Chapter 21

Estimation in Practice

Chapter 20 asked whether a fit means anything. This chapter asks the same question where the parameter is never observed at all — a variance, a regime, a factor driving a curve — and where the last decade of machine learning has made the boldest claims. Filtering recovers a latent state in principle and a network makes the recovery fast; neither changes what the data can identify, and telling the two apart is most of what follows. We work through what is genuinely hidden and what only looks that way, how long a regime takes to notice before it has already changed again, whether prices and a history can be fitted at once, the two parameters estimated constantly and understood least — a volatility and a correlation — and the oldest asymmetry in the subject: why implied volatility sits above realised, and what it costs to collect the difference.

21.1 What the Newer Methods Change

Everything in chapter 20 was classical. What the last decade of machine learning does and does not alter is worth setting out plainly, because the claims made for it are uneven.

21.1.1 Latent states: filtering, and the likelihood it hands over

Chapter 20 assumed throughout that the thing being estimated is observed. In fixed income it never is. Nobody sees the short rate, the variance, or the factors driving a curve; what is seen is a set of yields or option prices, each a known function of the state plus an error. That is a state-space model, and it changes what estimation means: there are now two problems, recovering the state and estimating the parameters, and the second cannot be attempted until the first is solved.

Definition 21.1 (State-space form)**.**

$$\begin{aligned} x_{t+1} &= f(x_t) + \text{noise} && \text{the state, unobserved} \\ y_t &= h(x_t) + \text{error} && \text{what is quoted} \end{aligned}$$

A *filter* computes $p(x_t \mid y_{1:t})$, the state given everything seen so far.

Remark (When it is exact: the Kalman filter)**.** If f and h are affine and both noises Gaussian, the filtering distribution is Gaussian and the recursion is exact and cheap.

Linearity is what makes the recursion *closed*: an affine map of a Gaussian is Gaussian, so a distribution described by a mean and a variance stays described by a mean and a variance, and the filter never has to represent anything else.

Gaussianity is what makes it *optimal among all estimators*. Drop it but keep linearity and the same recursion is still the minimum-variance estimator among linear ones, for any noise with finite variance — so the filter is not useless off its assumptions, it is merely no longer using everything in the data. What does break is the likelihood: the number the recursion emits is then a Gaussian quasi-likelihood rather than the true one, which is still usable for estimation but gives standard errors that are wrong unless corrected.

In the affine Gaussian case both hold. That case is not a toy here — it is precisely a Gaussian affine term structure model, where yields are affine in the state by construction, so chapter 8's models are estimated this way as a matter of course.

The recursion is two steps. *Predict* moves the state distribution forward by the dynamics, which widens it. *Update* conditions on the new observations, which narrows it, by a precision-weighted average of the prediction and the data — Bayes' rule for Gaussians and nothing more.

The by-product is what makes it valuable. Each step produces a predictive distribution for the next observation, and summing its log density over the history gives $p(y_{1:T} \mid \theta)$ *exactly*. So the filter does not merely estimate the state, it delivers the likelihood of the parameters — which is what makes maximum likelihood, or the posterior of the previous subsection, possible at all for a model whose state is never seen.[1]

Structure (The measurement error is not a nuisance term)**.** Five yields and a one-factor model is an over-determined system.

The difficulty is that the curve is not free. The loadings $b_i = (1 - e^{-\kappa\tau_i})/(\kappa\tau_i)$ are fixed by the parameters, and the parameters are held still while the filter runs; the only thing that moves from one date to the next is the scalar state. So the observation is not five points to be joined but a five-vector,

$$y_t \;=\; a + x_t\, b, \qquad b \in \mathbb{R}^5 \text{ fixed},$$

required to lie on a *line* in $\mathbb{R}^5$ — one dimension inside five. Five numbers, one unknown, and generically no x_t reproduces them. The measurement error is what makes the problem well posed.

That also settles an apparent contradiction with chapter 8, which fits today's curve exactly with a one-factor model. It does, and it buys the fit with a time-dependent drift: a free function, in effect one number per maturity, spent entirely on the first date. It does nothing on any date after, where the same fixed loadings apply and only x has moved. A one-factor model can match any single curve and cannot match a panel of them, and it is the panel the filter sees.

Two consequences. Assuming the yields are cleaner than they are makes the state estimate *worse*, because the filter then trusts each observation more than it deserves and chases the noise — which is measurable and is measured. And the fitted size of the error is a diagnostic worth reading: a model whose estimated measurement error comes out far above the bid-offer is telling you it cannot fit the cross-section, whatever its time series behaviour looks like. That is a specification test that costs nothing, since the number is estimated anyway.

[1] `estimation::AffineStateSpace`, with `the likelihood peaking at the true mean reversion` and the filtered state beating the best single yield by a third.

Remark (When it is not exact)**.** Affine and Gaussian is a strong assumption and the alternatives are ranked by how much of it they give up.

Extended and unscented filters keep the Gaussian approximation and handle a nonlinear h — the first by linearising, the second by propagating a small set of points through the nonlinearity and refitting a Gaussian. Both are cheap and both fail in the same circumstance, which is a nonlinearity strong enough that the filtered distribution is not Gaussian in shape. A square-root variance near zero is exactly that circumstance, so the models where one most wants a filter are the ones where these approximations are least safe.

Particle filters give up the Gaussian entirely: represent the distribution by a weighted sample, push each particle through the dynamics, reweight by how well it explains the new observation, and resample. Nothing is assumed about shape, which is why this is the method for a stochastic volatility model, where the state is a variance that is positive and skewed.

The cost is the usual one and it has a specific form. Weights degenerate — after a few steps one particle carries almost all of them — so resampling is required, and resampling makes the likelihood estimate noisy and discontinuous in the parameters, which breaks any optimiser that wants a gradient. The repair is to embed the particle filter's unbiased likelihood estimate inside a Markov chain Monte Carlo scheme rather than to optimise it, which is exact despite the noise and is why particle methods and the Bayesian treatment above are usually met together.

21.1.2 Particle filters, in more detail

Stochastic volatility has a problem the sections above sidestepped: v_t is not observed. Prices are, volatility is not, so the likelihood of a Heston model given a price history involves integrating over every possible volatility path — an integral of enormous dimension, which is why stochastic volatility models are so often calibrated to option prices instead of estimated from returns.

