Chapter 11

Local-Stochastic Volatility

In these notes we combine the two models of the previous chapters. Local volatility fits every European option exactly but moves the smile wrongly; stochastic volatility moves it sensibly but does not fit. We derive the condition under which a model can do both at once, find that it defines the model implicitly rather than explicitly, and describe how that circularity is broken in practice.

11.1 Taking the Good Half of Each

Two chapters have left us with two models and a complaint about each.

Local volatility fits the market's European options perfectly, because Dupire's formula constructs it to. Its dynamics are wrong: the backbone comes out at twice the skew, which is outside the range the market actually behaves in, so the delta is systematically wrong.

Stochastic volatility has sensible dynamics, because the backbone and the skew are controlled by different parameters and can be set independently. It does not fit: SABR with four parameters cannot pass exactly through a few dozen quoted strikes and expiries, and Heston with five does no better. For a desk that has to reprice the market to the penny — because it is marking a book against those very quotes — a fit that is close is not a fit.

The idea is to keep the fit of the first and the dynamics of the second by multiplying them together.

Definition 11.1 (Local-stochastic volatility)**.** A local-stochastic volatility model is

$$\begin{aligned} \frac{dS_t}{S_t} &= (r-q)\,dt + L(t,S_t)\,\sqrt{v_t}\,dW_t, \\ dv_t &= \alpha(t,v_t)\,dt + \beta(t,v_t)\,dZ_t, \qquad dW_t\,dZ_t = \rho\,dt, \end{aligned}$$

where v is whatever stochastic variance process one likes — Heston's, say — and L, a deterministic function of time and level, is called the *leverage function*.

The instantaneous volatility is now a product of two things. The factor $\sqrt{v_t}$ is random and carries the dynamics: the vol-of-vol, the correlation, the mean reversion, everything that decides

how the smile moves and what a forward smile looks like. The factor $L(t, S_t)$ is deterministic and carries the fit: it is a dial at each point of the plane, to be turned until the model's European prices agree with the market's.

The division of labour is the whole design. The question is whether it works — whether there really is a choice of L making the fit exact, whatever the stochastic part is doing. There is, and chapter 9 has already given us the tool to find it.

11.2 The Calibration Condition

Theorem 11.2 (The leverage function)**.** *A local-stochastic volatility model reproduces the market's European option prices at all strikes and expiries if and only if*

$$L^2(t,K)\,\mathbb{E}\big[\,v_t \,\big|\, S_t = K\,\big] \;=\; \sigma_{loc}^2(t,K), \tag{11.1}$$

where σ_{loc} is the Dupire local volatility of the market's own surface. Equivalently,

$$L(t,K) = \frac{\sigma_{loc}(t,K)}{\sqrt{\mathbb{E}\big[\,v_t \,\big|\, S_t = K\,\big]}}. \tag{11.2}$$

Proof. Chapter 9 established two facts and this is their intersection.

Gyongi's theorem says that a process with random volatility has the same one-dimensional distributions — and therefore the same European option prices — as the local volatility model whose squared volatility is the conditional expectation of the true squared volatility. Here the instantaneous volatility of S is $L(t,S_t)\sqrt{v_t}$, so the mimicking local volatility is

$$\tilde\sigma^2(t,K) = \mathbb{E}\Big[L^2(t,S_t)\,v_t \,\Big|\, S_t = K\Big] = L^2(t,K)\,\mathbb{E}\big[v_t \,\big|\, S_t = K\big],$$

where $L^2(t,S_t)$ comes out of the conditional expectation because, conditional on $S_t = K$, it *is* the number $L^2(t,K)$ — it is a deterministic function of the thing being conditioned on.

Dupire's formula says that a set of European prices determines exactly one local volatility surface, namely σ_{loc}. So the model matches the market if and only if its mimicking local volatility is the market's, which is (11.1). □

Dupire's local volatility is not the model's volatility. It is an average of the model's volatility, taken over all the ways the market could have arrived at the level K at time t. The leverage function is therefore a *correction factor*: at each point, take the local volatility the market demands, and divide by the volatility the stochastic part is already delivering there on average. Whatever is left over is what L must supply.

