Chapter 10

Smile Dynamics and Stochastic Volatility

In these notes we show that fitting every option in the market is not enough. We derive, rather than assert, the relationship between the slope of a smile and the way that smile moves, and find that a local volatility model gets it wrong by a factor of two in the wrong direction. We then introduce models in which volatility is genuinely random, and show that the market's quotes do not determine which one to use.

10.1 What Is Left to Get Wrong

Chapter 9 ended in an uncomfortable place. Dupire's formula fits every European option in the market exactly, with no fitting and no error, using a model in which volatility carries no randomness at all. If a model reproduces every price we can see, in what sense can it be wrong?

Gyongi's theorem already told us. A local volatility model matches the *marginal* distribution of the underlying at each date. It says nothing about the joint distribution across dates, and infinitely many joint distributions share a set of marginals. So the model is guaranteed to be right about anything determined by one date at a time, and is unconstrained about everything else.

- Anything path dependent. A barrier option cares whether the underlying visited a level, which is a statement about the whole path.

- Anything with an early exercise decision. A Bermudan swaption's value depends on what the smile will look like on each exercise date.

- Anything forward starting. A cliquet struck at the money in a year's time is a bet on volatility a year from now.

- *The hedge*. This one is easy to miss and is the most important, because it applies to the plain vanilla options themselves.

The last point deserves spelling out. Suppose we and the market agree exactly on the price of every option today. We then hold one and hedge it. The hedge ratio is

$$\Delta = \frac{dV}{dS},$$

the total derivative of the option's value with respect to the underlying — and V depends on S both directly and through the volatility we will use to price it after S has moved. Writing $V = \text{Black}(S, K, \sigma_{\text{imp}}(K, S), T)$ and differentiating,

$$\Delta = \underbrace{\frac{\partial \text{Black}}{\partial S}}_{\text{the textbook delta}} + \underbrace{\frac{\partial \text{Black}}{\partial \sigma} \cdot \frac{\partial \sigma_{\text{imp}}}{\partial S}}_{\text{the smile's contribution}} . \tag{10.1}$$

The first term is what chapter 5 gave us. The second is vega multiplied by the rate at which the implied volatility of *this* option changes when the underlying moves. Vega is large — for a one year at-the-money option it is roughly 0.4 of the underlying per unit of volatility — so the second term is not a refinement.

And $\partial\sigma_{\text{imp}}/\partial S$ is not observable today. It is a statement about the future, and therefore a statement the model has to make. Two models agreeing on every price today will disagree on the hedge unless they also agree about this. That is the subject of the chapter.

10.2 Two Rules of Thumb, and What They Would Mean

Before deriving what a model says, it helps to have the two things a trader might say, because they bracket the answer.

Definition 10.1 (Sticky strike)**.** The smile is sticky strike if the implied volatility of a *fixed strike* does not change when the underlying moves: σ_{imp} is a function of K alone.

Definition 10.2 (Sticky delta, or sticky moneyness)**.** The smile is sticky delta if the implied volatility depends only on the strike relative to the forward: $\sigma_{\text{imp}} = \varphi(K/F)$. The smile rides along with the market.

To compare them we need two quantities that are easy to confuse, so let us name them carefully. Both are measured at the money and in log-moneyness $y = \ln(K/F)$, which is the coordinate that makes them comparable.

Definition 10.3 (Skew and backbone)**.** The *skew* is the slope of today's smile in the strike direction, with the forward held still:

$$s = \left.\frac{\partial \sigma_{\text{imp}}}{\partial y}\right|_{y=0}.$$

The *backbone* is the rate at which the at-the-money volatility itself moves when the forward moves:

$$b = \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F}, \qquad \sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(K = F, F).$$

The skew is a photograph. The backbone is a motion. They are different numbers, they are measured on different days. The market lets us observe the first and not the second.

Before the rules, one identity, because these two numbers are constantly confused and the identity is what keeps them apart. There is a third derivative in play — how the volatility of *one fixed strike* moves — and all three are related by the chain rule.

Lemma 10.4 (The three derivatives)**.** *For any smile,*

$$b = s + \left.\frac{\partial \sigma_{imp}}{\partial \ln F}\right|_{K \text{ fixed}}, \tag{10.2}$$

both derivatives on the right evaluated at the money.

Proof. $\sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(F, F)$, so moving the forward moves both arguments. Differentiating,

$$b = \frac{d}{d \ln F}\, \sigma_{\text{imp}}(F, F) = \left.\frac{\partial \sigma_{\text{imp}}}{\partial \ln K}\right|_{K=F} + \left.\frac{\partial \sigma_{\text{imp}}}{\partial \ln F}\right|_{K=F},$$

and the first term is the skew by definition. □

Now the two rules.

Under **sticky delta**, $\sigma_{\text{ATM}} = \varphi(1)$ is a constant, so $b = 0$: the at-the-money volatility never moves, however far the market travels. From (10.1), with K fixed and $\sigma_{\text{imp}} = \varphi(K/F)$, the smile's contribution to delta is $\partial\sigma_{\text{imp}}/\partial \ln F = -s$ at the money.

Under **sticky strike**, $\sigma_{\text{imp}}(K)$ does not move, so as the forward slides along the fixed smile the at-the-money volatility takes the value of the smile at the new level: $\sigma_{\text{ATM}}(F) = \sigma_{\text{imp}}(F)$, and $b = s$. The smile's contribution to delta is zero, since the volatility of the option we hold does not change at all.

So the two conventions bracket the answer: $b = 0$ at one end and $b = s$ at the other. In an equity market with a downward skew, $s < 0$, so the two say the at-the-money volatility either stays put or falls a little when the market rises.

We are now in a position to ask what a local volatility model says — and the answer is outside the bracket.

