Portfolio Statistics and Optimization

chris bemis

August 11, 2023

Chapter 1

Introduction

Equity anomalies are abundant. This statement is both terribly, and purposely, imprecise. What defines an equity anomaly? In what sense are they anomalous? And just how abundant is abundant?

Eugene Fama in 1970 wrote *Efficient Capital Markets* [9], a pronouncement that the stock market efficiently disseminates every piece of information available about every stock that is traded into their respective prices. Asserting that no amount of analysis can separate future winners from losers, for, all available information has already been incorporated into the current price.

Broadly speaking, the above is the efficient market hypothesis (EMH). In the EMH world, no amount of fundamental analysis (reading balance sheets, income statements, and cash flow statements) or technical analysis (identifying patterns, real or perceived) will generate excess returns in the long run. Any contradiction to this is an anomaly. Again we are left with an imprecise statement as to what is meant by excess returns.

Consider, for example, the chart in Figure 1. Here a one dollar investment is made in each of two securities; the blue line being a market proxy, Standard and Poor's S&P 500 Index. At the end of a ten year period, there is clear 'excess.' But to what end? There is, even without a proper statistical analysis something unappealing in the wild swings and variations in the purported winner. Our sense of caution is not wrong, either, since in this case the winner is simply a leveraged version of the market–returns are $2\times$ the un-levered version. This surely can't be what is meant by the impossibility of generating excess returns, and in fact it isn't.

However, a somewhat technical treatment is needed. Referring to the figure once more, there is something *riskier* in the leveraged series. Defining risk and its requisite properties has become a seminal issue over the past few decades. At the outset, volatility was tantamount to risk. And as we will see in subsequent chapters, much of the theoretical machinations supporting the efficient market hypothesis will imply that excess returns are a positive function of risk (read volatility). The example above notwithstanding, this is antithetical to what is observed empirically and what will be shown in subsequent chapters. In fact

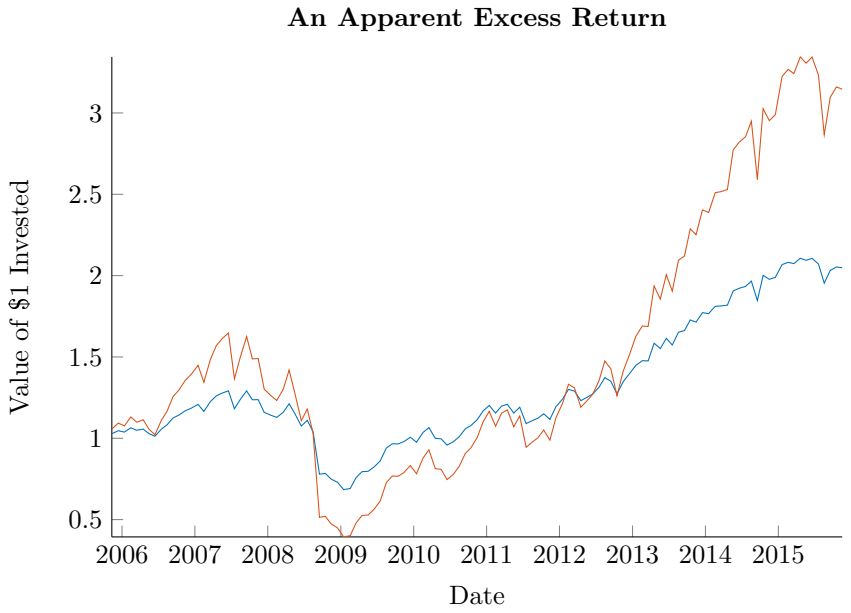

Figure 1.1: Value of $1 invested in the S&P 500 [25], including dividends and corporate actions, over a ten year period in blue. The red line shows the value of $1 invested in a related security. Performance is superior, but not without drawbacks.

we will exhibit that the market, rather than compensating for risk in the equity markets, punishes risk-holders across almost every measurement.

Staying within the confines of the EMH world, then, what should one do if choosing a good stock is proscribed by theoretical fiat? Buy them all, of course! One of the pivotal contributions to the field in fact belies this very statement. At its heart is a treatment of systemic and idiosyncratic risk which culminates in a familiar refrain: you can't beat the market, and diversification is the key to reducing idiosyncratic risk.

On the one hand, then, we have a yet-to-be defined model for determining anomalous returns, and on the other, an again yet-to-be defined model to reduce risk across a set of securities.

We will often attempt to gain a deeper understanding of the central themes of modern finance through the lens of the so-called *value* anomaly. We do this for two reasons: first, we choose to take most market standard models as providing a taxonomy rather than as prescriptive of actual market behavior; and second, while much of our work centers on using and utilizing statistical features of past price performance, we lean towards a belief that these do not encapsulate all of the information in the market. To quote Andy Redleaf, the founder of Whitebox Advisors, in his book, *Panic* [26]:

> *A market in which traders can predict future price moves solely on the basis of past price moves* ***is*** *pathetic. Traders can only do this because public financial markets are price-obsessed and price-paranoid. Financial markets are awash in price information (how many times a second does some price on some board somewhere change?) and relatively devoid of knowledge about value and cost to compare it to. Traders are desperate to know what everyone else is paying precisely because there is so little information about what anyone* ***should*** *be paying.*

We arrive next at the focus of the present work: statistics and optimization, with an emphasis on interpretation and allocation. The formulation of the efficient market hypothesis in this book will be a statistical one, with some modification of the standard literature. While we will focus primarily on Merton and Sharpe's Capital Asset Pricing Model [22, 31] and the subsequent Fama-French three factor model [10], we will often eschew their language of risk. Instead, we will use these two models much as the Black-Scholes' option formula is in practice: for an interpretive tool to understand asset allocation and for uniform comparisons.

The Capital Asset Pricing Model (or simply CAPM) is, in essence, simply a relationship between a given stock's returns and the market's returns (modulo something called the risk-free rate). Formally it says

$$r_t - r_f = \beta(m_t - r_f) + \epsilon_t,$$

or that the excess return for a given asset at time t, r_t, over the risk free rate, r_f, is a constant multiple β times the excess return of the market, m_t, over the risk free rate, with some allowance for random variation, ϵ_t. There are several theoretical consequences to the model, many of which will be discussed later. Presently, though, we obtain that the model gives a single variable by which to compare assets.

Returning to our first example of a leveraged market portfolio outperforming the market portfolio, we may, without calculations conclude that the former had a β of 2, while the latter had a β equal to 1, capturing quite nicely the intuition used to construct the example. The effects of leverage, correlation and risk-as-volatility are summarily presented in one metric, allowing uniform comparisons and a common language across any number of securities.

Caveats abound in the model above, however. Notably that equity returns are not stationary; i.e., the statistical properties needed to obtain a quantity like β above are not the same over various time slices. Calculating a stock's β from 1990 to 2000 will be different than from 2000 to 2010. Or worse, six months prior to, say, the financial crisis of 2008 will differ from six months after. Worse yet, even as we will develop the framework for understanding the statistical distribution of a variable like β (noting that in fact most financial data should be taken as ranges and not point estimates), the movements of the market can violate the assumptions for well-behaved estimators. Put another way, even if

we make an expected range for β, the market's moves could go outside this range in a heartbeat.

As David Viniar, the CFO at Goldman Sachs during the Quant Crisis of August 2007 said, "[We saw] things that were 25-standard deviation moves, several days in a row." [15] We aren't yet to the point where we can evaluate the sheer magnitude of this statement of improbability, but suffice it to say, it'll defy comprehension on a universal scale.

The empirical facts of the last thirty years drive our emphasis of interpretive abilities rather than on the normative modeling that has been a feature of mathematical finance since its inception. To quote Karl Popper

> No amount of observations of white swans can allow the conclusion that all swans are white, but the observation of a single black swan is sufficient to refute the conclusion.

There have been a few black swans of note, to the point that they deserve primacy in our applications of financial models. The trend within the field seems to be to accommodate new findings in ever expanding complexity, however.

An example–but one that is meritus–is the Fama-French extension to CAPM [10]. Here, finding that small companies outperform larger companies, and that higher value stocks outperform lower, they expand CAPM to

$$r_t - r_f = \beta_m(m_t - r_f) + \beta_h h_t + \beta_v v_t + \epsilon_t,$$

with h_t being the returns to a long-short portfolio that is long small companies and short large, and v_t similarly constructed from high book-to-price companies versus low book-to-price. Not surprisingly, the anomalies (relative to CAPM) that Fama and French noted are subsumed in their model, and *ipso facto*, size and value are no longer anomalies.

An implication in the standard literature is that these new factors are *risk factors*. Our treatment will not subscribe to this view as there is little empirical evidence to justify the term. Instead, we note the power of understanding an asset or a weighted portfolio of assets in terms of highly meaningful and understandable exposures within the market. We will also see that models like Fama-French and CAPM will lend themselves readily to portfolio allocation models where understanding the co-movements of assets is rife with estimation error.

Once we have gained an understanding of several statistical methods through the lens of mathematical finance, we will be in a position to study portfolio optimization. After developing some of the main tools for unconstrained and constrained optimization, we will find a direct connection between CAPM and Modern Portfolio Theory: Merton's 1952 Mean-Variance Optimization (MVO) [22],

$$\begin{aligned} \min_w \quad & \frac{1}{2} w'\Sigma w \\ & \mu' w = \mu^* \\ & 1' w = 1. \end{aligned}$$

where w is a vector of portfolio weights, Σ is the covariance of returns for the assets in question, μ is their expected return, and μ^* is some required return. No parameters are given from on high, however, so that while Σ and μ are neatly expressed above, obtaining reasonable estimates that are *stable through a future time horizon* is no easy task. Rather, for most models presented, the necessary stationarity requirements are assumed, in contrast to what is observed in the markets.

A proper treatment of the allocation problem in modern portfolio theory therefore necessitates a statistical treatment as well – as has already been alluded to vis a vis the previous model discussions. Here, understanding the structure of the covariance of returns becomes paramount, and we will visit and revisit this topic throughout the text.

With a solid underpinning of statistics and Modern Portfolio Theory in place, we will also consider more recent developments in portfolio optimization; namely, coherent measures of risk, and, as will be the tone of the text, generally, develop tools for implementing several such measures in practice.

Our approach in every topic will be to arduously highlight axioms before obtaining results, with the belief that implemented models require the practitioner to understand exactly which of the underlying assumptions are wrong (but perhaps useful), and which are dangerous, and we strive to develop the tools to do just this in the pages that follow.

Chapter 2

Distributions and Summary Statistics

The normal distribution (and its close cousin the log-normal distribution) dominate much of the financial landscape. One feature we believe should be front and center in understanding these objects is that in any discussion where investors make decisions solely on expected return and variance, there is a belief that other statistical shape measures aren't of concern. In this chapter, we work with density functions to better understand a few distributions that are useful for our future work, common in industry and the literature, and which may be useful in avoiding a myopia of expectation and variance as sole features for inspection.

We begin by briefly discussing random variables, and develop the mathematics of cumulative and probability density functions with examples [6]. We look at various statistics including the mean, variance, and percentiles of a cumulative density function, esp. the median, and then continue to example distributions. In so doing, we will compare estimated probabilities based on various density functions to empirical frequencies observed in equity returns, and note many so-called stylized features of equity returns as discussed by Rama Cont [8]; e.g., fat-tails, asymmetry of extreme gains and losses

Afterwards, we will show the impact of modifying a random variable on its density function as well, and prove the weak law of large numbers. Finally, we will work with multivariate random variables and generalize previous results as needed. The concepts of independence and correlation are also developed.

2.1 Univariate Distributions

We focus on real valued probability spaces, $\mathbb{P}$, and begin with the univariate case

$$(\mathbb{R}, \mathcal{B}, \mathbb{P})$$

with $\mathcal{B}$ the Borel σ algebra on $\mathbb{R}$, the real line. While the formalism of the statement is necessary, the usage will be limited, with the focus instead on methods that are much closer to calculus and real analysis.

Next, let X be a random variable on $(\mathbb{R}, \mathcal{B}, \mathbb{P})$, then the *cumulative distribution function* of X is given by

$$F(x) = \mathbb{P}\left(X \in (-\infty, x]\right). \tag{2.1}$$

We will assume that there are no point masses for F. The cumulative distribution function captures all of the distributional information about the random variable X; e.g.,

$$\begin{aligned} \mathbb{P}(X \in (a, b]) &= \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a), \\ \mathbb{P}(X > a) &= 1 - \mathbb{P}(X \leq a) = 1 - F(a), \end{aligned}$$

and if X is symmetric about 0, for instance, and $a > 0$, $\mathbb{P}(|x| > a) = 2 \cdot F(-a)$. In this case we also have

$$\begin{aligned} F(0) &= \frac{1}{2}, \\ F(x) &= 1 - F(-x). \end{aligned}$$

Notice, too, that we may consider the impact of shift and scale on the cumulative distribution function; e.g., let $Y = aX + b$ for $a > 0$ and $b \in \mathbb{R}$, and let F_Y and F_X be the cumulative distribution functions of Y and X, respectively, then

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(aX + b < y) \\ &= \mathbb{P}\left(X < \frac{y - b}{a}\right) \\ &= F_X\left(\frac{y - b}{a}\right). \end{aligned}$$

The case with $a < 0$ is treated similarly.

We say that X has a *probability density function*, $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$, if

$$F(x) = \int_{-\infty}^{x} f(s)ds. \tag{2.2}$$

At times we will simply refer to f as the density function or pdf of X. We note immediately that for f to be a density function, it is necessary that f is nonnegative, integrable, and satisfies

$$\int_{-\infty}^{\infty} f(s)ds = 1.$$

For X with a probability density function, the *expected value* of X, denoted $\mathbb{E}(X)$, is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} sf(s)ds. \tag{2.3}$$

Notice that we have not precluded infinite expectations so that $\mathbb{E}(X) \in \mathbb{R} \cup \infty$. We will often call the expected value of X the mean and denote it by μ.

The *variance* of X, $Var(X)$, is found by integrating the squared distance from the mean according to the density f. That is,

$$Var(X) = \int_{-\infty}^{\infty} (s-\mu)^2 f(s)ds \tag{2.4}$$

if μ is finite, otherwise $Var(X)$ is infinite as well. Here again, then, we have not disallowed infinite variances. For reasons that will become clear shortly, we routinely denote the $Var(X)$ as σ^2.

We may also define variance without resort to the density function of X, via

$$Var(X) = \mathbb{E}\left((X-\mu)^2\right), \tag{2.5}$$

which gives, after some manipulation, that

$$Var(X) = \mathbb{E}\left(X^2\right) - \mu^2.$$

Variance indicates some level of dispersion of the random variable X. The units of the mean and variance of X are not the same, however. Oftentimes we will want to construct statistics relating the mean to a dispersion metric. In this vein, the *standard deviation*, or *volatility* is defined as the square root of variance, σ.

The mean and standard deviation are often described as position and dispersion parameters, respectively. They are also known as the first and second moments of the distribution.

Considering once more the impact of shift and dilation to the expected value and variance of a random variable, X, we have that for any scalars a and b,

$$\begin{aligned} \mathbb{E}(aX+b) &= a\mathbb{E}(X)+b & (2.6)\\ Var(aX+b) &= a^2 Var(X). & (2.7) \end{aligned}$$

Each of these results may be quickly derived using the probability density function approach (the exercise is left to the reader). We shall see later that the expectation operator is linear in combinations of random variables, and the generalization of variance, covariance, is a bilinear form. Here we simply note that the mean is scale and shift sensitive, while variance is independent of changes in position.