Example 11.1 (The two extremes)**.** Two sanity checks, both of which should be reassuring.

If the variance is not random at all — say $v_t \equiv 1$ — then $\mathbb{E}[v_t \mid S_t = K] = 1$ and (11.2) gives $L(t,K) = \sigma_{\text{loc}}(t,K)$. The leverage function is the local volatility surface, and the model is the local volatility model of chapter 9. Nothing has been gained and nothing lost.

If instead the stochastic volatility model already fits the market on its own, then by the same theorem its own mimicking local volatility is already σ_{loc}, so $\mathbb{E}[v_t \mid S_t = K] = \sigma^2_{\text{loc}}(t,K)$ and $L \equiv 1$. The leverage function switches itself off when it is not needed.

Between the two, L measures exactly the part of the market's smile that the stochastic volatility model failed to produce by itself. A well-chosen stochastic part leaves L close to one and smooth; a badly chosen one leaves L doing all the work, which puts us back in chapter 9 with extra steps.

The one thing in (11.2) that is easy to read past is the word *conditional*. It would be much more convenient if the denominator were $\mathbb{E}[v_t]$, the plain average variance, which one could compute once and be done with. It is not, and the difference is not small. The following example is small enough to compute by hand and shows why.

Example 11.2 (Where the conditioning bites)**.** Take the crudest possible stochastic volatility: at time zero a coin is tossed, and the forward is lognormal for the rest of its life with volatility either 15% or 35%, each with probability one half. So v is random but constant along each path, and $\mathbb{E}[v] = \frac{1}{2}(0.15^2 + 0.35^2) = 0.0725$, whose square root is 26.9%.

Now ask for $\mathbb{E}[v \mid S_T = K]$ at $T = 1$, with the forward at 100. Observing where the underlying ended is evidence about which coin came up, and Bayes' rule turns that evidence into a posterior. Writing ϕ_{lo} and ϕ_{hi} for the two lognormal densities,

$$\mathbb{E}\big[v \mid S_T = K\big] = \frac{v_{\text{lo}}\,\phi_{\text{lo}}(K) + v_{\text{hi}}\,\phi_{\text{hi}}(K)}{\phi_{\text{lo}}(K) + \phi_{\text{hi}}(K)}.$$

Evaluating:

K	60	80	100	125	160
$\mathbb{P}(\text{high} \mid S_T = K)$	0.98	0.51	0.30	0.51	0.96
$\sqrt{\mathbb{E}[v \mid S_T = K]}$	34.7%	27.1%	22.9%	27.1%	34.4%

Read the middle row first. If the forward has barely moved, it was probably a quiet path, so the low volatility state is more likely than it was a priori. If the forward has moved a long way, it was almost certainly the noisy one.

The bottom row is what (11.2) divides by, and it ranges from 22.9% to 34.7% — while the unconditional figure is a flat 26.9% everywhere. Using the unconditional average would leave the leverage function wrong by four volatility points at the money and eight in the wing, in opposite directions. The model would then miss the market by roughly that much, which is a hundred times any bid-offer spread.

Notice also what the conditional expectation is shaped like: a hump, low in the middle and rising in both wings. So L, which divides by its square root, is high in the middle and low in the wings. That is the general shape, and it has a plain reading — the stochastic part already supplies plenty of volatility in the tails, because that is what a random volatility does, so the leverage function has less work to do there and more in the middle.

Remark (Why the mixing matters)**.** It is tempting to conclude that since L can always be found, the choice of the stochastic part does not matter. The opposite is true, and the reason is the point of the whole construction.

Everything about the model's *dynamics* — the backbone, the forward smile, the vega — comes from the stochastic part, because L is deterministic and contributes no randomness at all. And the fit is exact regardless. So the stochastic part is now a free choice that is invisible in the calibration and decisive for everything else, which is chapter 10's lesson repeating one level up. Desks parametrise it explicitly with a *mixing weight*: dial it to zero and the model degenerates to pure local volatility, dial it up and the vol-of-vol takes over, and the vanilla prices do not move either way. What moves is every exotic in the book.

