Table of Contents

last update: June 24, 2016

B. Market Risk Measurement

B01. Value at Risk – Overview

Under the Basel regime for banking regulation, Value-at-risk (VaR) is the de-facto risk model for the computation of regulatory capital requirements. However, VaR has been severely criticized, especially after the 2008 credit crisis, because many assumptions behind the model failed during that stressful period; it was found VaR understated risks at exactly the times that it was needed most.

Due to Basel's endorsement, the use of VaR and VaR-like models is widespread in many areas in banking regulation. In the banking book, we have the IRB (internal ratings based) model for credit risk and the IRRBB (interest rate risk in banking book) model for interest rate risk. In the trading book, we have the VaR, stressed VaR, and IRC (incremental risk charge) models for most trading operations, and the specific risk model and CRM (comprehensive risk model) for correlation trading, and in operational risk, we have the OpVaR model. These actuarial based or "VaR-like" models always involve sampling an empirical distribution or specifying an assumed distribution. VaR is a statistical measure of risk based on the loss quantile of such distributions. It measures risk based on a total portfolio basis taking into account diversification.

Generally, banks use VaR models for two purposes: First, VaR is a model for the calculation of regulatory minimum capital requirements. Since this so-called "minimum" capital is required to protect against crisis events that would threaten the survival of the bank, it seems prudent to set the confidence level to be extremely high, indeed typically at 99% or higher. Second, VaR is used for day-to-day risk management control, setting of risk limits and risk attribution analysis, whereby a lower confidence level is acceptable and indeed desirable since it will afford a more precise estimation of VaR. It is important to understand that this is not only a risk scaling exercise, as different asset classes will be impacted differently by those quantile changes. As a general rule, the further one goes into the (high quantile) tail, the riskier appear say AAA-rated securities when compared to B-rated one's.

Definition of VaR

VaR is (an estimate of) one single number representation of risks for the entire distribution. Specifically, it is a quantile. This simplification makes risk management and reporting easier. We define VaR an estimate of the loss from a fixed set of trading positions over a fixed time horizon (typically 1 day or 10 days) that would be equal or exceeded with a given probability. Mathematically, the VaR is defined by the probability statement

$$\mathbb{P}(V_T - V_0 \le \text{VaR}) = 1 - \alpha \tag{1}$$

where α is the confidence level, V_t the value of a portfolio under review at time t. We use $V_T - V_0$ to denote the portfolio gain or loss (or P&L, short for 'profit and loss') over the specified time horizon T. For example, if a portfolio has a 10-day 99% confidence level VaR of -$1 million (note that the VaR here is assumed to be negative, i.e. the smaller the VaR is, the higher the risk; often people also quote VaR as a positive number) as of today, it tells that there is a probability $(1 - \alpha)$ of

1% that the portfolio may lose by \$1 million *or more* over the next 10 days. Several details of the VaR definition are worth mentioning:

1. VaR is an estimate, not a uniquely defined value. In theory, the value of any VaR estimate depends on the stochastic process that is driving the random realizations of market data. For more sophisticated VaR models, this data generating process has to be identified (or modeled) and its specific parameters calibrated. This requires resorting to historical experience which raises many practical issues such as the length of the historical sample used and whether more recent events should be weighted more heavily than those further in the past. In practice, market data is also not generated by stable, long running random processes because there are what is often referred to as "regime changes". For example, the market state during the crisis period is different from that of the post or pre crisis period. A model that is not able to capture the dynamic nature of the market will be "too little, too late" in capturing risks. Differing methods for dealing with the uncertainty surrounding changes in regimes are at the heart of why VaR estimates are seldom unique.
2. The trading positions are assumed to be fixed over the forecast risk horizon (say 10 days for regulatory VaR reporting). This can be unrealistic in an investment banking or trading portfolio setting, where trades are bought and sold at a high turnover rate on a daily basis. In practice, simple scaling rules are used to scale estimates 1-day VaR to the longer risk horizon. Without this simplification, it will be necessary to model what happens within the specified time horizon and make behavioral assumptions relating to trading strategies during the period.
3. VaR does not address the distribution of potential losses on those occasions when the VaR estimate is exceeded i.e. it is oblivious to the tail losses beyond VaR. Hence VaR is not the 'worst-case loss' or 'expected loss'. In fact, VaR is the *minimum* loss given the probability.

While the VaR figure itself is the primary focus for regulatory reporting and limits monitoring, banks also look at other VaR related measures to assess how additional/component positions affect risks and diversification. We will discuss this risk decomposition next.

Marginal VaR

A trading portfolio may contain many component positions of different instruments and products. In order to analyze VaR and its risk contribution from each component, risk managers often attempt to "breakdown" the portfolio VaR in a process that is called *VaR decomposition*. For example, we have a portfolio whose value as of today is V_p. The portfolio consists of N components in which the value of the i-th component asset is V_i. Risk managers often like to know the contribution of risk from specific components to the portfolio total VaR. The marginal VaR is one of such measures. It is defined as the partial derivative of portfolio VaR with respect to the component asset value

$$\mathrm{MVaR}_i \equiv \frac{\partial \mathrm{VaR}_p}{\partial V_i} \tag{2}$$

The marginal VaR describes the change in the total VaR resulting from “one” dollar change of the component value.

Before we discuss further about marginal VaR, this is a good place to introduce the basic notion of *modern portfolio theory* (MPT) and see its relationship to VaR. The portfolio return R_p for a given time horizon is the weighted sum of component asset returns R_i for $i = 1, \cdots, N$

$$R_p = \sum_{i=1}^{N} w_i R_i \tag{3}$$

where $w_i = V_i / V_p$ is the weight of the i-th component asset and R_i is the i-th component asset return. The summation formula above can also be written succinctly in matrix notation.

$$R_p = w^T R \tag{4}$$

where w is a column vector of the weights and R is a column vector of the returns. In the above formula, we use a superscript ‘*T*’ to denote a matrix transpose operation. MPT assumes that the returns of component assets follow a multivariate Gaussian distribution, which implies that the portfolio returns follow a Gaussian distribution as well. The following formula gives the the variance of the portfolio return σ_p^2

$$\sigma_p^2 = w^T \Sigma w = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \sigma_{ij} \tag{5}$$

where Σ is the covariance matrix of returns of component assets. The entry σ_{ij} of matrix Σ is the covariance term between the returns of the i-th and j-th component assets and it becomes a variance term if $i = j$. Assuming the portfolio returns follow indeed a Gaussian distribution, the VaR can be easily derived from the return variance σ_p^2 that is

$$\mathrm{VaR}_p = z \sigma_p V_p \tag{6}$$

where z is the $(1 - \alpha)$ percentile of a standard normal (note that VaR tends to exclude the Expected Loss component, hence for simplicity we assume a zero mean for our Gaussian distribution). For example, if the confidence level $\alpha = 99\%$, then $z = -2.326$. You can compute this using the Excel function NORMSINV$(1 - \alpha)$.

Given the relationship in (6), we can derive the marginal VaR from (2) as

$$\mathrm{MVaR}_i \equiv \frac{\partial \mathrm{VaR}_p}{\partial V_i} = \frac{\partial \left(z \sigma_p V_p \right)}{\partial \left(w_i V_p \right)} = \frac{z \partial \sigma_p}{\partial w_i} \tag{7}$$

where we use $V_i = w_i V_p$ in the denominator. The partial derivative in (7) can be further derived as follows. Firstly we differentiate (5) with respect to vector w, this gives a first derivative in a row vector form written as

$$\frac{\partial \sigma_p}{\partial w} = \frac{1}{\sigma_p} w^T \Sigma \tag{8}$$

Its i-th component is calculated as

$$\frac{\partial\sigma_p}{\partial w_i} = \frac{1}{\sigma_p}\sum_{j=1}^{N} w_j\sigma_{ji} = \frac{\sigma_{ip}}{\sigma_p} \tag{9}$$

where we denote the term

$$\sigma_{ip} = \sum_{j=1}^{N} w_j\sigma_{ji} \tag{10}$$

the covariance between the returns of the i-th component asset and the portfolio. Since the covariance can be regarded as a product of two volatilities along with the correlation between them, we can write it as

$$\sigma_{ip} = \sigma_i\sigma_p\rho_{ip} \tag{11}$$

where σ_i is the return volatility of the i-th component asset and ρ_{ip} is the correlation of returns between the i-th component asset and the portfolio. After substituting (9) and (11) into (7), the marginal VaR reads

$$\mathrm{MVaR}_i = z\frac{\sigma_{ip}}{\sigma_p} = z\sigma_i\rho_{ip} = \frac{\mathrm{VaR}_i}{V_i}\rho_{ip} \tag{12}$$

We follow the *capital asset pricing model* (CAPM) and define a sensitivity term β_{ip} between the portfolio and its i-th component asset as the ratio between the component covariance and the portfolio variance

$$\beta_{ip} = \frac{\sigma_{ip}}{\sigma_p^2} \tag{13}$$

the marginal VaR in (12) can then be expressed as

$$\mathrm{MVaR}_i = z\sigma_p\beta_{ip} = \frac{\mathrm{VaR}_p}{V_p}\beta_{ip} \tag{14}$$

The above equations show that the marginal VaR of a component is the product of (a) the percentage VaR of the overall portfolio, and the component beta against the full portfolio. Note that the component beta defined in (13) can be negative (zero) in case the component has offsetting risks with (is uncorrelated to) the portfolio. Generally, the beta of a long and a corresponding short position will be of equal size but of opposite sign so those component risks will cancel out in the overall risk position.

When we use (14) to calculate marginal VaR, we implicitly assume the asset returns follow a multivariate Gaussian distribution. While the assumption of Gaussian asset returns is questionable, the assumption that portfolio returns follow a Gaussian distribution is more defensible, at least in quiet markets: because of the central limit theorem the average of a large number of similar sized independent random variables with finite variance converges to a Gaussian distribution. However, during a crisis there are two of assumptions that can be violated: (a) the underlying variable might not be of finite variance, and (b) the co-dependence structure can the strongly non-Gaussian and/or correlations are so large that the CLT does not apply (yet). If the assumption does not hold, the MVaR calculated by (12) or (14) gives only approximate values.

Incremental VaR

Besides the marginal VaR, there is another risk measure called 'incremental VaR', which measures the change in VaR due to a new position added into the portfolio. Mathematically, it is defined as

$$\mathrm{IVaR}_a \equiv \mathrm{VaR}_{p+a} - \mathrm{VaR}_p \tag{15}$$

where VaR_{p+a} is the VaR of a portfolio p plus position a, and VaR_p is the VaR of the portfolio before including the position a. In the notation above, all the VaR values are negative. Clearly when $\mathrm{IVaR}_a > 0$, then the position a contributes by increasing overall diversified VaR by the amount IVaR_a (making the VaR number less negative). In other words, if $\mathrm{IVaR}_a > 0$, the added position is risk reducing, it hedges some of the portfolio risk by the amount IVaR_a. Conversely, if $\mathrm{IVaR}_a < 0$, then the position is risk increasing.

Marginal and Incremental VaR are closely related: the Marginal VaR is the Incremental VaR for the proverbial *one-dollar-increment*. This is meant in percentage terms: if the MVaR of a position is say 5% then the IVaR for a small increase will be 5% as well – dollar amounts scale with the size of the increase. IVaR is always smaller (ie, more negative) than MVaR: the reason for that is that the IVaR effectively uses the *average beta* of the position over the range of the increment, and beta is a strictly increasing function of the component size. This is easy to see: the bigger the component, the more important it becomes as part of the portfolio; in the limit of very large sizes the beta of any component will be plus one, even for a position that started out at a negative beta.

Component VAR

A third risk measure is the component VaR (CVaR) which has the nice feature of being additive (unlike the above measures)—the component VaRs add up to the portfolio VaR. This technique of risk decomposition is thus often used for management information and capital allocation because it is more intuitive, convenient and easier to understand. For example, this is often used by the treasurer to allocate trading budget and to measure risk-adjusted performance of individual desks. However, this method should not be used for risk management and hedging as it ignores the reality of risk diversification effects in portfolios. Component VaR, in this case, partitions the portfolio VaR into parts that add up to the total diversified VaR. By definition, the component VaR can be expressed in terms of the marginal VaR we discussed previously

$$\mathrm{CVaR}_i \equiv \mathrm{MVaR}_i V_i = \frac{\partial \mathrm{VaR}_p}{\partial V_i} V_i \tag{16}$$

Again if we assume asset returns follow multivariate normal distribution, we can use the marginal VaR expression in (14) to write the component VaR as

$$\mathrm{CVaR}_i = \mathrm{VaR}_p w_i \beta_{ip} \tag{17}$$

Since we know the portfolio variance is a weighted sum of component covariances

$$\sigma_p^2 = \sum_{j=1}^{N} w_i \sigma_{ip} \tag{18}$$

and by the definition of β_{ip} in (13), we must have

$$\sum_{j=1}^{N} w_i \frac{\sigma_{ip}}{\sigma_p^2} = \sum_{j=1}^{N} w_i \beta_{ip} = 1 \tag{19}$$

and therefore

$$\sum_{i=1}^{N} \mathrm{CVaR}_i = \mathrm{VaR}_p \tag{20}$$

It shows the component VaRs sum to the total VaR of a portfolio.

Basic Concepts and Definitions

The VaR system used in a bank is nothing more than an aggregation engine to find the portfolio level (or joint) P&L distribution, and then compute the quantile loss at the portfolio level. However, even before the risk aggregation step, many upstream preprocesses need to happen. Such steps include market data capture and cleaning, risk factor mapping, positional capture, full revaluation of positions under various scenarios, etc. Before exploring the various VaR methodologies or systems, let's first introduce these relevant preprocessing concepts.

Mapping Position to Risk Factors

The first step in building a risk management system is the mapping of risk factors—in which a superset of risk drivers is identified and mapped to a *subset* of risk factors. Why do we need to reduce (and in some cases simplify) the number of risk factors that goes into the VaR model? A portfolio P&L is the sum of P&L of all the deals in the portfolio. The P&L of each deal can be derived from observing the daily changes in market prices of the deals (i.e. by marking-to-market). And so, in theory, one can analyze the risk of a portfolio by looking at changes in P&L contributed by each deal. In practice, banks analyze the risk of a portfolio by looking at risk factors that drive the changes in portfolio P&L. In other words, the P&L that is used for VaR and risk management is computed theoretically as a function of risk factors, instead of marking the portfolio to market. Given a set of risk factors, most trade-booking systems used by banks have pricing libraries that allow the computation of the present value (PV) and by extension the P&L of each deal. Thus, we project our positions onto a relatively small set of risk factors. This process of describing positions in terms of standard risk factors is known as 'risk factor mapping'. Modeling risk by risk factors has many advantages:

1. It allows proxying. We might not have enough historical data for some positions. For instance, we might have an emerging market security that has a very short data history or we might have a bespoke OTC (over the counter) instrument that has no price observation at all. In such circumstances it may be necessary to map our security to some comparable index or proxy asset, which does have sufficient data for modeling purposes.

2. It reduces the dimension of the problem. A typical bank may have hundreds of thousands of deals mapped to a smaller subset of risk factors. This greatly reduces the necessary computer time to perform risk simulations. In effect, reducing a highly complex portfolio to a consolidated set of risk-equivalent positions in basic risk factors simplifies the problem, allowing simulations to be done faster. The reduction in the dimension also improves the precision of the tail measures such as VaR.
3. It is more natural for the purpose risks analysis to decompose portfolio risks in terms of its risk factors. Often these risk factors are uniquely driven by changes in specific macroeconomic factors. For example, central bank policy actions and expectations have a direct impact on interest rates risk factors. Other risk factors may be less affected here.

As an example we consider an FX option. According to the Black-Scholes valuation formula, an FX option is a function of mainly the following risk drivers: FX spot rate, interest rates and volatility. Hence, even if a portfolio contains a thousand option deals (of the same currency) they can all be represented (or mapped) to just these three types of risk factors. This provides a great deal of efficacy to the business of risk management.

As a note, while it is true that every forex option can be correctly priced using its *implied volatility* (i.e. the volatility that, when plugged into the Black Scholes formula, backs out the correct price). However, this leaves us with as many implied vols as we have options. The big step to take here is to only model a small (but sufficient) number of volatilities across the entire options portfolio.

Scenario Generation

VaR is derived from a distribution of P&L, and is its quantile. In practice this distribution is made up of scenarios sampled from history in the so-called "historical simulation VaR" approach. In contrast, *parametric* approaches exist in which a *theoretical* distribution of P&L is assumed (but this is far less popular in banks, see next section for the VaR methods). The scenarios are generated from historical time series of risk factors where each risk factor comes in the form of a daily time series of level data i.e. prices or rates. Since VaR relies on *returns* scenarios to form the P&L distribution, the series of level data must be transformed into series of returns.

As a typical example, we can choose a 500-business day rolling observation period (or window) representing two calendar years. So each risk factor has a return series represented by a scenario vector of length 500. Once we have derived the return scenarios, we can apply the return series to estimate a parametric distribution for P&L and then compute the VaR analytically. The estimated parameters could be the moments of the distribution for example. Alternatively, we can use the scenario vector to generate P&L distribution empirically (non-parametric way) and then take its quantile as VaR. For instance, we can use a scenario to shift the current levels (or base levels) of risk factors to the shifted levels. Assets in a portfolio are then revalued at the current levels and the shifted levels respectively. The difference between the two valuations is then the P&L for that scenario. In summary, the set of scenarios computes and gives a P&L distribution; we often call this the 'P&L vector'. The VaR is then the empirical quantile of the resulted P&L vector.

Importantly, there are three common types of returns that can be computed from the level data, the results are the scenarios derived from history: 1) absolute return, 2) relative return and 3) log return. The absolute return takes the difference between two levels

$$\text{absolute\_return}(i) = \text{level}(i) - \text{level}(i-1) \tag{21}$$

where $i = 1, \cdots, 500$ is the scenario number. The relative return is a percentage change given by

$$\text{relative\_return}(i) = \frac{\text{level}(i)}{\text{level}(i-1)} - 1 \tag{22}$$

The log return takes the natural logarithm of the ratio of the two levels

$$\text{log\_return}(i) = \ln \frac{\text{level}(i)}{\text{level}(i-1)} \tag{23}$$

The relative return asymptotically converges to the log return as the two levels get closer to each other. They are also approximately the same for small perturbations. Relative return or log return are suitable for assets that trade based on price (e.g. exchange rates, price indices, and stock market indices, the big advantage being that in this case prices never get negative. For assets that trade on yield (e.g. interest rates, bond yields) often absolute return is a better representation.

For example, the S&P 500 would be modelled using relative or log returns, whilst bond yields might be modelled using absolute returns (a caveat being that this gives non-zero probability to negative yields, but in practice this probability is usually very small so this is acceptable; besides, as of 2014 some government bonds are trading on negative yields).

Risk Sensitivities (Greeks)

Risk sensitivity is an important topic for risk management, not just because they are used for limits monitoring/ control of risk taking, but also because they are used in parametric methods of computing VaR. Risk sensitivities or 'Greeks' measure the change in the present value of a position due to a specified change in the risk factor that the position is exposed to. They are often used in risk management and control, e.g. to hedge portfolios, and to set and assess risk limits. They are also often used in VaR calculations as they allow to approximate the price changes of a portfolio under the VaR scenarios in a computationally efficient manner. Here we will consider a few basic types of sensitivities, which are listed in Table 1.

Table 1. Risk Sensitivities

Sensitivity	*Type*	*Definition*	*Application*
Delta (δ)	First Order	P&L due to a small change in price	All derivatives based on assets that trade on price (e.g. equities, FX)
Gamma (γ)	Second Order	Second order P&L correction due to a small change in price	

Vega ($\mathcal{V}$)	First Order	P&L due to change in volatility (typically 1 point change)	All (non-linear) derivatives
PV01	First Order	P&L due to +1 basis point change in rate	All derivatives based on assets that trade on yield (e.g. swaps, bonds)
Convexity	Second Order	Second order P&L correction due to +1 basis point change in rate	
CR01	First Order	P&L due to +1 basis point change in credit spread	All derivatives based on credit assets

If we have a scenario we want to know the P&L generated by each of the positions in our portfolio. We fundamentally have two options for this: (1) we can run a full revaluation of every single position based on those new parameters, or (2) we approximate the P&L impact by developing the P&L using a Taylor expansion (note that Greeks are very closely related to partial derivatives, mathematically speaking). For example, if we have a 10-year fixed coupon bond with semi-annual coupon payments, we may price the bond using the usual cost-of-carry formula

$$V(y) = p\left(1+\frac{y}{2}\right)^{-10} + \sum_{i=1}^{10} c\left(1+\frac{y}{2}\right)^{-i} \tag{24}$$

where p is the par value paid upon maturity, c is the fixed coupon cashflow and y is the bond yield rate. The (symmetric) first order risk sensitivity, PV01 of the bond, is defined as

$$\text{PV01} = \frac{V(y+0.5\text{bp}) - V(y-0.5\text{bp})}{1\text{bps}} \tag{25}$$

Note that the actual perturbations used are not always 1bp wide, but they are generally normalized back to this level. The second order sensitivity (convexity)[6] of the bond (with a +- 1bp perturbation) can be computed using the following second order central difference formula

$$\text{CONVEXITY} = \frac{V(y+1\text{bp}) - 2V(y) + V(y-1\text{bp})}{(1\text{bp})^2} \tag{26}$$

In first order, the bond P&L can be approximated using

$$\text{P}\&L \approx \text{PV01} \times \Delta y \tag{27}$$

where Δy is the scenario change of yield rate in basis points (bps). However, a linear approach is only reliable when the products' payoff is linear or close to linear. For examples, forwards, futures and swaps have values that are almost linearly dependent on the values of the underlying assets (i.e. the risk factors). If our positions have considerable optionality or other nonlinear features, such in the case of options or exotic products, linear approximations can be very unreliable. In this case, we can try to accommodate nonlinearity by also including the second-order (gamma or convexity) term of the Taylor expansion. This is sometimes called the delta-gamma approximation. For the bond example, we can get better P&L approximation by including the convexity correction term

$$\text{P\&}L \approx \text{PV01} \times \Delta y + \frac{1}{2} \times \text{CONVEXITY} \times (\Delta y)^2 \tag{28}$$

Using the sensitivity approach in our P&L calculation means that we do not perform full-revaluations of a deal using its pricing formula repeatedly, which may incur a heavy computation load. Instead we get an approximation of the P&L by multiplying the deal's risk factor sensitivity (which just needs to be computed once) with the corresponding risk factor's scenario return. In fact – we even have to only compute the Greeks once at a portfolio level, and then from this point onwards we can ignore the actual number of positions in our portfolio.

