Implied Volatility: Statics, Dynamics, and Probabilistic Interpretation

Roger W. Lee*

November 22, 2002

Forthcoming in *Recent Advances in Applied Probability*

Abstract

Given the price of a call or put option, the Black-Scholes *implied volatility* is the unique volatility parameter for which the Bulack-Scholes formula recovers the option price. This article surveys research activity relating to three theoretical questions: First, does implied volatility admit a probabilistic interpretation? Second, how does implied volatility behave as a function of strike and expiry? Here one seeks to characterize the shapes of the implied volatility skew (or smile) and term structure, which together constitute what can be termed the *statics* of the implied volatility surface. Third, how does implied volatility evolve as time rolls forward? Here one seeks to characterize the *dynamics* of implied volatility.

1 Introduction

1.1 Implied volatility

Assuming that an underlying asset in a frictionless market follows geometric Brownian motion, which has constant volatility, the Black-Scholes formula gives the no-arbitrage price of an option on that underlying. Inverting this formula, take as given the price of a call or put option. The Black-Scholes *implied volatility* is the unique volatility parameter for which the Black-Scholes formula recovers the price of that option.

This article surveys research activity in the theory of implied volatility. In light of the compelling empirical evidence that volatility is *not* constant, it is natural to question why the inversion of option prices in an "incorrect" formula should deserve such attention.

To answer this, it is helpful to regard the Black-Scholes implied volatility as a *language* in which to express an option price. Use of this language does

*Department of Mathematics, Stanford University; and Courant Institute of Mathematical Sciences, NYU. This work was partially supported by an NSF Postdoctoral Fellowship. I thank Peter Carr and Jim Gatheral for their comments.

not entail any belief that volatility is actually constant. A relevant analogy is the quotation of a discount bond price by giving its yield to maturity, which is the interest rate such that the observed bond price is recovered by the usual *constant* interest rate bond pricing formula. In no way does the use or study of bond yields entail a belief that interest rates are actually constant. As YTM is just an alternative way of expressing a bond price, so is implied volatility just an alternative way of expressing an option price.

The language of implied volatility is, moreover, a *useful* alternative to raw prices. It gives a metric by which option prices can be compared across different strikes, maturities, underlyings, and observation times; and by which market prices can be compared to assessments of fair value. It is a standard in industry, to the extent that traders quote option prices in “vol” points, and exchanges update implied volatility indices in real time.

Furthermore, to whatever extent implied volatility has a simple interpretation as an average future volatility, it becomes not only useful, but also *natural.* Indeed, understanding implied volatility as an average will be one of the focal points of this article.

1.2 Outline

Under one interpretation, implied volatility is the market’s expectation of future volatility, time-averaged over the term of the option. In what sense does this interpretation admit mathematical justification? In section 2 we review the progress on this question, in two contexts: first, under the assumption that instantaneous volatility is a deterministic function of the underlying and time; and second, under the assumption that instantaneous volatility is stochastic in the sense that it depends on a second random factor.

If instantaneous volatility is not constant, then implied volatilities will exhibit variation with respect to strike (described graphically as a *smile* or *skew*) and with respect to expiry (the *term structure*); the variation jointly in strike and expiry can be described graphically as a *surface.* In section 3, we review the work on characterizing or approximating the shape of this surface under various sets of assumptions. Assuming only absence of arbitrage, one finds bounds on the slope of the volatility surface, and characterizations of the tail growth of the volatility skew. Assuming stochastic volatility dynamics for the underlying, one finds perturbation approximations for the implied volatility surface, in any of a number of different regimes, including long maturity, short maturity, fast mean reversion, and slow mean reversion.

Whereas sections 2 and 3 examine how implied volatility behaves under certain assumptions on the spot process, section 4 directly takes as primitive the implied volatility, with a view toward modelling accurately its time-evolution. We begin with the no-arbitrage approach to the direct modelling of stochastic implied volatility. Then we review the statistical approach. Whereas the focus of section 3 is cross-sectional (taking a “snapshot” of all strikes and expiries) hence the term *statics*, the focus of section 4 is instead time-series oriented, hence the term *dynamics.*

1.3 Definitions

Our underlying asset will be a non-dividend paying stock or index with nonnegative price process S_t. Generalization to non-zero dividends is straightforward.

A call option on S, with strike K and expiry T, pays $(S_T - K)^+$ at time T. The price of this option is a function C of the contract variables (K,T), today's date t, the underlying S_t, and any other state variables in the economy. We will suppress some or all of these arguments. Moreover, sections 2 and 3 will for notational convenience assume $t = 0$ unless otherwise stated; but section 4, in which the time-evolution of option prices becomes more important, will not assume $t = 0$.

Let the risk-free interest rate be a constant r. Write

$$x := \log\left(\frac{K}{S_t e^{r(T-t)}}\right)$$

for *log-moneyness* of an option at time t. Note that both of the possible choices of sign convention appear in the literature; we have chosen to define log-moneyness to be such that x has a *positive* relationship with K.

Assuming frictionless markets, Black and Scholes [8] showed that if S follows geometric Brownian motion

$$dS_t = \mu S_t dt + \sigma S_t d\tilde{W}_t$$

then the no-arbitrage call price satisfies

$$C = C^{BS}(\sigma),$$

where the Black-Scholes formula is defined by

$$C^{BS}(\sigma) := C^{BS}(S_t, t, K, T, \sigma) := S_t N(d_1) - K e^{-r(T-t)} N(d_2).$$

Here

$$d_{1,2} := \frac{\log(S_t e^{r(T-t)}/K)}{\sigma\sqrt{T-t}} \pm \frac{\sigma\sqrt{T-t}}{2},$$

and N is the cumulative normal distribution function.

On the other hand, given $C(K,T)$, the *implied [Black-Scholes] volatility* for strike K and expiry T is defined as the $I(K,T)$ that solves

$$C(K,T) = C^{BS}(K, T, I(K,T)).$$

The solution is unique because C^{BS} is strictly increasing in σ, and as $\sigma \to 0$ (resp. ∞), the Black-Scholes function $C^{BS}(\sigma)$ approaches the lower (resp. upper) no-arbitrage bounds on a call.

Implied volatility can also be written as a function $\tilde{I}$ of log-moneyness and time, so $\tilde{I}(x,T) := I(S_t e^{x+r(T-t)}, T)$. Abusing notation, we will drop the tilde on $\tilde{I}$, because the context will make clear whether I is to be viewed as a function of K or x.

The derivation of the Black-Scholes formula can proceed by means of a hedging argument that yields a PDE to be solved for $C(S,t)$:

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC = 0, \tag{1.1}$$

with terminal condition $C(S,T) = (S-K)^+$. Alternatively, one can appeal to martingale pricing theory, which guarantees that in the absence of arbitrage (appropriately defined – see for example [16]), there exists a "risk-neutral" probability measure under which the discounted prices of all tradeable assets are martingales. We assume such conditions, and unless otherwise stated, our references to probabilities, distributions, and expectations will be with respect to such a pricing measure, not the statistical measure. In the constant-volatility case, changing from the statistical to the pricing measure yields

$$dS_t = rS_t dt + \sigma S_t dW_t.$$

So $\log S_T$ is normal with mean $(r-\sigma^2/2)(T-t)$ and variance $\sigma^2(T-t)$, and the Black-Scholes formula follows from $C = e^{-r(T-t)}\mathbb{E}(S_T-K)^+$.

2 Probabilistic Interpretation

In what sense is implied volatility an average expected volatility? Some econometric studies [11, 13] test whether or not implied volatility is an "unbiased" predictor of future volatility, but they have limited relevance to our question, because they address the empirics of a far narrower question in which "expected" future volatility is with respect to the *statistical* probability measure. Our focus, instead, is the theoretical question of whether there exist natural definitions of "average" and "expected" such that implied volatility can indeed be understood – provably – as an average expected volatility.

2.1 Time-dependent volatility

In the case of time-dependent but nonrandom volatility, a simple formula exists for Black-Scholes implied volatility.

Suppose that

$$dS_t = rS_t dt + \sigma(t)S_t dW_t$$

where σ is a deterministic function. Define

$$\bar{\sigma} := \left(\frac{1}{T}\int_0^T \sigma^2(u)du\right)^{1/2}.$$

Then one can show that $\log S_T$ is normal with mean $(r-\bar{\sigma}^2/2)T$ and variance $\bar{\sigma}^2 T$, from which it follows that

$$C = C^{BS}(\bar{\sigma}).$$

and hence

$$I = \bar{\sigma}.$$

Thus implied volatility is equal to the quadratic mean volatility from 0 to T.