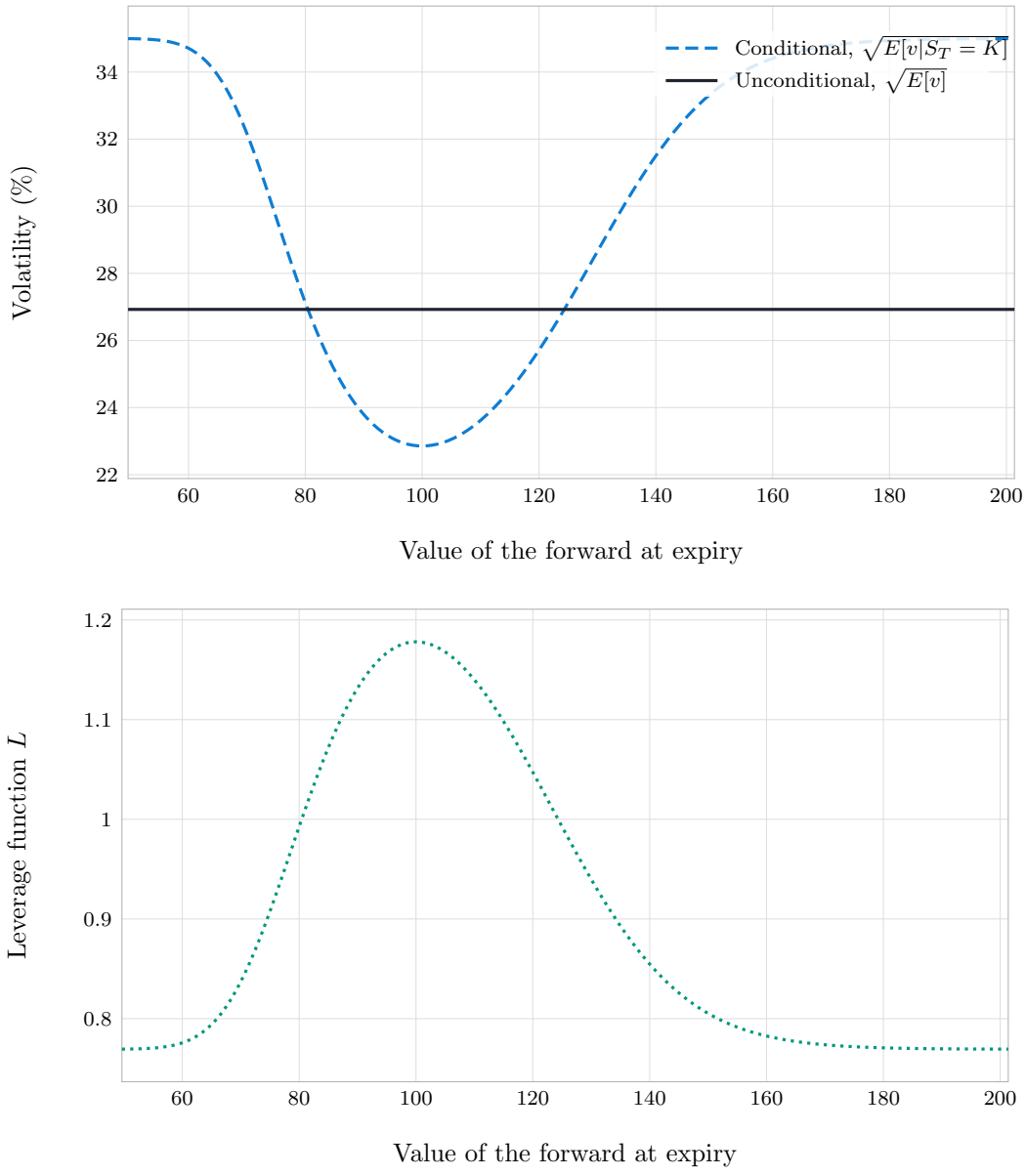

Figure 11.1: Above: the conditional volatility $\sqrt{\mathbb{E}[v \mid S_T = K]}$ against the unconditional $\sqrt{\mathbb{E}[v]}$ that it is tempting to use instead. Below: the leverage function this forces, against a flat target local volatility. The two differ by four volatility points at the money and eight in the wing, in opposite directions, which is the difference between a model that reprices the market and one that misses it by a hundred bid-offer spreads.

