The Mathematics of Risk Measurement

By Malcolm Kemp, 14 March 2020

1. Introduction

The purpose of this note is to provide a summary of the main mathematical components underlying risk measurement. Some of the mathematical ideas explored in this note are illustrated numerically in spreadsheets available via the Nematrian website. Readers should bear in mind that risk and uncertainty are only partly quantifiable and therefore only partly amenable to mathematical analysis. Readers interested in more general material on risk measurement and management may wish to refer to the enterprise risk management pages of the Nematrian website, at www.nematrian.com/erm.aspx.

This note does not in the main address how we might select between alternative approaches to managing risk. A mathematical treatment of this problem generally involves some form of portfolio optimisation or related technique. Nor does it seek to cover derivative pricing theory in any great depth or to cover how such theory might be implemented in practice, e.g. via Monte Carlo simulation techniques.

2. Portfolio risk measures

2.1 Value-at-Risk

Several types of (quantitative) measures of risk are used in the financial industry and increasingly in other sectors of the economy. Perhaps the best known is Value-at-Risk or VaR. For a portfolio (of risks, investments, …) it is the loss which will be exceeded on some fraction, α, of occasions if the portfolio is held for a given length of time, i.e. for a given time horizon, say T.

Suppose a portfolio consists of monetary amounts $\mathbf{a} = (a_1, \ldots, a_n)^T$ invested in n exposures. Let x_i be the loss (i.e. negative payoff) on the i'th exposure, and $\mathbf{x} = (x_1, \ldots, x_n)^T$. Let $L = \mathbf{a}.\mathbf{x} = \sum_{i=1}^{n} a_i x_i$ be the total portfolio loss.

Definition 2.1: *For a portfolio with total losses over a holding period T equal to a random variable L, the Value-at-Risk with confidence level α ($0 < \alpha < 1$), denoted VaR_α is defined as:*

$$VaR_\alpha = \inf\{z: Pr(L \geq z) \leq \alpha\}$$

For a continuous distribution VaR_α is implicitly defined by

$$Pr(L \geq VaR_\alpha) = \alpha$$

Or, if the probability density function (pdf) of payoff X is $p(x)$ (remember losses are negative payoffs) then VaR_α is defined implicitly as the value k such that:

$$VaR_\alpha(X) = k \quad where \int_{-\infty}^{-k} p(x)dx = 1 - \alpha$$

Points to note include:

(a) VaR_α is mathematically equivalent to the $(1-\alpha)$-quantile of the payoff distribution, or in mathematical notation $VaR_\alpha(X) = -F^{-1}(1-\alpha)$ where $F^{-1}(x)$ is the quantile function, also called the inverse cumulative distribution function or just the inverse function of the distribution with density $p(x)$. This indicates that when we are estimating and using VaR_α we may draw on an extensive body of statistical knowledge relating to quantile estimation.

(b) We've given above the usual definitions of VaR_α but sometimes the sign is flipped and/or α is replaced by $1-\alpha$ (so if someone refers to a 99% VaR and someone else refers to a 1% VaR then they may be referring to the same thing as both are likely to be referring to the *downside* tail).

(c) VaR_α has a natural connection with the amount of capital a firm should hold. Capital is (usually) defined as the excess of assets over liabilities and nowadays the tendency is to value assets and liabilities in such calculations by reference to economically relevant market values rather than, say, book or purchase costs. If a firm holds (market-value) capital equal to VaR_α (calculated for a holding period of T) then it should experience losses exceeding its capital with probability α (if it does not alter its portfolio or its asset or liability bases in the meantime).

(d) Strictly speaking we need some axioms to apply for the mathematics underlying these computations to work. In particular, we need to assume that the numerical value we ascribe to a loss satisfies uniqueness, additivity and scalability, i.e. here that L is well defined (in the above we are referring to the value that we place on the loss, typically its monetary value), and that if we have two losses x_1 and x_2 then $k(x_1 + x_2) = kx_1 + kx_2$

(e) In nearly all cases mathematical risk measurement theory concentrates on the *market value* of the exposures or some reasonable economic proxy. This is partly because such values, if suitably defined, should adhere to the axioms in (d) if we adopt the principle of no arbitrage. A corollary is that a good broad understanding of how to make values placed on exposures market consistent is very important for effective (financial) risk measurement and management. More specifically, a good broad understanding of option pricing theory is generally important if the types of exposures present include material optionality.

(f) We may conceptually split VaR_α into two parts, the expected loss, $EL = E(L)$, and the unexpected loss, $UL = VaR_\alpha - EL$. Some commentators argue that, say, banks should only hold capital equal to the *UL* on the grounds that the expected losses on, say, a bank's loan portfolio should be offset by anticipated profit margins included in loan pricing. The potential flaw in this logic is that firms do not necessarily estimate *EL* correctly (or necessarily price the loan 'correctly' in relation to the EL even if they have estimated EL accurately, e.g. their pricing may be driven by market forces). Also, the *EL* may change through time (as economic conditions change) but loan rates may not move in tandem, and some capital is potentially required to protect against this risk.

Ways in which financial firms (especially banks) use VaR_α include:

(i) In the context of decisions about appropriate levels of capital. Judging the right amount of capital for a firm implicitly involves calculating how much capital might be wiped out with a given probability, so is closely allied with VaR_α.

(ii) Working out appropriate measures of remuneration for dealers

(iii) Managing risk

(iv) Conveying information to the market about the riskiness of a firm's operations

(v) As part of the supervisory process that financial firms are typically nowadays subject to.

2.2 Relative VaR

Although VaR is usually defined by reference to (monetary) losses many types of risk measurement, particularly in the investment management sector, actually involve losses relative to some benchmark outcome. The 'loss' measured by the relevant VaR may then be reexpressed relative to the benchmark and the VaR in question may then be referred to as a 'relative' VaR if there is a wish to distinguish it from an 'absolute' VaR that is not benchmark related. Even an 'absolute' VaR for say a bank (particularly in a regulatory context) is actually normally a 'relative' VaR, since the VaR being calculated usually relates to movements in Assets minus Liabilities, rather than merely Assets in isolation, or is dependent on a particular monetary *numeraire* (e.g. dollars rather than euros) so is 'relative' to that numeraire.

If we are measuring returns relative to those generated by a benchmark, $\mathbf{b} = (b_1, \dots, b_n)^T$ then we might defined the *relative* VaR_α by reference to losses L given by $L = (\mathbf{a} - \mathbf{b}).\mathbf{x} = \sum_{i=1}^{n}(a_i - b_i)x_i$ where the losses per unit exposure of each underlying instrument are $\mathbf{x} = (x_1, \dots, x_n)^T$.

However, it is worth noting that relative returns can be calculated mathematically in a variety of ways. The simplest is merely the *arithmetic* difference between the two returns, i.e. $r(\mathbf{a}) - r(\mathbf{b})$ as above. But returns compound geometrically rather than arithmetically (which is one reason why returns are often assumed to follow a geometric Brownian motion rather than an arithmetic Brownian motion). So for non-infinitesimal time period lengths there are alternative and potentially preferable ways of defining relative returns. These include (if we are trying to calculate the return r_1 relative to r_2, each expressed as fractions) geometric relative returns, i.e. $r_{geometric\ relative} = (1 + r_1)/(1 + r_2) - 1$, or logarithmic relative returns, i.e. $r_{logarithmic\ relative} = \log(1 + r_1) - \log(1 + r_2)$, rather than arithmetic relative returns $r_{arithmetic\ relative} = r_1 - r_2$.

2.3 VaR for normally distributed random losses

Suppose that losses are distributed according to a normal distribution $L \sim N(\mu, \sigma^2)$. Then we have:

$$VaR_\alpha(L) = \mu + \sigma N^{-1}(1 - \alpha)$$

where $N^{-1}(q)$ is the inverse (standard) cumulative normal distribution function.

If losses for given unit exposures are multivariate normally distributed, i.e. $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$, and if the exposures per unit instrument are $\mathbf{a} = (a_1, \dots, a_n)^T$ then $L = \mathbf{a}.\mathbf{x} \sim N(\mu, \sigma^2)$ where $\mu = \mathbf{a}.\boldsymbol{\mu}$ and $\sigma^2 = \mathbf{a}^T\mathbf{V}\mathbf{a}$.

2.4 Marginal VaR, the Euler capital allocation principle and Incremental VaR

For some of the purposes mentioned in 2.1 above, it is desirable to apportion capital between exposures, i.e. to identify how individual risks contribute to the overall Value-at-Risk. The issue is that Value-at-Risk is not additive so some more sophisticated apportionment approach is required. Usually this involves *marginal Value-at-Risk*.

Suppose we have the same total loss as per definition 2.1, i.e. $L = \mathbf{a}.\mathbf{x} = \sum_{i=1}^{n} a_i x_i$ where the amount of the i'th exposure is a_i and the loss arising from a unit amount of the i'th exposure is x_i.

Definition 2.2: *The marginal Value-at-Risk (with confidence level α and time horizon T) for the i'th exposure is denoted $MVaR_{\alpha}^{(i)}$ where:*

$$MVaR_{\alpha}^{(i)} = \frac{\partial}{\partial a_i}\left(VaR_{\alpha}\left(\sum_i a_i x_i\right)\right)$$

As risks arising from individual positions interact there is no universally agreed way of subdividing the overall risk into contributions from individual positions. However, a commonly used way is to define the *Contribution to Value-at-Risk*, c_i, of the i'th position, a_i to be as follows:

$$c_i = a_i MVaR_{\alpha}^{(i)}(\mathbf{a})$$

Conveniently the c_i then sum to the overall VaR:

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} a_i MVaR_{\alpha}^{(i)}(\mathbf{a}) = \sum_{i=1}^{n}\left(a_i\mu_i + N^{-1}(1-\alpha)\frac{1}{\sigma}\left(a_i\sum_{j=1}^{n} a_j V_{ij}\right)\right)$$

$$\Rightarrow \sum_{i=1}^{n} c_i = \mathbf{a}.\boldsymbol{\mu} + N^{-1}(1-\alpha)\frac{\sigma^2}{\sigma} = \mathbf{a}.\boldsymbol{\mu} + \sigma N^{-1}(1-\alpha) = VaR_{\alpha}(\mathbf{a})$$

This is a special case of a more general result called Euler's capital allocation principle that applies to any risk measure that is homogeneous (of order 1), including VaR.

Definition 2.3: *a function $f(u_1, \ldots, u_n)$ is said to be homogeneous of order q (constant) if it satisfies:*

$$f(ku_1, \ldots, ku_n) = k^q f(u_1, \ldots, u_n)$$

More specifically, a function $f(u_1, \ldots, u_n)$ is said to be homogeneous (or homogeneous of order 1) if it satisfies:

$$f(ku_1, \ldots, ku_n) = kf(u_1, \ldots, u_n)$$

Suppose a firm has used a risk model to calculate its overall required economic capital. The Euler Principle is one general way in which this can be translated in a fair way into economic capital for individual business units or risk categories for any risk measure that is homogeneous of order 1. Suppose there are n units with associated loss variables $L_1, \ldots, L_n$ and total loss $L = L_1 + \cdots + L_n$. A coherent risk measure ρ (see section 2.9 for further details of coherent risk measures) will satisfy positive homogeneity, i.e. $\rho(kL) = k\rho(L)$ for any $k > 0$. More generally, a function of n variables that is homogeneous of order q is one that satisfies the relationship $f(ku_1, \ldots, ku_n) = k^q f(u_1, \ldots, u_n)$ for some constant q. Euler's homogeneous function theorem states that such a function satisfies the

following (where the vertical bar and subscript means that each partial derivative is evaluated at $(ku_1, \dots, ku_n)$ etc.:

$$u_1 \frac{\partial f}{\partial u_1}\Big|_{(ku_1,\dots,ku_n)} + \cdots + u_n \frac{\partial f}{\partial u_n}\Big|_{(ku_1,\dots,ku_n)} = qk^{q-1} f(u_1, \dots, u_n)$$

If there are p_i units of each loss L_i and $L(p) = p_1 L_1 + \cdots + p_n L_n$ then we may express the risk measure as $r(p) = \rho\big(L(p)\big)$. If ρ is homogeneous then r is also homogeneous so:

$$\rho(L) = r(1) = \frac{\partial r}{\partial p_1}\Big|_{p=1} + \cdots + \frac{\partial r}{\partial p_n}\Big|_{p=1}$$

where $p = 1$ means $(p_1, \dots, p_n) = (1, \dots, 1)$.

Definition 2.4: *a capital (or risk) allocation that satisfies the Euler (capital allocation) principle involves subdividing total economic capital, EC, into individual capital amounts for individual business lines/exposures using the following formula:*

$$EC_i = \frac{\partial r}{\partial p_i}\Big|_{p=1}$$

Sometimes attention is focused on an alternative measure called incremental VaR that does not (in general) add to the total portfolio VaR. It is the change in the VaR if the whole of a given position is removed from the portfolio, i.e.

Definition 2.5: *The incremental Value-at-Risk (with confidence level α and time horizon T) for the i'th exposure is denoted $IVaR_\alpha^{(i)}$ where:*

$$IVaR_\alpha^{(i)} = VaR_\alpha\left(\sum_i a_i x_i\right) - VaR_\alpha\left(\sum_{j,j\neq i} a_j x_j\right)$$

Confusingly, a small number of writers define marginal VaR in line with the definition for incremental VaR given above and vice-versa (perhaps because one leading software vendor uses this alternative terminology).

