SOUTH AFRICAN ACTUARIAL JOURNAL
SAAJ **14** (2014) 19–72
Pre-print publication online

ESTIMATING LONG-TERM VOLATILITY PARAMETERS FOR MARKET-CONSISTENT MODELS

By EJ Flint, ER Ochse and DA Polakow

Submission received 27 May 2013
Accepted for publication 10 June 2014

ABSTRACT
Contemporary actuarial and accounting practices (APN 110 in the South African context) require the use of market-consistent models for the valuation of embedded investment derivatives. These models have to be calibrated with accurate and up-to-date market data. Arguably, the most important variable in the valuation of embedded equity derivatives is implied volatility. However, accurate long-term volatility estimation is difficult because of a general lack of tradable, liquid medium- and long-term derivative instruments, be they exchange-traded or over the counter. In South Africa, given the relatively short-term nature of the local derivatives market, this is of particular concern. This paper attempts to address this concern by:

—providing a comprehensive, critical evaluation of the long-term volatility models most commonly used in practice, encompassing simple historical volatility estimation and econometric, deterministic and stochastic volatility models; and
—introducing several fairly recent nonparametric alternative methods for estimating long-term volatility, namely break-even volatility and canonical option valuation.

The authors apply these various models and methodologies to South African market data, thus providing practical, long-term volatility estimates under each modelling framework whilst accounting for real-world difficulties and constraints. In so doing, they identify those models and methodologies they consider to be most suited to long-term volatility estimation and propose best estimation practices within each identified area. Thus, while application is restricted to the South African market, the general discussion, as well as the suggestion of best practice, in each of the evaluated modelling areas remains relevant for all long-term volatility estimation.

KEYWORDS
Long-term volatility modelling; market-consistent valuation, historical volatility, deterministic volatility models, GARCH, stochastic volatility, break-even volatility, canonical valuation

CONTACT DETAILS
Mr Emlyn Flint, Peregrine Securities, 4th floor, Montclare Place, 21 Main Road, Claremont, Cape Town, 7708; Tel: +27 (0)11 722-7556; E-mail: emlynf@peregrine.co.za
Edru Ochse, Peregrine Securities, Montclare Place, 21 Main Road, Claremont, Cape Town, 7708
Daniel Polakow, School of Actuarial Sciences, University of Cape Town and the African Collaboration for Quantitative Finance and Risk Research (ACQuFRR)

1. INTRODUCTION

1.1 Since the inception of modern asset pricing models, starting as far back as Bachelier (1900), there has been considerable interest in volatility research. The body of literature on financial volatility is vast and encompasses a wide range of fields, both financial and other. However, there is a noticeable dearth of research on the forecasting and analysis of long-term volatility. This is largely because 'long-term' in the general field of equity volatility research refers to terms of one or two years. This is in contrast to the actuarial convention of 'long-term' meaning greater than 10 or 15 years. One struggles to find mention of long-term volatility estimation—let alone theoretical or empirical analysis of the same—outside of the literature on market-consistent valuation. Given the large quantity of life policies written with embedded investment derivatives as well as the current proclivity of many long-term insurers to continue to write similar policies, this should be a material concern for market-consistent valuation. Yet, even within this field, only a handful of academic papers and professional reports address this issue, most of these somewhat obliquely.

1.2 Current legislative and advisory practice notes (APN)[1] recommend the use of market-consistent models to set financial reserves for all embedded investment derivatives. 'Market-consistent' in this case refers to any model that "reproduces the market prices of tradable assets as closely as possible".[2] Whilst market-consistent models can take several different forms, without exception they all require a volatility surface defined across strike and term as an input. This is an acute problem given that the term of the embedded investment derivatives is usually far longer than any traded derivative contract. APN 110[3] makes allowance for this in the following manner:

1 Actuarial Society of South Africa. APN 110: Allowance for Embedded Investment Derivatives, Version 4. Advisory Practice Note, 2012. Actuarial Society of South Africa. Market Consistent Calibration in South Africa. APN 110 sub-committee presentation, 2010
2 APN 110 (2012): p.2
3 supra: p.3

> Where there are no traded market instruments from which to calibrate the market-consistent model, the actuary may apply alternative methods and judgement provided that he/she can argue that such derived values used to calibrate the model are probable in the market.

1.3 The situation outlined above typifies the current South African derivative market for any term beyond two or three years. Thus the above allowance actually provides a large element of subjectivity in market-consistent long-term volatility estimation. Figure 1 displays exactly how much subjectivity is allowed by giving a number of constructed implied volatility term structures that would all be considered market-consistent as per APN 110. The methods used to construct the respective volatility curves are indicative of those presently used in practice and are discussed in the ensuing sections. Clearly, the differences between the curves are substantial.

1.4 A 2010 survey of several long-term insurers conducted by the APN 110 sub-committee showed that of all market variables used in economic scenario generators, the highest relative importance was given equally to implied volatility on equity indices and the term structure of nominal interest rates. Implied interest-rate volatility and asset-class correlations were also shown to be of secondary importance. Although this paper focuses largely on estimation of implied volatility on equity indices, the ideas outlined below are, in certain cases, directly applicable to each of the variables highlighted above.

1.5 The contents of this paper are arranged as follows. Section 2 outlines the South African market data used in the analyses. Given the empirical nature of the paper and the long-term focus of the estimation, the particular choice and subsequent handling of data are of particular importance. Sections 3 to 6 provide a comprehensive critical review of those long-term volatility models most commonly used in practice:

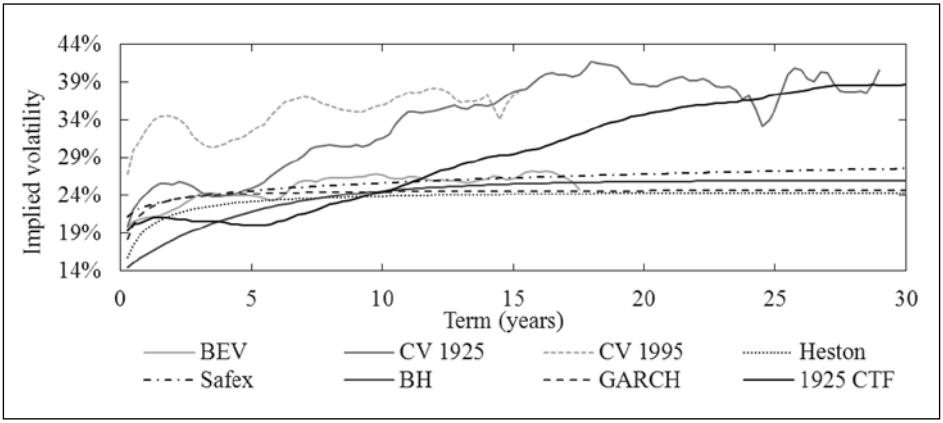

Figure 1. Long-term volatility estimates under a variety of market-consistent methods

- Section 3 reviews the estimation of historical and realised volatility, which can be used either directly to create a pseudo-implied volatility surface or as a means of creating a long-term volatility parameter for stochastic or deterministic volatility models.
- Section 4 assesses the use of econometric volatility models, specifically focusing on the GARCH family of models. The choice of different model specifications and innovation, or error, distributions is considered.
- Deterministic volatility models are outlined in Section 5, with a specific focus on the formulations suggested by the South African Futures Exchange (Safex) and Barrie & Hibbert.[4]
- Section 6 discusses the use of stochastic models for long-term volatility estimation. This includes both practical application—using the Heston (1993) model—and theoretical discussion.

Sections 7 and 8 introduce two fairly recent, compelling nonparametric alternatives for creating market-consistent long-term volatility estimates:

- Section 7 considers Dupire's[5] break-even volatility, which uses only historical data to calculate an implied volatility surface that makes delta-hedging a zero-sum game. Theoretical issues are discussed and practical application is given.
- Section 8 presents the nonparametric method of Stutzer's (1996) canonical valuation (CV) and constructs implied volatility surfaces via relative entropy and risk-neutralised historical distributions. This method has gained recent attention both locally and internationally because of its algorithmic tractability, financial flexibility—in terms of asset class and number of underlying assets—and solid statistical and economics foundation. Given the originality of the extended CV method presented here and the fact that many readers will be unfamiliar with the initial method, a significant portion of the nonparametric part of the paper is used to develop the ideas surrounding this nonparametric pricing method and several practical applications are given.

Section 9 concludes with a suggestion of best practices for long-term volatility estimation.

1.6 Because of the scope of the models, techniques and ideas discussed below, the technical detail inherent in each is inevitably condensed. However, in all cases the authors have endeavoured to provide the reader with suitable reference material so as to ensure accurate replication of all reported results. Two major works, which independently cover many of the volatility subfields reviewed in sections 3 to 6 and are worthy of initial citation, are Alexander (2008a; b; c; d) and Andersen et al. (2006).

4 See A Kotzé & A Joseph (unpublished). Constructing a South African Index Volatility Surface from Exchange Traded Data, JSE Technical Report, 2009, and D Roseburgh & C Holmes (unpublished). MC calibration to SA equity market. Barrie & Hibbert Calibration Note, 2006, respectively.

5 B Dupire (unpublished a). Fair skew: break-even volatility surface. Bloomberg LP White Paper, 2006

1.7 Alexander's (op. cit.) four-volume set is vast, rigorous and particularly practicable, and is considered a fundamental work in the greater risk-management literature. Andersen et al. (2006) provide a comprehensive survey of the most important theoretical and empirical literature in the field of volatility research, focusing specifically on forecasting. For further technical information on any specific implementation given here, the reader is welcome to contact the authors.

2. SOUTH AFRICAN MARKET DATA

2.1 EQUITY DATA

2.1.1 The various analyses in the paper make use of a number of different equity time series. All analyses are based on either the FTSE/JSE All-Share Index or the FTSE/JSE All-Share Top40 Index, referred to below as the ALSI and Top40 respectively. For long-term empirical analyses, the authors make use of Firer & Macleod's (1999) and Firer & Staunton's (2002) ALSI total monthly equity return series from January 1925 to February 2013, a total of 1 058 observations. The data are changed to capital returns using the assumption that linear returns are a linear sum of capital returns and (linear) dividend yield. The dividend yield is not readily available before January 1976, so the authors use the average yield as a simple proxy for the period 1925 to 1976. Total monthly ALSI returns are available from INet back to January 1976, totalling 446 observations. Accurate capital returns can be calculated for this period by using the reported monthly dividend yield.

2.1.2 Daily price and dividend data for the ALSI and Top40—capital and total-return indices—for the period 30 June 1995 to 28 March 2013—4 436 observations—are collected from INet. Intra-day Top40 data including opening, high, low and closing prices are available from 13 May 2002 onwards.

2.1.3 The dataset used in the analyses usually refers to the starting year, sampling frequency and underlying index unless the dataset choice is clear from the context or the specific analysis is found to be robust to the choice of dataset.

2.2 INTEREST-RATE DATA

2.2.1 A number of different data sources were amalgamated to construct 30-year yield curves back to January 1925. Firer & Macleod (1999) give a single annual interest rate, which is used for the period January 1925 to January 1965. Subsequently, basic yield curves were constructed using the three-month Treasury Bill rate and the Firer & Macleod rate and Hagan & West's (2008) raw interpolation method. The three-, six-, nine- and twelve-month negotiable certificate of deposit (NCD) rates are introduced in the raw interpolation method from January 1987, whilst the rand overnight deposit (RODI) rate is included from January 1999 onwards. Though simplistic, all instantaneous forward rates produced by this method are positive by construction, thus ensuring an arbitrage-free yield curve.

2.2.2 Comprehensive daily 30-year yield curves were provided by Old Mutual Specialised Finance for the period 28 August 2006 to 28 March 2013.

2.2.3 The term-specific yield-curve data used in the analyses below correspond to the period and frequency of the equity data outlined above.

3. HISTORICAL AND REALISED VOLATILITY

3.1 HISTORICAL VOLATILITY AND MARKET-CONSISTENT VALUATION

3.1.1 The estimation and measurement of empirical asset volatility is of central importance in most areas of finance. In recent years, there have been a number of significant improvements on the classic statistical methods used to measure an asset's return variation over time. Consequently the subfield of historical volatility measurement has blossomed. For an overview of this area of research, see Brandt & Kinlay,[6] Poon (2005) and Andersen et al. (2006).

3.1.2 APN 110 suggests the use of historical volatility analysis for estimating the most appropriate long-term volatility parameter to be used in a particular stochastic volatility model in the case where traded derivatives are not available. As noted above, in the South African market this essentially refers to any volatility estimate for a term greater than two to five years, allowing for direct bank-quoted prices. As an example, APN 110 suggests estimating term-specific realised volatility over some suitable period, comparing this estimate to the available implied volatility term structure and finally extrapolating this relationship to determine a suitable long-term stochastic volatility model parameter. According to the 2010 APN 110 survey, this type of estimation framework is used by all market participants. Therefore, the accurate measurement of historical volatility and its relation to the available implied volatility term structure is of particular importance and worthy of discussion.

3.1.3 Most textbooks define historical volatility as the standard deviation of past asset returns (Hull, 2009; Alexander, 2008a). However, this definition is naïve, leaving much unsaid. A better definition of historical volatility would be: the *ex-post* variation of an asset's returns taken at a particular frequency over a particular period. This succinctly connects the three fundamental variables latent in any volatility calculation:

- the specific functional form of the measured variation;
- the term of the asset returns; and
- the total period used for the estimation.

Each of these points is dealt with below. Moreover, the effect of underlying asset choice is also considered. This becomes an issue both when one is considering an index and when one is using the *ex-post* historical volatility estimate as an estimate of the *ex-ante* future realised volatility. Finally, the relationship between historical and implied volatility is considered.

3.2 MEASURING HISTORICAL VOLATILITY

3.2.1 Statistical Volatility Measures

3.2.1.1 The classic measure of historical volatility at time T, $\sigma_{C,T}$, is given by the common statistical standard deviation of n daily asset returns, $r_{t,1} = \ln\left(S_t^{\text{close}} / S_{t-1}^{\text{close}}\right)$,

6 MW Brandt & J Kinlay (unpublished). Estimating historical volatility. Investment Analytics Working Paper, 2005

with returns calculated from daily asset close prices at time t, S_t^{close}. Mathematically, one writes:

$$\sigma_{C,T} = \sqrt{\frac{252}{T-1}\sum_{t=1}^{T}\left(r_{t,1} - \left(\sum\frac{r_{t,1}}{T}\right)\right)^2}. \quad (1)$$

Financial convention is to quote volatility as an annualised standard deviation, where it is usually assumed that there are 252 business days a year. Note that $\sigma_{C,T}$ is always an *ex-post* measure. An occasional alternative to equation (1) is to assume that asset prices follow geometric Brownian motion and thus substitute the sample mean with the risk-neutral drift (Dupire[7]). This generally tends to increase the estimated volatility.

3.2.1.2 More commonly though, practitioners tend to remove the mean term altogether when calculating daily volatility. This is referred to as 'realised volatility', $\sigma_{R,T}$, and is calculated as:

$$\sigma_{R,T} = \sqrt{\frac{252}{T-1}\sum_{t=1}^{T} r_t^2}. \quad (2)$$

Realised volatility is the underlying asset for all traded variance and volatility derivatives and is an *ex-post* estimate of the asset return volatility over a particular period. Thus, while one can speak generally of historical volatility, one must be cognisant of the subtle differences between historical classical volatility as per equation (1) and historical realised volatility as per equation (2).

3.2.1.3 With the advent of high-frequency analysis, there has been a further classification of realised volatility. Specifically, if one assumes that the intra-day logarithmic-asset prices follow a general, continuous-time diffusion process, then, as Andersen et al. (2003) and Barndorff-Nielsen & Sheppard (2002) showed, the intra-day returns are normally distributed with mean and variance equal to the integral of the mean and variance process respectively over a continuous trading day. Andersen et al. (2003) showed that a consistent estimator for this 'integrated variance', $\sigma_{I,T}^2$, was given by the sum of squared intra-day logarithmic ("log") returns over the specified period. Integrated variance has since become the accepted standard for most accurately measuring empirical asset-return volatility. Because of this, integrated volatility is also sometimes referred to as 'realised volatility'. Whilst $\sigma_{I,T}^2$ is universally recognised as the best empirical estimate of asset-return volatility, intra-day asset tick data are readily available only for fairly short-term periods, thus limiting its current usefulness to the problem at hand. Therefore, for the remainder of this paper, realised volatility is exclusively defined by equation (2).

3.2.1.4 Over the last 30 years, a number of alternative, range-based volatility estimators have been put forward. These estimators use a combination of daily opening and closing prices together with intra-day high and low prices, and have been shown to have a much higher theoretical and empirical efficiency and thus lower bias than the common

7 B Dupire (unpublished b). Pricing Financial Derivatives. Bloomberg LP AFDC Presentation, 2006

standard-deviation estimator (Yang & Zhang, 2000; Brandt & Kinlay[8]). Although not in common use, those estimators developed by Parkinson (1980), Garman & Klass (1980), Rogers & Satchell (1991), and Yang & Zhang (op. cit.) have recently started to gain traction in practice, the Yang–Zhang estimator being the preferred practitioner's choice (Andersen et al., 2006). See Appendix A for mathematical definitions of these estimates. The appeal of range-based estimators is that they can account for intra-day volatility, time-varying drift of the underlying asset-price process, opening price gaps and market microstructure noise; all issues that historical and realised volatility unavoidably ignore. Figure 2 displays rolling five-year ALSI total-return volatility calculated by means of the different volatility estimators.

3.2.1.5 Notice the substantial spread between the estimators throughout the period. A common empirical finding of many studies is that classic standard deviation yields numbers higher than proposed alternative volatility estimators (Yang & Zhang, op. cit; Poon, op. cit.). Whilst this is not clearly apparent for the ALSI, classic historical volatility is one of the highest estimators. Obviously, use of these estimators is affected by the availability of daily opening, high, low and closing market prices. This is readily available only for the ALSI (and Top40) from June 2002 onwards, limiting the maximum term of analysis to approximately 11 years. This presents a problem for long-term volatility estimation. However, as the historical intra-day dataset increases, it is suggested that long-term historical volatility should, in the future, be estimated with a range-based estimator or with integrated variance.