Drawn by `TwoRegime::conditional_variance` and `TwoRegime::leverage`.

11.3 The Circularity, and How It Is Broken

Equation (11.2) looks like a formula for L. It is not, quite, and the reason deserves a clear look.

The right-hand side contains $\mathbb{E}[v_t \mid S_t = K]$, a conditional expectation taken under the law of the process (S, v). But the law of S depends on L, because L is in its diffusion coefficient. So L appears on both sides: it is defined in terms of a distribution that it itself determines.

Remark (What kind of equation this is)**.** An equation whose coefficients depend on the law of its own solution is called a McKean-Vlasov, or mean-field, stochastic differential equation. It is a genuinely different object from an ordinary SDE. The usual existence and uniqueness theorems do not apply, and for local-stochastic volatility in particular the question of whether a solution exists for a given market surface is, at the time of writing, not settled in general. Models are calibrated and traded anyway.

There are two standard ways to break the circle, and both work by stepping forward in time, so that at each step the conditional expectation is computed from a distribution that is already known.

The forward equation approach

Write $\psi(t, S, v)$ for the joint density of (S_t, v_t). It satisfies a two-dimensional Fokker-Planck equation whose coefficients involve L, and the conditional expectation we need is a ratio of integrals of it:

$$\mathbb{E}\big[v_t \mid S_t = K\big] = \frac{\int v\, \psi(t, K, v)\, dv}{\int \psi(t, K, v)\, dv}.$$

So march the density forward on a grid. At each time step, the density is known from the previous step; compute the conditional expectation from it by the ratio above; set L on that time slice by (11.2); use that L to take the density one step further. The circularity is broken because $L(t, \cdot)$ is only ever needed at a time when $\psi(t, \cdot, \cdot)$ has already been computed.

This is accurate and it is fast in two dimensions. It does not survive a large number of factors, because the grid does not.

The particle method

The alternative, due to Guyon and Henry-Labordere, replaces the density with a cloud of simulated paths.

1. Start N particles at (S_0, v_0). Take N large — tens of thousands.
2. At each time step t, having the particles' current positions (S_t^i, v_t^i):
   (a) estimate the conditional expectation from the cloud itself, by a kernel-weighted average over particles whose S is near K,

$$\widehat{\mathbb{E}}\big[v_t \mid S_t = K\big] = \frac{\sum_{i=1}^{N} \delta_h(S_t^i - K)\, v_t^i}{\sum_{i=1}^{N} \delta_h(S_t^i - K)},$$

   where δ_h is a smoothing kernel of bandwidth h;

(b) set $L(t,\cdot)$ on this time slice by (11.2);

(c) evolve every particle one step using that L.

3. The leverage function is the collection of the slices.

The usual first choice for the kernel is Gaussian,

$$\delta_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right), \tag{11.3}$$

so a particle sitting exactly at K counts fully, one a bandwidth away counts about six tenths as much, and one three bandwidths away counts for nothing that matters. Compactly supported kernels — Epanechnikov, $\delta_h(x) \propto (1 - x^2/h^2)^+$, or the quartic — are common in production for the plain reason that they let the sum skip every particle outside the window. The estimator is Nadaraya-Watson kernel regression, the standard nonparametric way of reading a conditional expectation off a scatter of points; nothing about it is specific to finance.

The bandwidth cannot be a constant, because the cloud spreads as it goes. A common choice scales as

$$h_t \;\propto\; \sigma S_0 \sqrt{t}\; N^{-1/5},$$

the $\sqrt{t}$ tracking the width of the cloud and the $N^{-1/5}$ being the rate at which a kernel bandwidth should shrink as the sample grows — slowly, because a bandwidth that shrinks too fast leaves too few neighbours to average over.

Structure (A grid moves the values, a cloud moves the nodes)**.** The two methods are not different in *what* they compute. Both need the law of (S_t, v_t) at each date, and both march it forward. They differ only in how they hold it, and chapter 3 already supplied both ways.