Using Greeks is a big computational advantage over the full revaluation approach, and can be used without loss of accuracy for moves that are small enough for the first- or second order approximation to be sufficiently precise. It is also possible (albeit complicated, especially if non-trivial cross sensitivities are involved) to use higher order terms. The main problem here is when payoffs are digital of nature (e.g., barrier options) where close to the boundary any Taylor expansion will fail.

Distributional Assumption and Volatility Estimation

When computing a tail risk measure such as VaR, it is necessary to make assumptions about the distribution of portfolio returns. A widespread assumption is that the (log) returns of the assets for any given period form a joint Gaussian distribution, and that they are independent from each other and identically distributed for non-overlapping periods. A one-dimensional Gaussian distribution can be uniquely described by only two parameters: its mean and its variance (for a Gaussian vector, the mean is a vector, and the covariance is a symmetric matrix).

However, looking at real financial time series, we often find that their distributions are heavy tailed and skewed – they are not Gaussian. The true probability of a very large return – especially on the downside, for assets where this notion makes sense – at the tails is greater than the one estimated under a Gaussian distribution with same mean and variance. This finding challenges the measurement and the use of VaR at high confidence levels. If VaR is calculated under the assumption of a Gaussian distribution and yet the actual markets are heavily tailed, then VaR will understate the true risk during crisis periods. Since VaR is used for the computation of required regulatory capital, the available capital of banks may be insufficient to withstand losses when disaster strikes. The naïve solution to this is to increase the multiplier relating capital requirement to the VaR measure which – especially for market risk VaR – is an arbitrary number anyway. The point though is that without having a detailed view on how *taily* a distribution is, it is difficult to impossible to understand whether or not that multiplier appropriately sized.

One area where the above assumptions fall short is that in real market big moves tend to follow big moves, meaning that are the very least there is a co-dependence of the variances (volatilities) at two adjacent points in time: high volatility follows high volatility and vice versa. The quickest way to identify volatility clustering is to plot the return series and to visually check for clusters (as appeared in Figure 1). More formally, one can test for autocorrelation of squared returns. During times of stress, financial data often exhibit significant positive autocorrelation in squared returns. Since an increase in volatility heightens the probability of large returns, it will make the empirical distribution of the return appear more heavily tailed. Under this view fat tails are not an intrinsic feature of the distribution, but are a result of a changing (or stochastic) volatility, but at any given point in time the instantaneous distribution is still Normal and independent (other than via its volatility parameters) from the instantaneous distribution at other points in time.

The volatility of risk factors determines the P&L distribution, and therefore has substantial influence on the VaR. To improve the forecasting power of the VaR, we must make use of historical observations that best represents the current market variation. The simplest measure for the volatility is the standard deviation s (we won't get into the discussion of whether to normalise with N or N-1 here – in practice this is largely irrelevant). Its square gives the variance defined by

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})^2 \tag{29}$$

which assigns equal weights to historical observations. In practice, this measure loses its forecasting power if the return distribution changes (i.e. is not constant) in time over the observation period.

Volatility clustering indicates that the asset returns are not independent. Although it is not possible to predict the direction of the returns based on historical returns, it is possible to predict their volatility. If we want to capture the volatility clustering phenomenon in VaR calculation, we can estimate the *conditional* volatility, that is, the volatility conditional on the recent past. We will discuss two widely used methods for this: 1) exponentially weighted moving average (EWMA) and 2) generalized autoregressive conditional heteroskedasticity (GARCH) models. The term "heteroskedasticity" here refers to non-constant volatility in a return series. The GARCH model is more sophisticated and difficult to implement but offers some potential advantages.

EWMA model was initially proposed by JP Morgan's Riskmetrics© in 1994 [1]. It quickly became a popular benchmark after Basel adopted VaR as the de facto model for risk capital under its 'internal models' approach. EWMA estimates the volatility by assigning heavier weights to recent observations than those from the distant past. The EWMA volatility forecast σ_i for day i is given by the recursive equation:

$$\sigma_i^2 = \lambda\sigma_{i-1}^2 + (1-\lambda)r_{i-1}^2 \tag{30}$$

where λ is the decay factor that determines how rapid the weight decays as an observation goes into the past. Notice that since λ is positive, today's variance will be positively autocorrelated with yesterday's variance, so we see that EWMA captures the idea of volatility clustering. The parameter λ may also be seen as a 'persistence' parameter. The higher the value of λ, the more persistently high (low) variance will lead to high (low) variance. Riskmetrics proposed $\lambda = 0.94$ in their daily volatility calculation for the stock market; this value of λ gives a volatility forecast closest to the realized ones in history.

On way of estimating the value of λ is via the maximum likelihood estimation (MLE) method. This method assumes a parametric distribution (e.g. normal or Student t distribution) for the return series. The idea is to try to find an optimal λ such that the computed σ series maximizes the probability of the realization of the observed return series. For example, if we assume normal distributions for the returns, i.e. $r_i \sim N(0, \sigma_i^2)$, the time-dependent conditional variances σ_i^2 are given recursively by EWMA model (30). The occurrence of the return series would have a probability that is proportional to the product of the probability density functions (PDF) of the series, which is called likelihood function $\mathcal{L}$. We can thus calibrate the parameter λ to match the observations in history. The likelihood function is given by

$$\mathcal{L} = \prod_{i=1}^{N} \varphi(r_i; \sigma_i^2) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{r_i^2}{2\sigma_i^2}\right) \tag{31}$$

where $\varphi(r_i; \sigma_i^2)$ is the PDF of a normal distribution with mean of zero and variance of σ_i^2. We have assumed a zero mean for returns for simplicity. The MLE finds the optimal λ such that the resulted σ_i series maximizes $\mathcal{L}$. This is equivalent to maximizing the natural logarithm of the likelihood function $\ln \mathcal{L}$ because log transformation is monotonous

$$\ln \mathcal{L} = \sum_{i=1}^{N} \left(-\ln \sigma_i - \frac{1}{2}\ln 2\pi - \frac{r_i^2}{2\sigma_i^2}\right) \tag{32}$$

GARCH models are similar to EWMA in that both are designed to address the issue of volatility clustering. They are natural extension of the autoregressive conditional heteroskedasticity (ARCH) models proposed by Engle (1982) [2] by including an autoregressive moving average (ARMA) model for the error variance. This generalization makes the GARCH model very flexible and the numerous free parameters allow the model to calibrate to various behaviors and characteristics of a particular market. All GARCH models share a common feature that is yesterday's risk is positively correlated with the today's risk, i.e. there is an autoregressive structure exists in risk. The GARCH model or more accurately GARCH(p,q) model has a general form

$$\sigma_i^2 = \omega + \sum_{k=1}^{p} \alpha_k \sigma_{i-k}^2 + \sum_{k=1}^{q} \beta_k r_{i-k}^2 \tag{33}$$

where we have the model parameters $\omega > 0$ and $\alpha_k, \beta_k > 0$ for $k > 0$ to ensure strong positivity of the conditional variance, and we also require that

$$\sum_{k=1}^{p} \alpha_k + \sum_{k=1}^{q} \beta_k < 1 \tag{34}$$

to ensure stationarity of the conditional process; otherwise the model is intractable (unsolvable). The lag lengths p and q in the model specification define the order of the dependence of current volatility on the past information. Hence the recursive definition in the model allows a non-constant volatility conditional on the volatilities and return realizations in the past. Again, we assume a zero mean for returns for simplicity. When we set $p, q = 1$, it gives the simplest GARCH model, known as GARCH(1,1) popular in financial applications

$$\sigma_i^2 = \omega + \alpha\sigma_{i-1}^2 + \beta r_{i-1}^2 \tag{35}$$

Our further discussion will mainly focus on this simple model. To estimate the parameters ω, α and β in GARCH(1,1) model, we can again use maximum likelihood estimation. For example, in most GARCH models, the returns are assumed to follow a normal distribution specified by mean of zero and the conditional variance series σ_i^2, that is $r_i \sim N(0, \sigma_i^2)$ for $i = 1, \cdots, 500$. By using MLE, the optimal estimation of model parameters is such that the resulting σ_i^2 series maximizes the likelihood function $\mathcal{L}$ as shown in (31) (or $\ln \mathcal{L}$ in practice). As one can see, the EWMA model in (30) is actually a special case of the GARCH(1,1) model in (33). GARCH(1,1) extends EWMA by adding a constant term ω and relaxing the constraint that the coefficients $(\alpha + \beta)$ has to sum to one. In fact, if the sum $(\alpha + \beta)$ is less than one (the more usual case), it does have an implication that the volatility is mean-reverting and the rate of mean reversion is inversely related to this sum. This means, unlike the EWMA model, the conditional variance in GARCH(1,1) model, in the absence of a market shock, will drift towards its long-term variance defined by

$$\sigma^2 = \frac{\omega}{1 - \alpha - \beta} \tag{36}$$

In the attached spreadsheet [VaR_Volatility_Models.xls], we demonstrate the calibrations of the EWMA and GARCH(1,1) model to the three years series (from 3 Jan 2011 to 31 Dec 2013) of daily log-returns of S&P500 equity index. The return series shown in Figure 1 revealed clustered volatilities across time. The models are calibrated to the return series by MLE where the Excel Solver is used to perform the optimization. The calibrated models are as follows (the numbers are the calibrated parameters):

EWMA model:

$$\sigma_i^2 = 0.9222 \times \sigma_{i-1}^2 + 0.0778 \times r_{i-1}^2$$

GARCH(1,1)model：

$$\sigma_i^2 = 0.000004334 + 0.8197 \times \sigma_{i-1}^2 + 0.1357 \times r_{i-1}^2 \tag{37}$$

The long-term (or steady-state) volatility in the GARCH(1,1) model is 0.986% as per equation (36), which is quite close to the sample standard deviation (of returns) of 1.048%. Using the estimated model parameters, we can make one-step forward predictions of the volatility also shown in Figure 1. Both models are able to capture the volatility clustering feature in the empirical data. In fact, the trends of the predicted volatilities are quite similar between the models. Since GARCH(1,1) model involves more free parameters, it shows more variations and responsiveness than the EWMA model.

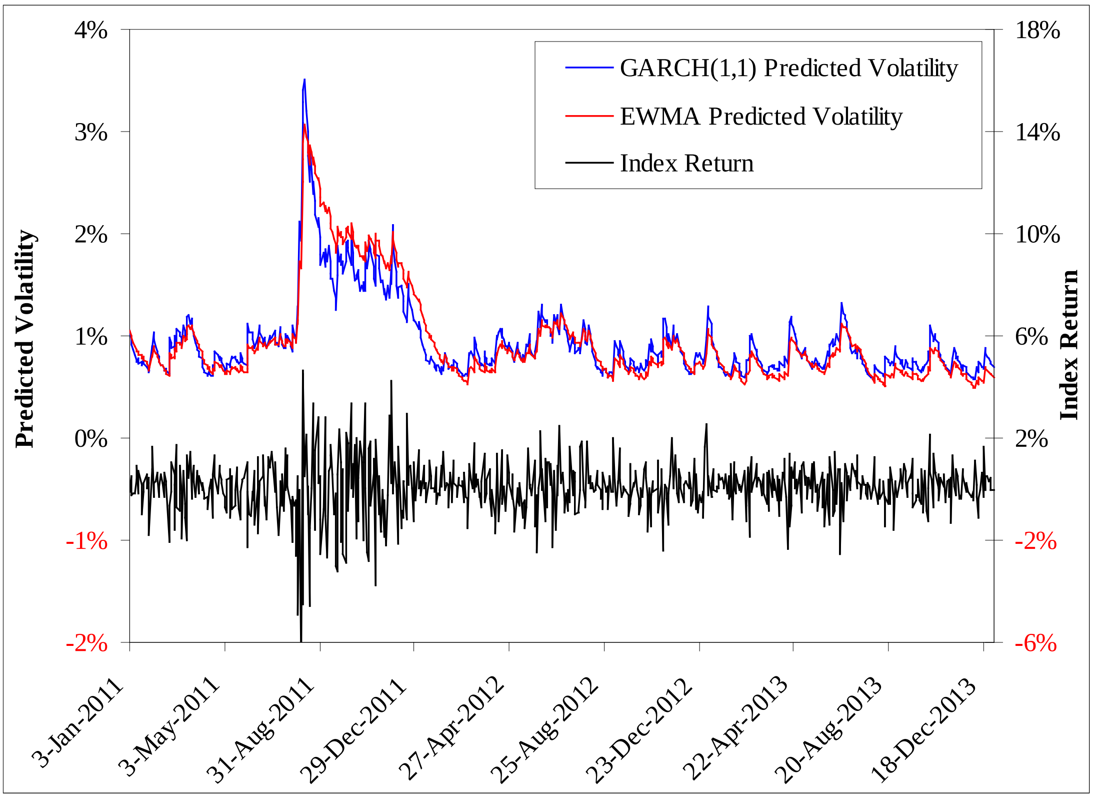


Figure 1. Volatility prediction: EWMA model vs. GARCH(1,1) model (calibrated to S&P500 equity index daily returns from 2011 to 2013)

At this stage it is important to note that there are no unique values for volatility and hence also for VaR. They are statistical estimates, which are dependent on the choice of models. Indeed, instead of estimating this number from past history, some analysts prefer to look at volatility implied from the option markets. This could be (arguably) forward looking as it gives an instantaneous poll of market's expectation via the market's price discovery mechanism.

Historical volatility and implied volatility

So far the models we have discussed in previous sections estimate volatility from realized past returns of a risk factor. We call it 'historical volatility' because it is a backward looking measure and statistically determined from history. It differs conceptually from another volatility measure, the so- called 'implied volatility'. The implied volatility is used in calculating the price of an option. It is the volatility of the underlying asset that when input into the option pricing model (e.g. Black-Scholes model) will return a theoretical value equals to the current market price of the option. In Black-Scholes model, the theoretical price of an option is a function of five parameters (for simplicity, dividend rate is ignored): spot price of underlying asset, strike price, expiry (time to maturity), volatility and cost-of-carry (e.g. Interest rates, storage cost, dividend rate, etc.). Given that the first four parameters are known for an option, the market price of the option gives a one-to-one translation to the volatility parameter. The value of the volatility backed out from traded price is the implied volatility. In fact, the market convention is to quote the implied volatility for trading. It is a forward looking and subjective volatility measure that reflects the market's expectation of the future dynamics of the underlying asset. Since options on an asset can be traded with different strike prices and expiries, the implied volatilities derived from the market prices form a volatility surface defined by the grid of strike price vs. expiry. For a given expiry, the implied volatilities at different strike prices usually show a 'smile'. This refers to a convex pattern of the implied volatility when plotted against the strike price. According to the Black-Scholes pricing model, the assumption of a normally distributed asset returns gives rise to a constant volatility across strikes. Hence, the phenomenon of volatility smile shows that the assumption is violated in practice. Interestingly, it can be shown mathematically that a distribution, which is fatter than normal (leptokurtic) and skewed, will give rise to a volatility smile which is slanted to one side. Indeed this is observed in almost all option markets. The volatility smile also contains other information such as the liquidity premium across different strikes; the premium typically increases for strikes that are further away from the current spot price.

In risk management, we are interested in the fluctuation of risk factors or scenarios. This is captured objectively using historical volatility and other quantile measures such as VaR. In contrast, implied volatility is actually a risk factor by itself when option products are in the portfolio. Since most banks trade in option products, the risk factor mapping will involve capturing the entire volatility surface. For example, if the FX option markets have 10 currencies, with 5 strikes and 10 maturity expiries, the bank will need to map 500 risk factors for implied volatility. The VaR will then involve measuring the fluctuation (or volatility) of these volatility risk factors to the extent that the bank has exposures to them.

VaR Specification

In the industry, most banks use their own in-house VaR systems to calculate VaR for risk management and reporting purposes. Different banks use different specifications for their VaR systems. The banks calculate a firm-wide VaR number, report it to the regulator, and periodically disclose it to investors. In order to make the comparison of VaR numbers meaningful and easy to understand, it is important to first specify the VaR system. A succinct way to do this is to use the format in Table 2.

Table 2. VaR system specification format

Item	*Possible Choices*
Product valuation	Full revaluation / delta approximation / delta-gamma approximation
VaR methodology	Parametric VaR / historical simulation / Monte-Carlo simulation
Observation window	250 days / 500 days / 750 days / 1000 days
Scenario weighting	Equally weighted / exponentially weighted
Confidence level	95% / 97.5% / 99%
Return calculation	Relative return / absolute return / log return

Return period	Daily / weekly / 10-day
Mean adjustment	Yes / no
Scaling (if any)	Scaled to 10-day / scaled to 99% confidence level

A VaR system specification basically defines how the VaR number is calculated. Some of the items have been mentioned previously. For example, we have discussed the advantages and disadvantages of full revaluation and approximation methods for product valuation. Considering the size and complexity of their trading portfolios, banks may choose either of the methods for their VaR systems as long as they are able to justify the appropriateness of its application. For example, if a portfolio consists of mainly linear products, the delta approximation is superior to the full revaluation in terms of computational speed while its accuracy is still acceptable. However, if we have a portfolio of options or other nonlinear products, the delta-gamma approximation or even the full revaluation method must be considered (in practice one might choose different revaluation methods for different products).

To choose a suitable length of observation window, we need to find a balance between the sensitivity of VaR to recent structural changes in the market and the accuracy of VaR estimation. A short window allows better sensitivity (or reaction) to the recent market moves, but it may not contain enough data to produce a reliable VaR estimate. On the other hand, a long window such as 1000-day means the VaR would hardly move even as the market enters a stressful situation. This would mean that the risk measure is late in registering risks at the onset of major crises. This dilemma can be addressed, for example, by using weighting schemes such as exponentially decaying weighted scenarios in the VaR model. This scheme weighs the scenarios further in the past by increasingly smaller weights, so that the resulting volatility figure is influenced more by recent fluctuations in prices that fluctuations far in the past. Thus, when the market enters a crisis mode, exhibiting large swings, the new information contributes to the volatility quicker, see Wong (2013) [3].

The calculation of returns can be different across risk factors. The choice of return calculation is determined by distributional type of the risk factor. As a general rule, if the size of the movements is independent of the value of the risk factor than often using absolute returns is appropriate, whilst if the size of the moves scales with the risk-factor then often using relative or log return is appropriate. Often a large portfolio may involve numerous risk factors. Hence a real world VaR system may contain a mixture of return definitions.

Basel requires banks to estimate their VaR for a time horizon of 10 days and at a confidence level of 99%. If a 10-day (non-overlapping) return period is used, we need a very long observation window to provide enough historical data for our VaR estimation. This not only reduces the predictability of VaR, but is also limited by data availability. Using over-lapping 10-day returns is not recommended as it introduces an auto-correlation bias in the VaR because every data point (daily return) is re-used ten times, see [4]. Basel allows banks to calculate the VaR using daily data and then scale the VaR to 10-day's by using the 'square root of time' rule. Strictly speaking this is valid only if the daily returns

$r_i, \cdots, r_n$, follow a Gaussian distribution and are independent from one another where the n-day return has a volatility of $\sqrt{n}$ times that of the daily returns.

Since quantile is proportional to volatility in a Gaussian model, this is equivalent to scaling a daily VaR by $\sqrt{n}$ to arrive at a n-day VaR. One has to be cautious that when the market is in a crisis and the distribution become fatter than normal, such scaling produces an understated VaR. There are advanced methods to deal with scaling for example using power law scaling, or by assuming returns following a Gaussian AR(n) autoregressive process. See Wong (2013) [3] for more details.

VAR methods

As mentioned, the VaR is no more than a loss quantile of the P&L distribution. There are many ways to generate the distribution and compute the quantile. Conventionally, the three most common methodologies used by banks are the parametric VaR, the Monte Carlo simulation VaR and the historical simulation VaR. A recent survey showed that most banks (about 73%) that report their VaR for regulatory purpose use the historical simulation VaR [5].

Table 3. Hypothetical Portfolio

Product Type	Equity	*Product Type*	Option	*Product Type*	Bond
Asset	SPX Index	*Asset*	SPX Index Call	*Asset*	5Y T-Bond
Risk Feature	Linear	*Risk Feature*	Nonlinear	*Risk Feature*	Nonlinear
Notional	$1,000,000	*Notional*	-$1,500,000	*Notional*	$500,000
Spot	1848.36	*Option Type*	Call	*Settlement Date*	12/31/2013
		Maturity (yrs)	1	*Maturity Date*	12/31/2018
		1Y Zero Rate	0.31%	*Coupon Rate*	2.00%
		Dividend Rate	0.00%	*Yield Rate*	1.74%
		Strike	1848.36	*Redemption Value*	100
		Volatility	15.23%	*Coupon Frequency*	2
		Spot	1848.36	*Day Count Basis*	Actual/360
Present Value	$1,000,000	*Present Value*	-$93,268	*Present Value*	$506,173
				Total Present Value of Portfolio	
					$1,412,905

In order to better explain the VaR estimation methods, we constructed a hypothetical portfolio as of date 31 Dec 2013 shown in Table 3. It consists of three component assets:

1. A long position on S&P500 equity index, i.e. denoted SPX Index, which is mapped to a single risk factor, 'SPX' spot;
2. A short position on equity index option, i.e. 1 year call options on SPX Index; its value given by the Black-Scholes option pricing formula, is mapped to

three risk factors: the 'SPX' spot, the '1-year at-the-money (ATM) volatility' and the '1-year zero rate' (or discount rate);

3. A long position on US Treasury bond, i.e. 5 year fixed coupon treasury bond, which is mapped to a single risk factor, '5-year yield rate'.

In the rest of the section, we will rely on this hypothetical portfolio to illustrate the three VaR methodologies.

As discussed earlier, although an asset's P&L distribution can be derived directly from price history of the asset, we need to perform risk factor mapping to reduce the dimensionality and simplify the calculation. For example, in the case of a Treasury bond, the market convention is to quote its price in the form of yield rate. Mapping the bond position to a risk factor, say, '5-year yield rate' is natural because bonds (of a specific issuer name and tenor) have a unique yield whereas their prices depend on the coupons.