We noted that both the mean and variance of a random variable, X, can be infinite. For a practical application, this is a disappointing feature. Especially, as we will see later, if we intend to understand risk via such metrics. An alternate approach is to look at the *percentiles* of the distribution of X. The $(100 \cdot$

p)th percentile of a random variable with cumulative density function F is the smallest number $\pi_p \in \mathbb{R}$ satisfying $F(\pi_p) = p$. That is, π_p satisfies

$$\pi_p = \min\{\pi \mid F(\pi) = p\}. \tag{2.8}$$

where we take the minimum of the set of all values satisfying our criterion. If F is invertible in a neighborhood about π_p, then we may solve for the pth percentile by solving $\pi_p = F^{-1}(p)$.

Percentiles of various distributions have gained prominence over the years as asymmetry of returns and so-called fat tails have become less surprising and more characteristic. Founding theory in the field leaned toward symmetric distributions with well behaved extrema. This yielded the volatility-as-risk paradigm which we will encounter later. Percentile based risk measures allow a much more flexible approach to risk modeling, but have their own trappings as well. At this point we lay the groundwork for future study by pointing out a few specific examples.

The *median* of a random variable X is defined as its 50^{th} percentile. This value may be defined (that is, not infinite) even when the mean may not be. In many cases that follow, for example any symmetrical distribution, the median and mean coincide, but it is by no means necessary.

When considering a random variable representing the loss of a portfolio, say, we might be concerned with documenting rare events. One method of doing this would be to report (assuming loss is a positive number) the 95^{th}, 99^{th}, or other extreme percentiles. This metric is the so-called Value at Risk (VaR) of the portfolio, and has found a place in institutional investing and in regulatory requirements. The methodology of constructing rare events and loss distributions is sometimes prescribed, but may be left to the practitioner's discretion as well. There are features of VaR that are not appealing in some lights, and an extension named Conditional Value at Risk (CVaR) ameliorates these issues in large part. Without defining this quantity exactly, the idea behind CVaR as a risk metric is to look at the average loss past the VaR threshold – that is, the center of mass of the area under the curve to the right of VaR. These ideas will be developed fully in later chapters, but they invariably have a place in early discussions about percentiles due to their conceptual simplicity, and practical applications.

Alternative extensions to better capture the empirical distribution of returns have looked at higher moments. In particular, for a random variable X with mean μ and variance σ^2, *skew* is defined as

$$\mathbb{E}\left(\frac{(X-\mu)^3}{\sigma^3}\right), \tag{2.9}$$

and *kurtosis* is defined by

$$\mathbb{E}\left(\frac{(X-\mu)^4}{\sigma^4}\right). \tag{2.10}$$

We will denote skew and kurtosis by γ and κ, respectively.

For symmetric X, skew is zero (consider the oddness of the integrand if a density is given). Generalizing this observation, skew measures a degree of asymmetry in the distribution of X; viz., if the left tail of X has more (less) weight than the right, X has negative (positive) skew.

Kurtosis is a measure to compare tail densities across random variables. It is theoretically deficient in its usual application which often takes kurtosis as a measure of 'tailedness.' Even so, it provides a metric which accounts for extreme observations. This will be especially important when comparing empirical distributions to the densities that have wide application in theory and practice.

We proceed to a few of these example densities.

2.1.1 Normal Distribution

The normal density, or the density function for a normal random variable, is a two parameter function $\phi_{\mu,\sigma^2}(\cdot)$ given by

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right). \tag{2.11}$$

If a random variable, X, has the above density, we say that it is normal, and denote this by $X \sim N(\mu, \sigma^2)$. The cumulative distribution of a normal random variable will be denoted Φ_{μ,σ^2}.

The usage of μ and σ above is not a coincidence.

Example 2.1.1. The expected value of X with density ϕ_{μ,σ^2} is μ.

We apply the integral in (2.3) directly to see

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx,$$

a change of variables using $u = \frac{x-\mu}{\sigma}$ gives

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \left(\sigma^2 u \exp\left(-u^2/2\right) + \sigma\mu \exp\left(-u^2/2\right)\right) du \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \left(\int_{-\infty}^{\infty} \sigma^2 u \exp\left(-u^2/2\right) du + \int_{-\infty}^{\infty} \sigma\mu \exp\left(-u^2/2\right) du\right).
\end{aligned}$$

Now, the first integrand is odd, and hence integrates to zero over $\mathbb{R}$. Therefore

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \sigma\mu \exp\left(-u^2/2\right) du \\
&= \mu \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-u^2/2\right) du.
\end{aligned}$$

We know from calculus that a change of variables to polar coordinates is all it takes to show that

$$\int_{-\infty}^{\infty} \exp\left(-u^2/2\right) du = \sqrt{2\pi},$$

so that finally we arrive at

$$\mathbb{E}(X) = \mu.$$

The variance is found to be σ^2 using integration by parts, and is left as an exercise.

The normal distribution is clearly symmetric about μ, and hence the median is also μ. When $(\mu, \sigma^2) = (0, 1)$, we say that X is the *standard normal*; we will denote the probability density and cumulative distribution functions of the standard normal by ϕ and Φ, respectively, dropping the subscripts of the general case.

Example 2.1.2. Consider the random variable Y defined by

$$Y = \sigma X + \mu,$$

with X a standard normal random variable. Then Y is a normal random variable with mean μ, and standard deviation σ.

We know immediately from (2.6) and (2.7) that $\mathbb{E}(Y) = \mu$ and $Var(Y) = \sigma^2$. We should also verify that the cumulative density function of Y is in fact the normal cumulative density. We have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(\sigma X + \mu < y) \\ &= \mathbb{P}\left(X < \frac{y-\mu}{\sigma}\right) \\ &= F_X\left(\frac{y-\mu}{\sigma}\right). \end{aligned}$$

Or that $F_Y(y) = \Phi(\frac{y-\mu}{\sigma})$. This is actually sufficient to show that Y is normally distributed, but we utilize the density function, ϕ, to make this more evident. We have that

$$\begin{aligned} F_Y(y) &= \Phi\left(\frac{y-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-\mu}{\sigma}} \exp\left(\frac{-s^2}{2}\right) ds \\ &= \text{(changing variables } x = \sigma s + \mu,\ \frac{1}{\sigma} dx = ds) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{y} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx, \end{aligned}$$

where the density function is now clearly seen to be ϕ_{μ,σ^2}, and so $Y \sim N(\mu, \sigma^2)$.

The above technique is not restricted to normal random variables. For a random variables with finite mean and variance the transformation

$$X \mapsto \frac{X-\mu}{\sigma}$$

Range	Probability
$x \in [\mu - \sigma, \mu + \sigma]$	0.6827
$x \in [\mu - 2\sigma, \mu + 2\sigma]$	0.9545
$x \in [\mu - 3\sigma, \mu + 3\sigma]$	0.9973
$x \in [\mu - 4\sigma, \mu + 4\sigma]$	0.9999

Table 2.1: Probability table for various standard deviation ranges for a normally distributed random variable

, will always produce a zero mean and unit standard deviation random variable. Such normalization is often employed to create scaled variables in modeling.

The skew and kurtosis of a normal random variable are 0, and 3 respectively. The somewhat odd appearance of the number 3 in a math text leads some authors to define kurtosis as $\kappa - 3$. With this modification, a normal random variable has zero kurtosis. We do not adopt this convention.

The normal distribution is (for many reasons) the de facto reference distribution. And if not the normal distribution, its close cousin the log-normal distribution. This is in spite of its lack of applicability in some scenarios. In particular, if we recall David Viniar, the CFO of Goldman Sachs during the Quant Crisis from the introduction [15], we may ask ourselves what the likelihood of moves observed during that particular crisis would be under the assumption of normality.

In the case of normally distributed variables, we need only look at the number of standard deviations from the mean to determine likelihood. This is evidenced in the exponential in the probability density function in (2.3) where, in fact, the motivation for normalization is straightforward. Table 2.1.1 shows various probabilities obtained by calculating $\Phi(k\sigma) - \Phi(-k\sigma)$ for various levels of k.

We may interpret some of this data in the following way: for a normally distributed random variable, seeing a four standard deviation observation is a 1 in 10,000 event. Looking at the table, it is clear that these probabilities are not linear in σ. A six sigma event, for instance, is a one in five hundred million event. Put in perspective, a six sigma event would occur one day every 1,388,455 million years. Five times longer than the appearance of anatomically modern humans in Africa.

Now, what about a 25 sigma event? The probability of such an event is 6.10×10^{-138} . In terms of years, this would occur one day in 4.49×10^{132} years. Here's the thing: the universe is only 13.82 billion years old, so that you shouldn't see this type of event but once across 3.2×10^{125} universes [16, 17].

Put another way, the surface area of the earth is 510 trillion m^2. Approximating a human hand at 0.01 m^2, it would be 3.2×10^{118} times more likely to catch a random ball dropped from space in the palm of your hand than to see a twenty five sigma event.

Returning to the cosmological scale, an upper bound for the number of particles in the universe is 1.0e85. Thus you'd be more likely to pick a single

particle from the universe than to witness a 25 sigma event. Much, much, much more likely.

We mentioned in the first chapter that during the Quant Crisis, David Viniar, the CFO of Goldman Sachs, was quoted as saying they were observing 25-standard deviation moves, several days in a row. Connecting this statement to the probabilities just calculated is not done in a pejorative sense. More, it is likely that they had very carefully constructed a very stable and attractive product that was blown out by overcrowding and liquidations. Unfortunately we: 1) may find ourselves inherently referring to the normal distribution and its shortcuts of standard deviation explaining everything even when this is inappropriate; 2) tend to think linearly so that a 25 standard deviation *feels* like six times a four standard deviation move via some heuristics; and 3) do not fully understand the dynamics of equity returns, which are subject to crises and other exogenous effects.

The normal distribution, even with the above in mind, has extremely nice properties, not the least of which is that it is overwhelmingly the most likely distribution to assume if you are only given the mean and standard deviation of a random variable. The analogy would be that the normal distribution acts as our first linear approximation to the underlying distributions that we may never know. This does not preclude its use as a tool, but should give us a moment to reflect, especially if leverage is determined via risk estimates based on its use.

2.1.2 Log-normal Distribution

The density function for a log-normal random variable, is again a two parameter function $ln_{\mu,\sigma^2}(\cdot)$ given by

$$ln_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}exp\left(\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right), \tag{2.12}$$

and is clearly only defined on $\mathbb{R}_+$. A log-normal random variable with parameters μ and σ^2 is such that its logarithm is distributed as $N(\mu,\sigma^2)$, and we will denote it by $LN(\mu,\sigma^2)$. That is, Y is log-normal if $\log(Y)$ is normal.

Example 2.1.3. In a manner similar to before we may obtain the cumulative distribution function for $Y \sim LN(\mu,\sigma^2)$; viz.,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(\exp(X) < y) \\ &= \mathbb{P}\left(X < \log(y)\right) \\ &= F_X(\ln(y)), \end{aligned}$$

so that $F_Y(y) = \Phi_{\mu,\sigma^2}(\ln(y)) = \Phi\left(\frac{\log(y)-\mu}{\sigma}\right)$.

From here, we may also derive the density function for $Y \sim LN(\mu, \sigma^2)$ as

$$\begin{aligned} F_Y(y) &= \Phi\left(\frac{\log(y)-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log(y)-\mu}{\sigma}} \exp\left(\frac{-s^2}{2}\right) ds \\ &= \text{(changing variables } s = \frac{\log(x)-\mu}{\sigma}, \frac{1}{\sigma \cdot x} dx = ds) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^y \frac{1}{x} \exp\left(\frac{-(\log(x)-\mu)^2}{2\sigma^2}\right) dx, \end{aligned}$$

which is exactly the density given in (2.12), as desired.

We shall see that the above technique generalizes, allowing us to derive density functions of random variables constructed from functions of random variables. We will call this a *change of variables theorem* – perhaps not surprisingly given the repeated application above.

The mean and variance of $Y \sim LN(\mu, \sigma^2)$ are given by

$$\begin{aligned} \mathbb{E}(Y) &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ Var(Y) &= \exp\left(2\mu + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right), \end{aligned}$$

and their derivations are left as exercises.

The log-normal distribution is terribly important, due in large part to the Black-Scholes option pricing formula which relies on the assumption of log-normality of future stock prices over any interval. For a stock price, a small increment in time δ, and initial and terminal times t and $T = t+\delta$, respectively, $\ln(S_T)$ is distributed as

$$\ln(S_{t+\delta}) \sim N\left(\ln(S_t) + \left(\mu - \frac{\sigma^2}{2}\right)\delta, \sigma^2\delta\right).$$

A related result is that so-called *log returns*,

$$r_t = \ln\left(\frac{S_{t+\delta}}{S_t}\right)$$

are normally distributed as

$$r_t \sim N\left(\left(\mu - \frac{\sigma^2}{2}\right)\delta, \sigma^2\delta\right).$$

The above relationship is not empirical as we shall soon see. It is, however, interpretive, as market participants calculate volatilities, σ, implied by market

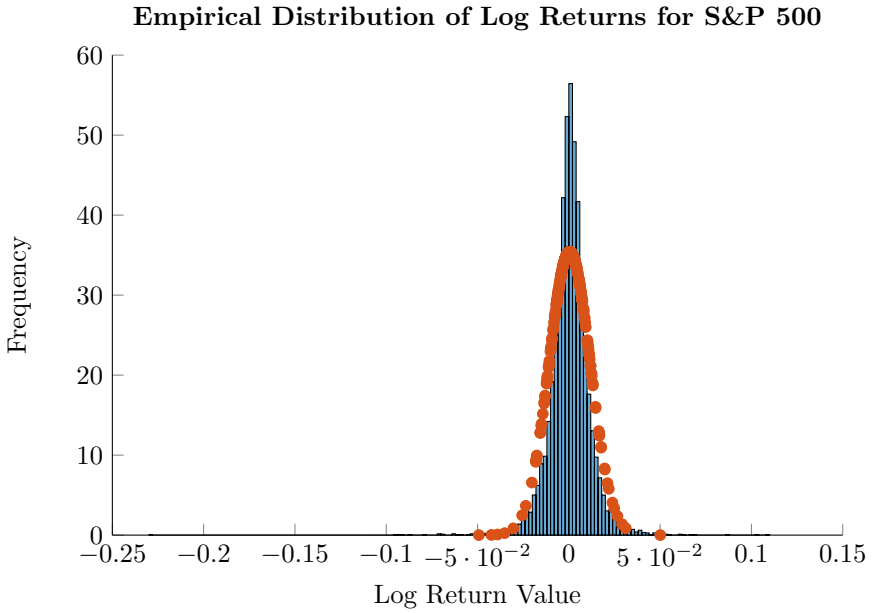

Figure 2.1: Empirical distribution of log returns of the S&P 500 from 1980 through 2015 with a fitted normal distribution in blue and red, respectively.

prices for various derivative contracts. Interestingly, the drift term, μ, is inconsequential in derivatives pricing as the drift may be hedged away in a frictionless world, but this is beyond the scope of the current work.

What we *can* look at is the empirical distribution of daily log returns for the S&P 500 Index over a long horizon, in this case from 1980 through 2015. Figure 2.1.2 shows a normalized histogram along with a fitted normal distribution. For the present, we will not be concerned about how to estimate μ and σ from data, even as there is likely ample familiarity with calculating both. Instead, we take as granted that a 'good' estimate of the mean and standard deviation are possible to obtain. We will have cause to revisit this notion when we notice that these estimates–even while they have good theoretical backing–are not stable through time in markets.

