Chapter 20
Financial Analytics

Krishnamurthy Vaidyanathan

1 Part A: Methodology

1.1 Introduction

Data analytics in finance is a part of quantitative finance. Quantitative finance primarily consists of three sectors in finance—asset management, banking, and insurance. Across these three sectors, there are four tightly connected functions in which quantitative finance is used—valuation, risk management, portfolio management, and performance analysis. Data analytics in finance supports these four sequential building blocks of quantitative finance, especially the first three—valuation, risk management, and portfolio management.

Quantitative finance can be dichotomized into two branches or subsets having mild overlaps. The first is the risk-neutral world or the Q-world, and the second, wherein data analytics is used extensively, is the risk-averse world or the P-world. Quant professionals in the Q-world are called the Q-quants, and those of the P-world are called P-quants. Before we delve into the methodology of data analysis in finance which we structure as a three-stage process in this chapter, we briefly highlight the processes, methodologies, challenges, and goals of these two quant-worlds and also look at the history and origins of these two dichotomized worlds of quantitative finance.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-319-68837-4_20) contains supplementary material, which is available to authorized users.

K. Vaidyanathan (✉)
Indian School of Business, Hyderabad, Telangana, India
e-mail: vaidya_nathan@isb.edu

B. Pochiraju, S. Seshadri (eds.), *Essentials of Business Analytics*, International
Series in Operations Research & Management Science 264,
https://doi.org/10.1007/978-3-319-68837-4_20

1.2 Dichotomized World of Quant Finance

One can paraphrase Rudyard Kipling's poem *The Ballad of East and West* and say, "Oh, the Q-world is the Q-world, the P-world is the P-world, and never the twain shall meet." Truth be told, Kipling's lofty truism is not quite true in the Quant world. The Q-world and P-world do meet, but they barely talk to each other. In this section, we introduce the P- and Q-worlds and their respective theoretical edifices.

1.2.1 Q-Quants

In the Q-world, the objective is primarily to determine a fair price for a financial instrument, especially a derivative security, in terms of its underlying securities. The price of these underlying securities is determined by the market forces of demand and supply. The demand and supply forces come from a variety of sources in the financial markets, but they primarily originate from buy-side and sell-side financial institutions. The buy-side institutions are asset management companies—large mutual funds, pensions funds, and investment managers such as PIMCO who manage other people's money—both retail and corporate entities' money. The sell-side institutions are market makers who make money on the margin they earn by undertaking market making. That is, they are available to buy (bid) a financial instrument for another market participant who wants to sell and make available to sell (offer) a financial instrument for somebody wanting to buy. They provide this service for a commission called as bid–offer spread, and that is how they primarily make their money from market making. The trading desks of large investment banks such as Goldman Sachs, JPMorgan, Citibank, and Morgan Stanley comprise the sell-side. The Q-quants primarily work in the sell-side and are price-makers as opposed to P-quants who work in the buy-side and are typically price-takers.

The Q-quants borrow much of their models from physics starting with the legendary Brownian motion. The Brownian motion is one of the most iconic and influential ingress from physics into math finance and is used extensively for pricing and risk management. The origins of the Q-world can be traced to the influential work published by Merton in 1969 (Merton 1969)—who used the Brownian motion process as a starting point to model asset prices. Later in 1973, Black and Scholes used the geometric Brownian motion (GBM) to price options in another significant work for which they eventually won the Nobel Prize in 1997 (Black and Scholes 1973). This work by Black and Scholes gave a fillip to pricing of derivatives as these financial instruments could be modeled irrespective of the return expectation of the underlying asset. In simple terms, what that meant was that even if I think that the price of a security would fall while you may think that the price of that security would increase, that is, the return expectations are different, yet we can agree on the price of a derivative instrument on that security. Another important edifice in the Q-space is the fundamental theorem of asset pricing by Harrison and Pliska (1981). This theorem posits that the current price of a security is fair only if there exists a stochastic process such as a GBM with constant expected value for all future points in time. Any process that satisfies this property is called a martingale. Because the expected return is the same for all financial instruments, it implies that there is no extra reward for risk taking. It is as if, all the pricing is done in a make-believe world called the risk-neutral world, where irrespective of the risk of a security, there is no extra compensation for risk. All financial instruments in this make-believe world earn the same return regardless of their risk. The risk-free instruments earn the risk-free rate as do all risky instruments. In contrast, in the P-world, the economic agents or investors are, for all intents and purposes, risk-averse, as most people are in the real world.

The Q-quants typically have deep knowledge about a specific product. So a Q-quant who, for instance, trades credit derivatives for a living would have abundant knowledge about credit derivative products, but her know-how may not be very useful in, say, a domain like foreign exchange. Similarly, a Q-quant who does modeling of foreign exchange instruments may not find her skillset very useful if she were to try modeling interest rates for fixed income instruments. Most of the finance that is done in the Q-world is in continuous time because as discussed earlier, the expectation of the price at any future point of time is equal to its current price. Given that this holds for all times, the processes used in the Q-world are naturally set in continuous time. In contrast, in the P-world, the probabilities are for a risk-averse investor. This is because in the real world, people like you and me need extra return for risk-taking. Moreover we measure returns over discrete time intervals such as a day, a week, a month, or a year. So most processes are modeled in discrete time. The dimensionality of the problem in the P-world is evidently large because the P-quant is not looking at a specific instrument or just one particular asset class but multiple asset classes simultaneously. The tools that are used to model in the P-world are primarily multivariate statistics which is what concerns data analysts.

1.2.2 P-Quants

We now discuss the P-world and their origins, tools, and techniques and contrast them with the Q-world. The P-world started with the mean–variance framework by Markowitz in 1952 (Markowitz 1952). Harry Markowitz showed that the conventional investment evaluation criteria of net present value (NPV) needs to be explicitly segregated in terms of risk and return. He defined risk as standard deviation of return distribution. He argued that imperfect correlation of return distribution of stocks can be used to reduce the risk of a portfolio of stocks. He introduced the concept of diversification which is the finance equivalent of the watchword—do not put all your eggs in one basket. Building on the Markowitz model, the next significant edifice of the P-world was the capital asset pricing model (CAPM) by William Sharpe in 1964 (Sharpe 1964). William Sharpe converted the Markowitz "business school" framework to an "economics department" model. Sharpe started with a make-believe world where all investors operate in the Markowitz framework. Moreover, all investors in the CAPM world have the same expectation of returns and variance–covariance. Since all the risky assets in the financial market must be held by somebody, it turns out that in Sharpe's "economics department" model, all investors end up holding the market portfolio—a portfolio where each risky security has the weight proportional to how many of them are available. At the time when William Sharpe postulated the model, the notion of market portfolio was new. Shortly thereafter, the financial industry created mutual funds which hold a diversified portfolio of stocks, mostly in the proportion of stocks that are available. It was as if nature had imitated art, though we all know it is almost always the other way around. In 1990, Markowitz and Sharpe won the Noble Prize in Economics for developing the theoretical basis for diversification and CAPM.

The next significant edifice came in 1970 by Eugene Fama called the efficient market hypothesis (EMH). The EMH hypothesizes that no trading strategy based on already available information can generate super-normal returns (Fama 1970). The EMH offered powerful theoretical insights into the nature of financial markets. More importantly, it lent itself to empirical investigation which was imperative and essential for finance—then a relatively nascent field. As a result, the efficient market hypothesis is probably the most widely and expansively tested hypothesis in all social sciences. Eugene Fama won the Nobel Prize in 2013 for his powerful insights on financial markets.

Another important contribution in the mid-1970s was the arbitrage pricing theory (APT) model by Stephen Ross (1976). The APT is a multifactor model used to calculate the expected return of a financial asset. Though both CAPM and APT provided a foundational framework for asset pricing, they did not use data analytics because their framework assumed that the probability distribution in the P-world is known. For instance, if financial asset returns follow an elliptical distribution, every investor would choose to hold a portfolio with the lowest possible variance for her chosen level of expected returns. Such a portfolio is called a minimum variance portfolio, and the framework of portfolio analysis is called the mean–variance framework. From a finance theory viewpoint, this is a convenient assumption to make, especially the assumption that asset returns are jointly normal, which is a special case of an elliptical distribution. Once the assumption that the probability distribution is already known is made, theoretical implications can be derived on how the asset markets should function. Tobin proposed the separation theorem which postulates that the optimal choice of investment for an investor is independent of her wealth (Tobin 1958). Tobin's separation theorem holds good if the returns of the financial assets are multinormal. Extending Tobin's work, Stephen Ross postulated the two-fund theorem which states that if investors can borrow and lend at the risk-free rate, they will possess either the risk-free portfolio or the market portfolio (Ross 1978). Ross later generalized it to a more comprehensive *k*-fund separation theorem. The Ross separation theorem holds good if the financial asset returns follow any elliptical distribution. The class of elliptical distribution includes the multivariate normal, the multivariate exponential, the multivariate Student *t*-distribution, and multivariate Cauchy distribution, among others (Owen and Rabinovitch 1983).

However, in reality, the probability distribution needs to be estimated from the available financial information. So a very large component of this so-called information set, that is, the prices and other financial variables, is observed at discrete time intervals, forming a time series. Analyzing this information set requires manifestly sophisticated multivariate statistics of a certain spin used in economics called as econometrics, wherein most of the data analytics tools come into play. In contrast to this, in the Q-world, quants mostly look at the pricing of a specific derivative instrument and get the arbitrage-free price of the derivative instrument based on the underlying fundamental instrument and other sets of instruments. However, in the P-world, we try to estimate the joint distribution of all the securities that are there in a portfolio unlike in the Q-world wherein we are typically concerned with just one security. The dimensionality in the Q world is small, but in the P-world it is usually a lot larger. So a number of dimension reduction techniques, mostly linear factor models like principal component analysis (PCA) and factor analysis, have a central role to play in the P-world. Such techniques achieve parsimony by reducing the dimensionality of the data, which is a recurring objective in most data analytic applications in finance. Since some of these data analytic techniques can be as quantitatively intense or perhaps more intense than the financial engineering techniques used in the Q-world, there is now a new breed of quants called the P-quants who are trained in the data analytic methodologies. Prior to the financial crisis of 2008, the Q-world attracted a lot of quants in finance. Many having PhDs in physics and math worked on derivative pricing. Till the first decade of the twenty-first century culminating in the financial crisis, all the way back from the 1980s when the derivatives market started to explode, quantitative finance was identified with the Q-quants. But in recent years, especially post-crisis, the financial industry has witnessed a surge of interest and attention in the P-world, and there is a decrease of interest in the Q-world. This is primarily because the derivatives markets have shrunk. The second-generation and third-generation types of exotic derivatives that were sold pre-crisis have all but disappeared. In the last decade, there has been a severe reduction in both the volume and complexity of derivatives traded. Another reason why the P-quants have a dominant role in finance is that their skills are extremely valuable in risk management, portfolio management, and actuarial valuations, while the Q-quants work mostly on valuation. Additionally, a newfound interest in data analytics, in general and mining of big data specifically, has been a major driver for the surge of interest in the P-world.

1.3 Methodology of Data Analysis in Finance: Three-Stage Process

The methodology of data analysis in the P-space in this chapter is structured as a three-part process—asset price estimation, risk management, and portfolio analysis. The first part, asset price estimation, is split into five sequential steps; the second part, risk management, into three steps; and the third part, portfolio analysis, into two steps. The methodology of the three-stage framework for data analysis in finance is shown in Fig. 20.1. The first among the estimation steps is the process of

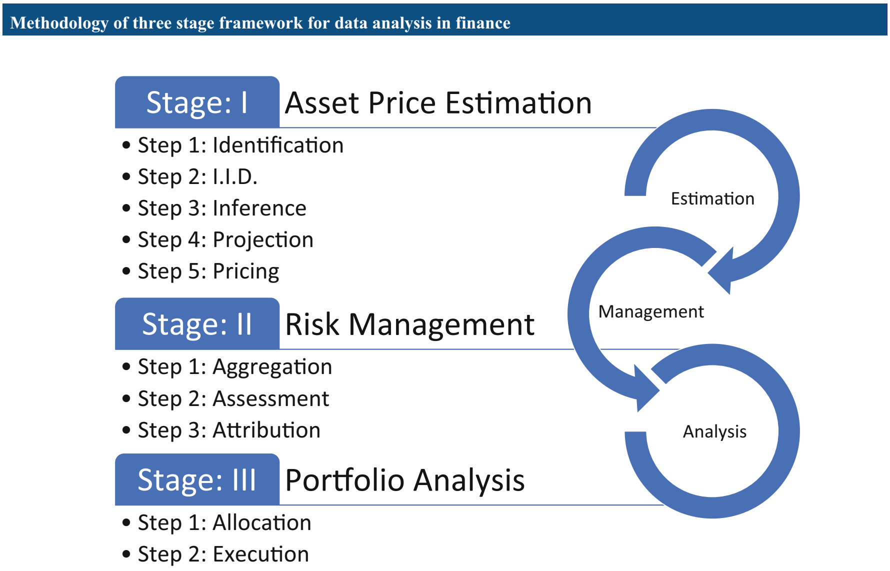

Fig. 20.1 Methodology of the three-stage framework for data analysis in finance

variable identification. Variable identification is an important first step because the variable to be modeled is different for different asset classes as shown in Fig. 20.1. The second step is to transform the identified variable into an independent and identically distributed (i.i.d.) process. The third step is to infer the joint distribution of these i.i.d. processes of multiple financial variables. The fourth step is to forecast the variable using past information, and the final fifth step in asset price estimation is to derive the forecasted price from the i.i.d. variables' joint distribution.

The next two stages in the modeling process constitute analysis for risk and portfolio management. The first three steps within the second part of analysis pertain to risk management, and the remaining two steps apply to portfolio management. Within risk management, we aggregate the entire portfolio (Stage II Step 1, Sect. 1.3.8) and then evaluate the portfolio (Stage II Step 2, Sect. 1.3.9) in terms of various risk metrics such as Value at Risk (VaR) and threshold persistence. Then, the return and risk of the portfolio is attributed to several risk factors using various decomposition techniques (Stage II Step 3, Sect. 1.3.10).

The final two steps constitute portfolio management where we will look at the design of the optimal portfolio or, from a firm-wide context, design of an optimal policy to allocate the portfolio of assets of a firm (Stage III Step 1, Sect. 1.3.12). In the next step, we execute and allocate assets according to the optimal portfolio benchmark determined in the previous step (Stage III Step 1, Sect. 1.3.12). This is done these days mostly programmatically using algorithms. Each of the steps are explained using simple stand-alone examples. Suitable references are provided at the appropriate steps. A comprehensive application of the framework that combines multiple steps is provided in Sect. 20.2. We now discuss the three-stage methodology starting with variable identification for different asset classes such as equities, foreign exchange, fixed income, credit, and commodities.

1.3.1 Stage I: Asset Price Estimation

The objective of the first stage is to estimate the price behavior of an asset. It starts with identification of the financial variable to model.

Step 1: Identification

The first step of modeling in the P-world is to identify the appropriate variable which is different for distinct asset classes. The basic idea is to find a process for the financial variable where the residuals are essentially i.i.d. The most common process used for modeling a financial variable x is the random walk:

$$x_t = x_{t-1} + \varepsilon_t$$

where ε_t is the error term and is random. The postulation where a financial variable follows a random walk is called as random walk hypothesis and is consistent with efficient market hypothesis (Fama 1965). What it means in simple terms is that financial variables are fundamentally unpredictable. However, if one looks at any typical stock price, the price changes in such a way that the order of magnitude of the change is proportional to the value of the stock price. This kind of behavior conflicts with homogeneity across time that characterizes a financial variable when it follows a random walk. As a way out, the variable that is actually modeled is the logarithm (log) of the stock price, and it has been observed that the log of the stock price behaves as a random walk. A simple random walk has a constant mean and standard deviation, and its probability distribution does not change with time. Such a process is called as a stationary process.

Similar to stock prices, the log of foreign exchange rates or the log of commodity prices behaves approximately as a random walk. The underlying variable itself, that is, the stock price, currency rate, or commodity price, is not exactly stationary, but the log of stock price, currency rate, or commodity price conducts itself just about as a haphazard random walk. A stationary process is one whose probability distribution does not change with time. As a result, its moments such as variance or mean are not time-varying.

However, choosing the right financial variable is as important as the modification made to it. For example, in a fixed income instrument such as a zero-coupon bond, the price converges to the face value as the bond approaches maturity. Clearly, neither the price itself nor its log can be modeled as a random walk. Instead, what is modeled as a random walk is the yield on bonds, called yield to maturity. Simply put, yield is the internal rate of return on a bond that is calculated on the cash flows

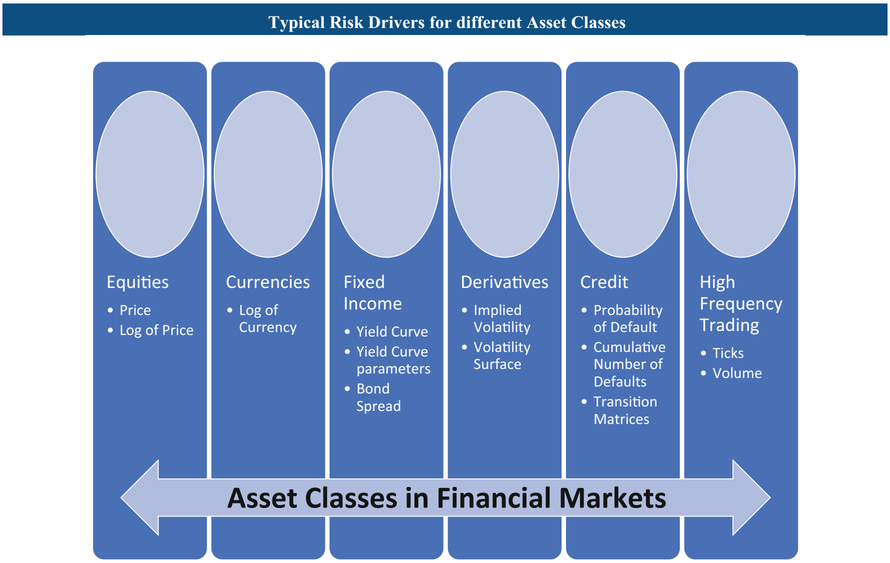

Fig. 20.2 Typical risk drivers for different asset classes

the bond pays till maturity. And this variable fits a random walk model adequately. The financial variables that are typically modeled in the different asset classes are shown in Fig. 20.2.