2.2 Time-and-spot-dependent Volatility

Now assume that

$$dS_t = rS_t dt + \sigma(S_t, t) S_t dW_t \tag{2.1}$$

where σ is a deterministic function, usually called the *local volatility*. We will also treat local volatility as a function $\tilde{\sigma}$ of time-0 moneyness x, via the definition $\tilde{\sigma}(x, T) := \sigma(S_0 e^{x+rT}, T)$; but abusing notation, we will suppress the tildes.

2.2.1 Local volatility and implied local volatility

Under local volatility dynamics, call prices satisfy (1.1), but with variable coefficients:

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2(S,t)S^2\frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} - rC = 0, \tag{2.2}$$

and also with terminal condition $C(S,T) = (S-K)^+$.

Dupire [20] showed that instead of fixing (K,T) and obtaining the backward PDE for $C(S,t)$, one can fix (S,t) and obtain a forward PDE for $C(K,T)$. A derivation (also in [9]) proceeds as follows.

Differentiating (2.2) twice with respect to strike shows that $G := \partial^2 C/\partial K^2$ satisfies the same PDE, but with terminal data $\delta(S-K)$. Thus G is the Green's function of (2.2), and it is the transition density of S. By a standard result (in [23], for example), it follows that G as a function of the variables (K,T) satisfies the adjoint equation, which is the Fokker-Planck PDE

$$\frac{\partial G}{\partial T} - \frac{\partial^2}{\partial K^2}\left(\frac{1}{2}\sigma^2(K,T)K^2 G\right) + r\frac{\partial}{\partial K}(KG) + rG = 0.$$

Integrating twice with respect to K and applying the appropriate boundary conditions, one obtains the Dupire equation:

$$\frac{\partial C}{\partial T} - \frac{1}{2}K^2\sigma^2(K,T)\frac{\partial^2 C}{\partial K^2} + rK\frac{\partial C}{\partial K} = 0, \tag{2.3}$$

with initial condition $C(K,0) = (S_0 - K)^+$.

Given call prices at all strikes and maturities up to some horizon, define *implied local volatility* as

$$L(K,T) := \left(\frac{\frac{\partial C}{\partial T} + rK\frac{\partial C}{\partial K}}{\frac{1}{2}K^2\frac{\partial^2 C}{\partial K^2}}\right)^{1/2}. \tag{2.4}$$

According to (2.3), this is the local volatility function consistent with the given prices of options. Define *implied local variance* as L^2.

Following standard terminology, our use of the term *implied volatility* will, in the absence of other modifiers, refer to implied *Black-Scholes* volatility, not implied *local* volatility. The two concepts are related as follows: Substituting

$$C = C^{BS}(I(S_0 e^{x+rT}, T)) \tag{2.5}$$

into (2.4) yields

$$L^2(x,T) = \frac{2TI\frac{\partial I}{\partial T} + I^2}{(1 - x\frac{\partial I}{\partial x}/I)^2 + TI\frac{\partial^2 I}{\partial x^2} - \frac{1}{4}T^2 I^2 \frac{\partial^2 I}{\partial x^2}}$$

See, for example, Andersen and Brotherton-Ratcliffe [1]. Whereas the computation of I from market data poses no numerical difficulties, the recovery of L is an ill-posed problem that requires careful treatment; see also [2, 9, 14, 26]. These issues will not concern us here, because our use of implied local volatility L will be strictly as a theoretical device to link local volatility results to stochastic volatility results, in section 2.3.1.

2.2.2 Short-dated implied volatility as harmonic mean local volatility

In certain regimes, the representation of implied volatility as an average expected volatility can be made precise. Specifically, Berestycki, Busca, and Florent ([7]; BBF henceforth) show that in the short-maturity limit, implied volatility is the *harmonic mean* of *local volatility*.

The PDE that relates implied volatility $I(x,T)$ to local volatility $\sigma(x,T)$ is, by substituting (2.5) into (2.3),

$$2TI\frac{\partial I}{\partial T} + I^2 - \sigma^2(x,T)\left(1 - x\frac{\partial I}{\partial x}\Big/ I\right)^2 \\ - \sigma^2(x,T)TI\frac{\partial^2 I}{\partial x^2} + \frac{1}{4}\sigma^2(x,T)T^2 I^2 \frac{\partial^2 I}{\partial x^2} = 0$$

Let $I_0(x)$ be the solution to the ODE generated by taking $T = 0$ in the PDE. Thus

$$I_0^2 - \sigma^2(x,0)\left(1 - x\frac{\partial I_0}{\partial x}\Big/ I_0\right)^2 = 0.$$

Elementary calculations show that the ODE is solved by

$$I_0(x) = \left(\int_0^1 \frac{ds}{\sigma(sx,0)}\right)^{-1},$$

A natural conjecture is that the convergence $I_0 = \lim_{T\to 0} I(x,T)$ holds. Indeed this is what Berestycki, Busca, and Florent [7] prove. Therefore, short-dated implied volatility is approximately the harmonic mean of local volatility, where the mean is taken "spatially," along the line segment on $T = 0$, from moneyness 0 to moneyness x.

The *harmonic* mean here stands in contrast to arithmetic or quadratic means that have been proposed in the literature as rules of thumb. As BBF argue, probabilistic considerations rule out the arithmetic and quadratic means; for example, consider a local volatility diffusion in which there exists a price level $H \in (S_0, K)$ above which the local volatility vanishes, but below which it is positive. Then the option must have zero premium, hence zero implied volatility. This is inconsistent with taking a spatial mean of σ arithmetically or quadratically, but *is* consistent with taking a spatial mean of σ harmonically.

2.2.3 Deep in/out-of-the-money implied volatility as quadratic mean local volatility

BBF also show that if local volatility is uniformly continuous and bounded by constants so that

$$0 < \underline{\sigma} \leqslant \sigma(x, T) \leqslant \overline{\sigma},$$

and if local volatility has continuous limit(s)

$$\sigma_{\pm}(t) = \lim_{x \to \pm\infty} \sigma(x, t)$$

locally uniformly in t, then deep in/out-of-the-money implied volatility approximates the *quadratic* mean of local volatility, in the following sense:

$$\lim_{x \to \pm\infty} I(x, T) = \left(\frac{1}{T} \int_0^T \sigma_{\pm}^2(s) ds \right)^{1/2}.$$

The idea of the proof is as follows. Considering by symmetry only the $x \to \infty$ limit, let $I_\infty(T) := (\frac{1}{T} \int_0^T \sigma_+^2(s) ds)^{1/2}$. Note that I_∞ induces, via definition (2.4), a local variance L^2 that has the correct behavior at $x = \infty$, because the denominator is 1 while the numerator is $\sigma_+^2(T)$.

To turn this into a proof, BBF show that for any ε one can construct a function $\overline{\psi}(x)$ such that $1 < \overline{\psi}(\infty) < 1 + \varepsilon$ and such that $I_\infty(T)\overline{\psi}(x)$ induces via (2.4) a local volatility that dominates L. By a comparison result of BBF,

$$\limsup_{x \to \infty} I(x, T) < I_\infty(T)\overline{\psi}(\infty) < (1 + \varepsilon) I_\infty(T).$$

On the other hand, one can construct $\underline{\psi}$ such that

$$\liminf_{x \to \infty} I(x, T) > I_\infty(T)\underline{\psi}(\infty) > (1 - \varepsilon) I_\infty(T).$$

Taking ε to 0 yields the result.

2.3 Stochastic volatility

Now suppose that

$$dS_t = r S_t dt + \sigma_t S_t dW_t,$$

where σ_t is stochastic. In contrast to local volatility models, σ_t is not determined by S_t and t.

Intuition from the case of time-dependent volatility does not apply directly to stochastic volatility. For example, one can define the random variable

$$\bar{\sigma} := \left(\frac{1}{T} \int_0^T \sigma_t^2 dt \right)^{1/2},$$

but note that in general

$$I \neq \mathbb{E}\bar{\sigma}.$$

For example, in the case where the σ process is independent of W, the mixing argument of Hull and White [31] shows that

$$\begin{aligned} C_0 &= \mathbb{E}e^{-rT}(S_T - K)^+ \\ &= \mathbb{E}(\mathbb{E}[e^{-rT}(S_T - K)^+ | \{\sigma_t\}_{0 \leqslant t \leqslant T}]) = \mathbb{E}C^{BS}(\bar{\sigma}). \end{aligned} \tag{2.6}$$

However, this is *not* equal to $C^{BS}(\mathbb{E}\bar{\sigma})$ because C^{BS} is not a linear function of its volatility argument. What we can say is that for the at-the-money-forward strike, C^{BS} is nearly linear in σ, because its second σ derivative is negative but typically small; so by Jensen $I < \mathbb{E}\bar{\sigma}$, but equality nearly holds.