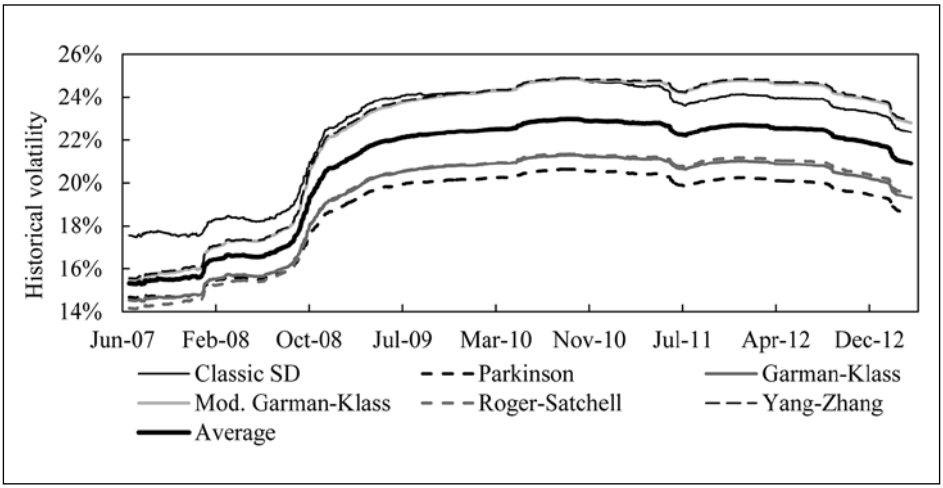

Figure 2. Daily five-year rolling ALSI total-return volatility estimates over the period June 2007 to April 2013

8 Brandt & Kinlay, supra

3.2.2 Return Term and Estimation Period

3.2.2.1 Apart from the actual function used to measure asset-return variation, one must also realise that historical volatility is keenly affected by the choice of sampling frequency and sampling period. 'Sampling frequency' here refers to the term, τ, over which returns are measured, leading to:

$$r_{t,\tau} = \ln\left(\frac{S_t}{S_{t-\tau}}\right) = \sum_{j=1}^{\tau-1} r_{t-j,1}\,. \tag{3}$$

One of the stylised facts—defined fully in section 4.1—identified by Cont (2001) was that of aggregational Gaussianity: the return distribution tends towards normality as the return term increases. While this fact has been analysed in several markets with varying results (Bingham & Kiesel, 2004; Flint, Chikurunhe & Seymour, 2012)[9] what is true is that historical volatility—and actually the complete return distribution—is heavily dependent on τ.

3.2.2.2 Sampling period also plays a large role in determining historical volatility. Firstly, estimation error is always a concern for any empirical analysis. It has been shown that, under certain asset-process conditions, sampling size plays a vital role in bounding theoretical estimation error (McAleer & Medeiros, 2008). Secondly, there is extensive literature showing that both the drift and volatility of postulated asset price processes is time-varying (c.f., e.g., Poon & Granger, 2003 and Brownlees, Engle & Kelly, 2011) and also that the market displays evidence of structural breaks (Hacker & Hatemi-J, 2006).

3.2.2.3 Figure 3 displays Top40 total-return volatility as measured by $\sigma_{C,T}$ over a τ-range of one day to one month (assuming 22 trading days per month), calculated

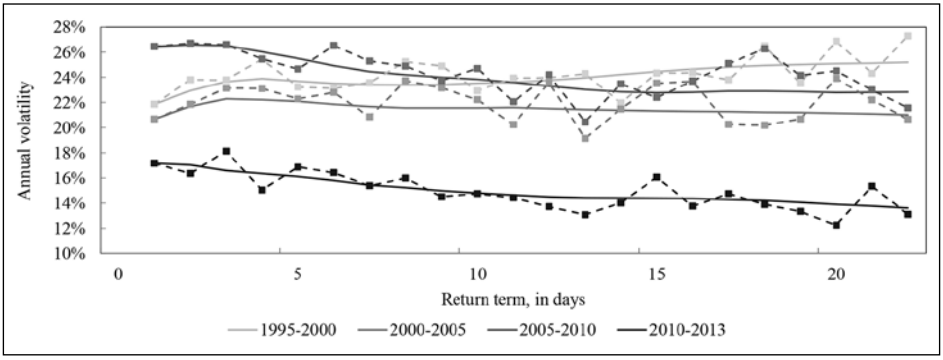

Figure 3. Top40 total return volatility estimates sampled over various frequencies and periods using non-overlapping return data (solid) and overlapping data (dotted)

9 The authors have also had sight of DA Polakow, DR Taylor & O Mahomed (in preparation). Aggregational gaussianity in the South African equity markets: implications for the pricing of risk

using the 1995 Top40 historical sample period and compared with that calculated from disjoint five-year periods. Historical volatility is calculated by means of both non-overlapping returns (solid lines) and τ–1 overlapping returns (dotted lines). Because of the general paucity of market data, one is forced to use overlapping returns as τ increases, in order to obtain a sufficiently large sample size. In both cases, differences across return term and period are readily apparent.

3.3 MEASURING HISTORICAL VOLATILITY ON THE CORRECT UNDERLYING DATA

3.3.1 THE IMPORTANCE OF THE ISSUE

3.3.1.1 While the identification of the correct underlying may, at first glance, appear somewhat obvious, it is actually of crucial importance. Perhaps contrary to one's general intuition, there is no definitively correct choice. To illustrate this, let us consider the following example.

3.3.1.2 An insurer has written a 30-year policy, which has an embedded minimum investment maturity guarantee. Investment performance is that of some balanced asset portfolio. For the purposes of our discussion, we focus exclusively on the equity portion. In a similar manner to that suggested by APN 110 (cf. ¶4.2), the insurer compares historical volatility with the available implied volatility term structure—inevitably short-term—and calculates a suitable historical implied-volatility scaling parameter. The insurer then estimates long-term historical volatility, multiplies this estimate by the imputed scaling factor and uses this as the fixed long-term volatility parameter in either a time-varying deterministic volatility model or a stochastic volatility model.

3.3.1.3 This example is actually quite close to market reality. Take note of just how many different volatility estimations, models, terms and types are inherent within this example process. Firstly, in practice, equities are usually modelled as a single asset class and guarantees are normally written on total return indices (APN 110 survey), implying that one should consider the total returns on either the ALSI or Top40. Furthermore, observe that the insurer has specifically written the guarantee on equity performance and not on forward or futures performance.

3.3.1.4 Secondly, the insurer compares historical asset volatility with implied volatility. South African exchange-traded options are written on Top40 futures with pre-specified maturities. Thus, the implied volatility term structure is really the implied volatility on futures options struck at the prevailing Top40 futures level at various maturities. For consistency, one should then really construct Top40 forward levels and measure the historical volatility of the constructed forward returns. An additional benefit is that historical volatility measured on index forwards latently accounts for the stochastic nature of interest rates and dividend yields, a feature also inherent in implied volatility.

3.3.1.5 Thirdly, the long-term implied volatility estimate is generally used as a fixed parameter in a specified volatility model. Whether the use of the implied volatility estimate as the fixed, long-term volatility parameter is suitable will depend on what type of model is specified. For instance, implied volatility is directly modelled by deterministic

models, whereas stochastic volatility prescribes dynamics for the underlying asset-price volatility. This distinction is subtle and is usually ignored (incorrectly) in practice.

3.3.1.6 On the basis of the discussion above, the authors advocate using historical volatility measured on the log returns of constructed Top40 forwards for the equity portion of the balanced portfolio. Whilst there is a slight mismatch throughout the life of the guarantee between performance of the underlying equity and equity forward, this is not an issue for the embedded European guarantees usually found in life policies. Furthermore, forwards by construction are investors' best estimates of the future level of the underlying asset price allowing for the inclusion of the stochastic risk-neutral drift and are thus prime candidates for estimating a forward-looking volatility estimate. Finally, the inherent inclusion of interest-rate and dividend-yield volatility in the drift term ensures further consistency with market-implied volatilities, which are also forward-looking. Therefore, the original question now becomes which forward to take as the underlying and over what period to measure log returns.

3.3.2 Constant-Maturity versus Floating-Maturity Forwards

3.3.2.1 The current forward level, $F_{t,s}$, represents the time t expected (in a risk-neutral sense) future value of the underlying at the specified time $T=t+s$, and S is the remaining time to maturity (the reason for not using τ to represent forward term is given below). Assuming that the yield curve, $y_{t,s}$, and the dividend yield, $\delta_{t,s}$, is stochastic, we can write:

$$F_{t,s} = S_t e^{(y_{t,s}-\delta_{t,s})s} \tag{4}$$

3.3.2.2 Forward prices are thus dependent on asset level, term-specific yield and dividend yield, and the remaining time to maturity. From equation (4) one is able to construct either a constant-term forward (CTF) price series, or a floating-term (FTF) series. We will define the CTF price series as $\{F_{t,s}\}$ and the FTF price series as $\{F_{t,T-t}\}$. Note that s and T are fixed but t increases through time, giving one the constant and floating terms as required. When calculating returns on forwards, there are essentially two terms to consider. The (backward-looking) term over which one measures the return is given by τ, while the (forward-looking) term of the forward is given by s and $T-t$ respectively. These two terms need not be equivalent, although the return term τ cannot be larger than the given forward term. Using the notation of equation (3), let us define $r^{\text{CTF}}_{t,\tau,s}$ and $r^{\text{FTF}}_{t,\tau,T-t}$ as the τ-period log returns from the CTF and FTF series' respectively. Mathematically, we can write:

$$r^{\text{CTF}}_{t,\tau,s} = \ln\left(\frac{F_{t,s}}{F_{t-\tau,s}}\right), \tag{5}$$

$$r^{\text{FTF}}_{t,\tau,T-t} = \ln\left(\frac{F_{t,T-t}}{F_{t-\tau,T-t+\tau}}\right). \tag{6}$$

3.3.2.3 Should one now calculate the historical volatility of, say, the daily rolling τ-period CTF returns and compare this directly with the τ-period implied volatility, or should one rather consider the average realised volatility of daily FTF returns over the τ-period life of the forward and compare this with τ-period implied volatility? Figure 4 depicts this dichotomy. Each line represents the possible cumulative returns over the life of a forward.

3.3.2.4 The annualised s-period volatility of the terminal forward return distribution at time T is given by $\sigma_{T,s}^{\text{CTF}}$, while the average annualised realised s-period volatility of the daily forward return distribution at time T is given by $\overline{\sigma_{T,s}^{\text{FTF}}}$. These volatilities are obviously dependent on the specified sample path. If the log forward price were perfectly defined by geometric Brownian motion (or, in fact any elliptically symmetric distribution), then terminal volatility would be equivalent to the average realised volatility scaled by the square root of the contract term. However, as is shown in section 4.1, this is not the case. Annualised realised volatility averaged across all sample paths is not the same as annualised terminal distribution volatility. So which of these two estimates is the most suitable historical volatility estimate? We consider first several theoretical points and then provide some empirical results.

3.3.2.5 Breeden & Litzenberger's (1978) seminal work proved that an implied volatility curve is simply another way of representing the underlying risk-neutral terminal distribution at a specific term. This result seems to favour CTF terminal distribution volatility over average daily realised FTF volatility. In addition, market consistency generally implies calibration only to vanilla option prices, which are solely based on the discounted expectation of the terminal payoff, again suggesting terminal volatility. However, there are no long-term options available within the market. Thus, one would have to rely on some sort of quasi-dynamic replication argument to hedge out any embedded guarantee exposure, which would necessarily be reliant on realised volatility over the period. This, contrastingly, suggests averaged realised volatility. However, as

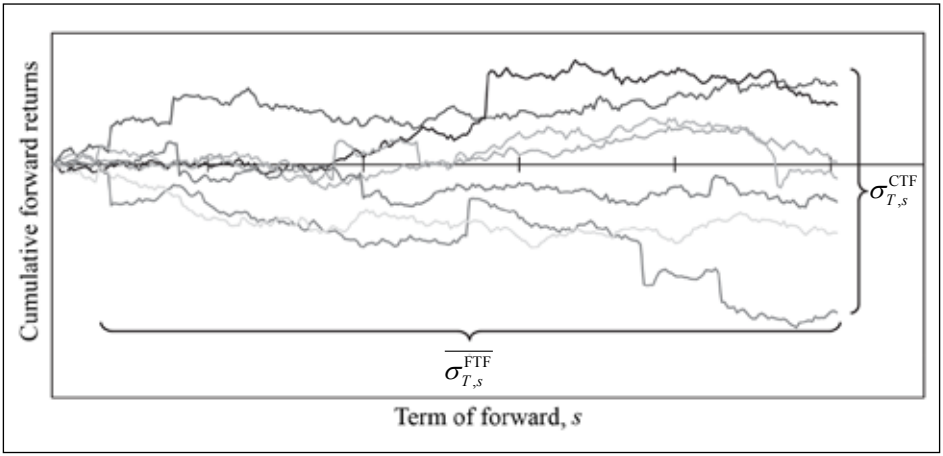

Figure 4. Terminal distribution CTF volatility vs. average realised FTF volatility

Sheldon & Smith (2004) note, market consistency seems to require implied rather than historical volatility, which would again suggest using terminal distribution volatility. On the whole then, it would appear that there may be stronger theoretical evidence supporting the use of terminal forward-return distribution volatility rather than averaged realised forward-return volatility.

3.3.2.6 A single sample path in Figure 4 represents the evolution of a constant-term forward over its historical contract life. The different sample paths are created by moving the start date of the constant-term forward through time. For ease of reference, we display each path beginning at the same start date. A simple schematic representation of this process is given in Figure 5, where, for simplicity, only six periods of history and terms up to three periods long are assumed, and abridged notation has been used. As a toy example, consider the terminal and average realised volatility of the three-period forward returns given below.

3.3.2.7 Using the spot and dividend-yield vectors over time, and the yield curve matrix across time and term, one can construct a CTF price matrix from which one can calculate CTF log returns. Using the notation introduced above, the first subscript denotes time (i.e. row number), the second subscript denotes the term of the return—all daily returns—and the third subscript denotes the term of the forward (i.e. the column number). We first consider the calculation of the average realised volatility of the three-period forward as at time period 5, $\overline{\sigma_{5,3}^{\text{FTF}}}$. In our example, we have three FTF return sample paths of a three-period forward, each displayed by a diagonal arrow. From each of these FTF sample paths, one calculates annualised realised volatilities, displayed on the left of the forward return matrix. Finally, the average of these volatilities represents is denoted $\overline{\sigma_{5,3}^{\text{FTF}}}$ and denotes the average realised three-period FTF volatility as at time 5.

3.3.2.8 Terminal forward return volatility is calculated from period-specific aggregated forward returns. Consider again the three-period case. The three-period

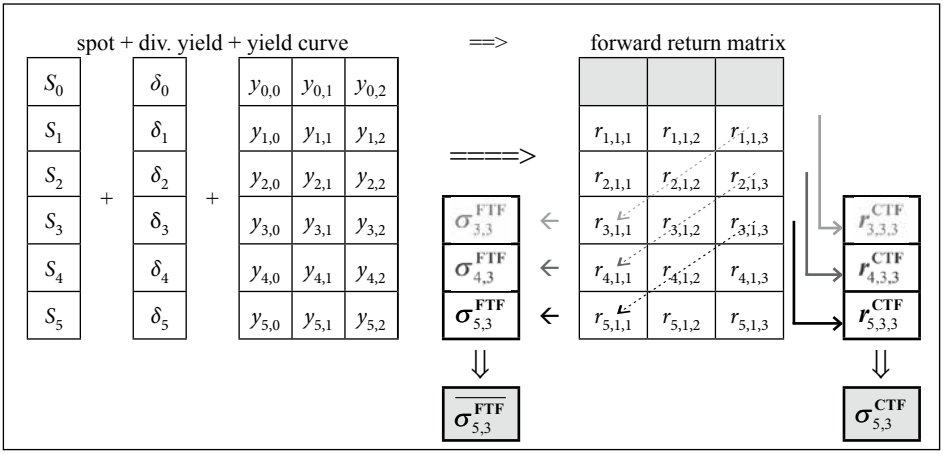

Figure 5. Schematic representation of terminal forward return volatility and average realised forward return volatilities

aggregated (log) return at time 3 is calculated by summing up the single period (log) returns at times 1, 2 and 3. The starting point is then moved one period forward and the process is repeated until the end of the dataset is reached. In our example, we have three three-period CTF returns, displayed on the right of the forward return matrix. From this CTF return series, we can calculate the three-period terminal volatility at time 5, $\sigma_{5,3}^{\text{CTF}}$. Very importantly, the length of the aggregating return period defines which forward return column to use. That is, we use the three-period historical forwards to calculate the three-period aggregated returns. This ensures that one is truly using the best estimate of the theoretical risk-neutral terminal distribution and thus measuring volatility as consistently as possible with implied volatility.

3.3.2.9 From the 1925 and 1976 ALSI monthly-return datasets, the 30-year historical volatility term structures were obtained, as displayed in Figure 6. The dataset used—1925 and 1976 monthly data—is represented by the line colours black and grey while the type of volatility—terminal CTF, average realised FTF and spot—is given by the type of line, i.e. solid, dashed and dotted. Also shown is the 17,75-year FTF volatility term structure calculated from the 1995 Top40 daily-return dataset.

3.3.2.10 Clearly, volatility is strongly dependent on both the method and dataset used. That being said, there is definitely a clear upwards-trend up to the 20-year mark irrespective of dataset or method. Furthermore, the 1925 CTF, 1976 FTF and 1995 FTF volatility series give fairly comparable results up to the 18-year mark. Contrastingly, the 1976 CTF volatility series is concave. It has a steeper slope than any other volatility, climbing up to a 20,25-year maximum value of 43,93 per cent, after which there is a significant downturn. However, this rather different general behaviour may simply be due to small sample size for longer terms.

3.3.2.11 If, as motivated above, one considers terminal-distribution CTF volatility to be the best historical estimate, then 15-year volatility is estimated either as 29,11% using the 1925 dataset, or as high as 39,04% for the 1976 dataset. Looking

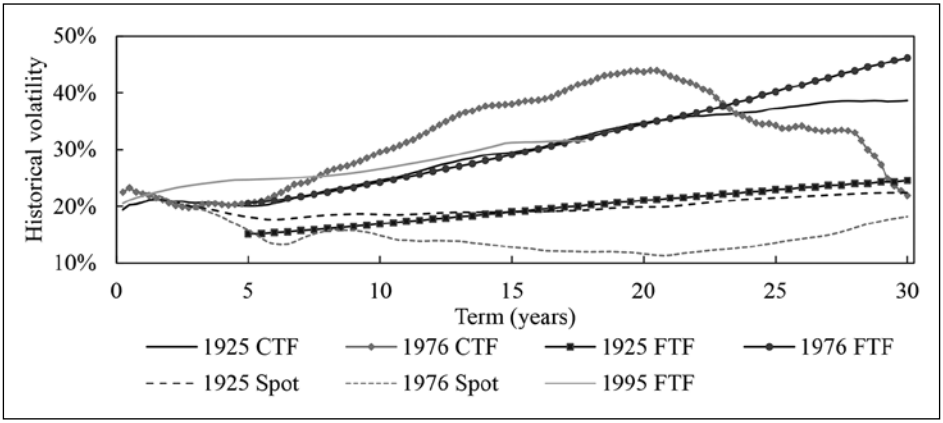

Figure 6. Terminal CTF volatility and average realised FTF volatility compared with spot volatility

further out to the 25-year point, one finds corresponding volatilities 34,28% and 37,25%. As discussed in section 5, these values are considerably higher than what is considered usual. However, this does not mean that they are unreasonable. As is shown in section 4.4.2, daily log returns of forwards comprise three distinct parts, namely, underlying asset return, change in dividend yield and change in yield curve. If one simply assumes that each component is independent, then CTF variance is merely the sum of the three component variances. So far in section 3 we have found that underlying asset volatility alone can be as high as 25%, and sometimes considerably higher. If one then adds yield-curve and dividend-yield volatilities, a 25-year CTF volatility of 35% seems empirically reasonable.