Sequential Monte Carlo solves it. Carry a population of particles, each a hypothesis about the current v_t; propagate them through the model's dynamics; reweight by how well each explains the newly observed return; resample. What comes out is the filtered distribution of the latent state, and — the part that matters here — the likelihood of the data as a by-product, which makes maximum likelihood estimation possible for models where it previously was not.

Two things about it matter before using it, because both are places it fails quietly.

Weights degenerate. After a few steps one particle carries almost all the weight and the rest carry none, so the effective sample size collapses and the filter is representing a distribution by one point. Resampling is the standard repair — discard the low-weight particles and duplicate the high-weight ones — and it trades the degeneracy for impoverishment, since the surviving particles share ancestors and the population loses diversity in the distant past. Neither problem is visible in the output unless the effective sample size is monitored, and it should be.

The proposal matters more than the particle count. The naive scheme propagates particles through the model's own dynamics and then reweights, which is wasteful when the observation is much sharper than the dynamics: nearly every particle lands somewhere the data rules out, and the few that survive determine the answer. Proposing from a distribution that already accounts for the new observation — the auxiliary and guided variants — fixes this, and the improvement is far larger than the one bought by multiplying the number of particles.

This is the principled answer to a real problem, it is genuinely used, and it has no fashionable name.

21.1.3 Reporting the valley: Bayesian calibration

Chapter 20 ended with a complaint: a calibration reports $\rho = -0.30$ when the evidence supports anything from -0.28 to -1. The Bayesian answer is direct — do not report a point. Put a prior on the parameters, condition on the quotes with a likelihood reflecting the bid-offer, and report the posterior.

The output is then the flat region itself rather than an arbitrary point inside it, and every downstream number inherits the uncertainty honestly. A risk figure computed across the posterior is a risk figure that knows the model was not fully determined, which is exactly the information a point estimate throws away.

The cost is computational and it is real, since the posterior must be sampled rather than optimised. But note that nothing has been added to the model. The information was always this thin; the Bayesian version merely declines to hide it.

Structure (A posterior over parameters is a posterior over prices)**.** The consequence is larger than better reporting of a calibration.

A price is a function of the parameters, $V = V(\theta)$. A posterior $p(\theta \mid \text{quotes})$ therefore pushes forward to a distribution of prices, obtained by pricing under each draw. That distribution is the object several later chapters have been approximating by hand.

It is the model reserve. Chapter 24 computes model risk by fixing a plausible range for an undetermined parameter, revaluing at each end, and taking the difference. That is a two point approximation to a push-forward, and it is crude in a specific way: it ignores the shape of the valley, ignores that two parameters may be jointly undetermined along a diagonal while each looks fine alone, and produces a number whose size depends on how wide somebody drew the range. A posterior quantile is the same quantity computed rather than sketched.

And it nets correctly across a book, which the crude version does not. Chapter 24 observes that model risk is not additive — two trades whose exposure to a parameter offsets have no joint model risk — and then has no machinery for computing the netted number. The push-forward has it for free: one draw of θ prices every trade in the book, so the distribution of the *portfolio* value is obtained by the same loop, and offsetting exposures cancel inside each draw. That is the correct portfolio-level reserve, and it comes out of the same computation as the trade-level one.

It gives relative value a decision rule. Chapter 22 insists that a model saying a spread is wide is not a signal. A posterior says how wide the spread would have to be to mean anything: if the observed spread lies inside the band the model's own uncertainty generates, the model has not identified a mispricing, it has failed to determine a price. Only a spread outside the band is a candidate — and even then the structural reason chapter 22 insists on is still required. This is a sharper filter than the usual one and it is available at no extra modelling cost.

Remark (What the band is really a statement about)**.** Along a direction the quotes constrain, the posterior is narrow and is driven by the data. Along a flat direction — and chapter 20's whole point is that these exist — the posterior is the *prior*, because conditioning on uninformative data returns what was assumed. So the width of the band in those directions measures the prior rather than the market.

That is not a defect, provided it is said. It converts an assumption that was previously invisible into a number somebody has to write down and defend, which is precisely what chapter 24 says the reserve process is for. A band that is wide because the prior was wide is an honest statement

that nobody knows; a point estimate in the same situation is a dishonest statement that somebody does.

The practical obstacle is cost, since sampling a posterior requires thousands of pricings of a model that may take minutes each. Which is where the next subsection's neural approximation earns its place: a fast surrogate for the pricing map does not resolve the identification problem, but it makes the calculation that *reports* the problem affordable. Speed does not buy knowledge; it buys the ability to say how little one has.

21.1.4 Speed, not knowledge: neural approximation

The honest and successful use of neural networks in derivatives is unglamorous: learn the map from parameters to prices offline, once, and then evaluate it in microseconds.

That pays off because the map is expensive for models like rough volatility where each price is a simulation, and calibration needs thousands of evaluations. Learning it turns an overnight calibration into an interactive one. The model is unchanged, the prices are unchanged, and the network is a lookup table with good interpolation — which is why it works and why it is safe: the approximation error can be measured against the true pricer on as many test points as one likes.

Remark (The limit, which is the point of this chapter)**.** A neural network approximating the pricing map does not, and cannot, resolve chapter 20's problem. The flat valley is a property of the map itself — of the fact that quotes near the money contain almost nothing about ρ. Learning that map faster reproduces the flatness faithfully, at speed. No architecture recovers a parameter the data does not contain.

Machine learning changes what is computationally feasible. It does not change what is identifiable, and identifiability is a property of the instruments quoted.

21.1.5 Learned models, and the two questions they do not answer

The more ambitious programme replaces the model rather than approximating it: neural stochastic differential equations, market generators, deep hedging, and network solutions of high-dimensional pricing equations. Those belong to chapter 19, since they are about computing a price rather than fitting one.

Two cautions belong here rather than there, because they are about fitting. The first needs stating carefully, because the usual version of it is wrong.

Whether a learned model is arbitrage-free depends entirely on what is learned. A network fitted to the price surface has no structural guarantee: chapter 4 established that a price system without a martingale measure admits a money pump, and nothing in a training loss enforces one. Constraints can be added as penalties, but a penalty is a preference and not a guarantee.