10.3 The Backbone of a Local Volatility Model

The one ingredient we need from outside is a short-maturity result about how a local volatility function turns into an implied volatility. It has a clean statement and a clean intuition.

Theorem 10.5 (Berestycki, Busca and Florent)**.** *In a local volatility model, as the expiry shrinks to zero the implied volatility converges to the harmonic mean of the local volatility along the path in log-space from the forward to the strike:*

$$\sigma_{imp}(K, F) \;\longrightarrow\; \left(\frac{1}{\ln K - \ln F} \int_{\ln F}^{\ln K} \frac{dz}{\sigma_{loc}(e^z)} \right)^{-1}. \tag{10.3}$$

Why it is true, before why it is a theorem. Think of the underlying as diffusing through a medium whose diffusivity varies from place to place. The option struck at K is asking how hard it is to get from F to K. In a medium of constant diffusivity σ, covering a log-distance d takes a characteristic time $(d/\sigma)^2$, so the natural measure of separation is d/σ, and in a varying medium the separation of F and K is the accumulated $\int dz/\sigma_{\mathrm{loc}}(e^z)$. Implied volatility is by definition the constant volatility reproducing the same option price, hence the same separation, so

$$\frac{\ln K - \ln F}{\sigma_{\mathrm{imp}}} = \int_{\ln F}^{\ln K} \frac{dz}{\sigma_{\mathrm{loc}}(e^z)},$$

which is (10.3) rearranged. Implied volatility is a harmonic average because it is an average of a *rate*, and rates average harmonically — for the same reason that driving one mile at 30 and one mile at 60 does not average to 45. □

Remark (What turning that into a proof requires)**.** The argument above is not a proof and the gap is not a technicality, so it pays to see where the work is. The rigorous chain has four links.

One. The short-time behaviour of the transition density is governed by a distance. Varadhan's theorem says

$$-2t \log p(t, x, y) \;\longrightarrow\; d(x, y)^2,$$

where d is the Riemannian distance in the metric $ds = dz/\sigma_{\mathrm{loc}}(e^z)$ — precisely the separation the heuristic wrote down. So the heuristic names the right object; what it does not do is establish the limit. The geometry that this metric carries is taken up later in the chapter.

Two. The same asymptotics transfer from the density to the option price, since an out-of-the-money call is an integral of the density over a region whose nearest point to the forward is the strike, and a Laplace-type estimate says an integral of $e^{-d^2/2t}$ is governed by the smallest d in the region.

Three. Black-Scholes with a constant volatility is the special case $\sigma_{\mathrm{loc}} \equiv \sigma_{\mathrm{imp}}$, where the distance is $|\ln K - \ln F|/\sigma_{\mathrm{imp}}$. Equating the two distances is the theorem.

Four. Step one is classically available for a uniformly elliptic operator with smooth coefficients on a compact space, and none of those three conditions holds here: the operator degenerates as $S \to 0$, the domain is not compact, and σ_{loc} coming out of a calibration is at best continuous. The route that survives this substitutes $u_t = -2t \log C$ directly into the pricing equation, which turns a linear parabolic equation into a nonlinear one, and the limit as $t \downarrow 0$ is a first-order Hamilton-Jacobi equation — an eikonal equation $|\nabla u| = 1$, whose solutions are exactly distance functions. Two facts about that equation then have to be supplied: that the limit satisfies it, and that it has only one solution with the given boundary behaviour. Neither is available in the classical sense, because a distance function is not differentiable where two geodesics meet. Viscosity solutions are the notion of solution designed for exactly that, and the comparison principle for them is what delivers uniqueness.

A degenerate limit of a linear problem became a nonlinear problem, and a nonlinear first-order equation with non-smooth solutions needs a weak notion of solution to be well posed at all. The same substitution and the same difficulty appear wherever a small-noise limit is taken; large deviations theory is this argument in general form, with the exponential rate playing the part of the distance here.

From this we get the fact we actually use.

Lemma 10.6 (The midpoint rule)**.** *To first order in* $\ln K - \ln F$*,*

$$\sigma_{imp}(K, F) = \sigma_{loc}\big(\sqrt{FK}\big) + O\big((\ln K - \ln F)^2\big),$$

that is, the implied volatility of a strike is the local volatility evaluated halfway between the forward and the strike, measured in log-space.

Proof. Write $g(z) = 1/\sigma_{\text{loc}}(e^z)$, and let

$$m = \frac{\ln F + \ln K}{2}, \qquad \delta = \frac{\ln K - \ln F}{2},$$

so the integral in (10.3) runs from $m-\delta$ to $m+\delta$ and the mean is taken over a width 2δ. Expanding g about the midpoint,

$$\frac{1}{2\delta}\int_{m-\delta}^{m+\delta} g(z)\,dz = \frac{1}{2\delta}\int_{-\delta}^{\delta}\Big(g(m) + u\,g'(m) + \tfrac{1}{2}u^2 g''(m) + \dots\Big)du = g(m) + \frac{\delta^2}{6}g''(m) + \dots,$$

the term linear in u integrating to zero by symmetry. So the harmonic mean is $1/g(m)$ up to a correction of order δ^2, and $1/g(m) = \sigma_{\text{loc}}(e^m) = \sigma_{\text{loc}}(\sqrt{FK})$. □

The midpoint rule is the whole story in one line: *implied volatility is local volatility, averaged over the journey.* Averaging halves a slope, and that is where the factor of two lives.

Remark (The same two as chapter 9)**.** This factor is not a new one. Chapter 9 established, statically, that the local volatility curve has twice the slope of the implied volatility curve — the rule of two — and for the same reason: implied volatility is an average of local volatility over the interval between the forward and the strike, and averaging a function over an interval halves its slope.