3.3.2.12 In summary, historical volatility should always be measured on the most appropriate underlying data series. The authors argue that this would either be the CTF forward-price series, representing the terminal forward distribution, or the FTF forward-price series, representing the realisation of each forward over time. These series can be constructed fairly simply from empirical data and the resultant terminal distribution CTF volatility and the average realised FTF volatility term structures calculated. Whilst the results are varied, there are common characteristics between both calculation method and dataset used, providing compelling evidence to suggest that a 25-year volatility of 35% is not unreasonable.

3.4 THE SOUTH AFRICAN IMPLIED–HISTORICAL VOLATILITY RELATIONSHIP

3.4.1 The relationship between historical and implied volatility has been extensively researched; a review of early work is given in Shu & Zhang (2001), while Eraker[10] outlines the more recent literature. Coined 'the volatility premium', average implied volatility on index options has consistently been shown to be higher than historical index volatility. Market participants try to take advantage of this mismatch through the use of various option strategies (Driessen & Maenhout[11]). Whilst there are several competing theories that attempt to justify the volatility premium, the focus here is on an empirical analysis of the implied–historical volatility relationship in South Africa.

3.4.2 Daily rolling terminal CTF volatility is calculated and compared with daily rolling term-specific implied volatility. Figure 7 shows the implied–historical volatility (IVHV) ratio since September 2005 for terms ranging from three to twelve months. Clearly, the IVHV ratio varies over time and shows strong signs of heteroscedasticity, or non-constant volatility. This finding is robust to the type of historical volatility estimation as well as the chosen type of return.

3.4.3 The effect of the subprime crisis is readily apparent, although somewhat lagged because of the *ex-post* nature of the historical volatility estimator. This lag is most pronounced for the 12-month IVHV ratio. This leads to significant negative

10 B Eraker (unpublished). The Volatility Premium. Working Paper, Duke University, 2008

11 J Driessen & P Maenhout (unpublished). The World Price of Jump and Volatility Risk. Working paper, Insead, 2006

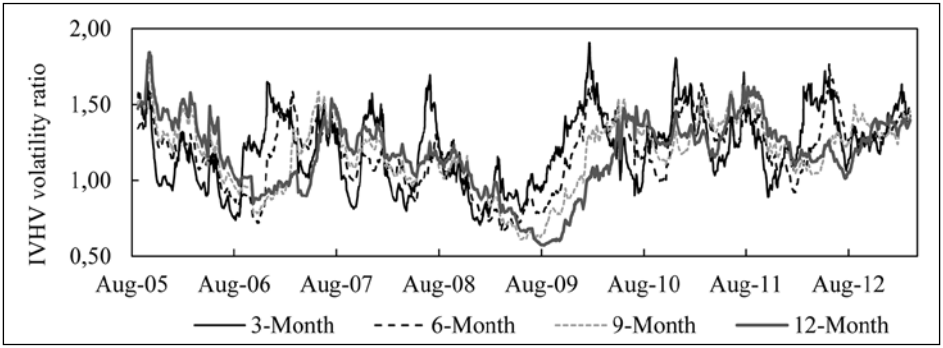

Figure 7. Daily IVHV ratios from 6 September 2005 to 22 March 2013 for terms of 3 to 12 months

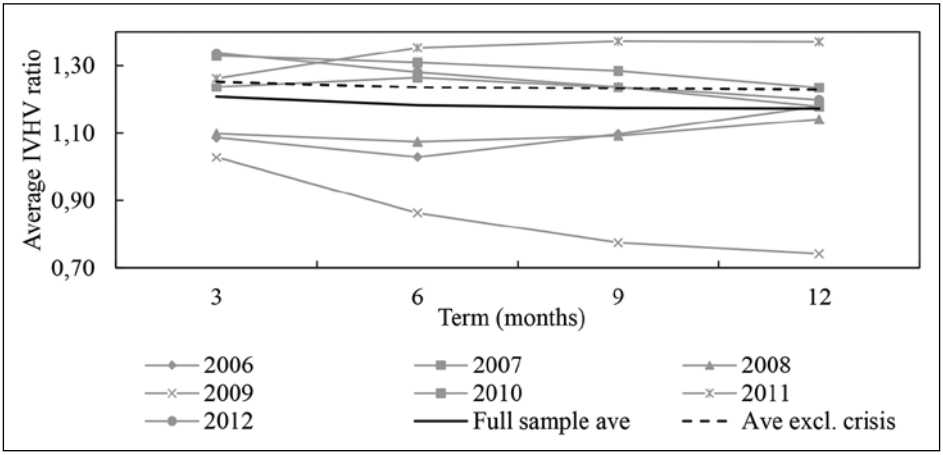

Figure 8. Average IVHV ratios for calendar years, as well as for the full sample, inclusive and exclusive of the subprime crisis

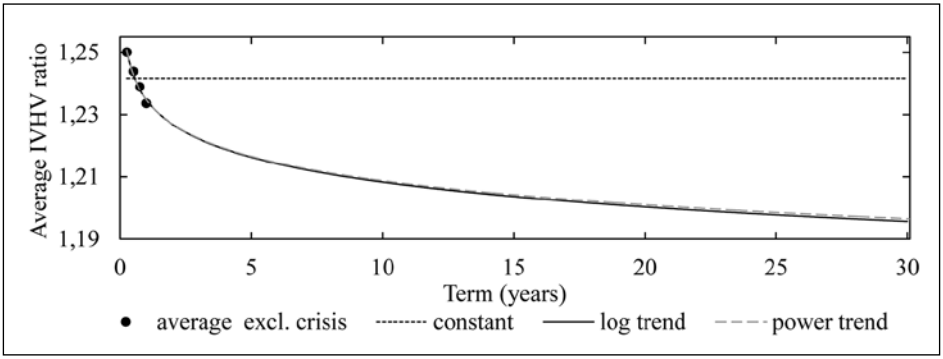

Figure 9. Extrapolated average IVHV ratios

skewness within the IVHV ratio distributions, barring the three-month series, which displays symmetry. In addition, the ratio distributions are all platykurtic, the shorter-term ratios displaying the lowest excess kurtosis. Figure 8 plots the average IVHV ratios (in grey) for each year from 2006 to 2012, the full sample average ratios and the average ratios calculated when the subprime crisis period is removed.

3.4.4 The 2009 average ratios are indicative of the subprime crisis and are clearly irregular. If these outlier IVHV ratios are removed, the sample average is increased by approximately 0,05 across all terms. A constant (1,244) or time-varying function can then be fitted to extrapolate this relationship out to the required term as shown in Figure 9.

4. ECONOMETRIC VOLATILITY MODELLING: THE GARCH FAMILY

4.1 STYLISED FACTS OF EMPIRICAL ASSET RETURNS

4.1.1 Empirical financial data are known to be characterised by several 'stylised facts', defined as statistical properties pervasive across a wide range of instruments, markets and time periods (Cont, op. cit.). Several of these market facts are of direct concern to any volatility modelling exercise:

(1) Heavy-tailed distributions: the tails of the conditional and unconditional returns distribution are most commonly modelled by a Pareto distribution with finite tail index between two and five.
(2) Skewed distributions: one observes larger individual losses than gains for stock and index returns, implying negatively skewed short-term return distributions.
(3) Volatility clustering: short-term volatility displays positive autocorrelation. This is technically referred to as conditional heteroscedasticity;
(4) Volatility feedback effect: asset volatility is generally negatively correlated with asset performance.

4.1.2 In order to adequately model the above stylised facts, one needs to use some time-varying function. The autoregressive conditional heteroscedasticity (ARCH) model introduced by Engle (1982), and subsequently generalised by Bollerslev (1986) to the GARCH model, has become the standard model for modelling such features.

4.2 GARCH VOLATILITY FORECASTING: BASIC THEORY

4.2.1 The standard GARCH(p,q) model for the return r_t during period t with conditional variance, h_t, takes the following form:

$$r_t = \mathbb{E}_{t-1}\left[r_t\right] + \varepsilon_t; \tag{7}$$

where:

$$\varepsilon_t = \sqrt{h_t} z_t;$$

$$h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 + \sum_{k=1}^{q} \beta_i h_{t-k}^2;$$

$\mathbb{E}_{t-1}\left[\bullet\right]$ is the expectation conditional on all information available at time $t-1$; and

z_t is a series of independent, identically distributed (iid) random variables with zero mean and unit variance.

The standard GARCH model assumes that z_t is standard-normally distributed. If $q=0$ in equation (7), then the model reduces to an ARCH(p) model. In most academic literature—and certainly in practice—a simple GARCH(1,1) specification is used to model the volatility of financial time series. This model has been shown to be highly robust and it is only with some difficulty that one can find an alternative model that shows consistent outperformance (Hansen & Lunde, 2005). We can thus rewrite h_t as:

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{8}$$

4.2.2 As a special case, equation 8 reduces to an exponentially weighted moving average (EWMA) when $\alpha_0=0$ and $\alpha_1=1-\beta_1=\lambda$. However, what makes the ARCH class of models so useful—in comparison to EWMA for example—is that one can optimally forecast volatility—as well as the full conditional density—using only equations 7 and 8. This is due to the embedded stochastic process $\{z_t\}$ within the conditional volatility function.

4.2.3 In particular, if one assumes that the conditional return expectation is zero and that $\alpha_1+\beta_1<1$, then the unconditional variance of the asset is

$$\sigma^2 = \frac{\alpha_0}{(1-\alpha_1-\beta_1)}, \tag{9}$$

and the optimal, *k*-step ahead, single-period variance forecast can be written as

$$h_{t+k} = \sigma^2 + (\alpha_1+\beta_1)^{k-1}(h_t - \sigma^2). \tag{10}$$

Therefore, as *k* increases, the forecasts will exponentially tend towards the long-run unconditional volatility at a rate that is governed by the process's persistence, $\alpha_1+\beta_1$.

4.2.4 Assuming that the correct GARCH model has been specified, one can appropriately forecast the variance term structure across *k*-period returns as the sum of the conditional variance forecast over the period:

$$h_{t:t+k} = j\sigma^2 + \frac{(h_t - \sigma^2)\left(1-(\alpha_1+\beta_1)^k\right)}{1-\alpha_1-\beta_1}. \tag{11}$$

For further information on GARCH theory, see Andersen et al. (2006) and Alexander (2008b).

4.3 THE EXTENDED GARCH FAMILY

4.3.1 While the basic GARCH model discussed in the previous section is by far the most ubiquitous in financial time series modelling, there are numerous other models that fall within the general ARCH family. Essentially one can classify a GARCH model by three characteristics:

(1) model type: the functional form of the conditional variance equation;
(2) innovation distribution: the assumed distribution for z_t; and
(3) model parameters: the number of lags within each component as well as the use of external covariates.

This section focuses on the first two characteristics of the model-distribution-parameter triplet and only addresses the base (1,1) parameter case. We limit the possible model set to the five candidates highlighted by Brownlees, Engle & Kelly (op. cit.) as both being tractable and having noteworthy volatility forecasting ability. The five candidates are described below in section 4.4.2. The possible innovation distributions are limited to the Gaussian and Student's *t* distributions. Model fitting is performed for each specification using the Gaussian and Student *t* likelihood functions respectively. The best-fitting model is defined by the model-distribution-parameter triplet that displays the minimum Bayesian Information Criterion (BIC) score, a commonly used model-selection statistic that essentially penalises a model's log-likelihood value for the number of estimated model parameters (Schwarz, 1978).

4.3.2 Extended GARCH Volatility Models

4.3.2.1 The standard GARCH(1,1) model outlined in section 4.2.1 has been criticised because of its symmetric treatment of positive and negative return shocks. In practice, it has been shown that negative return shocks increase conditional volatility more than positive return shocks of equal magnitude. This asymmetry is usually referred to as the 'leverage' or 'volatility feedback' effect (Andersen et al., 2006). While there are many extended GARCH models that account for this stylised fact, three models in particular have become prevalent.

4.3.2.2 The GJR or threshold GARCH (GJR-GARCH) specification of Glosten, Jagannathan & Runkle (1993) accounts for asymmetry by including an additional ARCH term conditioned by the sign of the previous innovation. Thus, GJR-GARCH(1,1) is written

$$h_t = \alpha_0 + \left(\alpha_1 + \gamma \mathbf{1}_{\{r_{t-1}<c\}}\right)\varepsilon_{t-1}^2 + \beta_1 h_{t-1}, \tag{12}$$

where **1** is the indicator function and c is a threshold return level, normally set to 0. The parameter γ controls the differential effect attributable to negative and positive return shocks.

4.3.2.3 Alternatively, the exponential GARCH (EGARCH) specification of Nelson (1991) models the logarithm of the conditional variance, and is given by

$$\ln\left(h_t\right) = \alpha_0 + \alpha_1\left(\left|\varepsilon_{t-1}\right| / \sqrt{h_{t-1}} - \sqrt{2/\,pi}\right) + \gamma\left(\varepsilon_{t-1} / \sqrt{h_{t-1}}\right) + \beta_1 \ln\left(h_{t-1}\right). \tag{13}$$

The leverage effect is again controlled by γ, with $\gamma < 0$ meaning that volatility increases more with negative-return shocks than with comparable positive shocks.

4.3.2.4 The fourth possible model is the nonlinear or power GARCH (NGARCH or PGARCH) specification of Higgins & Bera (1992). This model also attempts to capture the volatility asymmetry but it uses a slightly different form:

$$h_t^{\delta/2} = \alpha_0 + \alpha_1 \left|\varepsilon_{t-1}\right|^{\delta/2} + \beta_1 h_{t-1}^{\delta/2}. \tag{14}$$

The flexible δ allows one to capture more accurately the conditional volatility dynamics.

4.3.2.5 The final specification is the asymmetric power GARCH (APGARCH) specification of Ding, Granger & Engle (1993):

$$h_t^{\delta/2} = \alpha_0 + \alpha_1 \left|\left|\varepsilon_{t-1}\right| - \gamma\varepsilon_{t-1}\right|^{\delta/2} + \beta_1 h_{t-1}^{\delta/2}. \tag{15}$$

Similar to NGARCH above, APGARCH explicitly allows for the asymmetric volatility effect while also including flexible volatility dynamics. As Hentschel (1995) notes, the APGARCH specification latently nests a number of differing GARCH models.

4.4 GARCH VOLATILITY AND MARKET-CONSISTENT VALUATION

4.4.1 One can create a forward volatility term structure by taking the square root of equation (11). This provides one with a potential method for estimating long-term volatility in a market-consistent manner. In practice, Milliman, a large international actuarial consulting firm does exactly this when constructing their Milliman Guarantee Index. Based on a GARCH(1,1) model and coupled with market quotes where available, Milliman obtain a transparent, market-consistent 30-year volatility term structure from which expected hedging costs of variable annuity guarantees are published in the Milliman Hedge Cost Index, available on Bloomberg (MLHCINEW Index). We include this example to show that GARCH models are actively being used to obtain long-term market-consistent volatility estimates.

4.4.2 Practical GARCH Implementation Issues

4.4.2.1 As with historical volatility estimation, one should first consider which underlying return series to model and subsequently how to correctly use the forecast volatility term structure. The initial consideration includes sample size and sampling frequency as sub-issues. The second consideration refers to the manner in which GARCH can be used in a simulation, pricing or hedging framework.

4.4.2.2 Brownlees, Engle & Kelly (op. cit.) provide substantial empirical evidence that GARCH models perform best when using the longest available data series with frequent parameter updating. In terms of sample frequency, Alexander (2008b) and Poon (op. cit) state that GARCH models should ideally model daily (or intra-day) return data. Many of the effects that GARCH tries to capture are not readily apparent in monthly data because of the aggregation process and the fitted parameters are more likely to give spurious forecast results. Despite these misgivings, the authors fitted GARCH models to both daily and monthly return data, using the monthly results mostly for comparative analysis. Section 4.5 provides further detail on the empirical results.

4.4.2.3 Section 3.3 above highlights the importance of choosing the correct underlying-asset return series on which to measure historical volatility. One has to make a similar decision when fitting GARCH models. As discussed above, one would ideally

want to estimate the volatility of either the τ-period log returns of the CTF series, $\left\{r^{\text{CTF}}_{t,\tau,\tau}\right\}$, or the average volatility of the single-period log returns of the FTF series, $\left\{r^{\text{FTF}}_{t,1,\tau}\right\}$. Note that return term and forward term are equivalent. Using equations (4) and (5) we have:

$$\begin{aligned} r^{\text{CTF}}_{t,\tau,\tau} &= \ln\left(\frac{F_{t,\tau}}{F_{t-\tau,\tau}}\right) \\ &= \ln\left\{\frac{S_t \exp\left[\left(y_{t,\tau}-\delta_{t,\tau}\right)\tau\right]}{S_{t-\tau}\exp\left[\left(y_{t-\tau,\tau}-\delta_{t-\tau,\tau}\right)\tau\right]}\right\}, \end{aligned}$$

which, after some algebraic manipulation, gives:

$$\begin{aligned} r^{\text{CTF}}_{t,\tau,\tau} &= \tau\left[\left(y_{t,\tau}-y_{t-\tau,\tau}\right)-\left(\delta_{t,\tau}-\delta_{t-\tau,\tau}\right)\right]+r_{t,\tau} \\ &= \tau\sum_{i=1}^{\tau}\left(\Delta y_{t-i,\tau}-\Delta\delta_{t-i,\tau}+\frac{1}{\tau}r_{t-i,1}\right). \end{aligned} \tag{16}$$

Therefore, the τ-period log return of the fixed-maturity τ forward can be written as a linear sum of (1) the daily changes in constant-term yields—the natural time-series choice for fixed-income modelling (Meucci, 2005), (2) the daily changes in constant-term dividend yield; and (3) the daily single-period log returns of the underlying asset over the specified period τ. Equation (16) thus neatly partitions the requisite forward modelling exercise into three distinct sections:

- yield-curve forecasting;
- dividend-yield forecasting; and
- asset-price forecasting.