2.5 Marginal VaR for multivariate normally distributed (i.e. Gaussian) losses

If losses for given unit exposures are multivariate normally distributed, i.e. $\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{V})$, and if the exposures per unit instrument are $\mathbf{a} = (a_1, \dots, a_n)^T$ then the marginal VaR for the i'th exposure is:

$$MVaR_\alpha^{(i)}(\mathbf{a}) = \frac{\partial VaR_\alpha(\mathbf{a})}{\partial a_i} = \frac{\partial}{\partial a_i}\left(\mathbf{a}.\boldsymbol{\mu} + N^{-1}(1-\alpha)\sqrt{\mathbf{a}^T\mathbf{V}\mathbf{a}}\right)$$

$$\Rightarrow MVaR_\alpha^{(i)}(\mathbf{a}) = \frac{\partial}{\partial a_i}\left(\sum_{j=1}^{n} a_j\mu_j\right) + N^{-1}(1-\alpha)\frac{1}{2\sqrt{\mathbf{a}^T\mathbf{V}\mathbf{a}}}\frac{\partial}{\partial a_i}\left(\sum_{j=1}^{n}\sum_{k=1}^{n} a_j V_{jk} a_k\right)$$

$$\Rightarrow MVaR_\alpha^{(i)}(\mathbf{a}) = \mu_i + N^{-1}(1-\alpha)\frac{1}{\sigma}\left(\sum_{j=1}^{n} a_j V_{ij}\right)$$

The last part of this equation can be expressed in terms of the correlation between x_i and $\mathbf{a}.\mathbf{x}$ as follows. Suppose we view the x_i as corresponding to time series $x_{i,t}$ with T elements (which without loss of generality can be assumed to be de-meaned, i.e. to have their means set to zero) and $\mathbf{a}.\mathbf{x}$ as corresponding to a time series $y_t = \sum_{i=1}^{n} a_i x_{i,t}$. Then the correlation between x_i and $\mathbf{a}.\mathbf{x}$ would be:

$$Correlation(x_i, \mathbf{a}.\mathbf{x}) = \frac{\sum_{t=1}^{T} x_{i,t} y_t}{\sqrt{\sum_{t=1}^{T} x_{i,t}^2 \sum_{t=1}^{T} y_t^2}}$$

We also have:

$$V_{ij} = \sum_{t=1}^{T} x_{i,t} x_{j,t}$$

$$\sum_{t=1}^{T} x_{i,t} y_t = \sum_{t=1}^{T} x_{i,t} \sum_{j=1}^{n} a_j x_{j,t} = \sum_{j=1}^{n} a_j \sum_{t=1}^{T} x_{i,t} x_{j,t} = \sum_{j=1}^{n} a_j V_{ij}$$

$$\sum_{t=1}^{T} x_{i,t}^2 = V_{ii}$$

$$\sum_{t=1}^{T} y_t^2 = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j a_k \sum_{t=1}^{T} x_{i,t} x_{j,t} = \sum_{j=1}^{n} \sum_{k=1}^{n} a_j V_{jk} a_k = \mathbf{a}^T \mathbf{V} \mathbf{a} = \sigma^2$$

$$\Rightarrow Correlation(x_i, \mathbf{a}.\mathbf{x}) = \frac{\sum_{j=1}^{n} a_j V_{ij}}{\sqrt{V_{ii}}\sigma}$$

$$\Rightarrow \sum_{j=1}^{n} a_j V_{ij} = Correlation(x_i, \mathbf{a}.\mathbf{x})\sqrt{V_{ii}}\sigma$$

$$\Rightarrow MVaR_{\alpha,i}(\mathbf{a}) = \mu_i + N^{-1}(1-\alpha) Correlation(x_i, \mathbf{a}.\mathbf{x})\sqrt{V_{ii}}$$

2.6 Tail Value-at-Risk

Whilst VaR may be one of the more common ways of measuring risk, it is by no means the only one. A commonly proposed alternative is Tail Value-at-Risk. Analogous to marginal VaR it is also possible to define marginal TVAR (see here for details of its calculation for multivariate normally distributed variables).

Definition 2.6: *The Tail Value-at Risk or $TVaR_\alpha$ for a given confidence level α and time horizon T is defined as*:

$$TVaR_\alpha = E(L | L \geq VaR_\alpha)$$

Or, if the pdf of payoff X is $p(x)$ (remember again losses are negative payoffs) and $p(x)$ is continuous then $TVaR_\alpha$ is:

$$TVaR_\alpha(X) = E(-X | X \leq -VaR_\alpha) = -\frac{1}{1-\alpha} \int_{-\infty}^{-VaR_\alpha} x p(x) dx$$

The tail VaR is also sometimes called the Conditional VaR (because it involves a conditional probability). Occasionally TVaR (less commonly CVaR) is ascribed the same meaning as expected shortfall, see below, in which case the $1/(1-\alpha)$ factor is ignored, or is defined relative to some specific limit – k that in effect defines the α to be used in the above formula.

Whilst VaR has been the industry standard risk measure for some time now (at least in the banking industry, see below regarding the asset management industry), there seems to be some regulatory drive towards greater use of Tail VaR in the future. Reasons why TVaR may be preferred for this purpose include:

(a) VaR provides no guidance on how severe losses might be *beyond* the VaR cut-off point. This is potentially particularly important for some stakeholders (such as regulators, supervisors, customers and governments). If the VaR cut-off point is set at a level comparable with the point at which the firm defaults then losses up to the VaR cut-off point will (we might argue) be borne by shareholders. It is only when losses start to exceed this cut-off that costs fall to these wider stakeholders. So VaR in this sense can be viewed as overly shareholder-focused and insufficiently sensitive (as far as some other stakeholders are concerned) to magnitude of loss beyond the VaR cut-off.

(b) VaR does not in general exhibit desirable features we might expect a risk measure to exhibit in relation to diversification. In particular, it does not in general satisfy *sub-additivity* (one of the four requirements a risk measure needs to exhibit for it to be *coherent*, see below). For example, suppose there are two portfolios. One is (only) exposed to one risk that has a 0.3% chance of occurring and if it does then it will lose £1m. The other is exposed to five independent risks each of which has a 0.3% chance of occurring and each involves a loss of £0.2m. Then we would 'expect' a risk measure to show the second portfolio to be less risky than the first one, whereas the 99.5% VaR of the second portfolio is £0.2m which is *more than* the VaR of the first portfolio (which is 0 because its risk has a likelihood of occurrence less than 0.5%).

2.7 Expected Shortfall and Expected Policyholder Deficit

Another commonly used risk measure is:

Definition 2.7: *The expected shortfall, $ES(Q)$, usually with $Q = 0$ is normally defined in a manner akin to Definition 2.7 for expected policyholder deficit*:

$$ES(Q) = -\int_{-\infty}^{Q} xp(x)dx = EPD(Q)$$

Or more generally the expected shortfall below some trigger level Q is:

$$ES(Q) = -\int_{-\infty}^{Q} xp(x)dx = EPD(Q) + Q$$

However, some commentators define ES in a manner equivalent to the definition in Definition 2.6 for Tail Value-at-Risk.

Less commonly used (and then only in the context of insurance) is the following risk measure:

Definition 2.8: *Given an initial net worth or capital of W, the expected policyholder deficit, $EPD(W)$, is defined as*:

$$EPD(W) = -E\big((X-W)I(X<W)\big) = -\int_{-\infty}^{W}(x-W)p(x)dx$$

where $(X<W) = \begin{cases}1, & X<W\\ 0, & X\geq W\end{cases}$. Alternative notations for $I(x<W)$ are $I\{X<W\}$, $I_{X<W}$ $1\{X<W\}$ and $1_{X<W}$.

2.8 Tracking error

Although VaR and variants such as TVaR are probably the most commonly used risk measures in the banking world this is less true in the asset management world. Here, a particularly common risk measure is (ex-ante) tracking error.

Definition 2.9: *If X is a random variable (e.g. a portfolio return) with (assumed forward looking) pdf $p(x)$ then its ex-ante tracking error (if it exists) is σ where $\sigma^2 = var(X)$, i.e. the variance of the forward looking return*.

One reason some commentators do not like tracking error is that they assume that it is defined as a backward-looking statistic referring purely to how returns on a portfolio have behaved in the past, which is not necessarily a relevant risk measure for what might happen in the future (particularly if the portfolio positioning has changed materially). Such a tracking error is called an *ex-post* tracking error to differentiate it from an *ex-ante* tracking error which should refer to some assumed probability distribution for how portfolio returns might behave *in the future*.

The same distinction also technically arises with VaR; we can in principle refer to an ex-post VaR akin to an ex-post tracking error as well as an ex-ante VaR akin to an ex ante tracking error. However, ex-post VaRs are less commonly calculated or quoted (except to back test ex-ante VaR models), so the potential for confusion is less.

Implicit in a focus on (ex-ante) tracking error is the view that we should not when monitoring the risk characteristics of an actively managed portfolio take credit for any assumed (expected) outperformance the manager might deliver.

Tracking error also has a nice intuitive geometrical analogy. This arises because the formula for the (ex-ante) tracking error of the sum of two sets of exposures, i.e.:

$$\sigma^2_{\mathbf{a+b}} = \sigma^2_{\mathbf{a}} + 2\sigma_{\mathbf{a}}\sigma_{\mathbf{b}}corr(\mathbf{a},\mathbf{b}) + \sigma^2_{\mathbf{b}}$$

has a natural analogy with the relationship between the lengths of sides of a triangle two of whose sides are formed by vectors $\mathbf{a}$ and $\mathbf{b}$ if $corr(\mathbf{a},\mathbf{b})$ is associated with $\cos\theta$ where θ is the angle between these two vectors.

2.9 Coherent risk measures

We measure risks to help us manage them. Implicit in any risk management is selection between alternatives. Mathematically this can (usually) be framed as involving utility maximisation; we aim to select the 'best' strategy, given some suitable definition of 'best'.

One way of defining 'best' in this context might be that it involves the lowest value for some specific risk measure, e.g. VaR. If VaR is being used for this purpose then its undesirable characteristics regarding diversification noted above become problematic. Partly to resolve this sort of issue, Artzner et al. (1999) developed a set of 'reasonable' axioms that it was desirable for risk measure to exhibit. Risk measures that exhibit these axioms are called *coherent*.

Definition 2.10. *A risk measure $r(x)$ is coherent if it satisfies the following 4 axioms:*

(1) *Subadditivity*: for any pair of loss random variables, x_1 and x_2:

$$r(x_1 + x_2) \leq r(x_1) + r(x_2)$$

(2) *Monotonicity*: if $x_1 \leq x_2$ for all states of the world then:

$$r(x_1) \leq r(x_2)$$

(3) *Homogeneity*: for any constant $\lambda \geq 0$ and random losses x:

$$r(\lambda x) = \lambda r(x)$$

(4) *Translational invariance*: for any loss random variable x and constant c:

$$r(x + c) = r(x) + c$$

Points to note include:

(a) VaR is *not* in general coherent because it does not in general satisfy subadditivity. However it *is* coherent for normally distributed loss variables or more generally ones coming from a distribution from the elliptical family of distributions

(b) TVaR *is* coherent (if coming from a continuous distribution and if the threshold used is set appropriately, i.e. in line with the VaR).

(c) (Ex-ante) tracking error is *not* coherent because it does not satisfy translational invariance (it does not change if the loss variable is shifted in all states of the world by the same constant c). However, this does not in general present a problem when it comes to the mathematics of selecting between alternative strategies. It just means that the way we define utility needs to include a return component as well as a risk component, instead of both in effect being bundled up into a single overall 'risk' measure.

To demonstrate coherence of TVaR, Artzner et al. (1999) proved the following result:

Theorem 2.1. *Assume that the loss random variable x is defined on a sample space Ω. A risk measure $r(x)$ is coherent if and only if there exists a family Q of probability measures defined on Ω such that:*

$$r(x) = \sup_P \{E_P(x) | P \in Q\}$$

Here $E_P(x)$ is the expected value of x under the probability measure P.

This result can then be used to demonstrate that $TVaR_\alpha$ is coherent (for a continuous distribution). Suppose that x takes n different values $x_1 \leq x_2 \leq \cdots \leq x_n$ and that $\Omega = \{x_1, \ldots, x_n\}$. Let k be an

integer such that $k \leq n(1-\alpha) < k+1$, so $VaR_\alpha(x) = x_{k+1}$. Let B be the set of subsets of Ω that contain $n-k$ elements. For any $A \in B$ define a probability measure P_A as:

$$P_A(\omega) = \begin{cases} \frac{1}{n-k} & if\ \omega \in A \\ 0 & otherwise \end{cases}$$

Then from the above theorem $r(x) = \underset{A \in B}{sup}\{E_P(x)\}$ is a coherent risk measure.

Let K be the element of B consisting of the $n-k$ largest possible values of x, i.e. $\{x_{k+1}, \dots, x_n\}$. Then:

$$TVaR_\alpha(x) = E(x|x \geq VaR_\alpha(x)) = \frac{x_{k+1} + \cdots + x_n}{n-k} = E_{P_K}(x)$$

and for any other $A \in B$, $E_{P_A}(x) \leq E_{P_K}(x)$.

2.10 Strengths and weaknesses of VaR

Perhaps the most important advantage of VaR is that it requires little understanding of probability or statistics to grasp its essential meaning. However, it does have some disadvantages. We have already discussed above technical ones relating to coherence and also whether it gives the 'right' weight to extreme events for particular stakeholders. VaR is also basically a static measure of risk as it assumes a buy and hold approach to positions within the time horizon. It also relies on some assumptions about how markets operate and these can be invalidated when liquidity disappears. So far no one has come up with a really convincing way of incorporating liquidity risk into VaR.

2.11 Value-at-risk for linear claims using past data

A trading book will usually involve relatively rapid turnover of liquid positions. Applying VaR based on recent past market movements to such a book is usually reasonably straightforward as long as the instruments held are linear as we will typically have good time series data on the returns on the different exposures held within the portfolio, i.e. on the $\mathbf{x} = (x_1, \dots, x_n)^T$ and we will also typically have good knowledge of our exposures, i.e. the $\mathbf{a} = (a_1, \dots, a_n)^T$. We can then calculate a time series of total portfolio losses (defined as the negative of the return), i.e. of the the $\mathbf{a}.\mathbf{x}$, and apply univariate statistical techniques to estimate the $(1-\alpha)$-quantile of the loss distribution. The same sorts of techniques can also be used to estimate the $TVaR_\alpha$ but they will then be somewhat more involved.

The main techniques involve:

(a) *Parametric approach*: This involves estimating a parametric distribution for the portfolio return or loss. Two special cases involve:

  (i) Assuming that the distribution is a normal distribution, which some commentators refer to as the 'parametric' approach but is more usually called the 'variance-covariance' approach (over the short term any mean drift will generally be small relative to the contribution from variance or covariance terms); and

  (ii) Using Extreme Value Theory, which we discuss later.

(b) *Non-parametric approach*: This involves estimating quantiles of the return or loss distribution (particularly the $(1-\alpha)$-quantile) directly from quantiles of the empirical distribution of returns (perhaps with some smoothing applied).

Both (a) and (b) involve referring to past data, i.e. involve extrapolating the past into the future. An alternative approach, involving more contemporaneous data, is to estimate the VaR based on market implied variances and covariances.