Figure 2.1.2 clearly demonstrates 'peakedness', with a considerable amount of the mass of the empirical distribution exceeding the fit (blue versus red) for values near zero. In addition, we also see extreme behavior in the market returns, with the largest moves falling far outside the tails implied by the fit. This occurs on both sides of the mean; that is, for positive and negative returns, but the former are smaller in magnitude than the latter. We see, then, an example of the asymmetry of returns, with downside risk being greater than upside gains. The market attempts to price this phenomenon in the options market.

In addition to optics obtained from the figure, we may also calculate em-

Range	Empirical Frequency	Ratio of Empirical Extreme to Normal
$x \in [\mu - \sigma, \mu + \sigma]$	0.7889	0.665
$x \in [\mu - 2\sigma, \mu + 2\sigma]$	0.9534	1.02
$x \in [\mu - 3\sigma, \mu + 3\sigma]$	0.9859	5.22
$x \in [\mu - 4\sigma, \mu + 4\sigma]$	0.9942	92.14
$x \in [\mu - 5\sigma, \mu + 5\sigma]$	0.9974	4,609
$x \in [\mu - 6\sigma, \mu + 6\sigma]$	0.9983	837,128

Table 2.2: Empirical frequency table for daily log returns of the S&P 500 Index from 1980 through 2015. In addition to frequencies, we also calculate the ratio of the empirical frequency of moves outside $\mu \pm k\sigma$ to the fit probability of such moves.

pirical probabilities and compare those to the theoretical. To wit, we look at the percentage of observations within $\mu \pm k\sigma$ for various values of k. We also calculate how much more likely returns *outside* of k standard deviations from the mean are *relative to the fit normal density*. The results are in Table 2.1.2.

The table clearly exhibits our previous observations. Namely, there are many more observations within one standard deviation of the mean than implied by a normal fit. Two standard deviations seems to be about right, but as we go out to four, five, and six standard deviations, the empirical distribution completely loses track with the fit. For example, six-σ events are 837,128 were more likely to occur historically than would occur under a best fit normal distribution. This echoes our previous analysis regarding 25-σ events. But in the present case we are establishing what are termed *stylized features of returns*. In addition to the likelihood analysis, we may also look at the empirical skew and kurtosis of the S&P log returns. We find that the estimated skew is -1.151, indicating that there is more mass to the left of the mean than in a centered distribution; this is only indicative, however, as 51.07% of the log returns are positive, the result being influenced by the extreme moves in the left tail. The sample kurtosis is 29.43, a far-cry from the value of 3 obtained from the normal distribution. The implication of the large sample kurtosis is that the empirical distribution has many outliers, or more extreme outliers (relative to a normal density). This is indeed the case.

2.1.3 Student t Distribution

The Student t distribution found application (but not derivation) from William Gosset, working for Guiness Brewery in a paper in 1908 [13]. His paper describes the distribution as the "frequency distribution of standard deviations of samples drawn from a normal population." The distribution was later popularized by Robert Fisher. We will find widespread application of the Student t distribution when we encounter it again (somewhat more naturally) in our regression work when describing the distribution of estimation parameters. For now, we begin

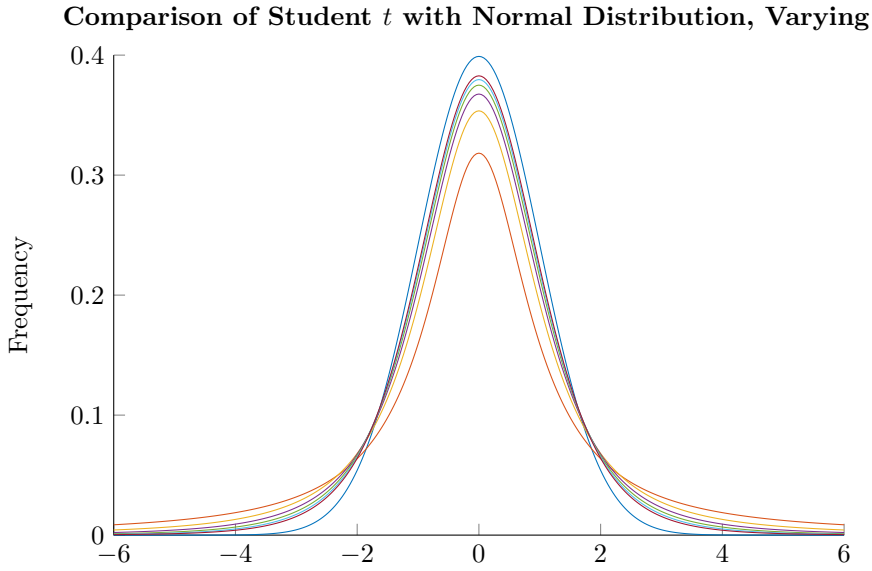

Figure 2.2: Comparison of the Student t density for $\nu = 1, 2, \ldots, 6$ with the standard normal density. The density with the lowest peak at $x = 0$ belongs to $\nu = 1$, with ν increasing with the height of the peak; the highest belonging to the normal distribution.

with the somewhat obtuse probability density function.

The Student t density is a three parameter function $\mathrm{st}_{\mu,\sigma^2;\nu}(\cdot)$ with $\nu \in \mathbb{N}_+$ given by

$$\mathrm{st}_{\mu,\sigma^2;\nu}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi\sigma^2}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}, \tag{2.13}$$

where ν is referred to as the *degrees of freedom* of the distribution, and $\Gamma(\cdot)$ is the gamma function (whose values at $n \in \mathbb{N}_+$ give $\Gamma(n) = (n-1)!$).

If a random variable, T, has the above density, we say that it is Student t with ν degrees of freedom and denote this by $T \sim \mathrm{St}(\mu, \sigma^2; \nu)$. We omit the cumulative distribution function here as it is nearly as intractable written down as not, involving a function called the regularized beta function. Modern programming languages are capable of determining the values of the CDF of the Student t distribution.

The t distribution with one degree of freedom is known as the *Cauchy distribution*, which has infinite variance and an undefined mean. It is simple to show, however, that for $\nu > 1$, $T \sim \mathrm{st}(\mu, \sigma^2; \nu)$ has expectation μ. For $\nu > 2, 3$,

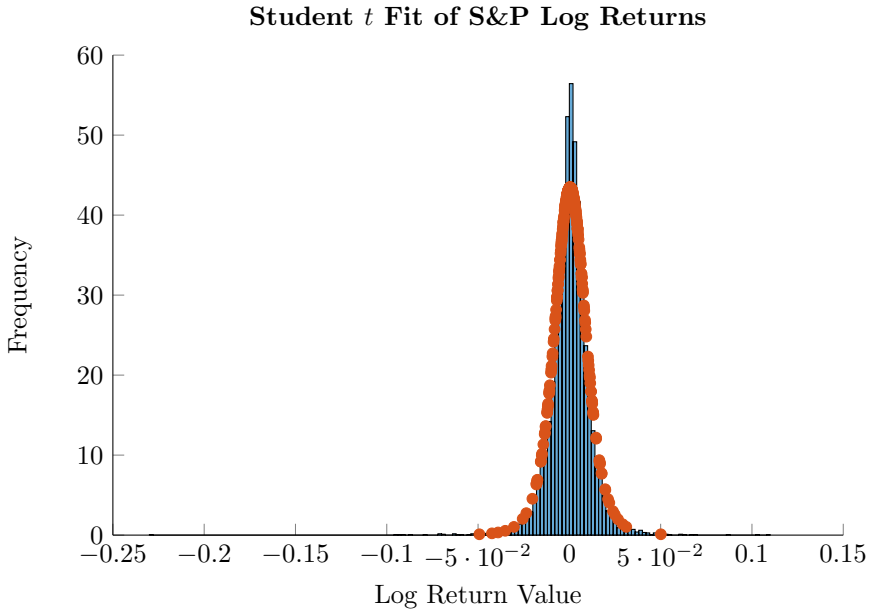

Figure 2.3: Empirical distribution of log returns of the S&P 500 from 1980 through 2015 with a fitted Student t distribution in blue and red, respectively. The Student t distribution used here has $\nu = 5$ degrees of freedom.

and 4, respectively, we have

$$\begin{aligned} Var(T) &= \frac{\nu}{\nu-2}\sigma^2 \\ \gamma(T) &= 0 \\ \kappa(T) &= 3 + \frac{6}{\nu-4}, \end{aligned}$$

where in the final equation we see that the Student t distribution has kurtosis greater than three for $\nu > 4$, a desirable property given previous discussions. The limit as $\nu \to \infty$ of $\text{st}_{\mu,\sigma^2;\nu}(\cdot)$ is the normal density, $\phi_{\mu,\sigma^2}(\cdot)$.

We compare $T \sim \text{St}(0, 1; \nu)$ and $X \sim N(0, 1)$ for various choices of ν in Figure 2.1.3. As can be seen, the Student t distribution has fatter tails than the normal density. However, this comes at the expense of losing some peakedness at the mean, another stylized feature we noticed in equity returns. This is mitigated, however, by noticing that the variance of $T \sim \text{St}(0, 1; \nu)$ is exactly $\frac{\nu}{\nu-2}$ for $\nu > 2$. That is, the peak would be higher if the variance of T had been set to a constant 1.

Meucci [23] has suggested using a parameter choice of $\nu = 5$ in fitting a Student t distribution to log returns. We do this and show the result as before in Figure 2.1.3. Optically, we see a much better fit to the data than in the normal fit to log returns.

In addition, we may go through the same exercise as before, comparing theoretical and empirical frequencies for fit values of μ and σ. This is left as an exercise for the reader. In doing this, one sees that both the center and the tails are much better explained by the Student t distribution than by the normal distribution. Extreme values are less than twice as likely to occur in our historical data set than the fit Student t density would imply as well: a vast improvement over the previous discrepancy of over 800,000.

2.1.4 Functions of Random Variables

In our previous examples, we encountered a few examples of random variables that were functions of random variables; e.g., $Y = e^X$ and $Y = \mu + \sigma X$, for X with a known distribution. In determining the cumulative distribution and density functions of these transformed random variables, we began with a manipulation of the CDF to a known distribution and proceeded to analyze the probability density function and integration bounds obtained. We then used a change of variables to determine the density of the transformed variable. Here we formalize the procedure, developing the *change of variable theorem*.

Theorem 2.1.1. If $g : \mathbb{R} \to \mathbb{R}$ is continuous and monotonically increasing and $X \in \mathbb{R}$ is a random variable, then the random variable $Y = g(X)$ defined by setting $y = g(x)$ for every realization of X has the cumulative distribution function

$$F_Y(y) = F_X\left(g^{-1}(y)\right), \tag{2.14}$$

and if X has a density, then the density of Y is given by

$$f_Y(y) = \frac{dg^{-1}(y)}{dy} f_x\left(g^{-1}(y)\right). \tag{2.15}$$

Proof: We know by the inverse function theorem that a local inverse of g exists, hence for any y, we may proceed as above,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(g(X) < y) \\ &= \mathbb{P}\left(X < g^{-1}(y)\right) \\ &= F_X\left(g^{-1}(y)\right). \end{aligned}$$

From here we may relate the density fucntion of X to the density function of Y. We get

$$\begin{aligned} F_Y(y) &= F_X\left(g^{-1}(y)\right) \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(s)ds \\ &\quad \text{(changing variables } s = g^{-1}(u),\ \frac{dg^{-1}(u)}{du}du = ds, \text{ by the chain rule)} \\ &= \int_{-\infty}^{y} \frac{dg^{-1}(u)}{du} f_x\left(g^{-1}(u)\right) du \end{aligned}$$

as desired. Notice that we took advantage of the monotonicity of g in our change of bounds, with $s = g^{-1}(y)$ implying $g^{-1}(u) = g^{-1}(y)$, and hence $u = y$.

We may also relax the original condition to require only that g is monotonic (and not necessarily increasing). Working through the proof in the same manner, but allowing for monotonically decreasing g gives

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_x\left(g^{-1}(y)\right). \tag{2.16}$$

2.2 Multivariate Distributions

We may generalize our work to include vectors of random variables,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}.$$

The cumulative distribution function then also generalizes to the *joint cumulative distribution function* of X, a function $F : \mathbb{R}^N \to [0, 1]$ which satisfies

$$F(x_1, x_2, \ldots, x_N) = \mathbb{P}\left(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N\right) \tag{2.17}$$

We say that the multivariate random variable, X, has a density,

$$f : \mathbb{R}^N \to \mathbb{R}_+,$$

if we may write

$$F(x_1, x_2, \ldots, x_N) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} f(s_1, \ldots, s_N) ds_N \cdots ds_1. \tag{2.18}$$

As before, we require that f be integrable, nonnegative, and integrate to one.

The expected value of the multivariate random variable X, $\mathbb{E}(X)$, is given component wise

$$\mathbb{E}(X) = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_N) \end{pmatrix}, \tag{2.19}$$

and the generalization of variance, *covariance*, denoted $Cov(X)$ is determined by

$$Cov(X) = \mathbb{E}\left((X - \mu)(X - \mu)'\right), \tag{2.20}$$

where $X \in \mathbb{R}^N$, $\mu = \mathbb{E}(X)$, and expectation is again componentwise in $Cov(X) \in \mathbb{R}^{N \times N}$. We will denote the (i, j) component of $Cov(X)$ as σ_{ij}, with $\sigma_{ii} = Var(X_i)$.

We will often denote the covariance matrix by Σ, and we will spend considerable time analyzing its properties both from a theoretical as well as empirical point of view (and oftentimes settling on the intersection of the two).

Example 2.2.1. The expectation operator is linear; viz.,

$$\mathbb{E}\left(w_1X_1 + w_2X_2\right) = w_1\mathbb{E}\left(X_1\right) + w_2\mathbb{E}\left(X_2\right). \tag{2.21}$$

for random variables, X_1 and X_2, and scalars w_1 and w_2. This generalizes to

$$\mathbb{E}\left(\sum w_iX_i\right) = \sum_{i=1}^{N} w_i\mathbb{E}\left(X_i\right).$$

We prove the result assuming that there exists a joint density function, $f(\cdot)$, but this is not necessary. We have in this case

$$\begin{aligned}
\mathbb{E}\left(w_1X_1 + w_2X_2\right) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(w_1X_1 + w_2X_2\right)f(x_1,x_2)dx_1dx_2 \\
&\quad \text{(by the linearity of the integral operator)} \\
&= w_1\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}X_1f(x_1,x_2)dx_1dx_2 + w_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}X_2f(x_1,x_2)dx_1dx_2 \\
&= w_1\mathbb{E}\left(X_1\right) + w_2\mathbb{E}\left(X_2\right)
\end{aligned}$$

as desired.

Example 2.2.2. The covariance matrix, Σ is positive semi-definite. We say that a matrix, $A \in \mathbb{R}^{N\times N}$, is *positive semi-definite* if for all $w \in \mathbb{R}^N$ not identically zero,

$$w'Aw \geq 0, \tag{2.22}$$

with $w'Aw = 0$ only when $w \equiv 0$.

Given a multivariate random variable, $X \in \mathbb{R}^N$, with $Cov(X) = \Sigma$, and a scalar vector $w \in \mathbb{R}^N$, the result follows by considering the variance of the random variable $Y = w'X = \sum_{i=1}^{N} w_iX_i$; viz.,

$$\begin{aligned}
Var(Y) &= Var(w'X) \\
&= \mathbb{E}\left((w'X - w'\mu)^2\right) \\
&= \mathbb{E}\left((w'(X-\mu))^2\right)
\end{aligned}$$

where $\mu = \mathbb{E}(X)$. Continuing, we have by the linearity of the expectation operator that

$$\begin{aligned}
Var(Y) &= \mathbb{E}\left(w'(X-\mu)(X-\mu)'w\right) \\
&= w'\mathbb{E}\left((X-\mu)(X-\mu)'\right)w \\
&= w'\Sigma w.
\end{aligned}$$

Now, we know from the definition of variance that $Var(Y) \geq 0$. The above relationship then implies $w'\Sigma w \geq 0$ as well, proving the result.