This is especially helpful, if our portfolio involves a large number of bonds with different maturities. For example, a bond matured in 4.75 years should be mapped to 4.75-year yield rate. However, the yield curve is made of discrete benchmark points. The 4.75-year yield rate is usually not included in the yield curve. Instead the standard tenors at 4-year and 5-year points are the closest adjacent points to 4.75-year. One way of dealing with this issue is to interpolate the curve and retrieve the 4.75-year yield and plug it into the calculations.

Another way – which is in practice faster and without much of a loss in terms of accuracy – is to split the bond into standard maturities: so instead of $100 of a 4.75 year bond we assume that we have $75 of a 5-year bond, and $25 of a 4-year bond. The big advantage of this method is that ultimately we only have one bond per standard tenor which greatly reduces the amount of calculation required.

Parametric VaR

Parametric VaR (pVaR) or variance-covariance (VCV) VaR was popularized by JP Morgan's Riskmetrics. The original methodology was published in 1994 and quickly took hold as the industry standard. Strictly speaking, Riskmetrics proposed the modeling of VaR using the normal distribution and the EWMA volatility measure [5]. There are two fundamental assumptions here. Firstly, the valuation method of pVaR is sensitivity based and assumes a linear dependence on the risk factors. This is equivalent to a Taylor expansion whereby all the second and higher order terms are ignored. For nonlinear products, e.g. options, if the risk factors involve large moves, this assumption may introduce inaccuracies to pVaR. Secondly, the pVaR method assumes the returns of risk factors follow a multivariate Gaussian distribution. This assumption is often violated when markets are under stress — in a crisis state distribution of risk factors can be fat tailed and drastically skewed. Since VaR deals with tail losses of the P&L distribution, when these two assumptions are violated, it could introduce large errors in VaR estimation.

Table 4. pVaR specification

Item	*Choice*
Product valuation	delta approximation
VaR methodology	parametric VaR
Observation window	500 days
Scenario weighting	equally weighted
Confidence level	97.5%
Return calculation	log return
Return period	daily
Mean adjustment	no
Scaling (if any)	scaled to 10-day

Table 5. Level and volatility of risk factors

	SPX	*1Y Zero*	*5Y Yield*	*SPX Vol*
Current Level	1,848.4	31.4	174.1	15.2
Unit	point	basis point	basis point	% point
Volatility (%)	0.75%	2.26%	4.10%	2.00%

Table 6. Correlation matrix of risk factors

	SPX	*1Y Zero*	*5Y Yield*	*SPX Vol*
SPX	1.00	0.14	0.12	-0.80
1Y Zero	0.14	1.00	0.00	-0.13
5Y Yield	0.12	0.00	1.00	-0.12
SPX Vol	-0.80	-0.13	-0.12	1.00

Table 7. Sensitivities to risk factors

	SPX	*1Y Zero*	*5Y Yield*	*SPX Vol*
Equity	$541	-	-	-
Option	-$437	-$71	-	-$5,956
Bond	-	-	-$240	-
Portfolio	$104	-$71	-$240	-$5,956

At portfolio level, there are four relevant risk factors: 'SPX', 'SPX ATM volatility', '1Y zero rate' and '5Y yield rate'. The specification for the pVaR calculation is given in Table 4. We want to calculate VaR at a confidence level of 97.5% with a 10-day time horizon. A 500-day observation window is used ending at 31 Dec 2013, the date on which the VaR is calculated. Referring to the spreadsheet [VaR_Methods.xls], the general steps in the pVaR calculation are:

1. For simplicity, we take daily log return $r_{i,t}$ for all the four risk factors, where $i = 1,\cdots,4$ indexes the risk factors and $t = 1,\cdots,500$ is the historical scenario number (time sequence) in the observation window.
2. We calculate risk factor volatilities (i.e. sample standard deviation) $\varsigma = (\varsigma_1,\cdots,\varsigma_4)$ and correlation matrix ρ from the return data. The results are shown in
3. Table 5 and
4. Table 6 respectively.
5. First order sensitivities to the risk factors are computed using central difference method for each asset in the portfolio and then summed across assets to the portfolio level. For example, if we denote $\delta = (\delta_1,\cdots,\delta_4)$ the portfolio level sensitivities, its first entry (sensitivity to 'SPX') is calculated as $\delta_1 = \$541 - \$437 + \$0 = \$104$. This is shown in
6. Table 7.
7. The P&L volatility with respect to the i-th risk factor is calculated as $\sigma_i = \delta_i\varsigma_i f_i$, where f_i is the current level of the i-th risk factor (e.g. the level on the date the VaR is estimated, i.e. 31 Dec 2013). For example, the P&L volatility with respect to 'SPX' is $\sigma_1 = \$104 \times 0.75\% \times 1{,}848.4 = \$1{,}443$. We end up with a vector of P&L volatilities $\sigma = (\sigma_1,\cdots,\sigma_4)$ for the portfolio.
8. Because of diversification effect among risk factors, the P&L volatilities must be aggregated via the correlation matrix ρ to yield a total P&L volatility of portfolio σ_p

   $$\sigma_p^2 = \sigma\rho\sigma^T = \sum_{i=1}^{4}\sum_{j=1}^{4}\sigma_i\sigma_j\rho_{ij} = \$3{,}328 \tag{38}$$

   which quantifies the volatility of the joint P&L distribution over the next day.
9. Given the assumption of normality, the quantile (or VaR) is related to the volatility by a simple factor. For example, the 1-day VaR at confidence level $\alpha = 97.5\%$ can be estimated as

   $$\begin{aligned}\mathrm{VaR}_{1d} &= \Phi^{-1}(1-\alpha) \times \sigma_p \\ &= -1.96 \times \$3{,}328 \\ &= -\$6{,}522\end{aligned} \tag{39}$$

   where $\Phi^{-1}(\cdot)$ is the inverse cumulative density function of standard normal distribution. On Excel, it is given by the function NORMSINV($\cdot$).
10. The 'square root of time rule' applies to derive the 10-day VaR

    $$\mathrm{VaR}_{10d} = \mathrm{VaR}_{1d} \times \sqrt{10} = -\$20{,}624 \tag{40}$$

Note that in step (4), the value of risk factor level f_i is involved in the formula to calculate the P&L volatility. This is because the volatility ς_i in our example is estimated based on log returns of the risk factor. This yield based (in the sense of relative changes) volatility must be converted to norm based (in the sense of absolute changes) by multiplying ς_i with f_i to account for volatility of absolute changes in risk factor levels. Hence this scaling by f_i is not necessary if absolute returns have been used. In this case, the P&L volatility is simply given by $\sigma_i = \delta_i \varsigma_i$. Note that we only use first order (delta) approximation for P&L valuation. The accuracy can be improved if we include the second order corrections; the approach is then known as the 'delta-gamma approach'.

Monte Carlo VaR

The pVaR imposes a strict assumption on the distribution of the risk factor returns, and it cannot handle non-Gaussian distribution properly. Moreover, it is limited in what kind of instruments it can treat – in general terms every product where the second order approximation is not sufficient is not a good fit (and even second order can be a stretch if a product exhibits unusual correlations amongst risk factors). An alternative is the Monte Carlo VaR ("mcVaR") where in principle any risk distribution can be simulated, and any valuation method can be applied. In the following, we will use an example to illustrate the mcVaR method step by step.

Table 8. mcVaR specification

Item	*Choice*
Product valuation	Full revaluation
VaR methodology	Monte Carlo Simulation VaR
Observation window	500 days
Scenario weighting	Equally weighted
Confidence level	97.5%
Return calculation	Log return
Return period	Daily
Mean adjustment	No
Scaling (if any)	N/A

Table 8 shows the VaR specification for our example. Our mcVaR calculation uses the same test portfolio as per Table 3 and the same observation window mentioned. For simplicity, we will still assume a joint Gaussian distribution for the risk factors. However, in actual implementation, more realistic distributional assumptions can be applied, for example, we may sample from a Student *t* distribution to model fat tails. Referring to the spreadsheet [VaR_Methods.xls], let's trace the following general steps in our mcVaR calculation:

1. We take daily log return $r_{i,t}$ for all the four risk factors, where $i = 1,\cdots,4$ indexes the risk factors and $t = 1,\cdots,500$ denotes the numbering across historical scenarios in the observation window.
2. We calculate risk factor volatilities (i.e. sample standard deviation) $\varsigma = (\varsigma_1,\cdots,\varsigma_4)$ and correlation matrix ρ from the return data, using the 500 days return data. The same results as shown in
3. Table 5 and
4. Table 6.
5. Multivariate normal random numbers $\epsilon_k = (\epsilon_{k,1},\cdots,\epsilon_{k,4})$ are generated using *Cholesky* decomposition of the correlation matrix ρ (technically, *Cholesky* decomposition is an algorithm in linear algebra that decomposes a symmetric positive definite matrix, e.g. a correlation matrix, into a product of a lower triangular matrix and its transpose. The spreadsheet shows how this is done.). See also the Professional Risk Manager's Handbook, Volume II. In our example, we draw 2,000 simulated scenarios such that $k = 1,\cdots,2000$. Obviously, the larger the number of simulations, the more precise the mcVaR becomes (not accounting for parameter estimation errors). Typically banks may simulate more than 10,000 scenarios.
6. Use ϵ_k to simulate the k-th scenario for the risk factors. We need to specify a model for this stochastic process. In our example, we assume our risk factors follow a geometric Brownian motion (GBM) process. In the case of GBM model (with no drift), the simulated shifted (or shocked) risk factor level is given by

   $$\hat{f}_{i,k} = f_i \exp\left(-\frac{1}{2}\varsigma_i^2\Delta t + \varsigma_i\epsilon_{k,i}\sqrt{\Delta t}\right) \tag{41}$$

   where Δt is the time increment, $i = 1,\cdots,4$ is the index of risk factors and f_i is the current level of the i-th risk factor (shortly we will discuss this formula in more details).
7. The portfolio P&L of the k-th scenario is then just the sum of the present values of the assets priced (using full revaluation or pricing function) at the *shifted* levels $\hat{f}_{i,k}$ of relevant risk factors *minus* the sum of the present values of the assets priced at *current* levels f_i. Repeat this for all k to obtain a portfolio P&L vector with 2000 entries.
8. We take the 0.025-quantile of this P&L vector to give the 97.5% confidence level daily VaR.

In our illustration, we have assumed all risk factors follow a GBM stochastic process which is a very common simplification practiced in the industry.

$$df_t/f_t = \mu dt + \sigma dW \tag{42}$$

where μ is a deterministic drift (or long-term rate of return) and σ is the volatility. The dW is known as a Wiener process, and can be written as $dW = Z\sqrt{dt}$, where Z is a random draw from a standard normal distribution. In other words, the dW

is normally distributed with mean of zero and variance of dt. Substituting for dW in (42), we get

$$df_t/f_t = \mu dt + \sigma Z\sqrt{dt} \tag{43}$$

The instantaneous rate of change in the risk factor, df_t/f_t, evolves according to its drift term μdt and the random term $\sigma Z\sqrt{dt}$. In practice, we would implement the model (or code it) in its discrete form. To reduce the numerical instability, (43) is often discretized in logarithmic form, i.e. given Δt is a small time increment, we write

$$\Delta \ln f_t = \mu\Delta t - \frac{\sigma^2}{2}\Delta t + \sigma Z\sqrt{\Delta t} \tag{44}$$

where $\Delta \ln f_t = \ln f_{t+\Delta t} - \ln f_t$ is the change in the logarithm of the risk factor over the time interval Δt (note: The term $-\frac{1}{2}\sigma^2\Delta t$ comes from the Ito lemma to offset the bias introduced by the logarithm transformation). The shifted level of the risk factor then becomes

$$f_{t+\Delta t} = f_t \exp\left(\mu\Delta t - \frac{1}{2}\sigma^2\Delta t + \sigma Z\sqrt{\Delta t}\right) \tag{45}$$

Note that (44) assumes that the log return of the risk factor is normally distributed. Hence our criticisms of parametric VaR with respect to the normality assumption will also apply to the mcVaR methodology, unless we employ a more realistic dynamics than the GBM model with constant volatility, e.g. stochastic volatility models.

At times we may need to simulate a risk factor f over a period of time T up to the time horizon of the VaR. We would usually divide T into a number N of small time increments Δt (i.e. we set $\Delta t = T/N$). We take the current level of f as a starting value and draw a random sample to update f using (44); this gives the change in the f over the first time increment, and we repeat the process again and again to evolve the changes of the f over all N increments up to the risk horizon T. This is one simulated scenario. We then repeat the exercise many times to produce as many simulated scenarios as needed. Since we have assumed constant μ and σ in our GBM model, dividing T into N small time increments is not necessary, the risk factor can be evolved in just *one* step with $\Delta t = T$. For a 10-day VaR, we can just let $\Delta t = 10$ in our Monte Carlo simulation given that our volatilities ς_i are estimated from daily returns (i.e. not annualized). Note that multi-step simulation is needed in some cases where the pricing of the product is dependent on the nature of the paths taken by risk factors.

Monte Carlo simulation VaR has many advantages over parametric VaR, such as:

1. With the help of suitable distributional assumptions, it can capture a wider range of market behavior.
2. Since the P&L is computed by full revaluation (and not sensitivity-based approximation) it can effectively handle nonlinear products, including exotic options and structured financial instruments.

3. With sufficient number of simulations, it can provide detailed insight into extreme losses that lie far out in the tails of the distributions, beyond the usual VaR cutoff level.

However, there are also considerable drawbacks to the mcVaR. Monte Carlo simulation is computer intensive not just because mcVaR uses full revaluation in P&L calculations, but also because of the large number of simulated scenarios needed to estimate the mcVaR precisely. In general, if we want to double the precision, we end up quadrupling the number of simulated scenarios (and this assumes Gaussian distributions; on more tailed distributions the scaling behavior is even worse). This square root convergence is slow and greatly limits the power of Monte Carlo simulation.

The mcVaR is also highly dependent on how the return distribution is modeled. The scenarios generated by Monte Carlo simulation must be consistent with the historical characteristics of the market. While we know the Gaussian distribution is too idealistic to hold, the real joint distribution of a dynamic market remains unknown and is difficult to model. In addition, the stochastic process that drives the risk factor needs to be modeled. Although a wide variety of models have been developed in the academia to describe the observed market dynamics of different asset classes, none of them is perfect; each model has its advantages and disadvantages. Moreover, with the exception of GBM, most models require calibration to determine their parameters. Unfortunately, such calibration is often limited by the data availability, and the parameters themselves are seldom constant, so re-calibration is periodically required. This means that in practice the banking industry often use simplistic Gaussian simulations because it is too cumbersome to model each riskfactor class separately. It is important that the student understands the limitation of simple models used in a realistic setting; and this often lead to understatement of risks.

Historical Simulation VaR

Historical simulation VaR (hsVaR) has gained popularity in recent years. Most banks that disclose their VaR method have reported using historical simulation [5]. The hsVaR is different from previous two methods in that it does not assume any distribution on returns. It can be regarded as a Monte Carlo simulation method that uses *historical* samples for its scenarios rather than samples drawn from *theoretical* distributions. It overcomes many drawbacks that plague pVaR and mcVaR. Firstly, by sampling from empirical or historical data, it avoids the need to make any distribution assumption. This takes care of fat-tails and skewness because we let the data themselves decide the shape of the distribution. Note that there will be one distribution for every risk factor. Secondly, hsVaR usually computes the P&L of each product using full revaluation. Given the complexity of today’s derivatives, most portfolios are non-linear. Full revaluation avoids the error arising from delta (or delta-gamma) approximation. Thirdly, risk aggregation is done by summation of P&L across products to form the portfolio P&L. The dependence structure among risk factors is accounted for in this way, i.e. you get the correlation for free. And there is no need to maintain a large correlation matrix as required in pVaR and mcVaR. This simple “cross summation” is easy to understand intuitively and allows for easy risk decomposition for the purpose of analysis.

Table 9. hsVaR specification

Item	*Choice*
Product valuation	full revaluation
VaR methodology	Historical Simulation VaR
Observation window	500 days
Scenario weighting	equally weighted
Confidence level	97.5%
Return calculation	log return
Return period	daily
Mean adjustment	no
Scaling (if any)	N.A.

Table 9 shows an hsVaR specification for our example. Our hsVaR calculation uses the same test portfolio as per Table 3 and the same observation window as before. Referring to the spreadsheet [VaR Methods.xls], let's summarize the steps in computing hsVaR:

1. A scenario vector is formed from the observation window using the historical daily log returns $r_{i,t}$ where $i = 1, \cdots, 4$ indexes the risk factors and $t = 1, \cdots, 500$ numbers the historical scenarios along the observation period.
2. The shifted levels of the risk factors are calculated for the t-th scenario, e.g. we have $\hat{f}_{i,t} = f_i \exp(r_{i,t})$ for log returns (or $\hat{f}_{i,t} = f_i + r_{i,t}$ for absolute returns and $\hat{f}_{i,t} = f_i(1 + r_{i,t})$ for relative returns), where f_i is the current level of the i-th risk factor.
3. For an asset, the P&L of the t-th scenario is just the present value priced (using full revaluation) at the *shifted* levels $\hat{f}_{i,t}$ of relevant risk factors *minus* the present value priced at the *current* levels f_i. Do this for all 500 scenarios to derive a P&L vector for that asset (a vector of 500 P&L's). Repeat this for all assets in the portfolio.
4. Sum the P&L vectors for all assets *by scenario* to derive the P&L vector for the entire portfolio, which is in fact a distribution with 500 points. We take the 0.025-quantile of this P&L vector to produce the 1-day VaR at a confidence level of 97.5%.

In essence we are really trying to project the P&L one step (1 day) into the future from today, but using random draws (or scenarios) from the past sample period. Thus, we always apply the historical returns to current risk factor levels and current portfolio positions. In step (3), hsVaR may also use approximation methods (sensitivity approach) to price asset for the purpose of reducing computational complexity. In step (4), the cross summation of asset P&L (by scenario) allows for a very flexible risk decomposition and diagnosis. For example, if we are interested in knowing the risk due to equity spot alone, we can shift only the equity spot risk factor and repeat the above steps. If we are interested in VaR breakdown by sub-portfolio, we can do the P&L vector summation for each sub-portfolio separately. If we are interested in the impact of a single asset, we can exclude the P&L vector of that asset from the summation, and look at the difference; this is known as the incremental VaR.

Although hsVaR has been widely used in industry, it is still subject to a few subtle weaknesses:

1. The returns are assumed independent for non-overlapping time periods, and the historical scenarios are assumed to provide a good guidance for tomorrow's risk. In reality, market risk parameters are not independent of time. They dynamically change, often suddenly, such as at the onset of a crisis. Hence, hsVaR lags major market shocks. Note this weakness (lateness) is generally true for any VaR methods which rely on past history.
2. The simulation uses a limited number of samples from historical scenarios; typically, 1-2 years of data. Hence there can be a large error in our estimated VaR. In fact, the higher the confidence level we set, the larger the error in the estimated VaR. As in the mcVaR, the precision can be improved by increasing the number of scenarios, but this is somewhat restricted for hsVaR because extending the window length reduces the sensitivity to detect regime changes in the market.
3. The aggregation of P&L vectors "by scenario" is affected by data quality. The price series must not be unusually quiet (or stale) and must not be too choppy (have spikes). Otherwise, the portfolio VaR becomes unstable. This is because the tail of an empirical distribution is not as smooth as a theoretical one used in pVaR and mcVaR. This problem is more pronounced for short observation period, high confidence level or small portfolio size. Thus, the sampling error (precision) of hsVaR is generally poorer.
4. Whilst this method looks objective, it is not: one still has to decide how a price change say from \$200 to \$220 compares to one say at \$400 (eg, to \$440, or \$420), and those choices have none trivial implications
5. A major drawback is that new asset classes often get very unrealistic risk figures; for example, in the run up to the 2008 crisis, ABS spreads had been constantly tightening, so much that even at a very high quantile the worst loss was a gain.

Simulation of Interest Rates

Interest rates markets are much more complex than markets for other asset classes such as equities, currencies or (most) commodities. Even for a single currency, there are many interest rate markets of different credit worthiness, for example the Treasury bond rates, inter-bank offered rates (Libor), corporate bond rates, mortgage rates and so on. Although these rates are affected by common macroeconomic factors, they do not move in perfect sync due to structural and credit differences. Additionally, we cannot use a single rate to describe the overall interest rate level of the market — interest rate is generally a function of term (or tenor). This function is often called 'interest rate term structure' or 'yield curve'. The yield curve is positively sloping during normal times, and its shape changes dynamically over time. The co-movements and dynamics of a yield curve complicate the risk measurement of a portfolio. Suppose a bank holds a bond portfolio, which contains a collection of various bonds with different maturities. Each bond has a yield rate associated with its maturity. The portfolio indeed has risk exposures along all the yield rates at various maturities. When we perform risk analysis, it become important and natural to look at the shapes and movements of the whole yield curve.

Term Structure of Interest Rates

Figure 1 shows a typical yield curve of US Treasury bonds observed on 31 Dec 2013. The yield curve is usually represented as a discrete set of benchmark rates of different standard terms (the points in the chart); these are often used for risk factor mapping. The shape of the yield curve indicates the current state of supply and demand and the cost of borrowing of that debt market. The longer dated the loan (or bond) is, the larger the credit and inflation risks; lenders thus offer long-term loans at higher interest rates than they offer for short-term loans. This acts as a premium to compensate for the default and inflation risk the lender is exposed to; hence the yield curve typically upward sloping. Occasionally, a yield curve can become inverted as long-term yields fall below short-term yields. This indicates the market is anticipating lower interest rate in the future, and borrowers are now seeking short term loans more aggressively than long-term loans. This is often seen as a forward indication of economic downturn.

A borrower's creditworthiness (represented by its credit rating) is another significant driver affecting the level of its yield curve. Governments often issue bonds in local currency. A series of such bonds make up the government bond yield curve or 'sovereign curve' or 'govies'. Governments are deemed to be risk-free and enjoy the highest credit rating in their own currencies since the central bank can always print local currencies to pay off maturing government debt. Thus 'govies' naturally have the lowest interest rates in the market. (Note that this working assumption does breakdown in exceptional cases such as the Russian bond default in 1998 and the Eurozone crisis in 2009). Banks with a high credit rating (e.g. AA) borrow money from each other at LIBOR rates (or the interbank rate). LIBOR curves are typically a little higher than government curves to account for the lower creditworthiness of banks. Corporate curves are another category of yield curves. They are constructed from the yields of bonds issued by corporations. Since corporations typically have higher perceived credit risk than governments and banks, they have to offer higher yields on their bonds to attract investors.