A network placed inside the coefficients of a stochastic differential equation is in a completely different position. Chapter 2 established that a continuous arbitrage-free price has no choice but to be an Itô diffusion, and chapter 4 that its discounted value must be a local martingale under the pricing measure — which fixes the drift and leaves only the diffusion coefficient free. So writing

$$\frac{dS_t}{S_t} = r\,dt + \sigma_\theta(t, S_t, v_t)\,dW_t \tag{21.1}$$

with σ_θ a neural network produces an arbitrage-free model *for every value of* θ. The constraint lives in the form of the equation rather than in the loss, so training cannot violate it, and no penalty

term is required. Nothing has been given up in flexibility either: by chapter 2 the equation is not a restriction on the model class, it is the model class.

That is the correct statement of what the SDE parametrisation buys, and it is a structural argument rather than an empirical one. The remaining requirements are mild — σ_θ positive and regular enough for (21.1) to have a solution — and the remaining problem is the one this chapter is about, which no amount of structure removes: a finite set of quotes does not determine a function.

And chapter 9's warning transfers exactly. A generative model that reproduces the historical distribution of surfaces perfectly is in the position local volatility was in: excellent fit, unconstrained dynamics, and no guarantee that the hedge ratios it implies are the ones that work. Fitting more flexibly does not turn an in-sample fit into evidence.

21.2 What Is Actually Latent

The previous section filtered a state on the grounds that it is not observed, and how much of that is true deserves asking.

Remark (Adapted to what?)**.** "Adapted" is not a property of a process but a relation between a process and a filtration, and three of them are in play whenever a latent state is posited: $\mathcal{G}_t$, generated by what is observed; $\mathcal{F}_t$, the model's own, containing whatever latent factors it carries; and the one generated by the driving Brownian motions. A process can be $\mathcal{F}$-adapted and not $\mathcal{G}$-adapted, and that gap is what the word "latent" is pointing at.

Two separate things force adaptedness, and they are usually conflated. The Itô integral needs a predictable integrand to exist at all — chapter 1's construction fails outright if the integrand sees the increment it multiplies — and since σ is that integrand, it must be predictable in whatever filtration the integral is taken in. That is well-posedness, not economics. Separately, no-arbitrage requires the price to be a semimartingale in the *market's* filtration, which is chapter 4's use of the Bichteler-Dellacherie theorem, and requires trading strategies to be predictable.

Put those together and positing a latent state does not escape adapted volatility, it only relabels it. If S is a semimartingale in $\mathcal{G}$, decompose it there; the continuous martingale part is $\int \hat{\sigma}\, d\hat{W}$ with $\hat{\sigma}$ $\mathcal{G}$*-adapted*, by chapter 2's representation theorem. Whatever hidden machinery is imagined upstream, the observable description always has an adapted volatility. What changes is which one — $\sqrt{v}$ becomes something like $\sqrt{\mathbb{E}[v \mid \mathcal{G}]}$.

Structure (Incompleteness is not unobservability)**.** There is a sharper statement, and it cuts against the usual picture of stochastic volatility as a model with something hidden in it.

Quadratic variation is a limit of sums of squared increments of the *observed* price, so $\langle S \rangle$ is measurable with respect to the price's own filtration. Where it has a density, $v_t = d\langle S \rangle_t / (S_t^2 dt)$ is therefore a function of the path being watched. And once v is known, its own innovations $dZ = (dv - \kappa(\theta - v)dt)/(\eta\sqrt{v})$ are known too. So continuous observation of a single price reveals the entire state of a standard stochastic volatility model. Nothing in it is hidden.

What remains true is that the market is *incomplete*: one traded asset cannot span two Brownian motions, so Z-risk cannot be hedged with S alone. That is a statement about spanning, and it is routinely confused with a statement about observation.

Which relocates the question. Genuine latency does not come from the model. It comes from the observation being discrete, from the microstructure noise that Calculation 21.5 shows drives realised variance away from the truth rather than towards it, and from jumps confounding the estimate.

The filtering literature is about imperfect observation, not about a metaphysically hidden state — and that is a more tractable problem, because the imperfection can be measured.

Remark (Two projections, and only one of them is a local volatility model)**.** Chapter 9 projects a stochastic volatility model onto $\sigma^2_{\mathrm{loc}}(t,x) = \mathbb{E}[v_t \mid S_t = x]$, conditioning on the current level and nothing else, and matches one-dimensional marginals. The projection onto what an observer knows conditions on the whole observed path,

$$\hat{\sigma}^2_t = \mathbb{E}\big[v_t \,\big|\, \mathcal{G}_t\big],$$

and these are not the same object. The second is not a local volatility model at all: $\hat{\sigma}_t$ depends on the entire history, which makes it a *path-dependent volatility* model.

That distinction explains a fact both chapter 10 and chapter 18 arrive at from other directions. Matching marginals leaves the joint law free, so two models with the same σ_{loc} agree on every European price and disagree completely about a forward smile. The path projection keeps what the level projection throws away, which is exactly the dependence across time.

It also names the family that modern work has moved towards. If the honest observable model is a functional of the path, then write one: volatility as a function of exponentially weighted moments of past returns, or in full generality as a functional of the path signature, which is dense in continuous path functionals. Such a model is fully observable by construction, and it has no latent state to filter.

Remark (What a neural stochastic differential equation does and does not escape)**.** A network in the coefficients of

$$dS_t = \mu_t\,dt + \sigma_\theta(t, S_t, v_t)\,S_t\,dW_t$$

is still an Itô diffusion with adapted volatility. It does not go beyond the form — chapter 2 showed the form is not a restriction — it goes beyond the *parametric families* within the form, which is a real gain and a different one.

The genuine extensions are three, and none is a change of equation. Richer path dependence, as above. Rough volatility, where v is driven by a fractional Brownian motion with H near 0.1: the volatility is then neither Markov nor a semimartingale, while the price remains a perfectly good semimartingale and no-arbitrage is untroubled. And jumps, which is the only one that changes the shape, and chapter 2 says what is obtained.