We will make stronger claims about the covariance matrix, Σ, in future chapters, refining our assumptions about the composition of X to ensure that

Σ is *positive definite*; that is, inner products with nonzero vectors are strictly nonzero. The proof will be identical, with only a few comments added.

We may also consider linear combinations of the form BX for a multivariate random variable $X \in \mathbb{R}^N$ and scalar matrix $B \in \mathbb{R}^{M \times N}$. We maintain linearity in expectation; viz.,

$$\mathbb{E}(BX) = B\mathbb{E}(X). \tag{2.23}$$

This can be shown directly since

$$BX = \begin{pmatrix} \sum_{i=1}^N b_{1i} X_i \\ \vdots \\ \sum_{i=1}^N b_{Mi} X_i \end{pmatrix}$$

so that

$$\mathbb{E}(BX) = \begin{pmatrix} \mathbb{E}\left(\sum_{i=1}^N b_{1i} X_i\right) \\ \vdots \\ \mathbb{E}\left(\sum_{i=1}^N b_{Mi} X_i\right) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N b_{1i} \mathbb{E}\left(X_i\right) \\ \vdots \\ \sum_{i=1}^N b_{Mi} \mathbb{E}\left(X_i\right) \end{pmatrix} = B\mathbb{E}(X).$$

Similarly, $\mathbb{E}(XB) = \mathbb{E}(X)B$ for scalar B with appropriate dimensions.

The covariance of a BX as above, $Cov(BX)$, is given by

$$Cov(BX) = BCov(X)B'. \tag{2.24}$$

This follows from the linearity of expectation since

$$\begin{aligned} Cov(BX) &= \mathbb{E}\left((BX - \mathbb{E}(BX))\,(BX - \mathbb{E}(BX))'\right) \\ &= \mathbb{E}\left(B\,(X - \mathbb{E}(X))\,(X - \mathbb{E}(X))'\,B'\right) \\ &= B\mathbb{E}\left((X - \mathbb{E}(X))\,(X - \mathbb{E}(X))'\right)B' \\ &= BCov(X)B'. \end{aligned}$$

Finally, if X and Y are univariate random variables taking their values in $\mathbb{R}$, we define

$$Cov(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right). \tag{2.25}$$

In terms of previous notation, then, if $X \in \mathbb{R}^N$, $Cov(X_i, X_j) = \sigma_{ij}$.

We may also consider the effect of arbitrary (continuous and injective) transformations on the distribution of a random variable, generalizing the change of variable theorem to the multivariate case.

Theorem 2.2.1. Let X be a multivariate random variable with $X \in \mathbb{R}^N$, and let

$$g : \mathbb{R}^N \to \mathbb{R}^N$$

be a one-to-one and continuous function. If $f_X(\cdot)$ is the density of X, then the density of

$$Y = g(X)$$

is

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \det\left(\nabla g^{-1}(y)\right) \tag{2.26}$$

where ∇h denotes the Jacobian of a function h, and $\det(\cdot)$ is the determinant function. The Jacobian, determinant, and change of variables in multivariate integration are discussed in the appendix.

The proof is exactly as before, accounting for the fact that we are working in $\mathbb{R}^N$ and hence encounter the Jacobian rather than the derivative of the inverse of g.

The *marginal distribution* of X_j is denoted as in the univariate case as

$$F_j(x_j) = \mathbb{P}(X_j \leq x_j)$$

and is given by

$$F\left(-\infty, \ldots, x_j, \ldots, \infty\right), \tag{2.27}$$

or in the case of densities, all upper bounds of integration are infinite except that related to the jth component, which is just x_j. When a density exists,

$$f_j(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(s_1, \ldots, s_{j-1}, x, s_{j+1}, , \ldots, s_N) ds_N \cdots ds_{j+1} ds_{j-1} \cdots ds_1 \tag{2.28}$$

is the jth *marginal density*.

Example 2.2.3. Figure 2.2 shows an example joint distribution using empirical data. Here, monthly log returns for both the S&P 500 and IBM are shown in a scatterplot. Contour lines are shown for the best fit multivariate normal density, and marginal distributions are given outside the scatterplot. We may make a few observations:

- The jointly normal distribution assumption seems to be a poor fit overall, but there seems to be evidence that an ellipsoidal distribution is a decent approximation for the joint distribution.
- There seems to be a fairly strong linear relationship between the monthly returns of IBM and the S&P 500 [11, 25].
- Extreme events seem more tightly clustered on the downside.

We will return to the second observation when we establish the Capital Asset Pricing Model. At this point, we can say something to the effect of 'IBM looks a lot like the market, with some random noise that oftentimes looks to be normally distributed,' and leave it at that.

Two random variables, X_1 and X_2, are said to be *independent* if

$$\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) = \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2).$$

Or equivalently,

$$F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2), \tag{2.29}$$

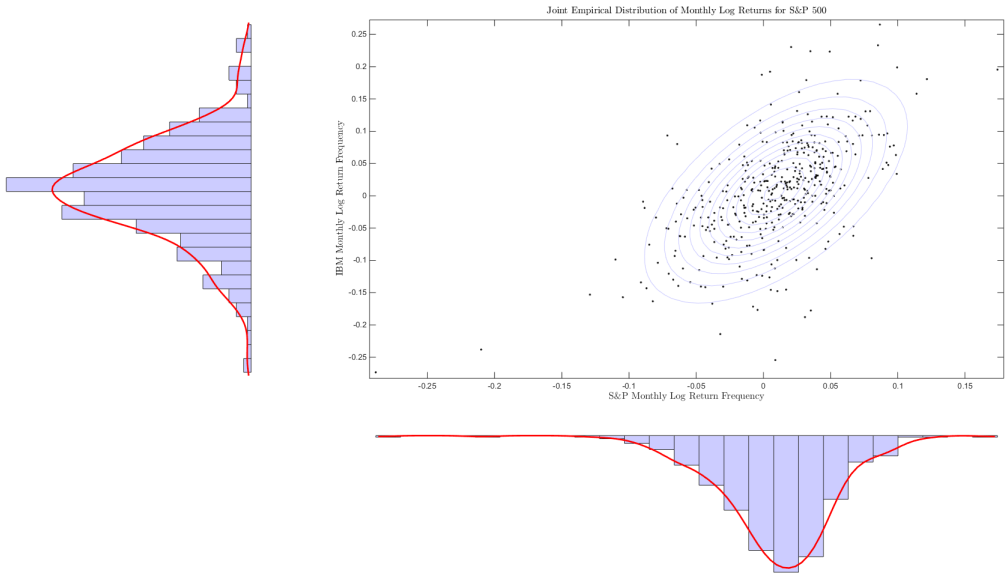

Figure 2.4: Empirical joint distribution of log returns of the S&P 500 and IBM from 1980 through 2015, including marginal distributions. The scatterplot includes contour lines of a best fit multivariate normal density.

where $F_i(\cdot)$ is the marginal CDF. Independence, then, captures the notion that the probability of independent events is the product of their marginal probabilities, the outcome of one does not impact the others. In terms of densities (should the joint cumulative distribution function allow a density), we have that independent random variables have density

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2), \tag{2.30}$$

where again $f_i(\cdot)$ is the marginal density of X_i.

Example 2.2.4. Returning to the return data used to construct Figure 2.2, we can verify that the empirical frequencies do not give an indication of independence between the two random variables. Our belief that these series should not be independent stems from our previous observation of the apparent relationship between the two. In particular, the ellipsoidal contour lines in Figure 2.2 bolster this claim.

We aren't in need of much theory, however: a little verification is all we need, utilizing Equation (2.30). We may look at the joint probability

$$\mathbb{P}(r_{IBM} < r_\tau, r_{S\&P} < r_\tau)$$

and compare this to

$$\mathbb{P}(r_{IBM} < r_\tau)\mathbb{P}(r_{S\&P} < r_\tau)$$

for various values of r_τ.

Looking at $r_\tau = 0$, for instance,

$$\begin{aligned} \mathbb{P}(r_{IBM} < r_\tau, r_{S\&P} < r_\tau) &= 0.333 \\ \mathbb{P}(r_{IBM} < r_\tau)\mathbb{P}(r_{S\&P} < r_\tau) &= 0.225, \end{aligned}$$

giving evidence from the empirical distribution that these are not independent, as expected. Notice that in this example independence will underweight the probability of joint downward movements.

A corollary to this observation is that risk metrics should account for joint as well as marginal distributions. The covariance matrix, Σ, is one method of capturing these joint dynamics, but as we see in the figure, the assumption of joint normality is likely not accurate. Extensions have been made to incorporate various joint distributions given marginal distributions as input. This is often done with *copula functions* which are a popular and flexible tool which we shall encounter later on.

The variance of the sum of univariate random variables X and Y with $\mathbb{E}(X) = \mu_X$ and $\mathbb{E}(Y) = \mu_Y$ is,

$$\begin{aligned} Var(X+Y) &= \mathbb{E}\left((X+Y-\mu_X-\mu_Y)^2\right) \\ &= \mathbb{E}\left((X-\mu_X)^2\right) + \mathbb{E}\left((Y-\mu_Y)^2\right) + 2\mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}. \end{aligned}$$

In the case of independent events the cross term, σ_{XY} is zero, thus if X and Y are independent $Var(X+Y) = Var(X) + Var(Y)$. This is left to the reader as an exercise.

This generalizes to the case of N jointly independent random variables, where, for instance the joint probability density function may be written as

$$f(x_1, \ldots, x_N) = \prod_{i=1}^{N} f_i(x_i), \tag{2.31}$$

the joint distribution function as

$$F(x_1, \ldots, x_N) = \prod_{i=1}^{N} F_i(x_i), \tag{2.32}$$

and joint probabilities as

$$\mathbb{P}(X_1 \leq x_1, \ldots, X_N \leq x_N) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq x_i) \tag{2.33}$$

with the same notation used previously. Intuitively, it is this final formulation that might resonate with the non-technical usage of independence: when one thinks of multiple coin flips, one instinctively multiplies probabilities to determine outcomes. Sequential coin flips are the prototypical example of independent and identically distributed random variables.

We say that a sequence of random variables $\{X_i\}_{i=1}^N$ are *independent and identically distributed,* or *iid*, if each X_i has the same distribution and all pairs, (X_i, X_j) are pairwise independent.

2.2.1 Multivariate Normal Distribution

The multivariate normal density, or the density function for a multivariate normal random variable, is a two parameter function, $\phi_{\mu,\Sigma}(\cdot)$, given by

$$\phi_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{\det(\Sigma)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \tag{2.34}$$

for $x \in \mathbb{R}^N$.

As before, if a random variable, X, has the above density, we say that it is multivariate normal, and denote this by $X \sim N(\mu, \Sigma)$. The cumulative distribution of a normal random variable will be denoted as before by $\Phi_{\mu,\Sigma}$. We state without proof that in such a case

$$\begin{aligned} \mathbb{E}(X) &= \mu \\ Cov(X) &= \Sigma. \end{aligned}$$

Implicit in the definition is that Σ is invertible, and therefore, since Σ is a covariance matrix and hence positive semidefinite, we must have that Σ is positive definite.

We also note that linear combinations of jointly normal random variables are again normal. In the context of portfolio management, if the log returns, say, of various assets are assumed to be jointly normal, then a portfolio (that is a weighted sum) of these log returns will also be normal, with tractable mean and variance. Some care is being taken here to say 'jointly normal.' This is due to the fact that it is possible for marginal distributions to be normal without the joint distribution being multivariate normal. We are precluding this possibility.

We have in the general case the if X is multivariate normal with $X \sim N(0, I)$, for I the identity matrix, 0 a vector of zeros, and $X \in \mathbb{R}^N$, then for $a \in \mathbb{R}^N$ and nonsingular $B \in \mathbb{R}^{N\times N}$, the random variable $Y = a + BX$ is distributed as $Y \sim N(a, BB')$. We may employ the change of variables theorem for the multivariate case, but we choose to follow the same method as was seen in the univariate case. We have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(a + BX < y) \\ &= \mathbb{P}\left(X < B^{-1}(y-a)\right) \\ &= F_X\left(X < B^{-1}(y-a)\right). \end{aligned}$$

This gives that $F_Y(y) = \Phi_{0,I}(B^{-1}(y-a))$. Looking now at the formulation in terms of the density of X, we have

$$\begin{aligned}
F_Y(y) &= \Phi_{0,I}\left(B^{-1}(y-a)\right) \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}} \int_{s \leq B^{-1}(y-a)} \exp\left(-\frac{1}{2}s's\right) ds \\
&= (\text{changing variables } x = Bs + a,\ \frac{1}{\det(B)} dx = ds) \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{\det(B)} \int_{x \leq y} \exp\left(-\frac{1}{2}(B^{-1}(x-a))'(B^{-1}(x-a))\right) dx \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{\det(B)} \int_{x \leq y} \exp\left(-\frac{1}{2}(x-a)'(B^{-1})'B^{-1}(x-a)\right) dx \\
&= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{\det(B)} \int_{x \leq y} \exp\left(-\frac{1}{2}(x-a)'(BB')^{-1}(x-a)\right) dx
\end{aligned}$$

where the density function is now clearly seen to be $\phi_{a,BB'}$, and so $Y \sim N(a, BB')$.

This example, too, indicates how we may simulate multivariate normal random variables. Given an arbitrary covariance matrix, Σ, and vector, μ, if one can write $\Sigma = \Lambda\Lambda'$, then $Y = \mu + \Lambda X$, with $X \sim N(0, I)$ is distributed as $Y \sim N(\mu, \Sigma)$.

In our proof, we required that B was invertible. A more general proof is available, resorting to so-called moment generating functions, but we do not cover these here. Suffice to say that the proof generalizes to the case $Y = a + BX$ for $a \in \mathbb{R}^M$, $B \in \mathbb{R}^{M \times N}$, and X multivariate normal of dimension N.

Example 2.2.5. Many paragraphs have already been spent discussing the potential lack of applicability of the normal distribution for equity returns. We continue this arc, again looking at figure 2.2, where an estimated $\hat{\mu}$ and $\hat{\Sigma}$ are used to draw the contour lines of the multivariate normal density best fitting the data. To do this, contour lines are simply determined by

$$\mathcal{I}_c = \left\{ x \,|\, (x - \hat{\mu})'\hat{\Sigma}^{-1}(x - \hat{\mu}) = c \right\} \tag{2.35}$$

for various values of c. The fact that these sets fix $\phi_{\hat{\mu},\hat{\Sigma}}$ is evident from the definition in (2.34).

Clearly there is, as in the univariate case, some reasonable fit, but extreme values and peakedness seem to still dominate, nonetheless.

2.2.2 Multivariate Log-Normal Distribution

The multivariate log-normal distribution is defined as before. We say that Y is a multivariate log-normal random variable if for some $X \sim N(\mu, \Sigma)$,

$$Y = \exp(X)$$

or, equivalently, Y is log-normal if its logarithm is normal. In the above, exponentiation and logarithms are componentwise; viz., for $X \in \mathbb{R}^N$,

$$\exp(X) = \begin{pmatrix} \exp(X_1) \\ \vdots \\ \exp(X_N) \end{pmatrix}.$$

The multivariate log-normal distribution is again described by two parameters, μ, and Σ, and denoted $Y \sim LN(\mu, \Sigma)$. The density function for lognormal Y taking values in $\mathbb{R}^N$ is given by

$$ln_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{\det(\Sigma)^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\log(x)-\mu)'\Sigma^{-1}(\log(x)-\mu)\right) \cdot \prod_{i=1}^{N} \frac{1}{x_i} \tag{2.36}$$

for $x \in \mathbb{R}^N_+$.