Step 2: I.I.D.

Once the financial variable that is of interest is identified, the next step in data preparation is to obtain a time series of the variables that are of interest. These variables should display a homogenous behavior across time, and as shown in Fig. 20.2, the variables are different for different asset classes. For instance, in equities, currencies, or commodities, it is the log of the stock/currency/commodity price. For fixed income instruments, the variable of interest may not be the price or the log of price. The variable of interest would be the yield to maturity of the fixed income security.

Once we have the raw data of the financial variable, we test if the financial variable follows a random walk using statistical tests such as the Dickey–Fuller test or the multiple variance ratio test (Dickey and Fuller 1979; Campbell et al. 1997). A lot of times they may follow less random processes and therefore may be predictable to some degree. This is known as non-random walk hypothesis. Andrew Lo and Craig Mackinlay, at MIT Sloan and Wharton, respectively, in their book *A Non-Random Walk Down Wall Street* present a number of tests and studies that validate that there are trends in financial markets and that the financial variables identified in Step 1 are somewhat predictable. They are predictable both in cross-sectional and time series terms. As an example of the ability to predict using cross-sectional data, the Fama–French three-factor model postulated by Eugene Fama and Kenneth French uses three factors to describe stock returns (Fama and French 1993). An example of predictability in time series is that the financial variable may display some kind of mean reversion tendency. What that means is that if the value of the financial variable is quite high, it will have a propensity to decrease in value and vice versa. For example, if yields become very high, there may be propensity for them to come back to long-run historical average levels. In such cases, the features that cause deviation from the random walk are extracted out so that the residuals display i.i.d. behavior. The models used would depend on the features displayed by the financial variable. For example, volatility clustering like mean reversion is a commonly observed feature in financial variables. When markets are extremely volatile, the financial variable fluctuates a lot, and there may be a higher probability of a large variability than otherwise. Techniques like autoregressive conditional heteroscedasticity model (ARCH) or generalized autoregressive conditional heteroscedasticity model (GARCH) are used to factor out volatility clustering (Engle 1982). If the variable displays some kind of mean reversion, one might want to use autoregressive moving average (ARMA) models if it is a univariate case or use vector autoregression (VAR) models in multivariate scenarios (Box et al. 1994; Sims 1980). These are, in essence, econometric models which can capture linear interdependencies across multiple time series and are fundamentally a general form of autoregressive models (AR) (Yule 1927). We could also use stochastic volatility models, and those are comparatively commonly used in volatility clustering. Long memory processes primarily warrant fractional integration models (Granger and Joyeux 1980). Fractional integration displays a long memory which principally means that the increments of the financial variable display autocorrelation. The increments therefore are not i.i.d., and these autocorrelations persist across multiple lags. For instance, the value of the random variable at time $t + 1$ is a function of time t, $t - 1$, $t - 2$, and so on. The lags decrease very gradually and are therefore called long memory processes. Such trends, be they long memory, volatility clustering, or mean reversion, are modeled using techniques such as fractional integration, GARCH, or AR processes, respectively. After such patterns are accounted for, we are left with i.i.d. shocks with no discernible pattern.

Step 3: Inference

The third step in estimation after the financial variable is identified and after we have gotten to the point of i.i.d. shocks is to infer the joint behavior of i.i.d. shocks. In the estimation process, we typically determine those parameters in the model which gets us to an i.i.d. distribution. We explain the first three steps using data on S&P 500 for the period from October 25, 2012, to October 25, 2017. The first step is to identify the financial variable of interest. We work with returns rather than with

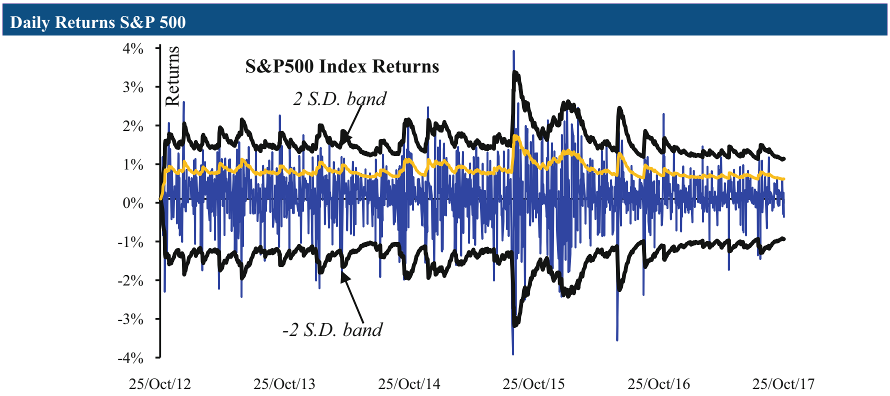

Fig. 20.3 Daily returns of the S&P 500

absolute index levels of S&P 500 for reasons mentioned in Step 1. From the daily index levels, the 1-day returns are calculated as follows:

$$r_t = \log(p_t) - \log(p_{t-1})$$

This return r_t itself is not distributed in an i.i.d. sense. Neither are the daily returns identical nor are they independent. We can infer from a quick look at the graph of the returns data in Fig. 20.3 that the returns are not i.i.d.

One may refer to Chap. 5, on data visualization, for a better understanding on how to interpret the graph. One of the things to observe from the graph is that if the return in a given day was either extremely high or low, it normally followed that return on the subsequent day was also quite high/low. That is, if the return was volatile at time t, the probability of it being more volatile is higher than it being stable at time $t + 1$. So in this case, the data seems to suggest that the financial variable is conditionally heteroscedastic, which means that the standard deviation is neither independent nor identical across time periods. To accommodate for conditional heteroscedasticity, we can use the GARCH(1,1) model (Bollerslev 1986). This model accounts for autocorrelation and heteroscedasticity, that is, for correlations among errors at different time periods t and different variances of errors at different times. The way we model variance ${\sigma_t}^2$ is:

$${\sigma_t}^2 = \omega + \alpha\, {r_{t-1}}^2 + \beta {\sigma_{t-1}}^2$$

We have to estimate the parameters ω, α, β. The estimation technique we use is maximum likelihood estimation (MLE). As the name implies, we maximize the likelihood of estimating the true values of the parameters ω, α, β. If GARCH(1,1) model is the right specification and if the parameters ω, α, β are estimated correctly, then $\frac{r_t}{\sigma_t}$ will be a sequence of random i.i.d. variables. In this case we assume the average daily returns to be zero. Using the Gaussian distribution, the likelihood or the probability of $\frac{r_t}{\sigma_t}$ being normally distributed is given by:

$$\frac{1}{\sqrt{2\pi{\sigma_t}^2}}e^{-\frac{1}{2}\left(\frac{r_t}{\sigma_t}\right)^2}$$

This completes Step 2 of reducing the variable to an i.i.d. process. The next step is to compute the joint distribution. Since the variables $\frac{r_t}{\sigma_t}$ across time are independent, the joint likelihood L of the sample is calculated as the product of the above likelihood function using the property of independence across the time series of n data points. Therefore:

$$L = \prod_{t=1}^{n}\frac{1}{\sqrt{2\pi{\sigma_t}^2}}e^{-\frac{1}{2}\left(\frac{r_t}{\sigma_t}\right)^2}$$

Since the above product would be a very small number in magnitude, the natural log of the above likelihood is maximized in MLE. This log-likelihood is given by:

$$\ln(L) = -\frac{1}{2}\sum_{t=1}^{n}\left\{\ln\left(2\pi{\sigma_t}^2\right) + \left(\frac{r_t}{\sigma_t}\right)^2\right\}$$

The joint log-likelihood is a function of the parameters ω, α, β. The value of the parameters that maximizes this joint likelihood is the correct estimate. The above steps are explained in the spreadsheet “Financial Analytics Steps 1, 2 and 3.xlsx” (available on the book’s website).

Step 4: Projection

The fourth step is projection. We explain this step using a simple example from foreign exchange markets. Let us say that the financial variable is estimated using a technique such as MLE, GMM, or Bayesian estimation (Hansen 1982). The next step is to project the variable using the model. Say the horizon is 1 year, and we want to calculate the expected profit or loss of a certain portfolio. A commonly used technique for this is the Monte Carlo simulation, which is another ingress from physics (Fermi and Richtmyer 1948). We project the financial variable in the Q-space using risk-neutral parameters and processes. This also helps us to understand how the P- and Q-worlds converge.

Let us say we want to project the value of the exchange rate of Indian Rupee against the US Dollar (USD/INR). USD/INR as of end October 2017 is 65. We assume that the returns follow a normal distribution characterized by its mean and standard deviation. The projection could be done either in the P-space or in the Q-space. In the P-space, the projection would be based on the historical average annual return and the historical annualized standard deviation, and we would use these first and second moments to project the USD/INR. The equivalent method in Q-world would be to calculate using the arbitrage-free drift.

To estimate the arbitrage-free drift, let us assume that the annualized USD interest rate for 1 year is 1.7% and that the 1-year INR rate is 6.2%. A dollar parked in a savings account in the USA should yield the same return as that dollar being converted to rupee, parked in India, and then reconverted to USD. This is needed to ensure no arbitrage in an arbitrage-free foreign exchange market. This implies that the exchange rate of USD/INR should depreciate at 4.5%. This can also be understood from the uncovered interest rate parity criterion for a frictionless global economy (Frenkel and Levich 1981). The criterion specifies that real interest rates should be the same all over the world. Let us assume that real interest rate globally is 0.5% and that the US inflation is 1.2%, implying nominal interest rate is 1.7%. Likewise, inflation in India is 5.7% implying a nominal interest rate of 6.2%. Inflation in India is 5.7% and that in the USA is 1.2%. Therefore, the currency in India (Rupee) should get depreciated against the US currency (Dollar) by the differential of their respective inflations 4.5% (=5.7%—1.2%). Let us assume that the standard deviation of USD/INR returns is 10%. Once we have the mean and standard deviation, we can run a Monte Carlo simulation to project the financial variable of interest. Such an exercise could be of interest if revenues are in dollars and substantial portion of expenditure is in dollars. For a more detailed reading of applicability of this exercise to the various components of earnings such as revenues and cost of goods sold in foreign currency, please refer to “Appendix C: Components of earning of the CreditMetrics™” document by JPMorgan (CreditMetrics 1999).

Monte Carlo Simulation

We first pick a random number from the standard normal distribution say x. We then scale (multiply) x by standard deviation and add average return to get a random variable mapped to the exact normal distribution of returns.

$$R = x^{*}\,(10\%)\,/\mathrm{sqrt}(365) + (4.5\%)\,/365$$

Note that the average return and standard deviation are adjusted for daily horizon by dividing with 365 and square root of 365, respectively. After scaling the variable, we multiply price of USD/INR at t with $(1 + R)$ to project the value of USD/INR to the next day. The above steps are explained in the spreadsheet “Financial Analytics Steps 4 and 5.xlsx” (available on the book’s website). An example of the above simulation with USD/INR at 65 levels at day 1 is run for seven simulations and 10 days in Table 20.1 and Fig. 20.4.

Once we have the projections of the currency rates in forward points in time, it is an easy task to then evaluate the different components of earnings that are affected by the exchange rate.

Table 20.1 Simulation of the dollar to rupee exchange rate

Date	1	2	3	4	5	6	7
27-10-2017	65.00	65.00	65.00	65.00	65.00	65.00	65.00
28-10-2017	65.08	65.25	64.89	65.59	64.72	65.13	64.86
29-10-2017	65.57	64.74	65.56	65.37	64.41	65.16	64.58
30-10-2017	64.78	65.09	65.52	65.58	64.31	65.58	64.41
31-10-2017	65.27	64.65	65.28	65.51	64.92	65.72	63.98
01-11-2017	65.04	64.15	65.08	65.80	64.30	65.71	63.47
02-11-2017	64.97	64.31	65.34	65.82	64.64	65.47	63.62
03-11-2017	65.76	64.49	65.39	65.98	64.43	65.26	63.56
04-11-2017	65.63	64.24	65.22	66.35	63.97	65.73	63.19
05-11-2017	65.64	64.46	65.41	66.32	64.12	65.04	63.25
06-11-2017	65.00	64.39	65.29	66.34	64.29	64.92	64.17
07-11-2017	64.59	64.26	65.97	65.87	64.46	64.96	64.82

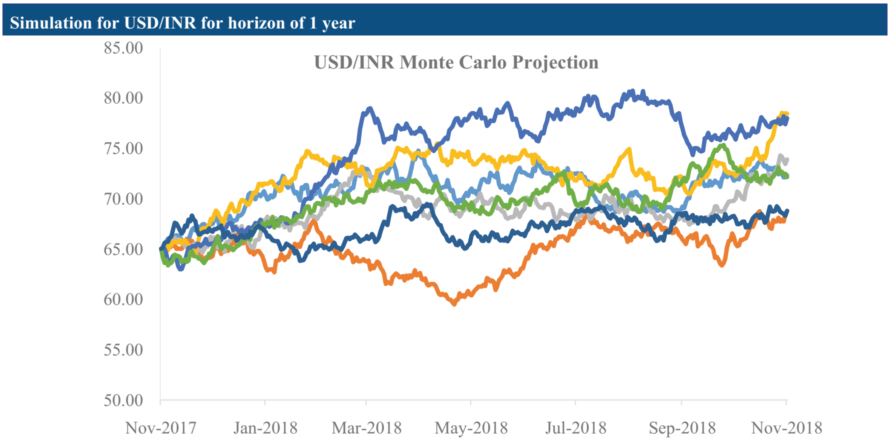

Fig. 20.4 Simulation of the dollar to rupee exchange rate

Step 5: Pricing

The fifth step is pricing which logically follows from projection. The example that we used in Step 4 was projection of USD/INR for a horizon of 1 year. What pricing allows us to do is arrive at the ex-ante expected profit or loss of a specific instrument based on the projections done in Step 4. In a typical projection technique like Monte Carlo, each of the steps is equally likely. So the probability of each of these steps is given by $1/n$, where n is the number of simulations done. The ex-ante profit or loss of the instrument is given by $1/n$ times the probability of profit or loss of the instrument. For instance, in the case of a forward contract on USD/INR that pays off 1 year from now, the payoff would be calculated at the end of 1 year as the projected value of USD/INR minus the forward rate, if it is a long forward contract and vice versa for a short forward contract. A forward contract is a contract between two parties to buy or sell an asset at a specified future point in time. In this case, the asset is USD/INR, and the specified future point in time is 1 year. The party that buys USD is supposed to be a "Long" forward contract, while the other party selling USD 1 year from now is a "Short" forward contract. The expected ex-ante payoff is the summation of the payoff in all the scenarios divided by the number of simulations. After pricing, we move on to the next stage of risk management.

1.3.2 Stage II: Risk Management

The second stage of data analytics in finance concerns risk management. It involves analysis for risk aggregation, risk assessment, and risk attribution. The framework can be used for risk analysis of a portfolio or even for an entire financial institution.

Step 1: Aggregation

The first of the three steps in risk management is risk aggregation. The aggregation step is crucial because all financial institutions need to know the value of the portfolio of their assets and also the aggregated risk exposures in their balance sheet. After we have priced the assets at the instrument level, to calculate the value of the portfolio, we need to aggregate them keeping in view the fact that the risk drivers are correlated. The correlation of the various risk drivers and the financial instruments' risk exposure is thereby aggregated. We exposit aggregation using one of the commonly used tools for risk aggregation called copula functions (Ali et al. 1978). Copula functions, especially Gaussian copulas, are used extensively by the financial industry and the regulators due to their analytical tractability. The Basel Committee on Banking Supervision relies exclusively on Gaussian copula to measure risk capital of banks globally. A copula function is a multivariate probability distribution for which the marginal distributions are known. Copula function illustrates the dependence between these correlated random variables. Copula in Latin means to link or to tie. They are widely used in both the P-world and the Q-world for risk aggregation and optimization. The underlying edifice for copula function is the Sklar's theorem (Sklar 1959). This theorem posits that any multivariate joint distribution of risk drivers can be described in terms of the univariate marginal distributions of the individual risk drivers. A copula function describes the dependence structure between these correlated random variables for the univariate marginal distributions. As usual, we discuss this step using an example. Let us say that there is a loan portfolio comprising N number of exposures. To keep the example computationally simple, we keep $N = 5$. So we have a bank which has lent money to five different corporates which we index $i = 1, 2, \ldots,$ 5. We assume for simplicity that Rs. 100 is lent to each of the five firms. So the loan portfolio is worth Rs. 500. The way we will go about aggregating the risk of this Rs. 500 loan portfolio is that we will first describe the marginal distribution of credit for each of the five corporates. We will then use the Gaussian copula function to get the joint distribution of the portfolio of these five loans.