4.4.2.4 A similar exercise for $r^{\text{FTF}}_{t,1,s}$ produces the same partitions as above—albeit in a clumsier expression. Whereas for a single constant-maturity forward one need only model the relevant fixed-term yield across the τ-period, one needs to model the entire yield curve up to term τ for a changing-maturity forward. Thus, the construction of a complete volatility term structure from either return series would necessitate modelling the entire yield curve.

4.4.2.5 The natural candidate series for GARCH modelling is thus the daily underlying asset log return series. However, this should always be coupled with the respective yield-curve and dividend-yield models to calculate correctly the relevant forward return series. This can be done either by simulation or, if closed-form solutions exist, by analytic forecasting.

4.4.2.6 One of the potential benefits of using a GARCH-based framework is that there exists a large body of work applying GARCH modelling direct to risk-neutral option pricing. For instance, Heston & Nandi (2000) and Duan, Ritchken & Sun's (2006) GARCH models are often used in practice as alternative option pricing models to, say,

stochastic-volatility models. This allows one to price, hedge and manage risk effectively under the same framework, increasing overall modelling tractability and minimising model incompatibility issues.

4.4.3 GARCH Long-Term Forecasting Caveats

Alexander (2008b) and Brownlees, Engle & Kelly (op. cit.) note that GARCH was not intended as a long-term forecasting model; at least in the actuarial sense. Rather, one finds that 'long-term' in the majority of GARCH literature refers to anything between one month and a year. Thus, one must always be aware that simply choosing the best fitting model may neither provide the best out-of-sample forecasts, nor the correct forecast dynamics. In fact, it is usually the long-term volatility parameter in GARCH models that is hardest to estimate when fitting. Alexander (2008c) notes that a common technique in practice is to fix the long-term volatility parameter before fitting the remaining model parameters to the data, a practice analogous to that advocated by APN 110.

4.5 DAILY TOP40 AND MONTHLY ALSI GARCH VOLATILITY FORECASTS

4.5.1 GARCH models were fitted to two datasets: the 1995 daily Top40 log returns and the 1976 monthly ALSI log returns. Model parameter estimation was done using the ARMAX-GARCH-K Toolbox in Matlab. The conditional return expectation in equation (7) is assumed to be constant. Using equations (10) and (11)—adjusted accordingly per model specification—a 30-year volatility term structure was forecast. Tables 1 and 2 provide summary fitting statistics for each dataset under the five candidate models and the two innovation distributions.

4.5.2 On the basis of the minimum BIC scores, the GJR-GARCH(1,1) model provides the best fit for the daily Top40 dataset, whilst the basic GARCH(1,1) model is chosen for the monthly Top40 dataset. Unsurprisingly, use of the Student's t distribution improves model calibration to both datasets in all cases bar one. Interestingly, choice of innovation distribution has a much larger effect on daily Top40 model performance than choice of model specification. The GARCH(1,1) model calculated using the Student's t distribution provides a much better fit than any daily or monthly model under the Gaussian distribution. See Kulikova & Taylor (2010) for a more rigorous investigation of the effects of distribution choice on GARCH models of South African indices.

4.5.3 Figure 10 displays the fitted daily Top40 volatility under the GARCH and GJR-GARCH models. Historical discrepancies between the different models are very slight. Whilst the benefits of time-varying GARCH modelling in comparison with constant volatility are reasonably clear, the real advantage of GARCH versus other time-varying estimation methods lies in its ability to forecast volatility. Figure 11 shows the forecast daily Top40 volatility term structures from the GJR-GARCH and GARCH models respectively under each innovation distribution. In comparison with Figure 10, the differences between the models are clearly visible in the volatility forecasts. The GJR-GARCH models have significantly lower unconditional volatilities than their GARCH counterparts because of the additional leverage parameter γ.

Table 1. Summary of GARCH models fitted to daily 1995 Top40 returns

Innovation distribution							
Gaussian							
Model	$\widehat{\alpha}_0$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\widehat{\gamma}/\widehat{\delta}_1$	$\widehat{\delta}$	DoF	BIC
GARCH	$0{,}20\cdot10^{-5}$	0,1088	0,8830	–	–	–	–27515,01
GJR-GARCH*	$0{,}21\cdot10^{-5}$	0,0479	0,8917	0,0958	–	–	–27574,40
EGARCH	–0,2870	0,2391	0,9678	–0,085	–	–	–27541,04
NGARCH	$0{,}25\cdot10^{-4}$	0,027	0,8171	–	0,4758	–	–27317,07
APGARCH	$0{,}10\cdot10^{-5}$	0,0012	0,8570	0,5941	1,1388	–	–27358,01
Student's *t*							
GARCH	$0{,}18\cdot10^{-5}$	0,1028	0,8893	–	–	8,5884	–27647,24
GJR-GARCH*	$0{,}19\cdot10^{-5}$	0,0491	0,8956	0,0863	–	9,3848	–27679,07
EGARCH	–0,2601	0,2277	0,9709	–0,0768	–	9,3427	–27645,19
NGARCH	$0{,}21\cdot10^{-8}$	0,0169	0,8653	–	0,5170	6,5658	–27501,67
APGARCH	$0{,}10\cdot10^{-5}$	0,0032	0,8939	0,7013	0,8546	8,1516	–27549,22

*The models with the lowest BIC score.

Table 2. Summary of GARCH models fitted to monthly 1976 Top40 returns

Innovation distribution							
Gaussian							
Model	$\widehat{\alpha}_0$	$\widehat{\alpha}_1$	$\widehat{\beta}_1$	$\widehat{\gamma}/\widehat{\delta}_1$	$\widehat{\delta}$	DoF	BIC
GARCH*	0,0003	0,1328	0,7919	–	–	–	–1229,91
GJR-GARCH	0,0005	0,1057	0,7009	0,1056	–	–	–1225,07
EGARCH	–0,7802	0,2600	0,8623	–0,0582	–	–	–1227,23
NGARCH	$0{,}34\cdot10^{-6}$	0,0537	0,8505	–	0,4287	–	–1228,96
APGARCH	0,0009	0,1257	0,8675	0,4875	0,0219	–	–1118,03
Student's *t*							
GARCH*	0,0003	0,1144	0,8195	–	–	10,653	–1232,57
GJR-GARCH	0,0004	0,0842	0,7740	0,0847	–	10,720	–1229,07
EGARCH	–0,6744	0,2569	0,8815	–0,0633	–	10,304	–1232,50
NGARCH	$0{,}13\cdot10^{-6}$	0,0769	0,8491	–	0,2919	9,6840	–1225,93
APGARCH	0,0025	0,1094	0,8662	0,4436	0,0813	10,089	–1227,73

*The models with the lowest BIC score.

4.5.4 Table 3 gives the unconditional volatility for the GJR-GARCH(1,1) and GARCH(1,1) models for the daily 1995 Top40 dataset under both distributions, as well as the unconditional GARCH(1,1) volatility for the monthly 1976 dataset. As discussed above in section 4.4.2, GARCH volatility—equivalent to the estimated realised asset volatility—merely represents one part of the three-stage volatility estimation procedure. Thus by using the IVHV scaling factor range of 1,204–1,244 found in section 3.4, we can estimate long-term implied volatility. Using the daily GJR-GARCH results, we find a long-term volatility estimate range of 23,9 to 24,7%. The monthly GARCH estimate

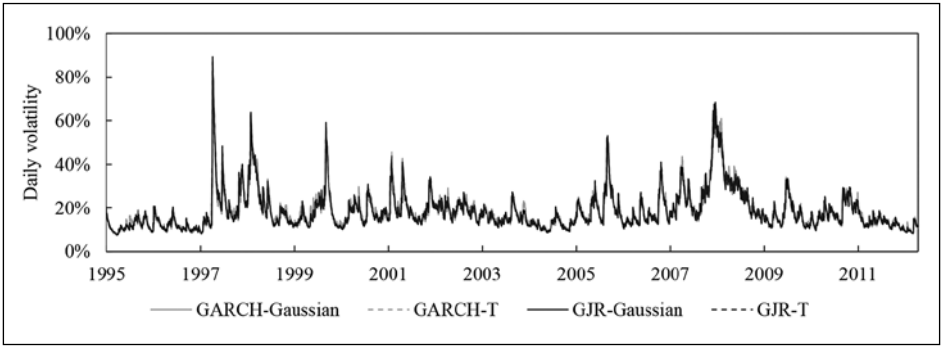

Figure 10. Daily Top40 GARCH(1,1) and GJR-GARCH(1,1) volatility

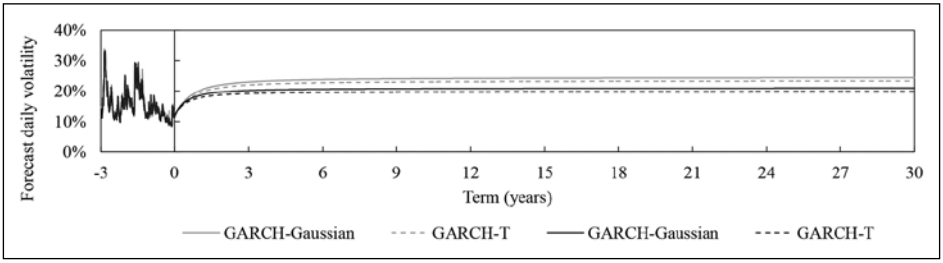

Figure 11. Daily Top40 volatility term structure: GARCH(1,1) and GJR-GARCH(1,1)

range of 25,7 to 26,6% is somewhat higher. That being said, one should always treat long-term GARCH estimates—that is, beyond a year—with particular caution and consider just how robust the forecasts are to model and distribution choice. In this case, use of the basic daily GARCH model leads to a much higher long-term volatility estimate of approximately 29%.

Table 3. Unconditional GJR-GARCH(1,1) volatility versus GARCH(1,1) equivalent

Dataset	Model	Innovation Distributions	Unconditional Volatility	IVHV Scaled Volatility
1995 – daily	GJR-GARCH	Gaussian	20,61%	24,81–25,64%
		Student's *t*	19,88%	23,93–24,73%
	GARCH	Gaussian	24,63%	29,65–30,64%
		Student's *t*	23,79%	28,64–29,59%
1976 – monthly	GARCH	Gaussian	21,82%	26,27–27,14%
		Student's *t*	21,36%	25,71–26,57%

5. DETERMINISTIC VOLATILITY MODELLING

5.1 According to the APN 110 sub-committee, all South African market participants in their 2010 survey used a time-varying deterministic volatility (TVDV) model for long-term implied volatility on equity indices at the time rather than a more sophisticated approach because of the lack of market data. Furthermore, all participants used an historical volatility estimate as the limiting long-term volatility parameter in the prescribed TVDV model. In this section, a brief overview of TVDV models and their calibration is given. Two contemporary TVDV models used within the South African context are highlighted below, the deterministic volatility-term-structure model used by Barrie & Hibbert and the deterministic volatility-surface model currently used by Safex. Another candidate TVDV model commonly used but not discussed here is Gatheral's (2006) stochastic volatility inspired model.

5.2 THE NATURE OF TIME-VARYING DETERMINISTIC VOLATILITY MODELS

5.2.1 It is a common misconception held by market practitioners that TVDV models give constant volatility across different moneyness levels, where moneyness is defined as option strike price over underlying asset price. In fact, TVDV models essentially fit separate curves to each traded maturity and then use a time-dependent function to link these curves in order to create a surface. Thus, a TVDV model is naturally split into two curve-fitting exercises: initial fitting across strike prices and subsequent fitting across time. These two exercises are linked during the total calibration exercise so as to ensure no butterfly-spread or calendar-spread arbitrage across the constructed surface.

5.2.2 In truth, the volatility surface from a TVDV model is not truly deterministic. Rather, TVDV is a deterministic function fitted to an underlying stochastic asset-price process. Thus, the TVDV surface remains stochastic because of its dependence on the underlying asset-price process. In this sense, local volatility is actually a nonparametric TVDV model. However, the reader should not infer that implied volatility coincides with local volatility; they are disparate. Rather, Dupire's (1994) equation provides a monotonic mapping between local and implied volatility. In contrast to deterministic models, stochastic volatility models assume that both the underlying asset-price and volatility processes are stochastic. See section 6 for more on stochastic volatility models.

5.3 THE BARRIE & HIBBERT MODEL

5.3.1 Barrie & Hibbert (BH) is a long-standing international financial consulting firm that provides comprehensive analytical support, particularly within the insurance sector. Their economic scenario generation (ESG) modelling platform is widely used in South Africa and the United Kingdom. Part of this platform is to provide accurate forecasts of long-term market-consistent valuation parameters. Specifically, the technical note by Roseburgh & Holmes (2006) outlines their approach to estimating long-term South African equity volatility by means of a simple TVDV model:

$$\sigma_{t,\tau} = \sigma_0 e^{-\alpha\tau} + \sigma_\infty \left(1 - e^{-\alpha\tau}\right). \quad (17)$$

The speed at which volatility converges to its long-run estimate σ_∞ is controlled by the α parameter, while the parameter σ_0 defines the instantaneous implied volatility.

5.3.2 During the BH quarterly calibration process, σ_∞ is fixed at 26% and the remaining two parameters are fitted to median, short-term (up to three years) implied volatility market quotes. The long-term volatility estimate of 26% was calculated by measuring the historical volatility of monthly equity returns over the 15-year period from 1989 to 2005 (21,3%) scaled up by an IVHV factor of 1,2 and rounded up to 26%.

5.3.3 Although the authors were unable to match exactly the BH historical volatility estimate, based on their estimations, 'monthly equity returns' most likely refers to monthly ALSI total returns. However, as discussed in section 3, the authors argue that, because of the choice of underlying return series and the sampling frequency of the returns, this is not the theoretically best justified method of measuring historical volatility. Use of what they suggest are the more correct historical volatility term structures given in Figure 6 leads to substantially higher long-term historical volatility estimates of around 35%. The substantial difference clearly has large potential balance-sheet implications.

5.3.4 In addition, section 3.4 suggests using a scaling factor of between 1,204 and 1,244 rather than 1.2. While this may seem a fairly trivial difference in comparison with the difference in historical volatility estimates, use of the upper bound of the scaling factor would increase the long-term volatility estimate by one percentage point, and would further affect scenario analysis and stress-testing ranges. Given that long-term volatility estimation is so important in the valuation of embedded investment guarantees, best estimation practice should be followed as a matter of course, even if this only means a change of one percentage point in the volatility.

5.3.5 Implementation of the BH Model

5.3.5.1 The authors implemented the BH model using market-volatility quotes obtained from three market makers given at quarterly intervals up to a year, and subsequently for two-, three- and five-year terms. As per the original BH calibration note, the model is initially calibrated to the volatility quotes ('basic' calibration). It is then calibrated including a 15-year dummy volatility point of 26% ('15-year' calibration) and, finally, using 26% as the σ_∞ parameter ('standard' calibration). This last calibration method is the standard BH method specified in the 2006 technical note. Figure 12 displays the BH model under the three different calibration methods. The black line represents the term structure calibrated under the standard BH calibration process. Given current market quotes, there is little difference between the standard and 15-year-dummy calibration methods.

5.4 THE SAFEX MODEL

5.4.1 An alternative TVDV model for the Top40 is that used by Safex, updated fortnightly. Based on the work of Kotzé & Joseph,[12] Safex implemented a

12 Kotzé & Joseph, supra

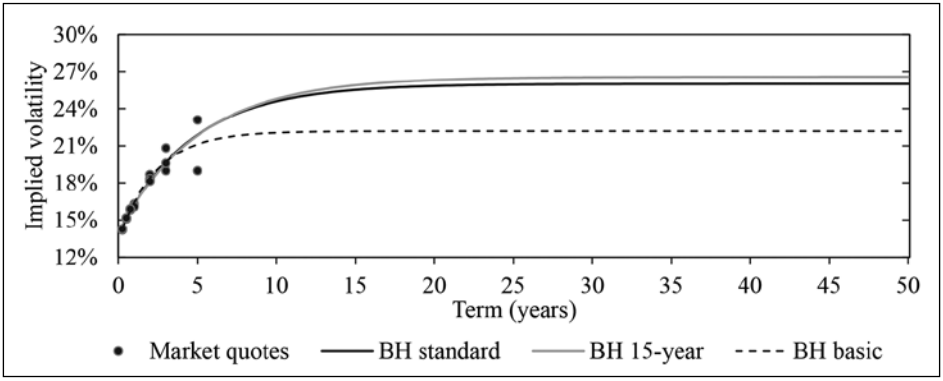

Figure 12. BH volatility term structures calibrated to March 2013 market volatilities

deterministic volatility model in which each (short-term) Top40 option maturity is modelled separately by a quadratic function. These maturity-specific curves are then linked across term in order to create a complete volatility surface. This is done by fitting an inverse power function to the estimated at-the-money (atm) volatility term structure. The final arbitrage-free surface is then a combination of the modelled volatility term structure and the floating volatility skews modelled from each quadratic function. Mathematically, this process can be described as follows:

$$\begin{aligned} \sigma_{\tau,K}^{\text{float}} &= \sigma_{\tau,K}^{\text{model}} - \sigma_{\tau}^{\text{atm}} \\ \sigma_{\tau,K}^{\text{model}} &= \beta_{0,\tau} + \beta_{1,\tau}K + \beta_{2,\tau}K^2 \\ \sigma_{\tau}^{\text{atm}} &= \beta_{0,\tau} + \beta_{1,\tau} + \beta_{2,\tau} \end{aligned} \tag{18}$$

$$\tilde{\sigma}_{\tau}^{\text{atm}} = \theta / \tau^{\lambda} \tag{19}$$

$$\sigma_{\tau,K}^{\text{surf}} = \sigma_{\tau,K}^{\text{float}} + \tilde{\sigma}_{\tau}^{\text{atm}} \tag{20}$$

In these equations, τ is the time to maturity in months, K is option moneyness and the parameter set $\left(\beta_{0,\tau}, \beta_{1,\tau}, \beta_{2,\tau}\right)$ control the shift, slope and curvature characteristics of each volatility curve respectively.

5.4.2 The term structure function given in equation (19) was initially postulated by Gatheral (op. cit.) as a deterministic counterpart to the discrete Heston (op. cit.) stochastic differential equation. In this vein, θ controls the short-term curvature whilst λ controls the slope of the term structure. Equation (19) can be used direct with current market quotes to estimate a volatility term structure in a similar manner to the BH implementation given in section 5.3.4. Unreported results of such a study lead to comparable findings. One interesting point to note is that the Safex term structure function produces a curve that tends to the long-term boundary at a slower rate. The Safex term structure still shows material curvature far beyond that given by the BH term structure, which generally flattens out between 15 and 20 years.