There is also a practical asymmetry to weigh, and it is this chapter's argument in a new setting. A latent-state neural model needs filtering inside the training loop, since the likelihood integrates over unobserved paths — which is why such models are fitted by variational approximation and why their parameters are so poorly determined. A path-dependent model has no latent state, so every conditioning variable is observed, the likelihood is direct, and the parameters are identified by data one actually has. If identification rather than compute is the binding constraint, flexibility is better spent on the path dependence than on a hidden state.

21.3 Regimes, and Whether They Can Be Caught

One observation motivates latent states more than any other: past prices are frequently a poor guide to future ones, in a way that looks less like noise than like something changing. A policy stance, a liquidity condition, information held by somebody else — none of it visible in the price

history until it acts. The standard response is a regime: volatility takes one of two values, the state switches at random times, and the state is not observed.

It is tempting to describe this as a parameter changing rather than a state being hidden, and the distinction does not survive inspection. A regime indicator *is* a state process; calling σ a parameter that changes is declining to model it, and modelling it turns the parameter into a state.

Remark (What actually separates a regime from a diffusive variance)**.** Something does separate them, and it is not what the word "latent" suggests.

Neither is hidden under continuous observation. If the regimes differ in volatility then $v_t = \sigma^2(R_t)$ is recoverable from the bracket by the argument above, so the regime is recoverable too — a two-state model is no more opaque than a square-root variance. The difference is in the *predictability of the change* rather than the observability of the level. A diffusive variance has continuous paths, so its next value is close to its current one and the near future is forecastable from the present. A regime switches at a totally inaccessible stopping time in the sense of chapter 2: no announcing sequence exists, so nothing in the observed past gives warning. It is fully observable and completely unforecastable, and those two properties are usually assumed to be incompatible.

Which leaves the practical difficulty exactly where the next section puts it. With continuous noiseless observation neither object is hidden, so everything that makes a regime hard is the discreteness and the noise of real observation, and (21.2) is the measure of what that costs.

There is a third case. Suppose the transition law is unknown, or the states cannot be enumerated — which is what people usually mean when they say the world changed. Then no filter applies, because a filter needs a likelihood for each state and a prior over the set of them, and neither exists. That is not a latent state but an unspecified model, and the honest response is chapter 24's reserve rather than an estimate. Distinguishing the two is worth more than distinguishing parameters from states: one is a computation and the other is an admission.

Remark (The regress, and where to stop)**.** Modelling a parameter as a process does not remove an arbitrary choice, it relocates one. A constant σ becomes a variance process with its own κ, θ and η; making those stochastic in turn introduces theirs. Each level pushes the arbitrariness up rather than eliminating it, and there is no level at which the regress terminates on its own.

So the stopping rule has to come from outside the modelling, and this chapter supplies it. Add a level only where the data can see it move faster than it moves — which is (21.2) again, applied to the new state rather than to a regime. A parameter promoted to a process whose variation the observations cannot resolve has not been modelled; it has been given somewhere to hide, and the fit will improve while nothing has been learned. That is the whole of this chapter in one criterion.

Remark (The Bayesian version is the natural one)**.** A regime is a discrete state, so the posterior over it is a probability rather than a distribution over a continuum, and the filter is correspondingly cheap. Predict by the switching probability, update by Bayes against the two observation densities, renormalise: two multiplications per observation, exactly.

What that buys is the object the Bayesian calibration above asked for. The output is not a regime but a probability of a regime, so every price computed from it is a mixture, and the width of that mixture is a confidence band that updates as each new price arrives. A desk holding a position can therefore report not only its value but how much of the value is a bet on the state estimate, which is the number that ought to size the position.

The question that decides whether any of this is worth building is not accuracy but *speed measured in data*. A regime that is detected reliably after two hundred observations, and lasts fifty, has been correctly identified and is of no use whatever. That comparison has an answer.

Calculation 21.2 (How long it takes to notice)**.** Sequential detection theory gives the expected delay to declare a change, subject to a false alarm rate α, as

$$\text{delay} \;\approx\; \frac{\ln(1/\alpha)}{D} \quad \text{observations}, \qquad D = \tfrac{1}{2}\big(r - 1 - \ln r\big), \quad r = \sigma^2_{\text{high}}/\sigma^2_{\text{low}}, \tag{21.2}$$

with D the Kullback-Leibler divergence per observation between the two regimes' return densities. The numerator is the evidence that must be accumulated; the denominator is the rate at which each observation supplies it.

Running the filter on simulated switches, with the threshold at a posterior of 0.9:[2]

Regime shift	D per observation	Measured delay	Predicted
15% → 30%	0.807	9.4	8.5
15% → 25%	0.378	16.3	17.3
15% → 18%	0.038	97.9	152.2

The measured delay tracks (21.2) for the two wide shifts and beats it for the narrow one, which is expected — the bound is asymptotic in small α and the filter is running at a false alarm rate of a third of a per cent.

Read the last row. A volatility regime moving from fifteen to eighteen takes about a hundred trading days to establish at ninety per cent confidence, which is five months. If such a regime lasts weeks, it is real, it is undetectable in time to act on, and a model containing it is describing something nobody can trade.

Structure (The screening criterion comes before the model)**.** Equation (21.2) turns a modelling question into an arithmetic one, and it should be asked before anything is built.

Is the expected life of the regime longer than $\ln(1/\alpha)/D$*?* If not, stop. No filter, Bayesian or otherwise, recovers information the observations do not contain, and the delay above is not a property of the algorithm — it is a property of the two distributions being told apart. A better estimator cannot beat it, because it is the rate at which the data distinguishes the hypotheses.

Three consequences follow.

The gap matters quadratically, not linearly. Expanding D for a small shift gives $D \approx (r-1)^2/4$, so halving the difference between the regimes quarters the information and quadruples the delay. Regimes that are subtle are not slightly harder to detect; they are out of reach.

Sampling faster helps, in calendar time, and only inside the model. The interval cancels out of D entirely — a return carries the same information about a volatility *ratio* whatever span it covers — so the delay in observations is fixed by the regimes alone and the delay in calendar time is that divided by the observation frequency.

The cancellation, however, needs the returns to be independent draws of variance $\sigma^2\Delta$, so that halving Δ doubles the count of observations and leaves each one's information intact. Three things go wrong with that as Δ shrinks.