Yield curves shifts and evolves from day to day reflecting the market reaction to news and perception of supply-demand. The yield curve exhibit various degrees of freedom in its motion — parallel shifts, changes in slope (e.g. flattening or steepening) and changes in curvature (e.g. bowing). In addition, short term rates are more volatile than long term rates because short rates are affected by central bank monetary policy actions/expectations and central bank actively manages this. The long end is driven mainly by structural factors such as inflation expectation, supply-demand of bonds. The super-long end (more than 10 years) is also dominated by long term ('buy and hold') institutional investors such as pension funds and insurance companies.

When the shape of the yield curve changes, it affects the present value of the portfolio, and this gives rise to P&L for positions. A measure of how sensitive the position P&L is to unit change in yield curve movement is given by its PV01 (See Table 1), but this only considers parallel movement in the yield curve. To study the full dynamics of the yield curve, the risk manager typically resorts to a common technique called Principal Component Analysis (PCA).

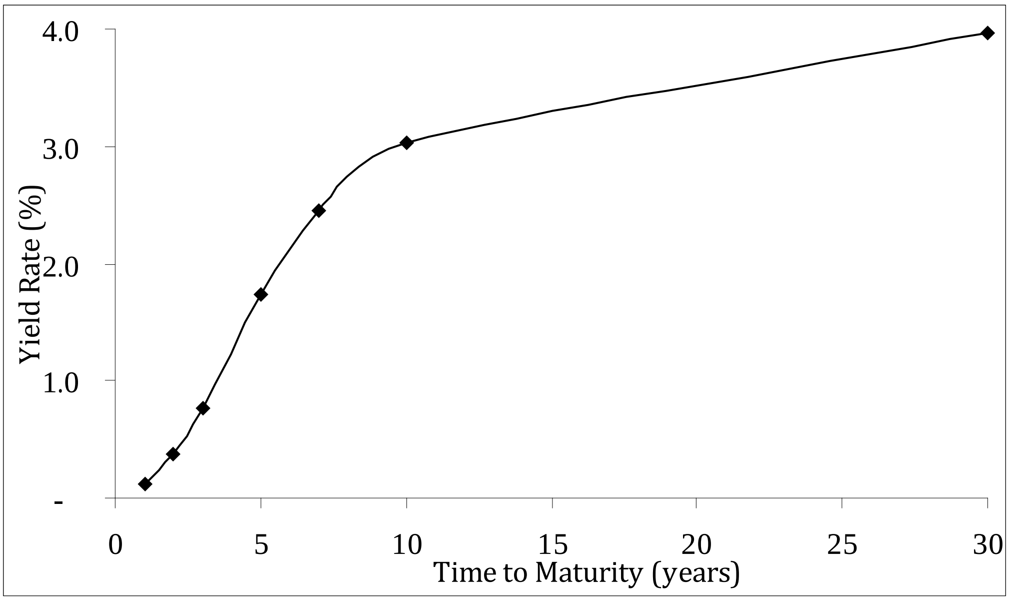


Figure 2. Yield curve of US Treasury bonds as of date 2013-12-31

Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical procedure that performs an orthogonal linear transformation that changes a Gaussian vector with a non-trivial covariance matrix into a standard Gaussian vector where variables are uncorrelated. Those new variables are known as Principal Components, and they are sorted by descending variance. What people do in practice to reduce the dimensionality of a problem is that they concentrate on say the first three risk factors that say capture 80% of the overall variance, effectively reducing a high dimensional problem to a three dimensional one. For example, when modeling yield curve the first factor tends to be a more or less parallel shift, the second factor tends to be a change in the slope, and the third one a change in the convexity of the curve. This describes the majority of the moves that are being seen in practice, and so the 20-tenor / 20-dimensional problem can be reduced to a 3-dimensional problem where all 20 tenors are reconstructed from those 3 factors.

There are two issues with this approach however: Firstly: 80% of the variance captured – which is a typical value – is not as impressive as this sounds because the standard deviations – which are much more relevant figures in this context – are the square roots of the variance. So in this example, if we explain

80% of the variance (and hence don't explain 20% of it) then this means for example that we have two random variables, one with volatility 100 and the other (independent) one with volatility 50, and we ignore the second one. The ignored risk is bigger than it sounds when we say "we explain 80% of the risk". The second issue is that ultimately it does not matter what percentage of the variance of our risk factors we explain, but we are interested in the portfolio variance. Let's stay in the interest rate space and let's assume we are using a two factor model, meaning we only explain level and slope. Any convexity product – eg long 1y and 5y, double short 3y – will show up as zero risk in this model which is evidently not right.

The PCA is best suited for highly correlated systems and systems that have common economic drivers. Typical examples include the stock market, interest rate yield curves, futures prices across tenors, and option volatility surfaces. The way PCA works in practice is the we start from a covariance matrix $\mathit{\Sigma}$. Since the $\mathit{\Sigma}$ is symmetric and positive (semi) definite, eigen-decomposition can be used to decompose the $\mathit{\Sigma}$ into

$$\mathit{\Sigma} = E\mathit{\Lambda}E^T \tag{46}$$

where $\mathit{\Lambda}$ is an diagonal matrix with non-negative entries being the eigenvalues in descending order and E is an orthogonal matrix with columns being the eigenvectors of a unit length. A non-zero (column) vector e is an eigenvector of a square matrix $\mathit{\Sigma}$ if and only if it satisfies the linear equation

$$\mathit{\Sigma}e = \lambda e \tag{47}$$

where λ is a non-zero scalar called the eigenvalue corresponding to the eigenvector e. In other words, the linear transformation by matrix $\mathit{\Sigma}$ merely elongates or shrinks the eigenvectors and the amount that they elongate or shrink by is the eigenvalue.

The orthonormal matrix E resulted from the eigen-decomposition can be used to transform the n risk factors into n latent variables, called the principal components, such that

$$p = rE \tag{48}$$

where $p = (p_1, \cdots, p_n)$ is a n-dimensional row vector. Since E is orthonormal, thinking geometrically, the transformation is just a simple rotation of coordinate system. The principal components $p_1, \cdots, p_n$ are uncorrelated and decreasingly responsible for the overall variation in the risk factors. For example, $p_1 = \sum_{i=1}^{n} r_i E_{i,1}$ is uncorrelated to $p_2 = \sum_{i=1}^{n} r_i E_{i,2}$ (i.e. the covariance between p_1 and p_2 is zero). The variance of p_1 is given by the first and the largest diagonal entry of matrix $\mathit{\Lambda}$, that is $\sigma_{p_1}^2 = \mathit{\Lambda}_{1,1}$, while the variance of p_2 is given by the second diagonal entry $\mathit{\Lambda}_{2,2}$, and so on. Figure 3 shows the two principal components derived from a 2-D joint normal distribution. Seen geometrically, the PCA transformation is a rotation of coordinate system from the chart's axes to an orthogonal system represented by the slanted "L". Referring to the spreadsheet [VaR_PCA.xls], the joint normal distribution is formed by two marginal distributions with volatility of 2.0 and 1.0 respectively, correlated by a coefficient of 0.8. Eigen-decomposition shows that the first principal has a variance of 4.69, and the second has a variance of only 0.31. In contrast, at correlation equal to zero, the two margins are already orthogonal; the principal components are the margins themselves (as shown in Figure 4).

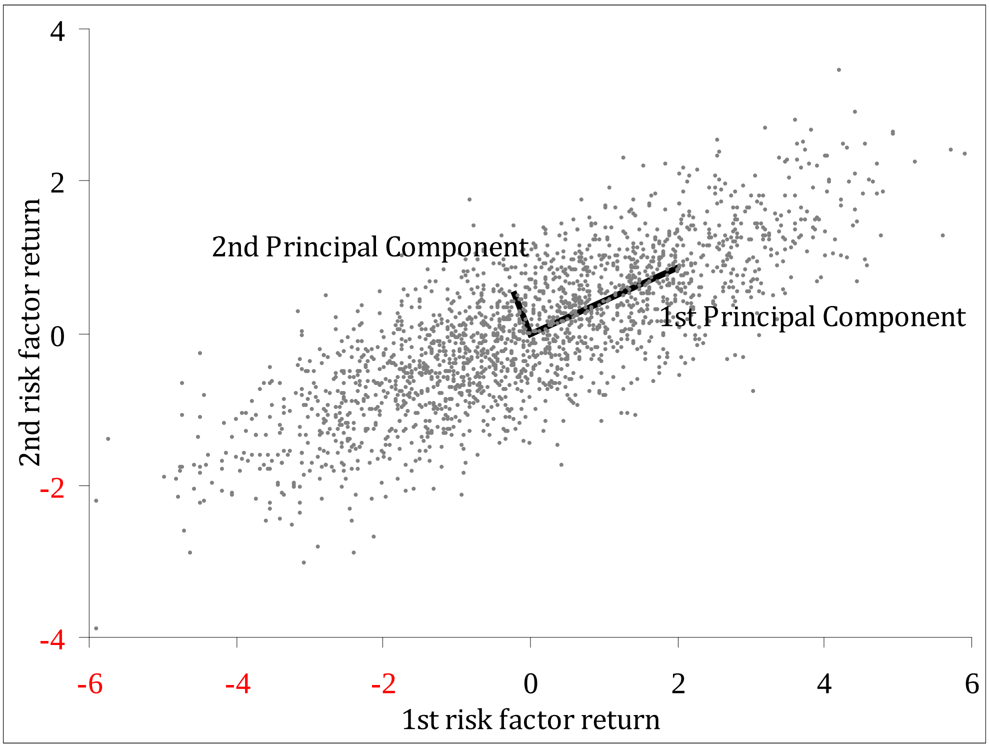


Figure 3. Principal components from a correlated 2-D joint normal distribution. The standard deviations of the cloud along the x-axis and y-axis give the volatilities of the two marginal distributions (2.0 and 1.0 respectively).

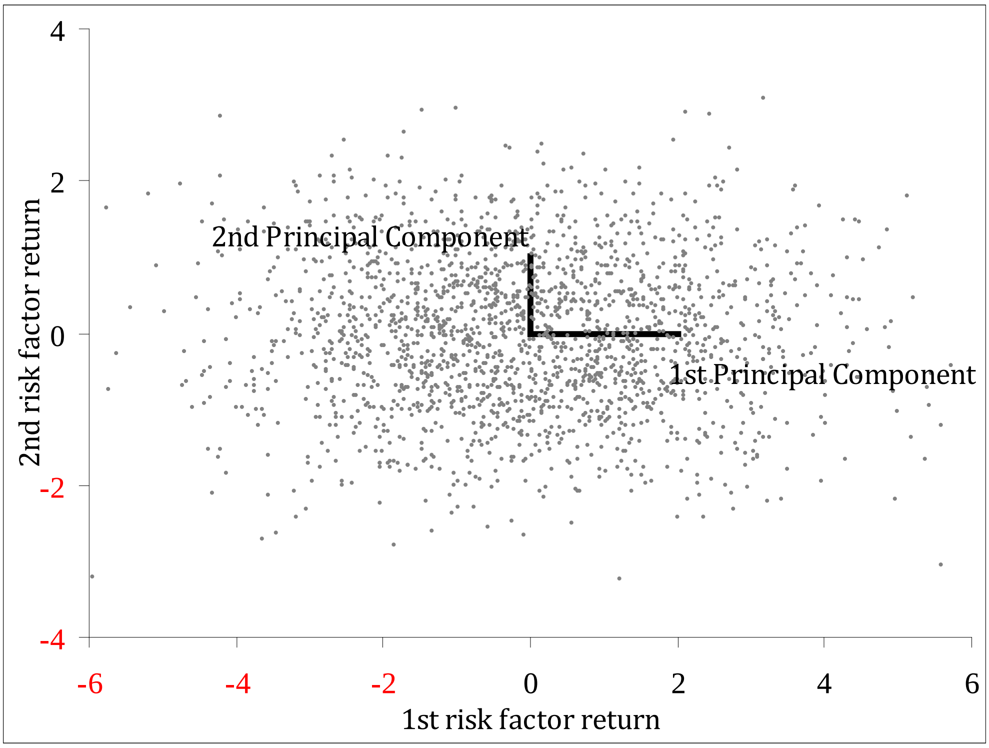


Figure 4. Principal components from an uncorrelated 2-D joint normal distribution

Reconstruction of the risk factors r from principal components p is straightforward. Since E is orthonormal, the inverse of E is its transpose. The principal components can then be transformed back to original risk factors by

$$r = pE^T \tag{49}$$

To reduce the dimensionality, we may form a vector $\hat{p}$ that includes only a few (but not all) of the most significant principal components, for example $\hat{p} = (p_1, \cdots, p_k, 0, \cdots, 0)$ includes only the first k principal components. The $\hat{p}$ is then transformed by matrix E^T to give an approximation $\hat{r} = \hat{p}E^T$ to the original risk factors r. Since the ignored principal components, $p_{k+1}, \cdots, p_n$, are insignificant, the $\hat{r}$ can still capture most of the variation in $\hat{r}$. In other words, we reduce the dimension of the problem from n to k without sacrificing much accuracy. This is why entire yield curve can be represented by just three components. In addition, PCA is also useful for scenario analysis and stress testing. Through reconstruction, we are able to generate scenarios of r consistent with its historical observations by introducing shocks to the main principal components. For example, we can separately quantify how large a parallel shift, a steepening and a bowing can be in history, and then shock the three principal components by these magnitudes.

PCA performs effectively when risk factors are strongly correlated. If the relationship is weak among risk factors, PCA does not work well to reduce dimensionality. In general, if most of the correlation coefficients are smaller than 0.3, PCA will not help. In the rest of the section, we will illustrate the performance of PCA under different conditions using an example. Referring to the spreadsheet [VaR_PCA.xls], we assume there are three risk factors forming a row vector $r =$

(r_1, r_2, r_3), each follows a univariate normal distribution with mean of zero and volatility of 4.0, 3.0 and 2.0 respectively. To illustrate the performance of PCA for a strongly correlated system, we simulate a joint distribution for the risk factors using a correlation matrix shown in Table 10. The correlation matrix shows r_1 and r_3 are both strongly correlated with r_2, whereas the correlation between them is weak. The correlation matrix along with the risk factor volatilities defines a covariance matrix (shown in Table 10), which PCA takes as input to perform eigen-decomposition. The resulted eigenvectors and eigenvalues of the covariance matrix are shown in Table 11. The first principal component given by the first eigenvector $(0.78,\ 0.58,\ 0.23)^T$ has a variance of 23.86 which is markedly greater than the second principal component and overwhelmingly larger than the third whose variance is only 0.07. Since the first two principal components capture most of the variation in the correlated system, we may just rely on them to reconstruct an approximation to the system. The omission of the third principal component is trivial as the covariance and correlation matrix of the reconstructed risk factors, shown in Table 12, only slightly differ from the original ones (in Table 10).

Table 10. Correlation and covariance matrix of risk factors (strongly correlated system)

Correlation Matrix			*Covariance Matrix*		
1.00	0.80	0.30	16.00	9.60	2.40
0.80	1.00	0.80	9.60	9.00	4.80
0.30	0.80	1.00	2.40	4.80	4.00

Table 11. Eigenvalues and eigenvectors of covariance matrix (strongly correlated system)

Eigenvalues, Λ			*Eigenvectors, E*		
23.86	-	-	0.78	-0.54	0.32
-	5.08	-	0.58	0.43	-0.69
-	-	0.07	0.23	0.72	0.65

Table 12. Covariance and correlation matrix of reconstructed risk factors using the first two principal components (strongly correlated system)

Correlation Matrix			*Covariance Matrix*		
1.00	0.80	0.30	15.99	9.61	2.39
0.80	1.00	0.81	9.61	8.97	4.83
0.30	0.81	1.00	2.39	4.83	3.97

In a weakly correlated system, PCA is generally less effective. Let's consider a correlation matrix, shown in Table 13, where risk factors are weakly correlated with each other. The eigen-decompostion of the covariance matrix shows all principal components (shown in Table 14) are significant, which means omission of even the least significant principal component in reconstruction may result in large errors. This can be seen from Table 15, where both the covariance and correlation matrix of the reconstructed risk factors markedly differ from the original ones (in Table 13). The impact is especially large for the third risk factor whose variance has been reduced from 4.00 to 0.45!

Table 13. Correlation and covariance matrix of risk factors (weakly correlated system)

Correlation Matrix			*Covariance Matrix*		
1.00	0.20	0.10	16.00	2.40	0.80
0.20	1.00	0.20	2.40	9.00	1.20
0.10	0.20	1.00	0.80	1.20	4.00

Table 14. Eigenvalues and eigenvectors of covariance matrix (weakly correlated system)

Eigenvalues, Λ			*Eigenvectors, E*		
16.84	-	-	0.95	-0.32	-0.02
-	8.44	-	0.30	0.93	-0.21
-	-	3.72	0.09	0.19	0.98

Table 15. Covariance and correlation matrix of reconstructed risk factors using the first two principal components (weakly correlated system)

Correlation Matrix			*Covariance Matrix*		
1.00	0.20	0.33	16.00	2.38	0.88
0.20	1.00	0.99	2.38	8.83	1.97
0.33	0.99	1.00	0.88	1.97	0.45

PCA in Interest Rate Simulations

To demonstrate the application of PCA for interest rate simulations, we look at the US Treasury bond curve with tenors 1Y, 2Y, 3Y, 5Y, 7Y, 10Y and 30Y. Daily returns for the rates are calculated using four years of historical data from 2010 to 2013. The covariance matrix derived from the return data is then subjected to the PCA analysis. The result is shown in Table 16. There are 7 principal components derived from the covariance matrix. They are ranked from left to right in terms of the amount of variance they explain. The first principal component is denoted as ‘PC1’, the second as ‘PC2’ and so on. Denoting the yield returns as $r_{1Y}, r_{2Y}, \cdots, r_{10Y}$, then the first principal component, for example, is formed by a linear combination of the returns with factor loadings shown in the column associated with PC1

$$PC_1 = 0.37 \times r_{1Y} + 0.53 \times r_{2Y} + 0.53 \times r_{3Y} + \cdots + 0.13 \times r_{10Y} \tag{50}$$

The rest of the principal components can be constructed in the same manner. The above formula says if we introduce a one basis point change in PC1, it corresponds to 0.37 basis points change in r_{1Y} and 0.53 basis points change in r_{2Y} and so on.

Table 16. Summary of PCA on US Treasury Bond Yield Curve

Summary of Principal Components

		PC1	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*
Variances, λ (bps)		92.37	21.97	6.95	1.83	0.64	0.23	0.07
Factor Loadings	*1Y*	0.37	-0.90	0.21	-0.02	-0.01	0.00	0.00
	2Y	0.53	0.05	-0.66	0.52	-0.08	-0.00	0.00
	3Y	0.53	0.19	-0.17	-0.72	0.37	-0.07	-0.01
	5Y	0.39	0.25	0.33	-0.10	-0.63	0.52	0.03
	7Y	0.29	0.21	0.41	0.18	-0.14	-0.73	0.33
	10Y	0.21	0.16	0.36	0.28	0.29	-0.02	-0.80
	30Y	0.13	0.11	0.29	0.31	0.60	0.43	0.50
							Total Variance (in basis points):	124.06

Percentage of Variance Explained

PC1	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*
74.45%	17.71%	5.60%	1.48%	0.51%	0.19%	0.06%

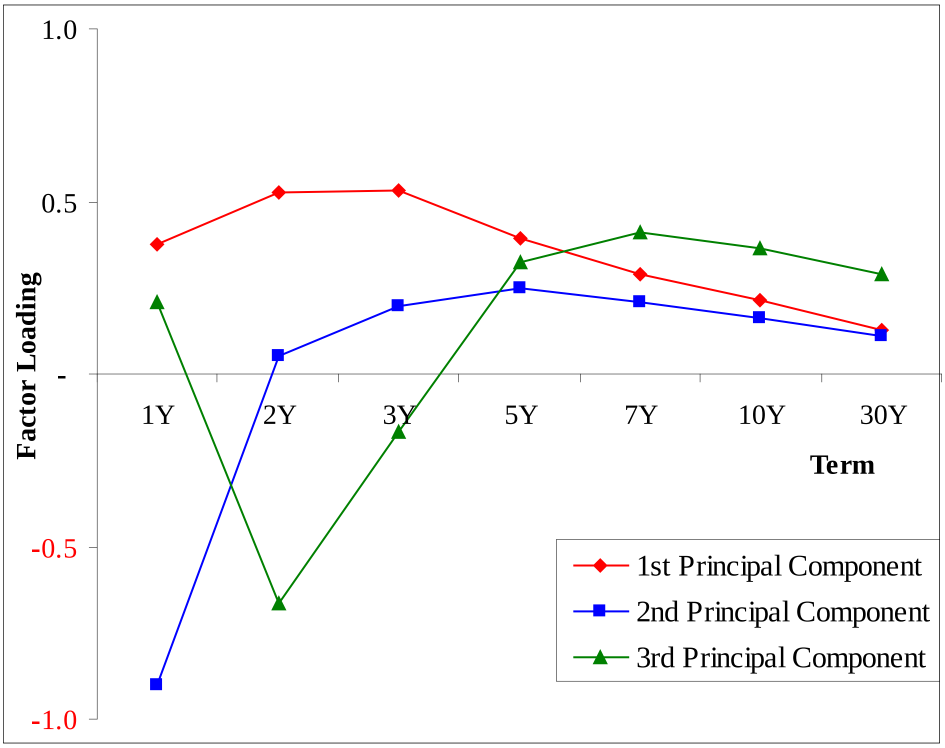


Figure 5. Factor loadings of the top 3 principal components

Figure 5 shows the factor loadings of the top 3 principal components. The PC1 has positive and relatively flat shape of factor loadings at all terms. This can be interpreted as a parallel shift of the yield curve (note that all coefficients of the first eigenvector always have the same sign). Table 16 shows that PC1 alone can explain 74.45% of the total variance observed in the yield curve (i.e. variance 92.37 of PC1 out of total variance 124.06 of the whole yield curve). It implies the yield curve does tend to move in parallel (as defined by the eigenvector, ie with moves in the short term roughly twice the size of the moves in the long term) most of the time. The factor loadings of PC2 form an upward sloping shape. They are negative at short terms and positive at longer terms. In other words, PC2 is negatively correlated with short term rates and positively correlated with long term rates, which corresponds to flattening or steepening movements of the yield curve. The factor loadings of PC3 are positive at both short and long terms but negative at mid-terms. This corresponds to a change of curvature in the yield curve.