Note that this $I \approx \mathbb{E}\bar{\sigma}$ heuristic is specific to one particular strike, that it assumes independence of σ_t and W_t, and that the expectation is under a risk-neutral pricing measure, not the statistical measure. We caution against the improper application of this rule outside of its limited context.

So is there some time-averaged volatility interpretation of I, that *does* hold in contexts where $I \approx \mathbb{E}\bar{\sigma}$ fails?

2.3.1 Relation to local-volatility results

Under stochastic volatility dynamics, implied local variance at (K, T) is the risk-neutral conditional expectation of σ_t^2, given $S_T = K$. The argument of Derman and Kani [17] is as follows. Let $f(S) = (S - K)^+$. Now take, formally, an Ito differential with respect to T:

$$\begin{aligned} d_T C &= d_T[e^{-rT}\mathbb{E}(S_T - K)^+] = \mathbb{E}d_T[e^{-rT}(S_T - K)^+] \\ &= e^{-rT}\mathbb{E}\left[f'(S_T)dS_T + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K)dT - (S_T - K)^+ dT \right] \\ &= e^{-rT}\mathbb{E}\left[rS_T H(S_T - K) + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K) - (S_T - K)^+ \right] dT \\ &= e^{-rT}\mathbb{E}\left[- rKH(S_T - K) + \frac{1}{2}\sigma_T^2 S_T^2 \delta(S_T - K) \right] dT, \end{aligned}$$

where H denotes the Heaviside function. Assuming that (S_T, σ_T^2) has a joint density p_{S_T,V_T}, let p_{S_T} denote the marginal density of S_T. Continuing, we have

$$\begin{aligned}\frac{\partial C}{\partial T} &= -rK\frac{\partial C}{\partial K} + \frac{1}{2}e^{-rT}\iint vs^2\delta(s-K)p_{S_T,V_T}(s,v)dsdv \\ &= -rK\frac{\partial C}{\partial K} + \frac{1}{2}e^{-rT}K^2\int vp_{S_T,V_T}(K,v)dv.\end{aligned}$$

So, by definition of implied local variance,

$$L^2(K,T) = \frac{\frac{\partial C}{\partial T} + rK\frac{\partial C}{\partial K}}{\frac{1}{2}K^2\frac{\partial^2 C}{\partial K^2}} = \frac{\int vp_{S_T,V_T}(K,v)dv}{p_{S_T}(K)} = \mathbb{E}(\sigma_t^2|S_T = K).$$

Consequently, any characterization of I as an average expected local volatility becomes tantamount to a characterization of I as an average conditional expectation of stochastic volatility.

Application 2.1. The BBF results in sections 2.2.2 and 2.2.3 can be interpreted, under *stochastic* volatility, as expressions of implied volatility as [harmonic or quadratic] average *conditional* expectations of future volatility.

2.3.2 The path-from-spot-to-strike approach

The following reasoning by Gatheral [25] provides an interpretation of implied volatility as average expected stochastic volatility, without assuming short times to maturity or strikes deep in/out of the money.

Fix K and T. Let

$$\Gamma^{BS} := \frac{\partial^2 C^{BS}}{\partial S^2}$$

be the Black-Scholes gamma function.

Assume there exists a nonrandom nonnegative function $v(t)$ such that for all t in $(0,T)$,

$$v(t) = \frac{\mathbb{E}[\sigma_t^2 S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))]}{\mathbb{E}[S_t^2 \Gamma^{BS}(S_t, t, \bar{\sigma}(t))]} \tag{2.7}$$

where

$$\bar{\sigma}(t) := \left(\frac{1}{T-t}\int_t^T v(u)du\right)^{1/2}.$$

Note that σ_t need not be a deterministic function of spot and time.

Define the function

$$c(S,t) := C^{BS}(S, t, \bar{\sigma}(t)),$$

which solves the following PDE for $(S,t) \in (0,\infty) \times (0,T)$:

$$\frac{\partial c}{\partial t} = -\frac{1}{2}v(t)S^2\frac{\partial^2 c}{\partial S^2} - rs\frac{\partial C}{\partial S} + rC. \tag{2.8}$$

We have

$$
\begin{aligned}
C(K,T) &= \mathbb{E}[e^{-rT}(S_T-K)^+] = \mathbb{E}[e^{-rT}c(S_T,T)] \\
&= c(S_0,0) + e^{-rT}\mathbb{E}\bigg[\int_0^T \frac{\partial c}{\partial t}(S_t,t)dt + \frac{1}{2}\sigma_t^2 S_t^2 \frac{\partial^2 c}{\partial S^2}(S_t,t)dt \\
&\qquad\qquad + \frac{\partial c}{\partial S}(S_t,t)dS_t - rc(S_t,t)dt\bigg] \\
&= c(S_0,0) + e^{-rT}\mathbb{E}\bigg[\int_0^T \frac{1}{2}(\sigma_t^2 - v(t))S_t^2\frac{\partial^2 c}{\partial S^2}(S_t,t)dt\bigg] \\
&= c(S_0,0) = C^{BS}(S_0,0,\bar\sigma(0)).
\end{aligned}
$$

using Ito's rule, then (2.8), then (2.7). Therefore

$$
I^2 = \bar\sigma^2(0) = \frac{1}{T}\int_0^T v(t)dt = \frac{1}{T}\int_0^T \mathbb{E}^{\mathbb{G}_t}\sigma_t^2 dt, \tag{2.9}
$$

where the final step re-interprets the definition (2.7) of $v(t)$ as the expectation of σ_t^2 with respect to the probability measure $\mathbb{G}_t$ defined, relative to the pricing measure $\mathbb{P}$, by the Radon-Nikodym derivative

$$
\frac{d\mathbb{G}_t}{d\mathbb{P}} := \frac{S_t^2\Gamma^{BS}(S_t,t,\bar\sigma(t))}{\mathbb{E}[S_t^2\Gamma^{BS}(S_t,t,\bar\sigma(t))]}.
$$

So (2.9) interprets implied volatility as an average expected variance. Moreover, this expectation with respect to $\mathbb{G}_t$ can be visualized as follows. Write

$$
\mathbb{E}^{\mathbb{G}_t}\sigma_t^2 = \int_0^\infty \mathbb{E}(\sigma_t^2|S_t=s)\kappa_t(s)ds, \tag{2.10}
$$

where the nonrandom function κ_t is defined by

$$
\kappa_t(s) := \frac{s^2\Gamma^{BS}(s,t,\bar\sigma(t))p_{S_t}(s)}{\int_0^\infty s^2\Gamma^{BS}(s,t,\bar\sigma(t))p_{S_t}(s)ds},
$$

and p_{S_t} denotes the density of S_t.

Thus $\mathbb{E}(\sigma_t^2|S_t=s)$ is integrated against a kernel $\kappa(s)$ which has the following behavior. For $t \downarrow 0$, the κ approaches the Dirac function $\delta(s-S_0)$, because the p_{S_t} factor has that behavior, while the $s^2\Gamma^{BS}$ factor approaches an ordinary function. For $t \uparrow T$, the κ approaches the Dirac function $\delta(s-K)$, because the $s^2\Gamma^{BS}$ factor has that behavior, while the p_{S_t} factor approaches an ordinary function. At each time t intermediate between 0 and T, the kernel has a finite peak, which moves from S_0 to K, as t moves from 0 to T.

This leads to two observations. First, one has the conjectural approximation

$$
\mathbb{E}^{\mathbb{G}_t}\sigma_t^2 \approx \mathbb{E}(\sigma_t^2|S_t=s^*(t)),
$$

where the non-random point $s^*(t)$ is the s that maximizes the kernel κ_t. By (2.10), therefore,

$$I^2 \approx \frac{1}{T}\int_0^T \mathbb{E}(\sigma_t^2|S_t = s^*(t))dt.$$

Second, the kernel's concentration of "mass" initially (for $t = 0$) at S_0, and terminally (for $t = T$) at K resembles the marginal densities of the S diffusion, *pinned* by conditioning on $S_T = K$. This leads to Gatheral's observation that implied variance is, to a first approximation, the time integral of the expected instantaneous variance *along the most likely path from* S_0 *to* K. We leave open the questions of how to make these observations more precise, and how to justify the original assumption.

Application 2.2. Given an approximation for local volatility, such as in [25], one can usually compute explicitly an approximation for a spot-to-strike average, thus yielding an approximation to implied volatility.

For example, given an approximation for local volatility linear in x, the spot-to-strike averaging argument can be used to justify a rule of thumb (as in [18]) that approximates implied volatility also linearly in x, but with one-half the slope of local volatility.