Example 2.2.6. We may derive the density $ln_{\mu,\Sigma}$ by use of the change of variable theorem. For $Y = \exp(X) = g(X)$, with X having density $\phi_{\mu,\Sigma}(\cdot)$, the inverse of g is given by

$$g^{-1}(Y) = \begin{pmatrix} \log(Y_1) \\ \vdots \\ \log(Y_N) \end{pmatrix} = \begin{pmatrix} g_1^{-1}(Y_1) \\ \vdots \\ g_N^{-1}(Y_N) \end{pmatrix}.$$

To determine the Jacobian of $g^{-1}(\cdot)$, we must calculate the partial derivatives $\frac{\partial g_i^{-1}}{\partial y_j}$. From the above, these are exactly given by

$$\frac{\partial g_i^{-1}}{\partial y_j} = \begin{cases} \frac{1}{y_j} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

The Jacobian, ∇g^{-1}, is therefore a diagonal matrix, whose determinant is found simply to be

$$\prod_{i=1}^{N} \frac{1}{y_i}.$$

We are now ready to derive the density function for Y:

$$\begin{aligned} ln_{\mu,\Sigma}(y) &= \phi_{\mu,\Sigma}\left(g^{-1}(y)\right) \left|\det\left(\nabla g^{-1}(y)\right)\right| \\ &= \phi_{\mu,\Sigma}\left(\log(y)\right) \prod_{i=1}^{N} \frac{1}{y_i} \end{aligned}$$

which is exactly (2.36).

The mean and covariance of $Y \sim LN(\mu, \Sigma)$ are given as

$$\begin{aligned} \mathbb{E}(Y)_i &= \exp\left(\mu_i + \frac{1}{2}\sigma_i^2\right) \\ Cov(Y)_{ij} &= \exp\left(\mu_i + \mu_j + \frac{1}{2}(\sigma_i^2 + \sigma_j^2)\right)\exp(\sigma_{ij} - 1) \end{aligned}$$

where σ_{ij} denotes the (i, j) component of the matrix, Σ as usual. We do not prove this result.

2.2.3 Multivariate Student t Distribution

The multivariate Student t distribution follows the same method of extension as the multivariate normal density. We say T is distributed as a multivariate Student t distribution with $\nu \in \mathbb{N}_+$ *degrees of freedom* if T has density

$$\mathrm{st}_{\mu,\Sigma;\nu}(x) = \frac{\Gamma\left(\frac{\nu+N}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{N/2}} \frac{1}{\det(\Sigma)^{1/2}} \left(1 + \frac{1}{\nu}(x-\mu)'\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+N}{2}}, \tag{2.37}$$

where $\Gamma(\cdot)$ is the gamma function.

It may be shown that the marginals of the multivariate Student t are themselves univariate Student t with density $\mathrm{st}_{\mu_i,\sigma_i^2;\nu}(x)$. As a result, when $\nu > 1$, then, $\mathbb{E}(T) = \mu$. Finally, for $\nu > 2$, $Cov(T) = \frac{\nu}{\nu-2}\Sigma$.

Example 2.2.7. Like the multivariate normal distribution, we may show that an affine transformation of $X \in \mathbb{R}^N$ with $X \sim St(\mu, \Sigma; \nu)$,

$$a + BX,$$

is again distributed as a Student t with ν degrees of freedom, with

$$a + BX \sim St(a + B\mu, B\Sigma B'; \nu).$$

The proof follows the multivariate normal example, and the proof for nonsingular B is left to the reader as an exercise.

Much as in the case for multivariate normals, then, we have that portfolios of Student t distributed random variables with ν degrees of freedom are again distributed as Student t with ν degrees of freedom. Further, the isocontours of the Student t distribution are similarly defined.

2.3 Convergence Results and Estimators

Many are familiar with calculating the *sample mean* from observations $\{x_i\}_{i=1}^N$ as

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.38}$$

The case in which this is unassailably correct, however, is when each x_i is drawn from a distribution X_i, and $\{X_i\}_{i=1}^N$ are iid. But the above is really an instance of a random variable; namely

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

If the $\{X_i\}_{i=1}^N$ are iid, with mean, μ, and standard deviation, σ, we know from (2.21) that

$$\begin{aligned}\mathbb{E}(M_N) &= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(X_i\right)\\ &= \mu.\end{aligned}$$

What we don't know–yet–is how M_N is distributed. That is, we know that the expected value is what we'd hope it to be, but we don't have an indication of just how big N should be to have a reasonable estimate of the mean, say.

In fact, N disappeared in our calculation above. What we would hope is that as N gets large, M_N is distributed more 'tightly' around μ. Or, put another way, we hope that the probability that M_N is too far away from μ as N gets large gets arbitrarily small. This second formulation is at the heart of the *Weak Law of Large Numbers*, and is descriptive of *convergence in probability.*

We say that a sequence of real valued random variables, $\{X_i\}$ converges in probability to X if for any $\epsilon > 0$,

$$\lim_{N\to\infty}\mathbb{P}\left(|X_N - X| > \epsilon\right) = 0. \tag{2.39}$$

Theorem 2.3.1. (Weak Law of Large Numbers) Let $\{X_i\}$ be iid random variables taking values in $\mathbb{R}$, with $\mathbb{E}(X_i) = \mu$, and $Var(X_i) = \sigma^2$ for each i. Then

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

converges in probability to μ.

The proof resorts to two lemmas: *Markov's Inequality* and *Chebyshev's Inequality.* We prove each in turn and then return to prove 2.3.1.

Lemma 2.3.1. (Markov's Inequality) Let X be a nonnegative random variable, and a some positive constant. Then

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}(X).$$

Proof. We begin by introducing the indicator random variable

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

Since X is a positive random variable, we have

$$\begin{aligned} \mathbb{E}(X) &\geq \mathbb{E}\left(X \cdot 1_{X \geq a}\right) \\ &\geq a\mathbb{E}\left(1_{X \geq a}\right) \\ &= a\mathbb{P}\left(X \geq a\right). \end{aligned}$$

A geometric interpretation can help orient these steps. Recall that for $f(x)$ the density of X, $f(x) \geq 0$ for all x, and

$$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x f(x) dx \\ &= \int_0^a x f(x) dx + \int_a^\infty x f(x) dx \\ &\geq \int_a^\infty x f(x) dx \\ &\geq a \int_a^\infty f(x) dx \\ &= a\mathbb{P}\left(X \geq a\right). \end{aligned}$$

The latter proof is only shown for clarity, however, as the former is a stronger result, not requiring an assumption that X has a density, f.

Lemma 2.3.2. (Chebyshev's Inequality) Let X be a random variable taking values in $\mathbb{R}$ with $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$. Then

$$\mathbb{P}\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2}.$$

Proof. Clearly, for positive a ,

$$\mathbb{P}\left(|X - \mu| \geq a\right) = \mathbb{P}\left(|X - \mu|^2 \geq a^2\right).$$

Now, $(X - \mu)^2$ is a nonnegative random variable, we we may apply Markov's Inequality with $\mathbb{E}\left((X - \mu)^2\right)$ and get

$$\mathbb{P}\left(|X - \mu|^2 \geq a^2\right) \leq \frac{1}{a^2}\mathbb{E}\left((X - \mu)^2\right).$$

We know that $\mathbb{E}\left((X - \mu)^2\right) = \sigma^2$, so, putting these results together, we obtain

$$\mathbb{P}\left(|X - \mu| \geq a\right) \leq \frac{\sigma^2}{a^2},$$

proving the result.

Notice that in both lemmas that proceeded, no distributional assumptions were used or needed. In fact, the existence of the mean and variance (with finiteness of the former) were the only statistical properties assumed.

We are now in a position to prove the Weak Law of Large Numbers. We again assume that $\{X_i\}$ are a sequence of univariate iid random variables with mean μ and variance σ^2. We have already shown that for

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

that $\mathbb{E}(M_N) = \mu$. Using the fact that the variance of the sum of independent random variables is the sum of their variances, we also have

$$\begin{aligned} Var(M_N) &= Var\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \\ &= \frac{1}{N^2} Var\left(\sum_{i=1}^{N} X_i\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var\left(X_i\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 \\ &= \frac{1}{N}\sigma^2. \end{aligned}$$

Notice that the mean remained fixed while the variance, a measure of dispersion of M_N, scaled by $\frac{1}{N}$.

Using Chebyshev's Inequality, we have for any $\epsilon > 0$,

$$\mathbb{P}\left(|M_N - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{N}\frac{1}{\epsilon^2}.$$

For a fixed ϵ, then, the right hand side goes to zero as $N \to \infty$, completing the proof.

The Weak Law of Large Numbers is presented here as a way in which to introduce the concept of *estimators*. Many of the quantities we take as granted (e.g., μ and Σ in Modern Portfolio Theory) rely on estimated parameters in practice. As such, we are best served understanding their statistical properties. The Weak Law gives some indication about the distribution of $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$, or at least gives a convergence result.

We might still want to know the actual distribution of M_N. This would of course be a stronger result, and requires *convergence in distribution*. We say that a sequence of real valued random variables $\{X_i\}$ with distribution functions $\{F_i(\cdot)\}$, respectively, converges in distribution to X if

$$\lim_{N\to\infty} F_N(X) = F(X), \tag{2.40}$$

where $F(\cdot)$ is the distribution function of X. Convergence in probability implies convergence in distribution in the sense that if real valued random variables $\{X_i\}$ with distribution functions $\{F_i(\cdot)\}$ converge in probability to X,

$$X_N \to X$$

then

$$F_N \to F.$$

The *Central Limit Theorem* looks at the distribution of M_N and is a remarkable result. It states that for $\{X_i\}$ iid real random variables with finite mean and variance, μ and σ^2, respectively. Then the cumulative distribution function, F_N, of the random variable

$$Z_N = \frac{\sqrt{N}(M_N - \mu)}{\sigma}$$

satisfies

$$\lim_{N\to\infty} F_N(x) = \Phi(x). \tag{2.41}$$

That is, Z_N has a limiting normal distribution. This is regardless of the distribution of the X_i's. The proof is outside the scope of our current work.

2.3.1 Estimators and Bias

An estimator of a given quantity is simply a rule for calculating that quantity based on observed data. If we denote the quantity being estimated as θ and the estimator as $\hat{\theta}$, we define the *bias* of the estimator $\hat{\theta}$ as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta. \tag{2.42}$$

An *unbiased estimator* has zero bias and satisfies

$$\mathbb{E}(\hat{\theta}) = \theta.$$

We have already encountered an estimator of the mean, μ, of a sequence of iid random variables, $\{X_i\}$, finding that

$$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

satisfied $\mathbb{E}(M_N) = \mu$ immediately in our investigation. We now may state that the estimator M_N is an unbiased estimator of the mean.

Example 2.3.1. We next give an unbiased estimator of the variance of iid random variables, $\{X_i\}$, each with mean μ and variance σ^2. Let

$$s_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \hat{\mu})^2, \tag{2.43}$$

with $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ as before. Then

$$\mathbb{E}(s_N) = \sigma^2.$$

The term $1/(N-1)$ is somewhat surprising on first pass. It will be slightly more intuitive when we encounter the estimate of variance again in ordinary least squares where degrees of freedom will be given more color as well. In the meantime, the current proof shows that we require $N-1$ in the denominator to ensure that our estimator is unbiased.

Proof. Let X be distributed as each of the X_i, and look at

$$\mathbb{E}\left(\sum_{i=1}^{N}(X_i - \hat{\mu})^2\right) = \sum_{i=1}^{N}\mathbb{E}\left((X_i - \hat{\mu})^2\right),$$

and notice that for each summand

$$\mathbb{E}\left((X_i - \hat{\mu})^2\right) = \mathbb{E}(X_i^2) - 2\mathbb{E}(X_i\hat{\mu}) + \mathbb{E}(\hat{\mu}^2),$$

so that the sum becomes

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{N}(X_i - \hat{\mu})^2\right) &= \sum_{i=1}^{N}\mathbb{E}(X_i^2) - 2\sum_{i=1}^{N}\mathbb{E}(X_i\hat{\mu}) + \sum_{i=1}^{N}\mathbb{E}(\hat{\mu}^2) \\
&= N\mathbb{E}(X^2) - 2\sum_{i=1}^{N}\mathbb{E}(X_i\hat{\mu}) + N\mathbb{E}(\hat{\mu}^2) \\
&= N\mathbb{E}(X^2) - 2\mathbb{E}(N\hat{\mu}\hat{\mu}) + N\mathbb{E}(\hat{\mu}^2) \\
&= N\mathbb{E}(X^2) - 2N\mathbb{E}(\hat{\mu}^2) + N\mathbb{E}(\hat{\mu}^2) \\
&= N\left(\mathbb{E}(X^2) - \mathbb{E}(\hat{\mu}^2)\right)
\end{aligned}$$

We look at $\mathbb{E}(X^2)$ and $\mathbb{E}(\hat{\mu}^2)$ in turn.

Now, we know that $Var(X) = \mathbb{E}\left(X^2\right) - \mu^2$, giving

$$\mathbb{E}\left(X^2\right) = \sigma^2 + \mu^2.$$

Looking at $\mathbb{E}(\hat{\mu}^2)$ next, we require the variance and mean of $\hat{\mu}^2$ and apply the same relationship. We have already established that

$$\begin{aligned}
\mathbb{E}(\hat{\mu}) &= \mu \\
Var(\hat{\mu}) &= \frac{1}{N}\sigma^2.
\end{aligned}$$

This gives that

$$\mathbb{E}\left(\hat{\mu}^2\right) = \frac{1}{N}\sigma^2 + \mu^2.$$

Putting the above together, we have

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{N}(X_i-\hat{\mu})^2\right) &= N\left(\mathbb{E}(X^2)-\mathbb{E}(\hat{\mu}^2)\right) \\
&= N\left(\sigma^2+\mu^2-\left(\frac{1}{N}\sigma^2+\mu^2\right)\right) \\
&= N\left(\frac{N-1}{N}\sigma^2\right) \\
&= (N-1)\sigma^2.
\end{aligned}$$

Dividing both sides by $N-1$ proves the result, and s_N is an unbiased estimator of the variance of the iid random variables X_i.

We have surreptitiously seen estimators in our work already. Whenever a fit of a distribution has been shown, we have estimated the mean and variance (and covariance in one case) of the underlying distributions using the above estimators. Implicit in this estimation is that the random variables under consideration are iid. Specifically, we repeatedly assumed, then, that daily log returns of the S&P are independent. This is actually less of a sure assumption than one might have guessed.

Amir Khandani and Andrew Lo [17] exhibit that a stock trading strategy that buys the last day's losers and sells the last day's winners is profitable – and their result is fairly robust within the context they study. If returns were independent, such a strategy should return something indistinguishable from zero. The fact that this is not the case implies that there is some dependence between yesterday's returns and todays. A simple codification of the finding would be something like

$$r_t = \phi r_{t-1} + \epsilon_t$$

for the most extreme winners and losers each day in the market. The equation above is an example of an *autoregressive* model, and it says that far from being independent, there is a structural form between daily returns.

