Chapter 24

Risk Management

In these notes we stop asking what a trade is worth and start asking how wrong we might be about it. We set out the axioms a risk measure ought to satisfy, show that the industry's standard measure fails one of them, derive the measure that repairs it, and then work through how such numbers are actually computed, tested, and set aside as capital — including for the model risk the previous four chapters kept uncovering.

24.1 A Different Question

Every chapter so far has answered the same question: what is this worth? The answer was always an expectation under a measure chosen so that no arbitrage is possible, and the work was in finding the measure and computing the expectation.

Risk management asks something else. Given that we have marked the book at those prices, how much could we lose, how likely is that, and how much capital should be held against it? The question is asked in the real world measure — the $\mathbb{P}^{\text{historic}}$ of chapter 4, the one that dropped out of every pricing formula — because we are now asking what will actually happen rather than what a hedged position costs.

This is a genuine reversal. The whole apparatus of chapters 4 through 12 was built to eliminate the historic measure from pricing. Risk management puts it back, and the two answers are not comparable: a position can be worth zero and be enormously risky, which is precisely what a hedged book is.

Three kinds of risk turn up:

- *Market risk*: the book loses money because prices move. The subject of most of this chapter.

- *Model risk*: the book was marked with a model, and the model was wrong. Chapters 9, 10 and 11 each ended by identifying a parameter that the calibration could not determine and the product depended on. This chapter says what to do with them.

- *Counterparty risk*: the trade was right and the other side did not pay.

Structure (A number without a decomposition is not a risk number)**.** A risk measure returns one number. That number is nearly useless on its own, because the position it describes was assembled by three quite different processes and only one of them is a live decision.

- *Risk the desk chose.* A view, taken deliberately, in a size somebody signed off. This is what a risk report is usually imagined to be about, and in a hedged book it is the smallest of the three.
- *Risk the flow chose.* Chapter 23 ends with a book nobody designed: its composition is the sum of what clients asked for. Nothing about it was selected, and the only decisions available are how much to quote and what to hedge.
- *Risk the model chose.* The hedge was computed from a model. If the model is wrong, the hedge is wrong, and the residual is a position the desk holds without knowing it holds. Chapter 10 measures one instance — a delta wrong by five per cent of notional, in a consistent direction — and §24.8 measures how large this term is in general.

Only the first can be changed by deciding to change it. The second is changed by quoting differently, which is a business decision with revenue on the other side. The third cannot be changed at all without changing the model, and no amount of capital held against it makes it smaller.

So the useful output is not $\rho(L)$ but the split, and a risk function whose answer cannot be attributed is not doing the job it is there for. That claim is the reason this chapter comes last.

24.2 What a Risk Measure Should Satisfy

Write L for the loss on a portfolio over some horizon — a random variable, positive when money is lost. A risk measure is a function ρ turning that random variable into a single number, to be read as the capital that should stand behind it.

Rather than propose a formula and defend it, the productive move is to write down what any such number ought to satisfy and see what survives. This is the approach of Artzner and coauthors.

Definition 24.1 (Coherent risk measure)**.** ρ is coherent if for all losses L, M and constants $c > 0$:

$$\begin{aligned} \text{(monotonicity)} \quad & L \le M \implies \rho(L) \le \rho(M), \\ \text{(translation invariance)} \quad & \rho(L + c) = \rho(L) + c, \\ \text{(positive homogeneity)} \quad & \rho(cL) = c\,\rho(L), \\ \text{(subadditivity)} \quad & \rho(L + M) \le \rho(L) + \rho(M). \end{aligned}$$

Monotonicity: a position that loses more in every state requires more capital. Hard to argue with.

Translation invariance: adding a certain loss of $\$c$ increases the required capital by exactly $\$c$. This is what makes the number readable as an amount of money — it says that holding $\rho(L)$ in cash against the position brings its risk to zero, since $\rho(L - \rho(L)) = 0$.

Positive homogeneity: doubling the position doubles the risk. This is the most arguable of the four, because in a crisis a position twice as large is more than twice as hard to sell, and a measure honouring that would be superadditive in size. It is retained for tractability.

Subadditivity: a merged portfolio is no riskier than the two apart. This is the important one. It is the statement that diversification cannot hurt, and it is what makes risk numbers add up sensibly across a firm — if it fails, the sum of the desks' risk can be less than the firm's, and a trader can reduce measured risk by moving a position to another book without changing anything real.

24.3 Value at Risk, and What Is Wrong With It

Definition 24.2 (Value at Risk)**.** At confidence α, the Value at Risk is the α-quantile of the loss:

$$\mathrm{VaR}_\alpha(L) = \inf\{x : \mathbb{P}(L \le x) \ge \alpha\}.$$

Read it as: on all but the worst $1-\alpha$ of days, the loss is no greater than this.