The market's verdict is shared too. Chapter 9 measured the ratio and found it near two only close to the money, decaying further out, because both statements are leading order. And the backbone measured in the market is not $2s$ either — which is the complaint this chapter is built around, arriving in a second form.

Theorem 10.7 (The local volatility backbone)**.** *In a local volatility model, the at-the-money volatility responds to a move in the forward at twice the rate of the skew:*

$$b = 2s. \tag{10.4}$$

Proof. The essential point is that σ_{loc} is a fixed function of the *level* of the underlying. It was calibrated once, to today's surface, and it does not move when the market does. Write it in log-space as $\ell(x) = \sigma_{\text{loc}}(e^x)$. By Lemma 10.6,

$$\sigma_{\text{imp}}(K, F) = \ell\bigg(\frac{\ln F + \ln K}{2}\bigg).$$

Both quantities we want are derivatives of this single expression, taken in different directions.

For the skew, hold F fixed and differentiate in $\ln K$, then set $K = F$:

$$s = \frac{\partial \sigma_{\text{imp}}}{\partial \ln K}\bigg|_{K=F} = \frac{1}{2}\,\ell'(\ln F).$$

For the backbone, set $K = F$ *first* — the at-the-money option is a different contract at each level of the forward — giving $\sigma_{\text{ATM}}(F) = \ell(\ln F)$, and then differentiate:

$$b = \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F} = \ell'(\ln F).$$

Dividing, $b = 2s$. □

The factor of two comes from a genuinely simple place. The skew moves only the strike, so it moves the midpoint by half as much; the backbone moves the forward and the strike together, so it moves the midpoint by the full amount. Chapter 9 met the same factor from the other side — its rule of two, that the local volatility curve extracted from a smile is twice as steep as the smile itself, is this statement read backwards. The two derivations are independent, and `local_volatility_has_twice_the_slope_of_the_implied_smile` in `quant/src/localvol.rs` checks the result numerically against a Dupire formula that shares none of the algebra.

Remark (Why this is bad, in numbers)**.** Put the three predictions side by side. Take a one year at-the-money option with $\sigma_{\text{ATM}} = 25\%$ and a skew of $s = -6$ volatility points per unit of log-moneyness, which is an ordinary equity index figure. Suppose the market rises 10%.

	backbone b	change in σ_{ATM}
sticky delta	0	0.00 points
sticky strike	s	-0.6 points
local volatility	$2s$	-1.2 points

Equity index markets behave much closer to the first row than the third. The local volatility model does not merely get the magnitude wrong; it predicts the at-the-money volatility falls twice as fast as even the more aggressive of the two rules of thumb, and it does so because it was fitted to today's skew. The steepness was not chosen. It was forced on the model by the requirement to match the smile.

Now feed that into (10.1). With vega around $0.4 \times S$ per unit volatility, an error of 1.2 volatility points per 10% move is a delta error of roughly $0.4 \times 0.012/0.10 = 0.048$, about five percent of the underlying's notional, on every option in the book, in the same direction. That is not a rounding error; it is a systematic position nobody chose to take.

Remark (No vega)**.** A second failure follows from the same source and can be stated quickly. Because there is no independent volatility factor, the model is complete in the sense of chapter 4: the underlying and the money market replicate everything, so there is no vega to hedge and nothing to hedge it with. A desk whose business is trading volatility needs volatility to be a risk factor, and here it is not one.

10.4 The Backbone the Market Actually Has

The backbone above is a modelling choice presented as a menu. A rate's volatility is taken proportional to some power of its level,

$$\sigma(r) = \alpha\, r^{\beta}, \tag{10.5}$$

and three values of β have names. At $\beta = 0$ the model is normal, so a yield at one per cent moves in basis points exactly as violently as a yield at ten. At $\beta = 1$ it is lognormal and moves ten times as violently. At $\beta = \frac{1}{2}$ it is the square-root middle that keeps the rate positive without letting the volatility scale fully. SABR carries the same exponent under the same letter, and it is usually not calibrated at all — it is set to one of the three by convention and the remaining parameters absorb the difference.

These notes can do better than convention here, because the underlying is observable. The curve is quoted every business day and has been for sixty-four years.

Remark (Why a physical measure estimate is admissible)**.** The objection to measuring β from history is the usual one: hedging happens under the pricing measure and a time series is drawn under the physical one, so the estimate appears to answer a different question.

It does not, and chapter 6 says why. Girsanov changes the drift of a process and leaves its diffusion coefficient exactly where it was, so (10.5) — which is a statement about the diffusion coefficient and nothing else — is the same equation under both measures. The premium lives in α's drift and in the level, not in the exponent. So β is one of the few quantities a pricing model needs that can be read off history without a change of measure.

Calculation 10.8 (The exponent, from sixty-four years of the ten-year yield)**.** Cut the history into non-overlapping months, take the realised volatility of daily changes within each and the average level across it, and regress one log on the other. Overlapping windows would quadruple the apparent sample without adding information and make the standard error a fiction.

Over the whole series the level ranges from 0.59% to 15.2%, a factor of twenty-six, and it is that range which identifies the exponent at all — a few months of data cannot see it, however many days they contain. The result is

$$\beta = 0.27 \pm 0.04,$$

on 768 monthly windows.[1] Every one of the three conventional choices is rejected: normal by 6.8 standard errors, square root by 5.8, and lognormal by 18.5.

Structure (What the number is saying, and what it is not)**.** The measured exponent is closer to normal than to anything else on the menu, and that is the empirical content of a convention the rates market adopted for a different reason. Normal quoting spread after 2008 because lognormal volatility is undefined at a negative rate and rates went negative; the argument was that the alternative had stopped working, not that it was right. The exponent says it was closer to right all along.