Note that the top three principal components combined can explain 97.76% (=74.45% + 17.71% + 5.60%) of total variance. Hence the dynamics of the yield curve can be well approximated using just three principal components. That is, for VaR estimation, we map the portfolio positions to the three principal components as risk factors rather than the seven yield rates, which effectively reduces the dimensionality of the risk factors without distorting the distribution and its risk features.

In pVaR, the remapping of risk factors is achieved by transforming the risk sensitivities with respect to the seven yield rates to the three principal components. This can be done using a similar formula shown in (50). For example, if we have the first order sensitivities (i.e. PV01) of a portfolio $\delta_{1Y}, \delta_{2Y}, \cdots, \delta_{10Y}$ corresponding to the yield rates, we can transform them into the sensitivities with respect to the principal components using the factor loadings. In the case of PC1, the transformation reads

$$\delta_{PC1} = 0.37 \times \delta_{1Y} \times y_{1Y} + 0.53 \times \delta_{2Y} \times y_{2Y} + \cdots + 0.13 \times \delta_{10Y} \times y_{10Y} \tag{51}$$

where δ_{PC1} is the PV01 with respect to PC1 and $y_{1Y}, y_{2Y}, \cdots, y_{10Y}$ are the current levels of the rates (e.g. at the time the VaR is estimated). Here we again scale the PV01 by the corresponding rate level because the principal components are derived using log returns; its sensitivity must be scaled to a yield based value. If absolute returns were taken, the scaling could be omitted. Since the principal components are uncorrelated with each other, the risk aggregation is just a simple summation. The portfolio P&L volatility is

$$\sigma_p^2 = \delta_{PC1}^2 \times \lambda_{PC1} + \delta_{PC2}^2 \times \lambda_{PC2} + \cdots + \delta_{PC3}^2 \times \lambda_{PC3} \tag{52}$$

where λ_{PC1} is the variance of PC1, λ_{PC2} is the variance of PC2, and so on. We know the top three principal components dominate the variance, so we can just calculate $\delta_{PC1}, \delta_{PC2}, \delta_{PC3}$ and ignore higher components terms in (52) to estimate an approximate portfolio P&L volatility $\hat{\sigma}_p$. The $\hat{\sigma}_p$ can then be used to estimate the pVaR using formula (39).

In hsVaR, we want to use historical scenarios of the principal components to simulate the yield curve evolution. This basically involves three steps: 1) construct return series for principal components; 2) use the return series of principal components to reconstruct return series for yield rates; 3) use the reconstructed yield rate returns to evolve the yield curve. The first step can be done using formula (50). For every historical scenario, we use the factor loadings to transform returns of yield rates into returns of principal components. We retain only the top three principal components and exclude the rest. We reconstruct the (approximated) rate returns from them using formula (49) for every scenario. The yield curve is then evolved using the reconstructed rate returns. For a single scenario date, the evolved yield curve can then be used to reprice the portfolio to give the full revaluation P&L. Repeating the repricing for all the historical scenarios, we will obtain a P&L distribution. The hsVaR is just the quantile of this distribution. The same strategy applies to mcVaR. In this case, normal distributions are assumed for the principal components. Hence we sample the principal components from independent normals with respective variances $\{\lambda_{PC1}, \lambda_{PC2}, \lambda_{PC3}\}$ rather than sample from the historical scenarios.

Weaknesses and Limitations of the Value-at-Risk Model

The 2008 global financial crisis was an expensive lesson to the banking industry, which revealed that the risk models used by many banks (in particular VaR) were found to be inadequate in capturing risks. The VaR model is criticized as being ‘too little, too late’ in that it is often underestimated and late in forecasting. It is also found that the model breaks down during crisis, and is useful only as a ‘peacetime’ tool. Numerous model weaknesses were revealed during that stressful period, and that led to the development of more sophisticated Basel III risks models, and exciting risks research in the academia in recent years. This section gives a high-level summary of weaknesses known about the VaR model.

Not All Risks are Modelable

After the crisis, banks were criticized for over-relying on models for risk management. Clearly not all risks can be modeled. A good working assumption is that any risk which is not of an actuarial nature (i.e. not statistically observable) cannot be modeled within VaR. Some of these risks include the risk of currency controls, global event risks (such as an abandonment of the dollar as a reserve currency), reputation risks, war and its impact on cross-border payment system, banking scandals, etc. For such risks, stress testing is a possible solution: see Chapter X in this Handbook (Rowe 2014).

Nevertheless, due to regulatory requirement, the banking industry (or to be exact certain qualified banks) still model VaR for operation risks, where the data set that supports its calculation is often scarce. This is not an ideal situation as it means that the measurement error for such "OpVaR" is huge (possibly in the same order of magnitude as the OpVaR itself).

Liquidity Effects

Liquidity risks are absent from the VaR input. These include the effects of bid-offer spreads dynamics and also price impact. The latter refers to impact of volume of trades on price stability. For example, the sale of a $1 billion trade at once will pressure the price to shift downwards more strongly as compared to the sales of a $100 million trade ten times over the course of a day. Before the crisis, there were research papers in the academia that tried to incorporate liquidity risk (bid-offers) into VaR, the so-called liquidity-VaR or LVAR. Post crisis, the Basel committee has somewhat influenced the course of model development by requiring that liquidity risk be categorized by product and modeled in terms of different risks horizons. Such Basel III models are still being developed (or fine-tuned) at major global banks and the industry has yet to agree on a standard best practice at the time of this writing.

Losses beyond VaR

VaR is the quantile/cut-off point at the left tail of the loss distribution. Thus, VaR is oblivious to the loss points larger than VaR. Indeed, two distributions with different tail shapes can have the same VaR value but have very different extreme risk profiles. Another measure that has been used in this context for a long time already is the expected shortfall, ie the expected loss, conditional being within a certain quantile in the tail.

Mapping Issues and Historical Data Limitations

A key issue to deal with when running market risk VaR models is the historical data. In a typical bank, the risk factor mapping can contain an order of magnitude of 100,000 risk factors, each of them being a time series. This is a tremendous amount of data to clean, process and maintain in the bank's system. Among the data issues faced, the most insidious is the problem of data asynchrony. That means different risk factors' time series are out of synch and hence the portfolio correlation structure is broken (and VaR blatantly incorrect). This could happen for a number of reasons:

1. Two data series are snapped at different times. For example, USD 10-year swap snapped at London close versus NY close. Using data from Jan 2011 to Dec 2012, the effect on 99% VaR of a typical portfolio can be 13% due to this effect alone.
2. Two data series are snapped at same time, but their market closes at different time zones. For example, a USD-denominated Asian bond closes at Tokyo close, but its hedge, a USD swap closes at NY time.
3. Two data series from markets from the same time zone is snapped at the same closing time, but one of them is less liquid and hence its quotes are updated less frequently. For example, a more liquid CDS hedging a less liquid cash bond of the same obligor name.

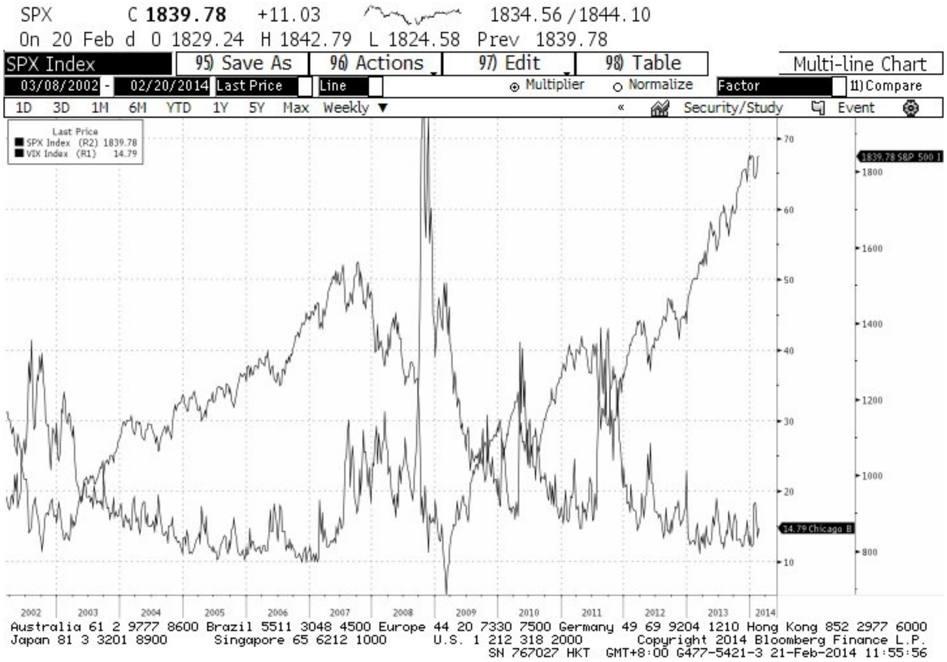
As seen in the above examples, the problem of data integrity often affects hedged positions and that causes the misrepresentation of basis risks. The LTCM debacle in 1998 which involved a consortium bail-out organized by the FED is a classic case illustrating the difficulty in modeling and managing basis risks. In that incident, the basis spread between long off-the-run bonds positions versus short more liquid benchmark bonds blew apart. It is often difficult to model such basis risk because of the absence of reliable data.

Furthermore, the more risk factors a bank include into its VaR model, the more imprecise the model becomes; this is due to the "curse of dimensionality". A typical VaR system in a bank uses 500 points (or 2-year worth of data) but runs more than 50,000 positions (aka dimensions). It is mathematically impossible a properly simulate this series without reducing the dimension: a covariance matrix in 50,000 dimensions has about 1.25bn coefficients, but we only have 25m data points to calculate them, so the system is greatly underdetermined and the resulting covariance matrix will have a lot of degenerated dimensions with zero variance.

Procyclical Risks

It is a well-known fact that VaR or volatility in general is low during a market rally and abruptly high during a market crash. This phenomenon is also known as the 'leverage effect' and can easily be seen by plotting the S&P 500 index against the VIX index (which is the volatility of the US equity market implied by the option market). See Figure 6.

Figure 6. Bloomberg screenshot of S&P 500 index vs. VIX index.



SPX C **1839.78** +11.03 1834.56 /1844.10
On 20 Feb d O 1829.24 H 1842.79 L 1824.58 Prev 1839.78

SPX Index | 03/08/2002 - 02/20/2014 | Last Price | Line | ⊙ Multiplier ○ Normalize | Factor

Australia 61 2 9777 8600 Brazil 5511 3048 4500 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 3201 8900 Singapore 65 6212 1000 U.S. 1 212 318 2000 Copyright 2014 Bloomberg Finance L.P.
SN 767027 HKT GMT+8:00 G477-5421-3 21-Feb-2014 11:55:56

The problem is that Basel regulation requires that a bank's capital be based on VaR. Hence, when the market is in a boom phase, VaR and capital requirement become benign (low), and this encourages balance sheet expansion of banks and the purchases of more risky assets. There is a degree of herding at work because most banks will want to maximize its usage of capital to compete with other banks. Conversely, during the bust phase, the market falls; VaR and capital requirement increase abruptly. This would cause banks to liquidate risky positions in order to manage required regulatory capital. In effect, the regulatory rules inadvertently cause banks to buy rallies and sell breaks; this herding behavior will pressure the markets and amplify the boom-bust cycle. This undesirable effect is called procyclical risk and has been getting a lot of attention since the crisis. Basel III responded somewhat to those concerns, for example by introducing countercyclical capital buffers and by introducing a stressed VaR figure whose risk parameters are frozen at stress-scenario values.

Extremistan and Black Swans

The idea of extremistan was made popular by Nassim Taleb, author of the New York Times best seller, The Black Swan (2007), and a strong critic of VaR models. Extremistan refers to a class of probability structures that are not measurable at the extreme tails of the distribution. Such distributions exhibit characteristic 'fat tails' and the rare events that make up the tails are atypical — meaning past occurrences offer no guidance on the magnitude of future occurrences — hence, not amenable to statistics. Examples of extremistan phenomena include destruction from flu pandemics, world wars, ponzi schemes, and the wealth creation of the super-rich, technological breakthroughs, etc. Taleb termed such unpredictable events Black Swans. In the social sphere, Black Swans are often caused by scalability and positive feedback of thinking participants. In today's world of electronic trading, banking interconnectedness, opaque derivatives markets and fiat money, the unconstrained use of leverage can create the dire possibility of massive financial losses and crisis contagion, as happened in 2008.

If one believes that such financial crises are extremistan events, then the use of VaR models to predict low probability events is questionable. Because of finite sampling, the risk modeler can always estimate VaR to the stated degree of confidence level, but in the presence of extremistan, such a number is misleading and dangerous because it gives the user a false sense of security; the true risk could be much more severe. Taleb's idea of extremistan provides a wise warning of the misuse of models and statistics.

Dangerous Nonlinearity

A common mistake in risk management is to assume that extreme losses that can threaten a bank's survival can only happen when the market is under stress. Actually, a bank is hurt by the payoff $g(x)$ of its positions instead of the market movement x. Since the payoff $g(\cdot)$ can be nonlinear or convex, the losses can be large even though the movement in the underlying market (x) is relatively regular.

The payoff of bonds for example is convex but this convexity is relatively mild except for extremely long-dated and high coupon bonds. More convex are the payoffs of option products especially exotic options with discontinuous payoffs. An option causes the P&L distribution to be skewed to one side and fatter than normal (leptokurtic); unfortunately, skewness and kurtosis are not well measured statistically, even for models such as historical simulation VaR. This is because of the scarcity of sampled data and because these moments are prone to data outliers. Hence, even if the market is in a regular (non-crisis) state, nonlinear risk is something that a risk manager should monitor closely using a variety of other tools such as stress scenario matrices, and gamma (or convexity) limits.

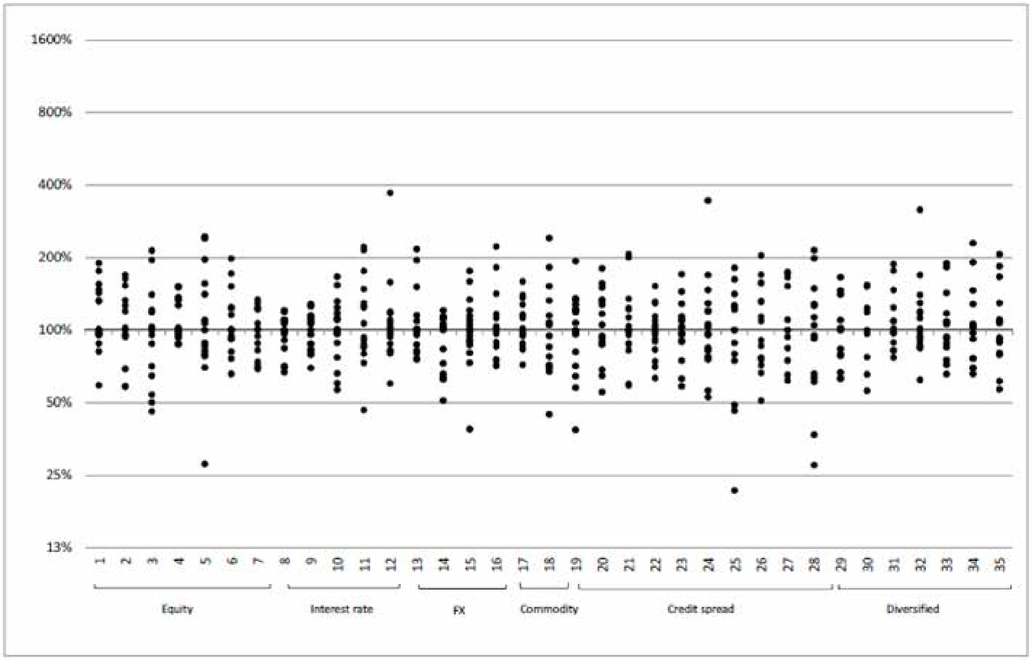
More dangerous is the nonlinearity that arises because of market impact. When a market is under stress or is illiquid, selling 100 blocks of securities at a go creates a much larger price impact on the market (and P&L swing), than selling 10 blocks of securities ten times over a period. Such dangerous nonlinearity occurs during crises (even for linear products) and will be completely missed by VaR models. See also the section "Liquidity effects" above.

Inconsistency in Estimation across Banks

The inability to accurately fit the distribution at the extreme loss tail of the sample, and the huge data challenges discussed above, meant that banks often obtain different VaR results for identical portfolios. In practice, banks build their own internal models (VaR) with different methodologies, parameters and calibrations — such as observation window length, weighting schemes, risk factor mappings, pricing models, parameter calibrations, system implementations, data sources, etc.

A study by Basel in Dec. 2013 shows that banks compute widely varying VaR results for the same test portfolios—see Figure Y.2. The analysis covers 17 banks of 6 jurisdictions. Each bank was given 35 different test portfolios covering a range of asset classes. The VaR from each bank is compiled, normalized to 100%, and plotted as a point in Figure 7.

Figure 7. Dispersion of normalized VaR results for all portfolios



Source: BCBS Dec 2013, "Regulatory consistency assessment program: Second report for RWA for market risk"

The figure revealed that the VaR across banks widely ranges from half to double the median value. Such inconsistency in reported VaR is worrying because it questions the efficacy of the metric and regulators cannot apply appropriate and consistent capital charges to safeguard banks.

Correlation Framework is Fragile

Linear (or Pearson's) correlation is a fundamental concept in Markowitz portfolio theory and is widely used in risk management to "sum" risks, but correlation is a minefield for the unaware. The key is to remember that the estimation of correlation involves "drawing the best straight line" across the bivariate scatter plot i.e. correlation measures the clustering around a straight line. This "drawing" is only good if the data "cloud" is elliptical, otherwise the estimation is biased, and indeed correlation will sometimes behave unintuitively. To describe the relationship between variables fully, one would normally at least need to use copulas (see the section on Copula) but due to the paucity of the data it is impossible to determine which copula is correct. Even where copulas are used, the industry typically uses the simplest copula (the Gaussian copula) in selected applications, such as to model default risks.

Systemic Risks and Breakdown of Model during Crisis

One of the key drawbacks of VaR model is that, by design, it does not capture the dynamics of feedback loops in prices; thus, during a crisis, VaR models will fail. The topic of systemic risk has gained a lot of attention after the 2008 crisis. It helps to first understand the mechanism of a market crisis. The financial markets are seldom in a state of equilibrium as suggested by classical economic theory. As early as 1987, George Soros introduced the idea of reflexivity in his book The Alchemy of Finance, which suggested that market prices (because of feedback loops) are always in an unstable state of overshooting; and it is the large and sudden corrective moves that constitute a crisis. Due to structural features of the market, participants will often build up crowded trades and exhibit herding mentality. For examples, the systemic involvement of global banks in credit derivatives in the years leading up to the credit crisis in 2008, and the collective speculation in Yen-carry trades by investment banks/ hedge funds in 1998. In the latter case, financial institutions purchased high yielding currencies (such as AUD dollar), funded by near zero Yen interest rates. Both ended badly in massive destruction of wealth and fallouts as participants rushed to exit trades when the market reversed.

Such panic selling gives rise to two effects: 1) contagion, the spillover effect among large players or institutions, 2) positive feedback loop within the market itself, panic selling which depresses prices which then leads to more panic selling. Contagion is worsened by the fact that financial institutions are often counterparties to each other, and hence, their balance sheet is interconnected. A contagion will likely cause institutions to tighten credit lines and margin requirements to other institution that are heavily exposed to the falling market. An institution may even sell into the same market to hedge itself; or to cut losses to meet margin calls. These actions give rise to feedback loop, which breaks the stationary assumption of VaR models and creates serial correlation in the data series. This string of losses or drawdowns is what hurts the banks most and is not well contained by VaR capital. Such a regime shift happens typically in a short span of time, which makes it difficult for VaR models to capture. As an extreme example, the flash crash on Thursday May 6, 2010, in which the Dow Jones index plunged about 1000 points (roughly 9%) only to recover within minutes, was an incident caused by high-frequency trading.

This sudden appearance of serial correlation gives rise to stochastic (or changing) volatility; it is mathematically shown that stochastic volatility causes fat-tail distributions and the phenomenon of volatility "smiles" reflected in option markets. Unfortunately, standard VaR models use a slowly moving observation window; that means the VaR metric will be late in picking up the new regime. At the portfolio level, a crisis regime shift is typically accompanied by a breakdown of correlation as well (i.e. relationships in the past no longer applies); markets tend to fall together and become highly correlated during a contagion.

Over the years, the academia has introduced conditional models (such as the GARCH model, 1986), which are more adept and responsive to changing volatilities. However, such models are difficult to implement for large portfolios typical at banks and are too complex to maintain and communicate.

Regulatory Impetus on Model Development

The BCBS is instrumental in guiding the development of the framework for banking regulation and models for capital calculations. A core element of supervision is the idea of capital adequacy—a bank has to hold sufficient capital for the business it is engaged in so as to buffer itself against unforeseen extreme losses.

The first VaR model was introduced into regulatory capital requirements in the 1996 Market Risk Amendment. The Basel 2 model then extended the VaR concept to credit risk and operational risk, with a different time horizon (1 year) and a different confidence level (99.9%). The trading book approach was not substantially changed from the 1996 Amendment.

The basic idea of Basel 2 was that sophisticated banks could use their own internal models (obeying certain constraints) for risk calculations on the credit side, as it was has already been case on the market risk side. Basel 2 thus was meant to provide economic incentives for banks to research and develop better models to measure risks, and to obtain more efficient ways to utilize capital.

Under Basel 2, market risk (MR) is divided into general market risk (GMR) and specific risk (SR). Based on 99% confidence level, 10-day risk horizon. For example, a bond issued by IBM would contain the interest rate risk (GMR) and the risk coming from the movement of the credit spreads of IBM, the issuer (SR).