[2] `estimation::RegimeFilter`, which measures the delay against (21.2) rather than assuming it.

Microstructure noise floors the estimate, by (21.4), so past some frequency the extra observations are measuring the spread.

The returns also stop being independent, since the same noise induces autocorrelation in them — and n dependent observations do not carry n times one observation's information, so the delay in observations stops falling as fast as the count rises.

The third is the one that is easiest to miss and hardest to repair. *The quantity being estimated changes with the scale.* Intraday volatility carries a pronounced daily pattern, opening and closing far above the middle of the session, and it responds to liquidity events that leave no mark on a daily series. So a regime in minute-by-minute volatility is not a finer measurement of the daily regime; it is a different object with its own dynamics. Sampling faster to detect a shift in daily volatility can converge quickly and confidently on a shift in something else.

Which gives the honest version of the criterion. The frequency that optimises the volatility measurement is a ceiling and not a target: sample as fast as (21.4) allows, but no faster than the scale on which the regime you care about is defined. If the regime is a monthly change in daily volatility, minute data does not shorten the delay for it, however many observations it supplies.

And the criterion generalises past regimes. Any latent structure — a switch, a jump in a parameter, a change in correlation — is worth carrying only if the data separates its states faster than the states change. That is a sharper version of this chapter's recurring complaint. Elsewhere the trouble was that the calibration set did not determine a parameter; here it is that the data cannot determine it *in time*.

21.4 Calibrating to Prices and to History at Once

The programme suggested by (21.1) is more ambitious than fitting a surface, and it is the natural one to attempt: learn σ_θ so that the model reprices the liquid instruments *and* reproduces the dynamics the price history actually exhibits. One model, two data sources, two terms in the loss.

The question is whether that is coherent, since the two datasets live under different measures. It is, and the reason is exact rather than approximate — which makes this one of the few places in the chapter where the news is good.

Remark (Girsanov decides what is shared)**.** A change of measure alters the drift of a diffusion and leaves its diffusion coefficient untouched. That is chapter 6's theorem, and read as a statement about estimation it says: σ is the same object under the historical measure and the pricing measure, while μ is not.

Under the pricing measure the drift is not a free parameter at all — it is r for a tradeable, or the drift condition of chapter 8 for a curve. So of the two coefficients in (21.1), one is determined by no-arbitrage and the other is shared between the two measures. The only genuinely free object linking them is the market price of risk $\lambda = (\mu^{\mathbb{P}} - \mu^{\mathbb{Q}})/\sigma$.

That much is standard. What makes the combination pay off is a fact about what a price history can and cannot measure.

Calculation 21.3 (What a history identifies, and what no amount of it does)**.** Observe a diffusion with drift 5% and volatility 20% over a fixed window of one year, sampled n times, and estimate both coefficients. Averaged over four hundred independent histories:[3]

[3] `estimation::identification_by_frequency`.

Samples in the year	Volatility error	Drift error
250	3.64%	0.158
2 500	1.12%	0.155
25 000	0.36%	0.159

The two columns do not merely differ in rate. The volatility error falls like $\sqrt{2/n}$ and can be driven as low as the data allows. The drift error does not move at all — not slowly, but exactly not at all — because the estimator $(X_T - X_0)/T$ is a function of the endpoints and the intermediate samples are not merely uninformative about the drift, they do not appear in it. Its standard error is $\sigma/\sqrt{T}$ regardless of n.

Only a longer window helps: sixteen times the history quarters the drift error, which is the same $1/\sqrt{T}$ arriving from the other direction.

One caveat, which the next section is about: the volatility column says what a *diffusion* gives up under sampling, and a price is not a diffusion when looked at closely enough. The improvement stops, and then reverses.

Structure (The division of labour is exact)**.** Put the two facts together and they interlock neatly.

The parameter a price history estimates *well* — arbitrarily well, given enough sampling frequency — is the diffusion coefficient. The parameter it estimates *not at all* over a fixed window is the drift. And by Girsanov the diffusion coefficient is precisely the one shared with the pricing measure, while the drift is precisely the one no-arbitrage has already determined there.

The payoff is a constructive answer to this chapter's own complaint. Every under-identified parameter these notes have found is a parameter of the *diffusion* structure — β in chapter 10, the mixing weight in chapter 11, the mean reversion in chapter 12, the correlation between forwards in chapter 13. All of them are measure-invariant. So all of them are in principle visible in the time series, which is a different dataset from the one that failed to determine them, and Calculation 21.3 says the time series is the good direction to look. When this chapter says the fix for an unidentified parameter is more information, this is where the information is.

Remark (Three reasons it is harder than that)**.** The argument above is correct and it is not a recipe. Three things stand between it and a working joint calibration.

The processes are latent. Quadratic variation identifies the volatility of something *observed*. The parameters that are unidentified belong mostly to processes that are not: the variance in a stochastic volatility model, the mixing weight, the state of a multi-factor curve. Estimating those from prices is a filtering problem rather than a sum of squared increments, which is why the particle filter appears earlier in this chapter, and the rates degrade accordingly.

Some of them are drift parameters of a latent process. A mean reversion is the drift of the variance or of the curve factor, so it inherits Calculation 21.3's bad column: it needs a long history rather than a finely sampled one, and this chapter has already measured that a mean reversion fitted to a finite sample comes out biased fast. Sampling faster does not touch it.

The over-identifying restriction can fail, and its failure is information. If both sources speak about σ, they can disagree — and they routinely do, because implied volatility exceeds realised volatility on average. That gap is not an error in either estimate; it is §21.7's subject. But it means a joint calibration that assumes the two agree will corrupt a σ the option prices had identified correctly, and one that does not assume it must model the discrepancy.

So the honest summary is narrower than the structure above suggests. Joint calibration adds genuine information about the diffusion structure, which is where these notes keep finding the gaps.

It buys that at the price of requiring a model for the risk premium, and it should be run as a *test* before it is run as a fit: fit to prices alone, then ask what the historical dynamics of the fitted model would look like and whether they resemble the ones observed.

21.5 Estimating a Volatility

The previous section leaned on volatility being the easy parameter. Four decisions have to be made before a number comes out, and each of them moves it by more than the precision anybody quotes.