Two limits on the claim, both of which the figure shows better than the number does. The scatter is wide: a single month's realised volatility is a poor estimate, and the regression is extracting a slope from a cloud rather than tracing a curve. And the sample crosses regimes that differ in more than the level — a decade of targeting money supply and a decade of targeting a rate are not the same experiment run at different heights. The exponent is what remains after averaging over all of that, and averaging over it is not the same as controlling for it.

[1] `measured`.

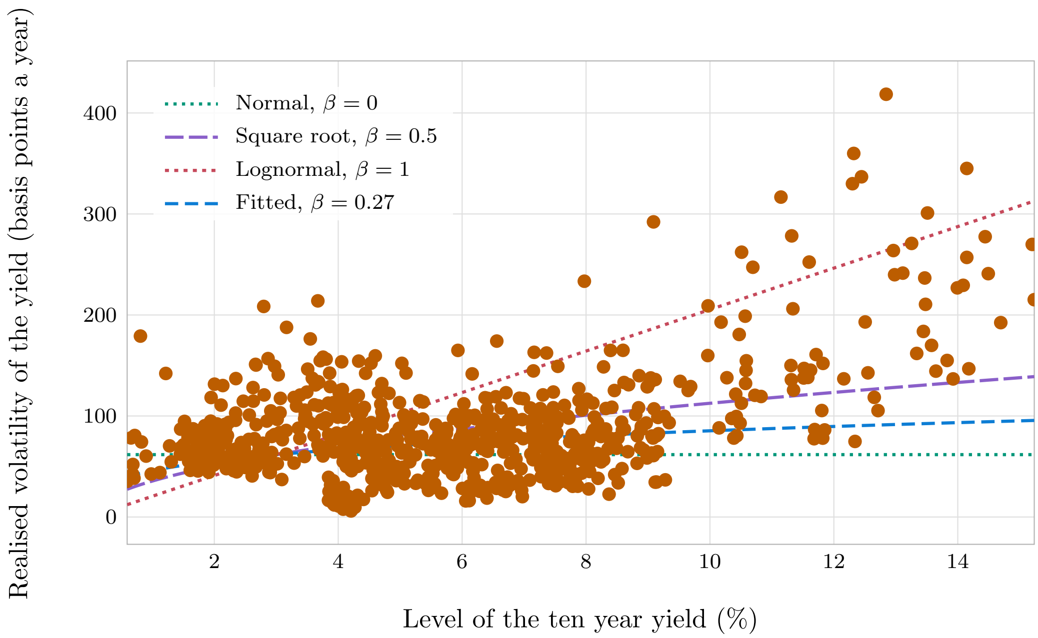

Figure 10.1: Realised volatility of the ten-year yield against its level, one point per non-overlapping month since 1962, with the three conventional exponents and the fitted one drawn through the sample's centre so that they are compared by slope rather than by height. The cloud is wide, which is why sixty-four years are needed and why the standard error matters more than the point estimate. What the picture shows is that the lognormal line is far too steep to pass through it — a yield at twelve per cent was not ten times as volatile as a yield at one, but roughly twice.

Drawn by `backbone`. Board of Governors of the Federal Reserve System (US), H.15 Selected Interest Rates, 10-year Treasury constant maturity (DGS10), retrieved from FRED, Federal Reserve Bank of St. Louis, 1962-01-02 to 2026-08-14 (constant maturity, bond equivalent yield, business days). Retrieved from https://fred.stlouisfed.org/graph/fredgraph.csv?id=DGS10.

10.5 The Forward Smile

The *forward smile* is the smile of an option on the ratio $S_{t+\tau}/S_t$, struck at the money at a future date t. Cliquets, ratchets and most structured equity payoffs are strings of them, so a model that gets the forward smile wrong misprices a whole product family without any vanilla price revealing it.

It helps to keep two term structures of skew apart, because the chapter turns on the difference.

- The *maturity* term structure: how today's skew varies with the expiry T. This is quoted, so we can read it off the market.
- The *forward* term structure: how the skew of a fixed tenor τ varies with the start date t. Nothing quoted pins this down, so the model supplies it.

The market says something definite about each. The first *decays*, roughly as $1/\sqrt{T}$. The second is roughly *stationary*: a one-year skew a year forward looks much like today's one-year skew, which is what one would expect of a market whose behaviour does not depend on the date in the calendar.

The question of this section is whether one model can do both.

10.5.1 A Local Volatility Model With No Calendar In It

Start with the plainest local volatility model that has a skew at all, constant elasticity of variance,

$$dS_t = \sigma_0 \left(\frac{S_t}{S_0}\right)^{\beta-1} S_t\, dW_t, \qquad \beta < 1, \tag{10.6}$$

in which volatility rises as the level falls, which is the equity shape. Simulating it gives both term structures at once.[2] At $\sigma_0 = 25\%$ and $\beta = 0.7$, neither one decays: the five-year skew stays close to the three-month skew, and the one-year forward skew starting in four years stays close to today's.

Structure. Both flat readings are the same fact. Equation (10.6) is *time homogeneous* — its coefficient depends on the level and not on the date — so the model looks identical from every starting date, and its forward smile is its spot smile. There is nothing in it that could distinguish one calendar date from another, so nothing that could make either term structure slope.

So CEV gets the forward term structure right and the maturity term structure wrong. It cannot be calibrated to a real surface at all.

10.5.2 What Calibration Puts In

Now calibrate. The one empirical fact we need is the $1/\sqrt{T}$ decay, so take the simplest surface with that shape,

$$\sigma_{\mathrm{BS}}(k,T) = \sigma + \psi(T)\, k, \qquad \psi(T) = \frac{\psi_1}{\sqrt{T+T_0}}, \tag{10.7}$$

with $k = \ln(K/F)$ and a short-maturity floor T_0, since no real skew is infinite at zero maturity. Nothing below depends on the smile being straight in k; only on ψ falling with T.