Three of the four axioms hold, and quickly. Monotonicity: if $L \le M$ everywhere then the distribution function of M is everywhere below that of L, so every quantile of M is at least that of L. Translation invariance and positive homogeneity are the corresponding properties of quantiles, which shift and scale with the variable.

Subadditivity fails. The standard reassurance — that it only fails for strange distributions — is not true, and the case where it fails is one banks hold a great deal of.

Example 24.1 (Diversification punished)**.** Take a corporate bond, notional \$100, which defaults over the horizon with probability 4% and is then worth nothing. The loss is \$100 with probability 0.04 and zero otherwise.

At $\alpha = 95\%$: the bond survives with probability 0.96, which is above 95%, so the 95% quantile of the loss is

$$\mathrm{VaR}_{0.95} = \$0.$$

The measure says this position needs no capital at all.

Now diversify. Hold two independent bonds of \$50 each, same issuer quality, same default probability. Then

$$\mathbb{P}(\text{neither defaults}) = 0.96^2 = 0.9216,$$
$$\mathbb{P}(\text{exactly one defaults}) = 2 \times 0.04 \times 0.96 = 0.0768.$$

Now $0.9216 < 0.95$, so the 95% quantile is no longer at zero loss but at the one-default outcome:

$$\mathrm{VaR}_{0.95} = \$50.$$

Splitting one position into two independent halves has taken the measured risk from nothing to half the notional. Writing A and B for the two halves, $\rho(A) = \rho(B) = 0$ but $\rho(A+B) = 50$, so subadditivity fails as badly as it can.

The failure is not a curiosity. It says that a firm managed against Value at Risk is being told that concentration is free and diversification is expensive, and the mechanism is completely general: the measure looks at a single quantile and is blind to everything beyond it. A position whose losses are rare and enormous — selling deep out-of-the-money options, writing insurance against a crisis, holding senior tranches — can be given a Value at Risk of zero, because the loss lives entirely in the tail the measure does not look at.

Remark. Notice which axiom is doing the work in the argument, and which is not. Nothing here depends on the losses being large or the probabilities being unrealistic; it depends only on the loss distribution having a jump near the quantile. Credit portfolios are made of jumps. So are books of digital and barrier options, whose payoffs are discontinuous by construction — and chapter 9 showed that a butterfly is the market's way of pricing exactly such a jump.

24.4 Expected Shortfall

The repair is to stop looking at one quantile and average over all of them beyond it.

Definition 24.3 (Expected shortfall)**.**

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)\,du. \tag{24.1}$$

In words: the average loss over the worst $1-\alpha$ of outcomes. Where Value at Risk asks how bad things get before the tail starts, expected shortfall asks how bad the tail is.

Remark (Not the conditional expectation)**.** It is often written as $\mathbb{E}[L \mid L \geq \mathrm{VaR}_\alpha]$, and for a loss with a continuous distribution the two agree. When the distribution has atoms they do not, and the difference is large.

Take the single bond of Example 24.1. Its 95% Value at Risk is zero, so the event $\{L \geq \mathrm{VaR}_{0.95}\}$ is the whole sample space, and the conditional expectation is the unconditional mean of \$4. Definition (24.1) instead averages the quantiles above 95%, of which the top 4% sit at a loss of \$100 and the rest at zero, giving

$$\mathrm{ES}_{0.95} = \frac{1}{0.05}\big(0.04 \times \$100\big) = \$80.$$

Eighty against four. The conditional form is not merely imprecise here, it is useless — and it is the form that fails to be coherent. When a distribution has jumps, and credit distributions are made of them, (24.1) is the definition.

Theorem 24.4. *Expected shortfall is coherent.*

Sketch. Monotonicity, translation invariance and homogeneity are inherited from the same properties of the quantiles being averaged in (24.1).

Subadditivity is the substantial one, and follows from a representation that matters on its own: expected shortfall is the worst expected loss over a family of scenarios,

$$\mathrm{ES}_\alpha(L) = \sup_{\mathbb{Q}\in\mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[L],$$

where $\mathcal{Q}$ is the set of measures with density at most $1/(1-\alpha)$ with respect to the real world measure. Given that, subadditivity is immediate: for any fixed $\mathbb{Q}$ expectation is additive, so

$$\mathbb{E}^{\mathbb{Q}}[L+M] = \mathbb{E}^{\mathbb{Q}}[L] + \mathbb{E}^{\mathbb{Q}}[M] \leq \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}[L] + \sup_{\mathbb{Q}} \mathbb{E}^{\mathbb{Q}}[M],$$

and taking the supremum on the left preserves the inequality. A supremum of linear functionals is convex; that is the whole argument. □

This representation says expected shortfall is the worst average loss across a set of stress scenarios, where the set is everything that does not distort the real world probabilities by more than a factor of $1/(1-\alpha)$. A coherent risk measure and a stress test are the same object, described two ways.