3 Statics

We examine here the implications of various assumptions on the shape of the implied volatility surface, beginning in section 3.1 with only minimal assumptions of no-arbitrage, and then specializing in 3.2 and 3.3 to the cases of local volatility and stochastic volatility diffusions. The term "statics" refers to the analysis of $I(x, T)$ or $I(K, T)$ for t *fixed.*

As reference points, let us review some of the empirical facts about the shape of the volatility surface; see, for example, [39] for further discussion. A plot of I is not constant with respect to K (or x). It can take the shape of a *smile*, in which $I(K)$ is greater for K away-from-the-money than it is for K near-the-money. The more typical pattern in post-1987 equity markets, however, is a *skew* (or skewed smile) in which at-the-money I slopes downward, and the smile is far more pronounced for small K than for large K. Empirically the smile or skew *flattens* as T increases. In particular, a popular rule-of-thumb (which we will revisit) states that skew slopes decay with maturity approximately as $1/\sqrt{T}$; indeed, when comparing skew slopes across different maturities, practitioners often define "moneyness" as $x/\sqrt{T}$ instead of x.

The theory of how I behaves under various model specifications has at least three applications. First, to the extent that a model generates a theoretical I shape that differs qualitatively from empirical facts, we have evidence of model misspecification. Second, given an observed volatility skew, analytical expressions approximating $I(x, T)$ in terms of model parameters can be useful in calibrating those parameters. Third, necessary conditions on I for the absence of arbitrage provide consistency checks that can help to reject unsound proposals for volatility skew parameterizations.

Part of the challenge for future research will be to extend this list of models and regimes for which we understand the behavior of implied volatility.

3.1 Statics under absence of arbitrage

Assuming only the absence of arbitrage, one obtains bounds on the slope of the implied volatility surface, as well as a characterization of how fast I grows at extreme strikes.

3.1.1 Slope bounds

Hodges [30] gives bounds on implied volatility based on the nonnegativity of call spreads and put spreads. Specifically, if $K_1 < K_2$ then

$$C(K_1) \geqslant C(K_2) \qquad P(K_1) \leqslant P(K_2) \tag{3.1}$$

Gatheral [24] improves this observation to

$$C(K_1) \geqslant C(K_2) \qquad \frac{P(K_1)}{K_1} \leqslant \frac{P(K_2)}{K_2}, \tag{3.2}$$

which is evident from a comparison of the respective payoff functions. Assuming the differentiability of option prices in K,

$$\frac{\partial C}{\partial K} \leqslant 0 \qquad \frac{\partial}{\partial K}\left(\frac{P}{K}\right) \geqslant 0.$$

Substituting $C = C^{BS}(I)$ and $P = P^{BS}(I)$ and simplifying, we have

$$-\frac{N(-d_1)}{\sqrt{T}N'(d_1)} \leqslant \frac{\partial I}{\partial x} \leqslant \frac{N(d_2)}{\sqrt{T}N'(d_2)},$$

where the upper and lower bounds come from the call and put constraints, respectively.

Using (as in Carr-Wu [12]), the *Mill's Ratio* $R(d) := (1 - N(d))/N'(d)$ to simplify notation, we rewrite the inequality as

$$-\frac{R(d_1)}{\sqrt{T}} \leqslant \frac{\partial I}{\partial x} \leqslant \frac{R(-d_2)}{\sqrt{T}}$$

Note that proceeding from (3.1) without Gatheral's refinement (3.2) yields the significantly weaker lower bound $-R(d_2)/\sqrt{T}$.

Of particular interest is the behavior at-the-money, where $x = 0$. In the short-dated limit, as $T \to 0$, assume that $I(0,T)$ is bounded above. Then

$$d_{1,2}(x = 0) = \pm I(0,T)\sqrt{T}/2 \longrightarrow 0.$$

Since $R(0)$ is a positive constant, the at-the-money skew slope must have the short-dated behavior

$$\left|\frac{\partial I}{\partial x}(0,T)\right| = O\left(\frac{1}{\sqrt{T}}\right), \qquad T \to 0. \tag{3.3}$$

In the long-dated limit, as $T \to \infty$, assume that $I(0,T)$ is bounded away from 0. Then

$$d_{1,2}(x=0) = \pm I(0,T)\sqrt{T}/2 \longrightarrow \pm\infty.$$

Since $R(d) \sim d^{-1}$ as $d \to \infty$, the at-the-money skew slope must have the long-dated behavior

$$\left|\frac{\partial I}{\partial x}(0,T)\right| = O\left(\frac{1}{T}\right), \qquad T \to \infty. \tag{3.4}$$

Remark 3.1. According to (3.4), the rule of thumb that approximates the skew slope decay rate as $T^{-1/2}$ cannot maintain validity into long-dated expiries.

3.1.2 The moment formula

Lee [37] proves the moment formula for implied volatility at extreme strikes. Previous work, in Avellaneda and Zhu [3], had produced asymptotic calculations for one specific stochastic volatility model, but the moment formula is entirely general, and it uncovers the key role of finite moments.

At any given expiry T, the tails of the implied volatility skew can grow no faster than $x^{1/2}$. Specifically, in the right-hand tail, for $|x|$ sufficiently large, the Black-Scholes implied variance satisfies

$$I^2(x,T) \leqslant 2|x|/T \tag{3.5}$$

and a similar relationship holds in the left-hand tail.

For proof, write $I^* := (2|x|/T)^{1/2}$, and show that $C^{BS}(I) < C^{BS}(I^*)$ for large $|x|$. This holds because the left-hand side approaches 0 but the right-hand side approaches a positive limit as $x \to \infty$.

Application 3.2. This bound has implications for choosing functional forms of splines to extrapolate volatility skews. Specifically, it advises against fitting the skew's tails with any function that grows more quickly than $x^{1/2}$.

Moreover, the tails cannot grow more *slowly* than $x^{1/2}$, unless S_T has finite moments of all orders. This further restricts the advisable choices for parameterizing a volatility skew. To prove this fact, note that it is a consequence of the moment formula, which we now describe.

The smallest (infimal) coefficient that can replace the *2* in (3.5) depends, of course, on the distribution of S_T, but the form of the dependence is notably simple. This sharpest possible coefficient is entirely determined by $\tilde{p}$ in the right-hand tail, and $\tilde{q}$ in the left-hand tail, where the real numbers

$$\tilde{p} := \sup\{p : \mathbb{E}S_T^{1+p} < \infty\}$$
$$\tilde{q} := \sup\{q : \mathbb{E}S_T^{-q} < \infty\},$$

can be considered, by abuse of language, the "number" of finite moments in underlying distribution. The moment formula makes explicit these relationships.

Specifically, let us write I^2 as a variable coefficient times $|x|/T$, the ratio of absolute-log-moneyness to maturity. Consider the limsups of this coefficient as $x \to \pm\infty$:

$$\beta_R(T) := \limsup_{x\to\infty} \frac{I^2(x,T)}{|x|/T}$$
$$\beta_L(T) := \limsup_{x\to-\infty} \frac{I^2(x,T)}{|x|/T}.$$

One can think of β_R and β_L as the right-hand and absolute left-hand slopes of the linear "asymptotes" to implied variance.

The main theorem in [37] establishes that β_R and β_L both belong to the interval $[0,2]$, and that their values depend only on the moment counts $\tilde{p}$ and $\tilde{q}$, according to the *moment formula*:

$$\tilde{p} = \frac{1}{2\beta_R} + \frac{\beta_R}{8} - \frac{1}{2}$$
$$\tilde{q} = \frac{1}{2\beta_L} + \frac{\beta_L}{8} - \frac{1}{2}.$$

One can invert the moment formula, by solving for β_R and β_L:

$$\beta_R = 2 - 4(\sqrt{\tilde{p}^2 + \tilde{p}} - \tilde{p}),$$
$$\beta_L = 2 - 4(\sqrt{\tilde{q}^2 + \tilde{q}} - \tilde{q}).$$

The idea of the proof is as follows. By the Black-Scholes formula, the tail behavior of the implied volatility skew carries the same information as the tail behavior of option prices. In turn, the tail growth of option prices carries the same information as the number of finite moments – intuitively, option prices are bounded by moments, because a call or put payoff can be dominated by a power payoff; on the other hand, moments are bounded by option prices, because a power payoff can be dominated by a mixture, across a continuum of strikes, of call or put payoffs.

In a wide class of specifications for the dynamics of S, the moment counts $\tilde{p}$ and $\tilde{q}$ are readily computable functions of the model's parameters. This occurs whenever $\log S_T$ has a distribution whose characteristic function f is explicitly known. In such cases, one calculates $\mathbb{E}S_T^{p+1}$ simply by extending f analytically to a strip in $\mathbb{C}$ containing $-i(p+1)$, and evaluating f there; if no such extension exists, then $\mathbb{E}S_T^{p+1} = \infty$. In particular, among affine jump-diffusions and Levy processes, one finds many instances of such models. See, for example, [19, 36].