[2] `localvol::Cev`, `the maturity ratio`, and `the forward-starting ratio`.

First, run Dupire's formula of chapter 9 on it. At zero rates that formula reads

$$\sigma^2_{\mathrm{loc}}(T,K) = \frac{2\,\partial C/\partial T}{K^2\,\partial^2 C/\partial K^2},$$

and writing $w(k,T) = \sigma^2_{\mathrm{BS}} T$ for the total implied variance and expanding to first order in the skew, the at-the-money slope of the local volatility comes out as

$$\left.\frac{\partial \sigma_{\mathrm{loc}}}{\partial k}\right|_{k=0} \;=\; \psi_1\,\frac{\frac{3}{2}T + 2T_0}{(T+T_0)^{3/2}}. \tag{10.8}$$

Two limits stand out. At $T = 0$ it is $2\psi_1/\sqrt{T_0}$, exactly twice the implied skew: the rule of two, recovered. At large T it is $\frac{3}{2}\psi_1/\sqrt{T}$, so the factor has fallen from 2 to $\frac{3}{2}$. Both ends, and the monotone passage between them, agree with a numerical Dupire to within five percent.[3]

Second, read (10.8) as a statement about *calendar* time. It is the local volatility at date T, and its skew falls as that date advances: by ten years it is a third of today's. The decay we fed in across the maturity axis has come back out along the time axis. The calibrated model is no longer time homogeneous, and could not be, because the surface it was asked to match is not.

Third, price the forward starting option. A short-dated option beginning at t diffuses through the slice of the surface at t, and there the rule of two applies in reverse: its implied skew is half the local skew at t. So

$$\psi_{\mathrm{fwd}}(t) \;\approx\; \tfrac{1}{2}\,\psi_1\,\frac{\frac{3}{2}t + 2T_0}{(t+T_0)^{3/2}} \;\longrightarrow\; \frac{3\psi_1}{4\sqrt{t}} \qquad \text{for large } t. \tag{10.9}$$

Remark (What that costs)**.** Equation (10.9) decays as $1/\sqrt{t}$, so doubling the forward start date multiplies the forward skew by $1/\sqrt{2}$. With an equity-shaped surface — $\sigma = 20\%$, $\psi_1 = -0.10$, T_0 one month — the forward skew is a fifth of today's after one year, an eighth after four, and under a tenth after ten. A cliquet is then priced off skews that have very nearly vanished, and because it is a string of them the error compounds along the string rather than cancelling.

Comparing the two subsections, a local volatility model has only one dial for both term structures. Dupire converts decay across maturities into decay along the date axis, so matching the observable one determines the unobservable one — and determines it wrongly.

10.5.3 Whether the Trade Is Forced

It is fair to ask whether any model can escape that, or whether decaying skew simply implies a decaying forward skew. It does not, and the cleanest demonstration uses the model of §10.8,

$$dS_t = \sqrt{v_t}\,S_t\,dW_t, \qquad dv_t = \kappa(\theta - v_t)\,dt + \eta\sqrt{v_t}\,dZ_t, \qquad dW_t\,dZ_t = \rho\,dt, \tag{10.10}$$

taken here only for the measurement; what its parameters do is §10.8's business.

The point is what (10.10) does *not* contain: any function of t. Every parameter is a constant, so the model is as time homogeneous as CEV was, and by the same argument its forward smile cannot depend on the start date once v_t has reached its stationary distribution. Yet its skew still decays

[3] `localvol::SkewSurface`.

with expiry — because the decay is produced by the variance mean reverting *over the option's own life*, which is a longer journey for a longer option, rather than by any parameter that knows the date.

Simulating (10.10) at $v_0 = \theta = 4\%$, $\kappa = 2$, $\eta = 0.3$ and $\rho = -0.7$ confirms both halves.[4]

	skew at 5y expiry (share of the 3m skew)	1y skew starting in 4y (share of today's)
market	about 22%	about 100%
CEV, equation (10.6)	near 100%	near 100%
calibrated local volatility	fitted exactly	11%
Heston, equation (10.10)	20%	97%

The market row uses the $1/\sqrt{T}$ rule, which gives $\sqrt{0.25/5} = 22\%$, and stationarity for the forward column. Heston's 20% landing beside it is a better agreement than the model deserves in general — its long-dated skew decays as $1/T$ rather than $1/\sqrt{T}$, and §10.9 has a separate complaint about its short end — but over this range the shape is right, and the forward column is the one that matters here.

Structure. The separation is possible because the two models locate the skew's decay in different places. Local volatility has one state variable, so the only place a term structure can live is in the coefficient's dependence on t — and a coefficient that depends on t is precisely what a forward starting option is exposed to. Stochastic volatility has a second state variable, and can put the term structure in the *dynamics of that variable* instead, where a forward starting option, which begins after the variance has settled, never sees it.

This is the sharper form of the completeness argument above. The second Brownian motion is not merely something to hedge vega with; it is somewhere to keep a term structure that would otherwise have to be written into the calendar.

One caveat on the mechanism, since it is the stationary distribution and not time homogeneity alone that does the work. Start Heston hot, at $v_0 = 9\%$ against $\theta = 4\%$, and the forward skew is *not* flat in the first year: today's one-year skew is diluted by a variance that has not yet reverted, and the skew has to travel before it settles. By one year it has, since $1/\kappa$ is six months, and successive start dates then agree to within a few percent. A model calibrated on a day when volatility is far from its long-run level therefore has a genuine forward term structure over the first few mean-reversion times, and only after that does stationarity take over.

10.6 Making Volatility Random

The diagnosis suggests the cure. The trouble was that volatility was pinned to the level of the underlying, so fitting the skew forced a backbone. Give volatility its own source of randomness and the two can be separated.

[4] `heston::Heston`. Spot and forward smiles come from the same routine, so no scheme difference stands between the two columns.