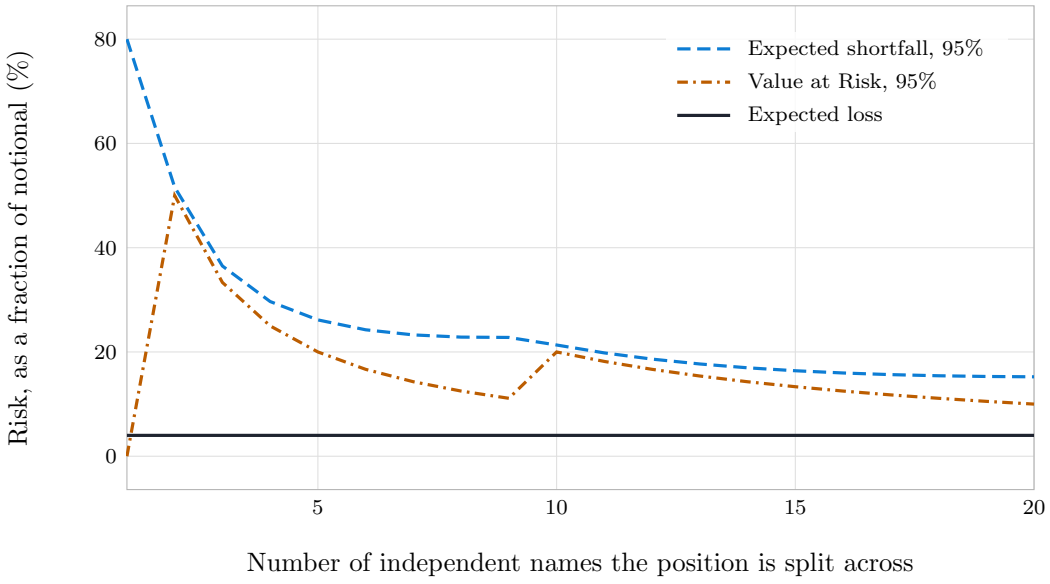

Figure 24.1: The two measures as a fixed notional is split across more independent names, each defaulting with probability 4%. Expected shortfall falls steadily towards the expected loss, which is what diversification ought to look like. Value at Risk starts at zero, jumps to half the notional at the first split, and then declines — and the jump is the failure of subadditivity, visible as the one place where diversifying makes the reported number worse.
Drawn by `Losses::value_at_risk` and `Losses::expected_shortfall`.

Exercise (Where the jump comes from)**.** In Example 24.1, find the confidence level α at which the single bond's Value at Risk jumps from \$0 to \$100, and the level at which the two-bond portfolio's jumps from \$0 to \$50. Deduce that the counterexample works for every α strictly between 0.9216 and 0.96, and that it disappears outside that window. This is why a measure defined by one quantile can be made to say almost anything by moving the quantile.

24.5 Computing It

Three methods, in increasing order of cost and fidelity. All three need the same input — a distribution of portfolio value changes — and differ in how they get it.

Variance-covariance

Assume the portfolio's loss is linear in a vector of risk factors X, and that those factors are jointly normal:

$$L = -\delta^{\top} \Delta X, \qquad \Delta X \sim N(0, \Sigma).$$

Then L is normal with variance $\delta^\top \Sigma \delta$, and both measures are available in closed form:

$$\mathrm{VaR}_\alpha = z_\alpha \sqrt{\delta^\top \Sigma \delta}, \qquad \mathrm{ES}_\alpha = \frac{n(z_\alpha)}{1-\alpha} \sqrt{\delta^\top \Sigma \delta},$$

with z_α the standard normal quantile and n its density.

The δ here is exactly the vector of sensitivities the pricing chapters produce: the DV01 buckets of chapter 7, the delta and vega of chapters 5 and 10. So the risk system consumes the same Greeks the trading desk hedges with, which is both convenient and the source of its main weakness.

The weakness is that the assumption of linearity is false for anything with optionality, and false in a direction that matters. A hedged option book has $\delta \approx 0$ and is not riskless at all — its risk is second order, in the gamma, and a measure linear in the risk factors reports zero. This is the same blindness as Example 24.1 wearing different clothes: the method cannot see a risk that is quadratic because it only asked about the first derivative.

Historical simulation

Take the last few hundred days of actual factor moves, apply each to today's portfolio, and read the quantile off the resulting distribution of losses. No distributional assumption at all, and the correlations are whatever they actually were.

Its weakness is its sample. A few hundred days contain very few tail events, so the estimate of a 99% quantile rests on a handful of observations, and the estimate of anything beyond it rests on none. And the window has to end somewhere: a crisis drops out of the sample on a particular day, and the reported risk of an unchanged portfolio falls that morning for no reason connected to the world.

Monte Carlo with full revaluation

Simulate factor paths, reprice the whole book on each, build the distribution. This is the only one of the three that is right for a book with optionality, because it reprices rather than approximating, and it is expensive for the same reason — a book of Bermudans repriced under fifty thousand scenarios is fifty thousand backward inductions.

