Volatility Surfaces:

Theory, Rules of Thumb, and Empirical Evidence

Toby Daglish

School of Economics and Finance

Victoria University of Wellington

Email: *toby.daglish@vuw.ac.nz*

John Hull

Rotman School of Management

University of Toronto

Email: *hull@rotman.utoronto.ca*

Wulin Suo

School of Business, Queen's University

Email: *wsuo@business.queensu.ca*

First version: March 2001

This version: August 2006

Abstract

Implied volatilities are frequently used to quote the prices of options. The implied volatility of a European option on a particular asset as a function of strike price and time to maturity is known as the asset's volatility surface. Traders monitor movements in volatility surfaces closely. In this paper we develop a no-arbitrage condition for the evolution of a volatility surface. We examine a number of rules of thumb used by traders to manage the volatility surface and test whether they are consistent with the no-arbitrage condition and with data on the trading of options on the S&P 500 taken from the over-the-counter market. Finally we estimate the factors driving the volatility surface in a way that is consistent with the no-arbitrage condition.

1 Introduction

Option traders and brokers in over-the-counter markets frequently quote option prices using implied volatilities calculated from Black and Scholes (1973) and other similar models. The reason is that implied volatilities usually have to be updated less frequently than the prices themselves. Put–call parity implies that, in the absence of arbitrage, the implied volatility for a European call option is the same as that for a European put option when the two options have the same strike price and time to maturity. This is convenient: when quoting an implied volatility for a European option with a particular strike price and maturity date, a trader does not need to specify whether a call or a put is being considered.

The implied volatility of European options on a particular asset as a function of strike price and time to maturity is known as the volatility surface. Every day traders and brokers estimate volatility surfaces for a range of different underlying assets from the market prices of options. Some points on a volatility surface for a particular asset can be estimated directly because they correspond to actively traded options. The rest of the volatility surface is typically determined by interpolating between these points.

If the assumptions underlying Black–Scholes held for an asset, its volatility surface would be flat and unchanging. In practice the volatility surfaces for most assets are not flat and change stochastically. Consider for example equities and foreign currencies. Rubinstein (1994) and Jackwerth and Rubinstein (1996), among others, show that the implied volatilities of stock and stock index options exhibit a pronounced "skew" (that is, the implied volatility is a decreasing function of strike price). For foreign currencies this skew becomes a "smile" (that is, the implied volatility is a U-shaped function of strike price). For both types of assets, the implied volatility can be an increasing or decreasing function of the time to maturity. The volatility surface changes through time, but the general shape of the relationship between volatility and strike price tends to be preserved.

Traders use a volatility surface as a tool to value a European option when its price is not directly observable in the market. Provided there are a reasonable number of actively traded European options and these span the full range of the strike prices and times to maturity that are encountered, this approach ensures that traders price all European options consistently with the market. However, as pointed out by Hull and Suo (2002), there is no easy way to extend the approach to price path-dependent exotic options such as barrier options, compound options, and Asian options. As a result there is liable to be some model risk when these options are priced.

Traders also use the volatility surface in an *ad hoc* way for hedging. They attempt to hedge against potential changes in the volatility surface as well as against changes in the asset price. Derman (1999) discusses alternative approaches to hedging against asset price movements. The "volatility-by-strike" or "sticky strike" rule assumes that the implied volatility for an option with a given strike price and maturity will be unaffected by changes in the underlying asset price. Another popular approach is the “volatility-by-moneyness” or “sticky delta” rule. This assumes that the volatility for a particular maturity depends only on the moneyness (that is, the ratio of the price of the underlying asset to the strike price).

The first attempts to model the volatility surface were by Rubinstein (1994), Derman and Kani (1994), and Dupire (1994). These authors show how a one-factor model for an asset price, known as the implied volatility function (IVF) model, can be developed so that it is exactly consistent with the current volatility surface. Unfortunately, the evolution of the volatility surface under the IVF model can be unrealistic. The volatility surface given by the model at a future time is liable to be quite different from the initial volatility surface. For example, in the case of a foreign currency the initial U-shaped relationship between implied volatility and strike price is liable to evolve to one where the volatility is a monotonic increasing or decreasing function of strike price. Dumas, Fleming, and Whaley (1997) have shown that the IVF model does not capture the dynamics of market prices well. Hull and Suo (2002) have shown that it can be dangerous to use the model for the relative pricing of barrier options and plain vanilla options.

This paper extends existing research in a number of ways. Similarly to Ledoit and Santa Clara (1998), Schönbucher (1999), Brace et al (2001), and Britten–Jones and Neuberger (2000), we develop no arbitrage conditions for the evolution of the volatility surface. We consider a general model where there are a number of factors driving the volatility surface and these may be correlated with stock price movements. We then investigate the implications of the no-arbitrage condition for the shapes of the volatility surfaces likely to be observed in different situations and examine whether the various rules of thumb that have been put forward by traders are consistent with the no-arbitrage condition. Finally we carry out two empirical studies. The first empirical study extends the work of Derman (1999) to investigate whether movements in the volatility surface are consistent with different rules of thumb. The second extends work of Kamal and Derman (1997), Skiadopoulos, Hodges, and Clewlow (2000), Cont and Fonseca (2002), and Cont, Fonseca and Durrleman (2002) to estimate the factors driving movements in the volatility surface in a way that reflects the theoretical drifts of points on the surface. One distinguishing feature of our work is that we use over-the-counter consensus data for our empirical studies. To the best of our knowledge we are the first derivatives researchers to use this type of data. Another distinguishing feature of our research is that, in addition to the sticky strike and sticky delta rules we consider a “square root of time” rule that is often used by traders and documented by, for example, Natenburg (1994).

The rest of the paper is organized as follows. Section 2 explains the rules of thumb. Section 3 develops a general model for the evolution of a volatility surface and derives the no-arbitrage condition. Section 4 discusses the implications of the no-arbitrage condition. Section 5 examines a number of special cases of the model. Section 6 considers whether the rules of thumb are consistent with the no-arbitrage condition. Section 7 tests whether there is empirical support for the rules of thumb. Section 8 derives factors driving the volatility surface in a way that is consistent with the no-arbitrage condition. Conclusions are in Section 9.

2 Rules of Thumb

A number of rules of thumb have been proposed about volatility surfaces. These rules of thumb fall into two categories. In the first category are rules concerned with the way in which the volatility surface changes through time. They are useful in the calculation of the Greek letters such as delta and gamma. In the second category are rules concerned with the relationship between the volatility smiles for different option maturities at a point in time. They are useful in creating a complete volatility surface when market prices are available for a relatively small number of options. In this section we explain three different rules of thumb: the sticky strike rule, the sticky delta rule and the square root of time rule. The first two of these rules are in the first category and provide a basis for calculating Greek letters. The third rule is in the second category and assists with the mechanics of "filling in the blanks" when a complete volatility surface is being produced.

We consider a European call option with maturity T, strike price K and implied volatility σ_{TK}. We define the call price as c, the underlying asset price as S, and the delta, Δ of the option as the total rate of change of its price with respect to the asset price:

$$\Delta = \frac{dc}{dS}.$$

2.1 The Sticky Strike Rule

The sticky strike rule assumes that σ_{TK} is independent of S. This is an appealing assumption because it implies that the sensitivity of the price of an option to S is

$$\frac{\partial c}{\partial S},$$

where for the purposes of calculating the partial derivative the option price, c, is considered to be a function of the asset price, S, σ_{TK}, and time, t. The assumption enables the Black–Scholes formula to be used to calculate delta with the volatility parameter set equal to the option's implied volatility. The same is true of gamma, the rate of change of delta with respect to the asset price.

In what we will refer to as the *basic sticky strike model*, σ_{TK} is a function only of K and T. In the *generalized sticky strike model* it is independent of S, but possibly dependent on other stochastic variables.

2.2 The Sticky Delta Rule

An alternative to the sticky strike rule is the sticky delta rule. This assumes that the implied volatility of an option depends on the S and K, through its dependence on the moneyness variable,

K/S. The delta of a European option in a stochastic volatility model is

$$\Delta = \frac{\partial c}{\partial S} + \frac{\partial c}{\partial \sigma_{TK}} \frac{\partial \sigma_{TK}}{\partial S}.$$

Again, for the purposes of calculating partial derivatives the option price c is considered to be a function of S, σ_{TK}, and t. The first term in this expression is the delta calculated using Black-Scholes with the volatility parameter set equal to implied volatility. In the second term, $\partial c/\partial \sigma_{TK}$ is positive. It follows that, if σ_{TK} is a declining (increasing) function of the strike price, it is an increasing (declining) function of S and Δ is greater than (less than) that given by Black-Scholes. For equities σ_{TK} is a declining function of K and so the Black-Scholes delta understates the true delta. For an asset with a U-shaped volatility smile Black-Scholes understates delta for low strike prices and overstates it for high strike prices.

In the most basic form of the sticky delta rule the implied volatility is assumed to be a deterministic function of K/S and $T - t$. We will refer to this as the *basic sticky delta model.* A more general version of the sticky delta rule is where the process for V_{TK} depends on K, S, T, and t only through its dependence on K/S and $T-t$. We will refer to this as the *generalized sticky delta model.* The implied volatility changes stochastically but parameters describing the process followed by the volatility (for example, the volatility of the volatility) are functions only of K/S and $T - t$.

Many traders argue that a better measure of moneyness than K/S is K/F where F is the forward value of S for a contract maturing at time T. A version on the sticky delta rule sometimes used by traders is that $\sigma_{TK}(S,t) - \sigma_{TF}(S,t)$ is a function only of K/F and $T-t$. Here it is the excess of the volatility over the at-the-money volatility, rather than the volatility itself, which is assumed to be a deterministic function of the moneyness variable, K/F. This form of the sticky delta rule allows the overall level of volatility to change through time and the shape of the volatility term structure to change, but when measured relative to the at-the-money volatility, the volatility is dependent only on K/S and $T-t$. We will refer to this model as the *relative sticky delta model.* If the at-the-money volatility is stochastic, but independent of S, the model is a particular case of the generalized sticky delta model just considered.