Definition 10.9 (SABR)**.** The SABR model is

$$\begin{aligned} dF_t &= \alpha_t F_t^{\beta}\, dW_t, \\ d\alpha_t &= \nu \alpha_t\, dZ_t, \qquad dW_t\, dZ_t = \rho\, dt, \end{aligned}$$

with $\alpha_0 = \alpha$, and $\beta \in [0, 1]$.

The forward is a martingale, as it must be under the measure of chapter 6, and its volatility α_t is itself a driftless lognormal process. Four parameters, and the reason the model is used on trading desks is that each one does a recognisable job:

- α sets the *level* of the smile.
- ν, the volatility of volatility, sets its *curvature*. With $\nu = 0$ the volatility is deterministic and the smile has no smile in it.
- ρ, the correlation, *tilts* it. Negative ρ means the volatility rises as the forward falls, which makes low strikes expensive.
- β also tilts it — and this is the interesting one, because it tilts it for a different reason.

What makes SABR usable at all is that Hagan and coauthors found an asymptotic formula for the implied volatility, so a vanilla can be priced without any simulation. At the money it reduces to something short:

$$\sigma_{\text{ATM}} \approx \frac{\alpha}{F^{1-\beta}} \left(1 + \left[\frac{(1-\beta)^2}{24} \frac{\alpha^2}{F^{2-2\beta}} + \frac{1}{4} \frac{\rho\beta\nu\alpha}{F^{1-\beta}} + \frac{2-3\rho^2}{24} \nu^2 \right] T \right). \tag{10.11}$$

The leading term is the whole of the intuition and the bracket is a correction that matters at longer expiries. The full expression, for all strikes, is `Sabr::implied_vol` in `quant/src/sabr.rs`.

Remark (It is an expansion, and expansions expire)**.** Hagan's formula is asymptotic in the time to expiry, not exact. For short and moderate expiries it is very accurate. For long ones it degrades, and at low strikes it can imply a negative probability density — which by chapter 9 is an arbitrage, sitting inside the formula the market uses to quote. This is well known and is patched in practice, either by replacing the wings below some strike or by using an arbitrage-free version of the model. The industry's standard smile formula is not arbitrage-free, and it matters to know exactly where it fails.

Now look at β properly, because it is the parameter that controls the thing this chapter is about. Set $\nu = 0$ for a moment, so volatility is deterministic. The leading term of (10.11) gives

$$\sigma_{\text{ATM}}(F) = \alpha F^{\beta - 1}, \qquad \text{so} \qquad \frac{\partial \sigma_{\text{ATM}}}{\partial \ln F} = (\beta - 1)\, \sigma_{\text{ATM}}.$$

So β *is* the backbone. At $\beta = 1$ the volatility of the forward is proportional to the forward, the model is lognormal, and the at-the-money volatility does not move at all when the market does — exactly sticky delta. At $\beta = 0$ the volatility is an absolute amount, so a rise in the forward means the same absolute movement is a smaller relative one, and the at-the-money volatility falls hard.

And ρ produces skew without producing backbone. So we have two dials that both tilt today's smile and disagree completely about tomorrow.

10.7 The Smile Does Not Determine the Dynamics

This is the punchline.

Fix a market: an at-the-money volatility of 25%, an ordinary downward skew, and a vol-of-vol of 0.35. Now fit the model with β held at various values, letting α and ρ adjust to reproduce that level and that skew. Every fit succeeds — there is always a correlation that reproduces any given skew — and the fitted smiles are nearly identical.[5]

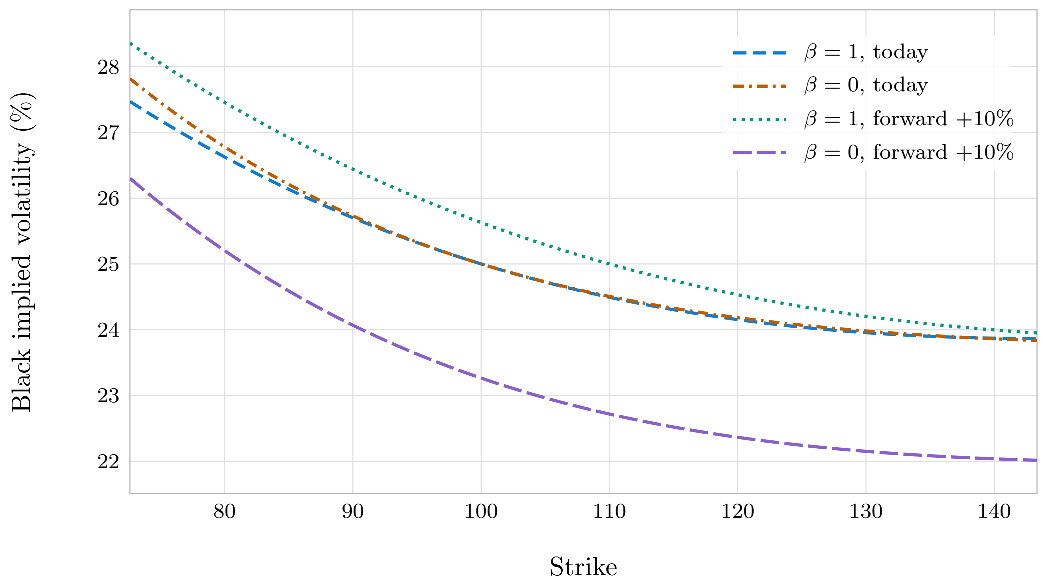

Figure 10.2: Two SABR models fitted to the same at-the-money volatility and the same skew. Today, their smiles lie on top of one another to within a fifth of a volatility point across the whole quoted range, so no set of vanilla quotes could distinguish them. Move the forward up by ten percent and they are more than two volatility points apart at the money. The parameter separating them, β, is invisible in today's prices and is the one that decides the hedge.