Application 3.3. The moment formula may speed up the calibration of model parameters to observed skews. By observing the tail slopes of the volatility skew, and applying the moment formula, one obtains $\tilde{p}$ and $\tilde{q}$. Combined with analysis of the characteristic function, this produces two constraints on the model parameters, and in models such as the examples below, actually *determines* two of the model's parameters. We do not claim that the moment formula alone can replace a full optimization procedure, but it could facilitate the process by providing a highly accurate initial guess of the optimal parameters.

Example 3.4. In the double-exponential jump-diffusion model of Kou [32, 33], the asset price follows a geometric Brownian motion between jumps, which occur at event times of a Poisson process. Up-jumps and down-jumps are exponentially distributed with the parameters η_1 and η_2 respectively, and hence the means $1/\eta_1$ and $1/\eta_2$ respectively. Using the characteristic function, one computes

$$\tilde{q} = \eta_2 \qquad \tilde{p} = \eta_1 - 1. \tag{3.6}$$

Thus η_1 and η_2 can be inferred from $\tilde{p}$ and $\tilde{q}$, which in turn come from the slopes of the volatility skew, via the moment formula.

The intuition of (3.6) is as follows: the larger the expected size of an up-jump, the fatter the S_T distribution's right-hand tail, and the fewer the number of positive moments. Similar intuition holds for down-jumps. Note that the jump *frequency* has no effect on the asymptotic slopes.

Example 3.5. In the normal inverse Gaussian model of Barndorff-Nielsen [6], returns have the NIG distribution defined as follows: Consider two-dimensional Brownian motion starting at $(a, 0)$, with constant drift (b, c), where $c > 0$. The $\mathrm{NIG}(a, b, c, d)$ law is the distribution of the first coordinate of the Brownian motion at the stopping time when the second coordinate hits a barrier $d > 0$. Then one calculates

$$\tilde{q} = \sqrt{b^2 + c^2} + b \qquad \tilde{p} = \sqrt{b^2 + c^2} - b - 1, \tag{3.7}$$

so b and c can be inferred from $\tilde{p}$ and $\tilde{q}$, which in turn come from the slopes of the volatility skew, via the moment formula.

This also has intuitive content: larger c implies earlier stopping, hence thinner tails and more moments (of both positive and negative order); larger b fattens the right-hand tail and thins the left-hand tail, decreasing the number of positive moments and increasing the number of negative moments. Note that the parameters a and d have no effect on the asymptotic slopes.

3.2 Statics under local volatility

Assume that the underlying follows a local volatility diffusion of the form (2.1). Writing $F := Se^{r(T-t)}$ for the forward price, suppose that local volatility can be expressed as a function h of F alone:

$$\sigma(S, t) = h(Se^{r(T-t)}).$$

Hagan and Woodward (in [28], and with Kumar and Lesniewski in [27]), develop regular perturbation solutions to (2.2) in powers of $\varepsilon := h(K)$, assumed to be small. The resulting call price formula then yields the implied volatility approximation

$$I(K,T) \approx h(\bar{F}) + \frac{1}{24}h''(\bar{F})(F_0 - K)^2, \tag{3.8}$$

where $\bar{F} := (F_0 + K)/2$ is the midpoint between forward and strike. The same sources also discuss alternative assumptions and more refined approximations.

Remark 3.6. The reasoning of section 2.3.2 suggests an interpretation of the leading term $h(\bar{F})$ in (3.8) as a midpoint approximation to the average local volatility along a path from $(F_0, 0)$ to (K, T).

3.3 Statics under stochastic volatility

Now assume that the underlying follows a stochastic volatility diffusion of the form

$$\begin{aligned} dS_t &= rS_t dt + \sqrt{V_t} S_t dW_t \\ dV_t &= a(V_t)dt + b(V_t)dZ_t \end{aligned}$$

where Brownian motions W and Z have correlation ρ. From here one obtains, typically via perturbation methods, approximations to the implied volatility skew I. Our coverage will emphasize those approximations which apply to entire *classes* of stochastic volatility models, not specific to one particular choice of a and b. We label each approximation according to the regime in which it prevails.

3.3.1 Zero correlation

Renault and Touzi [40] prove that in the case $\rho = 0$, implied volatility is a symmetric smile – symmetric in the sense that

$$I(x,T) = I(-x,T)$$

and a smile in the sense that I is increasing in x for $x > 0$.

Moreover, as shown in [4], the parabolic shape of I is apparent from Taylor approximations. Expanding the function $C^{bs}(v) := C^{BS}(\sqrt{v})$ about $v = \mathbb{E}\overline{V}$, we have

$$C = C^{bs}(I) \approx C^{bs}(\mathbb{E}\overline{V}) + (I^2 - \mathbb{E}\overline{V})\frac{\partial C^{bs}}{\partial V}.$$

Comparing this to a Taylor expansion of the mixing formula

$$C = \mathbb{E}C^{bs}(\overline{V}) \approx C^{bs}(\mathbb{E}\overline{V}) + \frac{1}{2}\mathrm{Var}(\overline{V})\frac{\partial^2 C^{bs}}{\partial V^2}$$

yields the approximation

$$I^2 \approx \mathbb{E}\overline{V} + \frac{1}{4}\frac{\mathrm{Var}(\overline{V})}{(\mathbb{E}\overline{V})^2}\left(\frac{x^2}{T} - \mathbb{E}\overline{V} - \frac{1}{4}(\mathbb{E}\overline{V})^2 T\right),$$

which is quadratic in x, with minimum at $x = 0$.

Remark 3.7. To the extent that implied volatility skews are empirically *not* symmetric in equity markets, stochastic volatility models with zero correlation will not be consistent with market data.

3.3.2 Small volatility of volatility, and the short-dated limit

Lewis [38] shows that the forward call price, viewed as a function of x, has a complex Fourier transform given by $\hat{H}(k, V, T)/(k^2 - ik)$, where k is the transform variable and $\hat{H}$ solves the PDE

$$\frac{\partial \hat{H}}{\partial T} = \frac{1}{2} b^2 \frac{\partial^2 \hat{H}}{\partial V^2} + (a - ik\rho b V^{1/2}) \frac{\partial \hat{H}}{\partial V} - \frac{k^2 - ik}{2} V \hat{H},$$

with initial condition $\hat{H}(k, V, 0) = 1$. In our setting, $\hat{H}$ can be viewed as the characteristic function of the negative of the log-return on the forward price of S.

Assuming that $b(V) = \eta B(V)$ for some constant parameter η, one finds a perturbation solution for $\hat{H}$ in powers of η. The transform can be inverted to produce a call price, by a formula such as

$$C = S - \frac{Ke^{-rT}}{2\pi} \int_{i/2-\infty}^{i/2+\infty} e^{ikx} \frac{\hat{H}(k, V, T)}{k^2 - ik} dk, \tag{3.9}$$

yielding a series for C in powers of η. From the C series and the Black-Scholes formula, Lewis derives the implied variance expansion

$$\begin{aligned} I^2 = \mathbb{E}\overline{V} &+ \eta \frac{J^{(1)}}{T} \left(\frac{x}{T\mathbb{E}\overline{V}} + \frac{1}{2} \right) \\ &+ \eta^2 \left[\frac{J^{(2)}}{T} + \frac{J^{(3)}}{T} \left(\frac{x^2}{2(\mathbb{E}\overline{V})^2 T^2} - \frac{1}{2T\mathbb{E}\overline{V}} - \frac{1}{8} \right) \right. \\ &\quad + \frac{J^{(4)}}{T} \left(\frac{x^2}{T^2(\mathbb{E}\overline{V})^2} + \frac{x}{T\mathbb{E}\overline{V}} - \frac{4 - T\mathbb{E}\overline{V}}{4T\mathbb{E}\overline{V}} \right) \\ &\quad \left. + \frac{(J^{(1)})^2}{2T} \left(- \frac{5x^2}{2T^3(\mathbb{E}\overline{V})^3} - \frac{x}{T\mathbb{E}\overline{V}} + \frac{12 + T\mathbb{E}\overline{V}}{8T^2(\mathbb{E}\overline{V})^2} \right) \right] + O(\eta^3), \end{aligned}$$

where $J^{()}$ are integrals of known functions.

Example 3.8. The short-time-to-expiry limit is

$$I^2(x, 0) = V_0 + \frac{1}{2} \frac{\rho b}{\sqrt{V_0}} x + \left[\left(\frac{1}{12} - \frac{11}{48} \rho^2 \right) \frac{b^2}{V_0^2} + \frac{1}{6} \frac{\rho b}{V_0} \frac{\partial(\rho b)}{\partial V} \right] x^2 + O(\eta^3). \tag{3.10}$$

The leading terms agree to $O(\eta)$ with the slow-mean-reversion result of section 3.3.5. We defer further commentary until there.