2.3 The Square Root of Time Rule

Another rule that is sometimes used by traders is what we will refer to as the "square root of time rule". This is described in Natenberg (1994) and Hull (2006). It provides a specific relationship between the volatilities of options with different strike prices and times to maturity at a particular time. One version of this rule is

$$\frac{\sigma_{TK}(S,t)}{\sigma_{TF}(S,t)} = \Phi\left(\frac{\ln(K/F)}{\sqrt{T-t}}\right),$$

where Φ is a function. An alternative that we will use is

$$\sigma_{TK}(S,t) - \sigma_{TF}(S,t) = \Phi\left(\frac{\ln(K/F)}{\sqrt{T-t}}\right). \tag{1}$$

We will refer to this version of the rule where Φ does not change through time as the *stationary square root of time model* and the version of the rule where the form of the function Φ changes stochastically as the *stochastic square root of time model.*

The square root of time model (whether stationary or stochastic) simplifies the specification of the volatility surface. If we know

1. The volatility smile for options that mature at one particular time T^*, and
2. At-the-money volatilities for other maturities,

we can compute the complete volatility surface. Suppose that F^* is the forward price of the asset for a contract maturing at T^*. We can compute the volatility smile at time T from that at time T^* using the result that

$$\sigma_{TK}(S,t) - \sigma_{TF}(S,t) = \sigma_{T^*K^*}(S,t) - \sigma_{T^*F^*}(S,t),$$

where

$$\frac{\ln(K/F)}{\sqrt{T-t}} = \frac{\ln(K^*/F^*)}{\sqrt{T^*-t}},$$

or

$$K^* = F^*\left(\frac{K}{F}\right)^{\sqrt{(T^*-t)/(T-t)}}.$$

If the at-the-money volatility is assumed to be stochastic, but independent of S, the stationary square root of time model is a particular case of the relative sticky strike model and the stochastic square root of time model is a particular case of the generalized sticky strike model.

3 The Dynamics of the Implied Volatility

We suppose that the risk-neutral process followed by the price of an asset, S, is

$$\frac{dS}{S} = [r(t) - q(t)]\ dt + \sigma\ dz, \tag{2}$$

where $r(t)$ is the risk-free rate, $q(t)$ is the yield provided by the asset, σ is the asset's volatility, and z is a Wiener process. We suppose that $r(t)$ and $q(t)$ are deterministic functions of time and that σ follows a diffusion process. Our model includes the IVF model and stochastic volatility models such as Hull and White (1987), Stein and Stein (1991) and Heston (1993) as special cases.

Most stochastic volatility models specify the process for σ directly. We instead specify the processes for all implied volatilities. As in the previous section we define $\sigma_{TK}(t,S)$ as the implied volatility at time t of an option with strike price K and maturity T when the asset price is S. For convenience we define

$$V_{TK}(t,S) = [\sigma_{TK}(t,S)]^2$$

as the implied variance of the option. Suppose that the process followed by V_{TK} in a risk-neutral world is

$$dV_{TK} = \alpha_{TK}\ dt + V_{TK} \sum_{i=1}^{N} \theta_{TKi}\, dz_i, \tag{3}$$

where $z_1, \cdots, z_N$ are Wiener processes driving the volatility surface. The z_i may be correlated with the Wiener process, z, driving the asset price in equation (2). We define ρ_i as the correlation between z and z_i. Without loss of generality we assume that the z_i are correlated with each other only through their correlation with z. This means that the correlation between z_i and z_j is $\rho_i\rho_j$. The initial volatility surface is $\sigma_{TK}(0, S_0)$ where S_0 is the initial asset price. This volatility surface can be estimated from the current ($t = 0$) prices of European call or put options and is assumed to be known.

The family of processes in equation (3) defines the multi-factor dynamics of the volatility surface. The parameter θ_{TKi} measures the sensitivity of V_{TK} to the Wiener process z_i. In the most general form of the model the parameters α_{TK} and θ_{TKi} ($1 \leq i \leq N$) may depend on past and present values of S, past and present values of V_{TK}, and time.

There is clearly a relationship between the instantaneous volatility $\sigma(t)$ and the volatility surface $\sigma_{TK}(t, S)$. The appendix shows that the instantaneous volatility is the limit of the implied volatility of an at-the-money option as its time to maturity approaches zero. For this purpose an at-the-money option is defined as an option where the strike price equals the forward asset price.[1] Formally:

$$\lim_{T \to t} \sigma_{TF}(t, S) = \sigma(t), \tag{4}$$

where F is the forward price of the asset at time t for a contract maturing at time T. In general the process for σ is non-Markov. This is true even when the processes in equation (3) defining the volatility surface are Markov.[2]

Define $c(S, V_{TK}, t; K, T)$ as the price of a European call option with strike price K and maturity T when the asset price S follows the process in equations (2) and (3). From the definition of implied volatility and the results in Black and Scholes (1973) and Merton (1973) it follows that:

$$c(S, V_{TK}, t; K, T) = e^{-\int_t^T q(\tau)\, d\tau} SN(d_1) - e^{-\int_t^T r(\tau)\, d\tau} KN(d_2),$$

where

$$d_1 \quad = \quad \frac{\ln(S/K) + \int_t^T [r(\tau) - q(\tau)] d\tau}{\sqrt{V_{TK}(T-t)}} + \frac{1}{2}\sqrt{V_{TK}(T-t)},$$

[1] Note that the result is not necessarily true if we define an at-the-money option as an option where the strike price equals the asset price. For example, when

$$\sigma_{TK}(t, S) = a + b\frac{1}{T-t}\ln(F/K)$$

with a and b constants, $\lim_{T\to t} \sigma_{TF}(t, S)$ is not the same as $\lim_{T\to t} \sigma_{TS}(t, S)$

[2] There is an analogy here to the Heath, Jarrow, and Morton (1992) model. When each forward rate follows a Markov process the instantaneous short rate does not in general do so.

$$d_2 = \frac{\ln(S/K) + \int_t^T [r(\tau) - q(\tau)] d\tau}{\sqrt{V_{TK}(T-t)}} - \frac{1}{2}\sqrt{V_{TK}(T-t)}.$$

Using Ito's lemma equations (2) and (3) imply that the drift of c in a risk-neutral world is:

$$\frac{\partial c}{\partial t} + (r-q)S\frac{\partial c}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 c}{\partial S^2} + \alpha_{TK}\frac{\partial c}{\partial V_{TK}} + \frac{1}{2}V_{TK}^2 \frac{\partial^2 c}{\partial V_{TK}^2}\left[\sum_{i=1}^{N}(\theta_{TKi})^2 + \sum_{i\neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j\right]$$

$$+SV_{TK}\sigma\frac{\partial^2 c}{\partial S \partial V_{TK}}\sum_{i=1}^{N}\theta_{TKi}\rho_i.$$

In the most general form of the model ρ_i, the correlation between z and z_i, is a function of past and present values of S, past and present values of V_{TK} and time. For there to be no arbitrage the process followed by c must provide an expected return of r in a risk-neutral world. It follows that

$$\frac{\partial c}{\partial t} + (r-q)S\frac{\partial c}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 c}{\partial S^2} + \alpha_{TK}\frac{\partial c}{\partial V_{TK}} + \frac{1}{2}V_{TK}^2 \frac{\partial^2 c}{\partial V_{TK}^2}\left[\sum_{i=1}^{N}(\theta_{TKi})^2 + \sum_{i\neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j\right]$$

$$+SV_{TK}\sigma\frac{\partial^2 c}{\partial S \partial V_{TK}}\sum_{i=1}^{N}\theta_{TKi}\rho_i = rc.$$

When V_{TK} is held constant, c satisfies the Black-Scholes (1973) and Merton (1973) differential equation. As a result

$$\frac{\partial c}{\partial t} + (r-q)S\frac{\partial c}{\partial S} = rc - \frac{1}{2}V_{TK}S^2\frac{\partial^2 c}{\partial S^2}.$$

It follows that

$$\frac{1}{2}S^2\frac{\partial^2 c}{\partial S^2}(\sigma^2 - V_{TK}) + \alpha_{TK}\frac{\partial c}{\partial V_{TK}} + \frac{1}{2}V_{TK}^2 \frac{\partial^2 c}{\partial V_{TK}^2}\left[\sum_{i=1}^{N}(\theta_{TKi})^2 + \sum_{i\neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j\right]$$

$$+SV_{TK}\sigma\frac{\partial^2 c}{\partial S \partial V_{TK}}\sum_{i=1}^{N}\theta_{TKi}\rho_i = 0,$$

or

$$\begin{aligned}\alpha_{TK} = & -\frac{1}{2\partial c/\partial V_{TK}}\left[S^2\frac{\partial^2 c}{\partial S^2}(\sigma^2 - V_{TK}) + \frac{\partial^2 c}{\partial V_{TK}^2}V_{TK}^2\sum_{i=1}^{N}(\theta_{TKi})^2\right. \\ & \left.+\frac{\partial^2 c}{\partial V_{TK}^2}V_{TK}^2\sum_{i\neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j + 2SV_{TK}\sigma\frac{\partial^2 c}{\partial S\partial V_{TK}}\sum_{i=1}^{N}\theta_{TKi}\rho_i\right]. \qquad (5)\end{aligned}$$

The partial derivatives of c with respect to S and V_{TK} are the same as those for the Black–Scholes model:

$$\begin{aligned}\frac{\partial c}{\partial S} &= e^{-\int_t^T q(\tau)\,d\tau}N(d_1), \\ \frac{\partial^2 c}{\partial S^2} &= \frac{\phi(d_1)e^{-\int_t^T q(\tau)\,d\tau}}{S\sqrt{V_{TK}(T-t)}}, \\ \frac{\partial c}{\partial V_{TK}} &= \frac{Se^{-\int_t^T q(\tau)\,d\tau}\phi(d_1)\sqrt{T-t}}{2\sqrt{V_{TK}}},\end{aligned}$$

$$\frac{\partial^2 c}{\partial V_{TK}^2} = \frac{Se^{-\int_t^T q(\tau)\,d\tau}\phi(d_1)\sqrt{T-t}}{4V_{TK}^{3/2}}(d_1 d_2 - 1),$$
$$\frac{\partial^2 c}{\partial S \partial V_{TK}} = -\frac{e^{-\int_t^T q(\tau)\,d\tau}\phi(d_1)d_2}{2V_{TK}},$$

where ϕ is the density function of the standard normal distribution:

$$\phi(x) = \frac{1}{2\pi}\exp\left(-\frac{x^2}{2}\right), \quad -\infty < x < \infty.$$

Substituting these relationships into equation (5) and simplifying we obtain

$$\begin{aligned}\alpha_{TK} &= \frac{1}{T-t}(V_{TK} - \sigma^2) - \frac{V_{TK}(d_1 d_2 - 1)}{4}\left[\sum_{i=1}^{N}(\theta_{TKi})^2 + \sum_{i \neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j\right] \\ &\quad + \sigma d_2\sqrt{\frac{V_{TK}}{T-t}}\sum_{i=1}^{N}\theta_{TKi}\rho_i. \qquad (6)\end{aligned}$$

Equation (6) provides an expression for the risk-neutral drift of an implied variance in terms of its volatility. The first term on the right hand side is the drift arising from the difference between the implied variance and the instantaneous variance. It can be understood by considering the situation where the instantaneous variance, σ^2, is a deterministic function of time. The variable V_{TK} is then also a function of time and

$$V_{TK} = \frac{1}{T-t}\int_t^T \sigma(\tau)^2 d\tau$$

Differentiating with respect to time we get

$$\frac{dV_{TK}}{dt} = \frac{1}{T-t}[V_{TK} - \sigma(t)^2],$$

which is the first term.