5.4.3 Kotzé & Joseph[13] show that equation (19) fits the term structure well. Furthermore, they give evidence that this functional form is a viable model for each β_i parameter over time. In this manner, Kotzé et al. (2013) showed that one can fully characterise an implied volatility surface using only six parameters:

$$\sigma_{\tau,K}^{surf} = \frac{\theta_0}{\tau^{\lambda_0}} + \frac{\theta_1}{\tau^{\lambda_1}} K + \frac{\theta_2}{\tau^{\lambda_2}} K^2, \; where \frac{\theta_0}{\tau^{\lambda_0}} > 0. \tag{21}$$

5.4.4 Figure 13 displays the long-term implied volatility surface calculated from equation (21) and the published Safex parameters as at 19 March 2013. The condition of no arbitrage is guaranteed within the construction process. See Kotzé & Joseph[14] and Kotzé et al. (2013) for full implementation details. The 50-year at-the-money implied volatility is 27,54%, the 50-year volatility curve ranging between 28,27% and 26,82%. The comparative 30-year values are 25,76%, and 26,79% to 24,76% respectively. Both term-structure point estimates and volatility curve ranges appear reasonable.

5.4.5 It can be argued that the Safex implied volatility surface given in equation (21) is the most market-consistent of all estimated surfaces, given that mark-to-market values of both vanilla and exotic options are calculated from this surface. However, one must realise that the Safex volatility model was constructed for explicitly modelling the short-term implied volatility surface. During the calibration, no preference is given to any specific long-term volatility estimate. Thus the estimated long-term volatility term structure can move substantially in a fairly short length of time. Figure 14 illustrates exactly this feature by plotting changes in the Safex 30-year term structure since December 2009. Within a period of three months, 30-year estimated volatility (shown in bold face) can change by as much as 7%, depending on changes in the short-

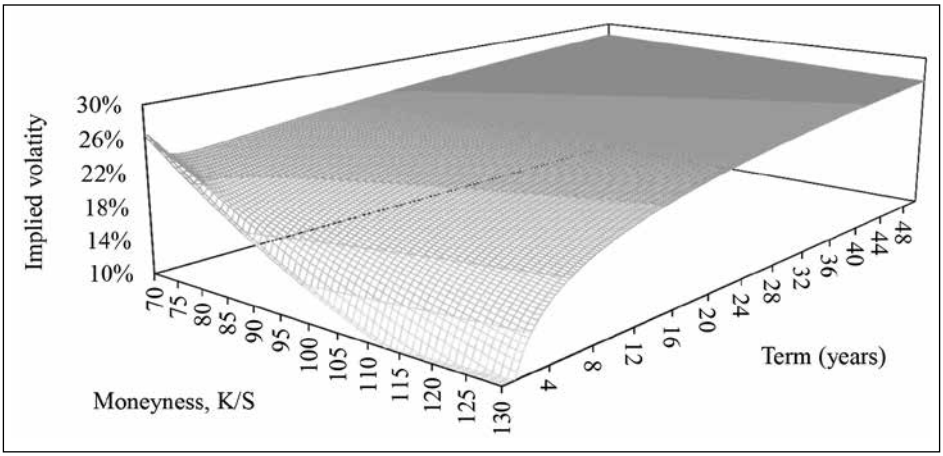

Figure 13. Safex Top40 implied volatility surface at 19 March 2013

13 Kotzé & Joseph, supra
14 supra

term options market. Over the full three-year period, long-term volatility itself shows high volatility, ranging between 20,66% and 35,41%.

5.4.6 The use of the direct Safex volatility surface parameters is therefore not viable. A better method for incorporating the changing Safex volatility surface would be to blend the current short-term Safex term structure with the average long-term Safex volatility term structure. A suitable exchange point is the five-year mark as this is generally the term limit on volatility quotes obtainable in the market. This builds on the general idea of Monte Carlo simulation, which approximates the expectation of a random variable by calculating the discrete average of numerous simulated outcomes or paths. In this instance, the random variable is the unobservable long-term volatility term structure and the historical Safex volatility surfaces are the simulated paths.

5.4.7 Figure 15 illustrates the mean volatility term structure using the Safex volatilities given in Figure 13. The median, minimum and maximum values are also displayed. Both the sample mean and median 50-year volatility estimates appear reasonable at 28,51% and 27,10% respectively. We also note that the mean and median term

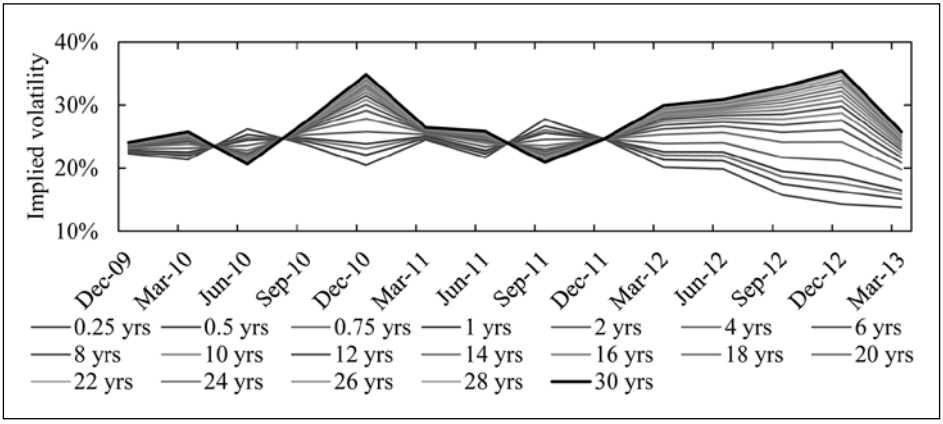

Figure 14. Safex volatility term structures from December 2009 to March 2013

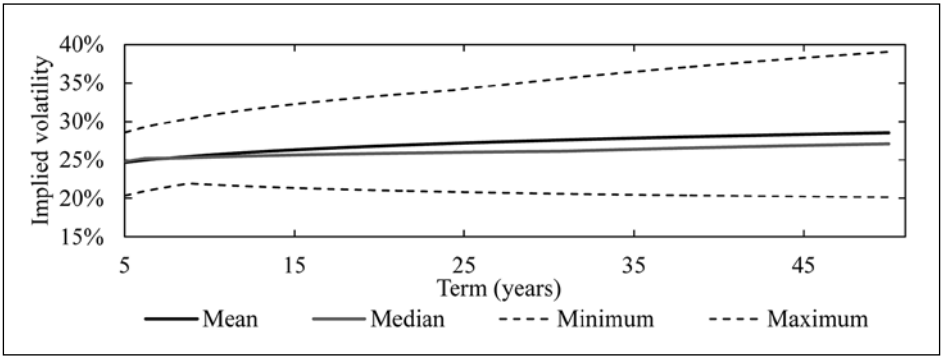

Figure 15. Average Safex volatility term structure since December 2009

structures are robust to outliers and can easily be further refined by considering more sophisticated weighting schemes.

5.4.8 A Viable Market-Consistent Long-Term Volatility Surface

5.4.8.1 Using equations (18) to (20) and the ideas outlined above, one can directly construct an implied volatility surface that is consistent with the current mark-to-market volatility surface at the short end, and which also provides reasonable and stable volatility estimates at the long end. The implementation issues and modelling complexity inherent in the IVHV method is neatly sidestepped. Moreover, a complete volatility surface is given rather than simply a volatility term structure. This surface has the added benefits of being arbitrage-free by construction, continuous and fully parameterised. These last two points are particularly useful if one wants to calibrate, say, a local volatility model to the implied volatility surface and to value exotic derivative structures.

5.4.8.2 A simple method to construct a viable market-consistent long-term volatility surface direct from Safex data—or any suitable TVDV model for that matter—is as follows:

(1) Compute the (weighted) average Safex volatility term structure, $\overline{\sigma^{\text{atm}}_{T,\tau}}$, using the published historical volatility parameter datasets, sampled quarterly:

$$\overline{\sigma^{\text{atm}}_{T,\tau}} = \sum_{t=1}^{T} w_{t,\tau}\tilde{\sigma}^{\text{atm}}_{t,\tau}. \tag{22}$$

(2) Using the most recent Safex parameter dataset, calibrate the volatility term structure, $\tilde{\sigma}^{atm}_{T,\tau}$, as per the usual Safex methodology but now including an X-year volatility fixed or dummy point, where $5 \leq X \leq 10$. The choice of X and type of point used allows one to optimise the short-term fit of the volatility term structure as well as the smoothness of the blended current-to-average term structures;

(3) Calculate the most recent floating volatility curves, $\sigma^{\text{float}}_{T,\tau,K}$, from equation (18), where $\sigma^{\text{model}}_{T,\tau,K}$ is calculated by means of equation (21).

(4) Construct a market-consistent long-term volatility surface from the floating volatility curves in step 3 and the blended term structure in step 2, using equation (20).

5.4.8.3 Quarterly sampling in step 2 helps avoid unnecessary effects on long-term estimates from short-term microstructure noise and also reduces autocorrelation in the sampled volatility-surface time series.

6. CONTINUOUS-TIME STOCHASTIC VOLATILITY MODELLING

6.1 MOTIVATING STOCHASTIC VOLATILITY

6.1.1 An alternative to the TVDV models given in section 5 is stochastic volatility (SV) models. A useful reference for SV modelling—and volatility modelling in general—is Gatheral (op. cit.). Sections 6.1 and 6.2 follow closely from that work. Within the SV family, both the underlying asset returns and volatility itself are considered to be random variables. Because of this assumption, SV models are able to explain why

volatility is a function of option strike and term to maturity in a self-consistent manner. From a practical perspective, SV models allow one to value exotic, path-dependent options more accurately because the dynamics of the volatility surface—the volatility smile—are embedded within the stochastic volatility process. Gatheral (op. cit.) notes that volatility is almost always modelled as a mean-reverting process. A simple rationalisation is that in the long-term, volatility cannot be negative, nor is it likely that volatility will be above 100%. Hence, mean reversion of volatility is established by necessity. Following from these observations, a general SV model is given by:

$$\begin{aligned} dS_t &= \mu_S\left(S_t, v_t, t\right)dt + \sqrt{v_t\left(S_t, v_t, t\right)}dZ_{1,t} \\ dv_t &= \alpha\left(S_t, v_t, t\right)dt + \eta\beta\left(S_t, v_t, t\right)dZ_{2t} \\ \mathbb{E}\left[dZ_{1,t}, dZ_{2,t}\right] &= \rho dt \end{aligned} \tag{23}$$

where S_t is the underlying asset price, μ_S is the instantaneous drift of the asset returns, v_t is the share-price variance, η is the volatility of volatility, ρ is the correlation between asset returns and changes in variance, and $dZ_{i,t}$ are Weiner processes. The functions $\alpha(\bullet)$ and $\beta(\bullet)$ control the variance dynamics and are left in general form for now.

6.1.2 Since the mid-1990s there has been a proliferation of SV models, each with different functional forms of $\alpha(\bullet)$ and $\beta(\bullet)$. The dynamics of the implied volatility surface are thus dependent on one's choice of SV model, with alternative models favoured in each asset class. The shape of the implied volatility surfaces generated from an SV model is not particularly dependent on the choice of model. That said, SV models provide a reasonable fit to the market-implied volatility surface—very short-term expirations are generally poorly fitted because continuous diffusion is unable to produce sufficient slope—and empirically display reasonably stable fitted parameters over time (Gatheral, op. cit.).

6.1.3 The phrase 'continuous-time' is attached to this section because, in truth, discrete-time SV models are discussed at length in section 4 under the more common moniker of ARCH and GARCH. Although GARCH models do describe the features of the joint asset and volatility processes in a simple and insightful manner, they do not—in general—directly address the challenges of pricing and hedging derivatives. In contrast, continuous-time SV models are able to do exactly that.

6.1.4 One of the most commonly used SV models is the specification given by Heston (op. cit.). For the reasons outlined in ¶6.1.2 and for the sake of brevity, this paper provides an empirical analysis based on the Heston model alone and that analysis is followed with a more general discussion about the potential advantages of extended SV models. As always, the analysis and discussion are based on a market-consistent, long-term viewpoint.

6.2 THE HESTON STOCHASTIC VOLATILITY MODEL

6.2.1 Since inception, the Heston (op. cit.) model has been the prevailing SV model of choice for the equity space, particularly within the South African market. Given

this prevalence, it is important to analyse whether the model provides reasonable estimates for long-term volatility. Although not especially realistic in terms of the dynamics of the variance process—a feature shared by a number of stochastic volatility models—its wide appeal is that it admits a quasi-closed-form solution for vanilla option pricing. This makes the Heston model computationally far more efficient than the majority of other SV model candidates. Further, according to Gatheral (op. cit.), in a world governed by the need for fast and efficient pricing of exotic derivatives under Monte Carlo methods, this feature is a prime reason for its continuing prevalence. This section gives a brief outline of the Heston model and of the role that each parameter plays before moving on to an analysis of long-term implied volatility surfaces calibrated to the South African market since December 2009.

6.2.2 Fundamental Theory of the Heston Model

6.2.2.1 Using the notation of equation (23), the Heston model is given as

$$\begin{aligned} dS_t &= \mu_t S_t dt + \sqrt{v_t} S_t dZ_{1,t} \\ dv_t &= \kappa(\theta - v_t)dt + \eta\sqrt{v_t} dZ_{2t} \\ \mathbb{E}\left[dZ_{1,t}, dZ_{2,t}\right] &= \rho dt, \end{aligned} \tag{24}$$

where κ, θ and η are strictly positive. Each parameter in the volatility stochastic differential equation above has an intuitive interpretation and effect on the overall surface:

- κ determines the speed of mean reversion, is largely responsible for the volatility term structure and also dampens any skew at longer terms.
- θ is the mean-reversion level and determines the long-term volatility that the surface will tend towards.
- η is the volatility of volatility, which adds convexity to the surface. This parameter is normally quite sizeable in order to accurately fit the market surface.
- ρ is the correlation between change in volatility and asset return and determines the short-term volatility skew. Normally, one needs $\rho < -0{,}7$ to accurately fit the short-term equity market volatility curve.

6.2.2.2 In order for the variance process to be greater than zero, one must satisfy the Feller condition $\kappa\theta > \frac{1}{2}\eta^2$. However, as noted by Jäckel,[15] this condition is often not satisfied for market-calibrated parameters. Thus, the Heston model imposes dynamics whereby volatility can (1) reach zero and stay there for a long period, and (2) stay very high or very low for long periods of time. Because of these problems, a great deal of research has gone into the creation of efficient and robust simulation algorithms for the Heston model—see, for example, Andersen.[16]

15 P Jäckel (unpublished). Stochastic Volatility Models: Past, Present & Future. Working Paper, 2008. Available at http://bfi.cl/papers/

16 L Andersen (unpublished). Efficient Simulation of the Heston Stochastic Volatility Model. Working Paper, 2007. Available from SSRN

6.2.2.3 One needs to be able to trade both the underlying asset and options of equal or longer term to the instrument in question in order to continuously hedge the specified exposure. In practice, this is not usually possible, especially for long-term instruments and thus one is actually then operating in an incomplete market. This usually leads to an optimal pure equity hedge ratio that is less than that given in the Black–Scholes (1973) framework. Whilst not directly relevant to the problem at hand, these factors may become relevant. That depends on how the model, and its latent long-term volatility estimate, is ultimately used.

6.2.3 Empirical Analysis of Heston-Implied Volatility Surfaces

6.2.3.1 The Heston model is calibrated to the observed South African volatility surface at each close-out maturity date back to December 2009. Parameters are calibrated by minimising the squared option pricing error using the GRG nonlinear algorithm within Excel's 'solver' add-in. The advantage of using this algorithm over, say, the commonly used Nelder–Mead simplex method, is that the GRG nonlinear method can directly accommodate constraints. A long-run variance-constrained calibration, where $\theta \equiv 0,26^2$, is compared with an unconstrained base case.

6.2.3.2 According to Jäckel,[17] calibrated Heston parameters tend to be stable over time. However, as shown in Figure 16 below, this is not really true for either the constrained (solid lines) or unconstrained parameters (dotted lines).

6.2.3.3 Of the four model parameters, only ρ shows high consistency both over time and between constrained and unconstrained cases. This is to be expected because the short-term market skew remains fairly constant over time and high $|\rho|$ values are an SV model's only mechanism for matching this empirical fact. Contrastingly, mean-reversion speed, κ, and volatility of volatility, η, show the largest deviations over time for both constrained cases. In particular, notice how high κ is pushed by imposing the constraint on the long-run variance. This is because when θ is fixed, the only parameter that allows the term structure to vary is the speed of mean reversion. In certain cases, this becomes unrealistically high in order to accommodate the short-term market surface. Finally, notice the extreme differences between the constrained and unconstrained long-term volatility, $\sqrt{\theta}$, over the three-year period. In general, one must always be cognisant of the severe effects that a constraint on long-run model variance has on the remaining model parameters.

6.2.3.4 Figure 17 gives the Heston-implied volatility term structure under the two calibrations. The instability in the parameters is clearly apparent in the short-term volatility differences. When one constrains the long-run variance though, notice how similar the models are at longer terms. Irrespective of the underlying short-term market surface, the 30-year constrained implied volatility lies between 23% and 25% and the term structure beyond the 10-year mark is remarkably similar. In contrast, both the ending points and curvature of the unconstrained term structures show significant variability. The 30-year volatility ranges between 22,1% and 37,8%.