Let us assume for simplicity that each corporate has a probability of default of 2%. Therefore, there is a 98% chance of survival of the corporate in a year. The horizon for the loan is 1 year. Assume that in the event of a default, the bank can recover 50% of the loan amount. The marginal distributions are identical in our example for ease of exposition, but the copula models allow for varying distributions as well. What we want to do is that based on the correlation structure, we want to calculate the joint distribution of credit of each of these corporates. We model this using a one-factor model. The single factor is assumed to be the state of the economy *M*, which is assumed to have a Gaussian distribution.

To generate a one-factor model, we define random variables x_i $(1 \leq i \leq N)$:

$$x_i = \rho_i M + \sqrt{1-\rho_i^2} Z_i$$

In the above equation, the single factor *M* and the idiosyncratic factor Z_i are independent of each other and are standard normal variables with mean zero and unit standard deviation. The correlation coefficient ρ_i satisfies $1 \leq \rho_i < 1$. The above equation defines how the assets of the firm are correlated with the economy *M*. The correlation between the assets x_i of firm *i* and assets x_j of firm *j* is $\rho_i \rho_j$.

Let *H* be the cumulative normal distribution function of the idiosyncratic factor Z_i. Therefore:

$$Probability\,(x_i < x|\,M) = H\left(\frac{x-\rho_i M}{\sqrt{1-\rho_i^2}}\right)$$

The assets of each of these corporates are assumed to have a Gaussian distribution. Note that the probability of default is 2%, corresponding to a standard normal value of −2.05. If the value of the asset standardized with the mean and its standard deviation is more than −2.05, the entity survives, else it defaults. The conditional probability that the i^{th} entity will survive is therefore:

$$S_i\,(x_i < x|\,M) = 1 - H\left(\frac{x-\rho_i M}{\sqrt{1-\rho_i^2}}\right)$$

The marginal distribution of each of the assets is known, but we do not know the joint distribution of the loan portfolio. So, we model the portfolio distribution using copulas based on the correlations that each of these corporates has. The performance of the corporate depends on the state of the economy. There is a correlation between these two variables. This can be explained by noting that certain industries such as steel and cement are more correlated with the economy than others like fast-

Table 20.2 State probabilities for one-factor Gaussian copula

States of the economy	Midpoint of range	Probability of state
1	−3.975	7.40435E-06
2	−3.925	9.02075E-06
3	−3.875	1.09626E-05
4	−3.825	1.32891E-05
5	−3.775	1.60692E-05
6	−3.725	1.93824E-05
7	−3.675	2.33204E-05
8	−3.625	2.79884E-05
9	−3.575	3.3507E-05
10	−3.525	4.00135E-05

moving consumer goods. Assume that the correlation of the first corporate with the economy is 0.2, the second is 0.4, the third is 0.5, the fourth is 0.6, and the fifth is 0.8. So the pairwise correlation can be calculated as the product of the two correlations to the single factor, which in our example is the economy. We model the state of the economy as a standard normal random variable in the range from −3.975 to 3.975 in intervals of 0.05. We take the mid-point of these intervals. Table 20.2 shows these values for the first ten states of the economy. The probability of the economy being in those intervals is calculated in column 3 of Table 20.2 using the Gaussian distribution. This is given by:

$$Prob\left\{m-\frac{\Delta}{2}\leq M\leq m+\frac{\Delta}{2}\right\}$$

where M follows the standard normal distribution, m is the mid-point of the interval, and Δ is the step size. The way to interpret the state of the economy is that when it is highly negative such as −2, then the economy is in recession. And if it is high such as greater than 2, the economy is booming, and if it is close to zero, then the health of the economy is average. Once we have the probabilities for the state of the economy (Table 20.2), we calculate the conditional probability of a corporate defaulting, and this again depends on the correlation between its asset values and the states of the economy.

Let $\pi(k)$ be the probability that exactly k firms default in the N-firm loan portfolio. Depending on the state of the economy, the conditional probabilities of M are independent. Therefore, the conditional probability that all the N firms will survive is:

$$\pi\ (0|M)=\prod_{i=1}^{N}S_i\ (x_i\ <x|\ M)$$

Similarly,

$$\pi\ (1|M) = \pi\ (0|M) \sum_{i=1}^{N} \frac{1 - S_i\ (x_i\ < x|\ M)}{S_i\ (x_i\ < x|\ M)}$$

Define

$$w_i\ = \sum_{i=1}^{N} \frac{1 - S_i\ (x_i\ < x|\ M)}{S_i\ (x_i\ < x|\ M)}$$

Conditioned on the state of the economy, the chance of exactly k firms defaulting is given by the combinatorial probability

$$\pi\ (k|M) = \pi\ (0|M) \sum_{i=1}^{N} w_{q(1)} w_{q(2)} \ldots w_{q(k)}$$

where $\{q(1), q(2), \ldots, q(k)\}$ is the combinatorial interpretation of the number of ways of k default among N firms $\{1, 2, \ \ldots, N\}$ and the summation is taken over the

$$q(k) = \frac{N!}{k!\,(N-k)!}$$

different ways in which k firms can default among N firms.

$\pi(k|\ M)$ is the combinatorial probability of k defaults, and $\sum_{i=1}^{N} w_{q(1)} w_{q(2)} \ldots w_{q(k)}$ represents summation over all possible combinations of k defaults among N firms. This is tabulated in Table 20.3 for the first ten states.

Table 20.3 Conditional survival probabilities for one-factor Gaussian copula

States of economy	Corporate_1	Corporate_2	Corporate_3	Corporate_4	Corporate_5
1	0.623	0.278	0.143	0.053	0.000
2	0.632	0.292	0.155	0.060	0.001
3	0.642	0.306	0.167	0.068	0.001
4	0.651	0.320	0.180	0.076	0.001
5	0.660	0.335	0.194	0.085	0.002
6	0.669	0.350	0.208	0.095	0.002
7	0.678	0.365	0.222	0.106	0.003
8	0.687	0.381	0.237	0.118	0.004
9	0.696	0.397	0.253	0.130	0.005
10	0.705	0.412	0.269	0.144	0.007

Table 20.4 Conditional joint survival probabilities for one-factor Gaussian copula

States of economy	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
1	5.37E-07	0.001336	0.035736	0.24323	0.134507	0.220673
2	9.99E-07	0.001746	0.041595	0.258263	0.136521	0.206578
3	1.83E-06	0.002267	0.048212	0.273353	0.138099	0.192794
4	3.3E-06	0.002926	0.055647	0.288374	0.139211	0.179355
5	5.84E-06	0.003755	0.063955	0.30319	0.139833	0.166295
6	1.02E-05	0.004791	0.073188	0.317651	0.139941	0.153647
7	1.75E-05	0.006079	0.08339	0.331596	0.139519	0.141441
8	2.96E-05	0.007672	0.094599	0.344857	0.138554	0.129705
9	4.93E-05	0.00963	0.106837	0.357256	0.13704	0.118463
10	8.08E-05	0.012025	0.120116	0.368611	0.134975	0.107738

Then we calculate the discrete joint distribution of survival of all the firms together. There are five possibilities—all firms survive, one firm fails, two firms fail, three firms fail, four firms fail, and all five firms fail. This is tabulated in Table 20.4 for the first ten states.

The above steps are explained in the spreadsheet "Financial Analytics Step 6.xlsx" (available on the book's website). For each outcome, we have the losses corresponding to that precise outcome. So using the copula functions we have effectively used the information on marginal distribution of the assets of each firm and their correlation with the economy, to arrive at the joint distribution of the survival outcomes of the firms. We thus are able to aggregate the risk of the portfolio even though as our starting point we only had the marginal probability distribution of only individual loans.

Step 2: Assessment

We now move on to the second step of risk management which is assessment of the portfolio. Assessment of the portfolio is done by summarizing it according to a suitable statistical feature. More precisely, assessment is done by calculating the ex-ante risk of the portfolio using metrics such as threshold persistence (TP) or value at risk (VaR) and sometimes sensitizing it using methods like stress-testing. Threshold persistence is defined as follows: Given the time frame for which a portfolio would remain constant and unchanged (T), the threshold level of cumulative portfolio return (β) and the horizon over which the cumulative return remains below the threshold β. VaR, on the other hand, is a measure of the risk of a portfolio under normal market conditions over a certain time horizon, typically a year, for most asset classes. VaR is used by regulators and firms to assess out how much loss can possibly happen in a certain portfolio and how much asset value is required to cover for this loss. Since VaR is intuitive and is comparable on an apple to apple basis across asset classes, it is widely popular both with the regulators and the market participants. VaR is defined for a given confidence level—usually 99%. This means that the risk manager can be confident that 99 times out of 100, the loss from the portfolio will not exceed the VaR metric. This metric is also used for financial reporting and for calculating the regulatory capital of financial institutions. VaR is an ex-ante assessment in the Bayesian sense—the VaR number is a value that is ex-ante assessed as the loss that can possibly result for the portfolio. It only incorporates information available at the time of computation. VaR is used for governance in pension plans, endowments, trusts, and other such risk-averse financial institutions where the investment mandate often defines the maximum acceptable loss with given probabilities. A detailed description of how Value at Risk has been used to calculate capital can be found in Chapter 3, "VAR-Based Regulatory Capital," of the book *Value at Risk: The New Benchmark for Managing Financial Risk* by Philippe Jorion. This particular measure incorporates the previous steps of portfolio aggregation. We will understand the step using an example. We will examine the VaR computation with a simple portfolio comprising 1 USD, 1 EUR, 1 GBP, and 100 JPY. The value of the portfolio in INR terms is Rs. 280 (1 USD = Rs. 64, 1 Euro (EUR) = Rs. 75, 1 Sterling (GBP) = Rs. 82, 100 Yen (JPY) = Rs. 59). We want to calculate at the end of 1 year what is the possible loss or gain from this particular portfolio. To aggregate the risk, we make use of the correlation matrix between the currencies as described in Table 20.5.

We will use Cholesky decomposition—which fundamentally decomposes the correlation matrix into a lower triangular matrix and an upper triangular matrix (Press et al. 1992). The only condition is that the correlation matrix should be positive definite Hermitian matrix. This decomposition is almost akin to computing the square root of a real number.

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^*$$

$\boldsymbol{A}$ is a positive definite Hermitian matrix, $\boldsymbol{L}$ is a lower triangular matrix, and $\boldsymbol{L}^*$ is the transpose conjugate of $\boldsymbol{L}$. The Cholesky decomposed matrix for the correlation matrix of Table 20.5 is shown in Table 20.6.

Table 20.5 Currency correlation matrix

	USD/INR	EUR/INR	GBP/INR	JPY/INR
USD/INR	1	0.9	0.5	0.5
EUR/INR	0.9	1	0.5	0.5
GBP/INR	0.5	0.5	1	0.2
JPY/INR	0.5	0.5	0.2	1

Table 20.6 Cholesky decomposed lower triangular matrix

	USD/INR	EUR/INR	GBP/INR	JPY/INR
USD/INR	1	0	0	0
EUR/INR	0.9	0.43589	0	0
GBP/INR	0.5	0.114708	0.858395	0
JPY/INR	0.5	0.114708	−0.07358	0.855236

Table 20.7 Simulation of portfolio gain/loss

Simulation number	Log prices				Prices in INR				Gain/loss
	USD	EUR	GBP	JPY	1 USD	1 EUR	1 GBP	1 JPY	
0	4.2	4.3	4.4	4.1	64.0	75.0	82.0	59.0	280.0
1	4.2	4.4	4.6	4.2	67.2	81.6	96.1	64.2	29.0
2	4.2	4.4	4.5	4.2	68.6	79.8	88.5	63.5	20.4
3	4.2	4.3	4.5	4.0	63.6	70.9	88.2	53.8	−3.6
4	4.3	4.5	4.6	4.2	74.3	89.6	95.8	64.3	44.0
5	4.2	4.4	4.4	4.2	69.1	77.6	83.7	65.6	16.0
6	4.2	4.4	4.4	4.3	66.2	78.6	83.4	71.4	19.6
7	4.2	4.3	4.4	4.0	63.7	74.7	81.0	55.3	−5.3
8	4.1	4.2	4.4	4.0	62.6	69.2	82.1	56.3	−9.9
9	4.2	4.3	4.4	4.2	64.6	71.4	83.4	63.8	3.2
10	4.2	4.2	4.4	4.0	65.8	69.1	81.4	56.6	−7.1

For each currency we then simulate a random number drawn from a standard normal distribution. These are independently drawn. This vector of independent draws can be converted to a vector of correlated draws by multiplying with the decomposed matrix.

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{X}$$

where $\boldsymbol{Y}$ is the vector of correlated prices and $\boldsymbol{X}$ is the vector of i.i.d. draws.

This process is repeated multiple times to arrive at a simulation of correlated draws. Using Step 4 we project the log of the prices of USD/INR, EUR/INR, GBP/INR, and JPY/INR. We price the exchange rate and aggregate the portfolio and subtract from the original value to get the portfolio loss or gain. These steps are repeated for a given number of simulations as shown in Table 20.7.

We then calculate the VaR at 99% level from the simulated gains or losses. The above steps are explained in the spreadsheet “Financial Analytics Step 7.xlsx” (available on the book’s website). For a simulation run 100 times on the above data, a VaR of −38 INR was obtained at 1% confidence level.

Step 3: Attribution

The third step in risk management analysis is attribution. Once we have assessed the risk of the portfolio in the previous step, we need to now attribute the risk to different risk factors. For instance, the combined risk of Rs. 38 of the portfolio in the previous example can be attributed to each of the individual assets. Like for a portfolio, this can be done at a firm level as well. What financial institutions typically do is to attribute risk along a line of business (LoB). This is because banks and financial institutions are interested in measuring the capital consumed by various activities. Capital is measured using the Value at Risk metric. VaR has become an inalienable tool for risk control and an integral part of methodologies that seek to allocate economic and/or regulatory capital. Its use is being encouraged by the Reserve Bank of India (RBI), the Federal Reserve Bank (Fed), the Bank for International Settlements, the Securities and Exchange Board of India (SEBI), and the Securities and Exchange Commission (SEC). Stakeholders including regulators and supervisory bodies increasingly seek to assess the worst possible loss (typically at 99% confidence levels) of portfolios of financial institutions and funds. A detailed description of how Value at Risk has been used to calculate capital can be found in Chapter 3, "VAR-Based Regulatory Capital," of the book *Value at Risk: The New Benchmark for Managing Financial Risk* by Philippe Jorion. There are three commonly employed measures of VaR-based capital—stand-alone, incremental, and component. It has been found that different banks globally calculate these capital numbers differently, but they follow similar ideas behind the measures.

Stand-Alone Capital

Stand-alone capital is the amount of capital that the business unit would require, if it were viewed in isolation. Consequently, stand-alone capital is determined by the volatility of each LoB's earnings.

Incremental Capital

Incremental capital measures the amount of capital that the business unit adds to the entire firm's capital. Conversely, it measures the amount of capital that would be released if the business unit were sold.

Component Capital

Component capital, sometimes also referred to as allocated capital, measures the firm's total capital that would be associated with a certain line of business. Attributing capital this way has intuitive appeal and is probably the reason why it is particularly widespread.

We use a simplified example to understand how attribution is done using metrics such as stand-alone, incremental, and component capital. Let us assume that there is a bank that has three business units:

- Line of Business 1 (LoB1)—Corporate Banking
- Line of Business 2 (LoB2)—Retail Banking
- Line of Business 3 (LoB3)—Treasury Operations

For ease of calculation, we assume that the total bank asset is $A =$ Rs. 3000 crores. We also assume for the sake of simplicity that each of the LoBs has assets worth $A_i =$ Rs. 1000 crores, $i = 1, 2, 3$. The volatility of the three lines of businesses is:

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\rho_{12}\sigma_1\sigma_2 + 2\rho_{23}\sigma_2\sigma_3 + 2\rho_{31}\sigma_3\sigma_1}$$

where σ_i is the volatility of the i^{th} line of business and ρ_{ij} is the correlation between the i^{th} and j^{th} LoB. The volatility of all three LoBs is calculated in Table 20.8, while that of each LoB is calculated in Table 20.9.

Table 20.8 Total capital calculation for the entire business

	Assets	Volatility	Capital
Bank	3000	4.534%	316.9142629

Table 20.9 Capital calculation attribution for LoB1, LoB2, and LoB3

	Assets	Volatility	Standalone capital	Incremental capital	Component capital
LoB1	1000	$\sigma_1 = 5\%$	116.50	51.25	69
LoB2	1000	$\sigma_2 = 7\%$	163.10	67.05	102
LoB3	1000	$\sigma_3 = 9\%$	209.70	89.81	146
Total	3000	4.53%	489	208	317
Unattributed			(172)	109	–

LoB1 is moderately correlated with that of LoB2 (ρ_{12}=30%) and less correlated to LoB3 (ρ_{31}=10%). LoB2 is uncorrelated to LoB3 (ρ_{23}=0).