2.11 Parametric estimation of VaR

One way of estimating the VaR of a portfolio is to fit a parametric distribution to time series data of losses and then to calculate the $(1-\alpha)$-quantile of this distribution. By 'losses' we nearly always mean here the losses that the portfolio would have incurred had its current positioning been replicated in the past rather than what it actually lost (the latter is of little relevance to what might happen in the future if the portfolio positioning has changed).

If the assumed distributional form is a correct representation of the stochastic behaviour of the losses then generally this sort of approach will be more accurate than a non-parametric approach (because it makes greater use of more relevant data). However, if the assumed distributional form is inconsistent with important features of the true loss distribution then problems may arise.

A simple example of parametric estimation is to assume that losses, over a given time horizon, are normally distributed, i.e. $x \sim N(\mu, \sigma^2)$. If the time series of losses are assumed to be independent and identically distributed then we could estimate μ and σ using the sample (unweighted) means and standard deviations. If we wanted to give greater credibility to, say, more recent data then we could instead use weighted sample means and standard deviations.

Three important problems arise with assuming that portfolio losses are i.i.d. normal:

(a) Distributions of returns/losses on financial data series, particularly relatively high frequency (e.g. weekly or daily) data often exhibit tails that are fatter-tailed than the normal distribution.

A random variable x is said to possess a fat-tailed or leptokurtic distribution if its (excess) kurtosis defined as follows exceeds zero, where $\mu = E(x)$ and $\sigma^2 = E((x-\mu)^2)$. The normal distribution has an (excess) kurtosis of zero.

$$(excess)\ kurt = E\left(\left(\frac{x-\mu}{\sigma}\right)^4\right) - 3$$

(b) Distributions of returns/losses on financial data series also often appear to be skewed, again a feature not exhibited by the normal distribution.

A random variable x is said to possess a skewed distribution if its skew defined as follows is non-zero. The normal distribution has a skew of zero.

$$skew = E\left(\left(\frac{x-\mu}{\sigma}\right)^3\right)$$

(c) Returns or losses often exhibit (conditional) heteroscedasticity by which we mean that squared deviations from the mean appear to show time dependency.

Non-zero skewness and excess kurtosis can arise for a range of reasons, some of which are explained further in Kemp (2009), Kemp (2010) and www.nematrian.com/erm.aspx. For example, high grade bonds may be expected to default relatively rarely but when they do they may experience significant market value declines. This makes their returns generally left skewed (i.e. skewed to the downside). There are a range of statistical tests that can be used to test for normality and/or for heteroscedasticity, see Book of Formulae, including some that refer just to skewness and kurtosis and others that are more sophisticated (and are capable of being used to test versus any specified distributional form).

If observations do not appear to be coming from a normal distribution then it becomes necessary to select a family from which the data does appear to be coming and then to fit parameters to select the particular member of the family deemed to fit the data best. The two main methods are maximum likelihood and the method of moments, see Book of Formulae and section 6.2 of this Appendix. With the maximum likelihood methodology it is also possible to estimate (asymptotic) standard errors, i.e. confidence levels, on the accuracy of the resulting estimates.

Sufficiently well behaved distributions will possess moments, i.e. $E\left(x^k\right)$ for $k = 1, 2, \ldots$. The mean and standard deviation can be viewed as corresponding to the first two moments (technically it is the variance, i.e. the square of the standard deviation, that relates to the second moment, and then the centred version, i.e. $E((x-\mu)^2)$. Likewise skew and (excess) kurtosis relate (in a suitably standardised sort of way) to the third and fourth moments respectively. It is in theory possible to identify asymptotic expansions that can characterise the cdf of a probability distribution by reference to skew, kurtosis and other higher moment analogues using the Cornish-Fisher asymptotic expansion, see Book of Formulae. However, if the skew and kurtosis used in this formula are estimated based on observed data it is worth noting that the sample skew and kurtosis give greater weight to how a distribution deviates from normality in the central part of the distribution than to how it deviates from normality in the tails (because most observations are in the centre of the distribution). Thus the Cornish-Fisher asymptotic expansion may not be a reliable way in practice of estimating the extent to which the tails of a distribution deviate from normality. Usually the quantile relevant in a VaR computation is towards the tail of a distribution rather than towards its centre.

2.12 Heteroscedasticity

The above methods assume that losses/returns are independent and identically distributed through time. Financial returns in many cases appear not to be independently distributed over time. In particular, volatility appears to be forecastable, in the sense that returns that are large in absolute magnitude today tend to be followed by returns that are large in absolute magnitude tomorrow (although less obvious is the sign of the return!). This does not mean that it is necessarily possible to create massively efficient trading strategies involving buying and selling volatility. Instead the term structure of option implied volatility as derived from option prices in effect incorporate market views about how strongly autocorrelated (implied) volatility might be in the future.

The standard approach to modelling this phenomenon involves use of Generalised Autoregressive Conditional Heteroscedasticity (GARCH) models. These are autoregressive in the volatility component (mainly) rather than in any mean drift component primarily because the latter, if they existed, would more naturally permit the construction of implausibly efficient trading strategies. Again, we see the underlying adoption of the premise that we should be wary when measuring and managing risk about assuming that active management strategies will add value relative to passive alternatives.

A simple example of a GARCH model is the GARCH(1,1) model, This would involve the following (in practice μ will slowly evolve as additional data is received):

$$x_{t+1} = \mu + \sigma_t \varepsilon_t$$
$$\varepsilon_t \sim N(0,1)$$

where for, say, a GARCH(1,1) model

$$\sigma_t^2 = \alpha + \beta(x_t - \mu)^2 + \gamma\sigma_{t-1}^2$$

Here α, β and γ are positive constants. If we rule out explosive behaviour by requiring $\delta = \beta + \gamma < 1$ we may derive the steady state (long-term) volatility of x_t as $\bar{\sigma} = \sqrt{\alpha/(1-\delta)}$ and the term structure of volatility, i.e. the conditional expectation of volatility, is:

$$v_{t,T}^2 = (T-t)\bar{\sigma}^2 + (\sigma_{t+1}^2 - \bar{\sigma}^2)\frac{(1-\delta)^{T-t}}{1-\delta}$$

So, after an unusually large shock, volatility will gradually revert to its long-run mean.

Even if ε_t are normally distributed (so returns are conditionally normal) such a process is unconditionally fat-tailed, i.e. the x_t will appear to have fatter-tails than a normal distribution viewed just as a single, non-time delineated, series. This arises because such a GARCH model then involves a distributional mixture of normal distributions all with the same mean but with different standard deviations. Observations in the tail of the (unconditional) distribution will tend to be drawn from normal distributions with relatively large standard deviations whilst observations in the centre of the (unconditional) distribution will tend be drawn from ones with relatively small standard deviations. This leads to a distributional form that is more peaked in the centre and more spread out in the tail than a normal distribution.

Perhaps the commonest relatively simple approach to parametric modelling that is actually used for trading book problems is a simplified version of the GARCH(1,1) model popularised by RiskMetrics (originally JP Morgan). This involves the losses being assumed to come from the following process (again in practice μ will slowly evolve as additional data is received):

$$x_{t+1} = \mu + \sigma_t \varepsilon_t$$
$$\varepsilon_t \sim N(0,1)$$
$$\sigma_t^2 = \frac{T}{T-1}\sum_{i=0}^{T-1} \lambda_i (x_{t-i} - \mu)^2$$

where the λ_i are weights applied to past squared deviations, i.e. the $\lambda_i \geq 0$ and $\sum_{i=0}^{T-1} \lambda_i = 1$. The term $T/(T-1)$ is a small sample adjustment so that the variance is not understated. Often the weights are constructed so that they form an exponentially declining series as one goes back in time, i.e. $\lambda_i = k\lambda^i$ where k is chosen so that $\sum_{i=0}^{T-1} \lambda_i = 1$. A typical value of, say, $\lambda = 0.97$ for daily data would result in observations lagged by more than a month being given little weight. More precise methods for introducing a small sample adjustment can be developed, see e.g. http://www.nematrian.com/WeightedMomentsAndCumulants.aspx. As with all use of past data for such purposes, some additional thought is generally needed to deal with incomplete data series and other situations where we do not have quite the data we would like.

2.13 Non-parametric VaR modelling

If there is no obvious parametric distributional form for the returns/losses we may prefer to use a non-parametric approach. Suppose we have a sample of n realisations of losses, $x_1, \ldots, x_n$ (again these

would normally be the losses had the current positions been prevailing at the time). We reorder the sample so that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. $x_{(r)}$ is called the r-th order statistic for this sample ($1 \leq r \leq n$).

An obvious non-parametric estimator for VaR_α is $x_{(k)}$ where $k - 1 < n(1-\alpha) \leq k$. For this to be a sensible estimator we must implicitly be assuming that the observations are not evolving conditionally through time.

It is possible to estimate standard errors for the resulting estimators but the process is somewhat more involved than for maximum likelihood. Suppose the observations are coming from a probability distribution with cdf $F(x)$ and pdf $f(x)$. Then the probability, P, that, in a sample of n observations, $r-1$ fall below y, 1 falls in the range $[y - dy/2\,, y + dy/2]$ (for small dy) and $n - r$ fall above y is:

$$F(y)^{r-1} f(y) dy \big(1 - F(y)\big)^{n-r}$$

So at $y = x_{(r)}$ it is, where $F_{(r)} = F\big(x_{(r)}\big)$:

$$F_{(r)}^{r-1} \big(1 - F_{(r)}\big)^{n-r} dF_{(r)}$$

Suppose we define $q = r/n$ and $p = 1 - q$ and suppose we identify the mode of the distribution of $x_{(r)}$. We can do this by taking logs of the above definition, taking derivatives and setting to zero. This gives:

$$(r-1)\frac{f_{(r)}}{F_{(r)}} + (n-r)\frac{f_{(r)}}{1 - F_{(r)}} + \frac{f'_{(r)}}{f_{(r)}} = 0$$

If $n \to \infty$ and if we hold q fixed this simplifies to:

$$\frac{q}{F_{(r)}} - \frac{p}{1 - F_{(r)}} = 0$$

So as expected, $F(\tilde{x}) = q$ where $\tilde{x}$ is the non-parametric estimator given above and so the probability that losses will exceed the mode of the distribution of $x_{(r)}$ in the limit equals the correct confidence level for VaR_α.

Consider now the distribution of $\tilde{x}$ in the neighbourhood of the mode of the distribution of $x_{(r)}$. Suppose $F\big(x_{(r)}\big) = q + \xi$. Then $P = (q+\xi)^{nq}(p-\xi)^{np} dF_{(r)}$. Taking logs and expanding we have:

$$nq \log\left(1 + \frac{\xi}{q}\right) + np \log\left(1 - \frac{\xi}{q}\right) = nq\left(\frac{\xi}{q} - \frac{\xi^2}{2q^2}\right) + np\left(-\frac{\xi}{q} - \frac{\xi^2}{2q^2}\right) + O(\xi^3) = -\frac{n\xi^2}{2pq} + O(\xi^3)$$

For large samples we may ignore higher order terms, so the distribution of ξ is proportional to $\exp\left(-\frac{n\xi^2}{2pq}\right) d\xi$, i.e. ξ is asymptotically normal with variance pq/n. The variance of $x_{(r)}$ therefore asymptotically satisfies:

$$\frac{pq}{n} \approx variance(\xi) \approx f^2\big(x_{(r)}\big) variance\big(x_{(r)}\big) \quad \Longrightarrow \quad variance\big(x_{(r)}\big) \approx \frac{pq}{n f_{(r)}^2}$$

To use this formula to determine the standard error of the estimate we need to calculate the density, f, of the loss random variable in the neighbourhood of $x_{(r)}$.

We may also calculate non-parametric confidence intervals for the VaR. The non-parametric VaR_α estimate, say ξ_q, corresponds to some order statistic $x_{(q)}$ where $q < n(1-\alpha) \leq q+1$. To construct a confidence interval we observe that:

$$Pr\big(x_{(r)} < \xi_q\big) = Pr\big(x_{(s)} \leq \xi_q\big) + Pr\big(x_{(r)} \leq \xi_q < x_{(s)}\big)$$

So if we take two order statistics, say $x_{(r)}$ and $x_{(s)}$, from our sample which bracket the VaR estimate we have:

$$Pr\big(x_{(r)} \leq \xi_q < x_{(s)}\big) = Pr\big(x_{(r)} \leq \xi_q\big) - Pr\big(x_{(s)} \leq \xi_q\big)$$

$$\Longrightarrow Pr\big(x_{(r)} \leq \xi_q < x_{(s)}\big) = \sum_{i=0}^{n-r} \binom{n}{r+i} q^{r+i}(1-q)^{n-r-i} - \sum_{i=0}^{n-s} \binom{n}{s+i} q^{s+i}(1-q)^{n-s-i}$$

$$\Longrightarrow Pr\big(x_{(r)} \leq \xi_q < x_{(s)}\big) = \sum_{i=r}^{s-1} \binom{n}{i} q^{i}(1-q)^{n-i}$$

2.14 Value-at-risk for non-linear claims

Additional challenges arise if the portfolio contains non-linear claims, e.g. options or other non-linear derivatives. The problem is that the price of the relevant derivative depends on a non-linear way on parameters such as volatility, maturity, strike price etc. and it is not possible to collect historical returns data for derivatives conditional on all these parameters.

However, what we can normally do is apply option pricing theory to model the prices of the derivatives as non-linear functions of underlying cash security prices and other risk factors. The portfolio return distribution can then be viewed as a weighted sum of non-linear functions of security prices (and possibly other series such as volatilities) for which there is historical time series data.

Some other instruments, e.g. bonds and swaps, come in so many varieties that we often also need to use a similar price function methodology for them.

Two problems may arise with the use of pricing models for this purpose:

(a) If the options are complicated then the price functions may be so complicated that in practice we may need to employ numerical techniques to calculate them; and

(b) The distribution of weighted sums of functions of random variables may be very difficult to calculate.

Both (a) and (b) can in principle be addressed using Monte Carlo or other numerical techniques but the computational cost may be excessive. It is therefore helpful to have some analytical approximations available, some of which are set out below.

2.15 The delta approach

A natural way of measuring the risks involved with an option is to view holding an option as approximately the same as holding a number of units of the underlying security equal to the delta of the option. If the only variable on which the price of the option depends that changes is the price of

the underlying, and if it only changes by a small amount then the delta measures the sensitivity to such changes.