17 Jäckel, supra

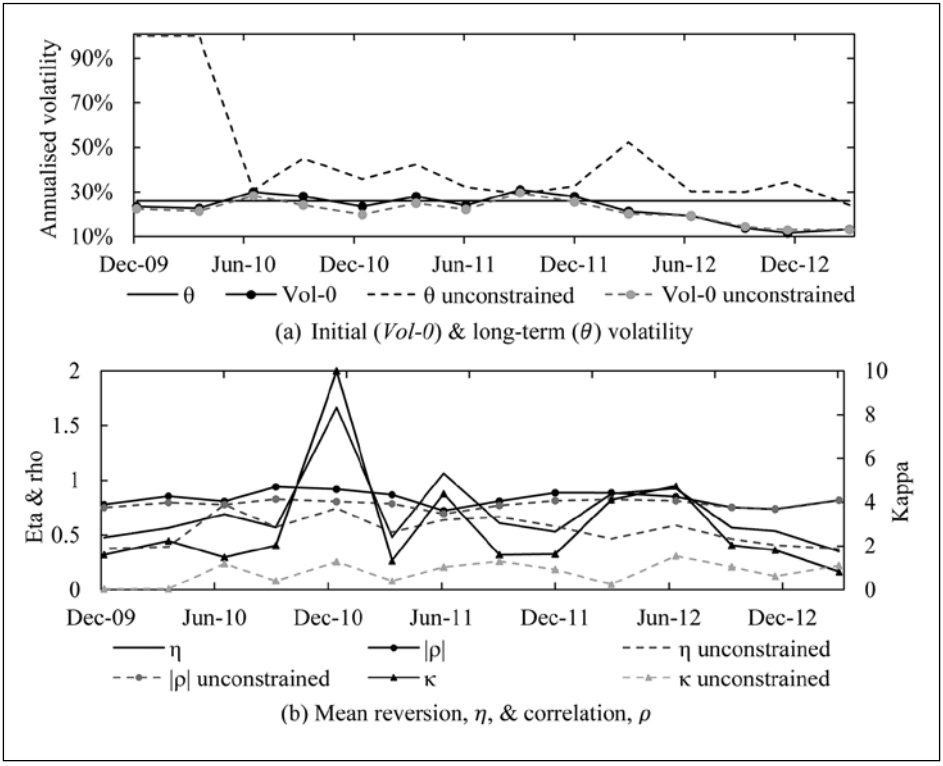

Figure 16. Constrained (solid) and unconstrained (dotted) Heston model parameters over time

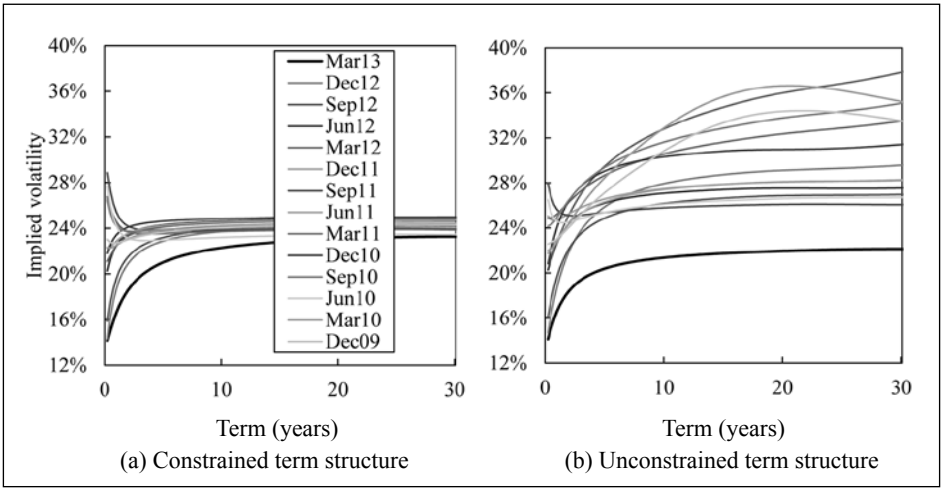

Figure 17. Heston model 30-year volatility term structures since December 2009

6.2.3.5 The unconstrained Heston term structures are actually quite similar to those calculated from the Safex TVDV model. This is not surprising, given that the Safex model uses a Heston-inspired function to fit the volatility term structure. This similarity suggests that the additional step of fitting a Heston model to a market surface has little marginal benefit over using the Safex model direct. We stress though, that this finding is strictly applicable only for the Safex and Heston models. Furthermore, in comparison with the Heston model, the Safex TVDV model arguably provides equal or better tractability and computational efficiency under simulation, finite-difference or tree-pricing methods.

6.3 EXTENDING STOCHASTIC VOLATILITY MODELS

6.3.1 JUMP DIFFUSION, SVJD AND SLV MODELS

6.3.1.1 Dupire[18] points out that there are two possible mechanisms for obtaining negative equity volatility skews:

(1) Model the negative relationship between the underlying asset price and volatility, either in the form of a deterministic dependence (TVDV or local volatility models) or as a negative correlation (SV models); the greater the dependence or correlation, the greater the negative skew.

(2) Model the discontinuity of asset prices by including jumps in the underlying asset process; a higher jump frequency and more probable downward jumps increase negative skew.

6.3.1.2 One of the problems with SV models is that they are unable to produce a negative skew great enough to fit the short-term market volatilities. This is an issue if one is trying to obtain a market-consistent, long-term volatility estimate, as the definition of market-consistency requires the specified stochastic model to accurately replicate short-term traded option prices. In order to ensure market consistency, one must then include jumps in the asset process as well as the standard continuous diffusion. Merton (1976) laid the foundation for these 'jump-diffusion' (JD) models by including jumps as an independent Poisson process with log-normally distributed jump size to the common Black–Scholes asset process. This neatly accounted for discontinuous stock prices and uncertain jump size whilst still maintaining a high level of tractability. The effect on the volatility surface is that one can now create a large negative skew at the short-term. However, this effect rapidly disappears with term as the aggregated diffusion volatility quickly overwhelms any effect from asset jumps.

6.3.1.3 The next obvious step was to link stochastic volatility with jump diffusion models. Such a model would be able to accurately capture the short-term skew and also account for the longer-term dynamics of the surface. Thus, stochastic volatility jump diffusion (SVJD) models were born, the Bates (1996) specification—a combination of the Heston and JD models—being the most ubiquitous in practice. However, as Gatheral (op. cit.) notes, SV and SVJD models essentially differ only at

18 Dupire (unpublished b), supra

very short terms—making the distinction in any long-term exercise fairly trivial—and the independence of asset jumps to volatility gives the counter-intuitive result that volatility remains constant following a jump. Therefore, the additional short-term market consistency obtained from the additional asset jumps has little effect on the estimated long-term volatility parameter.

6.3.1.4 That said, it would appear from empirical results (Andersen & Andreason, 2000; Duffie, Pan & Singleton, 2000; Gatheral, op. cit.) that the SVJD model fits the data better than most pure SV models and, importantly, this additional accuracy is not overly expensive in terms of tractability.

6.3.1.5 A recent paper by Manistre (2010) considers a special case of Merton's JD model derived under a cost-of-capital-inspired $\mathbb{C}$-measure rather than the usual risk-neutral $\mathbb{Q}$-measure. Manistre uses this model to "derive a long-term implied volatility assumption from first principles". Whilst novel in its derivation, the final model put forward is essentially an extended JD or SVJD model that explicitly allows for parameter shocks in the governing asset or asset-volatility processes. It does not strictly help one set a long-term volatility estimate. Rather, it enables one "to defend a long-term implied volatility assumption" by deconstructing the specified estimate into several cost-of-capital-inspired parameters. One is then able to assess the estimate's reasonability by somehow analysing these underlying parameters.

6.3.1.6 A final extension to the basic SV model is the stochastic local volatility (SLV) model class, widely used in foreign-exchange markets.[19] The possibility of jump processes is included in the SLV model class.[20] By incorporating features of local, stochastic and jump models one has the flexibility needed not only to calibrate to a market volatility surface, but also to accurately capture the correct surface dynamics. However, it remains difficult to set the long-run volatility parameter.

6.3.1.7 In summary, basic and extended SV models allow one to capture more and more of the empirical features of the volatility surface. However, what must always be remembered is that by picking a certain model, one is latently constraining the possible dynamics of the volatility surface. This is different from merely fitting a set of vanilla options maturing at a specific time. True modelling of the surface dynamics would require calibration to all existing derivative contracts, including path-dependent exotic derivatives. Secondly, for such long terms, the actual model specification becomes of secondary concern when one imposes a fixed long-term variance parameter. Thirdly, SV and extended SV models are not ideal candidates for estimating this long-run parameter. To paraphrase Rebonato (2004), one can define this problem as putting "the wrong parameters in the wrong *SV* formula to obtain the right price of plain vanilla options."

19 See, for example, C Alexander & L Nogueira (unpublished). Stochastic local volatility. Working paper, 2008. Available at SSRN 1107685

20 See, for example, Lipton (2002)

7. NONPARAMETRIC BREAK-EVEN VOLATILITY

7.1 INTRODUCING NONPARAMETRIC PRICING METHODS

7.1.1 Sections 2 to 6 highlight several different parametric classes of long-term volatility candidate models that can be calibrated to the market through a combination of sophisticated underlying process dynamics or distributional assumptions. As shown, this can lead to a practitioner (1) imposing material constraints on the underlying return distribution and (1) inferring the incorrect underlying dynamics because of the calibration process. Possibly a more fundamental approach is rather to ask: What should the implied volatility surface be, given only a history of underlying market data? Or in statistical parlance: Is there a nonparametric method capable of obtaining market-consistent implied volatility surfaces? In this section (section 7), Dupire's[21] break-even volatility is considered, while section 8 introduces Stutzer's (op. cit.) canonical valuation.

7.1.2 It is well-known that implied volatility on equity indices consists of:

(1) a theoretical curve, based on the market's expectation of future asset behaviour; and

(2) a supply-demand curve, based on the current microstructure noise and broader systemic make-up of the market.

7.1.3 Element (2) indicates the fluctuating risk premium and is strongly influenced by trading behaviour. On the other hand, (1) truly reflects the fair value of a traded option. Nonparametric pricing methods are directly focused on (1)—although (2) can easily be accommodated—and are thus indicative of the fair volatility surface. In this manner, nonparametric methods allow one to calculate market-consistent, fair surfaces for any underlying security—single counter or basket—that has a price history, irrespective of whether option information is available. This final feature makes nonparametric methods an ideal candidate for estimating market-consistent long-term volatility parameters.

7.1.4 The sole use of historical market data has several advantages. From a mathematical standpoint, the smoothness assumptions usually required by kernel-smoothed empirical distributions are not required. From a financial standpoint, the historical return distribution is a rich source of information, latently incorporating the stylised facts described in section 4.1. In addition, one easily incorporates stochastic interest rates and dividends, as well as multiple underlying assets.

7.2 INTRODUCING BREAK-EVEN VOLATILITY

7.2.1 Break-even volatility (BEV) is simply defined as the volatility level that gives a zero profit and loss for a delta-hedged option. It stems from the fact that option pricing is built around the concept of dynamic replication. Using small enough time steps, a portfolio of the underlying asset and cash can be made to replicate an option with arbitrary closeness conditionally on using the correct delta. Empirically, arbitrary closeness is not possible and so one obtains a profit and loss function, which is

21 Dupire (unpublished a), supra

dependent—amongst other variables—on the chosen volatility, $P\&L\left(\sigma_{imp}\right)$. Critically, this function always has a unique strictly positive root, or BEV.

7.2.2 Following standard Black–Scholes theory but in a discrete setting, the profit and loss of a delta-hedged option expiring at time T can be written:

$$P\&L\left(\sigma_{imp},t,T\right)=\frac{1}{2}\sum_{t=1}^{T}e^{-r_{t,T-t}(T-t)}\Gamma_t S_t^2\left(r_{t,1}^2-\sigma_{imp}^2\right)\Delta_t, \tag{26}$$

where Γ_t is the gamma of the option at time t and Δ_t is the annualised time-step. Equation (26) shows that BEV is essentially the gamma-weighted average of the quadratic return. Moreover, because gamma is a function of term and option strike, one can actually extract an entire volatility curve from a single historical path.

7.2.3 The BEV algorithm based on equation (26) is fairly simple to implement in practice. However, it can be challenging to find a smooth surface. Firstly, because of the circular dependence on implied volatility, BEV must be found by iterating through a fixed-point algorithm. Secondly, one must consider how to aggregate over time. According to Dupire,[22] one obtains a smoother surface if one solves for the implied volatility that cancels the average $P\&L$ over the different time periods, rather than taking the average of the volatilities that cancel the $P\&L$ within each time period. Thirdly, a moneyness framework guarantees that each time period can be used equivalently irrespective of absolute price-level changes over time. Finally, Dupire[23] notes that interest rates tend to have little effect on the resulting BEV surfaces. In this paper, for the sake of completeness, interest rates have been included in all calculations.

7.2.4 Given that BEV is calculated by re-weighting daily return volatility, as proxied by squared returns, and that, across all strikes at a specific term, gamma is equal to one, the corresponding historical volatility is actually equivalent to the average of the BEVs across all strike levels. Thus, the imputed volatility curve at each maturity is essentially dependent on the path that the returns take across each strike-specific gamma surface.

7.2.5 As with all methods, there are several caveats of which to be aware. Firstly, this method is data-intensive and requires a large amount of data for convergence. Secondly, the surface obtained is not where the market should be trading. The BEV surface solves for zero $P\&L$, which means that it assumes no volatility risk premium for any option writer. That is evidently false in practice. However, by purposefully excluding supply and demand of microstructure noise, one can get closer to a theoretically fair volatility surface. Should one wish, the complete risk-premium surface can then be measured as the spread between fair and market volatilities.

7.3 SOUTH AFRICAN BREAK-EVEN VOLATILITY SURFACES

7.3.1 Using the return series $\left\{r_{t,1}^{FTF}\right\}$ calculated from the 1995 daily dataset sampled at monthly intervals, a fair BEV surface was constructed across an 80–120 moneyness range and out to a term of 10 years. Because of computational time constraints

22 Dupire (unpublished a), supra
23 supra

the smaller surface is given here. Given that the BEV method is directly linked with dynamic replication, it is necessary to have at least daily data available. This obviously limits the maximum volatility term unless one considers bootstrapping methods to create a longer daily price history. Figure 18 displays the complete BEV surface, while Figure 19 gives the corresponding volatility term structure.

7.3.2 The short-term BEV curves show large negative slopes for the 80–105 moneyness range before noticeably flattening out and gently sloping up. Known as the volatility 'smirk', this pattern is a common empirical finding in short-term equity markets worldwide. The surface tends to flatten rather quickly across term, largely flat from the 4-year mark onwards. Even though there has been no calibration, the produced surface seems quite reasonable across all terms.

7.3.3 Considering its calculation method, BEV is most comparable with the average realised volatility of the FTF return series. While there are some similarities between the 1995 FTF and BEV volatilities, it is rather their differences that catch one's attention. BEV shows a much more gradual increase. In contrast, the 1995 FTF volatility term structure is upwards-trending, ending at a volatility around 32%. The BEV term

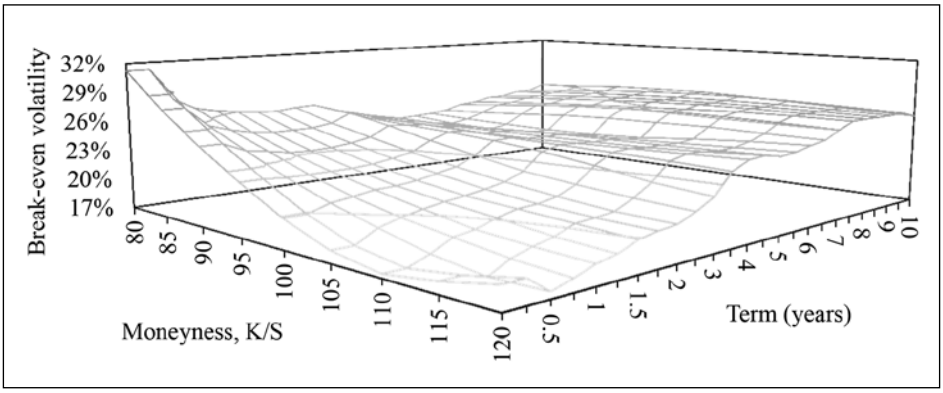

Figure 18. Fair BEV surface from the 1995 daily FTF Top40 dataset

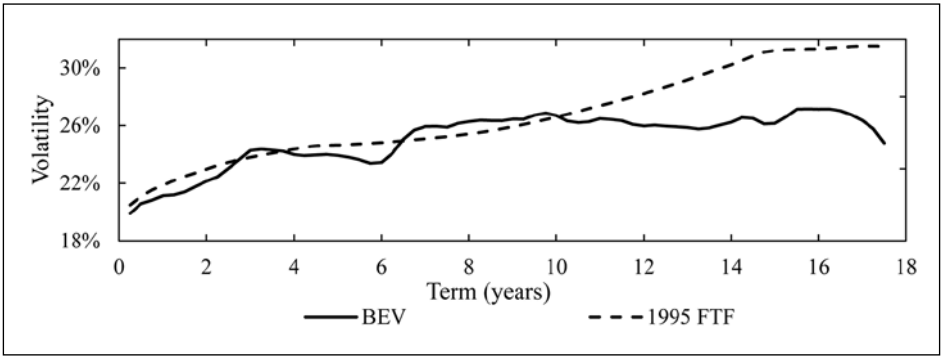

Figure 19. Fair BEV term structure versus the 1995 daily FTF volatility term structure

structure is also more uneven; a characteristic indicative of discrete hedging and also the small number of data points, particular beyond 15 years. Furthermore, Dupire[24] notes that the BEV approach was not specifically designed to create a volatility term structure because the average volatility at each maturity is simply equal to the historical volatility of the underlying return series.

7.3.4 This gives one direct insight into the link between the empirical return distribution and the option price. Consider a series of fixed-strike volatility lines over the term of the BEV surface. The difference between each of these lines and the average or historical volatility term structure is exactly caused by the variations in empirical return distribution across term, coupled with a deterministic mapping function (the gamma surface) that translates these distributional variations into corresponding option equivalent variations, specific to term and strike. Furthermore, this is all done consistently with arguments about dynamic replication arguments. In a market where long-term options are unavailable and synthesis by some form of replication is commonplace, this makes the BEV surface an ideal candidate for pricing or hedging

8. NONPARAMETRIC CANONICAL OPTION VALUATION

8.1 INTRODUCING CANONICAL VALUATION

8.1.1 An alternative, nonparametric approach is the canonical valuation (CV) method proposed by Stutzer (op. cit.) and further developed by Duan,[25] Alcock & Auerswald (2010), and Haley & Walker (2010). This pricing technique uses only historical market data and thus avoids the necessity of specifying underlying return dynamics. Stutzer normalised the historical return distribution via the principle of minimum relative entropy in order to find a risk-neutral option price. Entropy is a well-established concept in information theory and statistical mechanics, and is used extensively in a wide range of scientific fields. See section 8.2 for more detail. This method is very robust and can be easily altered to include multiple underlyings, empirical option-price data and early exercise for American options. Alcock & Gray (2005) extended Stutzer's (op. cit.) original work by developing the theory for a nonparametric, dynamic delta-hedging portfolio, which provided investors and traders with a tractable nonparametric valuation framework for European vanilla and basket options.