This is where the Markovian state of chapter 12 stops being an aesthetic preference. A model whose state is two numbers can be revalued in a nested simulation; one whose state is the whole forward curve cannot.

24.6 The Coordinates a Risk Report Uses

A book of ten thousand trades is not revalued from scratch against every scenario. It is reduced to a short list of sensitivities, and every number downstream — the value at risk, the limits, the capital — is computed from those. So the sensitivities are the coordinates in which the book exists as far as risk is concerned, and anything they do not capture is invisible to everything built on them.

Definition 24.5 (The first-order risk report)**.** For a book of value V depending on an underlying S, its volatility σ, time t and a rate r, the standard sensitivities are

$$\Delta = \frac{\partial V}{\partial S}, \qquad \Gamma = \frac{\partial^2 V}{\partial S^2}, \qquad \mathcal{V} = \frac{\partial V}{\partial \sigma}, \qquad \Theta = \frac{\partial V}{\partial t}, \qquad \rho_r = \frac{\partial V}{\partial r},$$

known as delta, gamma, vega, theta and rho. In a rates book the first is a vector — one sensitivity per curve point or per factor — and vega is a matrix, indexed by expiry and tenor.

Three observations, each of which the earlier chapters have already established and none of which is a detail.

The list is a *Taylor expansion*, so it is an approximation whose error is the terms omitted. Chapter 5 shows that Θ and Γ are not independent for a hedged book but two readings of one quantity, which is why a gamma limit and a theta budget constrain the same thing.

Every entry is *model-dependent*, including delta. Chapter 10 is a chapter about exactly this: two models agreeing on every price today disagree on Δ, because the total derivative includes vega multiplied by how the smile moves, and that is a modelling statement rather than an observable. So a risk report is not a measurement of the book. It is a measurement of the book according to a model.

And the list is *complete only if the factors are.* A one-factor rates model has one delta, so a curve twist is not represented in its coordinates at all. Chapter 12 measures the pricing cost of that; the next section measures the risk cost, which is larger and easier to miss.

24.7 Explaining the Profit and Loss

The sensitivities make a prediction, and it can be checked. At the end of a day the book has made or lost a definite amount, and the risk report says what it should have been:

$$\Delta P\&L_{\text{predicted}} = \Delta \cdot \delta S + \tfrac{1}{2}\Gamma\,(\delta S)^2 + \mathcal{V} \cdot \delta\sigma + \Theta\,\delta t + \dots \tag{24.2}$$

Comparing that with what happened is *profit and loss explain*, it is run daily on every derivatives desk, and under current bank capital rules a desk whose explain is poor loses the right to use its own model. What is left over after (24.2) is subtracted is the unexplained residual.

The residual deserves care, because it is easy to read as noise and it is not. Three things produce it, and only the first is innocent.

- *Higher-order terms.* Genuine and small, and shrinking as the moves shrink. The cure is more terms.
- *Sensitivities computed badly.* A bumped delta from an unstable numerical scheme, or a stale surface. An operational problem with an operational fix.
- *Risk factors that do not span the moves.* The book was exposed to something the report has no coordinate for. No number of extra Taylor terms helps, because the missing quantity is not a derivative with respect to anything in the list.

The third is model risk appearing as a measurement, and it is why the explain is run at all.

Calculation 24.6 (What a one-factor report cannot see)**.** Take a curve driven by two independent factors, a level and a slope, with sizes measured from a year of Treasury curves rather than assumed.[1] Annualised, the level moves about 184 basis points a year and carries 79% of the variance of daily changes, the slope about 69 and carries 11%, and the curvature about 38 and carries 3%. Let the risk system carry a sensitivity to the level alone.

For a book with exposures a to the level and b to the slope, the unexplained share of profit and loss variance is

$$\frac{b^2\sigma^2_{\text{slope}}}{a^2\sigma^2_{\text{level}} + b^2\sigma^2_{\text{slope}}}, \tag{24.3}$$

which a regression of realised against predicted profit and loss recovers exactly.[2] Evaluating it:

Book	**level exposure**	**unexplained**
outright	1.0	1.2%
duration hedged to a tenth	0.1	93%
exactly level neutral	0	100%

with the slope exposure held at 0.3 for the first row and 1.0 for the others. The middle row is the one to sit with: a book whose duration has been hedged down to a tenth has ninety-three per cent of its profit and loss outside what its risk report can describe.

Structure (Hedging is what breaks the explain)**.** Equation (24.3) says the unexplained share depends on the book's *exposures* and not only on how much of the curve's variance each factor carries, and that has an uncomfortable consequence.

Hedging removes exposure to the factor the model understands. It does not touch the others. So a duration hedge takes the numerator of (24.3) as it finds it and shrinks the denominator, and the explain deteriorates for the same reason the book got safer. A level-neutral curve trade — which is what most relative value positions are — has all of its profit and loss in factors a one-factor report cannot see, however small a share of curve variance those factors carry.