Basel 2.5 was the first BCBS policy response to the global financial crisis; released in July 2009, it was the precursor to Basel III, and the Basel 2.5 changes were eventually subsumed into Basel III. The new rules pertaining to market risk capital calculations and models are:

1. Incremental Risk Charge (IRC): applicable to any bank that has internal models approval for specific risks. Basel provided high level guidance on how this should be modeled:
    - Based on 99.9% confidence level, 1-year risk horizon, must have basic framework similar to the IRB model. Hence, the Vasicek model is implied.
    - Introduced multiple liquidity horizons for products of different liquidities. The most liquid products will have a liquidity horizon of at least 3 months.
    - Assumes a constant level of risks, i.e. positions are assumed to be rebalanced at every liquidity horizon (to ensure a constant VaR), and for many steps up to the risk horizon. Many banks have chosen to implement a multistep one-factor Gaussian copula model.
    - Rating migration, default risk, optionality and their cross correlations must be modeled.
    - Concentration risks must be reflected by having a granular classification/ differentiation of positions.
2. Credit securitization products (such as tranches, CDOs and credit correlation instruments) are excluded from the IRC and modeled instead using a Comprehensive Risk Model (CRM). The CRM in theory should capture the myriad of risk factors pertinent to these “toxic” products. Since 2008, most banks would have chosen to sell-off such legacy businesses over the years, instead of face the arduous task of designing/ maintaining the capital-punitive CRM.
3. Stressed VaR (SVaR): an additional capital requirement to cover the weakness that the standard VaR underestimates risks during a crisis, and

as a first buffer against procyclical risk. SVaR is just the 99%/10-day VaR calculated using a 1-year (fixed) observation period of high stress.

Basel III was finally released in Dec 2010. It contained a number of important changes, for example with respect to capital buffers and counterparty requirement, but it did not introduce any changes to the market risk framework other than those that had been introduced by Basel 2.5.

An ongoing process (as of 2014) with the BCBS is the fundamental review of the trading book. Key proposals in this respect are

1. A replacement of VaR with 97.5% expected shortfall which empirically is close to the 99% VaR on real-world distributions (ES). The latter would capture the shape of the tail of the loss distribution and is a coherent risk measure.
2. Use of stressed calibration for risk models for the purpose of capital. This recognizes the BCBS objective of reduction of cyclicality of risk measures.
3. Internal models will be approved at the more granular (desk) level rather than at bankwide level, and will be conditional on good backtesting and P&L attribution processes.
4. Additional charge to cover non-modelable risks, such as that arises from data issues.
5. Impose some constraints on diversification benefit in internal models. This recognizes the breakdown of correlation during crises.
6. Comprehensive incorporation of liquidity risk into ES. This uses the liquidity horizon construct (like the IRC) but whereby the horizons will be prescribed by BCBS based on asset classes.

B02. Advanced VAR Models - Univariate

Backtesting

From the previous section, it is obvious that banks could choose from many possible VaR methodologies and parameters settings such as observation periods and weighting schemes, etc. Furthermore, the accuracy of the VaR model critically depends on the quality of the data collected by the bank and the IT system implementation. In practice, many factors can lead to inaccurate, imprecise or inconsistent VaR numbers. Since VaR is used for computation of regulatory capital, there is an important implication on banks' safety buffer and cost of capital. Thus it is essential to perform internal testing, validation and review of VaR models

To ensure the VaR models are fit-for-purpose, regulators require that banks backtest their VaR models regularly. In the case of market risk there are two variations of backtesting that serve different purposes. These involve comparing ex-ante VaR estimates with ex-post values of (a) actual P&L in the applicable periods and (b) hypothetical P&L assuming *constant* positions for the applicable periods. For example, the VaR computed for positions at time T is compared against P&L of the same position at (T+1). This is done for all T in the sampled period. In the case of (a), the position changes across the sample, in the case of (b) the latest position is used and is assumed to hold every day in the past sample. The first approach is the one required of banks under the Basel rules.

The correctness of the actual recorded P&L *over various cycles* is what risk systems are ultimately designed to achieve. Hence comparing VaR estimates to *actual* P&L (based on the bank's changing positions) is a useful part of any backtesting process. A problem with such a test is that there can be many reasons why actual P&Ls may exceed the risk estimates more frequently than theoretically expected. This may occur not just because of weaknesses in the VaR model, but could also be due to contamination of the past P&L by factors such as intraday P&L and reserve adjustments that were not properly removed from the P&L for the purpose of backtesting, an old IT bug, etc.

But when backtesting reveals weaknesses in the VaR system, it is important to be able to diagnose the source of the problem as per the system and positions *today*. This is where the second form of backtesting is a useful complement. Furthermore, VaR assumes that the portfolio position remains static over the risk horizon, an idea consistent with the use of *hypothetical* P&L in backtesting. Put another way, we should be more interested in the correctness of the VaR model for today's positions than the correctness on the VaR model in the distance past.

Exception Measurement and Basel Rules

The main consideration in backtesting is whether the P&L series exceeds the corresponding ex-ante VaR estimate within the predicted frequency. This can and should be repeated for VaR with different confidence levels. In the simplest form, the backtesting procedure counts the number of times that the actual portfolio P&L breaches the VaR estimate, and compare that number to the confidence level used. For example, if a confidence level were 99% for a daily VaR, we would expect on average a 1% chance of a daily P&L breach of the VaR. The breach is also called exception or exceedance. We observe the number of exceptions that have occurred in the past. If the number of exceptions is approximately 1% of the total number of days in the observation period, we may conclude the VaR model is adequate. Otherwise, the VaR model is aggressive if the percentage of exception is evidently greater than 1% or conservative if it is evidently smaller than 1%.

Current regulatory guidelines require banks to use the so-called "traffic light approach (TLA)" as a standard method to backtest their VaR systems. This approach counts N, the number of 1% VaR violations (exceptions) in the previous 250 trading days. Regulators rely on this approach to justify the soundness of a bank's internal VaR model and to adjust the multiplier for capital charges accordingly. If a bank qualifies for Basel's internal models, its *market risk capital requirement* is given by:

$$MRC_t = \max\left(\mathrm{VaR}_t^{99\%} ,\ K_t \frac{1}{60}\sum_{i=0}^{59} \mathrm{VaR}_{t-i}^{99\%} \right) + SRC_t \tag{53}$$

which involves taking the larger of the most recent VaR and the 60-day average VaR. The multiplier K_t is determined by classifying the number N into three categories as follows

$$K_t = \begin{cases} 3.0 & \text{if } N \le 4 \quad \text{Green} \\ 3.0 + 0.2 \times (N - 4.0) & \text{if } 5 \le N \le 9 \quad \text{Yellow} \\ 4.0 & \text{if } N \ge 10 \quad \text{Red} \end{cases} \tag{54}$$

If the VaR is truly an unbiased estimate of the quantile, we would expect 2.5 violations in 250 days statistically by definition. The multiplier remains at its lowest level of 3 if the VaR exceptions are less than 4. If the VaR is violated more frequently, a bank is penalized more by higher capital charge via a bigger multiplier. In the red zone, the VaR model is deemed inaccurate (i.e. systematically understates risk) and immediate corrective actions are required to improve the VaR system. In contrast, too few violations (<2) would imply that the bank's model is overly conservative — it systematically overstates risk.

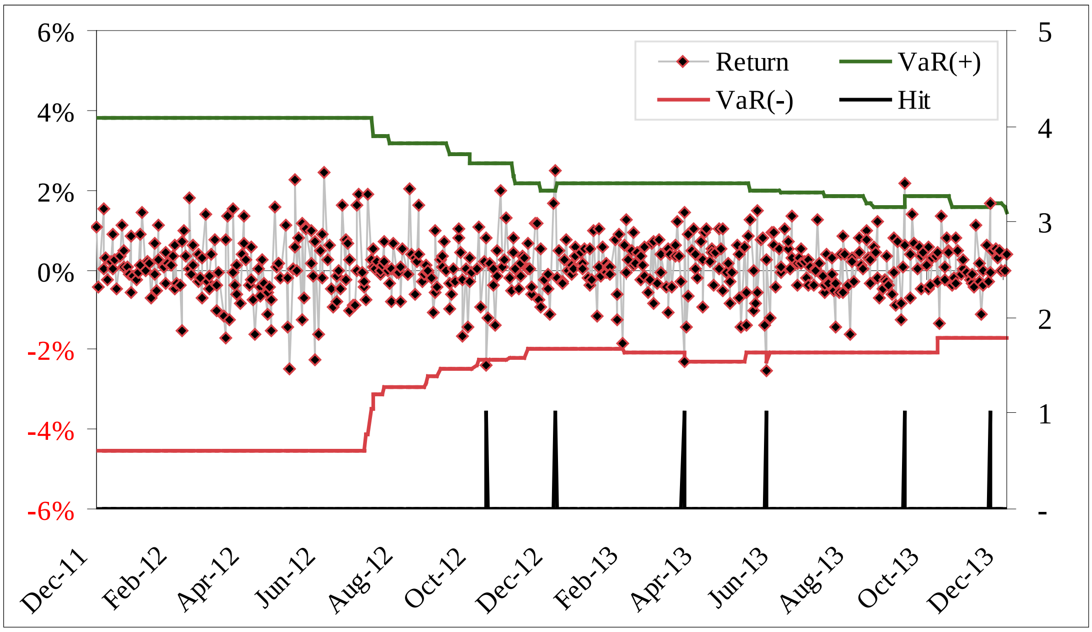


Figure 8. Backtesting using TLA for a simple equity portfolio

As an illustration, the spreadsheet [VaR Backtesting.xls] shows the TLA backtesting using a portfolio comprising an equity index asset (SPX index). The VaR calculation window and the backtesting window are both of 250-day. With 3-year P&L (expressed in % return) data from 2011 to 2013, we have approximately 500 VaR estimates. The results are shown in Figure 8. There are in total six VaR violations (including positive and negative violations) plotted as ‘hit’. Overall the TLA shows the VaR model falls into the green zone most of the time, but turns to yellow status briefly in late 2013.

Frequency Based Backtests

To determine if a VaR model is well specified and ‘good’, it suffices to test that it satisfies the hypothesis of unconditional coverage and independence. The hypothesis of unconditional coverage states that the expected frequency of observed exceptions does not differ significantly from the assumed probability p (e.g. $p = 1 - \alpha = 0.01$ for a 99%-VaR). The hypothesis of independence means the exceptions should occur independently across time. In other words, the exception sequence should be evenly spread out and not clustered together, otherwise serial dependence may arise in the P&L (i.e. in the market) that is not capture by the VaR model and hence the VaR model is mis-specified.

The traffic light approach is simple, but it is merely an empirical formula provided by the regulator that is only applicable to 99%-VaR. To improve this counting test, a few approaches have been developed based on statistical hypothesis testing. One of them is called ‘Kupiec test’ proposed by Kupiec (1995) [6], a cousin of TLA, which examines the null hypothesis of unconditional coverage rate p^*, e.g. $\mathcal{H}_0: p^* = 0.01$. Kupiec test is analogous to TLA as both focus on testing the proportion of exceptions (i.e. coverage rate). The difference is that TLA is a one-sided test with the alternative hypothesis $\mathcal{H}_1: p^* > 0.01$ while the Kupiec test is a two-sided test with alternative hypothesis $\mathcal{H}_1: p^* \neq 0.01$. This means, the

Kupiec test will fail if the VaR model is either too aggressive or too conservative whereas the TLA test fails only if the VaR model is too aggressive.

The idea of Kupiec test is to test whether the observed violation frequency is consistent with the frequency predicted by the model. For example, let's consider a daily series of ex post portfolio return r_t and a corresponding series of ex ante VaR forecast VaR_t^{1-p} with coverage rate p(e.g. $\mathrm{VaR}_t^{99\%}$ with $p = 0.01$). This yields a probability $\mathbb{P}\left[r_t < \mathrm{VaR}_t^{1-p}\right] = p^*$ at one sided $1-p$ confidence level. In particular, under the null hypothesis $\mathcal{H}_0\colon p^* = p$, the number of violations x follows a binomial distribution. Given that a total of n P&L observations, the probability of x violations can be calculated as

$$\mathbb{P}[x|n,p] = \binom{n}{x} p^x (1-p)^{n-x} \tag{55}$$

Rather than directly calculate probabilities from the discrete binomial distribution, Kupiec proposed a "proportion of failures" (POF) coverage test, a variant of likelihood ratio test, relying on the statistic $-2\ln\Lambda$, where the Λ is the likelihood ratio constructed as

$$\Lambda = \frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} = \frac{p^x(1-p)^{n-x}}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} \tag{56}$$

By assuming $\mathcal{H}_0$, the statistic $-2\ln\Lambda$ asymptotically follows a centered chi-squared distribution $\chi^2(1,0)$ with 1 degree of freedom as n increases. For a given significance level α (e.g. 95%), we can construct a non-rejection interval $[l,\ u]$, such that $\mathbb{P}[x < l] \le \frac{\alpha}{2}$ and $\mathbb{P}[x > u] \le \frac{\alpha}{2}$, where the interval bounds l and u are the two solutions of x making the $-2\ln\Lambda$ equal to the α quantile of the $\chi^2(1,0)$.

Both Kupiec and TLA methods focus only on testing the frequency, and ignores the potential serial dependence in VaR exceptions. The latter weakness can be detected by a 'duration based approach' developed by Christoffersen and Pelletier (2004) [7]. The idea is that if a 1-day VaR model is well specified for a frequency of p, then every day the *unconditional expected duration* should always be $1/p$ days. This defines the null hypothesis that the duration has no memory on the length it has already undergone and it always has an expected mean of $1/p$ days. Since *exponential* distribution is the only memory-free continuous random distribution, then under the null hypothesis the duration must follow an exponential distribution. In order to establish a statistical test for duration independence, an alternative hypothesis must be specified that allows for duration dependence. Hence, the test must be conducted on a distribution that nests the exponential distribution as a special case, e.g. *Weibull* distribution. A likelihood ratio test is then conducted to see whether the special case holds. If it holds, then the null hypothesis is passed and the VaR model has the duration independence.

There have been many other statistical tests based backtesting techniques developed in recent years. A good reference for these advance tests is Wehn (2008) [8].

Distributional equality based backtests

All the tests mentioned above check for frequency of the VaR exceptions, which are counts of rare events in general and are certainly less informative than the entire PL distribution. In order to fully utilize the entire P&L data, new approaches have been developed to backtest the whole distribution. Such tests first transform the realized P&L's in terms of their forecast probability CDF values and then use the transformed data to test the equality of the probability distributions. Because the tests are performed on the whole distributions, they have more diagnostic power than the frequency based approaches.

Full distribution backtesting exploits the *Rosenblatt transformation*, which is expressed as

$$u_t = F_t(x_t) \tag{57}$$

where x_t is the realized P&L at time t and $F_t(\cdot)$ is the CDF of the forecasting P&L distribution (i.e. the P&L distribution used to derive the VaR for day t; in hsVaR, it is the 250-day P&L vector up to day $t-1$). Since the true density function $F_t(\cdot)$ is unknown and is estimated by a VaR model ex-ante (i.e. using dataset one day earlier), we use the estimated $\hat{F}_t(\cdot)$ to perform the transformation, which gives a series of $\hat{u}_t$. For parametric VaR models, the functional form of $\hat{F}_t(\cdot)$ is known and can be used analytically. For non-parametric VaR models (e.g. hsVaR), the $\hat{F}_t(\cdot)$ is estimated as an empirical CDF, which is simply defined as

$$\hat{F}_t(x) = \frac{\text{number of elements in the P\&}L\ vector \le x}{\text{total number of elements in the P\&}L\ vector + 1} \tag{58}$$

We transform the series of x_t to $\hat{u}_t$ for the past n days (say $n = 500$). For the VaR model to be well behaved, the null hypothesis is the series of $\hat{u}_t$ must be uniformly distributed between 0 and 1. The uniformity in the transformed $\hat{u}_t$ series can be tested using the *Kolmogorov-Smirnov* statistics, which is defined as

$$\begin{aligned} D_n &= \sqrt{n} \max_{t=1,\cdots,n} |\hat{u}_t - u_t| \\ &= \sqrt{n} \max_{t=1,\cdots,n} \left|\hat{F}_t(x_t) - F_t(x_t)\right| \end{aligned} \tag{59}$$

The steps to construct the statistics are as follows:

1. Define a backtesting window of n-days such that $t = 1, \cdots, n$. For example, we let $n = 500$.
2. For each day t, we transform the realized P&L x_t to $\hat{u}_t$ using the estimated density function $\hat{F}_t(\cdot)$ of the P&L distribution (the empirical CDF in hsVaR or parametric CDF in pVaR using *ex-ante* (one day behind) dataset).
3. Sort the $\hat{u}_t$ in an ascending order to obtain the empirical uniform CDF. This gives a vector $\hat{u}_i$ for $i = 1, \cdots, n$ with $\hat{u}_1$being the smallest number closest to 0.0. For comparison, we define another vector u_i for $i = 1, \cdots, n$, the true uniform CDF such that

$$u_i = \frac{i-1}{n} \tag{60}$$

The *Kolmogorov-Smirnov* test is performed to examine the uniformity of the $\hat{u}_i$ series by benchmarking against the u_i series.

4. Kolmogorov-Smirnov statistic is defined as the largest absolute difference between the vector $\hat{u}_i$ and u_i multiplied by the square root of n. If the statistic is smaller than the critical value at a given confidence level, we say the null hypothesis is passed (or not rejected), the VaR model is considered to be well-behaved.

There are other statistics to test this as well, such as the *Anderson-Darling* test, etc. But this is outside the scope of this book.

The spreadsheet [VaR Backtesting.xls] illustrates this advanced backtest using a simple example. The data used in the example are 3-year log-returns of SPX index from 2011 to 2013. For simplicity, we assume a rolling observation period of 250-day returns (ending at date T) gives a good estimate of the P&L distribution (at date T) and use it to perform the out-of-sample Rosenblatt transformation of the P&L at date (T + 1). After deriving the $\hat{u}_t$ series (roughly 500 points), we plot a histogram (shown in Figure 9) to visualize its uniformity. It clearly shows a slightly lower than uniform frequency at both loss and gain tails. This indicates the P&L distribution in general forecasts a conservative (or overstated) VaR. On the other hand, if the tails of the frequency distribution is higher than the body, it suggests that any VaR forecast (that comes from that distribution) will likely be underestimated. A VaR which is unbiased will come from a P&L distribution in which its Rosenblatt-transformed distribution is reasonably uniform. Such underestimation or overestimation of VaR is likely caused by regime shifts in the dataset and stochastic volatility of the markets in general. We also draw the empirical uniform CDF vs. the theoretical uniform CDF as a QQ plot as per Figure 10. The discrepancy in the two tails also suggests the empirical tails are fatter than in theory — hence, computed risk measures such as VaR will be overestimated.

Figure 9. Out-of-sample Rosenblatt transformation

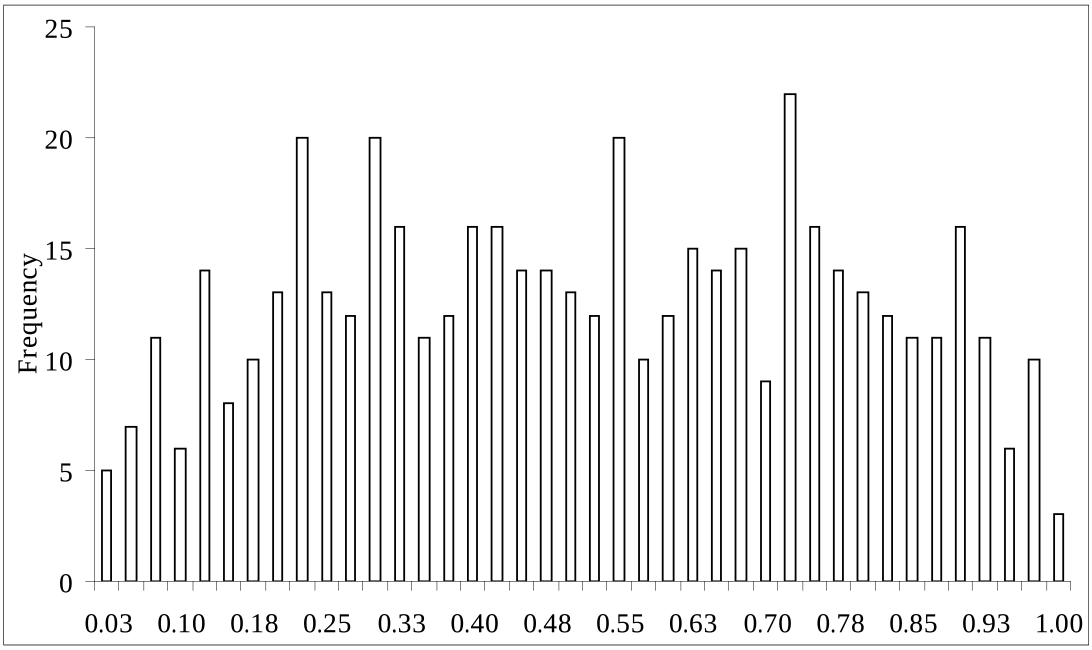
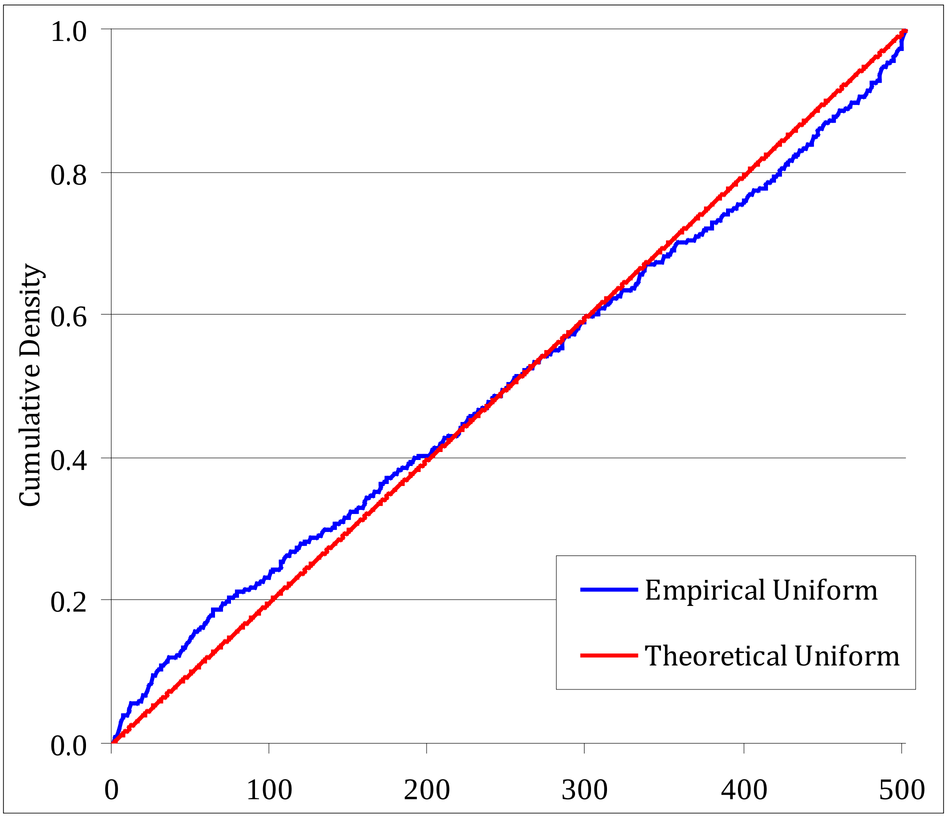


Figure 10. Empirical uniform CDF (via Rosenblatt transformation) vs. theoretical uniform CDF



Extreme Value Theory

Extreme Value Theory (EVT) is a field of applied statistics traditionally used in the insurance industry and now increasingly applied in the area of operational risk. It provides estimation techniques to forecast extreme events with low probability of occurring. The purpose of this section is to provide a brief introduction of what EVT can do in the area of VaR. We will end with a simple Excel example of EVT VaR. As we shall see, EVT does not attempt to model the tail process; hence, one can think of it as a tool to fit the tail distribution in a statistically correct manner. The reasoning is that if the rare event lies outside the range of available observation, it seems essential to rely on good fundamental methodology. The reason why EVT gained some acceptance in risk management is because return distributions in financial markets are severely fat-tailed during times of crises.