Before incorporating your new hedge fund in Delaware where you plan to exploit the strategy just exhibited, know this: it won't work. Short time scale strategies that center around daily close prices are virtually impossible to execute for two reasons in particular: 1) a massive amount of each day's trading volume is centered around the closing bell, so that achieving a fill on the exact close price is either very tricky or very expensive; and 2) a significant amount of the variation in stock prices comes between the close and the open (that is, when you can't trade) so that waiting till the open to get your fills downgrades the strategy to random noise.

Knowing that the strategy isn't viable in terms of execution does not negate its usefulness, however. For example, Lo and Khandani look at how this simplistic strategy did during the (wholly improbable) Quant Crisis we have alluded to previously – and it fared terribly. Performance over other periods was surprisingly stellar, and the crisis days in August 2007 were outliers for the strategy.

Their reversion strategy, then, provided a simple template to examine observations in the market through an interpretable lens.

While we won't present a solution to the apparent lack of independence of daily returns, we discuss these results to give some insight into the tension between estimation in practice, theory, and empirical findings.

2.3.2 Consequences and Toy Models: CAPM Lite

A couple observations are in order at this point. We have seen previously in examples like Figures 2.1.3 and 2.2 and their related discussions that understanding a random variable's distribution is no easy task. The Weak Law of Large Numbers and the Central Limit Theorem prove results about estimation and distribution of sequences of random variables. In the context of math finance or portfolio management, our mind wanders to an application of these powerful results.

Consider a model where each stock's returns, r_i, are iid with mean and standard deviation, μ_S and σ_S, respectively. The Central Limit Theorem implies that the mean and variance of the return of an evenweighted portfolio,

$$r_\Pi = \frac{1}{N}\sum_{i=1}^{N} r_i$$

will, in the limit as N gets large be normal with mean μ and standard deviation $\frac{1}{\sqrt{N}}\sigma$. That is, in this toy example we could maintain a fixed level of return, μ, while reducing uncertainty as measured by volatility to nearly zero. Of course, this isn't possible in practice since all stocks aren't identical, but as a prologue, we have some motivation to consider diversification as such when constructing a portfolio.

Figure 2.2 gives the next version of a toy model to consider. We might assume that a given stock's returns are comprised of a weighted systemic piece and an idiosyncratic piece, both random variables, with the systemic and idiosyncratic pieces being pairwise independent. Something like

$$r_i = \beta_i m + \epsilon_i, \tag{2.44}$$

where in the case of the figure, we are saying that IBM's returns (r) are linearly related to the returns to the S&P 500 (m), plus some error. We see that this is really just an approximation used for insight and not a model derived from empirical validation, but the implications are interesting.

For example, codifying what we mean by systemic and idiosyncratic, we may for instance assume that

$$\begin{aligned} \mathbb{E}(\epsilon_i) &= 0 \\ Var(\epsilon_i) &= \sigma_i^2 \\ Cov(m, \epsilon_i) &= 0 \\ Var(m) &= \sigma_m^2. \end{aligned}$$

An evenweight portfolio's returns in this case look like

$$\begin{aligned} r_\Pi &= \frac{1}{N}\sum_{i=1}^{N} r_i \\ &= \frac{1}{N}\sum_{i=1}^{N} \beta_i s + \epsilon_i \\ &= \frac{s}{N}\sum_{i=1}^{N} \beta_i + \frac{1}{N}\sum_{i=1}^{N} \epsilon_i. \end{aligned}$$

The mean and variance of r_Π, using (2.21) and (2.24) along with the above assumptions, are

$$\begin{aligned} \mathbb{E}(r_\Pi) &= \frac{\sum_i \beta_i}{N}\mathbb{E}(m) \\ Var(r_\Pi) &= \left(\frac{\sum_i \beta_i}{N}\right)^2 \sigma_m^2 + \frac{1}{N^2}\sum_{i=1}^{N} \sigma_i^2, \end{aligned}$$

so that if, say $\sum_i \beta_i \approx N$ (a reasonable assumption as we shall see later) we have that the evenweight portfolio has returns similar to the market, but the idiosyncratic variance is greatly decreased (imagine all of the $\sigma_i \approx \sigma$ for some σ, then the idiosyncratic variance is $\frac{1}{N}$ what it would be holding any single name).

Of course, the attractiveness of reducing idiosyncratic noise raises the question: Why not just own the market, then? We will encounter this question again when we have more tools under our belt.

Exercises

1. An odd function is such that $f(x) = -f(-x)$. Suppose that $f(\cdot)$ is integrable and odd. Show that
$$\int_{-\infty}^{\infty} f(x)dx = 0$$

2. For a random variable, X, show using the density function approach that
$$\begin{aligned} \mathbb{E}(aX+b) &= a\mathbb{E}(X)+b \\ Var(aX+b) &= a^2 Var(X) \end{aligned}$$
for scalars a and b.

3. From the relationship
$$Var(X) = \mathbb{E}\left((X-\mu)^2\right),$$
show that
$$Var(X) = \mathbb{E}\left(X^2\right) - \mu^2.$$

4. Prove that the variance of $X \sim N(\mu, \sigma^2)$ is σ^2.

5. Show using a change of variables to polar coordinates that
$$\left(\int_{\mathbb{R}} e^{-\frac{u^2}{2}} du\right)^2 = 2\pi.$$

6. Using the change of variables theorem directly, derive the log-normal density $ln(\mu, \sigma^2)$.

7. Show that the skew of a symmetric random variable with a density is 0.

8. Show that the mean and variance of $Y \sim LN(\mu, \sigma^2)$ are given by
$$\begin{aligned} \mathbb{E}(Y) &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ Var(Y) &= \exp\left(2\mu+\sigma^2\right)\left(\exp\left(\sigma^2\right)-1\right). \end{aligned}$$

9. If $T \sim \text{St}(0, 1; \nu)$, how is $\tilde{T} = \mu + \sigma T$ distributed for μ and σ scalars?

10. Prove that both the mean and median of $T \sim \text{St}(\mu, \sigma^2; \nu)$ are μ.

11. Using the S&P daily log return data, $\{r_t\}_{t=1}^N$, and estimating the mean and standard deviation of these log returns by
$$\begin{aligned} \hat{\mu} &= \frac{1}{N}\sum_{t=1}^{N} r_t \\ \hat{\sigma}^2 &= \frac{1}{N-1}\sum_{t=1}^{N}(r_t - \hat{\mu})^2, \end{aligned}$$

we may approximate the empirical distribution with a Student t distribution, $T \sim \text{St}(\hat{\mu}, \tilde{\sigma}; 5)$. What should $\tilde{\sigma}$ be so that $Var(T) = \hat{\sigma}$? Replicate Table 2.1.2, replacing the theoritcal density with the Student t distribution you just constructed.

12. Generalize (2.21), the linearity of the expectation operator, assuming the case of two random variables is already proved and using induction.

13. If X and Y are independent univariate random variables with a joint density, prove that $Cov(X, Y) = 0$, and hence $Var(X + Y) = \sigma_X^2 + \sigma_Y^2$. Generalize this result for N independent univariate random variables:

$$Var\left(\sum_{i=1}^{N} w_i X_i\right) = \sum_{i=1}^{N} w_i^2 Var\left(X_i\right)$$

14. Show that

$$\mathcal{I}_c = \left\{x \,|\, (x-\mu)'\Sigma^{-1}(x-\mu) = c\right\}$$

is an isocontour of $\phi_{\mu,\Sigma}(\cdot)$; that is $\phi_{\mu,\Sigma}(\cdot)$ is constant on $\mathcal{I}_c$.

15. Show that an affine transformation of $X \in \mathbb{R}^N$ with $X \sim St(\mu, \Sigma; \nu)$,

$$a + BX,$$

is distributed as a Student t with ν degrees of freedom, with

$$a + BX \sim St(a + B\mu, B\Sigma B'; \nu).$$

Chapter 3

Covariance

Ellipsoidal distributions like the multivariate normal and Student t distribution play a pivotal role in modeling financial data and at their heart is the covariance matrix, Σ. In this chapter, we discuss some key properties, empirical observations, and current research work.

The structure of the covariance matrix defines a language for the pairwise relationships in the market. We identify the mathematical framework for understanding correlation and covariance–namely the inner product–and utilize Cauchy-Schwarz to provide bounds on the former. We will also briefly discuss the relationship between correlation and independence and give an example regarding the normal distribution. Additional results follow from our interpretation of covariance as an inner product on the space of identified random variables including another Capital Asset Pricing Model (CAPM) vignette.

Some background in the theory of eigenvalues and eigenvectors for a general $N \times N$ matrix, and results specific to positive definite matrices are necessary. With regards to covariance matrices, we relate the sum of eigenvalues to what we define as total variance, and define the square root of a positive definite matrix and relate this result to Monte Carlo simulations, as alluded to in the previous chapter. Finally, we introduce the concept of the condition number and provide some motivation for its value.

Empirical examples of the covariance matrix will center around the frequency of observed eigenvalues for a correlation matrix at a given time for a large number of equities. These observations will echo the literature and can be summarized as follows: the largest eigenvalue explains an inordinate amount of total variance, and its associated eigenvector (read eigenportfolio) is a proxy for the market as such, giving credence to the simple CAPM model; several other eigenvalues are related to sector specific drivers, and are fairly large as well; and a bulk of eigenvalues are near zero and have characteristics of a random distribution.

Finally, we will present the theory of copulas, an essential tool for imposing a covariance structure given a set of prescribed marginals. Our work here is not extensive, but should provide the practitioner with the theory and tools

for implementation. We provide a few words of caution, however, related to the application of the method and some of its shortcomings during, say, the financial crisis of 2008.

3.1 Covariance and Correlation

As before (2.20) we define the covariance of a random variable $X \in \mathbb{R}^N$ with expectation μ as

$$Cov(X) = \mathbb{E}\left((X-\mu)(X-\mu)'\right),$$

giving an $N \times N$ real valued matrix in the case where $Cov(X)$ is defined, and we deonote the (i,j) component of $Cov(X)$ as σ_{ij}, and $\sigma_{ii} = Var(X_i)$.

In particular, for two real valued random variables, X and Y, we may introduce the notation $Cov(X,Y)$ where

$$Cov(X,Y) = \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right), \tag{3.1}$$

with μ_X and μ_Y being the expectation of X and Y, respectively. Relating this notation to the previous lines, we have

$$Cov(X,Y) = \sigma_{XY}.$$

It follows that there exists some ρ_{XY} such that,

$$\sigma_{XY} = \rho_{XY}\sigma_X\sigma_Y.$$

That is, the covariance between two random variables may be written as some scalar times the product of their standard deviations.

Far from being arbitrary, ρ_{XY} is a highly interpretable parameter called the *correlation* between X and Y. It is a measure of linear dependence, and is bounded between

$$-1 \leq \rho_{XY} \leq 1,$$

with the extremes being special cases in the relationship between X and Y.

To derive the above, we identify covariance as an *inner product*. For our purposes, we consider an inner product $(\cdot\,,\,\cdot)$ to be a function of the cross product of a real valued vector space, V, over $\mathbb{R}$ into $\mathbb{R}$,

$$(\cdot\,,\,\cdot) : V \times V \to \mathbb{R}$$

satisfying the following criteria for $X, Y, Z \in V$, and $a, b \in \mathbb{R}$:

- *symmetry*

$$(X,Y) = (Y,X) \tag{3.2}$$

- *bilinearity*

$$\begin{aligned} (aX,Y) &= a(X,Y) \\ (X+Y,Z) &= (X,Z)+(Y,Z) \end{aligned} \tag{3.3}$$

- *non-negativity*

$$(X, X) \quad \geq \quad 0 \tag{3.4}$$

  with equality only if $X \equiv 0$.

The definition also implies that for scalars $\{a_i\}_i$ and $\{b_j\}_j$, and $\{X_i\}_i$ and $\{Y_j\}_j$ in V,

$$\left(\sum_i a_i X_i, \sum_j b_j Y_j \right) = \sum_i a_i \sum_j b_j (X_i, Y_j).$$

This is left to the reader as an exercise.

Example 3.1.1. For $V = \mathbb{R}^N$, and $A \in \mathbb{R}^{N \times N}$,

$$(x, y)_A = x' A y$$

is an inner product when A is symmetric and positive definite. This immediately follows by checking the criteria of the definition. We have

$$\begin{aligned} (x, y)_A &= x' A y \\ &= (x' A y)' \\ &= y' A' x \\ &= y' A x, \end{aligned}$$

and for $a \in \mathbb{R}$,

$$\begin{aligned} (ax, y)_A &= a x' A y \\ &= a (x, y)_A \end{aligned}$$

and

$$\begin{aligned} (x + y, z)_A &= (x + y)' A z \\ &= x' A z + y' A z. \end{aligned}$$

Finally $(x, y)_A$ is positive definite exactly when A is positive definite.

We may consider the vector space of real valued random variables over $\mathbb{R}$. As in (2.25), we have

$$Cov(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y) \right).$$

We have already shown in (2.21) that the expectation operator is linear, so that we have

$$\begin{aligned} Cov(aX, Y) &= \mathbb{E}\left(a(X - \mu_X)(Y - \mu_Y) \right) \\ &= a \mathbb{E}\left((X - \mu_X)(Y - \mu_Y) \right) \\ &= a Cov(X, Y). \end{aligned}$$

Similarly, $Cov(\cdot\,,\,\cdot)$ is symmetric and bilinear.

We are left to show non-negativity. We have, clearly, that

$$Cov(X, X) = Var(X),$$

and we know that variance is a nonnegative number. In fact, looking at the definition of variance in (2.5), we see that the quantity is only zero when $X \equiv \mu_X$. That is, $Var(X) = 0$ if and only if X is a constant random variable. This gives that covariance is an inner product on the space of real valued random variables over $\mathbb{R}$ where constants are identified. We call such a space where a particular relation defines an identification a *quotient space*.

We may also identify when the covariance matrix of $X \in \mathbb{R}^N$ is positive definite. We have shown already that

$$Var(w'X) = Cov(w'X, w'X) = w'\Sigma w.$$

Hence Σ is positive definite only if $Var(w'X) = 0$ implies $w'X$ is a constant. We will say that the random variable X is *linearly independent*, if $w'X = c$, a constant, implies $w \equiv 0$. From this discussion, then, Σ is positive definite if and only if X is linearly independent.

As we shall soon see, this criterion is easily violated when constructing the sample covariance (i.e., an unbiased estimator of Σ from data). Consider the case of N assets and T observations. Whenever $N > T$, we will obtain semidefinite sample covariance matrices as our system is underdetermined. Other issues arise and we will get to them shortly.

With the proof that $Cov(\cdot\,,\,\cdot)$ is an inner product on the quotient space of real random variables in hand, we are ready to discuss correlation. The *Cauchy-Schwarz inequality* states that for an inner product $(\cdot\,,\,\cdot)$,

$$|(u, v)| \leq ||u|| \cdot ||v||, \tag{3.5}$$

with equality only if $u = av$ for a a scalar, and where

$$||u||^2 = (u, u) \tag{3.6}$$

is defined as the *norm* of u induced by the inner product $(\cdot\,,\,\cdot)$. The proof of the Cauchy-Schwarz inequality is in the appendix.

In general, a norm is a function $||\cdot|| : V \to \mathbb{R}_+$ satisfying, for all $X, Y \in V$, and scalars a,

- *absolute scalability*
$$||aX|| = |a| \cdot ||X|| \tag{3.7}$$
- *the triangle inequality*
$$||X + Y|| \leq ||X|| + ||Y|| \tag{3.8}$$
- *non-negativity*
$$||X|| \geq 0 \tag{3.9}$$
with equality only in the case that $X \equiv 0$.

We leave the proof that $||\cdot||^2 = (\cdot\,,\,\cdot)$ defines a norm to the reader.