The analysis can be simplified slightly by considering the variable $\hat{V}_{TK}$ instead of V where $\hat{V}_{TK} = (T-t)V_{TK}$. Because

$$d\hat{V}_{TK} = -V_{TK}dt + (T-t)dV_{TK},$$

it follows that

$$\begin{aligned}d\hat{V}_{TK} &= \left\{-\sigma^2 - \frac{\hat{V}_{TK}(d_1 d_2 - 1)}{4}\left[\sum_{i=1}^{N}(\theta_{TKi})^2 + \sum_{i \neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j\right]\right. \\ &\quad \left. + \sigma d_2\sqrt{\hat{V}_{TK}}\sum_{i=1}^{N}\theta_{TKi}\rho_i\right\}dt + \hat{V}_{TK}\sum_{i=1}^{N}\theta_{TKi}dz_i.\end{aligned}$$

4 Implications of the No-Arbitrage Condition

Equation (6) provides a no-arbitrage condition for the drift of the implied variance as a function of its volatility. In this section we examine the implications of this no-arbitrage condition. In the general case where V_{TK} is nondeterministic, the first term in equation (6) is mean fleeing; that is, it provides negative mean reversion. This negative mean reversion becomes more pronounced as the option approaches maturity. For a viable model the θ_{TKi} must be complex functions that in some way offset this negative mean reversion. Determining the nature of these functions is not easy. However, it is possible to make some general statements about the volatility smiles that are consistent with stable models.

4.1 The Zero Correlation Case

Consider first the case where all the ρ_i are zero so that equation (6) reduces to

$$\alpha_{TK} = \frac{1}{T-t}(V_{TK}-\sigma^2) - \frac{V_{TK}(d_1d_2-1)}{4}\sum_{i=1}^{N}(\theta_{TKi})^2$$

As before, we define F as the forward value of the asset for a contract maturing at time T so that

$$F = Se^{\int_t^T [r(\tau)-q(\tau)]d\tau}.$$

The drift of $V_{TK} - V_{TF}$ is

$$\frac{1}{T-t}(V_{TK}-V_{TF}) - \frac{V_{TK}(d_1d_2-1)}{4}\sum_{i=1}^{N}(\theta_{TKi})^2 - \frac{V_{TF}[1+V_{TF}(T-t)/4]}{4}\sum_{i=1}^{N}(\theta_{TFi})^2.$$

Suppose that $d_1d_2 - 1 > 0$ and $V_{TK} < V_{TF}$. In this case each term in the drift of $V_{TK} - V_{TF}$ is negative. As a result $V_{TK} - V_{TF}$ tends to get progressively more negative and the model is unstable with negative values of V_{TK} being possible. We deduce from this that V_{TK} must be greater than V_{TF} when $d_1d_2 > 1$. The condition $d_1d_2 > 1$ is satisfied for very large and very small values of K. It follows that the case where the ρ_i are zero can be consistent with the U-shaped volatility smile. It cannot be consistent with a upward or downward sloping smile because in these cases $V_{TK} < V_{TF}$ for either very high or very low values of K.

Our finding is consistent with a result in Hull and White (1987). These authors show that when the instantaneous volatility is independent of the asset price, the price of a European option is the Black–Scholes price integrated over the distribution of the average variance. They demonstrate that when $d_1d_2 > 1$ a stochastic volatility tends to increase an option's price and therefore its implied volatility.

A U-shaped volatility smile is commonly observed for options on a foreign currency. Our analysis shows that this is consistent with the empirical result that the correlation between implied volatilities and the exchange rate is close to zero (see, for example, Bates (1996)).

4.2 Volatility Skews

We have shown that a zero correlation between volatility and asset price can only be consistent with a U-shaped volatility smile. We now consider the types of correlations that are consistent with volatility skews.

Consider first the situation where the volatility is a declining function of the strike price. The variance rate V_{TK} is greater than V_{TF} when K is very small and less than V_{TF} when K is very large. The drift of $V_{TK} - V_{TF}$ is

$$\frac{1}{T-t}(V_{TK} - V_{TF}) - \frac{V_{TK}(d_1 d_2 - 1)}{4}\sum_{i=1}^{N}(\theta_{TKi})^2 - \frac{V_{TF}(1 + V_{TF}(T-t)/4)}{4}\sum_{i=1}^{N}(\theta_{TFi})^2$$
$$+\sigma d_2\sqrt{\frac{V_{TK}}{T-t}}\sum_{i=1}^{N}\theta_{TKi}\rho_i + \sigma\frac{V_{TF}}{2}\sum_{i=1}^{N}\theta_{TFi}\rho_i - \frac{V_{TK}(d_1 d_2 - 1)}{4}\sum_{i\neq j}\theta_{TKi}\theta_{TKj}\rho_i\rho_j$$
$$-\frac{V_{TF}(1 + V_{TF}(T-t)/4)}{4}\sum_{i=1}^{N}\theta_{TFi}\theta_{TFj}\rho_i\rho_j$$

When $K > F$, $V_{TK} - V_{TF}$ is negative and the effect of the first three terms is to provide a negative drift. For a stable model we require the last four terms to give a positive drift. As K increases and we approach option maturity the first of the last four terms dominates the other three. Because $d_2 < 0$ we must have

$$\sum_{i=1}^{N}\theta_{TKi}\rho_i < 0 \tag{7}$$

when K is very large. The instantaneous covariance of the asset price and its variance is

$$\sigma\sum_{i=1}^{N}\theta_{TKi}\rho_i.$$

Because $\sigma > 0$ it follows that when K is large the asset price must be negatively correlated with its variance.

Equities provide an example of a situation where there is a volatility skew of the sort we are considering. As has been well documented by authors such as Christie (1982), the volatility of an equity price tends to be negatively correlated with the equity price. This is consistent with the result we have just presented.

Consider next the situation where volatility is an increasing function of the strike price. (This is the case for options on some commodity futures.) A similar argument to that just given shows that the no-arbitrage relationship implies that the volatility of the underlying must be positively correlated with the value of the underlying.

4.3 Inverted U-Shaped Smiles

An inverted U-shaped volatility smile is much less common than a U-shaped volatility smile or a volatility skew. An analysis similar to that in Section 4.2 shows why this is so. For a stable model

$$\sum_{i=1}^{N}\theta_{TKi}\rho_i$$

must be less than zero when K is large and greater than zero when K is small. It is difficult to see how this can be so without the stochastic terms in the processes for the V_{TK} having a form that quickly destroys the inverted U-shaped pattern.

5 Special Cases

In this section we consider a number of special cases of the model developed in Section 3. This will enable us to reach conclusions about whether the rules of thumb in Section 2 can satisfy the no-arbitrage condition in Section 3.

Case 1: V_{TK} is a deterministic function only of t, T, and K

In this situation $\theta_{TKi} = 0$ for all $i \geq 1$ so that from equation (3)

$$dV_{TK} = \alpha_{TK}\ dt.$$

Also from equation (6)

$$\alpha_{TK} = \frac{1}{T-t}\left[V_{TK} - \sigma^2\right],$$

so that

$$dV_{TK} \quad = \quad \frac{1}{T-t}\left[V_{TK} - \sigma^2\right]\ dt,$$

which can be written as

$$(T-t)dV_{TK} - V_{TK}dt = -\sigma^2\ dt,$$

or

$$\sigma^2 = -\frac{d\left[(T-t)V_{TK}\right]}{dt}. \tag{8}$$

This shows that σ is a deterministic function of time. The only model that is consistent with V_{TK} being a function only of t, T, and K is therefore the model where the instantaneous volatility (σ) is a function only of time. This is Merton's (1973) model.

In the particular case where V_{TK} depends only on T and K, equation (8) shows that $V_{TK} = \sigma^2$ and we get the Black-Scholes constant-volatility model.

Case 2: V_{TK} is independent of the asset price, S .

In this situation ρ_i is zero and equation (6) becomes

$$\alpha_{TK} = \frac{1}{T-t}(V_{TK} - \sigma^2) - \frac{V_{TK}(d_1 d_2 - 1)}{4}\sum_{i=1}^{N}(\theta_{TKi})^2.$$

Both α_{TK} and the θ_{TKi} must be independent of S. Because d_1 and d_2 depend on S we must have θ_{TKi} equal zero for all i. Case 2 therefore reduces to Case 1. The only model that is consistent with

V_{TK} being independent of S is therefore Merton's (1973) model where the instantaneous volatility is a function only of time.

Case 3: V_{TK} is a deterministic function of t, T, and K/S.
In this situation

$$V_{TK} = G\left(T, t, \frac{K}{F}\right),$$

where G is a deterministic function and as before F is the forward price of S. From equation (4) the spot instantaneous volatility, σ, is given by

$$\sigma^2 = G(t, t, 1).$$

This is a deterministic function of time. It follows that, yet again, the model reduces to Merton's (1973) deterministic volatility model.

Case 4: V_{TK} is a deterministic function of t, T, S and K.
In this situation we can write

$$V_{TK} = G(T, t, F, K), \tag{9}$$

where G is a deterministic function. From equation (4), the instantaneous volatility, σ is given by

$$\sigma^2 = \lim_{T \to t} G(T, t, F, F).$$

This shows that the instantaneous volatility is a deterministic function of F and t. Equivalently it is a deterministic function of underlying asset price, S, and t. It follows that the model reduces to the implied volatility function model. Writing σ as $\sigma(S, t)$, Dupire (1994) and Andersen and Brotherton–Ratcliffe (1998) show that

$$[\sigma(K,T)]^2 = 2\frac{\partial c/\partial T + q(T)c + K[r(T) - q(T)]\partial c/\partial K}{K^2(\partial^2 c/\partial K^2)},$$

where c is here regarded as a function of S, K and T for the purposes of taking partial derivatives.