Example 3.9. In the case where

$$dV_t = \kappa(\theta - V_t)dt + \eta V_t^{\varphi} dW_t, \tag{3.11}$$

we have

$$\mathbb{E}\overline{V} = \theta + \frac{1 - e^{-\kappa T}}{\kappa T}(V_0 - \theta)$$

and $J^{(2)} = 0$, while

$$J^{(1)} = \frac{\rho}{\kappa}\int_0^T (1 - e^{-\kappa(T-s)})\Big[\theta + e^{-\kappa s}(V_0 - \theta)\Big]^{\varphi+1/2} ds$$

$$J^{(3)} = \frac{1}{2\kappa^2}\int_0^T \big(1 - e^{-\kappa(T-s)}\big)^2\Big[\theta + e^{-\kappa s}(V_0 - \theta)\Big]^{2\varphi} ds$$

$$J^{(4)} = \Big(\varphi + \frac{1}{2}\Big)\frac{\rho^2}{\kappa}\int_0^T \Big[\theta + e^{-\kappa(T-s)}(V_0 - \theta)\Big]^{\varphi+1/2} J^{(6)}(T,s) ds$$

$$J^{(6)} = \int_0^s (e^{-\kappa(s-u)} - e^{-\kappa s})\Big[\theta + e^{-\kappa(T-u)}(V_0 - \theta)\Big]^{\varphi-1/2} du.$$

In particular, taking $\varphi = 1/2$ produces the Heston [29] square-root model. In the special case where $V_0 = \theta$, the slope of the implied variance skew is, to leading order in η,

$$\frac{\partial I^2}{\partial x} = \frac{\rho\eta}{\kappa T}\Big(1 - \frac{1 - e^{-\kappa T}}{\kappa T}\Big),$$

which agrees with a computation, by Gatheral [25], that uses the expectations interpretation of local volatility.

3.3.3 The long-dated limit

Given a stochastic volatility model with a known transform $\hat{H}$, Lewis solves for $\lambda(k)$ and $u(k,T)$ such that $\hat{H}$ separates multiplicatively, for large T, into T-dependent and V-dependent factors:

$$\hat{H}(k,V,T) \approx e^{-\lambda(k)T}u(k,V), \qquad T \to \infty.$$

Suppose that $\lambda(k)$ has a saddle point at $k_0 \in \mathbb{C}$ where $\lambda'(k_0) = 0$. Applying classical saddle-point methods to (3.9) yields

$$C(S,V,T) \approx S - Ke^{-rT}\frac{u(k_0,V)}{k_0^2 - ik_0}\frac{\exp[-\lambda(k_0)T + ik_0x]}{\sqrt{2\pi\lambda''(k_0)T}}.$$

By comparing this to the corresponding approximation of $C^{BS}(I)$, Lewis obtains the implied variance approximation

$$I^2(x) \approx 8\lambda(k_0) + (8\mathrm{Im}(k_0) - 4)\frac{x}{T} - \frac{x^2}{2\lambda(k_0)T^2} + O(T^{-3}), \qquad T \to \infty.$$

The fact that $I(x,T)$ is linear to first order in x/T agrees with the fast-mean-reversion result of Fouque, Papanicolaou, and Sircar [21]. We defer further commentary until section 3.3.4.

Example 3.10. In the case (3.11) with $\varphi = 1/2$ (the square-root model), Lewis finds

$$k_0 = \frac{i}{1-\rho^2}\Big[\frac{1}{2} - \frac{\rho}{\eta}\Big(\kappa - \frac{1}{2}\sqrt{4\kappa^2 + \eta^2 - 4\rho\kappa\eta}\Big)\Big]$$

$$\lambda(k_0) = \frac{\kappa\theta}{2(1-\rho^2)\eta^2}\Big[\sqrt{(2\kappa - \rho\eta)^2 + (1-\rho^2)\eta^2} - (2\kappa - \rho\eta)\Big].$$

The sign of the leading-order at-the-money skew slope $(8\mathrm{Im}(k_0) - 4)/T$ agrees with the sign of the correlation ρ.

3.3.4 Fast mean reversion

Fouque-Papanicolaou-Sircar ([21]; FPS henceforth) model stochastic volatility as a function f of a state variable Y_t that follows a rapidly mean-reverting diffusion process. In the case of Ornstein-Uhlenbeck Y, this means that for some large α,

$$\begin{aligned} dS_t &= \mu_t S_t dt + f(Y_t) S_t d\tilde{W}_t \\ dY_t &= \alpha(\theta - Y_t)dt + \beta d\tilde{Z}_t \end{aligned}$$

under the statistical measure, where the Brownian motions $\tilde{W}$ and $\tilde{Z}$ have correlation ρ.

Rewriting this under a pricing measure,

$$\begin{aligned} dS_t &= rS_t dt + f(Y_t) S_t dW_t \\ dY_t &= [\alpha(\theta - Y_t) - \beta\Lambda(Y_t)]dt + \beta dZ_t, \end{aligned}$$

where the volatility risk premium Λ is assumed to depend only on Y. Let p_Y denote the invariant density (under the statistical probability measure) of Y, which is normal with mean θ and variance $\beta^2/(2\alpha)$. Let angle brackets denote average with respect to that density. Write

$$\bar{\sigma}_\infty^2 := \langle f^2 \rangle,$$

so that $\bar{\sigma}_\infty$ is the quadratic average of volatility with respect to the invariant distribution.

By a singular perturbation analysis of the PDE for call price, FPS show that implied volatility has an expansion with leading terms

$$I(x,T) = A\frac{x}{T} + B + O(1/\alpha),$$

where

$$\begin{aligned} A &:= -\frac{V_3}{\bar{\sigma}_\infty^3} \\ B &:= \bar{\sigma}_\infty + \frac{3V_3/2 - V_2}{\bar{\sigma}_\infty}, \end{aligned} \tag{3.12}$$

and

$$\begin{aligned} V_2 &:= \frac{\beta}{2\alpha}\langle(2\rho f - \Lambda)\phi\rangle \\ V_3 &:= \frac{\beta}{2\alpha}\langle\rho f \phi\rangle \\ \phi(y) &:= \frac{2\alpha}{\beta^2 p_Y(y)}\int_{-\infty}^{y}(f^2(z) - \langle f^2 \rangle)p_Y(z)dz. \end{aligned}$$

Remark 3.11. The fast-mean-reversion approximation is particularly suited for pricing long-dated options; in that long time horizon, volatility has time to undergo much activity, so relative to the time scale of the option's lifetime, volatility can indeed be considered to mean-revert rapidly.

Note that $I(x,T)$ is, to first order, linear in x/T. This functional form agrees with Lewis's *long-dated* skew approximation (3.3.3).

Remark 3.12. Today's volatility plays no role in the leading-order coefficients A and B. Instead, the dominant effects depend only on ergodic means. Intuitively, the assumption of large mean-reversion rapidly erodes the influence of today's volatility, leaving the long-run averages to determine A and B.

Remark 3.13. The slope of the long-dated implied volatility skew satisfies

$$\left|\frac{\partial I}{\partial x}(0,T)\right| \sim \frac{1}{T} \qquad T \to \infty.$$

As a consistency check, note that the long-dated asymptotics are consistent with the no-arbitrage constraint (3.4). Specifically, the $T \to \infty$ skew slope decay of these stochastic volatility models achieves the $O(T^{-1})$ bound.

Application 3.14. FPS give approximations to prices of certain path-dependent derivatives under fast-mean-reverting stochastic volatility. Typically, such approximations involve the Black-Scholes price for that derivative, corrected by some term that depends on V_2 and V_3.

To evaluate this correction term, note that the formulas (3.12) can be solved for V_2 and V_3 in terms of A, B, and $\bar{\sigma}$. FPS calibrate A and B to the implied volatility skew, and estimate $\bar{\sigma}$ from historical data, producing estimates of V_2 and V_3, which become the basis for an approximation of the derivative price.

For example, in the case of uncorrelated volatility where $\rho = 0$, FPS find that the price of an American put is approximated by the Black Scholes American put price, evaluated at the volatility parameter

$$\sqrt{\bar{\sigma}^2 - 2V_2},$$

which can be considered an "effective volatility."