From the above, utilizing (3.5) we have that

$$\begin{aligned} |Cov(X,Y)| &\leq \sqrt{Cov(X,X)}\sqrt{\cdot Cov(Y,Y)} \\ &= Var(X)\cdot Var(Y), \end{aligned}$$

or in the previous notation

$$|\sigma_{XY}| \leq \sigma_X\sigma_Y.$$

We may therefore define the *correlation between X and Y* as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}. \tag{3.10}$$

We may write $Corr(X,Y) = \rho_{XY}$ to emphasize the role of correlation as an operator. As one of the consequences of Cauchy-Schwarz, we know further that

$$-1 \leq \rho_{XY} \leq 1$$

with equality only in the case that $X = aY$ for some a in the quotient space of random variables over $\mathbb{R}$.

We say two random variables are positively (negatively) correlated if their correlation coefficient, ρ, is positive (negative). In the case that correlation is zero, we say that these random variables are *uncorrelated*. Notice, too, that correlation is dimensionless (whereas volatility has the dimensions of the original random variable). Finally, correlation is a measure of linear dependence as we shall see in the next example which revisits a version of CAPM once more.

Example 3.1.2. Consider the model

$$r = \beta m + \epsilon$$

where m, and ϵ are each random variables in $\mathbb{R}$ satisfying

$$\begin{aligned} Cov(m,\epsilon) &= 0 \\ Var(\epsilon) &= \sigma_\epsilon^2 \\ Var(m) &= \sigma_m^2. \end{aligned}$$

Letting $(\cdot\,,\,\cdot)$ denote $Cov(\cdot\,,\,\cdot)$, we may take the inner product of both sides with respect to m and obtain

$$(r,m) = \beta(m,m) + (\epsilon,m),$$

where we have used the bilinearity of the inner product on the right hand side. Simplifying based on our assumptions, we have

$$(r,m) = \beta(m,m),$$

or,

$$\frac{\sigma_{r,m}}{\sigma_m^2} = \beta$$

which yields

$$\beta = \rho \frac{\sigma_r}{\sigma_m} \tag{3.11}$$

with ρ the correlation between r and m.

The preceding model is the core definition of the Capital Asset Pricing Model. Notice that we haven't *estimated* any of these quantities or validated their applicability. Rather, we began with a model and obtained a consequence of our assumptions. Adding interpretation to assumptions, r represents a particular stock's returns, m the market's (whatever that may be) returns, and ϵ the so-called idiosyncratic component of that stock.

The model says, then, that stock returns have a linear relationship with the market's returns, with some variation defined by ϵ. Notice that we have not as yet prescribed a distribution for ϵ. We will. When we do, this will limit the scope of the model but open us to more interpretation and analysis.

We also have not discussed the relationship between multiple assets under the model. This too will come, but the astute reader can likely derive the consequences now by generalizing the above and applying the term idiosyncratic appropriately.

The above example highlights some of the features of a model based on CAPM. In addition, we saw in (3.11) that β, the linear coefficient of the model, was a scalar multiple of ρ, clearly establishing that correlation is a measure of linear dependence between two random variables. It is important to point out, then, that in practice, observations of small correlation (in absolute value) do not preclude a strong relationship between two random variables; viz., a near perfect quadratic relationship may exist with a close to zero correlation.

We have already seen an unbiased estimator of variance (2.43). For covariance, we have that

$$\hat{\sigma}_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_i - \hat{\mu}_X\right)\left(Y_i - \hat{\mu}_Y\right) \tag{3.12}$$

is an unbiased estimator as well (this is left as an exercise). An estimate of correlation can be obtained from these estimators by replacing population with sample moments in (3.10); viz.,

$$\hat{\rho}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}. \tag{3.13}$$

In general, the above technique is called the *method of moments*, and for our purposes can be summarized as an estimator which equates sample moments to population moments in a formula for a given distribution parameter.

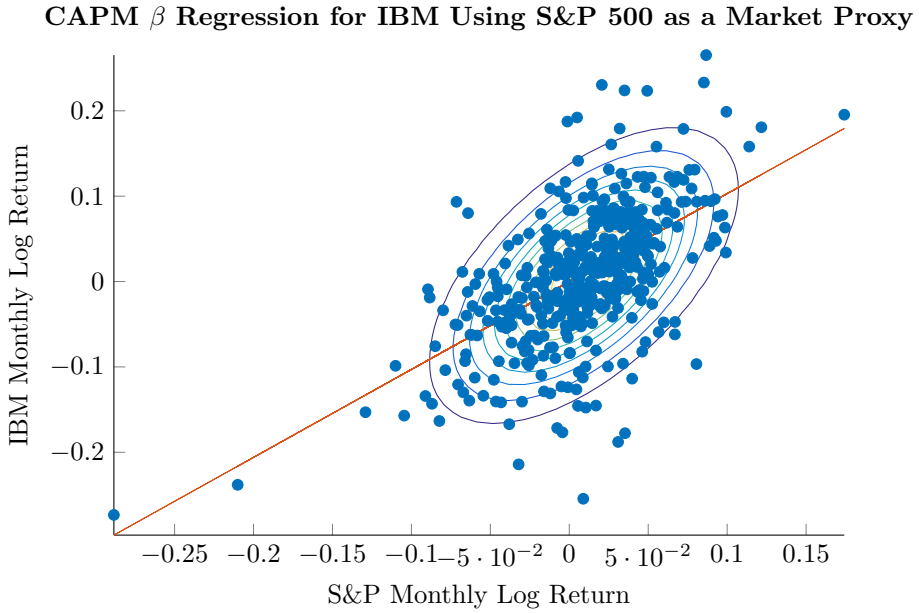

Figure 3.1: Empirical joint distribution of log returns of the S&P 500 from 1980 through 2015 [25] with estimated $\hat{\beta}$ of 1.03.

Example 3.1.3. The method of moments estimator of variance is

$$s^2_{m,N} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{\mu})^2,$$

which we know from (2.43) is a biased estimator. Even so, it is a *consistent estimator*; i.e.,

$$s^2_{m,N} \to \sigma^2$$

in probability.

The above example shows that a method of moments estimator may be biased (and likely is so), but consistent (and again, under light constraints likely is so). The correlation estimator presented here is both biased and consistent.

A method of moments estimator for β in the CAPM model above is given by

$$\hat{\beta} = \hat{\rho}\frac{\hat{\sigma}_r}{\hat{\sigma}_m}. \tag{3.14}$$

We shall see in our work on ordinary least squares (OLS) that this is exactly the OLS estimate of β, and that it is in fact unbiased as well.

Example 3.1.4. We have already seen a version of the relationship that motivates the CAPM approach in Figure 2.2. We may now calculate $\hat{\rho}$ and $\hat{\beta}$ using

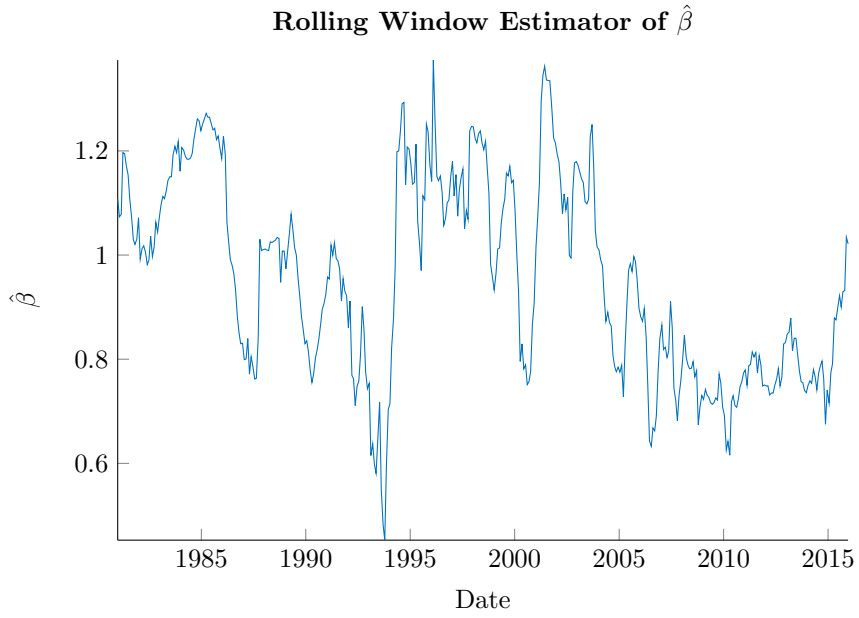

Figure 3.2: Rolling estimate of $\hat{\beta}$ for IBM returns based on S&P 500 from 1980 through 2015.

the whole sample. We update the original figure in Figure 3.1 to show the regression line given by the estimated values below:

$$\begin{aligned} \hat{\rho} &= 0.59 \\ \hat{\sigma}_{IBM} &= 0.0787 \\ \hat{\sigma}_{SPX} &= 0.0443 \\ \hat{\beta} &= 1.03. \end{aligned}$$

Notice that the slope of the regression line is not parallel to the major axis of the ellipses defined by the covariance matrix. While this is visually jarring, it is by design as we shall see that $\hat{\beta}$ is the result of minimizing squared errors in the y-axis dimension. Other questions arise as well:

- If we look at rolling time windows, will the results be substantially different?
- Does our estimate of $\hat{\beta}$ change through time?
- Is it correlation or the ratio of vols that drives the variation in $\hat{\beta}$?

We treat the first two points as purely optical. We see in Figure 3.1 what a rolling window using 252 trading days (approximately one year) of data yields for our estimate of $\hat{\beta}$.

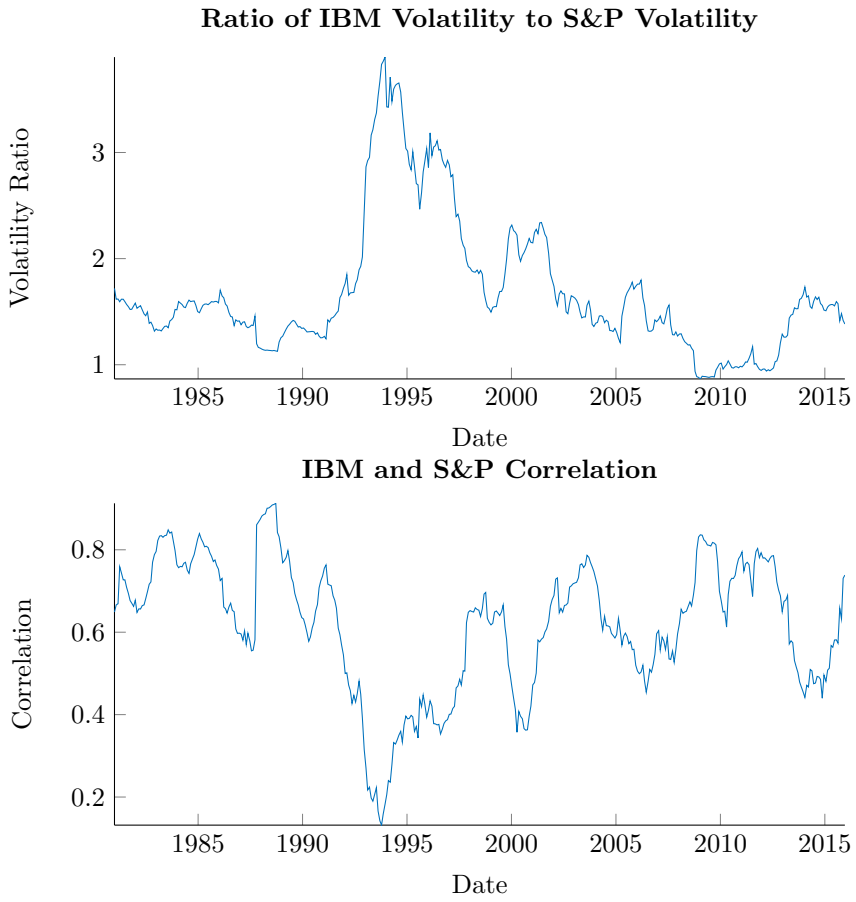

Figure 3.3: Empirical joint distribution of log returns of the S&P 500 from 1980 through 2015 with estimated $\hat{\beta}$ of 1.03.

We clearly identify that our estimate varies significantly through time. There is perhaps a mean-reverting component as well. In fact, services like Bloomberg offer up so-called shrinkage estimators of β that attempt to capture this very phenomenon, calculating their own estimator $\tilde{\beta}$ as

$$\tilde{\beta} = 0.67\hat{\beta} + 0.33.$$

Clearly this would be a biased estimator of β. The model reflects some desirable properties of the dynamics shown however. In particular, the dynamics implied in the above are that a company's β to the market should center around $\beta = 1$.

Interestingly, the bursting of the tech bubble in early 2000 is visible in the dramatic uptick of $\hat{\beta}$. This is less pronounced, however, in the financial crisis in 2008. A reasonable observation is that there may be sector-specific exposures that impact a company in addition to the market as such.

Focusing on these time periods a bit more, we have yet to distinguish whether the driver of $\hat{\beta}$ dynamics is correlation or volatility. The upper Figure 3.1 shows a plot of the ratio of estimated volatilities of IBM and the S&P, while the lower figure shows estimated correlation between the two. From a visual inspection, again, the two crises exhibit different behavior. The tech bubble shows an increase in the volatility ratio as well as an increase in correlation, while the financial crisis shows a decrease in the former and increase in the latter. In both cases, we see that in the event of a crisis, there is some evidence to expect that correlations increase between securities. This is yet another stylized feature of equity returns.

3.1.1 Correlation and Independence

Correlation and independence are related but distinct properties between random variables. It is easy to show that if two univariate random variables, X and Y, are independent, then they are uncorrelated; viz.,

$$\begin{aligned} \rho_{XY} &= \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right) \\ &= \mathbb{E}\left(X-\mu_X\right)\mathbb{E}\left(Y-\mu_Y\right) \\ &= (\mu_X-\mu_X)(\mu_Y-\mu_Y) \\ &= 0. \end{aligned}$$

The converse is not always true, however. One exception is in the case of jointly normal random variables. The proof is a bit more involved than the prior statement, however.

Theorem 3.1.1. (Multivariate Normality, Correlation, and Independence) If $X \sim N(\mu_X, \Sigma_X)$ and $Y \sim N(\mu_Y, \Sigma_Y)$ are each multivariate normal random variables which are jointly normal and uncorrelated, then X and Y are independent.

Proof. Let

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix},$$

with X and Y distributed as above. Then we have $Z \sim N(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix},$$

and

$$\Sigma = \begin{pmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{pmatrix}$$

since expectation is taken component-wise and the elements of X and Y are uncorrelated by assumption. We show the independence of X and Y by writing the probability density function of Z as the product of the densities for X and Y.

We have

$$\begin{aligned}\phi_{\mu,\Sigma}(z) &= \frac{1}{\det(\Sigma)^{1/2}}\frac{1}{(2\pi)^{N/2}}\cdot \\ &\quad \exp\left(-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right),\end{aligned}$$

where $N = dim(Z) = dim(X) + dim(Y)$. We will denote the dimensions of X an Y by N_X and N_Y, respectively.