6 Theoretical Basis for Rules of Thumb

We are now in a position to consider whether the rules of thumb considered in Section 2 can be consistent with the no-arbitrage condition in equation (6). Similar results to ours on the sticky strike and sticky delta model are provided by Balland (2002).

In the basic sticky strike rule of Section 2.1 the variance V_{TK} is a function only of K and T. The analysis in Case 1 of the previous section shows that the only version of this model that is internally consistent is the model where the volatilities of all options are the same and constant. This is the original Black and Scholes (1973) model. In the generalized sticky strike model V_{TK} is independent of S, but possibly dependent on other stochastic variables. As shown in Case 2 of the previous section, the only version of this model that is internally consistent is the model where the instantaneous volatility of the asset price is a function only of time. This is Merton's (1973) model.

When the instantaneous volatility of the asset price is a function only of time, all European options with the same maturity have the same implied volatility. We conclude that all versions of the sticky strike rule are inconsistent with any type of volatility smile or volatility skew. If a trader prices options using different implied volatilities and the volatilities are independent of the asset price, there must be arbitrage opportunities.

Consider next the sticky delta rule. In the basic sticky delta model, as defined in Section 2.2, the implied volatility is a deterministic function of K/S and $T-t$. Case 3 in the previous section shows that the only version of this model that is internally consistent is Merton's (1973) model where the instantaneous volatility of the asset price is a function only of time. As in the case of the sticky strike model it is inconsistent with any type of volatility smile or volatility skew.

The generalized sticky delta model, where V_{TK} is stochastic and depends on K, S, T and t only through the variable K/S and $T-t$, can be consistent with the no-arbitrage condition. This is because equation (6) shows that if each θ_{TKi} depends on K, S, T, and t only through a dependence on K/S and $T-t$, the same is true of α_{TK}.

The relative sticky delta rule and the square root of time rule can be at least approximately consistent with the no-arbitrage condition. For example, for a mean-reverting stochastic volatility model, the implied volatility approximately satisfies the square root of time rule when the reversion coefficient is large (Andersson, 2003).

7 Tests of the Rules of Thumb

As pointed out by Derman (1999) apocryphal rules of thumb for describing the behavior of volatility smiles and skews may not be confirmed by the data. Derman's research looks at exchange-traded options on the S&P 500 during the period September 1997 to October 1998 and considers the sticky strike and sticky delta rules as well as a more complicated rule based on the IVF model. He finds subperiods during which each of the rules appears to explain the data best.

Derman's results are based on relatively short maturity S&P 500 option data. We use monthly volatility surfaces from the over-the-counter market for 47 months (June 1998 to April 2002). The data for June 1998 is shown in Table 1. Six maturities are considered ranging from six months to five years. Seven values of K/S are considered, ranging from 80 to 120. A total of 42 points on the volatility surface are therefore provided each month and the total number of volatilities in our data set is $42 \times 47 = 1974$. As illustrated in Table 1, implied volatilities for the S&P 500 exhibit a volatility skew with $\sigma_{TK}(t, S)$ being a decreasing function of K.

[Table 1 about here.]

The data was supplied to us by Totem Market Valuations Limited with the kind permission of a selection of Totem's major bank clients.[3] Totem collects implied volatility data in the form shown in Table 1 from a large number of dealers each month. These dealers are market makers in the over-the-counter market. Totem uses the data in conjunction with appropriate averaging procedures to produce an estimate of the mid-market implied volatility for each cell of the table and returns these estimates to the dealers. This enables dealers to check whether their valuations are in line with the market. Our data consist of the estimated mid-market volatilities returned to dealers. The data produced by Totem are considered by market participants to be more accurate than either the volatility surfaces produced by brokers or those produced by any one individual bank.

Consider first the sticky strike rule. Sticky strike rules where the implied volatility is a function only of K and T may be plausible in the exchange-traded market where the exchange defines a handful of options that trade and traders anchor on the volatility they first use for any one of these options. Our data comes from the over-the-counter market. It is difficult to see how this form of the sticky strike rule can apply in that market because there is continual trading in options with many different strike prices and times to maturity. A more plausible version of the rule is that the implied volatility is a function only of K and $T-t$. We tested this version of the rule using

$$\sigma_{TK}(t,S) = a_0 + a_1 K + a_2 K^2 + a_3(T-t) + a_4(T-t)^2 + a_5 K(T-t) + \epsilon, \tag{10}$$

where the a_i are constant and ϵ is a normally distributed error term. The terms on the right hand side of this equation can be thought of as the first few terms in a Taylor series expansion of a general function of K and $T-t$. Here our analysis is similar in spirit to the "ad-hoc model" of Dumas, Fleming and Whaley (1998). As shown in Table 2, the model is supported by the data, but has an R^2 of only 27%. As a test of the model's viability out-of-sample, we fitted the model using the first two years of data and then tested it for the surfaces for the remaining 23 months. We obtained a root mean square error (RMSE) of 0.0525, representing a quite sizable error.

[Table 2 about here.]

We now move on to test the sticky delta rule. The version of the rule we consider is the relative sticky delta model where $\sigma_{TK}(t,S) - \sigma_{TF}(t,S)$ is a function of K/F and $T-t$. The model we test is

$$\begin{aligned}\sigma_{TK}(t,S) - \sigma_{TF}(t,S) \quad &= \quad b_0 + b_1 \ln\left(\frac{K}{F}\right) + b_2\left[\ln\left(\frac{K}{F}\right)\right]^2 + b_3(T-t) \\ &\quad + b_4(T-t)^2 + b_5 \ln\left(\frac{K}{F}\right)(T-t) + \epsilon \end{aligned} \tag{11}$$

where the b_i are constants and ϵ is a normally distributed error term. The model is analogous to the one used to test the sticky strike rule. The terms on the right hand side of this equation can be

[3]Totem was acquired by Mark-it Partners in May 2004

thought of as the first few terms in a Taylor series expansion of a general function of $\ln(K/F)$ and $T-t$. The results are shown in Table 3. In this case the R^2 is much higher at 94.93%.

[Table 3 about here.]

Here we repeat our out-of-sample exercise, in this case with more satisfactory results: a RMSE of 0.0073 is obtained, representing a fairly close fit. In Table 3 it should be noted that the estimates obtained from using only the first half of the sample are quite close to those obtained using the whole data set.

If two models have equal explanatory power, then the observed ratio of the two models' squared errors should be distributed $F(N_1, N_2)$ where N_1 and N_2 are the number of degrees of freedom in the two models. When comparing the sticky strike and relative sticky delta models using a two tailed test, this statistic must be greater than 1.12 or less than 0.89 for significance at the 1% level. The value of the statistic is 32.71 indicating that we can overwhelmingly reject the hypothesis that the models have equal explanatory power. The relative sticky delta model in equation (11) can explain the volatility surfaces in our data much better than the sticky strike model in equation (10).

The third model we test is the version of the stationary square root of time rule where the function Φ in equation (1) does not change through time so that $\sigma_{TK}(t,S) - \sigma_{TF}(t,S)$ is a known function of $\ln(K/S)/\sqrt{T-t}$. Using a similar Taylor Series expansion to the other models we test

$$\begin{aligned}\sigma_{TK}(t,S) - \sigma_{TF}(t,S) \quad = \quad & c_1 \frac{\ln(K/F)}{\sqrt{T-t}} + c_2 \left(\frac{\ln(K/F)}{\sqrt{T-t}}\right)^2 \\ & + c_3 \left(\frac{[\ln(K/F)]}{\sqrt{T-t}}\right)^3 + c_4 \left(\frac{\ln(K/F)}{\sqrt{T-t}}\right)^4 + \epsilon \qquad (12)\end{aligned}$$

where c_1, c_2, c_3 and c_4 are constants and ϵ is a normally distributed error term. The results for this model are shown in Table 4. We consider more restricted forms of the model where progressively c_4 and c_3 are set to zero, to render a more parsimonious model for the volatility smile. In this case the R^2 is 97.12%. This is somewhat better than the R^2 for the sticky delta model in (11) even though the model in equation (12) involves two parameters and the the one in equation (11) involves six parameters.

[Table 4 about here.]

Testing the out of sample performance of the three possible forms of the square-root model gives RMSEs of 0.0069, 0.0069 and 0.0073 for the four term, three term and two term forms of the model. We conclude that for both in-sample and out of sample performance, including a cubic term leads to improvement in performance while the addition of a quartic term has negligible effect. Accordingly, in the following analysis, when working with the square-root rule, we restrict our attention to the cubic version.

We can calculate a ratio of sums of squared errors to compare the stationary square root of time model in equation (12) to the relative sticky delta model in equation (11). In this case, the ratio is 1.373.[4] When a two tailed test is used it is possible to reject the hypothesis that the two models have equal explanatory power at the 1% level. We conclude that equation (12) is an improvement over equation (11).

In the stochastic square root of time rule, the functional relationship between $\sigma_{TK}(t, S) - \sigma_{TF}(t, S)$ and

$$\frac{\ln(K/F)}{\sqrt{T-t}}$$

changes stochastically through time. A model capturing this is

$$\sigma_{TK}(t, S) - \sigma_{TF}(t, S) = c_1(t)\frac{\ln(K/F)}{\sqrt{T-t}} + c_2(t)\left(\frac{\ln(K/F)}{\sqrt{T-t}}\right)^2 + c_3(t)\left(\frac{\ln(K/F)}{\sqrt{T-t}}\right)^3 + \epsilon \tag{13}$$

where $c_1(t)$, $c_2(t)$ and $c_3(t)$ are stochastic. To provide a test of this model we fitted the model in equation (12) to the data on a month by month basis. When the model in equation (13) is compared to the model in equation (12) the ratio of sums of squared errors statistic is 3.11 indicating that the stochastic square root of time model does provide a significantly better fit to the data that the stationary square root of time model at the 1% level.