The forward equation approach carries the density itself and pushes it with the adjoint $\mathcal{L}^*$, at nodes the modeller fixes in advance. The particle method carries a *sample* and pushes each member with the process whose generator is $\mathcal{L}$; the empirical measure

$$\hat{p}_t = \frac{1}{N} \sum_{i=1}^{N} \delta_{(S_t^i,\, v_t^i)}$$

converges to the same density the grid was computing. So the particles are not an approximation to something other than the density. They *are* the density, sampled instead of tabulated: a grid fixes where the nodes are and lets the values on them move, while a cloud fixes the values — every particle is worth $1/N$, always — and lets the nodes move.

That is the whole of the difference, and it is what decides which one survives a third factor. A grid has to cover wherever the process might go, so its cost grows exponentially in the number of factors, and most of that cost is spent on regions the process visits rarely. A cloud goes where the process actually goes, so it never spends anything on a region the process ignores, and its error stays $O(N^{-1/2})$ whatever the dimension. Trading accuracy per node for independence from dimension is the bargain every Monte Carlo method strikes; here it is the difference between a model with two factors and a model with more.

Notice that the particles interact. Each one's next step depends on where all the others currently are, through the conditional expectation — which is exactly the mean-field structure of the equation being solved, appearing in the numerical scheme as a set of particles that cannot be simulated independently. That is the price of the method and also the reason it is faithful to the problem.

Two practical points decide whether it works. The bandwidth h trades bias against noise in the usual way, and it must widen in the tails where particles are sparse. And in regions the process rarely visits, the denominator of (11.2) is estimated from a handful of particles, so there is noise; production implementations cap and floor it, and accept that the model's wings are the least trustworthy part of it — which is unfortunate, since the wings are what the exotics are usually sensitive to.

11.4 The Mixing Weight

Since the fit is exact whatever the stochastic part does, desks make the freedom explicit and give it a dial.

$$\begin{aligned} \frac{dS_t}{S_t} &= (r-q)\,dt + L_\lambda(t,S_t)\,\sqrt{v_t^\lambda}\,dW_t, \\ dv_t^\lambda &= \alpha(t,v_t^\lambda)\,dt + \lambda\,\beta(t,v_t^\lambda)\,dZ_t. \end{aligned} \tag{11.4}$$

Two things about (11.4). The parameter does not appear in the equation for S at all: it scales β, the diffusion of the *variance*, and nothing else. And L_λ carries a subscript because it is not chosen — it is whatever Theorem 11.2 returns once λ is fixed,

$$L_\lambda(t,K) = \frac{\sigma_{\text{loc}}(t,K)}{\sqrt{\mathbb{E}\big[\,v_t^\lambda \,\big|\, S_t = K\,\big]}}. \tag{11.5}$$

So λ and L are not two knobs. Turning the first determines the second, and the reason the fit survives is that the calibration condition rearranges to

$$L_\lambda^2(t,K)\,\mathbb{E}\big[\,v_t^\lambda \,\big|\, S_t = K\,\big] \;=\; \sigma_{\text{loc}}^2(t,K) \qquad \text{for every } \lambda, \tag{11.6}$$

whose right-hand side is the market's and contains no λ. That is the precise sense in which the weight tunes a balance: the product is pinned, and λ decides how it is divided between a factor that is random and a factor that is not.[1]

The two ends are then immediate. At $\lambda = 0$ the variance is deterministic, so the conditioning in (11.5) does nothing, $\mathbb{E}[v_t^0 \mid S_t = K] = v_t^0$, and $L_0^2 = \sigma_{\text{loc}}^2/v^0$ leaves the price equation reading $dS/S = \sigma_{\text{loc}}(t,S_t)\,dW_t$ exactly. At $\lambda = 1$ the stochastic part supplies as much of the smile as it can and L_1 is left correcting the remainder.

Calculation 11.3 (What the dial actually moves)**.** In the coin-toss model of *Where the conditioning bites* above, the vol-of-vol is the gap between the two volatilities, so λ closes that gap towards their common mean while holding the mean variance fixed — which isolates the randomness of the variance from its level, so that what moves is the split rather than the total.