Thus if the call price of the option is $C(S, K, t, \sigma)$ where S is the price of the underlying, K is the strike price, t is time and σ is volatility (here assumed constant) then the natural way to choose the equivalent number of units of the underlying is to perform a first order Taylor expansion in S, i.e. using:

$$\Delta C = C(S + \Delta S, K, t, \sigma) - C(S, K, t, \sigma) \approx \frac{\partial C}{\partial S} \Delta S$$

So, calculating the VaR (for a single position) is then roughly equivalent to calculating the $(1 - \alpha)$-quantile for dS and scaling by $\partial C / \partial S$.

In principle, we might also wish to model the mean change, i.e. drift, using:

$$\Delta C \approx \frac{\partial C}{\partial S} \Delta S + \frac{\partial C}{\partial t} \Delta t$$

However, for short holding periods the VaR is determined mainly by the martingale component of the price process rather than by its mean drift, so the drift would often be ignored.

More generally, if we have a portfolio of n instruments $\mathbf{a} = (a_1, \ldots, a_n)^T$ and each of these instruments has a price function v_i dependent on m factors $\mathbf{f} = (f_1, \ldots, f_m)^T$ with the price function of the overall portfolio then being $V(f_1, \ldots, f_m) = \sum_{i=1}^{n} v_i(f_1, \ldots, f_m)$ then the above generalises to calculating the VaR using the following:

$$\Delta \mathrm{V} = \sum_{i=1}^{n} a_i \frac{\partial V_i}{\partial t} \Delta t + \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \frac{\partial V_i}{\partial f_j} \Delta f_j = \mu_{\mathbf{a}} + \sum_{j=1}^{m} q_j \Delta f_j$$

where $\mu_{\mathbf{a}} = \sum_{i=1}^{n} a_i \frac{\partial V_i}{\partial t} dt$ and $\delta_j = \sum_{i=1}^{n} a_i \frac{\partial V_i}{\partial f_j}$. Set $\mathbf{q_a} = (q_1, \ldots, q_n)^T$.

The delta approach is biased upwards compared to the true VaR for long positions in standard puts or calls. This follows from Jensen's inequality because the prices of these types of claims when expressed as a function of the underlying security price are convex (upwards). For short positions the opposite is true.

2.16 Refinements

If a first order approximation does not work well then a natural generalisation is to employ a second order expansion, i.e. one involving both the option delta and the option gamma and hence both Δf_i and $\Delta f_i . \Delta f_j$

In principle, higher order terms can also be included. However, this is rarely done. This is partly because most analysis of the sensitivity of the price of a derivative to its underlying focuses on delta and gamma (and these, with a sensitivity to time, are the only sensitivities that appear in Ito's formula used in derivative pricing theory). More importantly, it misses the potential dependency of the price function to variables other than the price of the underlying. Recall, for example, that the Black-Scholes price of a call option involves a formula along the lines of:

$$C_t = S_t e^{-q(T-t)} N(d_1) - K e^{-r(T-t)} N(d_2)$$

where

$$d_1 = \frac{\log(S_t/K) + (r - q + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$

$$d_2 = \frac{\log(S_t/K) + (r - q - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}$$

and σ here is the option's implied volatility, r is an interest rate and q is a dividend yield.

In particular, we see that the price depends on the implied volatility, which is not necessarily constant through time (or, in reality, for options on the same underlying with different strikes and maturities).

A consequence is that if we want to develop effective hedging approaches to portfolios with options we need not only to hedge the option delta but also to hedge the option vega, i.e. the sensitivity of the option price to changes in the implied volatility. For the above sort of derivative, we might also in principle need to hedge its other sensitivities, e.g. to r and q. We may also want to hedge the option gamma (i.e. the sensitivity of the delta to changes in the price of the underlying) to minimise transaction costs incurred when dynamically hedging the position.

In a like fashion, a more exact VaR methodology needs to take account of the sensitivity of the price to changes in implied volatility (and to other factors on which the price depends). We can think of this as involving including extra terms in the equation for ΔV relating to sensitivity to implied volatility etc.

Some of the complexities involved are illustrated by material included in BCBS (2016), the latest Basel minimum capital requirements for market risk, also colloquially known as the Fundamental Review of the Trading Book.

3. Traditional portfolio market risk models

3.1 Introduction

Most equity portfolios and other relatively straightforward types of portfolio expressing market risk contain many different instruments each of which is expected to behave somewhat differently. A major aspect of *portfolio* risk models is to provide some simple but not overly simplistic way of aggregating the impact of these individual exposures. The most common way in which this is done is via factor models.

There are three main ways of estimating the factor structures underlying risk models using time series data:

(a) A *fundamental risk model* ascribes certain fundamental factors (such as price to book) to individual securities. These factor exposures are exogenously derived, e.g. by reference to a company's annual report and accounts. The factor exposures for a portfolio as a whole (and for a benchmark, and hence for a portfolio's active positions versus a benchmark) are the weighted averages of the individual position exposures. Different factors are assumed to behave in the future in a manner described by some joint probability distribution. The overall portfolio risk (versus its benchmark) can then be derived from its active factor exposures, this joint probability distribution and any additional variability in future returns deemed to arise from security specific idiosyncratic behaviours.

(b) An *econometric risk model* is similar to a fundamental model except that the factor exposures are individual security-specific sensitivities to certain pre-chosen exogenous economic variables, e.g. interest rate, currency or oil price movements. The sensitivities are typically found by regressing the returns from the security in question against movements in the relevant economic variables, typically using multivariate regression techniques.

(c) A *statistical risk model* eliminates the need to define any exogenous factors, whether fundamental or econometric. Instead we identify a set of otherwise arbitrary time series that in aggregate explain well the past return histories of a high proportion of the relevant security universe, ascribing to elements of this set the status of 'factors'. Simultaneously we also derive the exposures that each security has to these factors. This can be done using principal components analysis or other similar techniques, see below.

In practice these methods are less differentiated than first appears to be the case. This is because fundamental factors and/or factors identified via econometric modelling will generally only be considered valid if they also appear to have exhibited meaningful explanatory power, which means that in aggregate they should also largely coincide with factors derived from statistical analysis. Statistical analysis should identify the factors that *best* explain the past; one reason that we do not necessarily adopt purely statistical techniques all the time is because we think that incorporating more qualitatively driven factors may make the risk model *better* at explaining the future even if not then quite so good at explaining the past.

All three models generally in effect try to minimise the squared residuals $\sum \varepsilon_{j,t}^2$ (or some other appropriate loss function) involved in a model for the returns $r_{j,t}$ on the j'th instrument along the lines of the following equation (or with $x_{k,t}$ replaced by $f_j\left(x_{k,t}\right)$ for instruments that are expected to behave non-linearly in response to factor $x_{k,t}$, the f_j can then be viewed as akin to a pricing function for that instrument):

$$r_{j,t} = \alpha_j + \beta_{j,k} x_{k,t} + \varepsilon_{j,t}$$

For example, to create a fundamental factor model, we might:

(i) Identify fundamental characteristics, i.e. 'factors', that we believe a priori have some explanatory merit;

(ii) Calculate return series, $x_{k,t}$, that correspond to a unit amount of a given factor exposure (this is typically done by calculating a suitable 'average' return across all securities exhibiting this factor, after stripping away the impact of any exposures the securities have to any to other factors, so is actually done in tandem with step (iii));

(iii) Carry out a multiple regression analysis of $r_{j,t}$ versus $x_{k,t}$ based on the above equation, to identify the $\beta_{j,k}$. This simultaneously identifies the residuals remaining after the impact of the relevant factor exposures;

(iv) Impose some structure on the residuals (perhaps using 'blind' factors as per a statistical model);

(v) Identify the expected future behaviour of the factors in (i) and the residuals in (iv).

It is helpful to understand some of the inherent limitations that arise with risk models built up from past time series data. The most obvious is that the past is not necessarily a good guide to the future. However there are others that are more subtle in nature and can lead risk managers astray if they do not take proper account of them when using factor models. They are perhaps best appreciated by considering in more detail how a purely statistical risk model might operate.

3.2 Linear algebra and principal components

Suppose we have $i = 1, \dots, m$ data series (e.g. returns on different instruments) each with $j = 1, \dots, n$ observations, $X_{i,j}$, that are coincident in time across the different data series. Suppose the $m \times m$ covariance matrix of the (empirical) covariances between the different series is $\mathbf{V}$. The *eigenvalues* and *eigenvectors* of V are the values of λ (scalar) and associated $\mathbf{x}$ (vector) for which $\mathbf{V}\mathbf{x} = \lambda\mathbf{x}$. An $m \times m$ matrix has m (not necessarily distinct) eigenvalues and associated eigenvectors. Eigenvectors associated with distinct eigenvalues are *orthogonal*, i.e. $\mathbf{x}_i^T\mathbf{x}_k = 0$ for $i \neq k$. *Orthonormal* eigenvectors have $|\mathbf{x}_i| = \mathbf{x}_i^T\mathbf{x}_i = 1$ and $\mathbf{x}_i^T\mathbf{x}_k = 0$ for $i \neq k$. For any distinct eigenvalue the associated orthonormal eigenvector is unique up to a change of sign. If $q > 1$ eigenvalues all take the same value then it is possible to find q orthonormal eigenvectors corresponding to all of these eigenvalues. For empirical covariance matrices, $\mathbf{V}$ is symmetric non-negative definite (and positive definite if no two data series are perfectly correlated) and all of its m eigenvalues, λ_i, are greater than or equal to zero. One way of telling if a matrix is positive definite is to test whether it is possible to apply a Cholesky decomposition to it.

The eigenvalues and associated eigenvectors of an empirical covariance matrix may be sorted so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. The *first principal component* is the mixture of the underlying (de-meaned) series, i.e. the $r_j = \sum_{q=1}^{m} b_q\left(X_{q,j} - \bar{X}_i\right)$, that corresponds to the orthonormal eigenvector, $\mathbf{b}$, corresponding to the largest eigenvalue of V. This choice of $\mathbf{b}$ maximises $\mathbf{b}^T\mathbf{V}\mathbf{b}$ subject to $|\mathbf{b}| = 1$, Other (lesser) principal components correspond to orthonormal eigenvectors corresponding to smaller eigenvalues.

Eigenvectors form a basis for the relevant vector space. By this we mean that any data series (with zero mean) derivable from the m underlying data series can be formed as a linear combination of the eigenseries corresponding to the different eigenvectors.

Statistical factor models generally select factors that correspond to the most important principal components. Often an overall market factor is almost coincident with the first principal component. This is because it turns out that the principal components also, in a suitable sense, explain the most variability across the universe of instruments from which the principal components are derived. As the eigenvectors form a basis set, if we include in our factor model all of the non-zero eigenvectors and corresponding eigenseries then the entire variability in the dataset is captured.

Certain characteristics of principal components then take on added relevance:

(a) If we have n observations per data series then there are at most $n - 1$ non-zero eigenvalues even if there are many more data series, i.e. even if $m \gg n$. Suppose we have 60 months' worth of data. This means that we can only ever find non-zero 59 eigenvalues that in aggregate explain the entire variability of the dataset. A simple way to understand this is to note that linear combinations of the n series consisting of $(1,0,0, \dots)$, $(0,1,0,0, \dots)$, $(0,0,1,0, \dots)$, …, $(0,0, \dots,0,1)$ are able to replicate any series of n elements and there is in fact one spare if we also require the means of each series to be zero.

(b) For a typical value for n, say 60 or more, most of the observable eigenvalues that are non-zero generally do not appear to be statistically different to what we might expect were the data to be entirely random. Perhaps 5-10 appear meaningfully non-random in most cases.

(c) This means that most of the factor structure beyond a handful of relatively clear factors is subjective. There is no past data that can reliably differentiate between views on the finer structure of the factor model (i.e. the part of the structure dependent on eigenvectors outside the handful of more important factors where there is reasonable statistical evidence of their existence). The finer structure of portfolios optimised using virtually any portfolio optimisation technique will generally be heavily (if not entirely) dependent on our own views about properties this finer structure 'should' exhibit rather than on features that can be rigorously demonstrated by reference to past data.

(d) Another subtlety is that the principal components change if we change the weights given to different instruments, so our choice of universe will also influence any such analyses.

3.3 Market consistent risk measurement

In the above analysis we have focused on risk models that are derived to a large extent from past data. One way of circumventing some of the issues noted above is place greater weight on more contemporaneous data which we might expect to provide some guide to the future, e.g. implied volatilities derived from current option prices. Points to note include:

(a) There is a limit to the types of options traded in practice, so this does not provide a complete solution in practice to the fine structure issue noted in the previous section.

(b) There are certain philosophical merits to market consistent risk measurement. For example, by linking risk measures to option pricing theory we make it less likely that we bias our actions towards those that would have been effective in the past but which the market no longer believes will be effective in the future. For example, purely focusing on past data might suggest that we can sell large amounts of particular types of out-of-the money put options with impunity because the behaviours that would have triggered payouts on them have never happened in the past. However, a reason such options might have a material value is because the market is inferring that payouts might plausibly occur in the future. It is dangerous in such circumstances to assume that we know better than the market about what might happen in the future merely on the basis of analysis of past data (which all market participants should have access to).

3.4 Idiosyncratic risk

Given the inherent uncertainties in the finer structure of the factor structure (which in practice even an adoption of a fully market consistent focus does not fully circumvent) it is common to subdivide risk models into two parts, one involving a factor structure component that involves a relatively small number of factors and one involving idiosyncratic components that are specific to individual instruments. If our underlying model involves multivariate normal behaviour then this involves characterising the overall covariance matrix, $\mathbf{V}$, by a model in which we have factor exposures, f_{ij}, for the i'th instrument to the j'th factor (forming a matrix $\mathbf{F}$), a covariance matrix $\widehat{\mathbf{V}}$ between the factors and some idiosyncratic terms usually expressed via a diagonal (or nearly diagonal) matrix, $\mathbf{B}$, of idiosyncratic variances. Usually the number of factors is much less than the number of instruments, so this formulation is a much more parsimonious way of characterising the co-dependency between different exposures.

The ex-ante tracking error (and risk measures dependent on it, such as VaR) then involves a computation along the lines of the following:

$$\sigma^2 = \mathbf{a}^T\left(\mathbf{F}^T\hat{\mathbf{V}}\mathbf{F} + \mathbf{B}\right)\mathbf{a} = (\mathbf{Fa})^T\hat{\mathbf{V}}(\mathbf{Fa}) + \sum a_i^2\sigma_i^2$$

If a firm has a dual listing and therefore two separate equity instruments exist relating to essentially the same underlying assets but perhaps with different tax or other investor characteristics (or if a firm has a major subsidiary that is separately quoted) then the different securities involved would share common idiosyncratic characteristics. The matrix characterising idiosyncratic terms would then only be 'nearly' diagonal.