8.1.2 A large body of work has evaluated the pricing accuracy of CV relative to Black–Scholes under several different volatility regimes. Duan,[26] Gray & Newman (2005) and Alcock & Auerswald (op. cit.) show that the CV method performs arguably as well as the BS formula with an historical volatility input under a pure Black–Scholes framework. More importantly however, they show that under stochastic volatility, the nonparametric valuation method performs significantly better across the board and especially so for out-of-the-money options, which are notoriously difficult to value.

24 Dupire (unpublished a), supra

25 JC Duan (unpublished). Nonparametric option pricing by transformation. Working Paper, Rotman School of Management, University of Toronto, 2002

26 supra

8.1.3 Whilst several other nonparametric pricing approaches have been proposed, Alcock & Carmichael (2008) note that the majority of these approaches rely heavily on existing option prices.[27] In reality, these 'nonparametric' methods should be viewed more as numerical interpolation algorithms rather than true nonparametric option valuation theories.

8.1.4 CV pricing has been applied in several different areas. Zou & Derman[28] introduced the notion of strike-adjusted spread (SAS), defined as the spread between the observed BS-implied volatility and the BS-implied volatility imputed from the nonparametric CV option prices. SAS is, in essence, a one-dimensional metric ranking the relative richness of equity options across strike for a fixed option term, measured over time. Cakici & Foster (2001) followed on from this and used CV to price currency options, with encouraging fit statistics.

8.1.5 Cakici & Foster (op. cit.) provided—to the authors' best knowledge—the only case where the term structure of volatility has been evaluated. They used CV prices and the imputed volatility term structure to show that the observed term structure is well explained by their estimated forward distribution, without resorting to explanations based on market imperfections. They further concluded that the assumption of a specific functional form for returns (dynamic or otherwise) would imply severe pricing prejudices.

8.1.6 Another study exploring the link between the CV-implied volatility surface and the market-implied volatility surface is that of De Araujo & Maré (2006). Using the revised CV method proposed by Duan,[29] De Araujo & Maré (op. cit.) conducted a South African study on Top40 index options, which showed that the implied volatility surface obtained from the calculated CV option prices was similar to that implied by the market. Following from this insight, they motivated for the use of the CV method to generate volatility surfaces for illiquid single-stock options.

8.2 OVERVIEW OF CANONICAL VALUATION METHODOLOGY

8.2.1 The theory of option pricing is based on the proposition that, if no arbitrage opportunities exist within a market, there exists a risk-neutral return distribution $\mathbb{Q}$, such that the value $V_{t_i,T}$, of a contingent claim at time t_i on an underlying asset priced S_t is given by the discounted expectation of the payoff. Mathematically, we write

$$V_{t,T}\left(S_T\right)=e^{-r_{t,T}(T-t)}\mathbb{E}_{\mathbb{Q}}\left[f\left(S_T,T\right)|S_t,t\right], \tag{27}$$

where $\mathbb{E}_{\mathbb{Q}}\left[\bullet|\bullet\right]$ is the conditional expectation under the risk-neutral measure $\mathbb{Q}$ and $f\left(S_T,T\right)$ is the payoff function. Thus, if one knows $\mathbb{Q}$, one can calculate the value of

27 See Fengler (2005).

28 J Zou & E Derman (unpublished) Strike-adjusted-spread: a new metric for estimating the value of equity options. Goldman, Sachs Quantitative Strategies Research Notes, 1999. Available at http://ederman.com/new/docs

29 Duan, supra

any European option contract. In the Black–Scholes framework, the log-normal density function with given volatility is assumed to be the implied risk-neutral continuous distribution $\mathbb{Q}$. Other pricing theories assume different underlying distributions. Stutzer (op. cit.) challenged this assumption by considering the case where one does not want to assume a particular continuous-time process. Based on the fundamental option valuation theory above, Stutzer examined the estimation of $\mathbb{Q}$ direct from historical data via the following three-part nonparametric method:

(1) For a statistically relevant period, historical asset returns and risk-free rates are used to estimate the future real-world probability distribution, $\hat{\mathbb{P}}$ of the underlying asset price at time T. In this case, 'statistically relevant' refers to the descriptive statistics of the chosen period, and more specifically, the skewness and kurtosis of the return distribution.

(2) The estimated future real-world distribution is transformed into an estimated future risk-neutral density $\hat{\mathbb{Q}}$ of the equivalent martingale measure $\mathbb{Q}$ through the principle of minimising relative entropy.

(3) The derivative contract is valued by substituting $\hat{\mathbb{Q}}$ in equation (27). Fair CV volatility at time t for strike level K and option term τ, denoted $\sigma_{t,\tau,K}^{CV}$, is then defined as the BS-implied volatility imputed from using the CV option price.

8.2.2 The true difference between Stutzer's method and other return-probability-reweighting schemes was in the use of the relative-entropy divergence measure. For a proper mathematical treatment of CV pricing, please see Appendix B.

8.2.3 Motivating CV: Information, Uncertainty and Entropy

8.2.3.1 Information in financial markets plays an important role in shaping an investor's market view. If one is to believe the efficient markets hypothesis (EMH)—in whichever form—this role is nearly sacrosanct. In essence, asset prices are assumed to be subjective, functional transformations of all incoming admissible information, where 'admissible' is specified by the form of the EMH. The definition of information is derived within the rich scientific field of information theory. Here a simple, pedagogical example is described.

8.2.3.2 Consider an asset price X that either increases with probability p or decreases with probability $1-p$. If we knew a priori that $p=0{,}99$, then we would say that X is almost certain to increase and is thus almost perfectly predictable. Because of this, we learn fairly little when X does, in fact, increase. If however, X actually decreased, then we should gain additional information about the asset price process. Contrarily, if we knew a priori that $p=0{,}50$, then we would have maximal uncertainty about the future value of X, and in this case both an up- and a down-movement should provide us with the same amount of information. Following from these simple intuitions, we can define the information, $I(\bullet)$, obtained from the occurrence of a random event with assumed probability p:

$$I(p) = -\ln(p). \tag{28}$$

8.2.3.3 Stutzer (op. cit.) motivated the use of the uniform distribution for the estimated future real-world asset distribution $\hat{\mathbb{P}}$. Using the general assumption that asset returns are generated by an unknown, ergodic Markov chain, that author noted that the uniform distribution is an optimal nonparametric estimator of the unknown, invariant real-world distribution $\mathbb{P}$, given that its rate of convergence is the fastest among all such consistent estimators. In addition, Zou & Derman[30] provide a financial motivation for using a uniform prior probability distribution based on the fact that markets in equilibrium must, perforce, have equivalent supply and demand. This in turn implies that an equal number of investors think that a stock is both rich and cheap, thus implying that the expected return distribution must display maximum uncertainty.

8.2.3.4 Using equation (28), we define the Shannon–Gibbs–Boltzmann entropy, $H(\bullet)$, of the variable X, whose ith observation has probability p_i, as the expected value of information obtained across all possible observations:

$$H(X) = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{29}$$

8.2.3.5 For our example above, it is simple to show that $H(X)$ is maximised when p is equal to 0,50. Because all probabilities are less than 1, entropy is always positive. Higher expected values of information imply a greater spread of probabilities and thus greater uncertainty within the distribution. Entropy therefore measures the uncertainty of a probability distribution, maximum entropy implying maximum uncertainty within a distribution. In essence, the idea that entropy measures the uncertainty surrounding a series of observations or events corresponds to the idea that probability measures the uncertainty surrounding a single event. Through entropy, one is able to quantify the information gained from changing a distribution. Assume there is a prior probability distribution $\mathbb{P}$ for the random variable X. By incorporating new information, a posterior distribution $\mathbb{Q}$ is formed. By considering the notion of relative entropy, one is able to quantify the reduction in uncertainty. Relative entropy, also known as the Kullback–Leibler divergence and denoted by the function $f(\bullet)$, is the entropy difference between these two distributions:

$$f(\mathbb{P},\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[\ln \mathbb{Q} - \ln \mathbb{P}] = -\sum_{i=1}^{N} q_i \ln\left(\frac{p_i}{q_i}\right) \tag{30}$$

8.2.3.6 From equation (30) we know that $f(\mathbb{P},\mathbb{Q})$ is convex. By using Jensen's equality, Zou & Derman[31] show that $f(\mathbb{P},\mathbb{Q})$ is strictly non-negative and zero only for $\mathbb{P} \equiv \mathbb{Q}$. Using this fact, they motivate that relative entropy can be considered a 'distance' metric between prior and posterior distributions. Stutzer (op. cit.) intimates that by minimising the relative entropy between prior and posterior distributions—in this case the real-world and risk-neutral density estimates—one preserves maximum uncertainty—and thus market equilibrium—under the density transformation.

30 Zou & Derman, supra
31 supra

8.3 SOUTH AFRICAN CANONICAL-VALUATION VOLATILITY SURFACES

8.3.1 We construct two fair CV volatility surfaces using the τ-period return series $\left\{r_{t,\tau}^{CTF}\right\}$ calculated respectively from the 1925 monthly ALSI dataset and the 1995 daily Top40 dataset. The 1925 CV surface extends out to 30 years, while the 1995 daily CV surface extends out to 15 years. The respective term lengths are dictated by the amount of available data. Given that one is estimating the terminal risk-neutral distribution directly, CV volatility is most comparable with CTF terminal volatility. Figures 20 and 21 give the respective CV fair volatility surfaces. All given CV results are based only on the essential risk-neutral constraint rather than including further constraints to ensure calibration to short-term option prices, a straight-forward inclusion if desired.

8.3.2 Similarly to the BEV surfaces, both CV surfaces above have several appealing characteristics. Firstly, the short-term volatility smirk is readily apparent. Secondly, the surface also flattens out across moneyness as term increases, although to a

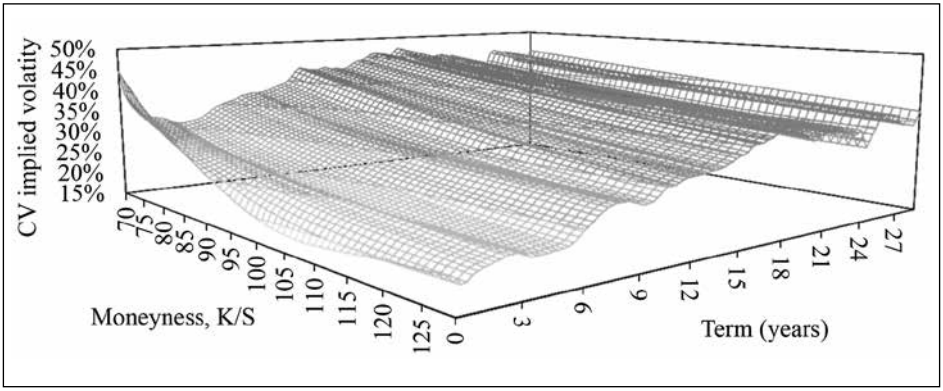

Figure 20. Fair CV volatility surface calculated from the 1925 CTF return datasets

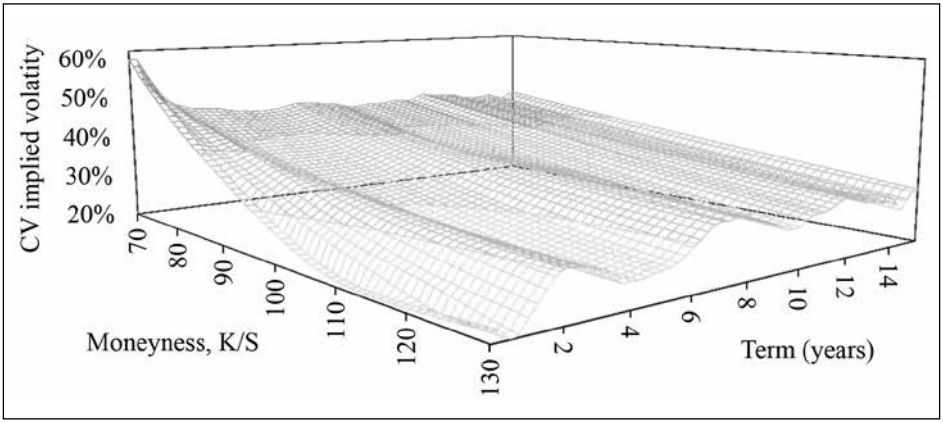

Figure 21. Fair CV volatility surface calculated from the 1995 CTF return datasets

much lesser extent than for the BEV surfaces. Finally, the absolute volatility levels are within what one would consider a reasonable range. Intuitively then, the CV surfaces are appealing candidates as fair estimates of long-term implied volatility surfaces.

8.3.3 Figure 22 displays the 1925 and 1995 CV term structures in comparison with the 1925 historical CTF volatility series. Although there is an absolute difference between the two CV volatility series, both display a remarkably similar pattern across term. The undulations are found at similar terms and are of comparable relative magnitudes.

8.3.4 Figures 20 and 22 suggest a 30-year fair volatility estimate close to 40%, a value generally in line with the comparative CTF historical long-term volatility estimate. The short-term volatility of both series is also fairly comparable although we notice that CV volatility tends to be greater than its CTF counterpart between start and end terms—close to the 1976 CTF volatility series.

9. CONCLUSION

9.1 This paper addresses the problem of accurately estimating long-term equity volatility in a market-consistent manner. This becomes a particularly difficult challenge in a market such as South Africa, where there is a lack of any medium- or long-term traded instruments. APN 110 caters for this by allowing the actuary the freedom to use alternative estimation methods and judgement conditional on some form of market consistency. However, this allowance is general and permits a broad range of models and methods to estimate long-term volatility, some of which are more justified than others.

9.2 According to a recent APN 110 survey, all market participants use a long-term historical volatility estimate as a base proxy for long-term implied volatility. This makes accurate historical volatility estimation extremely important. It is shown that historical volatility is strongly dependent on the function used to measure return variation, the data period used and the sampling frequency chosen. Each choice has a material impact on the final long-term volatility estimate. It is further shown that, for various theoretical and empirical reasons, historical volatility should be estimated as either the terminal

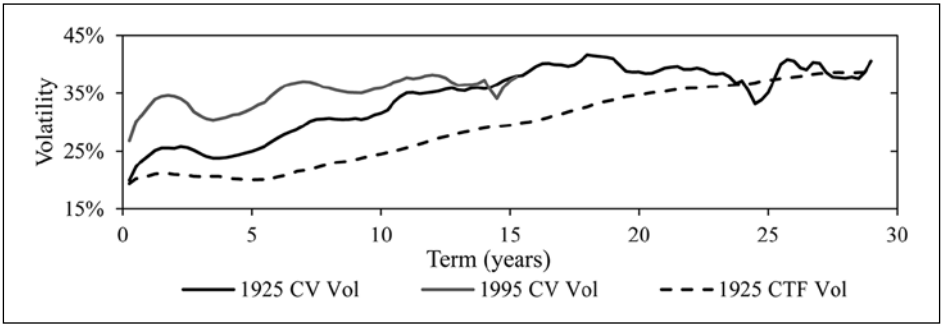

Figure 22. Fair CV volatility term structures in comparison to CTF historical volatility

distribution volatility of the historical constant-term forwards or the average realised volatility of the historical floating-term forwards. This is at odds with the current standard practice of purely considering equity volatility in isolation. When doing so, one finds that long-term historical volatility is materially higher than that obtained from the basic equity estimation method. In particular, compelling evidence is found to suggest that a 25-year historical futures volatility of 35% is not unreasonable.

9.3 Several econometric, deterministic and stochastic volatility models, where the long-term parameter is usually based on the historical volatility estimate, are reviewed and implemented. On the econometric side, the GARCH family of models is reviewed, focusing on several theoretical and practical implementation issues. For the South African market, the GJR-GARCH(1,1) model of daily Top40 returns with errors following a *t*-distribution appears to provide the best in-sample fit. Whilst GARCH models are able to forecast volatility, we stress that these models are not particularly suited for long-term forecasting and are very dependent on model specification and the chosen residual distribution.

9.4 On the deterministic side, two specifications are analysed and implemented; namely, the Barrie & Hibbert model and the Safex model. The Barrie & Hibbert model outputs only a volatility term structure, whereas the Safex model gives an entire volatility surface. From this analysis, the deficiencies latent in the generally implemented TVDV models are highlighted and a simple algorithm based on the historical Safex volatility surfaces is prescribed in order to create a smooth, fully parameterised long-term volatility surface.

9.5 On the stochastic side, the focus is on the Heston model, one of the most common stochastic models used in practice. It is demonstrated that constraint of the long-term volatility parameter has severe effects on the model parameters and essentially outputs equivalent term structures beyond the 10-year mark, irrespective of the short-term market surface. Several extensions of the basic stochastic model are discussed but it is shown that, extended or otherwise, these models should not be used *ex ante* to estimate long-term volatility. Rather, these models provide one with a means of fitting the current vanilla option prices given an existing assumption regarding long-term volatility.

9.6 A couple of recent nonparametric alternatives are introduced and discussed. Rather than impose constraints of the underlying return distribution and the volatility surface dynamics, nonparametric methods answer the question: What should the implied volatility surface be, given a history of underlying market history? These models are market-consistent because they are based on the underlying historical return data but are not influenced by short-term supply and demand factors. This means that nonparametric methods are able to estimate the fair volatility surface. Furthermore, because no options are needed, these methods can be applied to any underlying asset that has historical data. In this paper, we consider break-even volatility, which is the volatility that zeroes the

profit and loss of a delta-hedged position, and canonical-valuation volatility, which uses relative entropy techniques and risk-neutralised historical return distributions to construct an implied volatility surface. Helpfully, break-even volatility is comparable in method with the average realised floating-term forward historical volatility, while canonical-valuation volatility is similar to the terminal distribution constant-term forward historical volatility. In both cases, the constructed volatility surfaces provide compelling *ex ante* market-consistent long-term estimates.

9.7 This contribution provides a first attempt at systematically evaluating those models most commonly used and introduces several alternative models that may offer better solutions. The paper applies these various models and methodologies to South African market data, thus providing practical, long-term volatility estimates under each modelling framework whilst accounting for real-world difficulties and constraints. In so doing, the authors identify those models and methodologies they believe to be most suited to long-term volatility estimation and propose best estimation practices within each identified area. There is both substantial scope and a significant need for further research in this field. Each type of model reviewed in this paper can—and should—be further researched in the context of market-consistent, long-term estimation.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge insightful discussions with Anthony Seymour and, separately, Dylan Flint, both of whose ideas have influenced several sections of this paper. We are also grateful for thoughts from the participants at the 2013 ASSA Life Assurance Seminar. Finally, our sincere thanks go to the two anonymous referees as well as the SAAJ editorial staff for their many helpful comments and suggestions.