Drawn by `Sabr::implied_vol`, `Sabr::backbone` and `calibrate_alpha_rho`.

β	ρ	backbone b	b/s	σ_{ATM} after a 10% rise
1.0	−0.34	0	0	25.00%
0.7	−0.13	−0.00075	1.25	24.30%
0.5	+0.02	−0.00125	2.09	23.84%
0.0	+0.37	−0.00251	4.19	22.72%

All four rows price today's options the same. They disagree by more than two volatility points about where the at-the-money volatility goes after a move that markets make routinely, and therefore they disagree about the delta to put on today.

Remark (A check falling out of the table)**.** The row at $\beta = 0.5$ has $\rho \approx 0$, so its volatility is very nearly a deterministic function of the forward — it is essentially a local volatility model. Its backbone-to-skew ratio comes out at 2.09.

[5] Every column of the table below, at every row: `the_smile_dynamics_backbone_table_holds_at_every_beta`.

That is Theorem 10.7 arriving from a completely different direction. We derived $b = 2s$ from a short-maturity harmonic mean; here it emerges from Hagan's asymptotic expansion of a stochastic volatility model that happens to have no stochastic volatility left in it. Two independent routes to the same factor of two is the kind of agreement that makes a result trustworthy. The check is `the_smile_does_not_determine_the_dynamics`.

So how is β chosen, if the smile will not choose it? Not from vanilla prices, because they are silent. In practice it is fixed by judgement about the market — close to 1 for an equity index, lower for rates, near 0 where rates can be negative and an absolute volatility is the only sensible one — or it is inferred from the historical relationship between the at-the-money volatility and the level, which is precisely a measurement of the backbone. It is then usually held fixed, and α, ρ and ν are re-marked daily.

Remark (The general lesson)**.** Calibration is not the same as identification. A model can be fitted perfectly and still be undetermined in the direction that matters, because the instruments used to fit it are blind to that direction. Vanilla options are a set of one-date distributions, so they constrain the marginals and nothing else — which is Gyongi's theorem again, now doing damage rather than favours.

The practical consequence: if a parameter matters for your product and is invisible in your calibration set, then either the calibration set has to be enlarged — with forward starting options, or barriers, or historical dynamics — or the parameter has to be marked by hand and treated as a model risk, with a number attached.

10.8 Heston, and What a Term Structure Costs

SABR has one weakness that follows from its own definition: α_t is a driftless lognormal process, so it has no mean reversion and no long-run level. Over a long horizon it wanders arbitrarily far, and the model has nothing to say about a term structure of volatility. In practice this is handled by fitting a separate set of SABR parameters to every expiry, which prices every slice correctly and leaves the slices with no relationship to one another — fine for interpolating vanillas, useless for anything spanning two dates.

The other classical choice fixes exactly this.

Definition 10.10 (Heston)**.**

$$\begin{aligned} dS_t &= \mu S_t\,dt + \sqrt{v_t}\,S_t\,dW_t, \\ dv_t &= \kappa(\theta - v_t)\,dt + \eta\sqrt{v_t}\,dZ_t, \qquad dW_t\,dZ_t = \rho\,dt. \end{aligned}$$

The variance v_t is a mean-reverting square-root process: it is pulled towards a long-run level θ at a speed κ, and the $\sqrt{v_t}$ in its own diffusion keeps it non-negative, since the randomness switches off as it approaches zero. Those two features are what buy a term structure — θ sets where volatility goes in the long run, κ how fast it gets there — and they are what SABR lacks.

Heston is tractable for a reason that chapter 14 gives for any model being solvable at all. Its characteristic function

$$\phi(u) = \mathbb{E}\big[e^{iu\ln S_T}\big] = \exp\big(A(u,T) + B(u,T)v_0\big)$$

is available in closed form: the exponent is *affine* in the state, and finding A and B reduces to a pair of ordinary differential equations rather than a partial one. Option prices then follow by a single numerical integral. This is the affine structure that chapter 14 makes explicit and chapter 12 puts to work on interest rates.

The trade one is making between the two is roughly: SABR gives a better fit to one expiry's smile with parameters a trader can interpret, and Heston gives a consistent story across expiries at the cost of a stiffer fit to any one of them. Neither fits the whole surface exactly, and for a desk that has to reprice the market to the penny, that is a problem — which is where the next chapter starts.

10.9 The Short-Dated Skew, and Rough Volatility

Before leaving stochastic volatility there is one failure the two models share, and it is not a matter of degree. It concerns what happens to the skew as the expiry shrinks.

Measure the at-the-money skew s of §10.2 at a range of expiries. In SABR — and in Heston, and in classical stochastic volatility generally — it settles down to a constant as $T \to 0$. That is not an artefact of Hagan's expansion; it is a property of the model class. The skew comes from the correlation between the forward and its volatility, and over a short horizon the volatility simply has not had time to move far enough for that correlation to produce more than a bounded tilt.

The market disagrees, and not subtly. Observed at-the-money skews grow as the expiry shrinks, roughly as a power law, and short-dated options are far more skewed than any classical model will produce.

The fix that emerged in the last decade is to make the volatility *rough*.

Definition 10.11 (Rough volatility, informally)**.** Let the log variance be driven by a fractional Brownian motion with Hurst parameter H,

$$\frac{dS_t}{S_t} = \sqrt{v_t}\, dW_t, \qquad \log v_t \text{ driven by } B^H,$$

with H small — estimates from realised volatility put it near 0.1, against the $H = 1/2$ of an ordinary diffusion.