The coefficient, c_1, in the monthly tests of the square root of time rule is always significantly different from zero with a very high level of confidence. In almost all cases, the coefficients c_1, c_2 and c_3 are all significantly different from zero at the 5% level. We note, however, that in spite of this statistical significance, the most economically significant of the three terms is the linear term.[5]

Figure 1 shows the level of c_1, c_2 and c_3 plotted against the S&P 500 for the period covered by our data. We note that there is fairly weak correlation between c_1 and the level of the S&P 500 index. However, there is reasonably high correlation between c_2 and c_3 and the index, during the earlier part of our sample.

[Figure 1 about here.]

To provide an alternative benchmark for the square root of time rule, we now consider the possibility that volatility is a function of some other power of time to maturity. We consider both the stationary power rule model:

$$\begin{aligned}\sigma_{TK}(t, S) - \sigma_{TF}(t, S) &= d_1\left[\frac{\ln(K/F)}{(T-t)^{d_0}}\right] + d_2\left[\frac{\ln(K/F)}{(T-t)^{d_0}}\right]^2 \\ &\quad + d_3\left[\frac{\ln(K/F)}{(T-t)^{d_0}}\right]^3 + d_4\left[\frac{\ln(K/F)}{(T-t)^{d_0}}\right]^4 + \epsilon \end{aligned} \tag{14}$$

[4] The ratios for comparing the quadratic model and quartic model to (11) are 1.157 and 1.373 respectively

[5] The approximate linearity of the volatility skew for S&P 500 options has been mentioned by a number of researchers including Derman (1999).

and the stochastic power rule model:

$$\sigma_{TK}(t,S) - \sigma_{TF}(t,S) = d_1(t)\left[\frac{\ln(K/F)}{(T-t)^{d_0(t)}}\right] + d_2(t)\left[\frac{\ln(K/F)}{(T-t)^{d_0(t)}}\right]^2 + d_3(t)\left[\frac{\ln(K/F)}{(T-t)^{d_0(t)}}\right]^3 + d_4(t)\left[\frac{\ln(K/F)}{(T-t)^{d_0(t)}}\right]^4 + \epsilon \quad (15)$$

Estimating this model gives the results in Table 5. We note that the estimates of d_0 (the power of time to maturity which best explains volatility changes) are approximately equal to 0.44, and is statistically significantly different from 0.5. Performing an out of sample calculation of RMSE for the stationary power rule, we obtain 0.0070 for each model. All three seem to display reasonable parameter stability over time.

[Table 5 about here.]

Testing for significance between the performance of this power rule and the square-root rule, we obtain a ratio of squared errors of 1.04, insufficient to reject the hypothesis of equal performance. When the stochastic variants of each model are compared, the ratio becomes 5.39 showing that a considerable improvement in fit is obtained when the power relationship is allowed to vary with time. All our model comparison results are summarized in Table 6.

[Table 6 about here.]

For the model in equation (15) there is little or no relationship between the level of the index and the parameters of the rule. We do, however, note that in the later part of the data, there does seem to be evidence of the power rule being very close to the square root rule ($d_0 = 0.5$).

[Figure 2 about here.]

We conclude that there is reasonably strong support for the square-root rule as a parsimonious parametrisation of the volatility surface. There is some evidence that the rule could be marginally improved upon by considering powers other than 0.5.

8 Estimation of Volatility Factors

We also used the data described in the previous section to estimate the factors underlying the model in equation (3). We focus on implied volatility surfaces where moneyness is described as F/K rather than S/K, as originally reported. We create this data, and the attendant changes in log volatility, using interpolation.

The conventional analysis involves the use of principal components analysis applied to the variance-covariance matrix of changes in implied volatilities. The results of this are shown in Table

7 and Figure 3. This is a similar analysis to that undertaken by Kamal and Derman (1997), Skiadopoulos, Hodges and Clewlow (2000), Cont and Fonseca (2002) and Cont, Foseca and Durrleman (2002). Our study is different from these in that it focuses on over-the-counter options rather than exchange-traded options. It also uses options with much longer maturities.

8.1 Taking Account of the No-Arbitrage Drift

We now carry out a more sophisticated analysis that takes account of the no-arbitrage drift derived in Section 3. The traditional principal components analysis does not use any information about the drift of volatilities. In estimating the variance-covariance matrix, a constant drift for each volatility is assumed. From our results in Section 3 we know that in a risk-neutral world the drifts of implied volatilities depend on the factors. We assume that the market price of volatility risk is zero so that volatilities have the same drift in the real world as in the risk-neutral world. We also assume a form of the generalized sticky delta rule where each θ_{TKi} is a function K/F and $T-t$. Consistent with this assumption we write $V_{TK}(t)$ and $\theta_{TKi}(t)$ as $V_{(T-t),F/K,t}$ and $\theta_{(T-t),F/K,i}$.

Define $\hat{\alpha}$ as the drift of $\ln V_{(T-t),F/K,t}$. Applying Ito's lemma to equation (6) the process for $\ln V_{(T-t),F/K,t}$ is

$$d\ln V_{(T-t),F/K,t} = \hat{\alpha}dt + \sum_{i=1}^{N} \theta_{(T-t),F/K,i} dz_i,$$

where

$$\begin{aligned} \hat{\alpha}_{(T-t),F/K} &= \frac{1}{T-t}\left(1-\frac{\sigma^2}{V_{(T-t),F/K}}\right) + \sigma d_2 \frac{1}{\sqrt{V_{(T-t),F/K}(T-t)}} \sum_{i=1}^{N} \theta_{(T-t),F/K,i}\rho_i \\ & -\frac{(d_1 d_2+1)}{4}\sum_{i=1}^{N}\left[\theta^2_{(T-t),F/K,i} + \sum_{j\neq i}\rho_i\rho_j\theta_{(T-t),F/K,i}\theta_{(T-t),F/K,j}\right]. \end{aligned} \tag{16}$$

We can write the discretised process for $\ln V_{(T-t),F/K,t}$ as

$$\Delta \ln(V_{(T-t),F/K,t}) = \hat{\alpha}_{(T-t),F/K}\Delta t + \epsilon_{(T-t),F/K,t}. \tag{17}$$

Here, the term $\epsilon_{(T-t),F/K,t}$ (which has expectation zero) represents the combined effects of the factors on the movement of the volatility of an option with time to maturity $T-t$ and relative moneyness F/K, observed at time t.

The correlations between the ϵ's arise from a) the possibility that the same factor affects more than one option and b) the correlation between the factors arising from their correlation with the asset price. The covariance between two ϵ's is given by

$$E(\epsilon_{(T_1-t),(F/K_1),t}\epsilon_{(T_2-t),(F/K_2),t}) \quad = \quad \Delta t \sum_{k=1}^{N}\Big[\theta_{(T_1-t),(F/K_1),k}\theta_{(T_2-t),(F/K_2),k}$$

$$
\left. + \sum_{l \neq k} \theta_{(T_1-t),(F/K_1),k} \theta_{(T_2-t),(F/K_2),l} \rho_k \rho_l \right]
$$

$$
+ \begin{cases} \sigma_\epsilon^2 & \text{if } T_1 = T_2 \text{ and } K_1 = K_2. \\ 0 & \text{otherwise.} \end{cases} \tag{18}
$$

The analysis assumes no correlation between any volatility movements at different times. The σ_ϵ captures further effects over and above those given by the finite set of factors considered.

We carry out a maximum likelihood analysis to determine the values of the θ's, $\rho's$, and σ_ϵ that best describe the observed changes in option implied volatilities. As these parameters are changed, $\hat{\alpha}$ changes and there is a different covariance matrix for the errors.

For any set of parameters we back out an ϵ for each observation in our data set, using (17). Since we know that these ϵ's are jointly normally distributed, it is a standard result in econometrics[6] that the log likelihood function for this set of parameters is given by:

$$
L = \sum_{i=1}^{M} -\epsilon'_{t_i} (\Sigma_{t_i})^{-1} \epsilon_{t_i} - \ln \det(\Sigma_{t_i}) \tag{19}
$$

where ϵ_{t_i} is the vector of ϵ's for time t_i and Σ_{t_i} is the variance-covariance matrix described by (18) at time t_i.[7] To estimate the factors, we can thus use an optimization routine to maximize (19) by varying $\theta_{(T-t),F/K,j}$ (where $T-t$ and F/K are taken to be the varying levels of maturity and moneyness we observe in our volatility matrices), ρ_j and σ_ϵ^2. For an N factor model, describing a volatility matrix of M volatilities, this results in $N \times M + N + 1$ parameters. We used MATLAB's *fmincon* optimization routine (constraining the ρ's to lie between -1 and 1) and found that the routine is efficient for solving the problem, even for this large number of parameters.

Fitting a four factor model ($N = 4$) we find the factors given in Figure 4 and Table 8. To ease the comparison with the results from Figure 3 and Table 7, we have converted our factors from a collection of factors each correlated with stock price movements (and therefore each other) to a collection of five orthonormal factors, one of which (factor zero) is driven by the same Brownian Motion as in the underlying stock price. The normalization results in the factors having to be multiplied by a scaling factor, which allows the familiar analysis of considering the proportion of the covariance matrix of the surface which is explained by each factor.

Focusing on factor zero (the factor which is perfectly correlated with stock prices) we see that an increase in stock prices leads to an increase in the slope of implied volatilities. In the case of the typical skew observed, this would result in a decrease in stock prices causing the skew to steepen, rendering out of the money puts more valuable. The effect is most pronounced for short maturity options, and tapers off for longer maturity options.

[6] See, for example Greene (1997), page 89.

[7] It is worth observing that our model's intertemporally independent residuals allows us to consider the likelihood ratio as the sum of individual time steps' log likelihood ratios. This considerably speeds up the calculation of both the inner product term $(\epsilon'_{t_i} (\Sigma_{t_i})^{-1} \epsilon_{t_i})$ and $\ln \det(\Sigma_{t_i})$, which (for a large sample) could involve quite large matrices.

The most important of all the factors is the first one. This factor shows evidence of hyperbolic behaviour in the time dimension, and works to offset the explosive behaviour we noted in section 3. It is roughly level in the moneyness dimension, and can therefore be seen as representing moves up and down of the existing smile. Factors two and three show evidence of a square root effect, having roughly a square-root shape in the time dimension, but both showing evidence of increasing sensitivity for higher strike options. Factor four is less easy to interpret, seeming to be a factor which affects low moneyness, long maturity options.