The capital required at 99% (z = 2.33) calculated as Value at Risk is given by $2.33A_i\sigma_i$. The stand-alone capital required for the first line of business is 2.33 × Rs.1000 crores × 5% = 116.50 crores. The stand-alone capital required for the second line of business is 2.33 × Rs.1000 crores × 7% = 163.10 crores. The stand-alone capital required for the third line of business is 2.33 × Rs.1000 crores × 9% = 209.70 crores.

The total capital is given by:

$$C = 2.33A\sigma$$

$$= 2.33\sqrt{A_1^2\sigma_1^2+A_2^2\sigma_2^2+A_3^2\sigma_3^2+2\rho_{12}A_1A_2\sigma_1\sigma_2+2\rho_{23}A_2A_3\sigma_2\sigma_3+2\rho_{31}A_3A_1\sigma_3\sigma_1}$$

The incremental capital for LoB1 is calculated as the total capital less the capital of LoB2 and LoB3. It measures the incremental increase in capital from adding LoB1 to the firm. The incremental capital for LoB1 is therefore:

$$= 2.33\left\{\sqrt{A_1^2\sigma_1^2+A_2^2\sigma_2^2+A_3^2\sigma_3^2+2\rho_{12}A_1A_2\sigma_1\sigma_2+2\rho_{23}A_2A_3\sigma_2\sigma_3+2\rho_{31}A_3A_1\sigma_3\sigma_1}\right.$$
$$\left.-\sqrt{A_2^2\sigma_2^2 + A_3^2\sigma_3^2 + 2\rho_{23}A_2A_3\sigma_2\sigma_3}\right\}$$

The incremental capital for LoB2 is therefore:

$$= 2.33\left\{\sqrt{A_1^2\sigma_1^2+A_2^2\sigma_2^2+A_3^2\sigma_3^2+2\rho_{12}A_1A_2\sigma_1\sigma_2+2\rho_{23}A_2A_3\sigma_2\sigma_3+2\rho_{31}A_3A_1\sigma_3\sigma_1}\right.$$
$$\left.-\sqrt{A_1^2\sigma_1^2 + A_3^2\sigma_3^2 + 2\rho_{31}A_3A_1\sigma_3\sigma_1}\right\}$$

The incremental capital for LoB3 is calculated as:

$$= 2.33\left\{\sqrt{A_1^2\sigma_1^2+A_2^2\sigma_2^2+A_3^2\sigma_3^2+2\rho_{12}A_1A_2\sigma_1\sigma_2+2\rho_{23}A_2A_3\sigma_2\sigma_3+2\rho_{31}A_3A_1\sigma_3\sigma_1}\right.$$
$$\left.-\sqrt{A_1^2\sigma_1^2 + A_2^2\sigma_2^2 + 2\rho_{12}A_1A_2\sigma_1\sigma_2}\right\}$$

The component capital for LoB1 is calculated as:

$$A_1\sigma_1\frac{\partial C}{\partial\left(A_1\sigma_1\right)} = A_1\sigma_1\frac{\left(A_1\sigma_1 + \rho_{12}A_2\sigma_2 + \rho_{31}A_3\sigma_3\right)}{A\sigma}$$

This is because $\frac{\partial\sigma}{\partial\sigma_1} = \frac{(\sigma_1+\rho_{12}\sigma_2+\rho_{31}\sigma_3)}{\sigma}$.

Similarly, the component capital for LoB2 is calculated as:

$$A_2\sigma_2\frac{\partial C}{\partial\left(A_2\sigma_2\right)} = A_2\sigma_2\frac{\left(A_2\sigma_2 + \rho_{12}A_1\sigma_1 + \rho_{23}A_3\sigma_3\right)}{A\sigma}$$

Likewise, the component capital for LoB3 is:

$$A_3\sigma_3\frac{\partial C}{\partial\left(A_3\sigma_3\right)} = A_3\sigma_3\frac{\left(A_3\sigma_3 + \rho_{13}A_1\sigma_1 + \rho_{23}A_2\sigma_2\right)}{A\sigma}$$

The component capital of each LoB always sums to the total capital. Please refer to the spreadsheet "Financial Analytics Step 8.xlsx" (available on the book's website) for the specificities of the calculation. Readers interested in total capital calculation for the entire business may refer to the RiskMetrics™ framework developed by JPMorgan (RiskMetrics 1996).

1.3.3 Stage III: Portfolio Analysis

The third stage of data analytics in finance concerns portfolio risk management. It involves optimal allocation of risk and return as well as the execution required to move the portfolio from a suboptimal to an optimal level.

Step 1: Allocation

After having aggregated the portfolio, assessed the risk, and then attributed the risk to different lines of businesses, we move on to changing the portfolio for the entire firm, for a division or an LoB for optimal allocations. So if we continue with the previous example where we have three lines of business, the amount is essentially kept the same—Rs. 1000 crores. If we analyze the results from Step 3 of risk management, we find that risk attribution from all three metrics—standalone, incremental, and component capital—indicates that the lowest attribution of risk happens along the first line of business. If the Sharpe ratio (excess return as a proportion of risk) for LoB1 is the highest (followed by that of LoB2 and LoB3 respectively), then it is optimal for the firm to allocate more capital to the first line of business and then to the second line of business. LoB3 is perhaps the most expensive in terms of risk-adjusted return. Step 1 of portfolio analysis involves optimally allocating the assets such that the overall risk of the firm is optimal. Readers interested in optimal allocation of assets may refer to the RiskMetrics framework developed by JPMorgan (RiskMetrics 1996).

Step 2: Execution

The last step is execution. Having decided to change the portfolio from its current level to a more optimal level, we have to execute the respective trades for us to be able to get to the desired portfolio risk levels. Execution happens in two steps. The first step is order scheduling which is basically a planning stage of the execution process. Order scheduling involves deciding how to break down a large trade into smaller trades and timing each trade for optimal execution. Let us say a financial institution wants to move a large chunk of its portfolio from one block to the other. This is called as a parent order which is further broken down into child orders. The timescale of the parent order is in the order of a day known as volume time. In execution, the way time is measured is not so much in calendar time (called wall-clock time) but in what is called as activity time. Activity time behaves as a random walk. In this last step, we are coming back to Step 1 where we said that we need to identify the risk drivers. For execution, the variable to be modeled is the activity time. This behaves approximately as a random walk with drift and activity time as a risk driver in the execution world, especially in high-frequency trading.

There are two kinds of activity time—tick time and volume time. Tick time is the most natural specification for activity time on very short timescales which advance by 1 unit whenever a trade happens. The second type—volume time—can be intuitively understood by noting that volume time lapses faster when more trading activity happens, that is, the trading volume is larger. After the first step of order scheduling, the second step in order execution is order placement which looks at execution of child orders, and this is again addressed using data analytics. The expected execution time of child orders is of the order of a minute. The child orders—both limit orders and market orders—are based on real-time feedback using opportunistic signals generated from data analytic techniques. So, in order placement, the timescale of limit and market orders is of the order of milliseconds, and the time is measured by tick-time which is discrete. These two steps are repeated in execution algorithms after concluding the first child order called scheduling. It is executed by placing limit and market orders. Once the child order is fully executed, we update the parent order with the residual amount to be filled. We again compute the next child order and execute. This procedure ends when the parent order is exhausted. Execution is almost always done programmatically using algorithms and is known as high-frequency trading (HFT). The last step thus feeds back into the first step of our framework.

1.4 Conclusion

To conclude, the framework consists of three stages to model, assess, and improve the performance of a financial institution and/or a portfolio. The first five steps pertain to econometrical estimation. The next three steps concern risk management and help measure the risk profile of the firm and/or the portfolio. The last two steps are about portfolio management and help in optimizing the risk profile of the financial institution and/or the portfolio. Following these sequential steps across three stages helps us avoid common pitfalls and ensure that we are not missing important features in our use of data analytics in finance. That being said, not every data analysis in the finance world involves all the steps across three stages. If we are only interested in estimation, we may just follow the first five steps. Or if we are only interested in risk attribution, it may only involve Step 3 of risk management. The framework is all encompassing so as to cover most possible data analysis cases in finance. Other important aspects outside the purview of the framework like data cleaning are discussed in Sect. 20.2.

2 Part B: Applications

2.1 Introduction

This chapter intends to demonstrate the kind of data science techniques used for analysis of financial data. The study presents a real-world application of data analytic methodologies used to analyze and estimate the risk of a large portfolio over different horizons for which the portfolio may be held. The portfolio that we use for this study consists of nearly 250 securities comprising international equities and convertible bonds. The primary data science methods demonstrated in the case study are principal component analysis (PCA) and Orthogonal GARCH. We use this approach to achieve parsimony by reducing the dimensionality of the data, which is a recurring objective in most data analytic applications in finance. This is because the dimensionality of the data is usually quite large given the size, diversity, and complexity of financial markets. We simultaneously demonstrate common ways of taking into account the time-varying component of the volatility and correlations in the portfolio, another common goal in portfolio analysis. The larger objective is to demonstrate how the steps described in the methodology framework in the chapter are actually implemented in financial data analysis in the real world.

The chapter is organized as follows. The next section describes the finance aspects of the case study and its application in the financial world. Section 2.3 also discusses the metrics used in the industry for assessing risk of the portfolio. In Sect. 2.4, the data used and the steps followed to make the data amenable for financial analysis are described. Section 2.5 explains the principles of principal component analysis and its application to the dataset. Section 2.6 explains the Orthogonal GARCH approach. Section 2.7 describes three different types of GARCH modeling specific to financial data analysis. The results of the analysis are presented in Sect. 2.8.

2.2 Application of Data Science in the World of Investing

For most non-finance professionals, investments especially in hedge funds are shrouded in secrecy. The sensational stories of Ponzi hedge funds like that of Bernard Madoff make for great headlines and even greater storytelling. In fact, the chronicle of Bernie Madoff's Ponzi scheme is now a Hollywood movie called "The Wizard of Lies" starring Robert De Niro which got released in May 2017. But Hollywood movies do little to advance data analytics education or explain how data science can be used to investigate the portfolio risk of a hedge fund. Not all asset managers have the resources of a Harvard or Yale endowment fund to apply sophisticated models to detect market risk in hedge fund portfolios. This does not mean that we cannot use econometric models to estimate, measure, and assess market risk in portfolios, as we will demonstrate in this case study.

Before the advent of data science, measurement of market risk in relation to hedge funds was considered difficult, if not unmanageable. Large endowment funds like that of Harvard and Yale had the resource to engage econometricians to do the quantitative risk assessment and measurement, but it was mostly the preserve of a select few. Additionally, for a long time, hedge funds were engaged by "golden aged" investment managers who had no understanding of data science. These investment managers were mostly statistically challenged and therefore had more than their fair share of skepticism with regard to data science. They had the good old perspective that hedge fund risks and returns are based on fund managers' talent and that quantitative risk measures are not capable of measuring such complex risks. As a result, the most common risk assessment technique was extensive due diligence carried out by a dedicated set of risk professionals.

However, since the explosion of data science techniques and methodologies in the last decade, there has been a tectonic shift in how data science is viewed in the investment management world. If baseball matches and election outcomes can be predicted using data science, surely hedge fund risks too can be assessed using econometric tools. Another practical challenge facing the investment management industry has been the increase in the size and number of the hedge funds. As per Bloomberg estimates, there are more than 10,000 hedge funds available for investment. It is humanly impossible to carry out due diligence of more than 10,000 hedge funds by any one asset management company (AMC).

Apart from developments in data science and the vastness of hedge fund universe, another important driver in the use of data analytics in asset management has been the advancements in robust risk quantification methodologies. The traditional measures for risk were volatility-based Value at Risk and threshold persistence which quantified downside deviation. These risk metrics are described in the next section. The problem with a simple volatility-based Value at Risk is that it assumes normality. So the assumption made is that financial market returns distribution is symmetrical and that the volatility is constant and does not change with time. It implicitly assumes that extreme returns, either positive or negative, are highly unlikely. However, history suggests that extreme returns, especially extreme negative returns, are not as unlikely as implied by the normal distribution. The problem with downside measures such as threshold persistence is that, although they consider asymmetry of returns, they do not account for fat tails of distributions. These criticisms have resulted in the development of robust risk measures that account for fat tails and leverage such as GJR and EGARCH (see Sect. 2.7.5). So, nowadays all major institutional investors who have significant exposure to hedge funds employ P-quants and use data analytic techniques to measure risk. The exceptions of the likes of Harvard and Yale endowment funds have now become the new norm. Consolidation of market risk at the portfolio level has become a standard practice in asset management. In this chapter, we present one such analysis of a large portfolio comprising more than 250 stocks (sample data in file: tsr.txt) having different portfolio weights (sample data in file: ptsr.txt) and go through the steps to convert portfolio returns into risk metrics. We use Steps 1–6 of the data analysis methodology framework. We first identify the financial variable to model as stock returns. We reduce the dimensionality of the data using principal component analysis from 250 stock returns to about ten principal components. We then use GARCH, GJR, and EGARCH (described in Step 3 of "Part A—Methodology") to make suitable inference on portfolio returns. We estimate the GARCH, GJR, and EGARCH parameters using maximum likelihood estimation. We then project the portfolio returns (Step 4 of the methodology) to forecast performance of the hedge fund. We finally aggregate the risks using Step 6 of the framework and arrive at the key risk metrics for the portfolio. We now describe the risk metrics used in the investment management industry.

2.3 Metrics for Measuring Risk

As described in the "Financial Analytics: Part A—Methodology," two metrics are used for measuring the risk of the portfolio: value at risk and threshold persistence.

2.3.1 Value at Risk (VaR)

Value at Risk has become one of the most important measures of risk in modern-day finance. As a risk-management technique, Value at Risk describes the loss in a portfolio that can occur over a given period, at a given confidence level, due to exposure to market risk. The market risk of a portfolio refers to the possibility of financial loss due to joint movement of market parameters such as equity indices, exchange rates, and interest rates. Value at Risk has become an inalienable tool for risk control and an integral part of methodologies that seek to allocate economic and/or regulatory capital. Its use is being encouraged by the Reserve Bank of India (RBI), the Federal Reserve Bank (Fed), the Bank for International Settlements, the Securities and Exchange Board of India (SEBI), and the Securities and Exchange Commission (SEC). Stakeholders including regulators and supervisory bodies increasingly seek to assess the worst possible loss (typically at 99% confidence levels) of portfolios of financial institutions and funds. Quantifying risk is important to regulators in assessing solvency and to risk managers in allocating scarce economic capital in financial institutions.

2.3.2 Threshold Persistence (TP)

Given a threshold level of return for a given portfolio, traders and risk managers want to estimate how frequently the cumulative return on the portfolio goes below this threshold and stays below this threshold for a certain number of days. Traders also want to estimate the minimum value of the cumulative portfolio return when the above event happens. In order to estimate both these metrics, two factors specify a threshold, namely, financial market participants define a metric called threshold persistence.

Threshold persistence is defined as follows: Given the time frame for which a portfolio would remain constant and unchanged (T), two factors specify a threshold, namely, cumulative portfolio return (β) and the horizon over which the cumulative return remains below the threshold β. For the purposes of this chapter, we label this threshold horizon as T'. The threshold persistence metrics are defined as:

(a) The fraction of times the net worth of the portfolio declines below the critical value (β) vis-à-vis the initial net worth of the portfolio and remains there for T' days beneath this critical value
(b) The mean decline in the portfolio net worth value compared to the initial critical level conditional on (a) occuring

To clarify the concept, consider the following example. Say $T = 10$ days, $\beta = -5\%$, $T' = 2$ days, and the initial net worth of the portfolio is Rs. 100. We simulate the portfolio net worth (please refer to Step 4 of the methodology framework to understand how simulation is performed), and, say, we obtain the following path (Table 20.10):

Table 20.10 Threshold persistence example

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
102	98	**94**	**90**	**93**	96	98	90	95	97

The pertinent progression here for calculating (a) and (b) are the net worth of the portfolio in days 3, 4, and 5 since the net worth of the portfolio is lower than Rs. 95 on all these three days. Observe that the decline to Rs. 90 on Day 8 would not be reckoned as an applicable occurrence here since $T' = 2$ and the net worth of the portfolio came back above the critical value on Day 9 (the critical time span is 2 days, and it reverted above the critical level before 2 days). Let us suppose that we simulate ten paths in all and in not one of the remaining paths of the simulation does the portfolio value dip below the critical value and stays below the critical value over the 2-day horizon. Therefore, the proportion of times the value of the portfolio goes below the critical value is 1/10. Given that such a dip happens over the critical time period of over 2 days, the drop would be -10%.

2.4 Data

The data that is normally available from secondary sources are the prices of the various securities in the sample portfolio. The prices would be in local currencies—US securities in US dollars, Japanese equity in Japanese yen, and so on. In the case study, there is data from ten different currencies.

The data that is available from financial information services providers such as Bloomberg or Thomson Reuters (the two largest providers in the global financial markets), more often than not, is not "ready-made" for analysis. The foremost limitation in the data made available by financial information services providers is that they require considerable data cleaning before the data analytic methodologies can be applied. The data cleaning process is usually the most time-consuming and painstaking part of any data analysis, at least with financial data. The portfolio that we use for this study consists of nearly 250 securities. We use the data in the context of the study to describe in general the steps taken to make the data amenable for financial analysis:

- The prices of almost all securities are in their native currencies. This requires conversion of the prices into a common currency. Globally, the currency of choice is US dollars, which is used by most financial institutions as their base currency for reporting purposes. This is a mandatory first step because the prices and returns converted into the base currency are different from those in their native currencies.
- Securities are traded in different countries across the world, and the holidays (when markets are closed) in each of these countries are different. This can lead to missing data in the time series. If the missing data is not filled, then

this could manifest as spurious volatility in the time series. Hence, the missing data is normally filled using interpolation techniques between the two nearest available dates. The most common and simplest interpolation methodology used in financial data is linear interpolation.