Over what window

Volatility is not constant, which is the premise of five chapters of these notes, so a sample average over a window is an average of something that moved. A long window has small sampling error and estimates a blend of regimes; a short one tracks the current regime and is noisy. There is no correct answer, only a stated one, and the honest practice is to report the window alongside the number and to check that the conclusion does not depend on it.

The choice interacts with the use. A volatility feeding a risk model over a ten day horizon and a volatility feeding a five year swaption are not the same quantity even if both are called annualised volatility, and using one for the other is the commonest way a risk number and a pricing number come to disagree for reasons nobody can locate.

At what resolution, and the answer is the hedging frequency

The sampling interval looks like a purely statistical choice and it is not. Chapter 5 showed that the profit and loss of a delta hedged position over an interval is

$$\tfrac{1}{2}\Gamma S^2 \left(\left(\frac{\Delta S}{S} \right)^2 - \sigma^2 \Delta t \right),$$

so what a hedger is paid, interval by interval, is the difference between the *squared move over their rebalancing interval* and the variance they sold. Accumulated over the life of the option, the quantity that determines whether the hedge made or lost money is the realised variance sampled at the rebalancing frequency.

That answers the question. A desk rebalancing daily is exposed to daily realised variance and should estimate it from daily data, and one rebalancing hourly to hourly variance. The two are not the same number — if they were, the signature plot below would be flat — and the difference is not an estimation error to be minimised but a genuine difference in what is being hedged.

The square root is not free

Almost every estimator computes a variance and roots it. The sum of squared increments is unbiased for the variance; the square root is concave, so by Jensen the volatility that comes out is biased *low*.

Calculation 21.4 (The correction, in closed form)**.** For n increments of a driftless Gaussian, $n\hat{\sigma}^2/\sigma^2$ is chi-squared with n degrees of freedom, so

$$\mathbb{E}[\hat{\sigma}] = \sigma\sqrt{\frac{2}{n}}\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \;\approx\; \sigma\left(1 - \frac{1}{4n}\right). \tag{21.3}$$

At twenty increments — a month of daily data — the factor is 0.98758, so the volatility comes out 1.24% too low, which on a 20% volatility is a quarter of a point.[4]

The bias is one-signed, so averaging across months does not remove it, and it is available in closed form, so there is no reason to carry it. It is the smallest of the four difficulties here and the only one with an exact answer.

And a price is not a diffusion when looked at closely

The serious difficulty is the one that contradicts the previous section's arithmetic. What is observed is not the efficient price but the price at which a trade happened, which alternates between the bid and the offer. Write the observation as $Y = X + \epsilon$ with ϵ independent noise of scale η, roughly the half spread. Each observed increment carries two noise draws, and the noise does not shrink with the sampling interval, so

$$\mathbb{E}\big[\text{realised variance over } n \text{ samples}\big] = \sigma^2 T + 2n\eta^2. \tag{21.4}$$

The bias is *linear in the sampling frequency*. Sampling faster does not converge on the truth, it diverges from it.

Calculation 21.5 (The volatility signature plot)**.** A 20% volatility over one trading day, observed with a one basis point half spread.[5]

Samples in the day	Measured volatility	Error
6 (hourly)	19.09%	−4.5%
26 (quarter hourly)	19.86%	−0.7%
78 (five minutes)	20.05%	+0.3%
390 (one minute)	20.48%	+2.4%
1 560 (fifteen seconds)	21.87%	+9.3%
7 800 (two seconds)	28.16%	+40.8%

At two second sampling the measured volatility is half as large again as the truth, and it is not converging — it is measuring the spread.

Structure (Two biases pointing opposite ways)**.** The table is not monotone, and the reason is the useful part.

At coarse sampling (21.3) dominates: few increments, so the square root pulls the estimate down, and hourly sampling is four and a half per cent low. At fine sampling (21.4) dominates and the estimate runs away upwards. The two biases point in opposite directions and there is a

[4] `estimation::sqrt_bias_correction`, with `a simulation confirming that dividing by it removes the bias` rather than merely that the formula is what it is.

[5] `estimation::realised_variance_with_noise`, measured against (21.4) rather than against intuition.

frequency where they cancel — here around a quarter of an hour, and the five minute sampling that the literature recommends is within a fraction of a per cent of the truth for the same reason.

So the standard advice to sample every five minutes is not a rule of thumb about market conventions. It is where two known biases of known size happen to cross for typical parameters, and (21.3) and (21.4) say how it moves: a wider spread pushes the optimum coarser, a shorter horizon pushes it finer. A desk trading an instrument with a ten basis point spread and using five minute sampling because that is what the papers say is making an error of a hundred times the size the papers were addressing.

It also settles the previous section's overreach. The claim that a fixed history identifies volatility arbitrarily well is a statement about a diffusion. Against real observations the error falls, reaches a floor set by the spread, and then rises, so the identification argument holds down to the noise and no further — which is still a great deal better than the drift, whose error never falls at all.

Remark (What is done about it)**.** Three responses, in increasing order of effort. Sample at the frequency where the biases cross and accept the residual, which is what most desks do and which Calculation 21.5 shows is defensible. Subsample and average — compute the realised variance on several offset grids and average them, which reduces the variance without changing the bias. Or use an estimator built for the problem: two-scale realised volatility differences the estimate at two frequencies to cancel the noise term in (21.4), and realised kernels do the same by weighting the autocovariances that the noise induces.

The third route is the correct one: the noise is a known functional form with an unknown scale, so it can be estimated and removed rather than avoided. Equation (21.4) has η in it, and anything with a parameter in it can be fitted.

21.6 Estimating a Correlation

Chapter 13 produced an uncomfortable result: the correlation between rates moves a swaption volatility by percentage points, while the choice between two whole model families moves it by hundredths. If that is right — and it is measured — then the correlation estimate is the most consequential number in the whole apparatus, and it deserves more care than it usually gets.

It is also the hardest thing in this chapter to estimate, for a reason that is dimensional rather than statistical.