With four factors, we can explain 99.6% of the variation in implied volatilities, with both the last two factors making a significant contribution.

Contrasting the results of our consistent factor analysis with the simpler approach originally followed, we note that we obtain a richer factor structure. The principal components approach gives two dominant factors which explain 98.8% of the variation. The first factor is similar to that uncovered by the drift consistent approach, while the second factor shows some evidence of a square-root behaviour in the maturity dimension.

[Figure 3 about here.]

[Figure 4 about here.]

[Table 7 about here.]

[Table 8 about here.]

9 Summary

It is a common practice in the over-the-counter markets to quote option prices using their Black–Scholes implied volatilities. In this paper we have developed a model of the evolution of implied volatilities and produced a no-arbitrage condition that must be satisfied by the volatilities. Our model is exactly consistent with the initial volatility surface, but more general than the IVF model of Rubinstein (1994), Derman and Kani (1994), and Dupire (1994). The no-arbitrage condition leads to the conclusions that a) when the volatility is independent of the asset price there must be a U-shaped volatility smile and b) when the implied volatility is a decreasing (increasing) function of the asset price there must be a negative (positive) correlation between the volatility and the asset price. We outline how these no-arbitrage conditions can be incorporated into a factor analysis, resulting in no-arbitrage consistent factors. Our empirical implementation of this analysis suggests that more factors may in fact govern volatility movements than would be thought otherwise.

A number of rules of thumb have been proposed for how traders manage the volatility surface. These are the sticky strike, sticky delta, and square root of time rules. Some versions of these rules are clearly inconsistent with the no-arbitrage condition; for other versions of the rules the no-arbitrage condition can in principle be satisfied.

Our empirical tests of the rules of thumb using 47 months of volatility surfaces for the S&P 500 show that the relative sticky delta model (where the excess of the implied volatility of an option over the corresponding at-the-money volatility is a function of moneyness) outperforms the sticky strike rule. Also, the stochastic square root of time model outperforms the relative sticky delta rule.

A Proof of Equation (4)

For simplicity of notation, we assume that r and q are constants. Define $F(\tau)$ as the forward price at time τ for a contract maturing at time $t + \Delta t$ so that

$$F(t) = S(t)e^{(r-q)\Delta t}.$$

The price at time t of a call option with strike price $F(t)$ and maturity $t + \Delta t$ is given by

$$c(S(t), V_{t+\Delta t, F(t)}, t; F(t), t + \Delta t) = e^{-q\Delta t}S(t)[N(d_1) - N(d_2)],$$

where in this case

$$d_1 = -d_2 = \frac{\sqrt{\Delta t}}{2}\sigma_{t+\Delta t, F(t)}(t, S(t)).$$

The call price can be written

$$c(S(t), V_{t+\Delta t, F(t)}, t; F(t), t + \Delta t) = e^{-q\Delta t}S(t)\int_{-d_1}^{d_1} \phi(x)\ dx,$$

where

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

For some $\bar{x}$

$$\int_{-d_1}^{d_1} \phi(x)\ dx = 2d_1\phi(\bar{x})$$

and the call price is therefore given by

$$c(S(t), V_{t+\Delta t, F(t)}, t; F(t), t + \Delta t) = 2e^{-q\Delta t}S(t)d_1\phi(\bar{x}),$$

or

$$c(S, V_{t+\Delta t, F(t)}, t; F(t), t + \Delta t) = e^{-q\Delta t}S(t)\phi(\bar{x})\sigma_{t+\Delta t, F(t)}(t, S(t))\sqrt{\Delta t}. \tag{20}$$

The process followed by S is

$$\frac{dS}{S} = (r - q)\,dt + \sigma\,dz.$$

Using Ito's lemma

$$dF = \sigma F\,dz.$$

When terms of order higher that $(\Delta t)^{1/2}$ are ignored

$$F(t+\Delta t) - F(t) = \sigma(t)F(t)[z(t+\Delta t) - z(t)].$$

It follows that

$$\lim_{\Delta t \to 0} \frac{1}{\sqrt{\Delta t}} E[F(t+\Delta t) - F(t)]^+ = \lim_{\Delta t \to 0} \frac{1}{\sqrt{\Delta t}} \sigma(t)F(t)E[z(t+\Delta t) - z(t)]^+,$$

where E denotes expectations under the risk-neutral measure.

Because

$$E[F(t+\Delta t) - F(t)]^+ = e^{r\Delta t} c(S(t), V_{t+\Delta t, F(t)}, t; F(t), t+\Delta t),$$

and

$$\lim_{\Delta t \to 0} \frac{1}{\sqrt{\Delta t}} E[z(t+\delta t) - z(t)]^+ = \frac{1}{\sqrt{2\pi}},$$

it follows that

$$\lim_{\Delta t \to 0} \frac{1}{\sqrt{\Delta t}} e^{r\Delta t} c(S(t), V_{t+\Delta t, F(t)}, t; F(t), t+\Delta t) = \frac{\sigma(t)F(t)}{\sqrt{2\pi}}.$$

Substituting from equation (20)

$$\lim_{\Delta t \to 0} e^{r\Delta t} e^{-q\Delta t} S(t)\phi(\bar{x})\sigma_{t+\Delta t, F(t)}(t, S(t)) = \frac{\sigma(t)F(t)}{\sqrt{2\pi}}.$$

As Δt tends to zero, $\phi(\bar{x})$ tends to $1/\sqrt{2\pi}$ so that

$$\lim_{\Delta t \to 0} \sigma_{t+\Delta t, F(t)}(t, S(t)) = \sigma(t).$$

This is the required result.

References

Andersson, K. (2003) "Stochastic volatility," U.U.D.M. Project Report, Department of Mathematics, Uppsala University

Balland, P. (2002) "Deterministic implied volatility models," *Quantitative Finance*, 2, 31–44

Black, F. and M.S. Scholes (1973), "The pricing of options and corporate liabilities," *Journal of Political Economy*, 81, 637–659

Bates, D. S. (1996), "Jumps and Stochastic Volatility: Exchange Rate Process Implicit in Deutsche Mark Options," *Review of Financial Studies*, 9, 1, 69–107

Brace, A., B. Goldys, F. Klebaner, and R. Womersley (2001), "Market model of stochastic volatility with applications to the BGM model," Working paper S01-1, Dept of Statistics, University of New South Wales

Britten–Jones, M., and A. Neuberger (2000) "Option Prices, Implied Price Processes, and Stochastic Volatility," *Journal of Finance*, 55, 2 (2000), 839–866

Cont, R., J. Fonseca, and V. Durrleman (2002) "Stochastic Models of Implied Volatility Surfaces," *Economic Notes*, 31, 2 (2002), 361–377

Cont, R., and J. Fonseca (2002), "Dynamics of implied volatility surfaces," *Quantitative Finance*, 2, 45–60

Christie, A. A. (1982), "The stochastic behavior of common stock variances: Value, leverage, and interest rate effects," *Journal of Financial Economics*, 10, 4, 407–432

Derman, E. (1999), "Regimes of volatility," *Quantitative Strategies Research Notes*, Goldman Sachs, New York, NY. Also in *Risk Magazine*, April, 1999

Derman, E. and I. Kani (1994a), "The volatility smile and its implied tree," *Quantitative Strategies Research Notes*, Goldman Sachs, New York, NY

Dumas, B., J. Fleming and R.E. Whaley (1997), "Implied volatility functions: empirical tests," *Journal of Finance* 6, 2059–2106

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

Greene, W. (1997), *Econometric Analysis*, 3rd Edition, Prentice Hall, Upper Saddle River, NJ

Heath, D., R. Jarrow, and A. Morton (1992), "Bond Pricing and the Term Structure of the Interest Rates: A New Methodology," *Econometrica*, 60, 1, 77–105

Heston, S.L. (1993), "A closed-form solution for options with stochastic volatility applications to bond and currency options," *Review of Financial Studies*, 6, 327–343

Hull, J. (2002) *Options, Futures, and Other Derivatives*, 5th Edition, Prentice Hall, Englewood Cliffs, NJ

Hull, J. and W. Suo (2002), "A methodology for assessing model risk and its application to the implied volatility function model," *Journal of Financial and Quantitative Analysis*, 37, 2, 297–318

Hull, J. and A. White (1987), "The pricing of options with stochastic volatility," *Journal of Finance*, 42, 281–300

Jackwerth, J.C. and M. Rubinstein (1996), "Recovering probabilities from option prices", *Journal of Finance*, 51, 1611–1631

Kamal, M. and E. Derman, “The patterns of change in implied index volatilities,” *Quantitative Strategies Research Notes*, Goldman Sachs, New York, NY

LeDoit O. and P. Santa Clara (1998) “Relative pricing of options with stochastic volatility” Working Paper, Anderson Graduate School of Management, University of California, Los Angeles

Merton, R. C. (1973), “Theory of rational option pricing,” *Bell Journal of Economics and Management Science*, 4, 141–183

Natenberg S. (1994) *Option Pricing and Volatility: Advanced Trading Strategies and Techniques*, 2nd ed. McGraw-Hill

Rubinstein, M. (1994), “Implied binomial trees,” *Journal of Finance*, 49, 771–818

Schönbucher, P.J. (1999), “A market model of stochastic implied volatility,” *Philosophical Transactions of the Royal Society*, Series A, 357, 2071–2092

Skiadopolous, G., S. Hodges and L. Clewlow (2000), “Dynamics of the S&P500 implied volatility surface,” *Review of Derivatives Research*, 3, 263–82

Stein, E. and C. Stein (1991), “Stock price distributions with stochastic volatilities: An analytical approach,” *Review of Financial Studies*, 4, 727–752

List of Figures

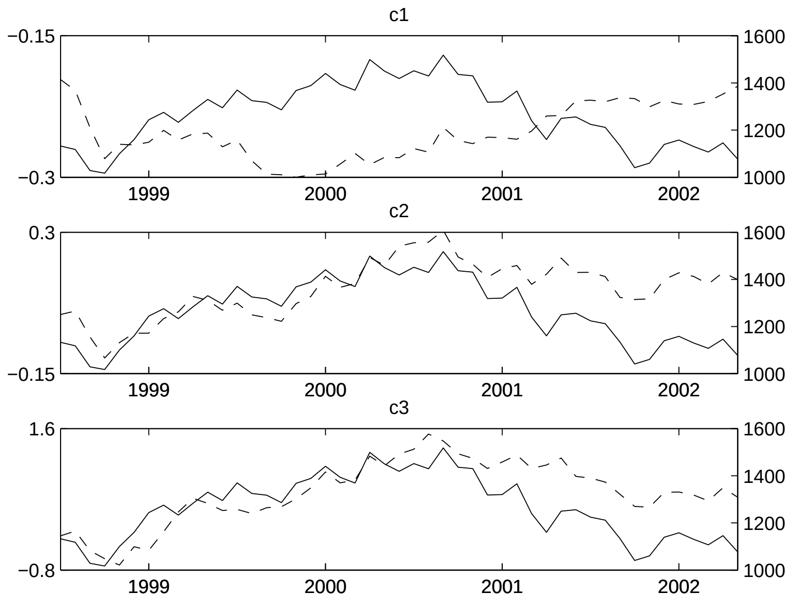

Figure 1: Plot of the level of the S&P 500 index (solid line, with scale on the right axis) and the estimates of c_1, c_2 and c_3 (dashed line, with scale on the left axes). Note the positive correlations between c_2 and c_3 and the index during the earlier part of the sample.