- Some securities may have no price quotes at all because even though they are listed in the exchange, there is no trading activity. Even when there is some trading activity, the time periods for which they get traded may be different, and therefore the prices that are available can vary for different securities. For instance, in the portfolio that we use for this study, some securities have ten years of data, while others have less than 50 price data points available. Those securities which do not have at least a time series of prices spanning a minimum threshold number of trading days should be excluded from the analysis. For the purpose of this case study, we use 500 price data points.
- While too few price points is indeed a problem from a data analysis perspective, many a times, a long time series can be judged to be inappropriate. This is because in a longer time series, the more historical observations get the same weights as the recent observations. Since recent observations have more information relevant to the objective of predicting future portfolio risk, a longer time series can be considered inappropriate. In the case study, the time series used for analysis starts in May 2015 and ends in May 2017 thus giving us 2 years of data (in most financial markets, there are approximately 250 trading days in a year) or 500 time series observations.
- Prices are customarily converted into continuously compounded returns using the formula $r_t = \ln (P_t/P_{t-1})$. As explained in Step 1 of the methodology in the "Financial Analytics: Part A—Methodology," we work with returns data rather than price data. Time series analysis of returns dominates that using prices because prices are considerably non-stationary compared to returns.
- Portfolio returns are computed from the security returns as discussed in Step 6 of the methodology framework. In the case study, two portfolios—an equally weighted portfolio and a value-weighted portfolio (calculated by keeping the number of shares in each of the security in the portfolio constant)—are used for the analysis.

2.5 Principal Component Analysis

2.5.1 Eigenvectors and Eigenvalues

For the purposes of the case study, readers need to understand PCA, the way it is computed, and also the intuition behind the computation process. We explain the intermediate steps and the concepts therein to make Sect. 2 of the chapter self-contained. Further discussion on PCA is found in Chap. 15 on unsupervised learning.

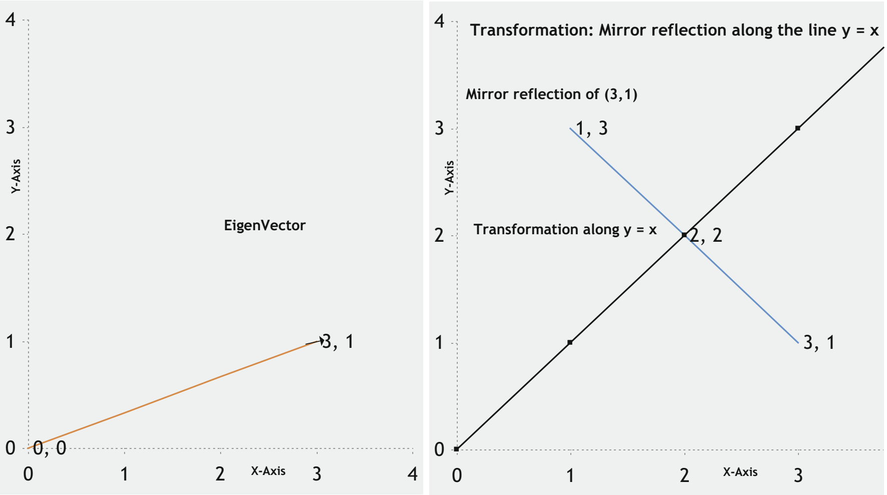

Fig. 20.5 Pictorial description of eigenvectors

From basic matrix algebra we know that we can multiply two matrices together, provided that they are of compatible sizes. Eigenvectors are a special case of this. Consider the two multiplications between a matrix and a vector below.

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$
$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

In the first multiplication, the resulting matrix $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$ is not an integer multiple of the original matrix $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$. In the second multiplication, the resulting matrix is a multiple (of 1) of the original matrix $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$. The first matrix is not an eigenvector of the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, while the second one is an eigenvector. Why is it so? The reason is that the eigenvector remains a multiple of itself after the transformation. It does not get transformed after multiplication like the first one.

One can think of the matrix $\begin{pmatrix} 1 \\ 3 \end{pmatrix}$ as a vector in two dimensions originating from (0,0) and ending at (1,3) as shown in Fig. 20.5.

For ease of visual imagination, we have employed the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ in the discussion above. This matrix can be thought of as the following transformation: reflection of any vector along the line $y = x$. For instance, a vector (1,3) after multiplication by this matrix becomes (3,1), that is, a reflection of the vector itself along the $y = x$ line. However, the reflection of the vector (2,2) would be the vector itself. It would be a scalar multiple of the vector (in this case, the scalar multiple is 1). Thus, an eigenvector even after transformation remains a scalar multiple of itself. The scalar multiple is called the eigenvalue "λ." In other words, an eigenvector remains itself when subject to some transformation and hence can capture a basic source of variation. When more than one eigenvector is put together, they can constitute a basis to explain complex variations.

In general, an $n \times n$ dimension matrix can have a maximum of n eigenvectors. All the eigenvectors of a matrix are orthogonal to each other, no matter how many dimensions they have. This is important because it means that we can represent the data in terms of these perpendicular eigenvectors, instead of expressing them in terms of the original assets. This helps to reduce dimensionality of the problem at hand considerably, which characteristically for financial data is large.

2.5.2 PCA Versus Factor Analysis

Having understood the mathematical intuition behind PCA, we are in a position to appreciate why PCA is a dominant choice compared to factor analysis in financial data analysis. The objective of PCA is to be able to explain the variation in the original data with as few (and important!) dimensions (or components) as possible. Readers could argue that the dimensions can be reduced through factor analysis as well. Why use PCA instead?

To illustrate graphically, in Fig. 20.6 the red dots are the data points for a hypothetical time series of two dimensions. Factors 1 and 2 together explain a major portion of the variation in the data, and also those two factors put together fit the data better than PCA1 and PCA2, but Factors 1 and 2 are not orthogonal as can be seen in the graph above. Hence, their covariance matrix would have non-zero diagonal elements. In other words, Factors 1 and 2 would covary. Therefore, we would not only have to estimate the factors but also the covariance between them, which in a time series of 200 odd securities (>200 dimensions) can be onerous (>20,000 covariances!). When these covariances have to be dynamically modeled, the procedure becomes considerably inefficient in most financial data analysis.

In contrast, PCA 1 and 2 explain equally the variation in the data, and yet they have zero covariance. In the analysis of time series of yield curves, for example, Factors 1 and 2 are thought of as duration and convexity. These two factors help explain a lot of variation in yield curves across time, but they are not orthogonal

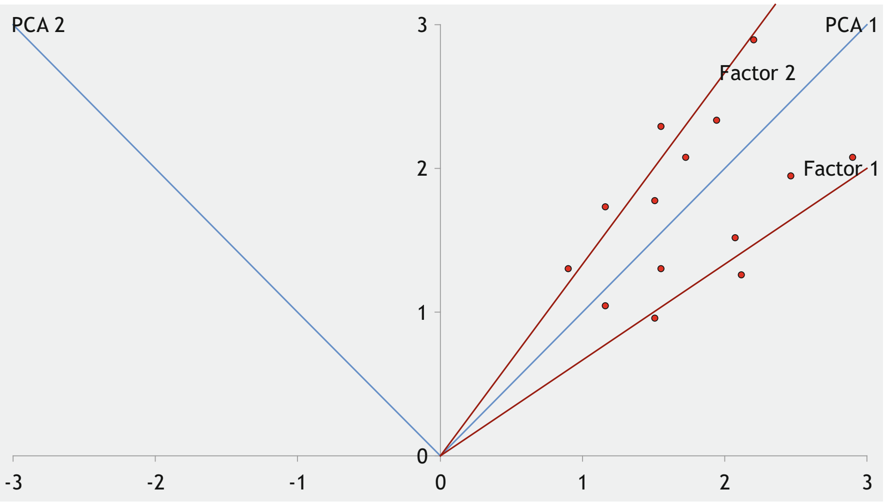

Fig. 20.6 Pictorial description of PCA and factor analysis

because convexity changes with duration. However, level of the yield curve and its slope are orthogonal components which can explain variation in bond prices equally well.

2.5.3 PCA Results

The results of the PCA for the portfolio of securities is detailed below. The first 21 principal components explain 60% nearly half of the variation as seen from Table 20.11.

Figure 20.7 shows that the first ten principal components capture the variation in the portfolio returns quite accurately. These ten principal components explain close to 50% of the variation in the security returns. As can be seen from Table 20.11, the first ten principal components explain 47.5% of the variation, while the next thirteen components explain less than 15% of the variation. Adding more principal components presents a trade-off between additional accuracy and the added dimensionality of the problem in most financial data analyses.

In the data in the portfolio that we study, the principal components from 11 onward each help explain less than 2% of the additional variation. However, adding one more principal component adds to the dimensionality by 10% and results in a commensurate increase in the computational complexity. Hence, we can limit to ten principal components for the subsequent analysis. This reduces the dimensionality of the data from 250 to 10.

As the histogram in Fig. 20.8 shows, the difference between the actual portfolio returns and the returns replicated using the ten principal components are, for the

Table 20.11 Contribution of various principal components

Principal components	Variance contribution	Cumulative variance
X1	16.03%	16.03%
X2	7.81%	23.85%
X3	5.11%	28.96%
X4	4.19%	33.15%
X5	3.50%	36.64%
X6	2.62%	39.27%
X7	2.38%	41.64%
X8	2.20%	43.85%
X9	1.96%	45.81%
X10	1.71%	47.53%
X11	1.64%	49.16%
X12	1.56%	50.72%
X13	1.36%	52.08%
X14	1.32%	53.40%
X15	1.23%	54.63%
X16	1.14%	55.76%
X17	1.07%	56.83%
X18	1.01%	57.84%
X19	0.99%	58.82%
X20	0.94%	59.76%
X21	0.93%	60.69%
X22	0.88%	61.57%
X23	0.86%	62.43%

most part, small. Hence, we can limit our subsequent analysis to ten principal components.

2.5.4 Stationarity of the Principal Components

To rule out spurious predictability in a time series, stationarity of the predicting variables is extremely important in most financial data analysis. For example, if a time series has a time trend to it, then it is rare that the time trend would repeat itself in the future. When a time series with a time trend is estimated, the time series would produce brilliant results (extremely high R-squared for the regression and high *t*-statistics for the coefficients of the predicting variables) leading a less careful data scientist to infer a well-fitted model capable of high levels of predictability. However, the predictability underlying such a time series is spurious and misleading. Thus, it is important to rule out such time trends in the predicting variables which are the principal components in our case. Figure 20.9 shows visually the absence of time trends in the ten principal components.

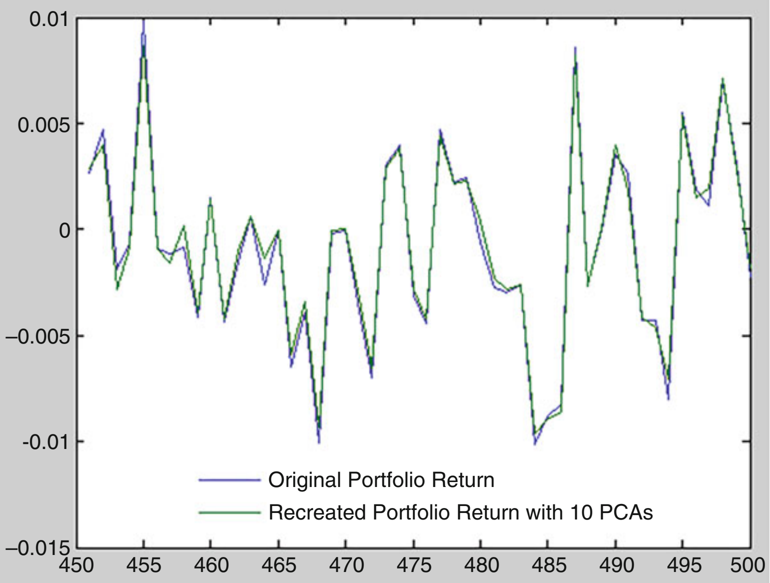

Fig. 20.7 Replicated time series using ten principal components vs original portfolio returns

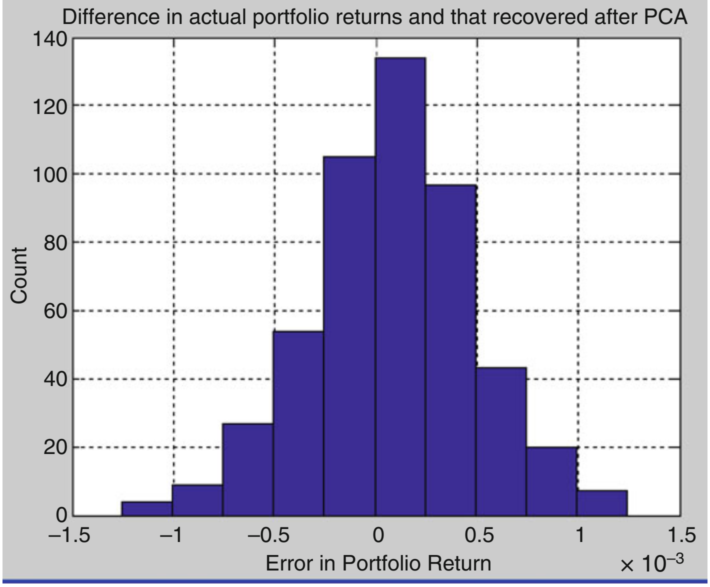

Fig. 20.8 Difference in actual portfolio returns vs PCA

Table 20.12 shows the results of the augmented Dickey–Fuller test (ADF) of stationarity for all the ten principal components. The ADF is used to test for the

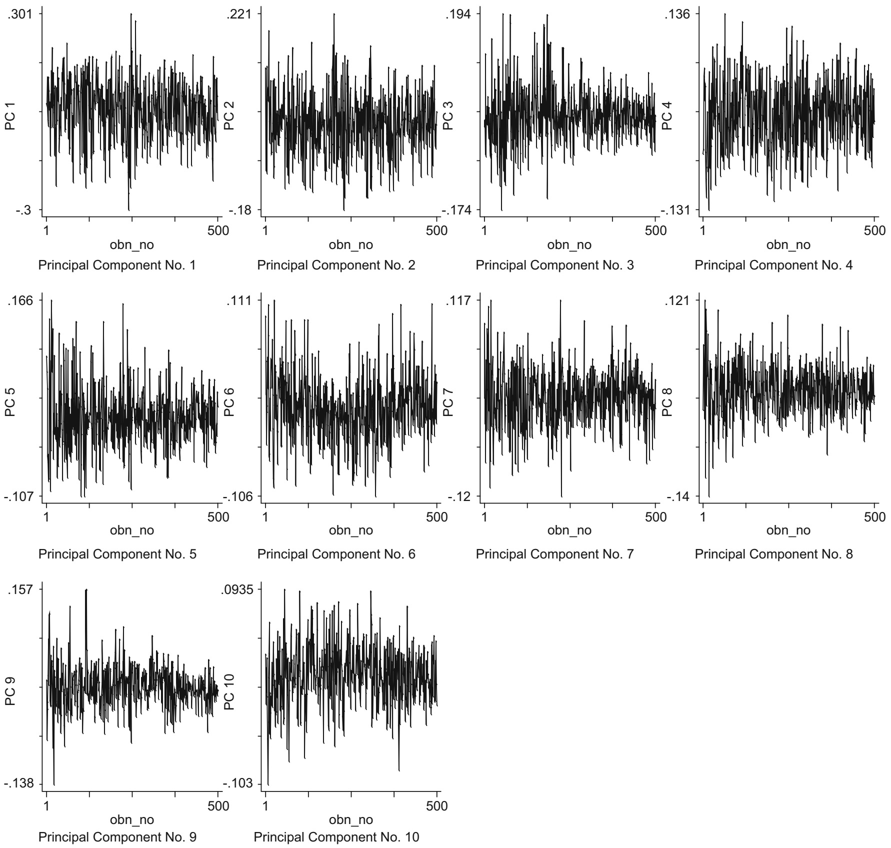

Fig. 20.9 Stationarity of principal components

presence of unit roots in the time series. If $y_t = y_{t\text{-}1} + e_t$, then the time series will blow up as the number of observations increases. Further, the variance of the time series will be unbounded in this case. In order to rule out the presence of unit roots, the augmented Dickey–Fuller test runs regressions of the following kind: $y_t - y_{t\text{-}1} = \rho\ y_t - {}_1 + e_t$. If the time series has a unit root, then ρ will be equal to zero. The ADF essentially tests the null hypothesis that $\rho = 0$ versus $\rho \neq 0$. As the results of the tests in Table 20.12 indicate, none of the principal components have a unit root. Also, as we examined earlier, they do not have a time trend either. So, predictability in the principal components is not spurious. This completes Step 2 of our framework in the chapter.

Let D.PCi indicate the first difference of the respective principal component. The absence of a unit root (which is the test of stationarity) is indicated by the coefficient of the lag of PC(i) being different from zero. Data scientists in the financial domain at times use the MacKinnon probability value to indicate the probability that the test statistic is different from the augmented Dickey–Fuller critical values.