Classical EVT

The fundamental model of EVT describes the behavior of the *maxima* of a distribution (the theory discussed here also applies to the minima because properties of the minima can be obtained from those of the maxima by a simple sign change). Consider a collection of n observed daily returns $\{r_i\}$ for $i = 1, \cdots, n$ where we ignore the sign of returns and express losses as positive numbers. In VaR we are interested in modeling extreme losses l_n. So consider the worst-case loss such that $l_n = \max\{r_i\}$. We denote the cumulative distribution function (CDF) of a random variable x as $F(x)$. Assuming the returns are i.i.d., the CDF of l_n, i.e. $F_n(x)$ can be easily derived:

$$\begin{aligned} F_n(x) &= \mathbb{P}[l_n < x] \\ &= \mathbb{P}[r_1 \le x, \cdots, r_n \le x] \\ &= \prod_{i=1}^{n} \mathbb{P}[r_i \le x] \\ &= \prod_{i=1}^{n} F(x) \\ &= F(x)^n \end{aligned} \tag{61}$$

However, this CDF becomes degenerated as n increases to infinity. Namely, it becomes a Heaviside step function translated to a value u, i.e.

$$F_n(x) \xrightarrow[n\to\infty]{} \begin{cases} 0 & \text{if } x < u \\ 1 & \text{if } x \ge u \end{cases} \tag{62}$$

Since the degenerated CDF is useless in practice, the extreme value theory tries to identify an asymptotic distribution of the *normalized* maximum

$$l_n^* = \frac{l_n - \mu_n}{\sigma_n} \tag{63}$$

as n goes to infinity, where $\{\mu_n\}$ and $\{\sigma_n\}$ are sequences of real numbers and $\sigma_n > 0$. The *Fisher–Tippett–Gnedenko* theorem states that if there exist such sequences, then

$$\mathbb{P}[l_n^* \le z] \xrightarrow[n\to\infty]{} G(z) \propto \exp\left[-(1 + \xi z)^{-1/\xi}\right] \tag{64}$$

where the term

$$z = \frac{l - \mu}{\sigma} \tag{65}$$

is normalized by the location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$. The parameter ξ governs the tail shape of the limiting distribution. In other words, the generalized extreme value (GEV) distribution has a cumulative density function

$$G(z) = \exp\left[-(1 + \xi z)^{-1/\xi}\right] \tag{66}$$

for $1 + \xi z > 0$. Depending on the value of ξ, the $G(z)$ belongs to one of the following distribution families:

1. if $\xi < 0$: *Weibull* family

$$G(z) = \begin{cases} \exp\left[-(1 + \xi z)^{-1/\xi}\right] & \text{if } z < -1/\xi \\ 1 & \text{otherwise} \end{cases} \tag{67}$$

2. if $\xi = 0$: *Gumbel* family (the function $G(z)$ takes the limit as $\xi \to 0$)

$$G(z) = \exp[-\exp(-z)], \qquad z \in \mathbb{R} \tag{68}$$

3. if $\xi > 0$: *Frechet* family

$$G(z) = \begin{cases} \exp\left[-(1 + \xi z)^{-1/\xi}\right] & \text{if } z > -1/\xi \\ 0 & \text{otherwise} \end{cases} \tag{69}$$

Figure 11 shows the probability density functions of the three types of the extreme value distributions

$$g(z) = \begin{cases} \exp\left[-(1 + \xi z)^{-1/\xi}\right](1 + \xi z)^{-1/\xi - 1} & \text{if } \xi \neq 0 \\ \exp[-\exp(-z)]\exp(-z) & \text{if } \xi = 0 \end{cases} \tag{70}$$

which are obtained by differentiating the $G(z)$ with respect to z.

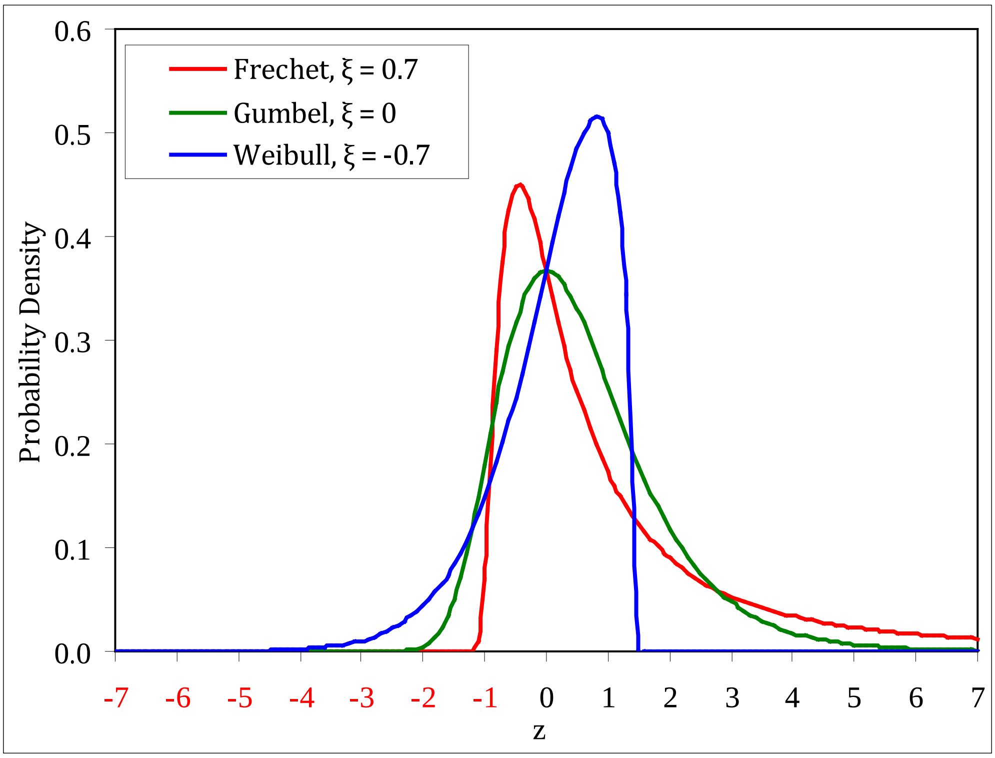


Figure 11. Probability density functions of extreme value distributions

The extreme value theorem claims that the *maximum* of a large but finite number of independent and identically distributed random variables asymptotically (if there is convergence at all) follows a generalized extreme value distribution regardless of the original distribution of the random variables. This has a good analogy with the better known central limit theorem, which states the *mean* of a large number of random variables asymptotically follows a normal distribution regardless of the original distribution of the random variables.

Block Maxima Approach

Classical EVT proves the existence of a limiting distribution for the maxima of i.i.d. random samples. Approaches have been developed for extreme value analysis based on the limiting distribution. One of them is called *Block Maxima* (BM) approach, which estimates the distribution for rare extremes by parametric modeling of maxima taken from large blocks of independent data. The idea here is to first divide the full dataset into equal sized blocks of data, and then to determine the maximum for each block. The GEV distribution (66) can then be fitted to the sampled maxima by say maximum likelihood estimation. The statistical analysis and inference can then be derived from the fitted GEV distribution.

It is obvious that for block maxima to become close to i.i.d., the blocks must be sufficiently long. Furthermore, because the statistical analysis is performed on maxima only, BM approach requires a large set of data, which is often unavailable for VaR estimation (for example, if a block is defined over 50 days to give a single maximum, a common VaR window, say 2 year of data, will only provide 10 usable points, far less than enough for a reliable calibration of BM analysis). For practical reasons, one has to consider other approaches that exploit data more cleverly.

Peak over Threshold Approach

The *Peak over threshold* (POT) is another popular method used in extreme value analysis. It focuses on parametric modeling of independent exceedances above a large threshold in a dataset.

It is convenient to illustrate this concept with an example. Suppose we have a set of data $\{x_i\}$ for $i = 1, \cdots, N$ that are i.i.d., taken from a cumulative distribution function $F(x)$. This can be the daily hypothetical P&L's of a portfolio from historical simulation over the past 500-day scenarios ($N = 500$). To simplify notation for the P&L data, positive numbers denote losses. Let's define a threshold u such that it is the 21th largest loss in the past 500 days. In other words, our analysis focuses on losses beyond the 4% tail. We define $y = x - u$ as the exceedance beyond the threshold. The CDF of the exceedances can then be derived from $F(x)$

$$\begin{aligned} E_u(y) &= \mathbb{P}[X < u + y | X > u] \\ &= \frac{\mathbb{P}[u < X < u + y]}{\mathbb{P}[X > u]} \\ &= \frac{F(u + y) - F(u)}{1 - F(u)} \end{aligned} \tag{71}$$

The asymptotic theorem for exceedance says that if the block maxima of $F(x)$ asymptotically converges to GEV distribution, then the distribution of exceedance $E_u(y)$ asymptotically converges to a *generalized Pareto distribution* (GPD) as u goes to infinity, that is

$$E_u(y) \xrightarrow[u\to\infty]{} P(y) = \begin{cases} 1-\left(1+\xi\dfrac{y}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1-e^{-y/\beta} & \text{if } \xi = 0 \end{cases} \tag{72}$$

where the shape parameters ξ in GPD is identical to that in GEV and the scale parameter β is related to the scale parameter σ in GEV. Differentiating (72) with respect to y, we get the probability density function of GPD

$$p(y) = \begin{cases} \dfrac{1}{\beta}\left(1+\xi\dfrac{y}{\beta}\right)^{-1/\xi-1} & \text{if } \xi \neq 0 \\ \dfrac{1}{\beta}e^{-y/\beta} & \text{if } \xi = 0 \end{cases} \tag{73}$$

Figure 12 shows the probability density functions of generalized Pareto distribution with various shape parameters.

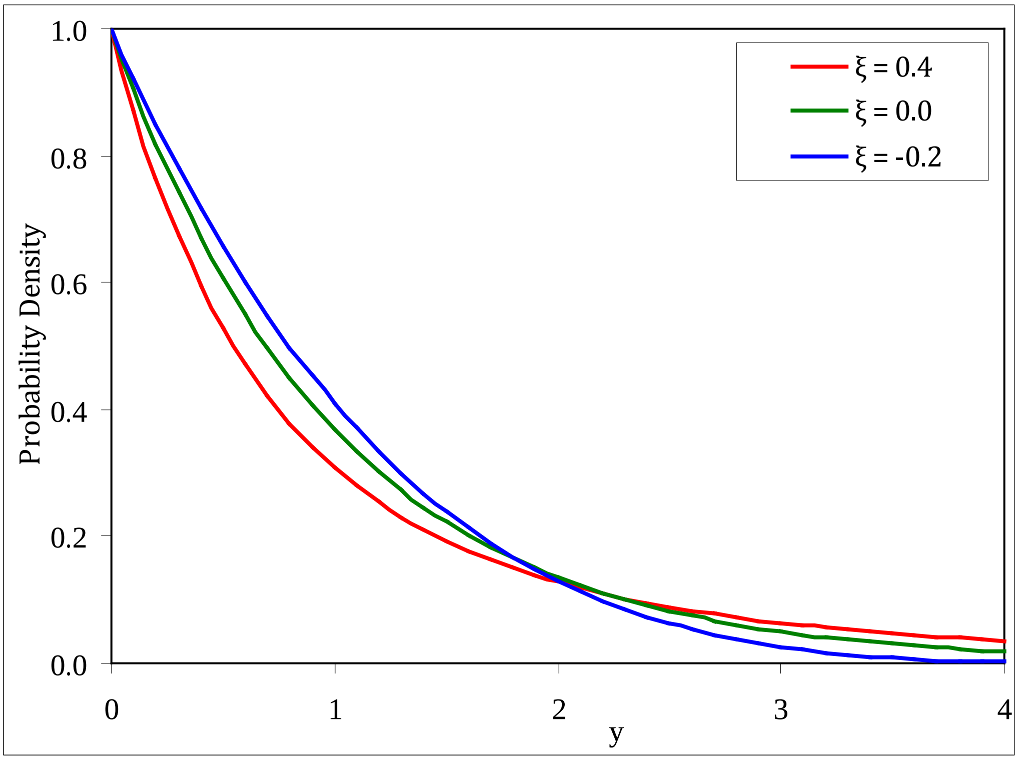


Figure 12. Probability density functions of generalized Pareto distribution (β = 1) with different shape parameters

To calibrate the GPD to our 500-day P&L data, we need to extract the exceedances from the dataset, e.g. there are 20 loss exceedances ($n = 20$) out of a total of 500 P&L's. The GPD distribution function $P(y)$ can be calibrated to the exceedances using MLE by maximizing the log-likelihood function

$$\ln \mathcal{L} = \begin{cases} -n\ln\beta - \left(1+\dfrac{1}{\xi}\right)\displaystyle\sum_{i=1}^{n} \ln\left(1+\beta\dfrac{y_i}{\xi}\right) & \text{if } \xi \neq 0 \\ -n\ln\beta - \dfrac{1}{\beta}\displaystyle\sum_{i=1}^{n}\ln y_i & \text{if } \xi = 0 \end{cases} \tag{74}$$

As an illustration, in spreadsheet [VaR_EVT.xls] we perform the calibration using the Excel Solver function, which yields estimates for parameters $\hat{\xi} = 0.031$ and $\hat{\beta} = 2{,}062$. Once the density function is calibrated, we can derive the conditional distribution

$$\begin{aligned} \mathbb{P}[X > u + y | X > u] &= 1 - \mathbb{P}[X < u + y | X > u] \\ &= 1 - P(y) \end{aligned} \tag{75}$$

Since $x = u + y$, the unconditional distribution is then given by

$$\begin{aligned} \mathbb{P}[X > x] &= \mathbb{P}[X > x | X > u] \cdot \mathbb{P}[X > u] \\ &= [1 - P(x-u)] \cdot \mathbb{P}[X > u] \end{aligned} \tag{76}$$

where $\mathbb{P}[\mathrm{X} > u]$ can be estimated as n/N from the empirical distribution. The VaR is then calculated from the unconditional distribution $\mathbb{P}[X > x]$. For example, if the confidence level is α, then by definition we have

$$\mathbb{P}[X > \mathrm{VaR}_\alpha] = [1 - P(\mathrm{VaR}_\alpha - u)]\frac{n}{N} = (1-\alpha) \tag{77}$$

where VaR_α in this case is a positive number denoting the size of a loss. After rearrangement we have

$$\begin{aligned} P(\mathrm{VaR}_\alpha - u) &= 1 - \left(1 + \xi\frac{\mathrm{VaR}_\alpha - u}{\beta}\right)^{-\frac{1}{\xi}} \\ &= 1 - (1-\alpha)\frac{N}{n} \end{aligned} \tag{78}$$

The last step is to solve for the VaR_α from (78), which gives the formula

$$\mathrm{VaR}_\alpha = u + \frac{\beta}{\xi}\left(\left[(1-\alpha)\frac{N}{n}\right]^{-\xi} - 1\right) \tag{79}$$

Referring to the spreadsheet, we use the formula in (79) to calculate 1-day VaR at confidence levels from 95.0% to 99.9%. The results are shown in Figure 13. As a comparison, the parametric VaR is also included in the plot where pVaR is estimated by assuming a normal distribution for the P&L. Clearly, the VaR estimated from EVT becomes increasingly fatter than normal as the confidence level increases.

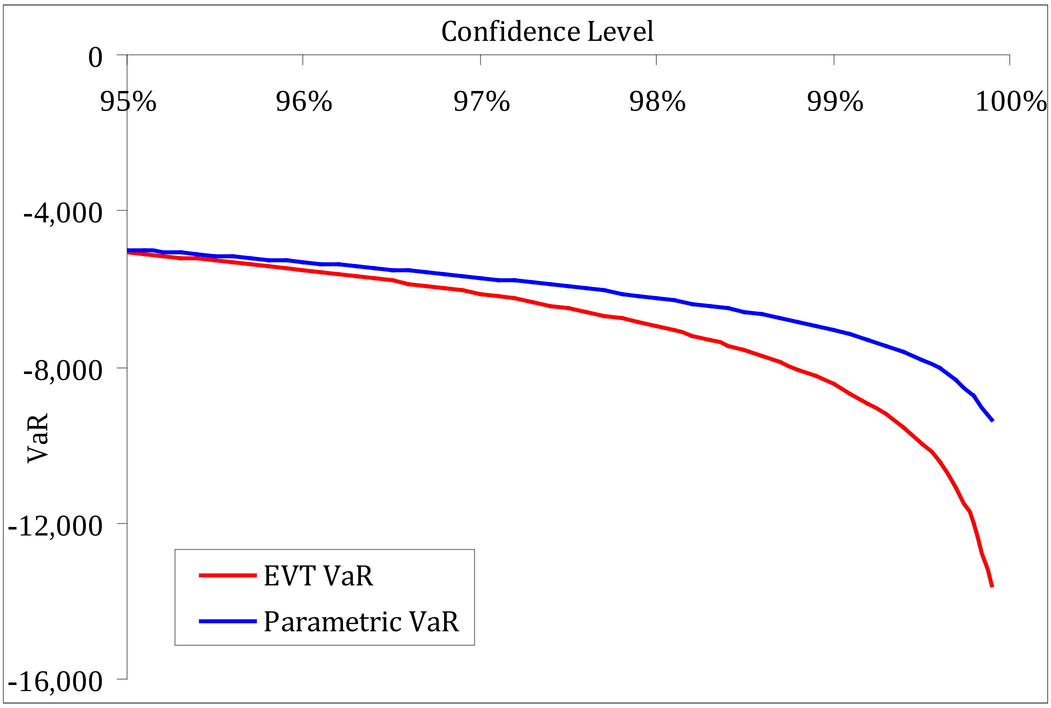


Figure 13. EVT VaR vs. Parametric VaR at different confidence levels

In our analysis, a natural question is: how do we choose the value for the threshold u? We want the u to be sufficiently large so that it meets the asymptotic criteria for POT; meanwhile we also want the u to be small so that the MLE will have enough exceedances to produce reliable model estimation. This is not an issue for large dataset. However in VaR applications, the size of dataset is almost always limited; we typically have only about 500 P&L data points to decide an optimal value for u. A practical 'rule of thumb' is to make the u equal to the 95% percentile of the empirical distribution as an initial guess. When we calibrate the GPD distribution, we need to ensure the estimated parameters $\hat{\xi}$ and $\hat{\beta}$ are positive. If the estimated $\hat{\xi}$ is negative, the reason could be i) the u is improperly chosen or ii) the empirical distribution does not have heavier tails than that of the normal distribution. This effect can be seen from our example, in fact if we choose u = 95% percentile (i.e. the 26th loss in our P&L data), the calibrated $\hat{\xi}$ turns out to be negative, as a result we may raise the u to 96% percentile to perform our EVT analysis.

Expected shortfall

The VaR metric is where a single number used to represent an entire distribution. Clearly, a scalar can never fully represent the detail features of the tail of a multi-variant distribution, but it does provide a great deal of convenience and intuition for risk management. In 1999, Artzner et al. introduced the concept of coherence, a list of desirable properties for any such risk measure. A risk measure F (here the F is assumed to be positive, i.e. the larger the F the higher the risk) is said to be coherent if it has the following four properties:

1. Monotonicity: if $X_1 \leq X_2$ then $F(X_1) \geq F(X_2)$

That means if a portfolio has values (X) lower than another under all possible scenarios, then its risk as measured by F must be larger.

2. Homogeneity: $F(aX) = aF(X)$

   That means increasing the size of the portfolio by a factor a will linearly scale its risk measure F by the same factor.

3. Translation invariance: $F(X + k) = F(X) - k$

   That means adding riskless assets or cash k to the portfolio will lower the risk measure F by k. This property justifies the practice of imposing regulatory capital buffers and reserves at banks.

4. Sub-additivity: $F(X_1 + X_2) \leq F(X_1) + F(X_2)$

   This last property reflects the intuition of risk diversification, such that the risk of a portfolio of securities should always be less than (or equal to) the sum of risks of its components, i.e. adding sub-portfolios together does not increase risks, and splitting portfolios does not decrease risks. If the latter was true, then a bank could simply book deals into separate portfolios and total risks will become lower; a clear fallacy.