Calculation 21.6 (A correlation matrix of pure noise)**.** Take forty independent series — the forwards of a ten year quarterly structure, if they had no correlation at all — and estimate their correlation matrix from one year of daily observations. The truth is the identity, so every eigenvalue is one. The estimate is not close.[6]

Series	Observations	Largest	Smallest	Ratio
40	250	1.84	0.39	4.7×
40	1 000	1.40	0.66	2.1×
40	4 000	1.19	0.83	1.4×

[6] `estimation::sample_correlation_spectrum`, against the Marchenko-Pastur edges rather than against an impression.

At a year of data the largest apparent factor looks nearly five times the smallest, in a matrix with no factors in it whatever. Anyone reading a scree plot of those eigenvalues would report a rich factor structure and would be reporting nothing.

Remark (The spread is predictable, which is what makes it usable)**.** The numbers above are not accidents of the simulation. Estimating p series from n observations means extracting $p(p-1)/2$ numbers from pn data points, and the ratio $q = p/n$ controls everything. The Marchenko-Pastur law says the sample eigenvalues of independent series fill the interval

$$\left[(1-\sqrt{q})^2,\ (1+\sqrt{q})^2\right], \tag{21.5}$$

and the measured columns above sit against those edges to within a few per cent.

That is more useful than a warning, because it supplies a null hypothesis. An eigenvalue *inside* (21.5) is consistent with noise and carries no information; one above the upper edge is a factor that is really there. So a scree plot becomes a test rather than an impression: draw the upper edge on it, and count what pokes out. For a curve, three or four eigenvalues clear the edge comfortably — which is chapter 24's level, slope and curvature, now established rather than asserted — and the rest of the spectrum is the noise bulk.

The remedy follows the diagnosis. Keep the eigenvalues above the edge, replace the bulk with its average, and rebuild the matrix; or shrink the sample matrix towards a structured target, which is the same operation performed continuously rather than by a cutoff. Both are ways of refusing to estimate what the data cannot support.

Structure (Three routes, and the market model takes the third)**.** There are only three things one can do about Calculation 21.6.

More data. It works and it works slowly: the excess of the top eigenvalue over one falls like $\sqrt{q}$, so sixteen times the history roughly quarters it. Sixteen years of daily data also spans several regimes, so what is gained in sampling error is lost in the assumption that the correlation was constant — which returns to the window problem of the previous section, now with sharper consequences.

Cleaning. Estimate the matrix, then remove the part attributable to noise, by eigenvalue clipping or by shrinkage towards a target. This is the honest general-purpose answer and it makes no structural assumption beyond the choice of target.

Parametrise. Assume a form. Chapter 13 uses $\rho_{ij} = e^{-\beta|T_i - T_j|}$, one parameter for the whole matrix, which cannot overfit because there is nothing to overfit with. The cost is that it is wrong in a specific way — real curve correlations do not decay to zero at long separations, they flatten out at a positive level, so the exponential understates the coupling between the long end and the very long end. A two parameter form $\rho_{ij} = \rho_\infty + (1-\rho_\infty)e^{-\beta|T_i - T_j|}$ fixes exactly that and is what is generally used.

The third route is the right one here and the reason is Calculation 21.6 rather than parsimony for its own sake. Forty forwards from a year of data cannot support a free matrix; the estimate would be noise wearing the shape of structure, and chapter 13 showed that this is the parameter the price is most sensitive to. When the data cannot identify a parameter and the answer depends on it, imposing a form is not laziness — it is the only alternative to letting the noise choose.

Remark (And it is not constant, in the direction that hurts)**.** One further problem, which no amount of cleaning addresses. Correlations move, and they move *up* in stress: rates that normally decouple stop decoupling when everything is being sold at once.

So an estimate from a calm sample understates the correlation in exactly the scenario a risk calculation is about, and chapter 18 has already made the general form of this point — the coupling in the tail is not the coupling in the body, and a single correlation number cannot carry both. For pricing, the calm number is arguably the right one, since it is the average over the option's life that matters. For risk it is not, and chapter 24's stress work has to use something else. Reporting one number for both uses is the error, rather than any particular estimate of it.

21.7 Why Implied Exceeds Realised

The gap is large, persistent and one-signed, so it is not a misfit to be calibrated away. Two explanations are usually offered and they are frequently presented as alternatives. They are not, and separating them matters for §21.4, because they imply different things about whether a joint calibration is well posed.

It is a risk premium, and it sits where Girsanov says it must

The first explanation is that variance is a priced risk factor. A position that is long volatility pays off when volatility spikes, and volatility spikes when everything else is going wrong, so it is a hedge — and a hedge earns a negative expected return, which is to say it costs a premium to hold. Options are that hedge, so they are expensive relative to what the underlying goes on to do.

This is formally the same object as the equity risk premium and belongs in the same place. Under the historical measure a stock drifts faster than a bond because holding it is uncomfortable; under the pricing measure it does not. Here the variance process drifts differently under the two measures for the same reason.

Calculation 21.7 (The premium, written down)**.** Take the Heston variance of chapter 10 under the historical measure, and let the market price of variance risk be proportional to volatility, so that changing measure adds $-\lambda v\,dt$ to the variance's drift. Then

$$\kappa^{\mathbb{Q}} = \kappa + \lambda\eta, \qquad \theta^{\mathbb{Q}} = \frac{\kappa\theta}{\kappa^{\mathbb{Q}}}, \tag{21.6}$$

and η is unchanged. Take a 20% long-run volatility, mean reversion $\kappa = 2$, vol-of-vol $\eta = 0.3$, and the variance starting at its own long-run level so that nothing below is confused with a transient:[7]

λ	$\kappa^{\mathbb{Q}}$	**One year implied**	**Premium**
-1	1.70	20.90%	0.90 pts
-2	1.40	21.89%	1.89 pts
-3	1.10	23.00%	3.00 pts

against a realised expectation of exactly 20%. The observed gap on an equity index is one to three volatility points, so a market price of variance risk of a few units reproduces it — the explanation is the right size, which is the first thing to check about any explanation.

Structure (The premium cannot be in the vol-of-vol)**.** If implied volatility carries a risk premium, it is tempting to suppose that vol-of-vol carries one too — that η is higher under the pricing measure

[7] `heston::variance_premium`, with the expected integrated variance in closed form rather than simulated.

the way θ is. It is not, and it cannot be. η is a diffusion coefficient, and a change of measure moves drifts and leaves diffusion coefficients alone. The premium has nowhere to go except into κ and θ.