The usual defence fails for exactly this reason. It is true that the slope factor carries an eighth of the level's variance, and tempting to conclude it can be neglected. What matters is the product of a factor's size with the book's exposure to it, and a hedged book has arranged for the other product to be small.

This is the same result as §24.8's, reached from the other end. There, hedging more often left model error as a larger share of the risk. Here, hedging more completely leaves unspanned factors as a larger share of the profit and loss. Both say that a better hedged book is not a book with less model risk but a book whose remaining risk is more purely model risk — and both are invisible to a risk measure applied to the model's own coordinates.

Remark (What the explain is therefore for)**.** Not for confirming that the greeks are right, which it does only weakly. If a book is marked and its sensitivities computed with the same model, the explain will look good even when the model is wrong, because both sides of the comparison share the error.

[1] `risk::curve_factors`, run on the history in `public/marketdata`. The first three components are the textbook ones and the code checks that they are: the first loads positively at every tenor, the second changes sign once across the curve and the third twice.

[2] `risk::ExplainTest`, both routes.

Its use is the opposite: the residual is an estimate of the risk the report has no coordinate for. A desk with a persistent unexplained term has measured a factor it is not carrying, and the useful response is to add the factor — which usually means a richer model — rather than to widen a limit. That is why the regulatory version withdraws model approval instead of demanding more capital: the finding is not that the book is risky but that the risk system is not describing it.

24.8 Attributing the Number

Everything so far computes $\rho(L)$. This section is about the claim made at the start, that the number is not useful until it is split.

Take a short one-year at-the-money call, delta hedged. It is the most thoroughly hedged position in this book: one factor, a liquid hedge, a known model. Whatever risk survives here is a floor for anything more complicated.

Calculation 24.7 (Three sources, on the same paths)**.** Hedge the option twice on *identical* paths — once with the volatility that is actually driving them, once with a volatility four points too high — so that the difference between the two is the mis-specification and nothing else. A four point error on a twenty vol is well inside the range a surface is marked to.[3]

Rebalances	correct model	wrong model	model part	ratio
50	0.98	1.12	0.55	1.15
250	0.44	0.71	0.55	1.60
1000	0.22	0.59	0.55	2.64

Read the columns rather than the rows. The first falls as the square root of the rebalancing frequency, which is chapter 4's hedging error behaving as it should. The third does not move at all. And the last therefore grows.

Structure (Hedging harder converts one risk into another)**.** The flat column is the content, and its consequence is not the obvious one.

Discretisation error is a sampling error: it comes from applying a correct hedge ratio at finitely many times, and trading more often reduces it without limit. Model error is a *bias*: it comes from applying a wrong hedge ratio, and trading more often applies the wrong ratio more often. Nothing about frequency touches it.

So a desk that improves its hedging operation does not reduce its risk proportionally. It converts discretisation risk into a floor, and the floor is set by the model. At fifty rebalances a year the mis-specification is a minor addition to the total; at a thousand it is nearly all of it. The better the operation, the larger the share of what remains that is nobody's decision.

This gives a second and less familiar reason to stop rebalancing. The usual one is transaction costs, which trade off against discretisation error and produce an optimal frequency in the manner of Leland. The reason here is different and applies even with free trading: past the point where model error dominates, additional rebalancing buys a reduction in a term that is no longer the binding one. A desk hedging continuously against a surface it has marked to the nearest vol point is being precise about the smaller half of its problem.

[3] `risk::hedge_risk_sources`.