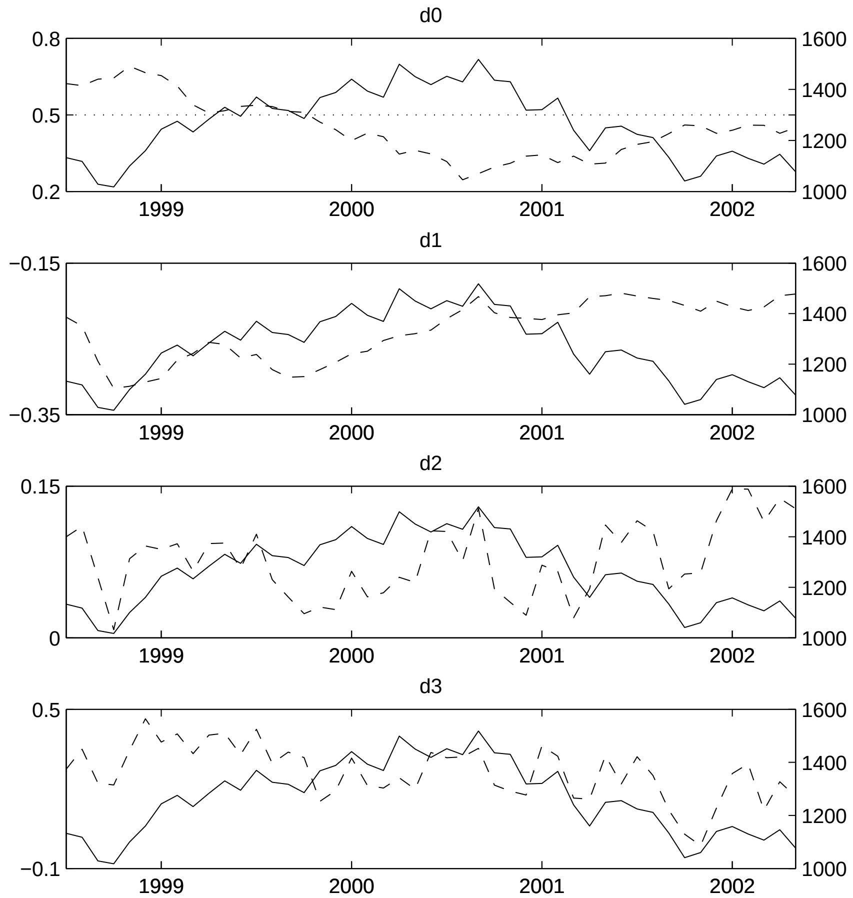

Figure 2: Plot of the level of the S&P 500 index (solid line, with scale on the right axis) and the estimates of d_0, d_1, d_2 and d_3 (dashed line, with scale on the left axes). Note the positive correlations between c_3 and the index. The dotted line on the first graph shows the level of d_0 (0.5) which would be consistent with the square-root rule.

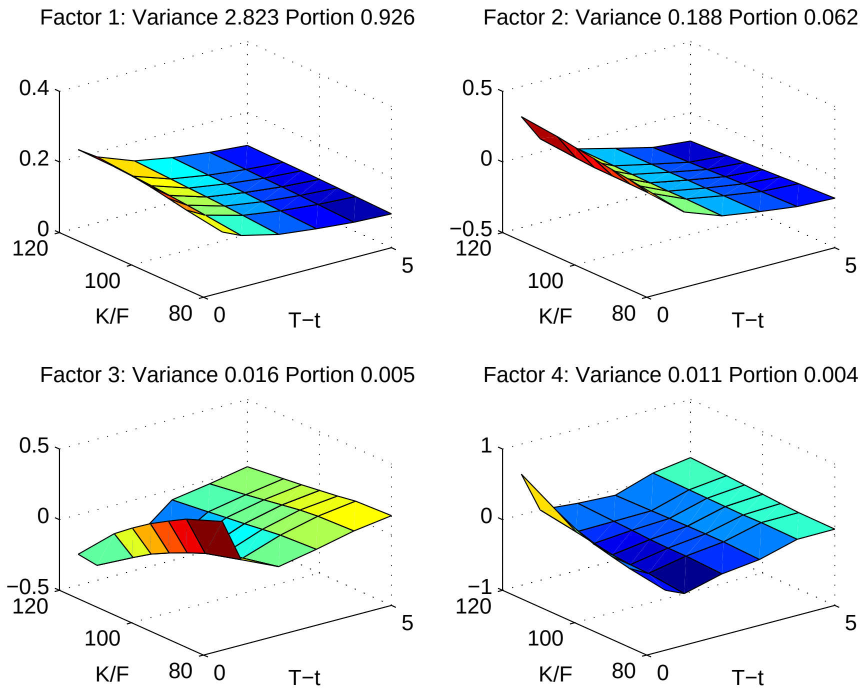

Figure 3: Principal Component factors for the Totem data. The height of the surface measures the normalised factor, for each maturity and moneyness combination.

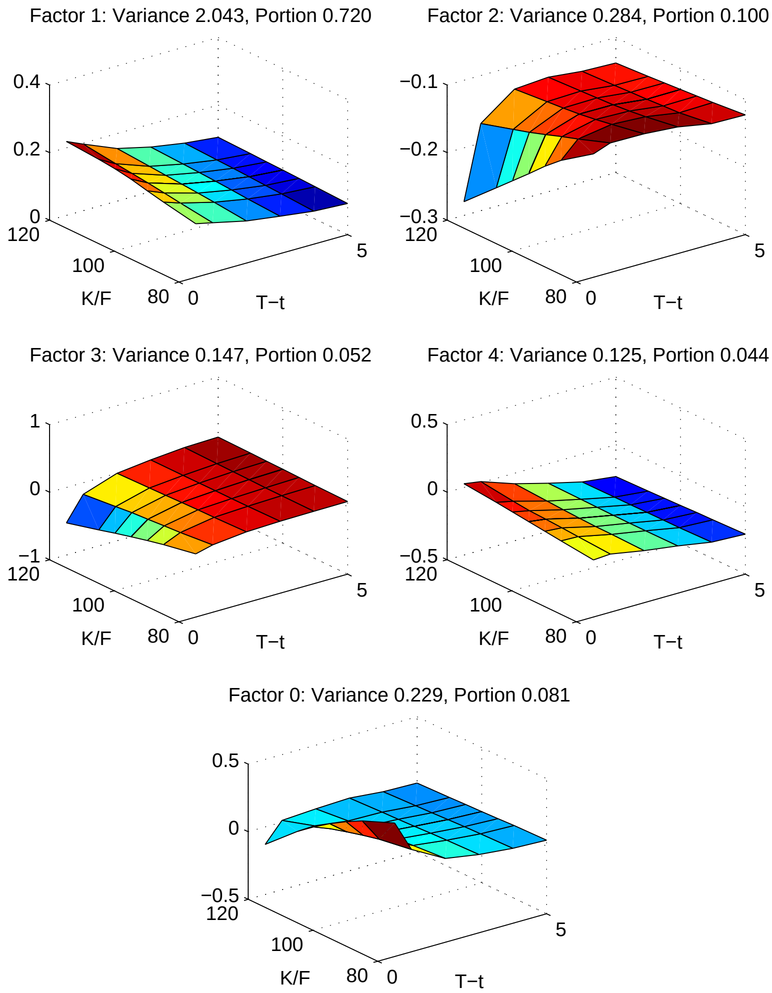

Figure 4: Estimated factors for the Totem data, accounting for no-arbitrage drift in volatilities. The height of the surface measures the value of the normalised factor for each maturity and moneyness combination. Factors 1,2,3 and 4 are independent of each other and the stock price, while factor 0 is perfectly correlated with the stock price.

List of Tables

	Time to Maturity (months)					
Strike	6	12	24	36	48	60
120	15.91%	18.49%	20.20%	21.03%	21.44%	21.64%
110	18.13%	20.33%	21.48%	21.98%	22.18%	22.30%
105	19.50%	21.26%	22.21%	22.52%	22.59%	22.70%
100	20.94%	22.38%	23.06%	23.14%	23.11%	23.07%
95	22.73%	23.71%	23.92%	23.73%	23.58%	23.53%
90	24.63%	24.99%	24.78%	24.40%	24.10%	23.96%
80	28.41%	27.71%	26.66%	25.83%	25.23%	24.83%

Table 1: Volatility matrix for June 1998. Time to maturity is measured in months while strike is in percentage terms, relative to the level of the S&P 500 index.

Variable	Estimate	Standard Error	t-statistic	Estimate from first 2 years
a_0	0.4438616	0.0238337	18.62	0.5405505
a_1	-0.0001944	0.000036	-5.40	-0.0002511
a_2	0.0000000	0.0000000	1.44	0.0000000
a_3	-0.0262681	0.0033577	-7.82	-0.0335188
a_4	-0.0006589	0.0003454	-1.91	-0.0011062
a_5	0.0000290	0.0000022	13.30	0.0000355
R^2	0.2672			
Standard error of residuals	0.0327			

Table 2: Estimates for the version of the sticky strike model in equation (10) where volatility depends on strike price and time to maturity.