Table 20.12 Augmented Dickey–Fuller test for stationarity of principal components

	D.PC1	D.PC2	D.PC3	D.PC4	D.PC5	D.PC6	D.PC7	D.PC8	D.PC9	D.PC10
Lag of PC(i)	−0.847*	−1.061*	−1.034*	−1.000*	−1.016*	−0.915*	−1.043*	−0.951*	−0.866*	−0.837*
	(19.11)	(23.76)	(23.06)	(22.31)	(22.73)	(20.62)	(23.28)	(21.24)	(19.49)	(18.91)
Constant	0.007	−0.003	0.004	−0.002	0.003	−0.008*	−0.001	0	0.008*	0.006*
	(1.80)	(0.89)	(1.85)	(0.74)	(1.44)	(4.57)	(0.74)	(0.18)	(5.51)	(4.38)
Obns	499	499	499	499	499	499	499	499	499	499
R-sqd	0.42	0.53	0.52	0.5	0.51	0.46	0.52	0.48	0.43	0.42
P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Absolute value of *t* statistics in parentheses; * indicates significance of coefficient at 1% level

2.6 Orthogonal GARCH

Given the large number of factors that typically affect the position of a large portfolio, estimating the risk of the portfolio becomes very complex indeed. At the heart of most data analytics models for estimating risk is the covariance matrix which captures the volatilities and the correlations between all the risk factors. Typically hundreds of risk factors encompassing equity indices, foreign exchange rates, and yield curves need to be modeled through the dynamics of the large covariance matrix. In fact, without making assumptions about the dynamics of these risk factors, implementation of models for estimating risk becomes quite cumbersome.

Orthogonal GARCH is an approach for estimating risk which is computationally efficient but captures the richness embedded in the dynamics of the covariance matrix. Orthogonal GARCH applies the computations to a few key factors which capture the orthogonal sources of variation in the original data. The approach is computationally efficient since it allows for an enormous reduction in the dimensionality of the estimation while retaining a high degree of accuracy. The method used to identify the orthogonal sources of variation is principal component analysis (PCA). The principal components identified through PCA are uncorrelated with each other (by definition they are orthogonal). Hence, univariate GARCH models can be used to model the time-varying volatility of the principal components themselves. The principal components along with their corresponding GARCH processes then capture the time-varying covariance matrix of the original portfolio. Having described principal component analysis and Orthogonal GARCH, we now illustrate the different variants of GARCH modeling.

2.7 GARCH Modeling

After the dimensionality of the time series is reduced using PCA, we now proceed to Step 3 of our framework in the chapter with modeling the covariance using GARCH on the principal components.

We first motivate the use of GARCH for measuring risk of a portfolio. Most common methodologies for estimating risk, through a Value at Risk calculation, assume that portfolio returns follow a normal distribution as shown in Fig. 20.10. This methodology of calculating VaR using normal distribution implictly assumes that the mean and standard deviation of the portfolio returns remain constant.

However, ample empirical evidence in finance shows that security returns exhibit significant deviations from normal distributions, particularly volatility clustering and fat tail behavior. There are certain other characteristics of equity markets which are not adequately accounted for in a normal distribution. Data scientists

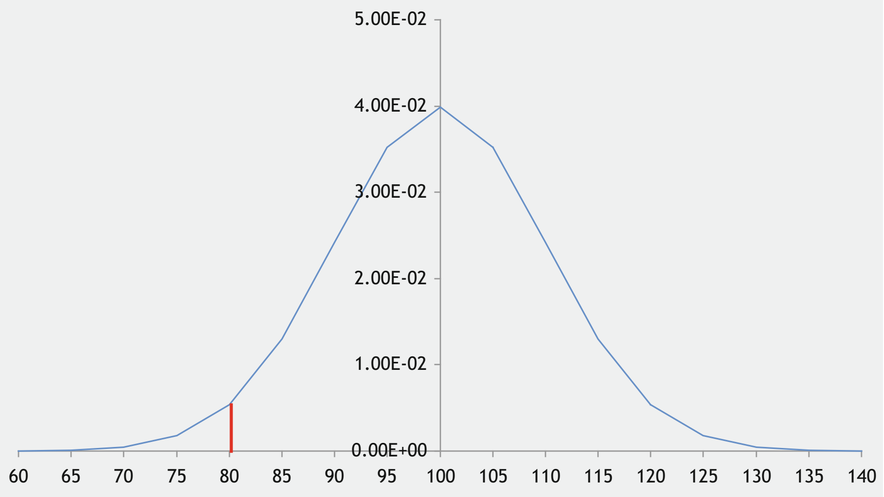

Fig. 20.10 Normal distribution

in finance therefore use GARCH models as they are devised to encapsulate these characteristics that are commonly observed in equity markets.

2.7.1 Volatility Clustering

Equity returns series usually exhibit this characteristic in which large changes tend to follow large changes and small changes tend to follow small changes. For instance, if markets were more volatile than usual today, there is a bias toward they being more volatile tomorrow than they typically are. Similarly, if markets were “quiet” today, there is a higher probability that they may be “quiet” tomorrow compared to they being unusually volatile. In both cases, it is difficult to predict the change in market activity from a “quiet” to a “volatile” scenario and vice versa. In GARCH, significant perturbations, either for good or for worse, are intrinsic part of the time series we use to predict the volatility for the next time period. These large perturbations and shocks, both positive and negative, persist in the GARCH model and are factored in the future forecasts of variance for future time periods. They are sometimes also called persistence and model a process in which successive disturbances, although uncorrelated, are nonetheless serially dependent.

An examination of the time series of principal components reveals that periods of high volatility are often clustered together. This has to be taken into account using a GARCH model.

Table 20.13 Shapiro–Wilk test

Variable	Shapiro–Wilk test statistic W	Prob. (normal)
pc1	0.99	0.01
pc2	1.00	0.63
pc3	0.97	0.00
pc4	1.00	0.39
pc5	0.98	0.00
pc6	0.99	0.00
pc7	1.00	0.12
pc8	0.99	0.00
pc9	0.97	0.00
pc10	1.00	0.44

2.7.2 Leverage Effects

Asset returns are often observed to be negatively correlated with changes in volatility. Meaning, markets tend to be more volatile when there is a sell-off vis-à-vis when markets rally. This is called leverage—volatility tends to rise in response to lower than expected returns and to fall in response to higher than expected returns. Asymmetric GARCH models are capable of capturing the leverage effect.

2.7.3 Fat Tails or Excess Kurtosis

The tail of distributions of equity returns are typically fatter compared to a normal distribution. In simple terms, the possibility of extreme fluctuations in returns is understated in a normal distribution, and these can be captured with GARCH models.

This lack of normality in our portfolio is tested by analyzing the distribution of the principal components using quantile plots as shown in Fig. 20.11. Fat tails are evident in the distribution of principal components as seen from the quantile plots since the quantiles at both the extremes deviate from the quantiles of a normal distribution. To further test whether the distributions for the principal components are normal, the Shapiro–Wilk test of normality is usually performed on all the principal components. The results of the Shapiro–Wilk test are provided in Table 20.13.

As is evident from Table 20.13, pc1, pc3, pc5, pc6, pc8, and pc9 exhibit substantial deviations from normality, while the remaining principal components are closer to being normally distributed. Since six of the ten principal components exhibit deviations from normality, it is important to model fat tails in the distribution of principal components. Figure 20.11 depicts quantiles of principal components plotted against quantiles of normal distribution (45% line). A look at the plot of the time series of the principal components reveals that periods of volatility are often clustered together. Hence, we need to take into account this volatility clustering using GARCH analysis.

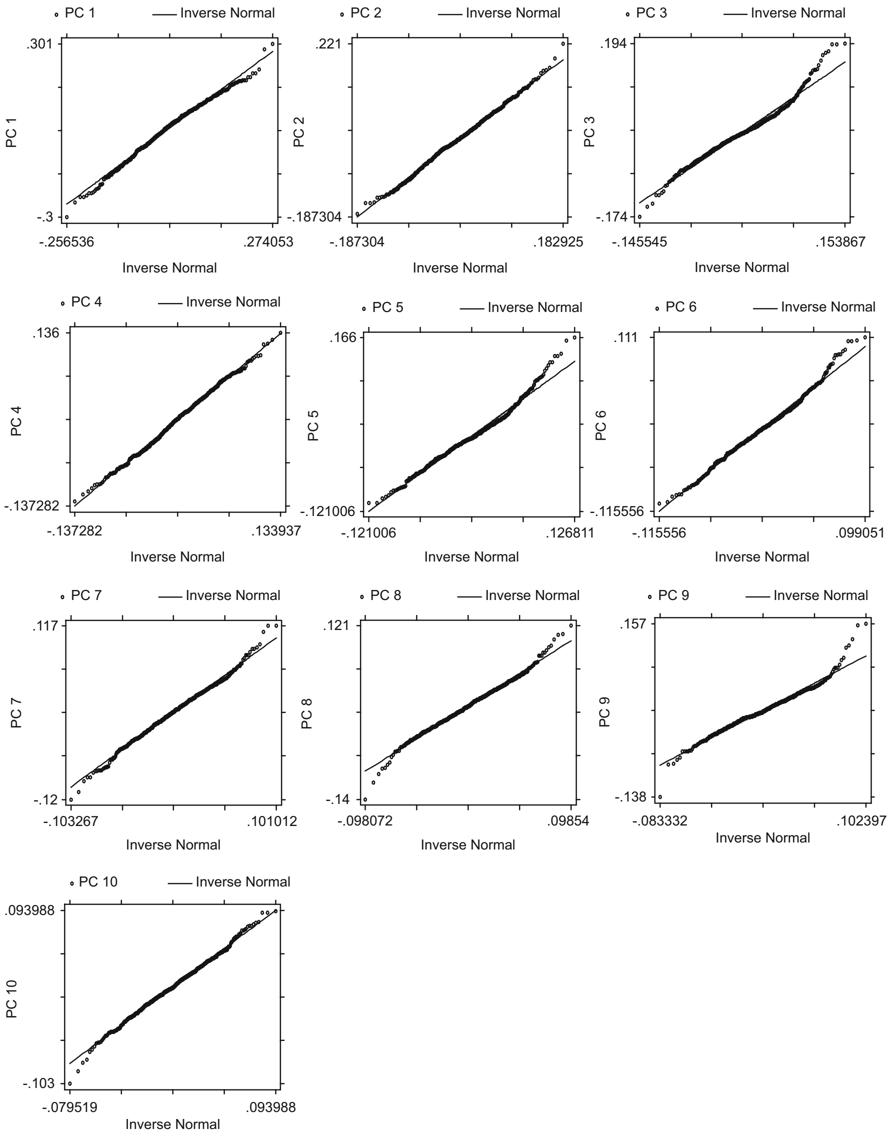

Fig. 20.11 Normality of principal components

Now that we have discussed why we use GARCH in financial data analysis, let us try to understand it conceptually. GARCH stands for generalized autoregressive conditional heteroscedasticity. Loosely speaking, one can think of heteroscedasticity as variance that varies with time. Conditional implies that future variances depend

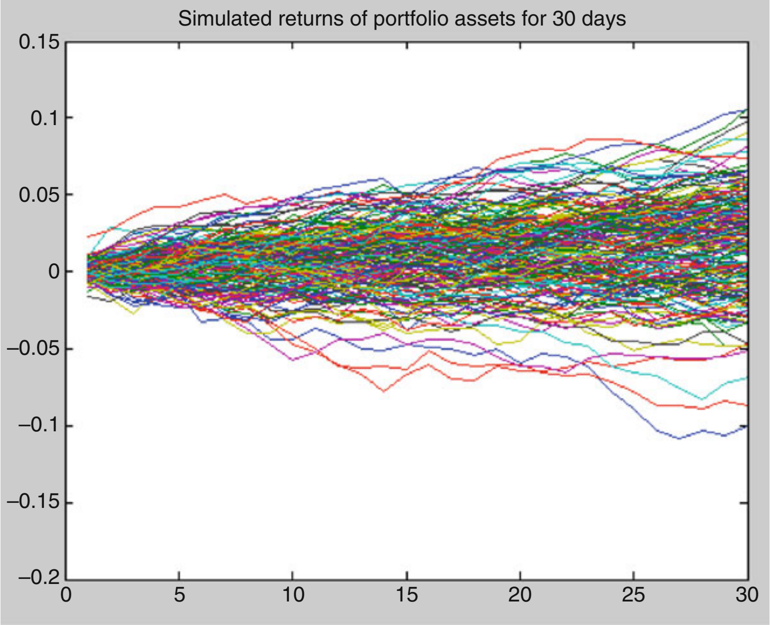

Fig. 20.12 Simulated returns of portfolio assets with GJR model, Gaussian distribution

on past variances. It allows for modeling of serial dependence of volatility. For the benefit of those readers who are well versed with econometric models and for the sake of completeness, we provide the various models used in the case study. Readers may skip this portion without losing much if they find it to be too mathematically involved.

2.7.4 Conditional Mean Model

This general ARMAX(*R*,*M*,*Nx*) model for the conditional mean applies to all variance models.

$$y_t = C + \sum_{i=1}^{R} \varphi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{M} \theta_j \varepsilon_{t-j} + \sum_{k=1}^{Nx} \beta_k X\,(t,k)$$

with autoregressive coefficients φ_i, moving average coefficients θ_j, innovations ε_t, and returns y_t. X is an explanatory regression matrix in which each column is a time series and $X(t,k)$ denotes the t^{th} row and k^{th} column.

In the case study, we implement the ARMA(1,1) model for all the principal components since a parsimonious model which captures both the autoregressive and moving average components of the conditional mean is desirable. Please see Chap. 12 for a brief description of the ARMA model and determining its parameters.

2.7.5 Conditional Variance Models

The conditional variance of the innovations, ${\sigma_t}^2$, is

$$\mathrm{Var}_{t-1}\left(y_t\right) = \mathrm{E}_{t-1}\left({\varepsilon_t}^2\right) = {\sigma_t}^2$$

The conditional variance quantifies the amount of variation that is left if we use the conditional expectation to predict y_t. The key insight of GARCH lies in the distinction between conditional and unconditional variances of the innovations process ε_t. The term *conditional* implies unambiguous dependence on a past time series of observations. The term *unconditional* is more concerned with long-term behavior of a time series and assumes no explicit knowledge of the past.

The various GARCH models characterize the conditional distribution of ε_t by imposing alternative parameterizations to capture serial dependence on the conditional variance of the innovations.

GARCH(*P*,*Q*) Conditional Variance

The general GARCH(P,Q) model for the conditional variance of innovations is

$${\sigma_t}^2 = K + \sum_{i=1}^{P} G_i \sigma_{t-i}^2 + \sum_{j=1}^{Q} A_j \varepsilon_{t-j}^2$$

GJR(*P*,*Q*) Conditional Variance

The general GJR(P,Q) model for the conditional variance of the innovations with leverage terms is

$${\sigma_t}^2 = K + \sum_{i=1}^{P} G_i \sigma_{t-i}^2 + \sum_{j=1}^{Q} A_j \varepsilon_{t-j}^2 + \sum_{j=1}^{Q} L_j S_{t-j}^{-} \varepsilon_{t-j}^2$$

where

$$S_{t-j}^{-} = \begin{cases} 1 & \varepsilon_{t-j} < 0 \\ 0 & \text{otherwise} \end{cases}$$

EGARCH(*P*,*Q*) Conditional Variance

The general EGARCH(P,Q) model for the conditional variance of the innovations with leverage terms and an explicit probability distribution assumption is

$$\log\left(\sigma_t{}^2\right) = K + \sum_{i=1}^{P} G_i \log\left(\sigma_{t-i}^2\right) + \sum_{j=1}^{Q} A_j \left[\frac{|\varepsilon_{t-j}\ |}{\sigma_{t-j}} - E\left\{\frac{|\varepsilon_{t-j}\ |}{\sigma_{t-j}}\right\}\right]$$
$$+\sum_{j=1}^{Q} L_j \left(\frac{\varepsilon_{t-j}}{\sigma_{t-j}}\right)$$

Models Used in the Case Study

We use the ARMA(1,1) model ($R = 1$, $M = 1$ in the equation for conditional mean) for conditional mean along with GARCH(1,1) and GJR(1,1) for our case study analysis. We employ the normal distribution and the Student's t-distribution to model the fat tails in the portfolio returns.

Although the above models are simple, they have several benefits. These represent parsimonious models that require estimation of at most eight parameters. The fewer the parameters to estimate, the greater the accuracy of these parameters. Complicated models in financial data analysis, more often than not, do not offer tangible benefits when it comes to predicting financial variables.

GARCH Limitations

While it is easy to be impressed with the mathematical expositions of the models, and the fact that GARCH models provide insights into a wide range of financial market applications, they do have limitations:

- The GARCH model at the end of the day is a parametric specification. The parameters remain stable only if the underlying market conditions are stable. GARCH models are good at capturing heterscedastic variances. That being said, they cannot capture tempestuous fluctuations in the market. Till now there does not exist a well-accepted model which can model market crashes as they are extremely unpredictable and unique.
- Asset returns have fat tails, i.e., large deviations from average are quite likely. GARCH models are not equipped to capture all of these fat tail returns that are observed in financial time series of returns. Time-varying volatility does explain a limited portion of this fat tail but, given the limitations of the normal distribution, cannot explain all of it. To offset for this constraint, data analysts more often than not implement Student's t-distribution in GARCH modeling.