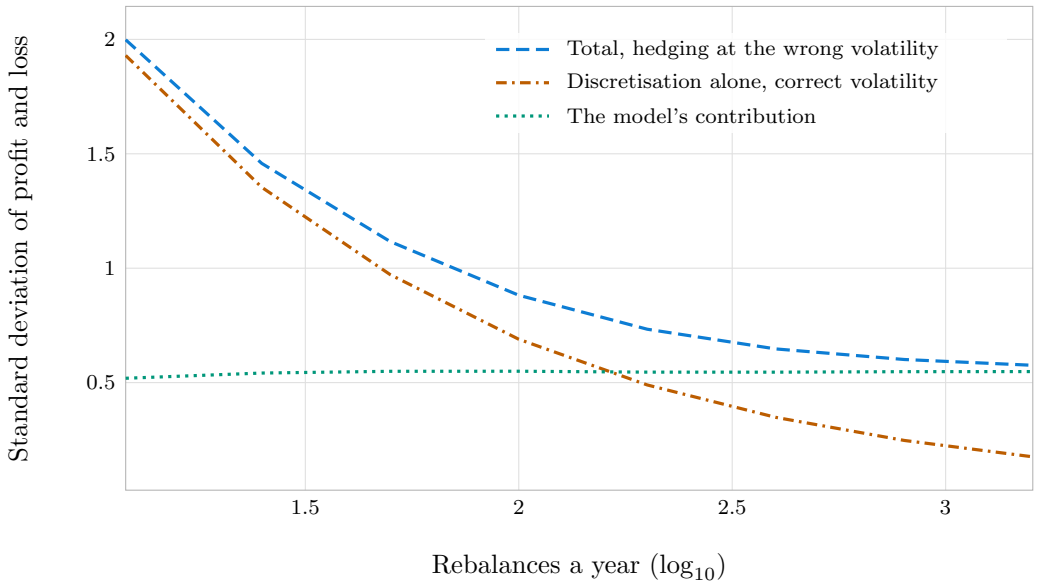

Figure 24.2: The same experiment across a range of rebalancing frequencies, on a logarithmic scale. The discretisation curve falls without limit, as the square root of the frequency; the model's contribution is flat, because it is a bias and not a sampling error. The two cross at about a hundred and sixty rebalances a year — roughly two a week — and past the crossing the total is set by the model rather than by the trading. A desk can move the falling curve by working harder and cannot move the flat one at all.

Drawn by `hedge_risk_sources`.

Remark (What a risk report on this book would say)**.** The number a risk system produces for the position above is the first column. It is computed from the book's stated sensitivities, which are the model's sensitivities, so it measures the risk the model believes the book has. The truth is the second column, and at a realistic frequency the report understates it by a factor of 1.6.

The understatement is not a modelling refinement. It is the difference between the risk of the position the desk thinks it holds and the risk of the position it holds, and no choice of ρ repairs it — expected shortfall computed from a wrong delta is a coherent measure of the wrong thing. That is why §24.10 treats model risk as a separate line rather than as an adjustment to this one.

Remark (The three sources across the book)**.** Calculation 24.7 isolates the model term on a position with only one. In a real book each of the three has an identifiable origin, and the point of naming them is that they are reduced by different actions.

Source	**Where it came from**	**How it is reduced**
chosen	a view, sized deliberately	by changing the view
flow	chapter 23's client requests	by quoting differently
unhedged by decision	the partial hedge, cost against variance	by paying the cost
discretisation	finite rebalancing	by trading more often
model	a wrong hedge ratio	not by any of the above

The last row is the one with no remedy in its own column, and there are two ways it arrives. One is a mis-specified parameter, as measured above. The other is chapter 19's: a valuation whose numerical error cannot be bounded, which has to be carried as model risk because a reserve requires a number and none is available. Both end in the same place, which is a provision made by judgement rather than by calculation.

24.9 Backtesting

A risk number is a prediction, and predictions can be scored. If the model is right, a loss exceeding VaR_α should happen on a fraction $1-\alpha$ of days, independently from day to day.

So count. Over n days, the number of exceptions N should be $\mathrm{Binomial}(n, 1-\alpha)$, and one can test that directly: at $\alpha = 99\%$ over 250 days the expected count is 2.5, and seeing eight is a one-in-a-thousand event under the null. This is the basis of the traffic-light system regulators use, in which too many exceptions raises a bank's capital multiplier.

Two subtleties matter here.

The first is that independence matters as much as the count. Exceptions arriving in a cluster — four in one week and none for a year — is consistent with the right total and inconsistent with the model, because it says the model is missing the volatility clustering that produced them.

The second is a genuine awkwardness with expected shortfall, and it takes a definition to state.

Definition 24.8 (Elicitable)**.** A statistic T of a distribution is *elicitable* if there is a scoring function $S(x, y)$ — a penalty for having forecast x and seen y — whose expected value is minimised by the truth:

$$T(F) \;=\; \operatorname*{argmin}_{x} \; \mathbb{E}_{Y\sim F}\big[S(x, Y)\big].$$

The idea is more familiar than the name. Squared error $S(x, y) = (x - y)^2$ is minimised in expectation by the mean, which is why least squares estimates a mean; absolute error is minimised by the median. For a quantile the corresponding penalty is the asymmetric one

$$S_\alpha(x, y) \;=\; \big(\mathbf{1}\{y \le x\} - \alpha\big)\,(x - y), \tag{24.4}$$

which charges α per unit of shortfall when the outcome lands above the forecast and $1 - \alpha$ per unit when below, and is minimised exactly at the α quantile. So Value at Risk is elicitable, with (24.4) the score.

Remark (Why that matters)**.** Two things follow from possessing such a score, and both are practical.

Forecasts become comparable. Each day a desk states its number, the outcome arrives, and (24.4) converts the pair into a penalty; averaging over a year ranks two models, or two banks, on a single figure. Without a scoring function "their risk numbers were better than ours" is not a well posed comparison.

And honesty becomes optimal. Because the truth minimises the expected score, a forecaster who believes one number and reports another expects to be penalised for it. A regulator scoring submissions this way does not have to ask anyone to be candid.