REFERENCES

Alcock, J & Auerswald, D (2010). Empirical tests of canonical nonparametric American option-pricing methods. *Journal of Futures Markets* **30**(6), 509–32

Alcock, J & Carmichael, T (2008). Nonparametric American option pricing. *Journal of Futures Markets* **28**(8), 717–48

Alcock, J & Gray, P (2005). Dynamic, nonparametric hedging of European style contingent claims using canonical valuation. *Journal of Financial Econometrics* **2**(1), 41–50

Alexander, C (2008a). *Market Risk Analysis, Volume I: Quantitative Methods in Finance*. Wiley.

Alexander, C (2008b). *Market Risk Analysis, Volume II: Practical Financial Econometrics*. Wiley.

Alexander, C (2008c). *Market Risk Analysis, Volume III: Pricing, Hedging and Trading Financial Instruments*. Wiley.

Alexander, C (2008d). *Market Risk Analysis, Volume IV: Value at Risk Models*. Wiley.

Andersen, L & Andreasen, J (2000). Jump-diffusion processes: Volatility smile fitting and numerical methods for option pricing. *Review of Derivatives Research* **4**(3), 231–62

Andersen, TG, Bollerslev, T, Christoffersen, P, & Diebold, FX (2006). Volatility and correlation forecasting. In Elliott, Granger & Timmermann (2006: 778–878)

Andersen, TG, Bollerslev, T, Diebold, FX & Labys, P (2003). Modeling and forecasting realized volatility. *Econometrica* **71**(2), 529–626

Bachelier, L (1900). Théorie de la speculation. *Annales Scientifiques de L'École Normale Supérieure* **17**, 21–86. (English translation by AJ Boness in Cootner (1964:17–75)

Barndorff-Nielsen, OE & Shephard, N (2002). Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B* **64**(2), 253–80

Bates, DS (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *Review of Financial Studies* **9**(1), 69–107

Bingham, NH & Kiesel, R (2004). *Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives*, 2nd edn. Springer Finance, Heidelberg

Black, F & Scholes, M (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* **81**(3), 631–59

Bollerslev, T (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**(3), 307–27

Breeden, DT & Litzenberger, RH (1978). Prices of state-contingeht claims implicit in option prices. *Journal of Business* **51**(4), 621–51

Brownlees, C, Engle, R & Kelly, B (2011). A practical guide to volatility forecasting through calm and storm. *The Journal of Risk* **14**(2), 3–22

Cakici, N & Foster, KR (2001). Risk-neutralized at-the-money consistent historical distributions in currency options pricing. *Journal of Computational Finance* **6**(1), 25–47

Cont, R (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* **1**(2), 223–36

Cootner, PH (1964). *The random character of stock market prices.* MIT Press, Cambridge, MA

De Araujo, M & Maré, E (2006). Examining the volatility skew in the South African equity market using risk-neutral historical distributions. *Investment Analysts Journal* 64, 15–20

Ding, Z, Granger, CWJ & Engle, RF (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance* **1**(1), 83–106

Duan, JC, Ritchken, P & Sun, Z (2006). Approximating GARCH-jump Models, Jump Diffusion Process, and Option Pricing. *Mathematical Finance* **16**(1), 21–52

Duffie, D, Pan, J & Singleton, KJ (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68**(6), 1343–76

Dupire, B (1994). Pricing with a smile. *Risk* 7, 18–20

Elliott, G, Granger, CWJ & Timmermann, A (eds.) (2006). *Handbook of Economic Forecasting*. North-Holland, Amsterdam

Engle, R (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society* **50**(4), 987–1007

Fengler, MR (2005). Semiparametric Modeling of Implied Volatility, Lecture Note in Finance. Springer Verlag, Heidelberg.

Firer, C & McLeod, H (1999). Equities, bonds, cash and inflation: historical performance in South Africa 1925–1998. *Investment Analysts Journal* 50, 7–28

Firer, C & Staunton, M (2002). 102 Years of South African financial market history. *Investment Analysts Journal* 56, 57–65

Flint, E, Chikurunhe, F & Seymour, A (unpublished). (Un)modelling the volatility surface: valuing South African volatility surfaces via risk-neutral historical return distributions, Peregrine Securities Report, 2012

Garman, M & Klass, M (1980). On the estimation of security price volatilities from historical data. *Journal of Business* **53**(1), 67–78

Gatheral, A (2006). *The Volatility surface: A practitioner's guide*. Wiley Finance.

Glosten, LR, Jagannathan, R & Runkle, D (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* **48**(5), 1779–801

Gray, P & Newman, S (2005). Canonical valuation of options in the presence of stochastic volatility. *Journal of Futures Markets* **25**(1), 1–19

Hacker, RS & Hatemi-J, A (2006). Tests for causality between integrated variables using asymptotic and bootstrap distributions: theory and application. *Applied Economics* **38**(13), 1489–500

Hagan, PS & West, G (2008). Methods for constructing a yield curve. *WILMOTT Magazine*, 70–81

Haley, MR & Walker, TB (2010). Alternative tilts for nonparametric option pricing. *Journal of Futures Markets* **30**(10), 983–1006

Hansen, PR & Lunde, A (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1). *Journal of Applied Econometrics* **20**(7), 873–89

Hentschel, L (1995). All in the family: Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics* **39**(1), 71–104

Heston, S (1993). A closed-form solution for options with stochastic volatility, with application to bond and currency options. *Review of Financial Studies* **6**(2), 327–43

Heston, SL & Nandi, S (2000). A closed-form GARCH option valuation model. *Review of Financial Studies* **13**(3), 585–625

Higgins, ML & Bera, AK (1992). A class of nonlinear ARCH Models. *International Economic* Review **33**(1), 137–58

Hull, J (2009). *Options, Futures and Other Derivative Securities*, 7th edn., Prentice Hall, Englewood Cliffs, NJ

Kotzé, A, Labuschagne, CAC, Nair, ML & Padayachi, N (2013). Arbitrage-free implied volatility surfaces for options on single stock futures. *North American Journal of Economics and Finance*, 26, 380–399

Kulikova, MV & Taylor, DR (2010). A conditionally heteroskedastic time series model for certain South African stock price returns. *Investment Analyst Journal* 72, 43–52

Lipton, A (2002). The volatility smile problem. *Risk*, 61–5

Manistre BJ (2010). A cost of capital approach to extrapolating an implied volatility surface. *Society of Actuaries*, 1–23

Merton, RC (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics* **3**(1), 125–44

Meucci, A (2005). *Risk and Asset Allocation*. Springer Finance.

McAleer, M & Medeiros, M (2008). Realized volatility: A review. *Econometric Reviews* **27**(1–3), 10–45

Nelson, DB (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**(2), 347–70

Parkinson, M (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business* **53**(1), 61–8

Poon, SH (2005). *A Practical Guide to Forecasting Financial Market Volatility*. JohnWiley & Sons Ltd, Chichester

Poon, SH & Granger, CWJ (2003). Forecasting financial market volatility: A review. *Journal of Economic Literature* **41**(2), 478–539

Rebonato, R. (2004) *Volatility and Correlation: the Perfect Hedger and the Fox*, 2nd edn., Wiley, Chichester

Rogers, LCG & Satchell, SE (1991). Estimating variance from high, low and closing prices. *Annals of Applied Probability* **1**(4), 504–12

Schwarz, GE (1978). Estimating the dimension of a model. *Annals of Statistics*, **6** (2), 461–464

Sheldon, TJ & Smith, AD (2004). Market consistent valuation of life assurance business. *British Actuarial Journal* **10**(03), 543–626

Shu, J & Zhang, JE (2001).The relationship between implied and realized volatility of S&P 500 Index. *Wilmott Magazine*, 83–91

Stutzer, M (1996). A simple nonparametric approach to derivative security valuation. *The Journal of Finance* **51**(5), 1633–52

Yang, D & Zhang, Q (2000). Drift independent volatility estimation based on high, low, open and close prices. *Journal of Business* **73**(3), 477–91

APPENDIX A
HISTORICAL VOLATILITY ESTIMATORS

A.1 MATHEMATICAL DEFINITIONS

A.1.1 Using the notation as per Garman & Klass (op. cit.), we have:

$$\begin{aligned} c_t &= \ln\left(S_t^{close} / S_t^{open}\right) \\ o_t &= \ln\left(S_t^{open} / S_{t-1}^{close}\right) \\ h_t &= \ln\left(S_t^{high} / S_{t-1}^{open}\right) \\ l_t &= \ln\left(S_t^{low} / S_{t-1}^{open}\right). \end{aligned} \tag{A.1}$$

A.1.2 From equation (A.1) and assuming n days within the periods, we have the Parkinson (op. cit.) variance estimator:

$$\sigma_{P,T}^2 = \frac{252}{4T\ln 2}\sum_{t=1}^{T}\left(h_t - l_t\right)^2. \tag{A.2}$$

The Garman-Klass (op. cit.) and modified Garman–Klass variance estimators are given by

$$\sigma_{GK,T}^2 = \frac{252}{T}\left[0.511\sum_{t=1}^{T}\left(h_t - l_t\right)^2 - \left(2\ln 2 - 1\right)\sum_{t=1}^{T}c_t^2\right] \tag{A.3}$$

$$\sigma_{GK^*,T}^2 = \frac{252}{T}\left[\sum_{t=1}^{T}o_t^2 + 0.511\sum_{t=1}^{T}\left(h_t - l_t\right)^2 - \left(2\ln 2 - 1\right)\sum_{t=1}^{T}{c_t}^2\right]. \tag{A.4}$$

The Rogers & Satchell (op. cit.) variance estimator is calculated as

$$\sigma_{RS,T}^2 = \frac{252}{T}\sum_{t=1}^{T}\left[h_t\left(h_t - c_t\right) + l_t\left(l_t - c_t\right)\right]. \tag{A.5}$$

Finally, the Yang–Zhang (op. cit.) variance estimator is

$$\sigma_{YZ,T}^2 = \sigma_{o,T}^2 + k\sigma_{c,T}^2 + \left(1-k\right)\sigma_{RS,T}^2, \text{ where } k = 0.34 \Big/ 1 + \frac{T+1}{T-1}. \tag{A.6}$$

APPENDIX B
TECHNICAL OUTLINE OF CANONICAL VALUATION

B.1 ESTIMATING THE FUTURE EMPIRICAL DENSITY

B.1.1 Arbitrage-free option pricing is governed by equation (27), which states that the fair value of a derivative is equal to the discounted expectation of its payoff under an appropriate risk-neutral measure. On this fundamental valuation theory, Stutzer used a normalisation method commonly used in discrete time models. At each future expiry time $\boldsymbol{T}$, the price process is discounted by the product of one-period, continuous risk-free interest rates $\boldsymbol{r_{t,1}}$ up to $\boldsymbol{T}$. Denoting the current price of the asset by $\boldsymbol{S_t}$ and current dividend payment by $\boldsymbol{D_t}$, the equivalent martingale probabilities $\boldsymbol{q}$ at time $\boldsymbol{T}$ must satisfy:

$$\begin{aligned} \boldsymbol{S_t} &= \mathbb{E}_{\mathbb{Q}}\left[\frac{\boldsymbol{S_t}+\boldsymbol{D_t}+\sum_{t=1}^{T}\boldsymbol{D_t}\prod_{s=1}^{T-1}\boldsymbol{e}^{\boldsymbol{r_{s,1}}}}{\prod_{t=1}^{T}\boldsymbol{e}^{\boldsymbol{r_{t,1}}}}\right] \\ &= \mathbb{E}_{\mathbb{P}}\left[\frac{\boldsymbol{S_t}+\boldsymbol{D_t}+\sum_{t=1}^{T}\boldsymbol{D_t}\prod_{s=t}^{T-1}\boldsymbol{e}^{\boldsymbol{r_{s,1}}}}{\prod_{t=1}^{T}\boldsymbol{e}^{\boldsymbol{r_{t,1}}}}\frac{\boldsymbol{dq}}{\boldsymbol{dp}}\right], \end{aligned} \tag{B.1}$$

where $\mathbb{P}$ denotes the real-world probability measure and $\boldsymbol{dq/dp}$ denotes the Radon–Nykodym derivative of the martingale measure with respect to $\mathbb{P}$ at time $\boldsymbol{T}$. Thus, one must be able to estimate the equivalent martingale measure satisfying the no-arbitrage constraint given in equation (B.1) in order to calculate the fair value of a European derivative claim from equation (27).

B.1.2 Given an historical sample of τ-period asset returns $\left\{r_{t,\tau}, t=1,\ldots T\right\}$, define the de-trended asset return as

$$Z_{t,\tau}=\frac{r_{t,\tau}-\bar{\mu}}{\bar{\sigma}}, \tag{B.2}$$

where $\bar{\mu}$ and $\bar{\sigma}$ are the sample mean and sample standard deviation respectively. The terminal τ-period asset prices are then given by

$$\begin{aligned} S_{t,\tau} &= S_0 e^{X_{t,\tau}} \\ X_{t,\tau} &= \left[\left(\mu_t+\sigma_t Z_{t,\tau}\right)-y_{t,\tau}-\delta_{t,\tau})\right]+\left(y_{T,\tau}-\delta_{T,\tau}\right) \end{aligned} \tag{B.3}$$

where S_0 is the current asset price, $y_{t,\tau}$ and $\delta_{t,\tau}$ are the τ-period historical risk-free and dividend rates respectively, $y_{T,\tau}$ and $\delta_{T,\tau}$ are the respective forward-looking, τ-period risk-free and dividend rates, and it is not necessary to have $\mu_t=\bar{\mu}$ and $\sigma_t=\bar{\sigma}$. By working with historical excess returns and adding back the current term-specific rates, one latently addresses the stochastic nature of interest and dividend rates.

B.1.3 The process $X_{t,\tau}$ has empirical distribution function $G(\bullet)$, which can be estimated as a step function:

$$\hat{G}(x) = \frac{1}{T}\sum_{t=1}^{T} I\left\{X_{t,\tau} \le x\right\}, \tag{B.4}$$

where $I\{\bullet\}$ is the indicator function.

B.1.4 This implies a finite support for the future terminal share-price distribution characterised by the minimum and maximum returns. Each possible future price $S_{t,T}$ has an estimated real-world probability $\hat{p} = 1/n$.

B.2 ESTIMATING THE RISK-NEUTRAL DENSITY VIA RELATIVE ENTROPY

B.2.1 The method used to transform $\hat{\mathbb{P}}$ into the estimated risk-neutral return distribution $\hat{\mathbb{Q}}$ is described here. Using the fact that $\hat{p} = 1/n$ and equations (B.2) and (B.3) (dropping subscripts for notational ease), the no-arbitrage constraint given in equation (B.1) can be simplified as

$$1 = \sum_{t=1}^{T}\left(e^{-r}X_{t,\tau}\right)\frac{q_t}{\hat{p}_t}\hat{p}_t, \tag{B.5}$$

where q_t is the risk-neutral probability of return $X_{t,T}$.

B.2.2 Stutzer showed that the solution to minimising the relative entropy given in equation (30) subject to the risk-neutral constraint presented in equation (B.5) is given by the following Gibbs canonical distribution:

$$\hat{q}_t = \frac{exp\left(\gamma^*\left(e^{-r}X_{t,\tau}\right)\right)}{\sum_{t=1}^{T} exp\left(\gamma^*\left(e^{-r}X_{t,\tau}\right)\right)}, \tag{B.6}$$

where γ^* is the Lagrange multiplier found by solving the following unconstrained minimisation problem,

$$\gamma^* = arg\min_{\gamma}\sum_{t=1}^{T} exp\left[\gamma\left(\left(e^{-r}X_{t,\tau}\right)-1\right)\right]. \tag{B.7}$$

B.2.3 Using the risk-neutral probabilities calculated in equation (B.7), the discounted expected payoff of the derivative contract can be computed. Thus, the price of a European call option C at time t with strike price K, expiring at time τ is given by

$$C_{t,\tau,K} = e^{-r_{t,T}(T-t)}\sum_{i=1}^{n}\max\left[S_{t,\tau}-K,0\right]\hat{q}_t. \tag{B.8}$$

The implied CV volatility, $\sigma_{t,\tau,K}^{CV}$, can then be solved for from the computed option price C.

B.3 COMMON EXTENSIONS OF CANONICAL VALUATION

B.3.1 We discuss two common extensions of CV here; namely, multiple underlying assets and incorporating known option prices.

B.3.2 It is easy to extend the CV method to incorporate multiple underlying assets. Consider a derivative contract written on M underlying assets. By including $M-1$ additional constraints in the form of equation (B.5) to the constrained relative entropy minimisation problem, the solution obtained is now given by the multivariate canonical distribution

$$\hat{q}_t = \frac{exp\left(\sum_{j=1}^{M}\gamma_j^*\left(e^{-r}X_{t,\tau,j}\right)\right)}{\sum_{t=1}^{T}exp\left(\sum_{j=1}^{M}\gamma_j^*\left(e^{-r}X_{t,\tau,j}\right)\right)}, \tag{B.9}$$

where $X_{t_i,\tau,j}$ denotes the i^{th} return of asset j for the term τ. In this case, the M-component vector satisfies

$$\gamma^* = arg\min_{\gamma}\sum_{t=1}^{T}exp\left[\sum_{j=1}^{n}\gamma_j^*\left(\left(e^{-r}X_{t,\tau,j}\right)-1\right)\right]. \tag{B.10}$$

B.3.3 Another common example of an additional constraint is to ensure that the at-the-money option price implied by the distribution $\hat{\mathbb{Q}}$ is equal to the at-the-money option price quoted in the market. For example, assume that there is a call option C^* with strike level $K \equiv S_0$, expiring at time t_n. In order to ensure the correct pricing of this option we need to include an additional constraint and thus solve for two multipliers, γ_1^* and γ_2^* that satisfy

$$\gamma^* = \arg\min_{\gamma}\sum_{t=1}^{T}\exp\left[\gamma_1\left(\left(e^{-r}X_{t,\tau}\right)-1\right)+\gamma_2\left(e^{-r}\max\left[S_{t,\tau}-K,0\right]-C^*\right)\right]. \tag{B.11}$$

B.3.4 Substituting the multiplier values into a bivariate canonical distribution, the estimated risk-neutral probabilities are given by

$$\hat{q}_t = \frac{exp\left[\gamma_1^*\left(e^{-r}X_{t,\tau}\right)+\gamma_2^*\left(e^{-r}\max\left[S_{t,\tau}-K,0\right]\right)\right]}{\sum_{t=1}^{T}exp\left[\gamma_1^*\left(e^{-r}X_{t,\tau}\right)+\gamma_2^*\left(e^{-r}\max\left[S_{t,\tau}-K,0\right]\right)\right]}. \tag{B.12}$$