3.5 Back-testing market risk models

As market risk models have become more sophisticated and particularly as they have become more commonly used for setting regulatory capital requirements greater attention has been placed on *back-testing* models to see whether they appear to be robust based on what they would have predicted had they been applied in the past.

All back testing suffers from the possibility of *look-back bias*, i.e. with likely reliability of the model being distorted by our knowledge of the past and therefore our ability to select models that appear reasonable based on that knowledge. There are ways of reducing look-back bias, e.g. using out-of-sample approaches in which we do not test the predictions of a single model applied to all past time periods but instead we test the predictions of a model formulation with the formulation as applied to a particular past time period only using data that would already have been available at that time. However, even this has some look-back bias as the model formulation will itself have been selected from a range of possibilities. This selection can be biased as it will inevitably be based on what we now know about the past. We do not attempt to identify in this section how to adjust mathematically for look-back bias.

3.6 Statistical background: maximum likelihood

Suppose we have a sample of n observations, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ each of which is an independent draw from a distribution $F(x|\theta)$ with pdf $f(x|\theta)$ where θ is a vector of unobserved parameters. The likelihood of the sample (strictly speaking the unweighted likelihood as in some cases we may want to give more weight to some observations than to others), L, is defined as the product of the densities:

$$L(\mathbf{x}|\theta) = f(x_1|\theta)f(x_2|\theta)\dots f(x_n|\theta)$$

We may then estimate θ by maximising L with respect to θ. This is called *maximum likelihood estimation*. So the maximum likelihood (ML) estimator is:

$$\hat{\theta} = \theta_{ML} \equiv \arg\max_\theta L(\mathbf{x}|\theta)$$

As $\log(.)$ Is a monotonically increasing function we have $\arg\max_\theta L(\mathbf{x}|\theta) = \arg\max_\theta \log L(\mathbf{x}|\theta)$ so commonly we actually maximise the log likelihood:

$$\log L(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

The ML estimator has some desirable properties:

(a) It is consistent, so $\lim_{n\to\infty} \theta_{ML} = \theta$.

(b) It is asymptotically Gaussian, in the sense that:

$$\theta_{ML} \sim (asympotically)\ N\left(\theta, I^{-1}(\theta)\right)$$

where $I(\theta)$ is the *information matrix* defined as

$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial\theta\partial\theta^T}\right)$$

(c) It is asymptotically efficient in the sense that it is the consistent, asymptotically Gaussian estimator with the smallest variance (meaning the covariance matrix of any other such estimator minus that of the ML estimator is positive definite).

The other main method of fitting distributions to data is the *method of moments*. With the method of moments applied to univariate data, we calculate some moments for the observed data (or equivalents such as centred moments or cumulants). For example, the first moment of a distribution is its mean $E(X)$ and the second moment is $E(X^2)$. The variance of a distribution is technically a centred moment and is $E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E(X)^2$. We generally calculate the first n moments of the distribution if there are n different parameters to estimate. We then equate the observed values of these moments with expressions that identify their values for any given set of parameters defining the distribution. The approach is relatively simple to implement but is not always robust. For example, it may result in inappropriate or impractical values for the parameters. It is also not always clear how to handle any small sample adjustments. For example, should we equate the second centred moment with $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ or with $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$?

An approach that is philosophically connected with method of moments and circumvents its weaknesses (but typically at the expense of losing its computational simplicity) is the *generalised method of moments* (GMM) approach. In this approach we select parameters that 'best' fit selected moments, given some criterion for best. 'Best' is typically defined as minimising overall divergence versus all targeted moments simultaneously, often using a (positive semi-definite) weighting matrix, W, i.e. it involves choosing as the GMM estimator the parameter vector $\hat{\theta}$ that minimises $E\left(g\left(X|\hat{\theta}\right)\right)^T W E\left(g\left(X|\hat{\theta}\right)\right)$ where $g(X|\theta)$ is the difference between the observed moment and the theoretically expected moment if the parameter vector were θ. Given suitable regularity criteria, the GMM estimator (for a given W) is consistent. We can also use the approach to test whether the observations appear to satisfy some moment constraint derived from economic theory.

If W is chosen appropriately (as the inverse of $\Omega = E\left(g(X|\theta_0)\right)^T E\left(g(X|\theta_0)\right)$ where θ_0 is the true underlying parameter set that $\hat{\theta}$ is aiming to estimate) then the GMM estimator is also asymptotically efficient. However, we cannot of course determine Ω^{-1} in advance because by definition we need to know the value of θ_0 to compute it. Practical implementations of GMM often involve two-step or iterated approaches, in which some simpler form is initially assumed for Ω which is replaced later on in the algorithm by a more accurate estimate derived from intermediate $\hat{\theta}$ estimates.

3.7 The likelihood ratio test

Using likelihoods we may construct hypothesis tests. Suppose we have a set of k restrictions (constraints) on the parameters of the form $C(\theta) = \mathbf{0}$ where $\mathbf{0}$ is a k-vector of zeros. If these restrictions actually applied then the restrictions should not reduce the likelihood. So, suppose we estimate the model with and without the restrictions. The ratio of the restricted likelihood, L_R, to the unrestricted likelihood, L_U, i.e. $LR = L_R/L_U$, should therefore provide a statistic about the validity or otherwise of the restrictions.

In fact it is possible to show that:

$$-2\log LR \sim (asymptotically)\ \chi_k^2$$

where χ_k^2 denotes the chi-squared distribution with k degrees of freedom.

As $LR \leq 1$ so $-2\log LR \geq 0$. If the restrictions bind hard then the likelihood ratio will be small and $-2\log LR$ will be large. So we can reject the restriction (at a confidence level of γ) if $-2\log LR > \chi_k^2(\gamma)$ where $\chi_k^2(\gamma)$ is the γ-quantile of the χ_k^2 distribution.

3.8 Binomial back-testing

The likelihood ratio test is perhaps most easily applied to VaR back-testing by supposing that we observe a time-series sample of losses, x_t, over n time periods and the corresponding VaR_α estimates from a portfolio VaR model relating to the same n time periods. From these we can create a variable y_t that indicates if the loss exceeded the VaR or not, i.e.:

$$y_t = \begin{cases} 1 & if\ x_t \geq VaR_\alpha \\ 0 & if\ x_t < VaR_\alpha \end{cases}$$

If the model is correct then the y_t should be a sample of independent binomial random variables taking the value 1 with probability α and 0 with probability $1-\alpha$.

Suppose the true probability were α then the likelihood of the sample is $L = \alpha^j(1-\alpha)^{n-j}$ where j is the number of observations where $y_t = 1$.

The maximum likelihood estimator is the value of α that maximises $\log\left(\alpha^j(1-\alpha)^{n-j}\right)$, i.e. has

$$0 = \frac{d}{d\alpha}(j\log\alpha + (n-j)\log(1-\alpha)) = \frac{j}{\alpha} - \frac{n-j}{1-\alpha} \quad \Rightarrow \quad \alpha_{ML} = \frac{j}{n}$$

We may therefore test whether the VaR model appears to have predicted an appropriate number of losses over the n time periods by considering the likelihood ratio between a restricted L_R in which we restrict α to some specified value α_0 (in which case L_R is constant irrespective of the data) and an unrestricted likelihood L_U using the ML estimate for α. The likelihood ratio test would then be:

$$LR = \frac{\alpha_0^j(1-\alpha_0)^{n-j}}{\left(\frac{j}{n}\right)^j\left(1-\frac{j}{n}\right)^{n-j}}$$

We would then reject the null hypothesis that $\alpha = \alpha_0$ (i.e. that the VaR model was appropriately specified) if $-2\log LR > \chi_1^2(\gamma)$ for some suitable confidence level γ.

This test was proposed by Kupiec and is used by regulatory frameworks to assess the accuracy of VaR models used by large banks. It tests whether one can reject the null hypothesis that the data involves independent binomial random variables with probability of α_0 of observing a VaR exception against the alternative of independent binomial random variables with a probability (j/n) of observing an exception.

3.9 Binomial back-testing with autocorrelation

If the VaR model does not appear to be correct then the exceptions might not be independent. It therefore makes sense to test the model against a more general alternative. Following an approach suggested by Christensen, we might assume the possibility of first order autocorrelation by assuming that the probability of an exception occurring also depends on the probability that an exception occurred in the previous period. So we might assume:

$$\pi_{12} \equiv Pr(y_t = 1|y_{t-1} = 0) \neq Pr(y_t = 1|y_{t-1} = 1) \equiv \pi_{22}$$

Suppose we have two states: (1) no exception and (2) exception. Suppose also that transitions between these states are generated by a Markov chain with transition matrix (where π_{ij} is the probability of moving from state i to state j):

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix}$$

The assumption of independence corresponds to the assumption that Π has the following form for some π:

$$\Pi = \begin{pmatrix} 1-\pi & \pi \\ 1-\pi & \pi \end{pmatrix}$$

The maximum likelihood estimator for the unrestricted transition matrix is:

$$\Pi_{ML} = \begin{pmatrix} \hat{\pi}_{11} & \hat{\pi}_{12} \\ \hat{\pi}_{21} & \hat{\pi}_{22} \end{pmatrix}$$

where $\hat{\pi}_{11} = \frac{n_{11}}{n_{11}+n_{12}}$, $\hat{\pi}_{12} = \frac{n_{12}}{n_{11}+n_{12}}$, $\hat{\pi}_{21} = \frac{n_{21}}{n_{21}+n_{22}}$, $\hat{\pi}_{22} = \frac{n_{22}}{n_{21}+n_{22}}$ and n_{ij} is the number of observations in state i one period and state j the following period.

Suppose we now calculate:

$$L_3 = likelihood\ under\ independence = (1-\hat{\pi})^{n_{11}+n_{21}}\hat{\pi}^{n_{12}+n_{22}}$$
$$L_4 = unrestricted\ likelihood = \hat{\pi}_{11}^{n_{11}}\hat{\pi}_{12}^{n_{12}}\hat{\pi}_{21}^{n_{21}}\hat{\pi}_{22}^{n_{22}}$$

We can test for independence using the following likelihood ratio, which under the null hypothesis is distributed as chi-squared with 1 degree of freedom:

$$LR_{independence} = -2\log(L_3/L_4)$$

We can also jointly test for independence and accuracy of the VaR model using a likelihood ratio akin to the likelihood formulae referred to previously.

4. Portfolio credit risk models

4.1 Introduction

There are three main types of model commonly used to assess portfolio credit risk. These are:

(a) Ratings-based models such as Riskmetric's *Creditmetrics*™ or Standard and Poor's *Portfolio Risk Tracker*™

(b) Equity-based models such as Moody's-KMV *Portfolio Manager*™

(c) Actuarial or so-called mixture models such as CSFB's *CreditRisk+*™.

The main aim of these models is to provide statistics relating to the portfolio value distribution (e.g. VaR) at some future date T given information at some initial date t. Each typically has two key components:

(i) A model for the stochastic evolution of credit quality between t and T; and

(ii) A model for valuing the credit exposures at the future date T conditional on their credit quality at that time.

4.2 Modelling the non-credit sensitive component of bonds

Portfolio credit risk models in general need to capture risks that are not credit related. In particular, they in general need to capture overall market interest rate risks, if these have not been fully hedged and if the instruments involved do express such risks (as is the case with bonds although largely not the case with credit default swaps). The evolution of the price of a credit risk-free bond will depend on the evolution of the relevant yield curve through time. These evolutions are often modelled in a manner conceptually akin to the non-linear claims covered in section 3, by formulating some 'factors' that are assumed to drive the behaviour of different bonds. In particular, we might determine factors that we think characterise how yields might move and then work out the sensitivity of bond prices to these factors.

It is relatively common to use a three-factor model when modelling overall market interest rate risks. The three factors then typically involve (1) a 'shift' involving the same yield move up or down across the entire yield curve, (2) a 'twist' to the yield curve (with the short end moving up and the long end down or vice versa) and (3) a 'butterfly' factor in which the two ends move the same way but each move in the opposite way to the middle of the yield curve. Whilst the same three factor structure may be used for each currency, the relevant factors for different currencies will not normally be assumed to be perfectly correlated, as yield curves in different currencies evolve differently through time. The reference market yield curves will often be associated with relevant government debt yield curves although strictly speaking this requires the assumption that such debt is itself credit risk free.

It is possible to apply similar types of factor model to the credit sensitive components of instrument values or to refine the three types of portfolio credit risk model referred to above in such a manner. This might perhaps be done if there appears to be a noticeable term structure to the credit spread of a specific issuer. We can think of this as involving refining the components in section 4.1(i) and (ii) so that credit quality is in part associated with a credit spread term structure (that might be instrument as well as name specific). More normally with portfolio credit risk modelling we adopt the simplifying assumption that the same credit quality applies to all instruments issued by the same name. In the

case of ratings-based models we also assume that credit quality is well characterised by the rating ascribed to the issuer by a credit ratings agency.

It is relatively common when analysing interest rate risk to calculate metrics for individual bonds such as the bond's duration. If a bond gives rise to cash flows c_t at time t then its (dirty) price, if its gross redemption yield (expressed annually) is i, is:

$$V = \sum_t \frac{C_t}{(1+i)^t}$$

Its duration and its modified (i.e. Macaulay) duration are then:

$$duration = \frac{1}{V}\sum_t \frac{tC_t}{(1+i)^t} \qquad modified\ duration = \frac{1}{V}\frac{dV}{di} = \frac{duration}{1+i}$$

A credit risk free bond's modified duration is closely allied to its DV01 (otherwise called PV01), i.e. the (dollar) value change for a 1 basis point change in its yield, or more precisely its interest rate DV01 or IRDV01.

For instruments that are credit sensitive life is more complicated. This is because some instruments like CDS have relatively little overall interest rate DV01 or duration (because their price is not sensitive to overall movements in, say, government bond yield curves). However, they may have much greater sensitivity to changes in credit spreads relative to such yield curves, which we might capture via a corresponding credit DV01 or CRDV01.