Expected shortfall has no such function, and the obstruction is structural rather than a failure to find one. An elicitable statistic must have convex level sets: if two distributions share a value of the statistic, so must every mixture of them, since the minimiser of an average of the two expected scores has to be the common minimiser. Quantiles pass this — if F and G both put mass α below x then so does any mixture — and expected shortfall fails it, because it depends on where the quantile sits as well as on the mass beyond it, and mixing moves the quantile.

What is available instead is that the *pair* is jointly elicitable: there is a scoring function of a Value at Risk forecast, an expected shortfall forecast and the outcome, minimised at the true pair together. So the measure can be backtested, and Basel's move to expected shortfall is not unbacktestable in practice. What is unavailable is attributing the score: a joint penalty says the pair was good or bad and does not decompose into a verdict on the tail average alone.

Structure (Two desiderata, pulling opposite ways)**.** This could not have been avoided by choosing more carefully because the two properties are about different things.

Coherence, in §24.2, is a property of the measure *as a summary of risk*: it asks whether the number behaves sensibly when portfolios are combined, and subadditivity is the part Value at Risk fails. Elicitability is a property of the measure *as a forecast*: it asks whether a prediction of the number can be scored against what happened, and it is the part expected shortfall fails.

Nothing connects them, and the two measures land on opposite sides of both. Value at Risk can be verified and cannot be aggregated; expected shortfall can be aggregated and cannot be verified on its own. The regulatory move from one to the other bought the diversification property at the cost of the single-measure backtest, and calling it a strict improvement misdescribes it. It was a choice about which of two failures a capital regime could better live with — and, given that the alternative was a measure telling banks that splitting a position reduces its risk, a defensible one.

24.10 Model Risk, and Turning It Into a Number

Four of the modelling chapters each ended in the same place. Chapter 10 found that the smile does not determine β, and β determines the hedge. Chapter 11 found that the vanilla surface does not determine the mixing weight, and the mixing weight determines the exotics. Chapter 12 found that European swaptions do not determine the mean reversion, and the mean reversion determines what a Bermudan is worth. Chapter 13 found that a hundred or so liquid swaptions do not determine a correlation matrix with hundreds of entries, and the correlation determines every product written on more than one rate.

In each case the calibration succeeded perfectly and left the answer undetermined. That is model risk in its purest form, and it is not addressed by calibrating harder.

What is done instead is to price the book across the range of the undetermined parameter and reserve the difference.

1. Fix a plausible range for the parameter — from history, from a related market, or by policy.
2. Revalue the portfolio at each end.
3. Take the difference as a reserve, held against the marked value.

This is what prudent valuation adjustments are, and the discipline it imposes is worth more than the number: it forces someone to write down what the range is and to defend it, which turns an invisible assumption into a line item that a committee can argue about.

Remark. The reserve should be computed at the level of the book, not the trade. Model risk is not additive — two trades whose exposure to β offsets have no joint model risk even though each has plenty — and computing it trade by trade and adding produces a number that is both wrong and, by the argument of Definition 24.1, incoherent in exactly the way Value at Risk is.

24.11 Counterparty Risk

Finally, the risk that the trade was right and the money did not arrive.

The exposure at a future date is what is owed to us, which is the value of the trade if positive and nothing if negative — so it is an option on our own portfolio, $\max(V_t, 0)$. Averaging over scenarios gives the expected exposure $\mathrm{EE}(t)$, and the credit valuation adjustment is that exposure weighted by the probability of default and the loss given it:

$$\mathrm{CVA} = \mathrm{LGD} \int_0^T \mathrm{EE}(t)\, d\,\mathbb{P}(\tau \le t).$$

Two observations connect this back.

The first is that CVA has volatility risk. A desk that hedges it holds vega against its own counterparties' exposure profiles, which is why counterparty risk ended up staffed by derivatives quants rather than by credit analysts.

The second is that this is where chapter 7's discussion of collateral pays off. A fully collateralised trade has an exposure that is reset to zero continuously, so the integral above nearly vanishes and the adjustment is small. That is the sense in which the collateral agreement was doing real work: it

converts a credit problem into a funding problem, and chapter 7 showed that the funding problem is solved by discounting on the collateral rate. Uncollateralised, the credit problem stays, and its price is this integral.

24.12 One Model or Many

A real book is priced by several models at once. Chapter 12's Bermudans sit in one, caps and floors in another, a CMS in the static replication of chapter 15 with no dynamic model at all, a callable bond behind the option-adjusted spread of chapter 7. Each was chosen because it fits the instruments that product is hedged in. The question this raises is whether the book as a whole is then incoherent, and whether pricing everything in one rich model would fix it.

The answer has two halves, and they point in opposite directions.

What multiple models genuinely break

Three things, and only the third is usually anticipated.