A smaller H means a rougher path: one that reverses direction more violently and has less memory of its recent level. And the payoff is exactly the quantity that was wrong. A model of this kind produces an at-the-money skew behaving like

$$s(T) \;\sim\; T^{H-1/2},$$

which for $H = 0.1$ is $T^{-0.4}$ — a power law that blows up as the expiry shrinks, which is what the market shows. Setting $H = 1/2$ returns the constant of the classical case, so the classical models are the corner of this family in which the volatility is exactly as smooth as the price.

Remark (Why this does not break anything)**.** Chapter 2 showed that fractional Brownian motion is not a semimartingale for $H \neq 1/2$, and that a *price* following one would admit arbitrage. To be clear, no such thing is happening here.

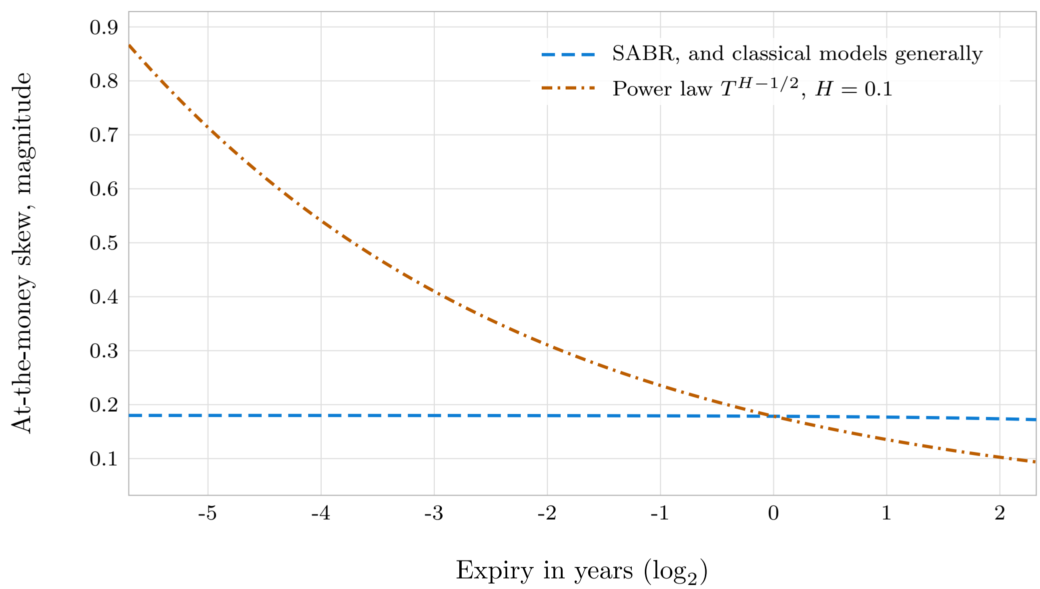

Figure 10.3: The at-the-money skew against expiry, from a week to five years. The SABR curve is computed; the power law beside it is the shape rough volatility predicts, drawn as a reference and anchored to agree with SABR at one year, so the comparison is about the term structure rather than the level. SABR moves from 0.180 to 0.172 across the whole range — it is, to the eye, a flat line. The power law spans nearly a factor of ten, and at a one week expiry it is five times as steep. No amount of recalibrating the classical model closes that gap, because the gap is in the shape.

Drawn by `Sabr::skew`.

The roughness is in v, and v is not a traded asset. The price is still

$$S_t = S_0 + \int_0^t \sqrt{v_s}\, S_s\, dW_s,$$

an ordinary continuous semimartingale, and chapter 2's representation theorem never asked the volatility to be anything but adapted with $\int \sigma^2 ds$ finite. Every result in these notes therefore survives unchanged. Rough volatility is not an alternative to the framework; it is a choice of σ within it.

Remark (What it costs)**.** Fractional Brownian motion has memory, so the model is not Markov in any finite set of state variables. Chapter 12 will meet exactly this difficulty in a different guise — the Heath-Jarrow-Morton framework is also non-Markov, for a different reason — and the consequence is the same in both cases: no low-dimensional grid, no backward induction, and therefore real difficulty with anything callable.

The escape is also the same in both cases. Chapter 12 recovers a Markov state by insisting the volatility kernel decay exponentially, since $e^{-\kappa(T-u)}$ is the one shape that factorises into a maturity piece and a history piece. Rough volatility recovers one by approximating its fractional kernel with a *sum* of exponentials, each of which is Markov by that same argument, giving a model with several ordinary factors that behaves like a rough one over the horizons that matter. The trick that makes rates models computable is the trick that makes rough volatility models computable.

There is a further parallel. Chapter 12's excursion into affine models explains why certain models reduce a partial differential equation to ordinary ones. The rough Heston model retains an affine structure of exactly that kind, except that its Riccati equation is fractional, so the same machinery applies with the ordinary derivative replaced by a fractional one.

References

- Hagan, P. S., Kumar, D., Lesniewski, A. S., & Woodward, D. E. (2002). Managing smile risk. *Wilmott Magazine*, 1, 84–108.

- Berestycki, H., Busca, J., & Florent, I. (2002). Asymptotics and calibration of local volatility models. *Quantitative Finance*, 2(1), 61–69.

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

- Hagan, P. S., Kumar, D., Lesniewski, A., & Woodward, D. (2014). Arbitrage-free SABR. *Wilmott*, 2014(69), 60–75.

- Gatheral, J., Jaisson, T., & Rosenbaum, M. (2018). Volatility is rough. *Quantitative Finance*, 18(6), 933–949.

- Bayer, C., Friz, P., & Gatheral, J. (2016). Pricing under rough volatility. *Quantitative Finance*, 16(6), 887–904.

- El Euch, O., & Rosenbaum, M. (2019). The characteristic function of rough Heston models. *Mathematical Finance*, 29(1), 3–38.

- Abi Jaber, E., & El Euch, O. (2019). Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 10(2), 309–349.