[1] `checked at every setting`.

Take a flat local volatility surface at 25%, so that any structure in L is the stochastic part's doing and not the market's. At $\lambda = 0$ the leverage function is flat at 0.93: with a deterministic variance there is nothing strike-dependent for it to do. At $\lambda = 1$ it runs from 1.09 at the money down to 0.71 in both wings.[2]

So the dial does not scale L; it changes its shape, from a constant to a hump with a range half its own height. That shape is the smile the stochastic part now generates by itself, being taken back out again — which is the division of labour of this chapter, made visible as a single function.

Recalibrating at each setting always succeeds, by Theorem 11.2.

- At $\lambda = 0$ the variance is deterministic, the conditional expectation in (11.2) is just its value, and L absorbs the whole surface. The model *is* local volatility: exact fit, backbone twice the skew, no vega.

- At $\lambda = 1$ the stochastic part produces most of the smile on its own and L is a modest correction near one. The dynamics are the stochastic model's.

- In between, the two are blended in a proportion the desk chooses.

Every setting of λ prices every vanilla option identically — to the penny, by construction. And the prices of the exotics move steadily across the range, often by several percent of premium between the ends. So λ is a parameter that changes the answer, does not change the calibration, and therefore cannot be determined by it.

This is the third time in three chapters that we have met the same shape: β in chapter 10, the choice of stochastic part here, and λ now. The right response is not to search harder for a way to pin it down from vanillas — Gyongi's theorem says there is none — but to treat it as what it is. Desks mark λ by policy, revalue the book at both ends of the range, and carry the difference as a reserve. A model risk that has been measured and provisioned for is a different thing from one that has been assumed away.

Remark (The forward smile, concretely)**.** The clearest place the difference shows is the forward smile.

Consider an option that will be struck at the money in one year and expire a year later — a forward starting option, the building block of a cliquet. Its value depends entirely on what the smile looks like in a year.

Under local volatility, conditional on the underlying reaching some level in a year, all remaining randomness is Brownian, and the smile from that point on is close to flat. Under a model with genuine vol-of-vol, the variance in a year is itself uncertain, and that uncertainty produces a smile then just as it does now. The prices differ by a lot, and the observable forward smiles in liquid markets are much closer to the second. Since L hides the difference in the vanillas, the forward starting option is one of the few instruments that could tell the two apart — which is exactly why it is used to calibrate λ where it trades.

[2] `measured`.

11.5 What Has Been Bought

Close the loop back to the complaint that started the chapter.

The model reproduces every European option price exactly, by construction, through L. It reproduces them for *any* choice of the stochastic part, so that choice is left entirely free.

That freedom is then spent on the things chapter 10 showed vanilla options cannot see. The correlation ρ and the vol-of-vol set the backbone, and therefore the delta. The mean reversion of the variance sets the maturity term structure — how fast the skew decays as the expiry lengthens — and, when today's variance sits away from its long-run level, how quickly the forward smile settles to its stationary shape. Both bear on the price of anything forward starting, though chapter 10's measurement is the thing to keep in mind here: in a model with no calendar in it the forward smile does not flatten as the start date recedes, it settles. And because volatility is now a genuine risk factor, the model is incomplete in the sense of chapter 4 — there is a vega, it is real, and it can be hedged with other options rather than being an artefact.

This is the standard equity and foreign exchange exotics model, and the pattern is more general than the setting. When a model must match a set of prices and also behave sensibly, the productive move is often to split it into a part that is forced by the fit and a part that is free, and then to be explicit that the free part is a choice that the calibration did not make for you.

Chapter 12 applies the same idea to interest rates, where the state is a curve rather than a number, and where the analogue of the leverage function is a local volatility function of the model's own state variables.

References

- Guyon, J., & Henry-Labordere, P. (2012). Being particular about calibration. *Risk*, 25(1), 88–93.

- Ren, Y., Madan, D., & Qian, M. Q. (2007). Calibrating and pricing with embedded local volatility models. *Risk*, 20(9), 138–143.

- Lipton, A. (2002). The vol smile problem. *Risk*, 15(2), 61–65.