2.8 Results

The calculation of Value at Risk for large portfolios presents a trade-off between speed and accuracy, with the fastest methods relying on rough approximations and the most realistic approach often too slow to be practical. Financial data scientists try to use the best features of both approaches, as we try to do in this case study.

Tables 20.14, 20.15, 20.16, 20.17, 20.18, 20.19, 20.20, and 20.21 show the calculation of Value at Risk and threshold persistence using four different models—GARCH with Gaussian distribution, GARCH with Student's *t*-distribution, GJR with Student's *t*-distribution, and EGARCH with Student's *t*-distribution. This is done for both the value-weighted portfolio and equi-weighted portfolio as below:

- Table 20.14 shows the calculation of Value at Risk and threshold persistence for market value-weighted GARCH model with a Gaussian distribution.
- Table 20.15 shows the calculation of Value at Risk and threshold persistence for market value-weighted GARCH model with a Student's *t*-distribution.
- Table 20.16 shows the calculation of Value at Risk and threshold persistence for market value-weighted GJR model with a Student's *t*-distribution.
- Table 20.17 shows the calculation of Value at Risk and threshold persistence for market value-weighted EGARCH model with a Student's *t*-distribution.
- Table 20.18 shows the calculation of Value at Risk and threshold persistence for equi-weighted GARCH model with a Gaussian distribution.
- Table 20.19 shows the calculation of Value at Risk and threshold persistence for equi-weighted GARCH model with a Student's *t*-distribution.

Table 20.14 Market value-weighted, GARCH, Gaussian

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−2.42723%	−2.40828%	−2.51901%	−2.43652%	−2.36555%
10	−3.48908%	−3.43362%	−3.36722%	−3.22902%	−3.27812%
20	−4.58238%	−4.71451%	−4.45732%	−4.47809%	−4.48455%
60	−7.28079%	−7.27208%	−7.30652%	−7.13126%	−7.28960%
125	−9.84899%	−9.21783%	−9.61347%	−9.95102%	−9.26511%
250	−11.90397%	−11.27439%	−12.28956%	−11.69521%	−11.58802%
	Percentage of times portfolio is below beta				
5	0.01000%	0.00000%	0.00000%	0.01000%	0.00000%
10	0.10000%	0.08000%	0.07000%	0.05000%	0.02000%
20	0.70000%	0.71000%	0.62000%	0.70000%	0.51000%
60	7.07000%	6.69000%	6.72000%	7.23000%	6.76000%
125	15.11000%	14.76000%	14.75000%	14.98000%	15.06000%
250	22.15000%	21.72000%	22.16000%	22.53000%	22.38000%
	Average drop in portfolio when level drops below beta				
5	−6.08684%	0.00000%	0.00000%	−5.99193%	0.00000%
10	−6.42005%	−5.82228%	−6.09030%	−5.79600%	−5.36812%
20	−6.19519%	−6.21074%	−6.33539%	−6.04260%	−5.97896%
60	−6.74850%	−6.73923%	−6.79577%	−6.65510%	−6.71048%
125	−7.48904%	−7.39024%	−7.52920%	−7.50248%	−7.34893%
250	−8.34098%	−8.25300%	−8.37886%	−8.26895%	−8.25156%

Table 20.15 Market value-weighted, GARCH, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−2.37854%	−2.39676%	−2.40350%	−2.38341%	−2.43340%
10	−3.34525%	−3.32570%	−3.40838%	−3.33190%	−3.32166%
20	−4.53003%	−4.56522%	−4.67216%	−4.49673%	−4.43257%
60	−7.28625%	−7.36338%	−7.19435%	−6.95804%	−7.09243%
125	−9.07297%	−9.50501%	−9.43058%	−9.32454%	−9.33755%
250	−11.33888%	−11.74615%	−11.71408%	−11.35302%	−11.60062%
	Percentage of times portfolio is below beta				
5	0.01000%	0.00000%	0.00000%	0.00000%	0.00000%
10	0.08000%	0.08000%	0.05000%	0.05000%	0.07000%
20	0.70000%	0.64000%	0.60000%	0.57000%	0.60000%
60	6.71000%	6.75000%	7.16000%	6.48000%	6.94000%
125	14.85000%	14.65000%	14.92000%	14.31000%	15.24000%
250	21.63000%	21.58000%	22.16000%	21.74000%	22.49000%
	Average drop in portfolio when level drops below beta				
5	−5.60456%	0.00000%	0.00000%	0.00000%	0.00000%
10	−6.17866%	−5.98915%	−5.65985%	−6.35440%	−6.02021%
20	−5.96225%	−6.06412%	−6.32215%	−6.26961%	−6.05488%
60	−6.71141%	−6.72747%	−6.74299%	−6.64512%	−6.64694%
125	−7.33353%	−7.37906%	−7.47784%	−7.33874%	−7.37931%
250	−8.15825%	−8.24734%	−8.27726%	−8.16879%	−8.22191%

Table 20.16 Market value-weighted, GJR, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−2.64120%	−2.67966%	−2.56970%	−2.72138%	−2.71597%
10	−3.82148%	−3.80847%	−3.73202%	−3.86924%	−3.82415%
20	−5.33825%	−5.41227%	−5.26804%	−5.26190%	−5.39392%
60	−8.50145%	−8.79663%	−8.54043%	−8.80623%	−8.73182%
125	−11.53589%	−11.68212%	−11.20231%	−11.61595%	−11.51297%
250	−15.96105%	−15.13785%	−14.87603%	−15.85603%	−14.83327%
	Percentage of times portfolio is below beta				
5	0.02000%	0.00000%	0.02000%	0.02000%	0.00000%
10	0.17000%	0.26000%	0.19000%	0.18000%	0.19000%
20	1.59000%	1.48000%	1.29000%	1.39000%	1.51000%
60	10.12000%	10.74000%	10.29000%	9.78000%	9.76000%
125	21.56000%	21.63000%	21.51000%	21.27000%	20.89000%
250	32.62000%	32.44000%	32.70000%	32.65000%	31.98000%
	Average drop in portfolio when level drops below beta				
5	−6.29212%	0.00000%	−6.46780%	−5.83254%	0.00000%
10	−7.07093%	−6.76246%	−6.49871%	−6.47768%	−6.24227%
20	−6.57822%	−6.59474%	−6.74424%	−6.55687%	−6.45223%
60	−7.25348%	−7.16626%	−7.14942%	−7.19248%	−7.19982%
125	−8.02682%	−7.96397%	−7.93543%	−7.98327%	−7.99186%
250	−9.18003%	−9.17600%	−9.04975%	−9.15641%	−9.10158%

Table 20.17 Market value-weighted, EGARCH, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−2.79077%	−2.85296%	−2.77644%	−2.69456%	−2.72527%
10	−3.93194%	−3.94460%	−3.83798%	−3.83507%	−3.87852%
20	−5.40504%	−5.28543%	−5.31962%	−5.33853%	−5.39825%
60	−8.58980%	−8.38608%	−8.37522%	−8.48768%	−8.71635%
125	−11.54664%	−11.21148%	−11.22397%	−11.52668%	−11.24505%
250	−14.78795%	−14.88799%	−14.44216%	−14.46602%	−14.44222%
	Percentage of times portfolio is below beta				
5	0.01000%	0.01000%	0.03000%	0.00000%	0.00000%
10	0.23000%	0.20000%	0.13000%	0.16000%	0.18000%
20	1.85000%	1.54000%	1.48000%	1.55000%	1.43000%
60	10.73000%	9.78000%	10.17000%	10.70000%	10.75000%
125	21.42000%	21.09000%	20.72000%	21.30000%	21.10000%
250	32.14000%	31.55000%	31.17000%	31.80000%	31.33000%
	Average drop in portfolio when level drops below beta				
5	−6.64370%	−6.70843%	−6.19737%	0.00000%	0.00000%
10	−6.16354%	−6.24120%	−6.27418%	−6.43948%	−6.07836%
20	−6.28625%	−6.43126%	−6.39376%	−6.50772%	−6.33898%
60	−7.08317%	−7.08325%	−7.08800%	−7.10921%	−7.19341%
125	−7.97817%	−7.88085%	−7.99144%	−8.00227%	−8.00791%
250	−9.05973%	−9.00021%	−9.05599%	−9.06243%	−9.00664%

Table 20.18 Equi-weighted, GARCH, Gaussian

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−3.02645%	−3.12244%	−3.10246%	−3.15969%	−3.08626%
10	−4.22080%	−4.34003%	−4.23744%	−4.39152%	−4.36571%
20	−5.77279%	−5.99089%	−5.89724%	−5.84402%	−6.00114%
60	−9.49140%	−9.72849%	−9.46143%	−9.57403%	−9.76131%
125	−13.00547%	−12.91606%	−13.16552%	−12.80902%	−13.03503%
250	−16.19272%	−17.10488%	−17.15668%	−16.31266%	−16.92225%
	Percentage of times portfolio is below beta				
5	0.01000%	0.01000%	0.01000%	0.01000%	0.04000%
10	0.27000%	0.45000%	0.29000%	0.38000%	0.36000%
20	2.24000%	2.77000%	2.45000%	2.50000%	2.63000%
60	14.98000%	14.71000%	14.69000%	14.50000%	15.09000%
125	26.76000%	26.31000%	26.23000%	26.49000%	27.08000%
250	35.85000%	35.38000%	36.05000%	35.87000%	35.99000%
	Average drop in portfolio when level drops below beta				
5	−6.30389%	−6.15856%	−5.35099%	−7.11069%	−5.63802%
10	−6.01000%	−6.16643%	−6.19892%	−6.46273%	−6.09211%
20	−6.35667%	−6.38641%	−6.30022%	−6.38837%	−6.44409%
60	−7.36676%	−7.40805%	−7.34800%	−7.37507%	−7.43454%
125	−8.36959%	−8.50622%	−8.45060%	−8.38819%	−8.41793%
250	−9.69286%	−9.85419%	−9.74382%	−9.63297%	−9.78937%

Table 20.19 Equi-weighted, GARCH, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−3.09375%	−3.14080%	−3.10339%	−3.07041%	−3.11094%
10	−4.36277%	−4.21170%	−4.38308%	−4.27252%	−4.39058%
20	−5.79307%	−5.92853%	−5.95257%	−5.78613%	−6.01026%
60	−9.49324%	−9.41315%	−9.67444%	−9.57806%	−9.44238%
125	−12.96079%	−13.42418%	−12.86986%	−13.34371%	−12.92972%
250	−16.55830%	−16.94585%	−16.56531%	−16.59782%	−16.18167%
	Percentage of times portfolio is below beta				
5	0.01000%	0.02000%	0.03000%	0.01000%	0.01000%
10	0.40000%	0.27000%	0.40000%	0.24000%	0.36000%
20	2.56000%	2.34000%	2.62000%	2.48000%	2.57000%
60	14.93000%	14.54000%	15.28000%	14.27000%	14.20000%
125	26.55000%	26.00000%	26.80000%	25.64000%	25.82000%
250	36.04000%	34.73000%	35.91000%	34.59000%	35.32000%
	Average drop in portfolio when level drops below beta				
5	−5.48361%	−6.01500%	−5.60246%	−5.88618%	−6.43174%
10	−6.00148%	−6.38804%	−6.26810%	−6.37318%	−6.13003%
20	−6.38101%	−6.47291%	−6.52129%	−6.41859%	−6.44267%
60	−7.31137%	−7.31075%	−7.43181%	−7.44125%	−7.33935%
125	−8.38976%	−8.49580%	−8.42735%	−8.48043%	−8.35742%
250	−9.67619%	−9.72951%	−9.68775%	−9.79860%	−9.54100%

Table 20.20 Equi-weighted, GJR, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−3.41649%	−3.37054%	−3.43304%	−3.35435%	−3.48986%
10	−4.89835%	−4.84086%	−4.86367%	−4.79133%	−5.00438%
20	−6.97760%	−6.83118%	−6.85524%	−6.72537%	−6.99981%
60	−11.39598%	−11.26982%	−11.43001%	−11.77783%	−11.78486%
125	−15.86655%	−16.52494%	−15.94209%	−15.45319%	−16.09934%
250	−20.67736%	−21.36881%	−22.31839%	−20.82328%	−21.46509%
	Percentage of times portfolio is below beta				
5	0.03000%	0.05000%	0.09000%	0.06000%	0.10000%
10	0.68000%	0.65000%	0.76000%	0.62000%	0.86000%
20	4.16000%	4.32000%	3.88000%	3.98000%	3.72000%
60	20.00000%	20.72000%	19.83000%	19.46000%	19.59000%
125	34.93000%	36.08000%	34.81000%	34.70000%	34.74000%
250	48.44000%	49.80000%	48.28000%	48.03000%	48.37000%
	Average drop in portfolio when level drops below beta				
5	−7.36444%	−7.72945%	−6.35243%	−5.81271%	−6.14997%
10	−6.56317%	−6.69730%	−6.58803%	−6.36042%	−6.67660%
20	−6.85652%	−6.77818%	−6.97909%	−6.77189%	−7.14428%
60	−7.87520%	−7.83506%	−7.91032%	−7.91935%	−7.88042%
125	−9.18330%	−9.18761%	−9.21253%	−9.17098%	−9.21382%
250	−10.85028%	−10.95301%	−11.02630%	−10.94284%	−10.95912%

Table 20.21 Equi-weighted, EGARCH, Student's *t*

Horizon	Value 1	Value 2	Value 3	Value 4	Value 5
	Value at Risk				
5	−3.57247%	−3.56529%	−3.50317%	−3.52430%	−3.41424%
10	−5.03825%	−4.94431%	−4.94004%	−4.95409%	−4.90723%
20	−6.80683%	−6.90082%	−6.80978%	−7.15053%	−7.04633%
60	−11.54241%	−11.56384%	−11.17854%	−11.59858%	−11.47896%
125	−16.02569%	−15.78560%	−15.09081%	−16.25870%	−15.86065%
250	−21.40973%	−21.42843%	−19.87466%	−20.98633%	−20.91555%
	Percentage of times portfolio is below beta				
5	0.05000%	0.05000%	0.03000%	0.04000%	0.05000%
10	0.87000%	0.74000%	0.67000%	0.83000%	0.76000%
20	4.18000%	4.10000%	4.18000%	4.42000%	3.96000%
60	19.67000%	20.03000%	20.09000%	19.69000%	20.00000%
125	33.91000%	34.78000%	34.46000%	34.42000%	34.52000%
250	47.81000%	48.03000%	46.65000%	47.48000%	47.87000%
	Average drop in portfolio when level drops below beta				
5	−6.45992%	−6.25018%	−7.06294%	−6.51616%	−6.25605%
10	−6.47289%	−6.46955%	−6.53686%	−6.39210%	−6.61915%
20	−6.77121%	−6.81445%	−6.71482%	−6.87959%	−6.89216%
60	−7.93299%	−7.95354%	−7.88962%	−7.96126%	−7.92641%
125	−9.23481%	−9.20317%	−9.10547%	−9.27140%	−9.19768%
250	−10.87354%	−10.92524%	−10.85239%	−10.95101%	−10.92249%

- Table 20.20 shows the calculation of Value at Risk and threshold persistence for equi-weighted GJR model with a Student's *t*-distribution.
- Table 20.21 shows the calculation of Value at Risk and threshold persistence for equi-weighted EGARCH model with a Student's *t*-distribution.

2.8.1 Value-Weighted vis-à-vis Equi-weighted

In general, the value-weighted portfolio has less dispersion than equi-weighted portfolio. This is to be expected because in general traders have a higher weightage for assets which have less volatility given similar expected returns. This is consistent with the results in the tables obtained for percentage of times the portfolio value hits the threshold level (−5.00%) and the average drop in the portfolio given that the portfolio hits this threshold. Both the values are lower for value-weighted portfolio vis-à-vis equi-weighted portfolio.

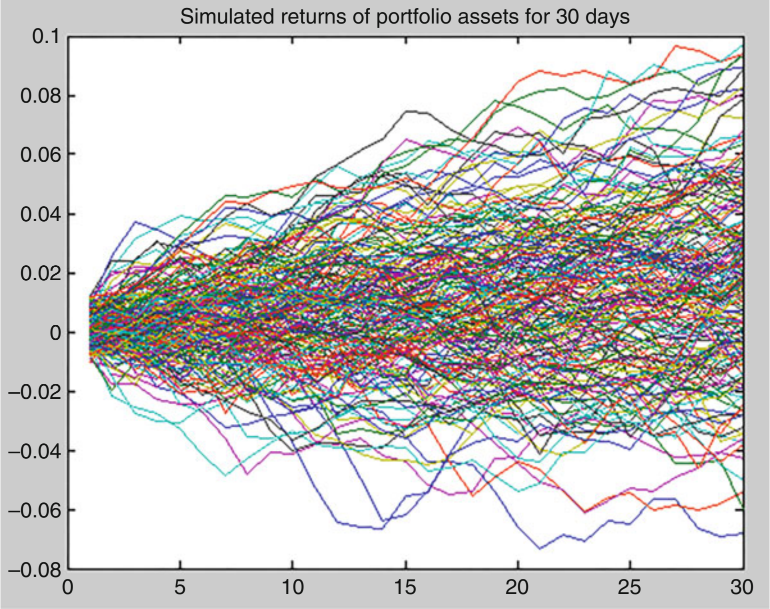

Fig. 20.13 Simulated returns of portfolio assets with GJR model, Student's *t*-distribution

2.8.2 Gaussian vis-à-vis Student's *t*-Distribution

For GARCH models, the difference in dispersion for the two distributions is small. However, the tables report the results for GJR and EGARCH using the two different-distributions, and, in general, as is to be expected, the Student's *t*-distribution tends to have a higher dispersion than in the case of Gaussian distribution.