3.3.5 Slow mean reversion

Assuming that for a constant parameter ε,

$$d\sigma_t = \varepsilon\alpha(V_t)dt + \sqrt{\varepsilon}\beta(V_t)dW_t,$$

Sircar and Papanicolaou [42] develop, and Lee [35] extends, a regular perturbation analysis of the PDE

$$\frac{\partial C}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + \sqrt{\varepsilon}\rho S\sigma\beta\frac{\partial^2 C}{\partial S \partial\sigma} + \frac{1}{2}\varepsilon\beta^2\frac{\partial^2 C}{\partial\sigma^2} + \varepsilon\alpha\frac{\partial C}{\partial\sigma} + rS\frac{\partial C}{\partial S} = rC$$

satisfied by the call price under stochastic volatility. This leads to an expansion for C in powers of ε, which in turn leads to the implied volatility expansion

$$\begin{aligned}I \approx \sigma_0 &+ \sqrt{\varepsilon}\left[\frac{\rho\beta}{2\sigma_0}x + \frac{\rho\sigma_0\beta}{4}T\right] \\ &+ \varepsilon\Bigg[\left(\left(\frac{\beta\beta'}{6\sigma^2} - \frac{5\beta^2}{12\sigma^3}\right)\rho^2 + \frac{\beta^2}{6\sigma^3}\right)x^2 + \left(\left(\frac{\sigma\beta^2}{12} + \frac{\sigma^2\beta\beta'}{24}\right)\rho^2 - \frac{\sigma\beta^2}{24}\right)T^2 \\ &\quad - \left(\frac{\beta^2}{24\sigma} - \frac{\beta\beta'}{6}\right)\rho^2 Tx + \left(\left(\frac{\beta^2}{24\sigma} - \frac{\beta\beta'}{6}\right)\rho^2 + \frac{\alpha}{2} - \frac{\beta^2}{12\sigma}\right)T\Bigg],\end{aligned}$$

where $\beta' := \partial\beta/\partial\sigma$. In particular, short-dated implied volatility satisfies

$$I(x,0) \approx \sigma_0 + \sqrt{\varepsilon}\frac{\rho\beta}{2\sigma_0}x. \tag{3.13}$$

Remark 3.15. The slow-mean-reversion approximation is particularly suited for pricing short-dated options; in that short time horizon, volatility has little time in which to vary, so relative to the time scale of the option's lifetime, volatility can indeed be considered to mean-revert slowly.

Note that (3.13) agrees precisely with the leading terms of Lewis's *short-dated* skew approximation (3.10).

Remark 3.16. In contrast to the case of rapid mean-reversion, the level to which volatility reverts here plays no role in the leading-order coefficients. With a small rate of mean-reversion, today's volatility will have the dominant effect.

Remark 3.17. For $\rho \neq 0$, the at-the-money skew exhibits a slope whose sign agrees with ρ. For $\rho = 0$ the skew has a parabolic shape.

Remark 3.18. In agreement with a result of Ledoit, Santa-Clara, and Yan [34], we have $I(x,T) \to \sigma_0$ as $(x,T) \to (0,0)$.

Application 3.19. In principle, given a parametric form for b, the fact that the short-dated skew has slope ρb gives information that can simplify parameter calibration. For example, if the modelling assumption is that $b = \beta f(V)$ for some constant parameter β and known function f, then directly from the short-dated skew and its slope, one obtains the product of the parameters ρ and β.

Application 3.20. Lewis observes, moreover, that this tool facilitates the *inference* of the functional form of b. Specifically, observe time-series of the short-dated at-the-money data pair: (implied volatility, skew slope). As implied volatility ranges over its support, the functional form of b is, in principle, revealed.

Remark 3.21. Note that the $T \to 0$ skew slope is $O(1)$, which is strictly smaller than the $O(T^{-1/2})$ constraint. To the extent that the short-dated volatility skew slope empirically seems to *attain* the $O(T^{-1/2})$ upper bound instead of the $O(1)$ diffusion behavior, this observed skew will not be easily captured by standard diffusion models. Two approaches to this problem, and subjects for further research, are to remain in the stochastic-volatility diffusion framework but introduce time-varying coefficients (as in Fouque-Papanicolaou-Sircar-Solna [22]); or alternatively to go outside the diffusion framework entirely and introduce jump dynamics, such as in Carr-Wu [12].

4 Dynamics

While traditional diffusion models specify the dynamics of the spot price and its instantaneous volatility, a newer class of models seeks to specify directly the dynamics of one or more implied volatilities. One reason to take I as primitive is that it enjoys wide acceptance as a descriptor of the state of an options market. A second reason is that the observability of I makes calibration trivial.

In this section, today's date t is not fixed at 0, because we are now concerned with the time evolution of I.

4.1 No-arbitrage approach

4.1.1 One implied volatility

Consider the time-evolution of a single implied volatility I at some fixed strike K and maturity date T. Schönbucher [41] models directly its dynamics as

$$dI_t = u_t dt + \gamma_t dW_t^{(0)} + v_t dW_t,$$

where W and $W^{(0)}$ are independent Brownian motions. The spot price has dynamics

$$dS_t = rS_t dt + \sigma_t S_t dW_t^{(0)},$$

where σ_t is yet to be specified.

Since the discounted call price $e^{-r(T-t)}C^{BS}(t, S_t, I_t)$ must be a martingale under the pricing measure, we have for all $I > 0$ the following drift restriction on the call price:

$$\frac{\partial C^{BS}}{\partial t} + rS\frac{\partial C^{BS}}{\partial S} + u\frac{\partial C^{BS}}{\partial I} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C^{BS}}{\partial S^2} + \gamma\sigma S\frac{\partial^2 C^{BS}}{\partial I \partial S} + \frac{1}{2}v^2\frac{\partial^2 C^{BS}}{\partial I^2} = rC^{BS}.$$

This reduces to a joint restriction on the diffusion coefficients of I, the drift of I, and the instantaneous volatility σ:

$$Iu = \frac{I^2 - \sigma^2}{2(T-t)} - \frac{1}{2}d_1 d_2 v^2 + \frac{d_2}{\sqrt{T-t}}\sigma\gamma. \tag{4.1}$$

Since S, t, and T are observable, we have that the volatility of I, together with the drift of I, determines the spot volatility. Other papers [10, 34] have arrived at analogous results in which one fixes not (strike, expiry), but instead some other specification of exactly which implied volatility is to be modelled, such as (moneyness, time to maturity).

Schönbucher imposes a further constraint to ensure that I does not blow up as $t \to T$. He requires that

$$(I^2 - \sigma^2) - d_1 d_2 (T-t) v^2 + 2 d_2 \sqrt{T-t}\,\sigma\gamma = O(T-t) \qquad t \to T, \tag{4.2}$$

which simplifies to

$$I^2\sigma^2 + 2\gamma x I \sigma - I^4 + x^2 v^2 = 0.$$

This can be solved to get expiration-date implied volatility in terms of expiration-date spot volatility. The solution is particularly simple in the zero-correlation case, where $\gamma = 0$. Then, suppressing subscripts T,

$$I^2 = \frac{1}{2}\sigma^2 + \sqrt{\frac{\sigma^2}{4} + x^2 v^2}.$$

Under condition (4.2), therefore, implied volatility behaves as $\sigma + O(x^2)$ for x small, but $O(|x|^{1/2})$ for x large. Both limits are consistent with the statics of sections 3.1.2 and 3.3.1.

Application 4.1. Schönbucher applies this model to the pricing of other derivatives as follows. Subject to condition (4.2), the modeller specifies the drift and volatility of I, and infers the dependence of instantaneous volatility σ on the state variables (S, t, I) according to (4.1). Then the price $C(S, t, I)$ of a non-strongly-path-dependent derivative satisfies the usual two-factor pricing equation

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + u\frac{\partial C}{\partial I} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + \gamma\sigma S \frac{\partial^2 C}{\partial S \partial I} + \frac{1}{2} v^2 \frac{\partial^2 C}{\partial I^2} = rC$$

with boundary conditions depending on the particular contract. Finite difference methods can solve such a PDE.

Care should be taken to ensure that I does not become negative.

4.1.2 Term structure of implied volatility

Schönbucher extends this model M different maturities. The implied volatilities to be modelled are $I_t(K_m, T_m)$ for $m = 1, \ldots, M$, where $T_1 < T_2 < \cdots < T_M$. Let

$$V_t^{(m)} := I_t^2(K_m, T_m)$$

be the implied variance. One specifies the dynamics for the shortest-dated variance $V^{(1)}$, as well as all "forward" variances

$$V^{(m,m+1)} := \frac{(T_{m+1} - t)V^{(m+1)} - (T_m - t)V^{(m)}}{T_{m+1} - T_m}.$$

The spot volatility σ_t and the drift and diffusion coefficients of $V_t^{(1)}$ are jointly subject to the drift restriction (4.1) and the no-explosion condition (4.2). Then, given the σ_t and $V_t^{(1)}$ dynamics, specifying each $V^{(m,m+1)}$ diffusion coefficient determines the corresponding drift coefficient, by applying (4.1) to $V^{(m+1)}$.