Variable	Estimate	Standard Error	t-statistic	Estimates from first 2 years
b_0	0.0058480	0.0003801	15.39	0.0067000
b_1	-0.2884075	0.0019565	-147.41	-0.3345461
b_2	0.0322727	0.0067576	4.78	0.0112730
b_3	-0.0075740	0.0003487	-21.72	-0.0091289
b_4	0.0015705	0.0000701	22.42	0.0018404
b_5	0.0414902	0.0009180	45.20	0.0496937
R^2	0.9493			
Standard error of residuals	0.0057			

Table 3: Estimates for the version of the relative sticky delta model in equation (11).

Variable	Estimate	Standard Error	t-statistic	Estimates from first 2 years
		Unrestricted		
c_1	-0.2529421	0.0022643	-111.7102965	-0.2702221
c_2	0.1423444	0.0124579	11.4260104	0.0675920
c_3	0.5101281	0.0572173	8.9156322	0.2231858
c_4	-0.2067651	0.2106077	-0.9817545	-0.1157022
R^2	0.9631			
Standard error of residuals	0.0049			
		$c_4 = 0$		
c_1	-0.2547240	0.0013537	-188.1644600	-0.2712324
c_2	0.1319096	0.0064984	20.2987718	0.0620709
c_3	0.5595131	0.0272658	20.5206944	0.2522889
R^2	0.9631			
Standard error of residuals	0.0049			
		$c_3 = c_4 = 0$		
c_1	-0.2402662	0.0012591	-190.8294624	-0.2654298
c_2	0.0565335	0.0058389	9.6821751	0.0240899
R^2	0.9562			
Standard error of residuals	0.0052			

Table 4: Estimates for the stationary square root of time rule in equation (12).

Parameter	Estimate	Standard Error	t-statistic	Estimates from first two years
		Unrestricted		
d_0	0.4392204	0.0052036	84.4076018	0.5114474
d_1	-0.2416622	0.0023361	-103.4488377	-0.2714956
d_2	0.0725670	0.0152348	4.7632480	0.0864143
d_3	0.3307833	0.0617382	5.3578373	0.2373512
d_4	0.3775721	0.2574542	1.4665602	-0.3039707
R^2	0.9652			
Standard error of residuals	0.0047			
		$d_4 = 0$		
d_0	0.4419942	0.0048206	91.6878322	0.5083725
d_1	-0.2394926	0.0018407	-130.1086545	-0.2733074
d_2	0.0912235	0.0076917	11.8600018	0.0674926
d_3	0.2581095	0.0390363	6.6120391	0.2901229
R^2	0.9651			
Standard error of residuals	0.0047			
		$d_4 = d_3 = 0$		
d_0	0.4220134	0.0034726	121.5254560	0.4825231
d_1	-0.2308670	0.0012548	-183.9941367	-0.2632187
d_2	0.0540911	0.0055657	9.7185927	0.0238527
R^2	0.9644			
Standard error of residuals	0.0048			

Table 5: Estimates of the power rule from equation (14).

Test	Statistic
Sticky Strike (10) versus Sticky delta (11)	32.71
Sticky Delta (11) versus (cubic) Stationary Square-root rule (12)	1.37
Stationary Square-root rule (12) versus Stochastic Square-root rule (13)	3.11
Stationary Square-root rule (12) versus Stationary power rule (14)	1.04
Stochastic Square-root rule (13) versus Stochastic power rule (15)	5.39

Table 6: Comparison of Models using the ratio of sums of squared errors statistic. In a two-tailed test the statistic must be greater than 1.123 (1.111) to reach the conclusion that the first equation to provide a better explanation of the data than the second equation at the 1% (5%) level

Factor 1			Variance=2.823464				Factor 2			Variance=0.187797			
			Portion=0.926429							Portion=0.061620			
			Time to Maturity (months)							Time to Maturity (months)			
Strike	6	12	24	36	48	60	Strike	6	12	24	36	48	60
0.80	0.17	0.15	0.12	0.11	0.10	0.10	0.80	0.15	0.03	-0.06	-0.10	-0.14	-0.15
0.90	0.19	0.16	0.13	0.12	0.11	0.10	0.90	0.18	0.06	-0.06	-0.11	-0.15	-0.16
0.95	0.20	0.16	0.13	0.12	0.11	0.10	0.95	0.20	0.06	-0.06	-0.11	-0.15	-0.17
1.00	0.21	0.17	0.14	0.12	0.11	0.10	1.00	0.22	0.06	-0.06	-0.11	-0.16	-0.18
1.00	0.21	0.17	0.14	0.12	0.11	0.10	1.00	0.22	0.06	-0.06	-0.11	-0.16	-0.18
1.05	0.21	0.17	0.14	0.12	0.11	0.10	1.05	0.24	0.06	-0.06	-0.11	-0.16	-0.18
1.10	0.22	0.18	0.14	0.12	0.11	0.10	1.10	0.26	0.07	-0.05	-0.11	-0.17	-0.19
1.20	0.22	0.19	0.15	0.13	0.11	0.11	1.20	0.29	0.10	-0.05	-0.11	-0.18	-0.20
Factor 3			Variance=0.015721				Factor 4			Variance=0.010684			
			Portion=0.005158							Portion=0.003505			
			Time to Maturity (months)							Time to Maturity (months)			
Strike	6	12	24	36	48	60	Strike	6	12	24	36	48	60
0.80	0.41	0.10	-0.02	0.04	0.10	0.14	0.80	-0.15	-0.27	-0.13	-0.07	0.07	0.08
0.90	0.31	0.03	-0.07	0.02	0.09	0.12	0.90	-0.07	-0.21	-0.11	-0.05	0.08	0.09
0.95	0.24	-0.02	-0.09	0.01	0.07	0.10	0.95	-0.02	-0.20	-0.11	-0.05	0.08	0.10
1.00	0.17	-0.08	-0.12	-0.01	0.05	0.08	1.00	0.05	-0.18	-0.10	-0.05	0.08	0.12
1.00	0.17	-0.08	-0.12	-0.01	0.05	0.08	1.00	0.05	-0.18	-0.10	-0.05	0.08	0.12
1.05	0.08	-0.14	-0.15	-0.03	0.03	0.07	1.05	0.15	-0.13	-0.09	-0.06	0.09	0.13
1.10	-0.02	-0.22	-0.20	-0.04	0.01	0.05	1.10	0.27	-0.08	-0.07	-0.07	0.09	0.15
1.20	-0.28	-0.39	-0.28	-0.07	-0.03	0.02	1.20	0.58	-0.00	-0.05	-0.08	0.10	0.18

Table 7: Factors obtained by Principal Components analysis applied to changes in the volatility surface.

Factor 1			Variance=2.043035 Portion=0.719536 Time to Maturity (months)			
Strike	6	12	24	36	48	60
0.80	0.16	0.15	0.12	0.11	0.10	0.09
0.90	0.18	0.16	0.13	0.12	0.10	0.10
0.95	0.19	0.17	0.14	0.12	0.11	0.10
1.00	0.20	0.17	0.14	0.12	0.11	0.10
1.00	0.20	0.17	0.14	0.12	0.11	0.10
1.05	0.21	0.18	0.15	0.13	0.11	0.10
1.10	0.21	0.18	0.15	0.13	0.11	0.10
1.20	0.22	0.20	0.16	0.13	0.12	0.11

Factor 2			Variance=0.283968 Portion=0.100011 Time to Maturity (months)			
Strike	6	12	24	36	48	60
0.80	-0.12	-0.11	-0.11	-0.11	-0.12	-0.12
0.90	-0.15	-0.13	-0.12	-0.12	-0.13	-0.13
0.95	-0.17	-0.13	-0.12	-0.12	-0.13	-0.13
1.00	-0.19	-0.14	-0.13	-0.13	-0.13	-0.13
1.00	-0.19	-0.14	-0.13	-0.13	-0.13	-0.13
1.05	-0.21	-0.15	-0.13	-0.13	-0.13	-0.13
1.10	-0.24	-0.16	-0.13	-0.13	-0.13	-0.13
1.20	-0.28	-0.17	-0.13	-0.13	-0.14	-0.14

Factor 3			Variance=0.146822 Portion=0.051709 Time to Maturity (months)			
Strike	6	12	24	36	48	60
0.80	-0.07	-0.00	0.05	0.07	0.07	0.08
0.90	-0.16	-0.05	0.03	0.06	0.07	0.08
0.95	-0.21	-0.07	0.03	0.06	0.08	0.09
1.00	-0.27	-0.08	0.02	0.06	0.08	0.09
1.00	-0.27	-0.08	0.02	0.06	0.08	0.09
1.05	-0.33	-0.10	0.02	0.06	0.09	0.10
1.10	-0.40	-0.12	0.01	0.06	0.09	0.10
1.20	-0.52	-0.17	-0.00	0.06	0.10	0.11

Factor 4			Variance=0.125169 Portion=0.044083 Time to Maturity (months)			
Strike	6	12	24	36	48	60
0.80	-0.08	-0.07	-0.11	-0.15	-0.19	-0.20
0.90	-0.05	-0.06	-0.11	-0.15	-0.20	-0.21
0.95	-0.04	-0.05	-0.11	-0.15	-0.20	-0.22
1.00	-0.02	-0.04	-0.10	-0.15	-0.20	-0.22
1.00	-0.02	-0.04	-0.10	-0.15	-0.20	-0.22
1.05	-0.01	-0.04	-0.10	-0.15	-0.20	-0.22
1.10	0.00	-0.02	-0.09	-0.15	-0.20	-0.23
1.20	0.03	0.01	-0.08	-0.14	-0.21	-0.24

Factor 0			Variance=0.229168 Portion=0.080711 Time to Maturity (months)			
Strike	6	12	24	36	48	60
0.80	0.49	0.26	0.12	0.08	0.05	0.04
0.90	0.39	0.22	0.10	0.07	0.05	0.04
0.95	0.32	0.19	0.09	0.06	0.04	0.03
1.00	0.25	0.16	0.08	0.06	0.04	0.03
1.00	0.25	0.16	0.08	0.06	0.04	0.03
1.05	0.17	0.12	0.07	0.05	0.03	0.02
1.10	0.08	0.09	0.05	0.05	0.03	0.02
1.20	-0.13	0.01	0.03	0.04	0.02	0.01

Table 8: Factors obtained by using maximum likelihood estimation, incorporating the no-arbitrage consistent drift. The factors are all orthogonal, with factors 1, 2, 3 and 4 being uncorrelated with the stock price Brownian Motion, and factor 0 being perfectly correlated with the stock price.