Figures 20.11 and 20.12 show the simulated paths for the equal-weighted portfolio over the 30-day horizon with GJR model with a Gaussian distribution and Student's *t*-distribution. As is clearly visible, the fat-tailed Student's *t*-distribution generates greater variation at the extremes than the normal distribution.

2.8.3 GARCH, GJR, and EGARCH

GARCH tends to underestimate the VaR and persistence measure vis-à-vis GJR and EGARCH. Again this is to be expected given that GJR and EGARCH factor in the leverage effect which GARCH fails to do. GJR and EGARCH return similar results which again is to be expected.

Five values are exhibited for each parameter to show the measure of dispersion. Standard errors for the estimates are computed as also the *t*-statistics, and both are found to be statistically acceptable.

Each value itself is an average of 10,000 paths. The horizon is mentioned in the first column. The threshold horizon is taken as 2 days for consistency across horizon.

VaR and persistence measures are also computed for horizons of 2 years, 3 years, 4 years, and 5 years. The range of Value at Risk is between 20 and 22% for these horizons. The probability of the portfolio remaining below the threshold level β for 2 or more days is about 42–47%, whereas the average drop in the portfolio given that this happens is about 11–13%.

2.8.4 Principal Component Analysis Results

- Table 20.22 shows the PCA analysis for GARCH model with a Gaussian distribution.
- Table 20.23 shows the PCA analysis for GARCH model with a Student's t-distribution.
- Table 20.24 shows PCA analysis for GJR model with a Student's t-distribution.
- Table 20.25 shows the PCA analysis for EGARCH model with a Student's t-distribution.

2.9 Conclusion

The case study provides an overview of statistical portfolio risk analysis as is practiced in the investment management world. The approach employs econometric modeling of stock price returns and explains the econometric theory behind the application so as to make the chapter self-contained. The risk measurement is formulated using industry standard risk metrics such as Value at Risk and threshold persistence. We use robust methods that account for fat tails and leverage such as GJR and EGARCH to measure risk. One fundamental implication of data analysis for financial markets is that risk regimes change. So a GJR or EGARCH may be apt for this data set but may not be universally appropriate for risk measurement of other kinds of financial data. Since risk regimes change in abrupt and unforeseen ways, a P-quant needs to understand and communicate the assumptions and limitations of data analytics to consumers of risk reports. For instance, it may not be out of place to keep reminding the consumers of risk reports that worst outcomes like Value at Risk and threshold persistence look singularly at the extreme left tail of the portfolio loss distribution. They are therefore less tractable and stable than a simpler metric like variance that is computed over a long-time horizon.

That said, risk measurement and reporting in financial institutions, in general, has moved away from long descriptive type discussions to providing more quantitative information so that risk professionals can make their own assessment. Additionally, the frequency of reporting has changed significantly. Chief risk officers (CROs) in AMCs typically receive reports that contain VaR and threshold persistence estimates

Table 20.22 GARCH, Gaussian

	Coeff C	Coeff AR	Coeff MA	Coeff K	Coeff GARCH	Coeff Arch	Coeff leverage	Log likelihood estimate
PCA 1	0.006191	−0.037002	0.252324	0.002363	0.581073	0.150042	0.000000	482.661472
PCA 2	0.003177	0.019312	−0.090477	0.000727	0.673067	0.159143	0.000000	664.383809
PCA 3	0.000307	1.000000	−0.832486	0.000062	0.533216	0.327323	0.000000	1275.084082
PCA 4	0.001868	−0.099412	−0.037937	0.000177	0.880774	0.042033	0.000000	812.619731
PCA 5	0.000127	−0.781957	0.813060	0.000005	0.965923	0.028736	0.000000	941.624211
PCA 6	−0.000491	0.790801	−0.808419	0.000618	0.416907	0.092563	0.000000	961.156845
PCA 7	−0.000602	−0.254338	0.235336	0.000008	0.978600	0.013823	0.000000	967.958283
PCA 8	−0.007667	−0.942045	0.994696	0.000029	0.906187	0.063275	0.000000	1026.846244
PCA 9	0.003197	−0.802868	0.771076	0.000042	0.854067	0.103361	0.000000	1035.681025
PCA 10	−0.000173	0.935406	−0.996544	0.000607	0.240479	0.096681	0.000000	1042.554168

Table 20.23 GARCH, Student's *t*

	Coeff C	Coeff AR	Coeff MA	Coeff K	Coeff GARCH	Coeff ARCH	Coeff leverage	Log likelihood estimate
PCA 1	0.006368	−0.044536	0.258300	0.002355	0.582574	0.149677	0.000000	482.593216
PCA 2	0.003167	0.024280	−0.095032	0.000733	0.671451	0.159680	0.000000	664.281732
PCA 3	0.000307	1.000000	−0.831038	0.000062	0.532482	0.328004	0.000000	1274.948566
PCA 4	0.001870	−0.097163	−0.040524	0.000177	0.880851	0.042276	0.000000	812.438789
PCA 5	0.001090	−0.735297	0.769962	0.000010	0.955841	0.036009	0.000000	946.274938
PCA 6	−0.000612	0.774787	−0.794976	0.000619	0.418905	0.096472	0.000000	964.726132
PCA 7	0.000830	−0.987641	1.000000	0.000009	0.974729	0.017016	0.000000	971.144066
PCA 8	−0.007482	−0.946004	0.994906	0.000034	0.888511	0.078636	0.000000	1028.985630
PCA 9	0.003049	−0.810527	0.778840	0.000042	0.853885	0.103836	0.000000	1035.893574
PCA 10	−0.004129	−0.569986	0.622669	0.000544	0.337492	0.085532	0.000000	1042.300629

Table 20.24 GJR, Student's *t*

	Coeff C	Coeff AR	Coeff MA	Coeff K	Coeff GARCH	Coeff ARCH	Coeff leverage	Log likelihood estimate
PCA 1	0.003418	−0.025396	0.239486	0.001970	0.659505	0.000000	0.221961	488.076790
PCA 2	0.003857	0.067171	−0.149177	0.000968	0.614463	0.235522	−0.156928	665.908440
PCA 3	0.000358	1.000000	−0.827733	0.000061	0.558873	0.367855	−0.141102	1276.062104
PCA 4	0.001946	−0.101657	−0.035439	0.000184	0.875252	0.051476	−0.013321	812.486859
PCA 5	0.000893	−0.737373	0.771080	0.000009	0.958213	0.028510	0.010643	946.341024
PCA 6	−0.003602	−0.445124	0.440511	0.000620	0.420793	0.127517	−0.069612	964.670930
PCA 7	−0.000854	−0.320778	0.305520	0.000011	0.970576	0.010650	0.018427	969.812064
PCA 8	−0.006819	−0.946771	0.995566	0.000037	0.883832	0.108470	−0.054267	1029.653034
PCA 9	0.000362	0.783815	−0.822747	0.000043	0.851227	0.101817	0.006893	1036.241450
PCA 10	−0.004473	−0.588531	0.639748	0.000432	0.455036	0.050880	0.075837	1042.590481

Table 20.25 EGARCH, Student's *t*

	Coeff C	Coeff AR	Coeff MA	Coeff K	Coeff GARCH	Coeff ARCH	Coeff leverage	Log likelihood estimate
PCA 1	0.003997	−0.064723	0.281886	−1.036871	0.782946	0.185203	−0.163616	484.578627
PCA 2	0.003834	0.092554	−0.174779	−0.994841	0.820174	0.252133	0.084304	667.115909
PCA 3	0.000333	1.000000	−0.830915	−0.415517	0.947283	0.386928	0.067537	1272.729986
PCA 4	0.002047	−0.094801	−0.046898	−0.443169	0.927416	0.105747	0.032806	813.616259
PCA 5	0.000785	−0.758420	0.793959	−0.001776	1.000000	0.041367	−0.003358	945.009720
PCA 6	−0.000627	0.773375	−0.792651	−3.274587	0.510104	0.231433	0.035319	965.878475
PCA 7	−0.000004	0.982168	−1.000000	−0.060656	0.991111	0.036622	0.001679	971.161642
PCA 8	−0.006313	−0.944539	0.995196	−0.282069	0.959395	0.154341	0.021652	1028.909113
PCA 9	0.000353	0.759004	−0.806882	−0.374390	0.946651	0.220938	−0.018483	1034.961690
PCA 10	−0.004564	−0.590409	0.646866	−2.425124	0.652703	0.153982	−0.082664	1043.202155

for 1-Week, 1-Month, and 1-Year horizon. For instance, from Table 20.14 the CRO can infer that if stock returns follow approximately Gaussian distribution, there is a 1% chance (Value at Risk at 99% confidence level) that the portfolio might lose around 12% of its value over a 1-year horizon. Using the metric of threshold persistence, the CRO can infer that over a 1-year horizon, there is a 22% chance of the portfolio dipping below 5%. And given that such a dip happens over the critical time period of over 2 days, the drop in the portfolio value would be approximately 8%. The other tables quantify risk of portfolio when asset returns have excess kurtosis or when there are causal mechanisms at play between returns and volatility such as leverage effects.

Some CROs and investment management boards prefer to receive only summary risk reports. The summary report is typically short so as to make it less likely that the risk numbers will be missed by the board members. Most P-quant CROs choose to receive both the summary and detailed risk reports. It is not usual for the modern-day CROs to receive daily MIS (management information system) reports that contain analysis from Table 20.14 to Table 20.21 on a daily basis. In the last few years, most CROs come from the P-world and are quantitatively well equipped to understand and infer risks from the detailed risk reports.

Apart from the senior management and the board, the other principal audience of risk reports are regulators. Regulators like the Fed and the RBI mandate all financial institutions that they regulate to upload their risk reports in a prescribed templete at the end of each business day. Regulators normally prescribe templates for risk reporting so that they can do an apples-to-apples comparison of risk across financial institutions. Regulators themselves use systems to monitor the change in risk of a given financial insitution over time. More importantly, it helps them aggregate risk of all financial institutions that they regulate so as to assess the systemic risk in the financial industry. With rapid advances in data sciences, it is envisaged that application of analytics in finance would get increasingly more sophisticated in times to come.

Electronic Supplementary Material

All the datasets, code, and other material referred in this section are available in www.allaboutanalytics.net.

- Data 20.1: Financial Analytics Steps 1, 2 and 3.xlsx
- Data 20.2: Financial Analytics Steps 4 and 5.xlsx
- Data 20.3: Financial Analytics Step 6.xlsx
- Data 20.4: Financial Analytics Step 7.xlsx
- Data 20.5: Financial Analytics Step 8.xlsx
- Data 20.6: nifty50.txt
- Data 20.7: ptsr.txt
- Data 20.8: randomgaussian.txt
- Data 20.9: randomgaussiancurrency.txt
- Data 20.10: tsr.txt

Exercises

Ex. 20.1 Stage I Step 3 Inference: We consider the National Stock Exchange index NIFTY-50 recorded daily for the period November 1, 2016–October 31, 2017. Let p_t be the NIFTY-50 index and r_t be the log return $\{r_t = \log(p_t) - \log(p_{t-1})\}$. Load the data from the file nifty50.txt into Matlab.

- Draw graphs of the stock index, the log returns, and the squared log returns.
- Do the graphs indicate GARCH effects?
- Estimate the GARCH parameters (ω,α,β).

Ex. 20.2 Stage I Step 4 Projection: Load the data on random numbers from randomgaussian.txt into Matlab. The standard deviation is 20%, while the average return is 5%. Note that the average return and standard deviation should be adjusted for daily horizon by dividing with 365 and square root of 365, respectively. Project the value of USD/INR for a horizon of 1 year.

Ex. 20.3 Stage II Step 2 Aggregation: Assume a loan portfolio of Rs. 500 lent to five different corporates for Rs. 100 each. Aggregate the risk of this Rs. 500 loan portfolio using one-factor Gaussian copulas. Assume each corporate has a probability of default of 4%. The horizon for the loan is 1 year. Assume that in the event of a default the bank can recover 75% of the loan amount. Assume the single factor to be the economy. The correlation of each firm's asset to the economy is given in the table below. Calculate the joint distribution of credit of each of these corporates using a one-factor model.

	Corporate_1	Corporate_2	Corporate_3	Corporate_4	Corporate_5
C_i	1.00%	1.00%	1.00%	1.00%	1.00%
D_i	75.00%	75.00%	75.00%	75.00%	75.00%
a_i	0.200000	0.400000	0.500000	0.600000	0.800000
X_i	4.00%	4.00%	4.00%	4.00%	4.00%

Ex. 20.4 Stage II Step 2 Assessment: Load the data on random numbers from randomgaussiancurrency.txt into Matlab. Compute VaR for a portfolio of 1 USD, 1 EUR, 1 GBP, and 100 JPY. The value of the portfolio in INR terms is Rs. 280 (1 USD = Rs. 64, 1 EUR = Rs. 75, 1 GBP = Rs. 82, 100 JPY = Rs. 59). Calculate the possible loss or gain from this portfolio for a 1-year horizon. To aggregate the risk, use the correlation matrix below between the currencies:

	USD/INR	EUR/INR	GBP/INR	JPY/INR
USD/INR	1	0.9	0.6	0.4
EUR/INR	0.9	1	0.5	0.5
GBP/INR	0.6	0.5	1	0.2
JPY/INR	0.4	0.5	0.2	1

Ex. 20.5 Stage II Step 3 Attribution: A bank comprises three lines of businesses:

- Line of Business 1 (LoB1)—Corporate Banking
- Line of Business 2 (LoB2) —Retail Banking
- Line of Business 3 (LoB3) —Treasury Operations

LoB1 has a correlation of 0.5 with LoB2 and has a correlation of 0.2 with LoB3. LoB2 is uncorrelated to LoB3. The total bank assets are Rs. 6000 crores. Each of the LoBs has assets worth Rs. 2000 crores.

	Assets	Volatility
LoB1	2000	$\sigma_1 = 5\%$
LoB2	2000	$\sigma_2 = 7\%$
LoB3	2000	$\sigma_3 = 9\%$

(a) Determine the total economic capital for the bank.
(b) Attribute the capital consumed by LoB1, LoB2, and LoB3 on a stand-alone basis.
(c) Attribute the capital consumed by LoB1, LoB2, and LoB3 on incremental basis.
(d) Attribute the capital consumed by LoB1, LoB2, and LoB3 using a component approach.

Ex. 20.6 For multiple horizons of 5, 10, 20, 60, 125, and 250 trading days, for a market value-weighted portfolio, calculate the following:

(a) The Value at Risk of the portfolio using GJR model with Gaussian distribution
(b) The Value at Risk of the portfolio using EGARCH model with Gaussian distribution
(c) The percentage of times portfolio is below beta using GJR model with Gaussian distribution
(d) The percentage of times portfolio is below beta using EGARCH model with Gaussian distribution
(e) The average drop in portfolio when level drops below beta using GJR model with Gaussian distribution
(f) The average drop in portfolio when level drops below beta using EGARCH model with Gaussian distribution

Ex. 20.7 Repeat Exercise 20.6 for an equi-value weighted portfolio.

References

Ali, M. M., Mikhail, N. N., & Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis, 8*, 405–412.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy, 81*(3), 637–654.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.

Box, G., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control*. Upper Saddle River, NJ: Prentice-Hall.

Campbell, J. Y., Lo, A. W., & MacKinlay, A. C. (1997). *The econometrics of financial markets*. Princeton: Princeton University Press.

CreditMetrics. (1999). *Technical document* (1st ed.). New York: J.P. Morgan.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4), 987–1007.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*, 3.

Fama, E. F. (1965). Random walks in stock market prices. *Financial Analysts Journal, 21*(5), 55–59.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance, 25*(2), 383–417.

Fermi, E., & Richtmyer, R. D. (1948). Note on census-taking in Monte Carlo calculations. *LAM, 805*(A).

Frenkel, J. A., & Levich, R. M. (1981). Covered interest arbitrage in the 1970s. *Economics Letters, 8*(3).

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis, 1*, 15–30.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), 1029–1054.

Harrison, J. M., & Pliska, S. R. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and their Applications, 11*(3), 215–260.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77–91.

Merton, R. C. (1969). Lifetime portfolio selection under uncertainty: The continuous-time case. *The Review of Economics and Statistics, 51*(3), 247–257.

Owen, J., & Rabinovitch, R. (1983). On the class of elliptical distributions and their applications to the theory of portfolio choice. *Journal of Finance, 38*, 745–752.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (p. 994). Cambridge: Cambridge University.

RiskMetrics. (1996). *Technical document* (4th ed.). New York: J.P. Morgan.

Ross, S. A. (1978). Mutual fund separation in financial theory—The separating distributions. *Journal of Economic Theory, 17*, 254–286.

Ross, S. (1976). The arbitrage theory of capital asset. Pricing. *Journal of Economic Theory, 13*, 341–360.

Sharpe, W. (1964). Capital asset prices: A theory of market. equilibrium under conditions of risk. *Journal of Finance, 19*, 425–442.

Sims, C. (1980). Macroeconomics and reality. *Econometrica, 48*(1), 1–48.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris, 8*, 229–231.

Tobin, J. (1958). Liquidity preference as behaviour towards risk. *Review of Economic Studies, 25*, 65–86.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London, Series A, 226*, 267–298.