Application 4.2. To price exotic contracts under these multi-factor dynamics, Schönbucher recommends Monte Carlo simulation of the spot price (which depends on simulation of implied volatilities). Upon expiry of the T_1 option, the T_2 option becomes the "front" contract; at that time $V^{(2)}$ coincides with $V^{(1,2)}$, and at later times its evolution is linked to spot volatility via the drift *and* the no-explosion conditions. Similar transitions occur at each later expiry.

Care should be taken to avoid negative forward variances.

4.2 Statistical approach

Direct modelling of arbitrage-free evolution of an entire implied volatility *surface* remains largely unresolved. Unlike traditional models of spot dynamics, direct implied volatility models face increasing difficulty in enforcing no-arbitrage conditions, when multiple strikes are introduced at a maturity.

Instead of demanding no-arbitrage, the modeller may have a goal more statistical in nature, namely to describe the empirical movements of the implied volatility surface. According to Cont and da Fonseca's [15] analysis of SP500 and FTSE data, the empirical features of implied volatility include the following:

Three principal components explain most of the daily variations in implied volatility: one eigenmode reflecting an overall (parallel) shift in the level, another eigenmode reflecting opposite movements (skew) in low and high strike volatilties, and a third eigenmode reflecting convexity changes. Variations of implied volatility along each principal component are autocorrelated, mean-reverting, and correlated with the underlying.

To quantify these features, Cont and da Fonseca introduce and estimate a d-factor model of the volatility surface, viewed as a function of moneyness m and time-to-maturity τ. The following model is specified under the statistical probability measure:

$$\log I_t(m,\tau) = \log I_0(m,\tau) + \sum_{k=1}^{d} y_t^{(k)} f^{(k)}(m,\tau),$$

where the eigenmodes $f^{(k)}$, such as the three described above, can be estimated by principal component analysis; the coefficients $y^{(k)}$ are specified as mean-reverting Ornstein-Uhlenbeck processes

$$dy_t^{(k)} = -\lambda^{(k)}(y_t^{(k)} - \bar{y}^{(k)})dt + v^{(k)} dW_t^{(k)}.$$

Remark 4.3. If one takes $y_t^{(k)} = 0$ for all k, then $I(m,\tau)$ does not vary in time. This corresponds to an ad-hoc model known to practitioners as "sticky delta." Balland [5] proves that if the dynamics of S are consistent with such a model (or even a generalized sticky delta model in which $I_t(m,\tau)$ is time-varying but determinstic), then assuming no arbitrage, S must be the exponential of a process with independent increments.

Application 4.4. A natural application is the Monte Carlo simulation of implied volatility, for the purpose of risk management.

However, this model, unlike the theory of section 4.1, is not intended to determine the consistent volatility drifts needed for martingale pricing of exotic derivatives. How best to introduce the ideas from this model into a no-arbitrage theory remains an open question.

References

[1] Leif B. G. Andersen and Rupert Brotherton-Ratcliffe. The equity option volatility smile: an implicit finite-difference approach. *Journal of Computational Finance*, 1(2):5–37, Winter 1998.

[2] Marco Avellaneda, Craig Friedman, Richard Holmes, and Dominick Samperi. Calibrating volatility surfaces via relative-entropy minimization. *Applied Mathematical Finance*, 4(1):37–64, 1997.

[3] Marco Avellaneda and Yingzi Zhu. A risk-neutral stochastic volatility model. *International Journal of Theoretical and Applied Finance*, 1(2):289–310, April 1998.

[4] Clifford A. Ball and Antonio Roma. Stochastic volatility option pricing. *Journal of Financial and Quantitative Analysis*, 29(4):589–607, December 1994.

[5] Philippe Balland. Deterministic implied volatility models. *Quantitative Finance*, 2:31–44, 2002.

[6] Ole E. Barndorff-Nielsen. Processes of normal inverse type. *Finance and Stochastics*, 2:41–68, 1998.

[7] Henri Berestycki, Jérôme Busca, and Igor Florent. Asymptotics and calibration of local volatility models. *Quantitative Finance*, 2:61–69, 2002.

[8] Fisher Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–659, 1973.

[9] Ilia Bouchouev and Victor Isakov. The inverse problem of option pricing. *Inverse Problems*, 13(5):L11–L17, October 1997.

[10] Alan Brace, Ben Goldys, Fima Klebaner, and Bob Womersley. Market model of stochastic implied volatility with application to the BGM model. Working Paper S01-1, Department of Statistics, University of New South Wales, 2000?

[11] Linda Canina and Stephen Figlewski. The informational content of implied volatility. *Review of Financial Studies*, 6:659–681, 1993.

[12] Peter Carr and Liuren Wu. The finite moment logstable process and option pricing. *Journal of Finance*, 2002. Forthcoming.

[13] Bent Christensen and Nagpurnanand Prabhala. The relation between implied and realized volatility. *Journal of Financial Economics*, 50:125–150, 1998.

[14] Thomas F. Coleman, Yuying Li, and Arun Verma. Reconstructing the unknown local volatility function. *Journal of Computational Finance*, 2(3):77–102, Spring 1999.

[15] Rama Cont and Jose da Fonseca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2:45–60, 2002.

[16] Freddy Delbaen and Walter Schachermayer. A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 2(4):61–73, Summer 1994.

[17] Emanuel Derman and Iraj Kani. Stochastic implied trees: Arbitrage pricing with stochastic term and strike structure of volatility. *International Journal of Theoretical and Applied Finance*, 1(1):61–110, 1998.

[18] Emanuel Derman, Iraj Kani, and Joseph Z. Zou. The local volatility surface: Unlocking the information in index option prices. *Financial Analysts Journal*, pages 25–36, July/August 1996.

[19] Darrell Duffie, Jun Pan, and Kenneth Singleton. Transform analysis and option pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

[20] Bruno Dupire. Pricing with a smile. *Risk*, 7:18–20, January 1994.

[21] Jean-Pierre Fouque, George Papanicolaou, and K. Ronnie Sircar. *Derivatives in Financial Markets with Stochastic Volatility.* Cambridge University Press, 2000.

[22] Jean-Pierre Fouque, George Papanicolaou, Ronnie Sircar, and Knut Solna. Maturity cycles in implied volatility. Preprint, 2002.

[23] Avner Friedman. *Partial Differential Equations of Parabolic Type.* Prentice-Hall, 1964.

[24] Jim Gatheral. The volatility skew: Arbitrage constraints and asymptotic behaviour. Merrill Lynch, 1999.

[25] Jim Gatheral. Lecture 2: Fitting the volatility skew. Case Studies in Financial Modelling course notes, Courant Institute, 2001.

[26] Henryk Gzyl and Minaya Villasana. A perturbative approach to implied volatilities. Preprint, 2002.

[27] Patrick Hagan, Deep Kumar, Andrew Lesniewski, and Diana Woodward. Managing smile risk. Preprint, 2002.

[28] Patrick Hagan and Diana Woodward. Equivalent black volatilities. *Applied Mathematical Finance*, 6(3):147–157, April 1999.

[29] Steven L. Heston. A closed-form solution for options with stochastic volatility and applications to bond and currency options. *Review of Financial Studies*, 6(2):327–343, 1993.

[30] Hardy M. Hodges. Arbitrage bounds on the implied volatility strike and term structures of European-style options. *Journal of Derivatives*, pages 23–35, 1996.

[31] John Hull and Alan White. The pricing of options on assets with stochastic volatilities. *Journal of Finance*, 42(2):281–300, June 1987.

[32] Steven G. Kou. A jump diffusion model for option pricing. Preprint, 2002.

[33] Steven G. Kou and Hui Wang. Option pricing under a double exponential jump diffusion model. Preprint, 2001.

[34] Olivier Ledoit, Pedro Santa-Clara, and Shu Yan. Relative pricing of options with stochastic volatility. Preprint, November 2001.

[35] Roger W. Lee. Implied and local volatilities under stochastic volatility. *International Journal of Theoretical and Applied Finance*, 4(1):45–89, 2001.

[36] Roger W. Lee. Option pricing by transform methods: Extensions, unification, and error control. Stanford University and Courant Institute, 2001.

[37] Roger W. Lee. The moment formula for implied volatility at extreme strikes. Stanford University and Courant Institute, 2002.

[38] Alan L. Lewis. *Option Valuation under Stochastic Volatility.* Finance Press, 2000.

[39] Riccardo Rebonato. *Volatility and Correlation in the Pricing of Equity, FX and Interest Rate Options.* John Wiley & Sons, 1999.

[40] Eric Renault and Nizar Touzi. Option hedging and implied volatilities in a stochastic volatility model. *Mathematical Finance*, 6(3):279–302, July 1996.

[41] Philipp L. Schönbucher. A market model for stochastic implied volatility. Bonn University, 1998.

[42] K. Ronnie Sircar and George Papanicolaou. Stochastic volatility, smile & asymptotics. *Applied Mathematical Finance*, 6(2):107–145, June 1999.