during the unprecedented wave of mortgage refinancing of 2002, when pre-payment models were constantly 're-calibrated' by trading houses, and the coupon on every outstanding issue was higher than the current par coupon) it reached values well over 100 basis points.

The introduction of the OAS links in an interesting way derivatives pricing when perfect replication is possible with more classic asset-pricing techniques. When pricing is absolute and not relative, i.e. when we are pricing primitive and not derivatives securities, the standard prescription, in fact, is to discount future uncertain cash flows using a rate 'appropriate to the riskiness of the cash flows'. Adding an OAS to the riskless forward rates used to discount the cash flows that the pricing model assumes to be riskless (because replicable) directly introduces an explicit recognition of the real uncertainty associated even with the best hedging strategy.

Why has the equivalent of an OAS not been developed in the derivatives area? Apart from issues of product homogeneity, liquidity and standardization, I believe that an important reason has been the different 'starting points' for the two markets. Even the first MBSs (pass-throughs) have always been perceived as being patently complex. This was due both to the difficulty in estimating the dependence of pre-payments on interest rates (the theoretically perfectly hedgeable part), and because of the inherent difficulty in hedging the non-interest-rate-related risk factors ('media effect', housing mobility, etc.). The appearance of more complex products (IOs, POs, sequentials, PACs, etc.) therefore simply added to an existing substantial modelling complexity, and stressed the relatively poor ability to hedge. Assuming perfect replication, in other terms, was never a realistic working hypothesis.

First-generation derivatives products (such as caplets, simple stock options, etc.), on the other hand, were relatively simple to hedge effectively, and, given the well-known robustness of the Black-and-Scholes model *vis-à-vis* reasonable mis-specification of the input volatility, payoff replicability (which ultimately justifies the risk-less discounting) was a very reasonable working assumption. As new products have been introduced, each

incremental increase in complexity has not been so big as to require a totally new and fresh pricing approach. The cumulative effect of this process, however, has been to give rise to products of considerable complexity: some of the instruments that received quite a lot of (unwanted) attention from Warren Buffet in 2002 (such as power-reverse-dual swaps) require the simultaneous modelling of compound optionality arising from the evolution over 30 years or more of two yield curves, of their volatilities and correlations, of the correlations among the forward rates of the two currencies, of the spot FX rate, and of its correlation with the interest forward rates. Most remarkably, one of the 'underlying' instruments behind power-reverse-dual swaps (extremely long-dated FX options) was literally created because of the introduction of the more complex product. At the same time, parallel pricing developments in related areas (credit derivatives, and *n*th-to-default swaps in particular) have brought about similarly difficult modelling challenges. One can therefore argue that these products have become no simpler, and their payoff replication not any easier, than the first mortgage-backed pass-throughs. None the less, due to the *incremental* process of adding relatively small elements of added complexity, no equivalent of the OAS has been introduced in the pricing of these assets, and the paradigm of risk-neutral valuation still reigns supreme. Model reserves are sometimes applied when recognizing the book value of these products, but this has not affected the 'mid' marking to model. The reasons for this, I believe, can be traced to the power of a robust and elegant conceptual framework (the Black-and-Scholes replication insight) and the self-sustaining nature of the 'inertial momentum' that a successful modelling framework generates.

If this analysis is correct the implications for derivatives pricing are not that the approaches described in the rest of this book are of little use: even in the MBS arena state-of-the-art models are continuously refined and developed for the diversifiable risk factors, and the interest-rate models of choice have closely followed the evolution of the (perfect-replication-based) LIBOR market model. What is required, I believe, is a re-assessment of the limitations of the pure-replication-based pricing philosophy, and the introduction in the price-making process of explicit recognition of the existence of substantial unhedgeable components. Because of the unavoidable presence of market imperfections, I will argue in this book that the *qualitative*, 'digital' distinction between complete and incomplete markets, or between replicable or non-replicable payoffs is not the most important characterization of a market or of a set of products. The *quantitative* differences in degrees of replicability are, in my opinion, more important and more relevant to the practice and to the theory of pricing.

Perhaps the equivalent of a 'LIBOR-OAS' could be arrived at in a coherent and theoretically robust manner by following one of the approaches (see, for example the 'no-to-good-deal' approach by Cochrane and Saa-Requejo (2000)) recently introduced in the literature to account for this very state of affairs. I can appropriately close this section by quoting Cochrane (2001):

> Holding [an] option entails some risk, and the value of that option depends on the 'market price' of that risk – the covariance of the risk with an appropriate discount factor. Nonetheless we would like not to [...] go back to 'absolute' methods that try to price all assets. We can [...] still form an approximate hedge based on [...] a portfolio of basis assets 'closest to' the focus payoff. [...]. Then the uncertainty about the option value is reduced only to figuring out the price of the residual.

1.6 Using Models

A few more comments about the use of models are in order before closing this 'foundation' chapter. I will be spending a lot of time discussing models whose conceptual foundation rests on the idea of perfect payoff replicability. At the same time I will argue that the conditions for applicability of these results are never met in practice, and that noticeable 'violations' appear even in very 'benign' settings. Am I being inconsistent, or, worse, am I wasting the reader's time?

I don't think so. Models, *qua* models, are always 'wrong' in the sense that they must leave out some features of the phenomenon they attempt to explain. Recognizing that market frictions 'spoil', to some extent, the Black-and-Scholes results is no different than observing that mechanical friction spoils the Newtonian result that a free ball will roll forever on an ideally smooth surface at constant velocity. Aristotelian physics, by the way, seems to produce, in this case, an answer more similar to 'reality', in that it postulates the need for an engine to keep the ball rolling at constant speed. Yet we find Newtonian mechanics more useful than Aristotelian mechanics for most problems. If we begin to consider objects moving at very high speed, special relativity gives better predictions. If we are in a rapidly varying gravitational field we will have to invoke general relativity. Every 'model' has a domain of applicability beyond which it ceases to be useful. The skill of the researcher is to gauge up to what point a certain modelling framework can be used, and to search for a more complex explanation as soon as, but no sooner than, it ceases to produce useful outputs.

So, yes, we will spend a lot of time looking at models based on perfect replication, and, yes, perfect replication is never possible in practice. There is however no contradiction in the approach, and it would be foolish to discard completely the insight and the power of the replication approach. Much as Newtonian dynamics is contained as a limiting case in special relativity, unique pricing by no-arbitrage can be seen as a limiting case (of vanishing variance for the stochastic discount factor) of Cochrane's no-good-deal approach.