4.3 Ratings-based portfolio credit risk models

A ratings-based model typically assumes that there is a portfolio of n credit sensitive exposures. Each is assumed to have a credit rating and these ratings are assumed to evolve according to a Markov chain defining the probability of moving from rating i to rating j over a given period (say a year) as, say, π_{ij}. We might for example have K states with the K'th state involving default, so π_{iK} is the one-period likelihood of default for an i-rated exposure. Estimates of π_{ij} might be obtained from historical data on ratings transitions as prepared by a credit ratings agency. We generally assume that default is an absorbing state, i.e. once an exposure has defaulted it never reemerges from the default state (or if it can do the portfolio would no longer have any exposure in the outcome).

The model then identifies some method of valuing the exposure at the horizon of the risk calculation. As explained in section 4.2, the component of value that is not credit sensitive (e.g. exposure to general interest rates) would be valued using standard bond pricing approaches and the risk characteristics probably modelled as above, usually via some sort of factor modelling process. So a key additional component needed by a ratings-based model is some way of valuing the credit sensitive component.

If we assume that changes in credit quality and interest rates are independent then a promise by an obligor (that is r-rated and not defaulted at time t) to pay a stream of m cash flows c_i at times $t_i > t$ may be priced at time t as:

$$V_t^{(r)} = \sum_{i=1}^{m} c_i P_{t,t_i} \exp\left(-s_{t,t_i}^{(r)}(t_i - t)\right)$$

where $P_{t,u}$ is the price at t of a default-free zero coupon bond paying 1 for sure at time $u \geq t$ and $s_{t,u}^{(r)}$ is the credit spread at t for a payment at time u from an r-rated obligor. We are here assuming that the credit spread is quoted in a geometric manner with continuous compounding.

The $P_{t,u}$ correspond to the part of the bond value that is not credit sensitive, and is handled as above (or is assumed to be constant, but this is only sound if we assume that interest rate risk has been hedged). Quite commonly ratings-based credit risk models assume that $s_{t,u}^{(r)}$ are also known ex ante but again this is only sound if spread risk has in aggregate been hedged. If this is not the case then strong assumptions are needed for such an approach to be valid, e.g. that spreads on defaultable debt of a given rating are constant.

We also need to identify the prices of claims in the case of default. Commonly, it is assumed that the value of the claims is then equal to a random fraction, γ, of either (i) the par value for the claim or (ii) the value of a default free security with the same contractual cash flows. γ is known as the recovery rate. Alternatively, a simpler approach may be adopted in which γ is set to a constant, e.g. 40%. If γ is random then it might be assumed to be an independent draw from a beta distribution. If alternative (ii) is adopted then the value in the defaulted state is:

$$V_t^{(K)} = \sum_{i=1}^{m} c_i P_{t,t_i} \gamma$$

By compounding up the transition matrix to the relevant time horizon we may calculate the likelihood that an instrument from an obligor currently r-rated is j-rated at the relevant time horizon, T. Given the values at time T conditional on different possible future ratings we can work out the univariate distributions of the (future) values of each exposure.

To identify the distribution for the overall portfolio value we also need to allow for correlations (or more generally co-dependencies) between ratings changes for the different exposures. Commonly we assume that for the k'th exposure there is a latent (i.e. not directly observable random variable) R_k that drives its ratings transitions over a given period. Without loss of generality we can standardise the R_k so that their (marginal) distributions are unit Normal and we assume that if the rating at time 0 is i then the rating at the end of period 1 is j if R_k lies in the interval $\left(Z_{i,j-1}, Z_{i,j}\right)$ where the $Z_{i,j}$ are chosen to be consistent with π_{ij}. This requires (where $N(x)$ is the unit normal cdf, $Z_{i,K} = +\infty$ and $Z_{i,0} = -\infty$):

$$\pi_{i,j} = N\left(Z_{i,j}\right) - N\left(Z_{i,j-1}\right)$$

Solving these equations recursively we have:

$$Z_{i,j} = N^{-1}\left(\pi_{i,1} + \pi_{i,2} + \cdots + \pi_{i,j}\right) \qquad (1 \leq j \leq K-1)$$

This type of approach to ratings transitions is called an ordered probit model and is relatively commonly used in discrete choice econometric analysis. Its main advantage here is that we can now make some relatively simple assumptions about correlations (or more generally, co-dependencies) of credit rating evolutions of different obligors by incorporating assumptions about the behaviour of the latent variables.

Commonly we assume that the latent variables are jointly Gaussian distributed (i.e. multivariate normal). We are then left with choosing the correlations within the relevant correlation matrix. Often these are derived from the correlations of the equity returns of the individual obligors in question.

Alternatively, and especially for obligors that do not have equity issues, we might use a weighted average of industry and country indices corresponding to the obligor in question, together with an idiosyncratic term, c.f. traditional market risk models. This approach is particularly simple when there is a single index or factor which is the same for all obligors, plus an idiosyncratic term. We may then express the (log) return of the n'th firm's equity as:

$$r_n = \alpha_n f + \varepsilon_n \sqrt{1 - \alpha_n^2}$$

where f and ε_n correspond to the return on the market and the obligor idiosyncratic return component and are assumed to be independent unit normal random variables.

4.4 Equity-based portfolio credit risk models

For the ratings-based credit risk model to have theoretical validity we need the rating ascribed by a ratings agency to a particular obligor to provide meaningful information about the likely future evolution of the credit spread and default probability applicable to that obligor. If we are not confident that credit ratings are sufficiently reliable for this purpose then we will need to think of alternative ways of modelling credit risk exposures.

Perhaps the most common alternative is an equity-based portfolio credit risk model. This type of model relies on the underlying observation that, under limited liability, the value of a firm's equity may be seen as a type of call option written on the firm's underlying assets with the strike price being its liabilities whilst the (market) value of the firm's debt can be viewed as another derivative dependent on these two inputs since both equity and debt form parts of the same overall firm capital structure. For example, the seminal works on option pricing by Black and Scholes (1973) and Merton (1974) both stressed the potential application of derivative pricing to modelling corporate debt.

Suppose a firm issues bonds (without intermediate coupons) that pay D at time T and that the value of the firm's assets (under pure equity financing) follows a process V_t. Then the payoff to bond-holders and equity holders are:

$$Bond\ payoff = \max(0, \min(D, V))$$
$$Equity\ payoff = \max(V - D, 0)$$

Pricing these claims is in principle straightforward if we make appropriate assumptions about the process for V_t. For example, we might merely assume that it follows a geometric Brownian motion as per Black-Scholes:

$$dV_t = \mu V_t + \sigma V_t dB_t$$

A more plausible model might involve the following features:

(a) There might be coupon payments on the D so that it follows a process D_t and some dependency with overall market movements which are assumed to follow a process M_t. There might also be dividend flows to equity holders, so we might have for the i'th firm (where the μ_V, ρ etc. might depend on i):

$$dV_t = \mu_V V_t dt + \sigma_V V_t dW_{V,t}$$
$$dD_t = \mu_D D_t dt$$
$$dM_t = \mu_M M_t dt + \sigma_M M_t dW_{M,t}$$

$$dW_{V,t}dW_{M,t} = \rho dt$$

(b) We might also assume that insolvency occurs when the ratio of assets to liabilities $k_t \equiv V_t/D_t$ falls below an exogenously defined trigger level, k^*, and that if this does occur then equity-holders receive nothing (i.e. there is an instantaneous decline to zero in any remaining value of the firm).

We can then value the equity using barrier option pricing techniques.

In principle, we could extend such models to price the debt too. However, it is often more convenient to think of the firm's assets and liabilities as determining the default probability and to value the debt issue as if it is small compared to the firm's broader balance sheet. The survival probabilities would then be derived from the probability that k_t has not yet hit k^* as above by time T (or earlier maturity of the debt).

Implementation of an equity-based portfolio credit risk model thus requires the following:

(1) We need to estimate the processes driving the k_t for each exposure, as well as their correlations.

(2) We need to determine the cut-off points k^* for each exposure.

(3) Having estimated the various parameters we then need to simulate a vector of correlated k_t up to the horizon and value the debt conditional on the levels that the k_t has reached.

(4) We need to sum across all exposures to obtain the portfolio value

(5) We need to repeat such a simulation many times to build up an estimate of the overall probability distribution and hence portfolio statistics such as VaR.

In principle we could use, say, 'pure' ML estimates to estimate each of the parameters on which the above depends. However, as pricing depends on the correlation, ρ, this means that we would need to estimate a joint model for changes in each k_t with M_t (and if we believe that there are multiple market factors driving different obligor evolution then there will be several M_t to consider, as with traditional market risk models). Moreover, the underlying likelihood expression is very complicated.

In practice, therefore, commercial approaches tend to be much less 'pure'. For example, we might invert observed equity-liability ratios to obtain time series of the presumed k_t for each firm and then estimate the correlation matrix of the Brownian motions driving different firms' underlying asset processes. In practice it is also common to adjust how k^* are derived from, say, the combination of short and long term debt from published accounts so that the simulations more closely replicate historically observed default rates.

4.5 Actuarial (or so-called mixture) portfolio credit risk models

These types of model focus principally on losses associated with defaults rather than on losses that occur when credit standing deteriorates without default (i.e. they focus on default risk rather than credit spread risk per se). They are often developed using probability generating functions, see e.g. here.

Conceptually similar sorts of techniques can also be used to analyse n'th to default credit default swaps and other basket credit derivatives but these are beyond the scope of this note. These can be formulated as continuous time models (although often then solved numerically after discretising the timeline) in which case the probability of default per unit time is called the *default intensity*.

4.6 Analytical results in the 'fine grained' limit

Ratings-based models employed in practical applications are generally solved using Monte Carlo techniques. However, for some purposes it is helpful to solve special cases analytically. An important example which has been widely applied by regulators and others involves unpublished work by Vasicek (1991).

In the Vasicek approach we assume that a given obligor defaults when a latent variable falls below a threshold. We also assume that there is a single common risk factor, so conditional on this defaults are independent Bernoulli random variables and the number of defaults is thus binomial, so we can identify probabilities of k defaults out of a set of n obligors. The key further step is to note that as the number of obligors goes to infinity (i.e. as the portfolio becomes better and better diversified) the randomness associated with idiosyncratic risk disappears because of the law of large numbers. The loss distribution then converges to a simple analytical form.

The derivation is as follows. Suppose the i'th obligor defaults over a given time period if a latent variable Z_i falls below a cut-off point $-c$. Suppose that exposures are either in default or not. Assume that the Z_i satisfy a factor structure:

$$Z_i = \sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_i$$

where the X and ε_i are independent standard normals. Default then occurs when

$$\sqrt{\rho}X + \sqrt{1-\rho}\varepsilon_i < -c$$

As $Z_i \sim N(0,1)$ we have default probability $q = N(-c)$ where $N(x)$ is the unit normal cdf.

Conditional on the common factor X defaults are independent across individual obligors. The probability of observing k defaults out of n obligors is $P(k,n)$ is where:

$$P(k,n) = \binom{n}{k} \int_{-\infty}^{\infty} \left(N\left(\frac{-c-\sqrt{\rho}x}{\sqrt{1-\rho}} \right) \right)^k \left(1 - N\left(\frac{-c-\sqrt{\rho}x}{\sqrt{1-\rho}} \right) \right)^{n-k} dN(x)$$

Adopting the change of variables:

$$s(x) = N\left(\frac{-c-\sqrt{\rho}x}{\sqrt{1-\rho}} \right)$$

we have:

$$P(k,n) = -\binom{n}{k} \int_{-\infty}^{\infty} s^k (1-s)^{n-k} dN\left(-\frac{\sqrt{1-\rho}N^{-1}(s) + c}{\sqrt{\rho}} \right)$$

But $-dN\big(f(s)\big) = dN\big(-f(s)\big)$ so:

$$P(k,n) = -\binom{n}{k}\int_{-\infty}^{\infty} s^k(1-s)^{n-k}dW(s)$$

where:

$$W(s) \equiv N\left(\frac{\sqrt{1-\rho}N^{-1}(s) - N^{-1}(q)}{\sqrt{\rho}}\right)$$

Consider now what happens as $n \to \infty$. Let θ be the fraction of the pool of obligors that defaults. Then:

$$\lim_{n\to\infty}\sum_{k=1}^{int(n\theta)} P(k,n) = \int_0^1 \left(\lim_{n\to\infty}\sum_{k=1}^{int(n\theta)}\binom{n}{k}s^k(1-s)^{n-k}\right)dW(s)$$
$$= \int_0^1 I_{s<\theta}dW(s) = W(\theta) - W(0) = W(\theta)$$

where $I_{s<\theta}$ is the indicator variable that takes the value 1 if $s < \theta$ and 0 otherwise.

Hence the loss distribution in this 'fine grained' limit is:

$$W(\theta_t) = N\left(\frac{\sqrt{1-\rho}N^{-1}(\theta_t) - N^{-1}(q)}{\sqrt{\rho}}\right)$$

This also means that the transformed loss rate $\tilde{\theta}_t = N^{-1}(\theta_t)$ is Gaussian and satisfies:

$$\tilde{\theta}_t \equiv N^{-1}(\theta_t) \sim N\left(\frac{1}{\sqrt{1-\rho}}N^{-1}(q), \frac{\rho}{1-\rho}\right)$$

5. Modelling operational risk

5.1 Introduction

Operational risk is difficult to define precisely but for a financial firm is often taken to be all risks faced by a firm other than market and credit risks. Much of the focus of operational risk management is on implementing effective controls and other incentives that minimise (in a cost-effective way) the likelihood of an operational risk occurring, because operational risks often provide little or no upside compared to their potential downside. We introduce a simple mathematical model of fraud in a hierarchy which provides insights into what characteristics we might expect an effective operational risk mitigation approach to exhibit. We also briefly consider more mathematical ways in which operational risks can be measured and therefore, we hope, managed more effectively.

5.2 A simple model of fraud in a hierarchy

We may think of a business as involving a chain of command, with more senior staff reviewing work undertaken by more junior staff. Most or all businesses are exposed to fraud risk so it is desirable to understand at a very high level what might encourage or discourage fraud.

Suppose in a hypothetical firm we have an infinite chain of individuals with no start and end, represented by a chain $-\infty, \dots, x-1, x, x+1, \dots, +\infty$. We index the chain so that any individual in the chain, say x, has an immediate superior $(x-1)$ and an immediate subordinate $(x+1)$. An individual chooses either to exert effort in monitoring his subordinate(s), or if he does not monitor, he chooses whether to defraud the firm himself (we assume for simplicity that an individual is unlikely to monitor his subordinate if he himself is committing fraud, e.g. because it might attract additional scrutiny).