Coherence is often discussed in the academia because it is found that VaR is not sub-additive. This is rarely an issue in practice. Firstly, sub-additivity is rarely violated on actual bank portfolios, and even when it is violated, on a large portfolio its effect may not be obvious (or material) enough for a risk manager to take notice. Violation of sub-additivity typically happens when the portfolio contains concentrated long-short positions, which are hedged, and the market goes into a stressful period. Under such a situation, a risk manager may get a nonsensical VaR decomposition.

In 1999, Artzner et al. (1999) [9] proposed an alternative measure called expected shortfall, which satisfies all the criteria of coherence. This is sometimes called conditional VaR (cVaR), expected tail loss (ETL), or Tail Average (TA). It is defined as the expectation of the loss at the tail of the distribution beyond the VaR quantile:

$$ES = \mathbb{E}[-X|X \leq -VaR] \tag{80}$$

Practically, this is computed as the average of all the loss points in the sample distribution which is larger than VaR, and can thus be easily incorporated into existing VaR architectures especially if a bank uses historical simulated VaR. For example, the ES at 97.5% confidence using a 250-day observation period is given by the average of the 6 (rounded from 6.25) largest losses in the left tail. The ES in some sense take into account the “shape” of the loss tail. In contrast VaR is oblivious to the tail beyond the quantile level. Thus, the latest BCBS consultative paper “fundamental review of the trading book” (2013) calls for the adoption of expected shortfall as a replacement for VaR. In particular, the 99% VaR will be replaced by 97.5% ES at a 10-day risk horizon.

The expected shortfall is a special case of a spectral risk measure (SRM) introduce by Acerbi in 2002 [10]. The return sample is first ranked; then each observation (each quantile $L(u)$) is multiplied by a weighting function $w(u)$:

$$SRM = \int_0^1 w(u)L(u)du \tag{81}$$

To qualify as a SRM, $w(u)$ must be function that strictly increases over $[0, 1]$ and integrates (sums up) to one. The expected shortfall (of confidence level $(1-\alpha)$) is a special case of a SRM, where the weights are equal beyond a certain quantile α, and beneath α, the weights are zero, ie it is the average loss which occurs for all the losses that are above VaR. Every reasonable weighting function can be approximated by a step function, and a spectral risk measure based on a step function is simply a weighted average of the cVaRs at the jump points, so spectral risk measures don't add much over an beyond what can be achieved with - intuitively easier to grasp - cVaR measures.

B03. Advanced VaR Models - Multivariate

Risk factors are usually co-dependent amongst themselves, and the dependence structure can strongly influence the total risk at the portfolio level due to diversification effects or the lack thereof. In previous sections, we have often assumed a Gaussian model. This assumption plays a very important role in financial risk modeling, mostly because of its simplicity, computational efficiency and because practitioners think that it is "good enough", especially given the paucity of the underlying data that often does not allow to fit more complex models adequately. In a Gaussian model, the co-dependence structure can be completely described by the covariance matrix (or correlation matrix plus variances, which is the same). This makes the risk aggregation of a portfolio simple and straightforward.

However, the financial crisis of 2008 was yet another example of the shortcomings of a Gaussian model in a crisis environment. Firstly, it fails to fit to the empirical distributions of risk factors, notably their fat tails and skewness. Secondly a single covariance matrix is insufficient to describe the fine co-dependence structure among risk factors. For example, it cannot capture things like non-linear dependencies, or tail-correlations.

To overcome such issues, a more advanced approach to portfolio risk analysis is required. A well-known method is called the ‘copula function’, which is a useful extension and generalization of approaches for modeling joint distribution of risk factors.

A word of warning here: already determining the correct covariance matrix is impossible for high-dimensional data (the amount of data available scales linearly with the number of dimensions whilst the number of coefficients in a covariance matrix scales quadratically). We are introducing additional degrees of freedom here so this problem becomes worse, and choice of a good (implicit) prior is essential for obtaining reasonable results.

Joint Distribution of Risk Factors

First we need to introduce some basic concepts related to joint distribution of random variables. Let’s consider a d-dimensional vector of variables $(f_1, f_2, \cdots, f_d)$, e.g. returns of risk factors. Their randomness (hence risks) is characterized by a joint distribution. For example, the joint distribution for a pair of continuous random variables (X, Y) can be expressed by a 2-D cumulative density function

$$F_{X,Y}(x,y) = \mathbb{P}[X \le x, Y \le y] \qquad \text{and} \qquad F_{X,Y}(x,y) \in [0,1] \tag{82}$$

We define a probability density function $f_{X,Y}(x,y)$ for the distribution, which by definition, must satisfy (i.e. probabilities sum up to one)

$$\int_x \int_y f_{X,Y}(x,y)dy\,dx = 1 \qquad \text{where} \qquad f_{X,Y}(x,y) \ge 0 \tag{83}$$

In (82) and (83) we have denoted the random variables by capital letters and their values by lower case letters. By definition the joint CDF can be obtained from the joint PDF via integration

$$F_{X,Y}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{X,Y}(x,y)dy\,dx \tag{84}$$

Marginal and Conditional Densities

If we consider the probability density function $f_X(x)$ of random variable X alone, it is called *marginal* density function, and it gives the probability of variable X without referencing to (or independent of) the values of the other variable Y. The joint density is related to the marginal density through the equation

$$\begin{aligned} f_{X,Y}(x,y) &= f_{Y|X}(y|x)f_X(x) \\ &= f_{X|Y}(x|y)f_Y(y) \end{aligned} \tag{85}$$

where $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ are the *conditional* probability density functions that define the distribution of a random variable *given the occurrence* of the other random variable. In a special case that X is independent of Y, the $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$ both reduce to their respective margins and give

$$f_{X,Y}(x,y) = f_Y(y)f_X(x) \tag{86}$$

From (85) we see that the joint distribution among random variables can be described by their conditional PDF along with the marginal PDF.

Sklar's theorem

Sklar's theorem provides a theoretical foundation for the application of copulas in describing the dependence structure of random variables. It states that every continuous multivariate distribution $F(x_1,\cdots,x_n)$ of a random vector $(X_1,\cdots,X_n)$ can be decomposed into marginal distributions $F_i(x) = \mathbb{P}[X_i \le x]$ linked by a unique function C called 'copula', such that

$$F(x_1,\cdots,x_n) = C\big(F_1(x_1),\cdots,F_n(x_n)\big) \tag{87}$$

or equivalently

$$F\big(F_1^{-1}(u_1),\cdots,F_n^{-1}(u_n)\big) = C(u_1,\cdots,u_n) \tag{88}$$

where $x_i \in \mathbb{R}$, $u_i = F_i^{-1}(x_i) \in [0,1]$ and F_i^{-1} is the inverse function of the marginal CDF F_i.

In other words, an m-copula is an m-dimensional distribution function where all m univariate margins are uniformly distributed on [0, 1]. In practice, the data for marginal distributions of random variables is far easier to obtain than their joint distribution. Rather than only considering explicit multivariate distributions, it is wise to model the joint distribution by the margins along with a suitable copula function measuring their dependence structure. A prominent feature of the copula based approach is that it models the dependence structure independent of the margins. This is especially useful when the variables are not Gaussian (as is often the case in financial world). We may choose appropriate univariate distributions (e.g. fat-tailed Student-*t* distributions) for the margins and then link them by a copula function most suitable for the implied dependence structure.

Copula functions

Given the definition above, we are going to introduce a few copula functions. In fact, any multivariate distribution CDF can serve as a copula function. However, copula functions in high dimension are generally non-trivial and difficult to visualize. For illustrative purpose, we now consider a bivariate copula function, which is defined as $C(u,v) \in [0,1]\ \forall\, u,v \in [0,1]$. The copula function, by definition, must satisfy the following three properties:

1. The copula must be zero if one of the arguments is zero

$$C(u,0) = C(0,v) = 0 \tag{89}$$

2. The copula is equal to u if one argument is u and the other is 1

$$C(1,v) = v \qquad \text{and} \qquad C(u,1) = u \tag{90}$$

3. The copula is m-increasing, i.e. when $u_1 < u_2$ and $v_1 < v_2$

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0 \tag{91}$$

The simplest example of copula function is the *product copula*. For the bivariate case, it is written as

$$C_{Prod}(u,v) = uv \tag{92}$$

It is easy to check that the product copula satisfies the three properties mentioned above. It describes the case where there is no co-dependence amongst the random variables, and therefore is the generalization of the "zero correlation" situation into non-Gaussian marginal variables.

Another example is *Gaussian copula* that has been widely used in financial applications. It is defined as

$$C_{Gauss}(u_1,\cdots,u_n;\rho) = \Phi(\phi^{-1}(u_1),\cdots,\phi^{-1}(u_n);\rho) \tag{93}$$

where $\phi^{-1}(u_i)$ is the inverse of the univariate standard Gaussian CDF and $\Phi(x_1,\cdots,x_n;\rho)$ is the joint standard Gaussian CDF with an $n \times n$ correlation matrix ρ . The term ‘standard’ here means that: for a univariate normal distribution it has a mean of zero and a variance of one, and for a joint normal distribution it has a mean of zero and a covariance of ρ (i.e. all marginal normals have a unit variance of 1). The standard joint Gaussian CDF reads

$$\begin{aligned}&\Phi(x_1,\cdots,x_n;\rho)\\&= \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\frac{1}{\sqrt{(2\pi)^n|\rho|}}\exp\left(-\frac{1}{2}z^T\rho^{-1}z\right)dz_n\cdots dz_1\end{aligned} \tag{94}$$

in which, we write variable $z = (z_1,\cdots,z_n)^T$ as a column vector and $|\rho|$ is the determinant of the correlation matrix. Since the copula function behaves like a CDF, we can differentiate (93) to derive an equivalent PDF for the Gaussian copula

$$c_{Gauss}(u_1,\cdots,u_n;\rho) = \frac{1}{\sqrt{|\rho|}}\exp\left[-\frac{1}{2}x^T(\rho-I)^{-1}x\right] \tag{95}$$

where I is an $n \times n$ identity matrix and x is a column vector given by $x = \left(\phi^{-1}(u_1),\cdots,\phi^{-1}(u_n)\right)^T$. In the bivariate case, the Gaussian copula reduces to

$$C_{Gauss}(u,v;\rho) = \int_{-\infty}^{\Phi^{-1}(u)}\int_{-\infty}^{\Phi^{-1}(v)}\varphi_{X,Y}(x,y)dy\,dx \tag{96}$$

where $\varphi_{X,Y}(x,y)$ denotes the PDF of a bivariate standard Gaussian. If X and Y are independent, we can split the double integral and the Gaussian copula in (96) further reduces to a product copula, such that

$$C_{Gauss}(u,v;I) = \int_{-\infty}^{\Phi^{-1}(u)}\varphi(x)dx\int_{-\infty}^{\Phi^{-1}(v)}\varphi(y)dy = uv \tag{97}$$

where I is an 2×2 identity matrix (implying zero correlation between X and Y) and $\varphi(x)$ is the univariate standard Gaussian PDF.

In financial markets, the occurrence of concurrent bad events is not rare. A good copula function must be able to model such extreme co-movements regardless of the shape of margins. Embrechts et al. [11] introduced the coefficient of tail dependence to measure the probability of joint extreme events among random variables. They have shown that Gaussian copula has zero tail dependence regardless of high correlation we choose. This says Gaussian copula is not suitable for modelling joint distribution with strong tail dependence. A better choice is to use the Student t copula. It differs from Gaussian copula by assuming t distributions for the joint and marginal densities, which has the form

$$C_t(u_1,\cdots,u_n;\nu,\rho) = t(t_\nu^{-1}(u_1),\cdots,t_\nu^{-1}(u_n);\nu,\rho) \tag{98}$$

where $t_\nu^{-1}(u)$ is the inverse of a univariate standard student t CDF with degree of freedom ν, and $t(x_1,\cdots,x_n;\nu,\rho)$ is a multivariate standard student t CDF with mean of zero, correlation matrix of ρ and degree of freedom ν (note that although we still call it a correlation matrix, the ρ in a joint student t distribution has a different meaning from that in a joint normal distribution)

$$\begin{aligned} &t(x_1,\cdots,x_n;\nu,\rho) \\ &= \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\frac{1}{\sqrt{(\nu\pi)^n|\rho|}}\frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{z^T\rho^{-1}z}{\nu}\right)^{-\frac{\nu+n}{2}}dz_n\cdots dz_1 \end{aligned} \tag{99}$$

where $\Gamma(\cdot)$ denotes the well-known Gamma function. Similarly differentiating (98), we are able to derive a PDF for the Student t copula

$$\begin{aligned} &c_t(u_1,\cdots,u_n;\nu,\rho) \\ &= \frac{1}{\sqrt{|\rho|}}\frac{\Gamma\left(\frac{\nu+n}{2}\right)\Gamma\left(\frac{\nu}{2}\right)^{n-1}}{\Gamma\left(\frac{\nu+1}{2}\right)^n}\frac{\left(1+\frac{z^T\rho^{-1}z}{\nu}\right)^{-\frac{\nu+n}{2}}}{\prod_{i=1}^n\left(1+\frac{z_i^2}{\nu}\right)^{-\frac{\nu+1}{2}}} \end{aligned} \tag{100}$$

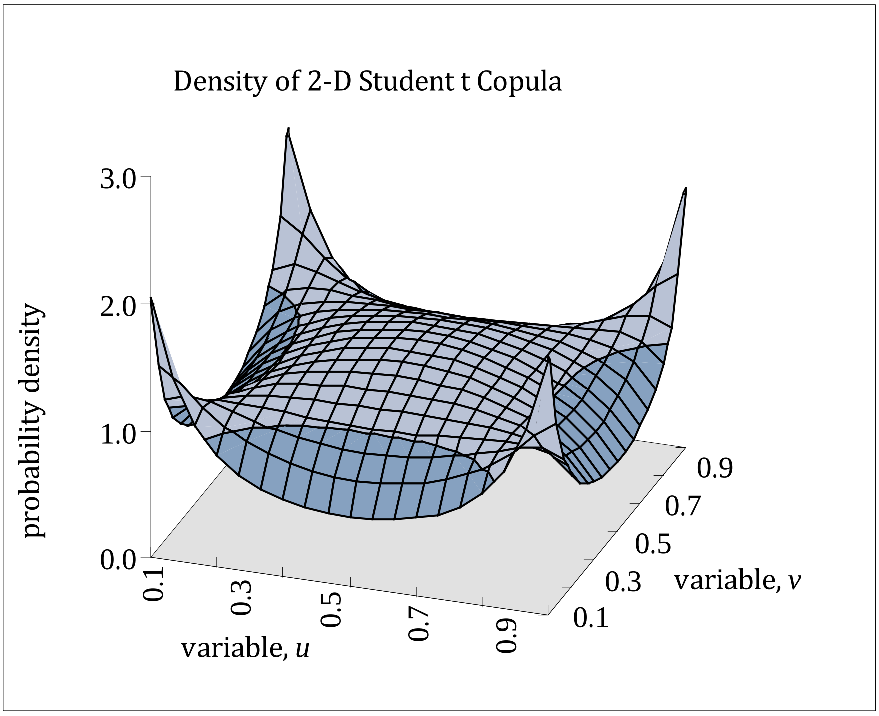


Figure 14. Density of 2-D Student *t* Copula with correlation of 0 and degree of freedom of 2.0

To compare the two copula functions, we can look at their density functions with correlation set to 0. The density of Gaussian copula (95) in this case becomes uniform. The tail dependence is now determined only by the tail density of margins, which can vanish quickly. In contrast, the density of Student *t* Copula (100), shown in Figure 14, has increasingly large density in tails, which elevates the overall tail dependence.

Two points are noteworthy. Firstly, the Student *t* copula converges to Gaussian copula as the degree of freedom v goes to infinity. When the degree of freedom is properly chosen (usually calibrated from historical data), the Student *t* copula gives a better fit to empirical joint distributions than the Gaussian copula. Secondly, in our derivation, we have assumed that all the associated margins have the same degree of freedom. Although this simplification may limit the flexibility of the Student *t* copula in modelling the tail dependence, it greatly reduces the complexity of calibration and simplifies its application in VaR estimation.

VaR Estimation using Copula

The dependence structure of risk factors plays a crucial role in determining the VaR of portfolios. In practice, the joint distribution of risk factors can be far from a multivariate normal, a single correlation matrix estimated from historical data may not be sufficient to capture the full dependence structure in risk factors. Sklar's theorem suggests that we can use a unique copula function along with a set of marginal density functions to describe the joint distribution. In this section, we are going to use an example referring to spreadsheet [VaR Copula.xls] to demonstrate the application of copula functions in VaR estimation.

Suppose Gaussian copula is used to describe the dependence structure. Then, based on (87) and (93), the true joint distribution F can be approximated by

$$\hat{F}(x_1,\cdots,x_n) = \Phi\big(\phi^{-1}\big(F_1(x_1)\big),\cdots,\phi^{-1}\big(F_n(x_n)\big);\rho\big) \tag{101}$$

where we use a 'hat' on F to denote an estimation. The first step is to calibrate the copula model (101). Ideally, we can assume parametric marginals for $F_i(x)$ and then use an estimation method such as MLE to calibrate the correlation matrix ρ and the parameters in $F_i(x)$ simultaneously to historical returns of risk factors. However the calibration process may be overwhelmed by the large number of parameters. Instead we can use empirical cumulative density function $\hat{F}_i(x) \in [0,1)$ defined by

$$\hat{F}_i(x) = \frac{\text{number of elements in the sample} \leq x}{\text{total number of elements in the sample} + 1} \tag{102}$$

as estimations of the marginals $F_i(x)$ and calibrate the correlation matrix ρ only. It can be proven that the MLE of ρ for the Gaussian copula is just the correlation matrix estimated from historical returns of risk factors after mapping to standard normal random space. For instance, in our previous example, we have daily returns $r_{i,t}$ for risk factor $i = 1,\cdots,4$ and historical scenario $t = 1,\cdots,500$. For the i-th risk factor, we have a vector of 500 returns that defines an empirical distribution $\hat{F}_i(\cdot)$. The empirical distribution is used to transform the 500 returns to a vector of decimal numbers $u_{i,t} = \hat{F}_i\big(r_{i,t}\big)$ valued between 0 and 1. The transformation is similar to the *Rosenblatt* transformation mentioned earlier. However, the difference is that in this case the value being transformed belongs to the sample that forms the empirical distribution (i.e. we are performing in-the-sample transformation). After the transformation, the series of $u_{i,t}$ is then further mapped to standard Gaussian random space by the inverse of the standard Gaussian CDF $x_{i,t} = \phi^{-1}\big(u_{i,t}\big)$. Repeating the transformation for all the risk factors, we can then use the resulted data (e.g. four vectors of $x_{i,t}$ for $i = 1,\cdots,4$) to estimate the correlation matrix ρ for the Gaussian copula.

With the Gaussian copula calibrated, we can estimate the VaR using Monte Carlo simulation. The idea is to draw samples from the calibrated copula function, which are then inversed by the empirical CDF of the corresponding marginal to generate simulated risk factor returns. The detailed procedure is as follows:

1. Sample from a multivariate standard Gaussian distribution $\Phi(0,\rho)$ to obtain $z = (z_1,\cdots,z_4)^T$. Again, the *Cholesky* decomposition of the correlation matrix ρ is used here to correlate the independent standard normal samples.
2. Apply standard Gaussian CDF ϕ to each component of z to obtain $v = (v_1,\cdots,v_4)^T = \big(\phi(z_1),\cdots,\phi(z_n)\big)^T$ such that each entry of v is valued between 0 and 1. Basically the sample v follows a distribution given by the Gaussian copula $C_{Gauss}(v_1,\cdots,v_n;\rho)$.

3. Apply respective inverse function of the marginal empirical cumulative density to each component of v to obtain $\hat{r} = \left(\hat{F}_1^{-1}(v_1), \cdots, \hat{F}_1^{-1}(v_n)\right)^T$. The $\hat{r}$ is a vector of simulated risk factor returns.
4. Calculate portfolio P&L using the simulated $\hat{r}$ by full revaluation. This is the P&L resulted from *one* simulated scenario.

We repeat the above simulation many times to derive a P&L vector that characterizes the P&L distribution, from which we can estimate the portfolio VaR.

The Student t copula can be applied in VaR estimation in a similar manner. An example of Student t copula in VaR estimation is also included in the spreadsheet [VaR Copula.xls]. Due to additional parameters in the model, it requires more sophisticated calibration and sampling procedures, which are outside the scope of this book. The interested readers can refer to Trivedi and Zimmer (2007) [12] for more information.

References

1. J.P. Morgan/ Reuters (1996), “Riskmetrics technical document, 4th ed".

2. Engle, R. (1982), “Autorregressive Conditional Heteroskedasticity with Estimates of United Kingdom Inflation”, *Econometrica*, 50, 987-1008.

3. Wong, M. (2013), “Bubble Value at Risk: A countercyclical risk management approach”, Wiley Finance, Singapore.

4. Sun, H.; Nelken, I.; Han, G.; Guo, J. (2009), “Error of VaR by overlapping intervals,” *Risk Magazine*, 4/2009, 50-55.

5. Riskmetrics (2001), “Return to Riskmetrics: The evolution of a standard”.

6. Kupiec, P. H. (1995), “Techniques for verifying the accuracy of risk management models”, *Journal of Derivatives*, 3, 73–84.

7. Christoffersen, P. F. and Pelletier, D. (2004), “Backtesting value-at-risk: a duration-based approach.” *Journal of Financial Econometrics* 2, 84–108.

8. Wehn, C.S. (2008), “Looking forward to back testing”, *Risk*, 21, 5, 90-95

9. Artzner, P., F. Delbaen, J. Eber and D. Heath (1999), “Coherent Measures of Risk.” *Mathematical Finance* 9 (3): 203_228.

10. Acerbi, C. (2002), “Spectral Measures of Risk: A Coherent Representation of Subjective Risk Aversion.” *Journal of Banking and Finance* 26 (7): 1505_1518.

11. Embrechts P.; McNeil, A. and Straumann, D. (2002), “Correlation and dependence in risk management: properties and pitfalls”, In *Risk management: value at risk and beyond*, edited by Dempster M., published by Cambridge University Press, Cambridge.

12. Trivedi, P. K. and Zimmer, D. M. (2007), "Copula Modeling: An Introduction for Practitioners," *Foundations and Trends in Econometrics*, now publishers, 2007