Internal arbitrage. Two models calibrated to the same vanillas disagree about an exotic — that is the whole content of §24.10. If two desks price adjacent products with different models, there is a trade between them that both books mark as profitable. Nothing has been created; the firm has written down the difference between two models twice, with opposite sign, and called it revenue. This is not hypothetical and it is the strongest practical argument for a single valuation model within any set of products that trade against each other.

Joint distributions rather than marginals. Chapter 18 is the general statement: agreeing on every marginal constrains the joint distribution not at all. Two models that price every instrument in the book correctly on their own can still disagree completely about what happens to the book, because a portfolio's value depends on the joint law of its inputs. A netted position that one model says is flat is not flat under the other.

The explain report absorbs the inconsistency. §24.7 decomposes the day's profit into factor moves and a residual. With one model the residual is model error plus hedging error. With several, it also contains the disagreement between them, and those three are no longer separable — which removes the diagnostic that the report exists to provide.

What does not break, and why

First-order risk aggregation.

The reason is §24.5's observation read forwards. A risk system does not store models or parameters; it stores sensitivities to *traded instruments* — bucketed curve deltas against the calibrating swaps, a vega grid indexed by expiry and tenor, sensitivities to the basis quotes of chapter 7. Those are derivatives of value with respect to observable prices, and derivatives with respect to the same quantity add up regardless of what computed them. A Bermudan's vega to the ten year ten year swaption and a cap's vega to the same instrument net exactly, whatever models produced them, because both are answers to the same question about the same price.

So the common language is the market, not the model. That is the same principle as key rate durations being reported against the instruments rather than against the curve's internal parametrisation, and it is why a multi-model book is not the operational disaster it sounds like.

Where a single model is genuinely required

Two calculations cannot be assembled this way, and they are the reason global models were built.

The first is portfolio-level scenario analysis. Revaluing a book under a joint move of the curve and the volatility surface requires a consistent way of moving every input together, which is a joint distribution over inputs — and a joint distribution over inputs is a model. Sensitivities cannot supply it, because the whole point of a stress is that it is large enough for the Taylor expansion to fail.

The second is counterparty exposure. The expected exposure of the previous section is

$$\mathrm{EE}(t) = \mathbb{E}\big[\max(V_t, 0)\big],$$

where V_t is the value of the *whole netting set* at a future date. The maximum does not distribute across trades, so this cannot be computed trade by trade and summed; every trade in the set must be simulated forward under one measure, on one set of paths. That is a hard requirement rather than a preference, and it is what forced the industry to build global models after 2008 — not a desire for elegance, but the fact that a netted exposure is not a sum of exposures.

Structure (The two requirements want different models)**.** Put the two halves together and the apparent conflict dissolves, because the calculations that need local accuracy and the calculations that need global consistency are not the same calculations.

Pricing and marking need to be exactly right on the instruments the product is hedged in, and need to be right about that product's own second-order behaviour. They tolerate having nothing to say about the rest of the book, because the hedge is executed in that product's own market.

Exposure, capital and stress need one consistent measure across everything and tolerate being coarse. An expected exposure profile is an average of an average; a stress is a scenario chosen by judgement to within a wide band. Neither is improved much by a model that prices each trade to the last basis point, and chapter 19's argument applies with force — a global model rich enough to price everything well has more parameters than the market identifies, so its extra richness buys consistency of an arbitrary choice rather than accuracy.

So the architecture that desks actually run — a model per product for pricing, and one coarser model for the whole netting set — is not a compromise between two ideals. It is the correct answer, because the two jobs have different requirements and no single model is best at both. The failure mode is not having two systems; it is using each for the other's job: marking a book off the exposure model, or computing a netted exposure by adding per-trade sensitivities.

Remark (Consistency is not correctness)**.** One argument for the single global model deserves to be answered directly, because it is the one usually made and it is weaker than it sounds.

A single model does give consistent hedges. What it does not give is correct ones. If the model is wrong — and §24.10 is a list of parameters the market does not determine — then every hedge in the book is wrong in the same direction, and the book has no way of finding out. Two models disagreeing is unpleasant, but the disagreement is a *measurement*: it is the model reserve of §24.10, arrived at for free by having built the second model. Unifying on one model does not remove that risk. It removes the instrument that was detecting it.

There is a second, quieter cost. A model calibrated to every market at once fits none of them exactly, so the misfit is spread across all products rather than concentrated where it can be reserved against. The Bermudan desk's model that fits swaptions exactly has a known, bounded error —

it prices its hedges right and its exotic wrong by an amount §24.10 can bound. The global model prices everything slightly wrong, including the hedging instruments, and the error in the hedge is the one that costs money continuously rather than once.

The reasonable position is therefore narrower than either slogan. Unify within a set of products that trade against each other, because internal arbitrage is real. Do not unify across sets that do not, because the consistency bought is between quantities nobody trades, and it is paid for with accuracy on quantities everybody does.

References

- Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.

- Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7), 1487–1503.

- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

- Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4), 1680–1707.

- Gregory, J. (2020). *The xVA Challenge*. Wiley.