Let F be the utility gain of fraud, C be the cost of being caught or associated with fraud, G be the utility gain from detecting fraud and M be the cost of monitoring. F, C, G and M are exogenous parameters, i.e. presumed fixed externally (and to be constant throughout the chain). We will also assume that F, C, G and M are all positive. We suppose that individuals who have not monitored are only held responsible for frauds by their immediate subordinates, so an individual only needs to worry about being monitored by his immediate superior or about monitoring his immediate subordinate.

Let f be the likelihood that any individual commits fraud. Let m be the probability (conditional on not committing fraud) that any individual monitors. f and m are endogenous parameters. To solve the problem, we will need to derive the probability, p_f, that an individual commits an undetected fraud and the probability, p_m, that the individual will uncover a fraud.

There are two possible types of equilibrium that the arrangement can exhibit (if individuals behave rationally), see here. In a 'corner' equilibrium all individuals commit fraud and no-one monitors. In an 'interior' equilibrium each individual commits fraud and monitors with some non-zero probability (the same for all individuals in this simplified model), so fraud occurs intermittently at any given point in the hierarchy, the equilibrium involving:

$$f = \frac{M}{G} \qquad m = \left(\frac{F}{F+C}\right)\left(\frac{G}{G-M}\right)$$

Assuming the interior solution applies we may conclude that to reduce likelihood of fraud we should:

- Increase G, e.g. by raising the payoff to whistleblowers and/or
- Reduce M, i.e. the cost of monitoring

The important lesson here is that reducing fraud (and more generally other types of operational risk) is facilitated by effective and relatively cheap monitoring. This perhaps explains why so much of operational risk management seems to concentrate on implementation of effective processes and systems for monitoring operational risk events and exposures. Four components of a typical system seen in a financial firm as described by Chapelle are:

- Incident reporting
- Dashboards
- Key risk indicators
- Risk and control self-assessment

5.3 Statistical modelling of operational risk

We can view operational risk as risks that disrupt the smooth running of operational processes present within the organisation in question. This is a potentially useful starting point because there is a very extensive body of operations research literature that has already been applied to manufacturing processes across many different industrial sectors. The typical approach is therefore to think of the firm (or just a specific process within the firm) as akin to a machine with a set of components, each potentially subject to failure. We might then model time or duration until failure, τ, by supposing that it has a distribution F with density f. Certain ways of viewing the distributional form then take on additional appeal, including use of the 'hazard' of a distribution, i.e.:

$$h(\tau) \equiv \frac{f(\tau)}{1 - F(\tau)}$$

By suitable choice of $F(\tau)$ we can, say, arrange for the hazard rate to fall (or rise) as τ increases depending on our intuition regarding the way in which the incidence of the risk might change through time.

If we have N machines whose failures are independent then we can simulate number of failures out of the N machines by time T by evaluating $\tau_i = F^{-1}(u_i)$ where the u_i are independent random draws from a uniform distribution. To allow for correlation between machines / machine components we might assume that the $F(\tau)$ depend on a common stochastic factor, say ρ. We would then need to evaluate $\tau_i = F^{-1}(u_i|\rho)$. Many approaches used in credit risk modelling can also be adapted to operational risk, because they involve a specific type of hazard rate, i.e. the default rate.

One important potential differentiator between credit risk and operational risk is that with credit risk modelling we usually know the number and magnitude of our exposures. The main focus is therefore on how many of a known number of these exposures might default and if so how much we might lose. In contrast, with operational risk it is less clear how large is the number of our exposures or their potential maximum magnitude. Usually it is desirable to split the overall loss into frequency and magnitude components. If both can be reliably estimated then this should reduce the variance of the overall forecasted loss. In particular, we might expect frequency of operational losses to have some reasonable correlation with business volumes if defined sufficiently widely. Moreover, better understanding (and hence potentially better monitoring) of the risks, which subdivision between frequency and severity should provide, should also help us manage the risks better, see previous section.

Another source of inspiration for mathematical analysis is the set of mathematical tools used in general (i.e. non-life) insurance, since many of the operational risks a firm faces have similarities with the risks that a non-life insurer might insure.

6. Copulas

6.1 Introduction

If random variables are multivariate normal with zero means then their co-dependency, i.e. the way in which they move in tandem, is entirely specified by their covariance matrix or equivalently by their individual (marginal) standard deviations and their correlation matrix. However, this can be a restrictive assumption and, as we shall see below, requires us to have a particular view about how likely it might be for two or more of the random variables each to take extreme values.

It is therefore important in some circumstances to have methodologies that cater for more general types of co-dependency. The most important and general of these methodologies involves *copulas*.

6.2 Definition

A copula is a multivariate cumulative distribution function for an n dimensional random vector $U = (U_1, \dots, U_n)^T$ in the unit hypercube ($[0,1]^n$) that has uniform marginals, U_i, each distributed according to $U(0,1)$ but not in general independent of each other. Let $u = (u_1, \dots, u_n)^T$ also be restricted to the unit hypercube $[0,1]^n$. Then a copula is defined as a function of the form:

$$C(u) = C(u_1, \dots, u_n) = Pr(U_1 \leq u_1, \dots, U_n \leq u_n)$$

Equivalently $C(u_1, \dots, u_n)$ is the joint cumulative distribution function for the random vector $U \in [0,1]^n$.

The *copula density* (for a continuous copula) is the pdf for which the cdf is the copula.

6.3 Properties

In the bivariate case ($n = 2$) for a general function $C(u_1, u_2)$ to be a copula it must satisfy the following properties:

1. $C(u, 1) = u = C(1, u)$ for all $0 \leq u \leq 1$
2. $C(u_1, u_2)$ must be increasing in both u_1 and u_2
3. $C(b_1, b_2) - C(a_1, b_2) - C(b_1, a_2) + C(a_1, a_2) \geq 0$ for all $0 \leq a_1 < b_1 \leq 1$ and $0 \leq a_2 < b_2 \leq 1$
4. $C(u_1, u_2) \leq \min(u_1, u_2)$
5. $C(u_1, u_2) \geq \max(u_1 + u_2 - 1, 0)$

There are equivalent generalisations when the copula relates to more than two random variables.

6.4 Sklar's theorem

A key feature of any copula is that in combination with the marginals it provides a complete characterisation of the joint probability distribution. This follows from Sklar's theorem

Theorem (Sklar's theorem). *If F is a joint (cumulative) distribution with marginal cdf's $F_1, F_2, \dots, F_n$ then there exists a copula C which maps the unit hypercube $[0,1]^n$ onto the interval $[0,1]$ such that for all $x_1, \dots, x_n$ we have:*

$$F(x_1, \dots, x_n) = C\big(F_1(x_1), \dots, F_n(x_n)\big)$$

Moreover, if the F_i are continuous functions then the copula is unique and

$$C(u_1, \dots, u_n) = F\Big(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n)\Big)$$

Conversely, suppose $C(u_1, \dots, u_n)$ is a copula and $F_1(x_1), \dots, F_n(x_n)$ are univariate cdf's. Then the function $F(x_1, \dots, x_n) = C\big(F_1(x_1), \dots, F_n(x_n)\big)$ is a joint distribution function with marginal cdf's $F_1, F_2, \dots, F_n$.

6.5 Example copulas

Random variables with continuous marginals are independent if and only if their copula is the Independence or Product copula, i.e.

$$C(u_1, \dots, u_n) = \prod_{i=1}^{n} u_i$$

A Gaussian copula is one that corresponds to the cumulative distribution function of a multivariate normal (i.e. Gaussian) distribution and is specified by the correlation matrix of the corresponding multivariate normal distribution. All multivariate normal distributions with the same correlation matrix have the same copula as the means and marginal standard deviations drive the marginal distributions rather than the copula. The independence copula is a special case of the Gaussian copula with the correlation matrix being the identity matrix.

The Archimedean family involves copulas of the following form, where $\varphi: [0,1] \to [0,\infty)$, $\varphi(0) = \infty$, $\varphi(1) = 0$, φ is continuous and strictly decreasing and $(-1)^k \, d^k \varphi^{-1}(t)/dt^k \geq 0 \quad \forall k = 0,1, \dots$

$$C(u_1, \dots, u_n) = \varphi^{-1}\big(\varphi(u_1) + \cdots + \varphi(u_n)\big)$$

Special cases include the Clayton copula which has $\varphi(t) = t^{-\theta} - 1$ (for some suitable value of θ) and the independence or product copula which has $\varphi(t) = -\log t$.

Exchangeable copulas are ones where we can permute the $u_1, \dots, u_n$ without altering the form of the copula. The Archimdean family are exchangeable. So is the bivariate Gaussian copula (since the correlation between x_1 and x_2 if (x_1, x_2) is bivariate normally distributed is the same as the correlation between x_2 and x_1). However, in general the multivariate normal distribution for $n > 2$ is not exchangeable. The correlation between x_1 and x_3 is not necessarily the same as the correlation x_2 and x_3 or the correlation between x_1 and x_2 for a trivariate normal distribution.

Exchangeability in effect corresponds to a one factor model for random variables that are multivariate normally distributed. A set of random variables with unit variance is said to possess a factor structure if their (symmetric non-negative definite) correlation matrix, V, is of the form:

$$V = AA^T + B$$

where V is an $m \times m$ matrix, A is an $m \times k$ matrix and B is a diagonal matrix. If the model is exchangeable then $correlation\big(x_i, x_j\big) = \rho$ for all $i \neq j$. The ρ also clearly needs to be in the range $[0,1]$. Setting:

$$A = \begin{pmatrix} \sqrt{\rho} \\ \vdots \\ \sqrt{\rho} \end{pmatrix} \qquad B = \begin{pmatrix} 1-\rho & \cdots & 1-\rho \\ \vdots & \ddots & \vdots \\ 1-\rho & \cdots & 1-\rho \end{pmatrix}$$

we obtain a factor structure with a single factor, so exchangeability implies a one-factor factor structure.

6.6 Tail dependence

Perhaps the most important reason in practice why use of copulas might be preferred over, say, merely using traditional multivariate distributions such as the Gaussian distribution is because it

allows us a richer range of models about how likely might be joint extreme events. This is generally introduced by reference to tail dependency.

If X_1 and X_2 are continuous random variables with copula $C(u_1, u_2)$ then their coefficient of (joint lower) tail dependence (if it exists) is:

$$\lambda \equiv \lim_{u \to 0} \frac{C(u,u)}{u}$$

For continuous random variables X and Y each with lower limit of $-\infty$ the coefficient of (lower) tail dependence is also:

$$\lambda = \lim_{z \to -\infty} Pr(Y < z | X < z) = \lim_{z \to -\infty} \frac{Pr(Y < z, X < z)}{Pr(X < z)}$$

For multivariate normally distributed random variables $\lambda = 0$ (unless the random variables are perfectly correlated in which case $\lambda = 1$ in the relevant joint directions). This is problematic if we are particularly concerned about the impact of joint extreme and e.g. believe that in extreme circumstances "all correlations might go to unity". Other, more general copulas allow a wider range of λ.

Please bear in mind the inherent uncertainty involved in the extrapolation involved in deriving coefficients of tail dependency (which mirror similar uncertainties arising with Extreme Value Theory, see next section of this Appendix). For example, λ may not exist. It may also differ according to the (joint) direction of the tail so e.g. the (joint) upper tail dependence $\lim_{u \to 0} \frac{1 - C(1-u, 1-u)}{u}$ may differ from the (joint) lower tail dependence $\lim_{u \to 0} \frac{C(u,u)}{u}$ and these two may both also differ from the tail dependences where u_1 and u_2 go to opposite extremes.

6.7 Simulating copulas

Correlated Gaussian (i.e. multivariate normal) random variables (i.e. random variables with a Gaussian copula and Gaussian marginals) can be generated using Cholesky decomposition.

For random variables that have a Gaussian copula but non-normal marginal (with cdfs $F_1, \ldots, F_n$) we can generate a vector $(x_1, \ldots, x_n)^T$ of correlated Gaussian random variables as above and then transform as per $y_i = F_i^{-1}(x_i)$.

In general, for non-Gaussian copulas we may need to generate a vector of unit uniform random variables $(u_1, \ldots, u_n)^T$ and then transform them using $u_1^* = u_1$, $u_2^* = C^{-1}(u_2 | u_1)$ etc.

6.8 Combining risk exposures using tail dependence

A natural potential use of copulas is to combine risk amounts for different risk types. If we were not using copulas then the aggregation would typically involve application of a correlation matrix (which technically involves use of a Gaussian copula), e.g. if the capital required to support the i'th risk is s_i then the total capital required to support all the risk combined is deemed to be $s = \sqrt{\sum_i \sum_j s_i c_{ij} s_j}$ where the c_{ij} are the assumed correlations between risks. The special case where $c_{ij} = 1 \ \forall i, j$ involves the total capital requirement being merely the sum of the individual capital requirements (and hence ignores any diversification benefits).

When combining risk capital amounts using copulas it is important to realise that in such computations capital is usually being set by reference to quantiles of an underlying probability distribution. We therefore need to combine the copula with some assumed marginal distributions, i.e. the copula in isolation provides insufficient information to permit a full aggregation of risk capital amounts. We might carry out the relevant calculations using simulations and then determine, say, the VaR from the combination of exposures based on these simulations. Please bear in mind that overall VaR calculated in such a manner may not respect usual axioms regarding diversification (see section 2) because the overall joint distribution may not then be coming from the elliptical family of distributions.

6.9 Selecting between copulas

If we decide to use copulas then it will generally be necessary to select between them. Typically a particular copula family deemed likely to be representative of the data is chosen with different members of the family being parameterised in a suitable fashion. It then becomes possible to use standard statistical distribution fitting techniques to choose which member of the family best fits the observed data. Two points to note are:

(a) If the fitting of the copula is separated from the fitting of the marginals then the copula will be wholly driven by the ranking of the observations, rather than by their magnitude. This means that non-parametric measures of correlation are more relevant than the traditional (Pearson) correlation coefficient. The two most commonly used are Spearman's rank correlation coefficient and Kendal's tau which conveniently happen to provide maximum likelihood estimators for certain specific (bivariate) copula family parameters. These measures of co-dependency are defined as follows for two underlying (paired) data series x_t and y_t each consisting of n observations:

Spearman's rank correlation coefficient:

$$\rho_{Spearman} = \frac{\sum_{t=1}^{n}(q_t - \bar{q})(r_t - \bar{r})}{\sqrt{\sum_{t=1}^{n}(q_t - \bar{q})^2 \cdot \sum_{t=1}^{n}(r_t - \bar{r})^2}} \quad where\ \bar{q} = \frac{1}{n}\sum_{t=1}^{n} q_t \ \ etc.$$

where q_t and r_t are the ranks within x and y of x_t and y_t respectively

Kendall's tau:

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)}$$

where computation is taken over all $i = 1,2,\dots,n$ and $j = 1,2,\dots,n$ with $i \neq j$ and (for the moment ignoring ties) a concordant pair is a case where $\left(x_i > x_j\ and\ y_i > y_j\right)\ or\ \left(x_i < x_j\ and\ y_i < y_j\right)$ and a discordant pair is a case where $\left(x_i > x_j\ and\ y_i < y_j\right)\ or\ \left(x_i < x_j\ and\ y_i > y_j\right)$.