That is §21.4's division of labour applied one level down, and it has a consequence that is directly useful. Vol-of-vol is measure-invariant, so it is *jointly identified*: an η estimated from the time series of realised volatility and an η implied by the curvature of the smile are estimates of the same number, and they can be compared. If they disagree, something is wrong with the model rather than with the premium, because no premium can explain a discrepancy in a diffusion coefficient. This is one of the few genuinely testable cross-measure restrictions in derivatives pricing, and it is almost never tested.

It is one-sided flow, and dealers cannot arbitrage it away

The second explanation makes no reference to risk aversion. Banks are structurally short options, because their clients — corporates hedging liabilities, insurers hedging guarantees, funds buying protection — want to own them. A dealer who cannot lay the position off must hold it, and chapter 23 derives what happens next: a market maker holding inventory it did not choose quotes to be paid for holding it. The premium is then a price for absorbing one-sided demand, in exactly the sense chapter 22 requires of a trade — somebody is on the other side for a reason that is not an opinion.

This explanation is not a rival to the first. It is a second premium, paid for a different service, and both are collected by the same position.

And liquidity, which is why neither gets arbitraged away

A premium of one to three volatility points invites the obvious question: why does nobody sell it until it is gone? The answer is the third component, and it is not a separate explanation so much as the reason the first two survive.

Harvesting the premium means selling options and delta hedging to expiry. The round trip costs the option's bid-offer, which on a liquid index swaption is a fraction of a volatility point and on anything else is more, plus the accumulated cost of rebalancing the delta, which chapter 23 prices and which grows with how often one rebalances — and Calculation 21.5 has just shown that rebalancing frequency is not a free choice either. A premium smaller than the cost of collecting it is not an arbitrage; it is a fee for a service that is genuinely expensive to provide.

That produces a floor. *A model error smaller than the bid-offer spread is not a model error anyone can trade*, so it cannot be arbitraged away and there is no force driving it to zero. Chapter 13 used exactly this to settle the choice of coordinates: an approximation two orders of magnitude below the spread is free. The same reasoning says the variance premium, being larger than the spread but not by much, is exactly in the range where it persists — big enough to be real, small enough that collecting it is a business rather than an arbitrage.

Liquidity also enters as a premium in its own right, and in the same direction. An instrument that cannot be exited cheaply trades at a discount, which is to say a higher expected return, and chapter 22's on-the-run spread is that discount measured directly. For options the effect concentrates where it is least convenient: far strikes and long tenors, which have the widest spreads, the thinnest volume and the least reliable marks — and which are also where a model is least constrained by quotes. So the part of the surface where calibration has the least to say is the part where the price is least trustworthy, and the two failures compound rather than offset.

Remark (Volume is a poor proxy and it is the one everybody uses)**.** Since liquidity is not directly observable, it is proxied — usually by volume, sometimes by quoted spread, occasionally by the price impact measures of chapter 23. Volume is the worst of the three and the most used, because it is high in two opposite circumstances: when an instrument is easy to trade, and when everybody is trying to get out of it at once. A liquidity signal built on volume therefore reads its highest exactly when liquidity is about to be worst, which is the failure mode chapter 22's six trades all share.

Quoted spread is better and still incomplete, since a quoted spread is for a quoted size and the size is what matters in a stress. The honest measure is the one chapter 23 builds, price impact per unit of quantity, and it requires data most participants do not have.

Remark (How to tell them apart, and why it matters here)**.** The two have different signatures, and the distinction is testable rather than philosophical.

A risk premium is compensation for an exposure, so it should be roughly proportional to the quantity of variance risk borne and should have a term structure fixed by the model — Calculation 21.7 shows it emerging over the mean reversion time, so it grows with horizon in a shape κ determines. It should be present in every market where variance is risky, in proportion to how risky.

A flow premium is compensation for absorbing inventory, so it should track *dealer capacity* rather than risk: largest where client demand concentrates, which is downside strikes and specific tenors rather than uniformly across the surface; widening when balance sheet is constrained, at quarter ends and after dealer losses; and varying across markets with who happens to be buying rather than with how volatile they are.

Both signatures are visible in the data, which is the honest answer to which explanation is right: there is a persistent baseline consistent with a risk premium, and a strike- and tenor-dependent component that moves with demand.

For §21.4 that distinction is decisive rather than academic. If the gap is a risk premium, λ is a smooth function of the state with few parameters, it can be modelled, and a joint calibration to prices and history is well posed — one estimates the shared diffusion structure and the premium together. If the gap is flow, then the quantity separating the two measures depends on dealer positioning, which appears in *neither* dataset. A joint fit will then attribute a flow effect to dynamics, and the dynamics is what it was trying to learn. So the programme's viability rests on how much of the gap is which, and that question has to be answered before the fit rather than by it.

21.8 What To Take Away

Three things carry over from chapter 20, sharpened by what a latent state adds.

First, speed is not knowledge. A neural surrogate for a pricing map, a particle filter, a posterior sampled instead of optimised — each turns a slow calculation into a fast one, and none of them extracts information the quotes or the history did not contain. The flat valley in a SABR fit is exactly as flat once the fit is instant.

Second, not everything travels between the two measures, but more does than is usually assumed. A diffusion coefficient is the same object under $\mathbb{P}$ and $\mathbb{Q}$, so it can be estimated from history and checked against what is implied; a drift, a mean reversion, or the risk premium in the

gap between implied and realised volatility cannot make that trip, because a change of measure is precisely what moves them.

Third, a latent state is usually a smaller problem than it is made out to be, and a real one is a race. Continuous observation of a price reveals its own quadratic variation and, with it, whatever a model called hidden; what remains genuinely unknown is what discreteness and noise destroy, and a change worth modelling is one the data can tell apart from noise faster than the change itself unfolds.

The tools for computing have changed enormously. The questions have not.

References

- Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

- Horvath, B., Muguruza, A., & Tomas, M. (2021). Deep learning volatility. *Quantitative Finance*, 21(1), 11–27.

- Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8), 1271–1291.

- Han, J., Jentzen, A., & E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *PNAS*, 115(34), 8505–8510.