There are various possible ways of handling ties in these two non-parametric measures of correlation (ties should not in practice arise if the random variables really are continuous).

(b) If the reason that copulas are being used is primarily to have the 'right' joint tail behaviour then it is important to be aware that fitting copulas based on measures such as the ones referred to above may give too much weight to the structure of the copula in the 'wrong' part of the distribution. This mirrors issues raised in section 2 with Cornish-Fisher expansions.

6.10 Further comments

Copulas have become the conventional more sophisticated way of handling co-dependency in cases where a multivariate Normal distribution seems inappropriate. In practice, however, they involve a considerable leap in technical complexity, making it harder to estimate their parameter inputs reliably. Regulatory frameworks often therefore revert to covariance (i.e. correlation) based methodologies with the covariance / correlation matrix adjusted in a prudent manner to achieve a desired outcome.

For example, the standard formula SCR for Solvency II involves a covariance-based risk aggregation (except for Operational Risk, the charge for which is added to the covariance-based combination of the others). A further reason for avoiding undue complexity in the step that incorporates co-dependency is that the individual risk capital charges may themselves be derived via nested stress tests (e.g. they may involve adopting whichever is the more onerous of an up or a down shift in the relevant input element). Thus, the inputs may no longer be easily translated into the marginal of a multidimensional probability distribution (except a rather artificial one), making the theoretical arguments for combining them using a copula somewhat weaker.

An over-focus on copulas also has other possible disadvantages, as explained in Kemp (2010):

- Visually, they have some undesirable consequences, in that it is not easy to tell from them where any divergence from Normality is most pronounced.

- Whilst it is mathematically valid to split a multivariate distribution into its marginal and its copula, there is no intrinsic reason to expect the impact on fat-tailed behaviour and hence propensity for extreme events to split in the same manner. Thus, they do not necessarily encourage the 'holistic' analysis of exposures that is a cornerstone of ERM.

7. Extreme value theory

7.1 Introduction

The topic of extreme events is particularly important for students and appliers of Enterprise Risk Management, particularly if their role has a strong focus on downside risk mitigation. Typically, the events that are most problematic and thus most likely to stick in the minds of the bosses or clients of such individuals are 'extreme' events, i.e. ones that are unusually severe. It is widely accepted that these occur more often than would be the case if the world behaved 'normally'.

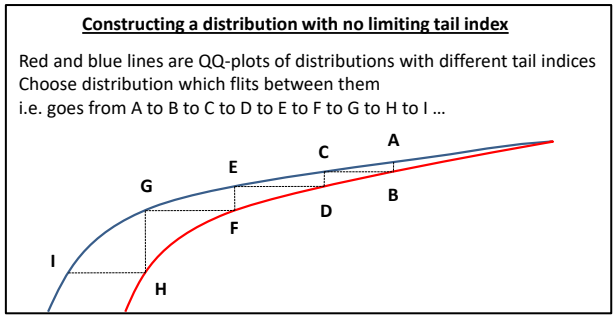
Extreme value theory (EVT) is a well-established branch of statistics that has been employed in insurance problems for many years but has only more recently been applied in a risk management context. The main contributions of EVT are to define and describe a set of limiting distributions which, if certain regularity conditions apply, characterise the limiting behaviour of the tail of a distribution.

Unfortunately, these regularity conditions are rather more restrictive than they appear at first sight. This means that EVT needs to be approached with some caution for some of the tasks to which it may be applied within the field of risk management. The reasons are explored further in Kemp (2010) and essentially arise because application of EVT involves *extrapolation* into the tail of a distribution.

If there is some inherent underlying physical process at work which we can reasonably assume operates stably through time then this type of extrapolation may have a strong theoretical underpin. For example, we might assume that the distribution of magnitudes of large earthquakes does exhibit

some sort of regularity given the geological processes involved. However, if the assumption of time stability is suspect, as might be the case with extreme events in financial markets driven primarily by human behavioural biases then the theoretical grounds for believing that EVT is a reliable extrapolation tool may be weaker. This of course depends on whether you think that the impacts of such behavioural biases are predictable in terms of their impact on financial markets.

There are certain other flaws with how EVT is typically treated in texts relating to ERM. For example, EVT is often developed by describing the three different EVT distributional forms (Gumbel, Fréchet and Weibull or, for VaR-style risk management purposes, more commonly corresponding variants of the generalised Pareto distribution) and then developing ways of identifying which one is relevant for the distribution in question. However, the tail behaviour may not actually converge at all, a point rarely highlighted in such texts. We show below a stylised way, involving quantile-quantile plots, of creating a distributional form which has no limiting tail behaviour of the sort required for EVT to apply. For traditional EVT to apply we also need the tail behaviour to converge in a specific way since it is also possible to extrapolate using any selected distributional family and not just the generalised Pareto distribution typically viewed as relevant within EVT, see e.g. Kemp (2013).



In what follows we will assume that (traditional) EVT can validly be applied to the problem in hand, i.e. that the relevant regularity conditions are satisfied.

7.2 EVT variants

There are two main variants of EVT.

(a) The first, involving *block maxima*, describes the behaviour of, say, the largest daily loss over a period such as a month and indicates how these block maxima converge asymptotically to a distribution with a relatively simple form.

(b) The second, involving *peaks-over-thresholds* (also called *threshold exceedances*) indicates that the distribution of losses over some threshold also converges, as the threshold is pushed out into the tail of the distribution, to a relatively simple form. This is the type of EVT most usually applied to VaR-style risk management problems.

More formally, suppose we have a set of portfolio losses, x_t, measured over time (and assumed to be independent). EVT provides two closely related sets of results relating to:

(1) Distributions of *block maxima*, i.e. m_n which is a random variable corresponding to the block maximum for blocks of x_t of length n, so the first realisation of m_n is given by $m_{n,1} = \max(x_1, \dots, x_n)$, the second by $m_{n,2} = \max(x_{n+1}, \dots, x_{2n})$ etc; and

(2) Distributions of *excesses*, i.e. $y_j \equiv x_j - u$ where u is a predetermined high threshold and the x_j are realisations that exceed u.

7.3 Block maxima results

The main result for block maxima is that if i.i.d. random variables x_i have cdf $F(x)$ and there also exist sequences $\{c_n\}, c_n > 0$ and $\{d_n\}$ and a cdf $H(x)$ such that:

$$\lim_{n\to\infty} Pr\left(\frac{m_n - d_n}{c_n} \leq x\right) = H(x)$$

where m_n is the random variable corresponding to the block maximum for blocks of such variables of length n then then F is said to be in the *maximum domain of attraction* (MDA) of H, written $F \in MDA(H)$ and (the Fisher-Tippett theorem) H comes from the generalised extreme value (GEV) family of distributions.

GEV distributions in general have three parameters, $GEV(\xi, \mu, \sigma)$, including a tail index, a location and a scale parameter. However if we replace c_i by $\tilde{c}_i = \sigma c_i$ and d_i by $\tilde{d}_i = d_i + \mu c_i$ we find that $F \in MDA(H_\xi)$ and $H = GEV(\xi)$, the one-parameter variant with $\mu = 0$ and $\sigma = 1$.

The cdf of $GEV(\xi, \mu, \sigma)$ is:

$$F(x) = \exp\left(-\left(1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right)$$

(or if $\xi = 0$ the limit of the above as $\xi \to 0$).

ξ defines the tail behaviour of the distribution. The sub-families defined by $\xi = 0$ (Type I), $\xi > 0$ (Type II) and $\xi < 0$ (Type III) correspond to the Gumbel, Frechét and Weibull families respectively.

Most well-known statistical distributions have well defined tail behaviours that are characterised by EVT. For example, the extreme values of normally and lognormally distributed random variables converge to Gumbel random variables (i.e. have $\xi = 0$), while Student's t and uniform random variables converge to the Frechét and Weibull distributions respectively (i.e. have $\xi > 0$ and $\xi < 0$ respectively).

7.4 Peaks over thresholds results

The second set of EVT results (relating to distribution of excesses) is probably more directly applicable to estimation of VaRs and the like. The main result here is the Pickands-Balkema-de Haan theorem.

Theorem. *Let F_u be defined as $F_u(y) = Pr(x - u < y | x > u)$ where $y = x - u$ for those cases $x > u$ and let x_F be the maximum limiting value of the random variable X then we can find a function $\beta(u)$ such that*

$$\lim_{u\to x_F}\left(\sup_{0\leq y<x_F-u}\left|F_u(y) - G_{\xi,\beta(u)}(y)\right|\right) = 0$$

if and only if $F \in MDA(H_\xi)$, i.e. under the same hypotheses as applied to the block maxima results.

Here $G_{\xi,\beta(u)}$ is the generalised Pareto distribution with zero mean, i.e. $G_{\xi,0,\beta}$and has:

$$G_{\xi,\beta} = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \xi = 0 \end{cases}$$

7.5 Estimating tail distributions using maximum likelihood

To use the results of the previous section to estimate VaRs we first need to fit a GPD to the data. We will suppose that the data comes from a Fréchet distribution (the Weibull distribution has a maximum upper limit which will not normally be consistent with financial market data of the sort typically used for VaR purposes). We therefore approximate the data using:

$$F_u \approx G_{\xi,\beta}$$

for some suitable $\xi \geq 0$ and $\beta > 0$.

One way of proceeding using maximum likelihood estimation (and assuming that the tail distribution is in line with a GPD) is as follows. Suppose we have n observations of which n_u is the number of excesses which have x_i exceeding the threshold u (for which y_i is then defined by $y_i = x_i - u$. We assume that the excesses are distributed according to $G_{\xi,\beta}$ and set $\tau = -\xi/\beta$. Then ML estimates are found by solving:

$$\hat{\xi} = \frac{1}{n_u} \sum_{i=1}^{n_u} \log(1 - \hat{\tau} y_i)$$

$$\frac{1}{\hat{\tau}} + \frac{1}{n_u}\left(\frac{1}{\hat{\xi}} + 1\right) \sum_{i=1}^{n_u} \frac{y_i}{1 - \hat{\tau} y_i} = 0$$

$$\hat{\beta} = -\hat{\xi}/\hat{\tau}$$

Suppose we define the complement of a cdf as $\bar{F}_u(y) \equiv 1 - F_u(y)$. We note that $\bar{F}(x) = \bar{F}(u)\bar{F}_u(y)$. Having estimated $\bar{F}_u(y)$ via maximum likelihood we are left with estimating $\bar{F}(u)$. We can do this by using the empirical tail, i.e.:

$$\bar{F}(u)_{ML} = \frac{n_u}{n}$$

Combining the estimates we find:

$$\bar{F}(x)_{ML} = \bar{F}(u + y)_{ML} = \frac{n_u}{n}\left(1 + \hat{\xi}\frac{y}{\hat{\beta}}\right)^{-1/\hat{\xi}}$$

From this we can estimate the p-quantile as:

$$\hat{x}_p = u + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{n}{n_u}(1-p)\right)^{-\hat{\xi}} - 1\right)$$

The only remaining problem is to choose the threshold u, ideally sufficiently far into the tail that the limiting tail distribution has largely been reached but not so far that we have little or no observations left in the tail. Typically this is done empirically using e.g. graphical methods. One such method relies on the mean excess function, $e(u) \equiv E(x-u|x>u)$. For $G_{\xi,\beta}$ this has $e(u) = \frac{\beta+\xi u}{1-\xi}$ so in the tail should be linear. We can thus use the empirical mean excess function defined below to investigate the choice of u:

$$e_n(u) = \frac{1}{n_u} \sum_{i \in S_u(u)} (x_i - u)$$

where $S_u(u) = \{i: x_i > u\}$.

7.6 Hill estimator techniques

An alternative and somewhat mathematically simpler approach is based on the Hill estimate of the tail parameter $\alpha = 1/\xi$. We again assume that the limiting tail distribution is the Fréchet distribution and we assume that the GPD can be approximated by the Pareto distribution, $F(z) = 1 - x^{-\alpha}$. The maximum likelihood estimator for α, known in this situation as the Hill estimator, is:

$$\hat{\alpha}^{(H)} = \hat{\alpha}^{(H)}_{k,n} = \left(\frac{1}{k} \sum_{j=1}^{k} \left(\log x_{(j)} - \log x_{(k)} \right) \right)^{-1}$$

where the $x_{(j)}$ are the ordered observations and there are k $(= n_u)$ observations in the tail of the distribution (k depends on u).

An estimator for the p-quantile is then:

$$\hat{x}_p^{(H)} = x_{(k)} \left(\frac{n}{k}(1-p) \right)^{-1/\hat{\alpha}^{(H)}_{k,n}}$$

As with the direct maximum likelihood approach, a crucial choice to make in computing the estimator is the threshold u and hence the value of $k = n_u$. One graphical approach to the selection of u is to examine so-called Hill plots which involve plotting $\hat{\alpha}^{(H)}_{k,n}$ against k and selecting the 'optimal' k (that best trades-off bias versus variance of the estimator) as the largest value for k for which the $\hat{\alpha}^{(H)}_{k,n}(k)$ seems to be constant.

7.7 Generalisations

EVT as described above relies on specific regularity conditions which do not always apply in practice. If it is known that the tail can be approximated by some suitable member of a given distributional family then the tail behaviour can be derived by selecting the relevant family member using standard statistical criteria such as maximum likelihood. Traditional EVT as above can in effect be viewed as a special case of this approach, using the generalised Pareto (or the generalised extreme value) distributional family. If members of the selected distributional family can be manipulated relatively easily (e.g. via a suitable numerical package) then there may be little practical benefit in limiting oneself to the GPD, if the regularity conditions needed for it to apply are suspect. For further details see e.g. Kemp (2013).