By the structure of Σ, we have that the determinant is given by

$$\det(\Sigma) = \det(\Sigma_X)\cdot\det(\Sigma_Y)$$

and the inverse is given by

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_X^{-1} & 0 \\ 0 & \Sigma_Y^{-1}. \end{pmatrix}$$

Putting this together, we have, referring to the components of the density of Z in turn,

$$\frac{1}{\det(\Sigma)^{1/2}} = \frac{1}{(\det(\Sigma_X)\det(\Sigma_Y))^{1/2}} = \frac{1}{\det(\Sigma_X)^{1/2}}\frac{1}{\det(\Sigma_Y)^{1/2}},$$

$$\frac{1}{(2\pi)^{N/2}} = \frac{1}{(2\pi)^{(N_X+N_Y)/2}} = \frac{1}{(2\pi)^{N_X/2}}\frac{1}{(2\pi)^{N_Y/2}}.$$

The pattern continues. Looking now at the exponential, we have

$$\exp\left(-\tfrac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right) =$$

$$\exp\left(-\tfrac{1}{2}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right)'\Sigma^{-1}\left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right)\right) =$$

$$\exp\left(-\tfrac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}'\Sigma^{-1}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right)$$

$$\exp\left(-\tfrac{1}{2}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}'\begin{pmatrix} \Sigma_X^{-1} & 0 \\ 0 & \Sigma_Y^{-1}. \end{pmatrix}\begin{pmatrix} x-\mu_x \\ y-\mu_y \end{pmatrix}\right)$$

$$\exp\left(-\tfrac{1}{2}\left((x-\mu_X)'\Sigma_X(x-\mu_X) + (y-\mu_Y)'\Sigma_Y(y-\mu_Y)\right)\right)$$

$$\exp\left(-\tfrac{1}{2}(x-\mu_X)'\Sigma_X(x-\mu_X)\right)\exp\left(-\tfrac{1}{2}(y-\mu_Y)'\Sigma_Y(y-\mu_Y)\right).$$

Therefore,

$$\begin{aligned}\phi_{\mu,\Sigma}(z) = \phi_{\mu,\Sigma}(x,y) &= \frac{1}{\det(\Sigma_X)^{1/2}}\frac{1}{\det(\Sigma_Y)^{1/2}}\frac{1}{(2\pi)^{N_X/2}}\frac{1}{(2\pi)^{N_Y/2}}\cdot \\ &\quad \exp\left(-\frac{1}{2}(x-\mu_X)'\Sigma_X^{-1}(x-\mu_X)\right)\cdot \\ &\quad \exp\left(-\frac{1}{2}(y-\mu_Y)'\Sigma_Y^{-1}(y-\mu_Y)\right) \\ &= \phi_{\mu_X,\Sigma_X}(x)\phi_{\mu_Y,\Sigma_Y}(y),\end{aligned}$$

proving X and Y are independent.

The astute reader may have noticed that care was taken in the above example to explicitly say *jointly* normal random variables. In fact, it is not remarkable to have random variables X_1 and X_2, for instance which are each normally distributed but whose joint distribution is not normal. In fact, due to a result by Sklar [32], we know that for a set of univariate random variables, $\{X_i\}_{i=1}^N$, with marginal distribution functions, $F_i(\cdot)$, any joint distribution may be constructed which respects the marginal distributions prescribed. So, for instance, it is possible to have normal marginals with a Student t joint distribution [13]. Further, this is a constructive procedure which we establish in the next section on copulas.

3.2 Copulas

A *copula* is the joint distribution of random variables, $\{U_i\}_{i=1}^N$, each of which is uniformly distributed on $[0,1]$. We say that a univariate random variable is uniformly distributed on $[0,1]$ if

$$\mathbb{P}(U \leq x) = x \tag{3.15}$$

for $x \in [0,1]$, and denote this by $U \sim U([0,1])$. Clearly in this case, $F_U(x) = x$, and the probability density is simply $f_U(x) = 1$. We will denote a copula by C, and based on the above, we must have

$$C(u_1, \ldots, u_N) = \mathbb{P}(U_1 \leq u_1, \ldots, U_N \leq u_N). \tag{3.16}$$

As mentioned previously, *Sklar's Theorem* states that for any random variables, $\{X_i\}_{i=1}^N$ with marginals $F_i(\cdot)$ and joint distribution $F(\cdot)$, there exists a copula, C, such that

$$F(x_1, \ldots, x_N) = C\left(F_1(x_1), \ldots, F_N(x_N)\right), \tag{3.17}$$

and that if the F_i's are unique, then so is C. This is a powerful result, but one that can be proven readily in the continuous case. We leave to the reader to prove the fact that

$$F_i(X_i) \sim U([0,1]). \tag{3.18}$$

With this result in hand, we simply write the joint distribution in Sklar's Theorem in terms of a copula, C, as

$$\begin{aligned} F(x_1, \ldots, x_N) &= \mathbb{P}(X_1 \leq x_1, \ldots, X_N \leq x_N) \\ &= \mathbb{P}\left(F_1(X_1) \leq F_1(x_1), \ldots, F_N(X_N) \leq F_N(x_N)\right) \\ &= \mathbb{P}\left(U_1 \leq F_1(x_1), \ldots, U_N \leq F_N(x_N)\right) \\ &= C\left(F_1(x_1), \ldots, F_N(x_N)\right). \end{aligned}$$

For a specified, $F(\cdot)$, then, and continuous, $F_i(\cdot)$, we may construct a copula as

$$C\left(u_1, \ldots, u_N\right) = F\left(F_1^{-1}(u_1), \ldots, F_N^{-1}(u_N)\right). \tag{3.19}$$

The above result says that if we so desire, we may divorce the joint and marginal densities because the linking may be done entirely through some copula. Or, as so happens in practice, we may specify marginal distributions and a joint distribution separately.

Example 3.2.1. Let X_1 and X_2 be distributed as standard normal random variables, and let $St_{\mu,\Sigma;\nu}(\cdot)$ be the joint distribution of a two dimensional Student t distribution with ν degrees of freedom. Then

$$C(u_1, u_2) \quad = \quad St_{\mu,\Sigma;\nu}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right)$$

or

$$C(\Phi(x_1), \Phi(x_2)) \quad = \quad St_{\mu,\Sigma;\nu}\left(x_1, x_2\right)$$

exhibits a pair of random variables, jointly Student t, and with marginals that are standard normal.

Even more, for ellipsoidal distributions, we may focus solely on correlation, disregarding both position and scale; i.e., expectation and variance. This is due to the fact that copula's have a so-called *rank-invariant property*, namely, if $g_i(\cdot)$ are each strictly increasing functions

$$g_i : \mathbb{R} \mapsto \mathbb{R},$$

for $i = 1, \ldots, N$, and C is the copula of $\{X_i\}$ as in (**??**), then C is also the copula of $\{g_i(X_i)\}$.

Proof. Let $F(\cdot)$ be the joint distribution function of X, a multivariate random variable and $g_i(\cdot)$ strictly increasing functions from $\mathbb{R}$ to $\mathbb{R}$. We know by the change of variable theorem (2.14), that the CDF of $g_i(X_i)$ is

$$\tilde{F}_i(\cdot) = F_i \circ g_i^{-1}(\cdot) = F_i\left(g_i^{-1}(\cdot)\right)$$

whose inverse is

$$\tilde{F}_i^{-1}(\cdot) = g_i \circ F_i^{-1}(\cdot) = g_i\left(F_i^{-1}(\cdot)\right).$$

Now, denoting the CDF of $(g_1(X_1), \ldots, g_N(X_N))'$ by $F_g(\cdot)$, we have

$$\begin{aligned} C\left(u_1, \ldots, u_N\right) &= F\left(F_1^{-1}(u_1), \ldots, F_N^{-1}(u_N)\right) \\ &= \mathbb{P}\left(X_1 \leq F_1^{-1}(u_1), \ldots, X_N \leq F_N^{-1}(u_N)\right) \\ &= \mathbb{P}\left(g_1(X_1) \leq g_1\left(F_1^{-1}(u_1)\right), \ldots, g_N(X_N) \leq g_N\left(F_N^{-1}(u_N)\right)\right) \\ &= F_g\left(g_1\left(F_1^{-1}(u_1)\right), \ldots, g_N\left(F_N^{-1}(u_N)\right)\right). \end{aligned}$$

The result of this property is that we may strictly consider correlation structures when modeling ellipsoidal joint distributions via a copula rather than covariance structures; consider

$$g_i : X_i \mapsto \frac{X_i - \mu_i}{\sigma_i}$$

for finite mean and standard deviation, μ_i and σ_i.

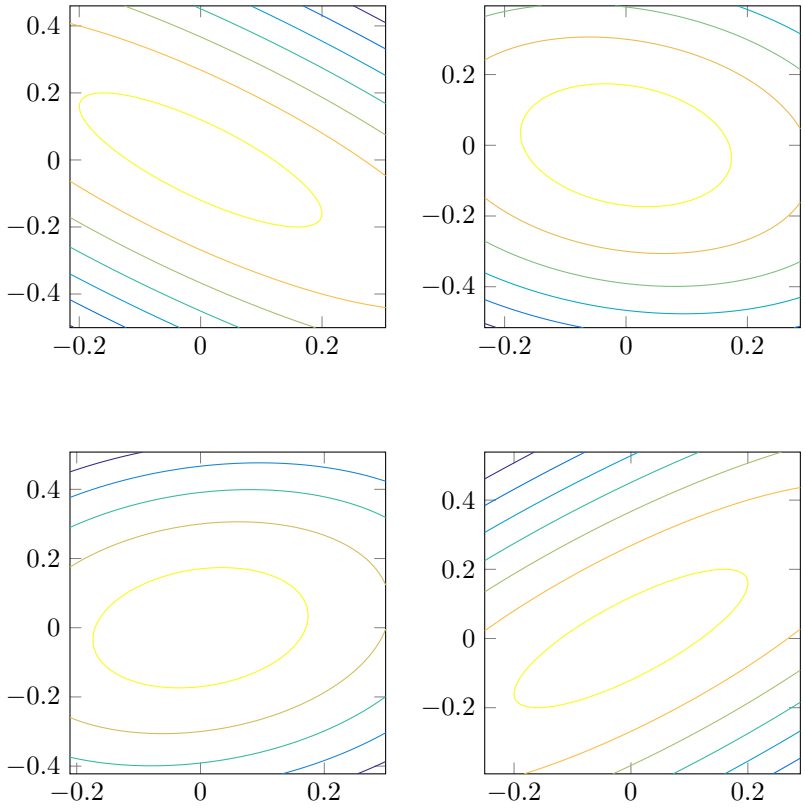

Figure 3.4: Simulated joint distributions using a copula approach, assuming a multivariate normal distribution with correlation parameter ranging from -0.8, -0.2, 0.2, and 0.8, clockwise from upper left subplot. In each case, the marginal distribution is fixed.

Example 3.2.2. Revisiting the example of random variables, jointly Student t, with standard normal marginals, we may use the preceding result to write

$$\begin{aligned} C(u_1, u_2) &= St_{\mu,\Sigma;\nu}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\right) \\ &= St_{0,R;\nu}\left(\frac{\Phi^{-1}(u_1) - \mu_1}{\sigma_1}, \frac{\Phi^{-1}(u_2) - \mu_2}{\sigma_2}\right) \end{aligned}$$

for R the correlation matrix obtained from Σ.

Example 3.2.3. In practice, $F(\cdot)$ and marginals, $\{F_i(\cdot)\}$ are determined *a priori*; i.e., in a manner fixing a model. Oftentimes, copulas are used to simulate data with these prescribed distributions. Here, we look at a simple case of

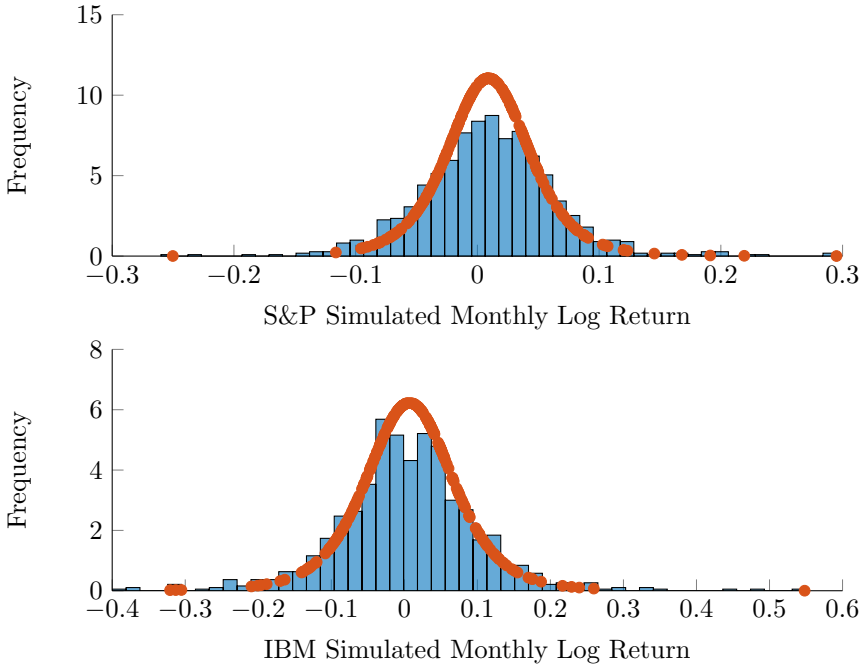

Figure 3.5: Simulated marginal distributions fit to a Student t distribution with $\nu = 5$ degrees of freedom. The marginals shown here were input into various copula functions in the preceding figure.

simulating jointly normal data with Student t marginals. Specifically, let

$$\begin{aligned} X_1 &\sim St(\mu_1, \sigma_1^2; 5) \\ X_2 &\sim St(\mu_2, \sigma_2^2; 5), \end{aligned}$$

and let $F(X) = \Phi_{\mu,\Sigma}$. We know from the above that, without loss of generality, we may construct our copula function $C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2))$; or, if we would like, directly from the correlation matrix.

Since the joint distribution is assumed to be normal, we may first simulate $N \times 2$ independent samples from a standard normal random variable to obtain a matrix $\hat{Z}_0 \in \mathbb{R}^{N\times 2}$. We know from previous work that

$$\hat{Z} = \Lambda \hat{Z}_0$$

will then be sampled according to $N(0, \Sigma)$ when $\Sigma = \Lambda\Lambda'$. Notice that each column of $\hat{Z}$ is sampled according to a standard normal. Letting $\hat{Z}_i$ be the ith column of $\hat{Z}$, we set

$$\hat{U}_i = \Phi(\hat{Z}_i)$$

with evaluation occurring componentwise. From the above, each $\hat{U}_i$ is now sampled according to a uniform random variable on $[0, 1]$ with the joint distribution

specified by $F(\cdot)$. Taking inverses of the sampled uniform distributions, then,

$$\begin{aligned}\hat{X}_1 &= St^{-1}_{\mu_1,\sigma_1^2;5}\left(\hat{U}_1\right)\\ \hat{X}_2 &= St^{-1}_{\mu_2,\sigma_2^2;5}\left(\hat{U}_2\right),\end{aligned}$$

we obtain our desired result: each of X_1 and X_2 has a Student t distribution with specified mean and variance, and the joint distribution is multivariate normal with correlation specified by Σ.

Figure 3.2 shows the isocontours of various joint distributions, with correlation varying over -0.8, -0.2, 0.2, and 0.8. In each case, the marginals are sampled from the same distribution. Representative histograms are shown in Figure 3.2.3.

Copulas are capable of modeling a considerable amount of information more than simple correlation Even so, the applications that dominate the field involve using copula models to imbue marginal distributions with a given correlation structure. As our primary focus is on ellipsoidal distributions, their introduction within the chapter on covariance is not an accident and is in line with the view of most market practitioners.

The copula approach found wide appeal in credit derivatives markets due to a paper published by David Li in the Journal of Fixed Income [21]. *On Default Correlation: A Copula Function Approach* modeled default correlation in a novel way, linking marginal default risks obtained from credit default swap (CDS) pricing through a copula with a very simple structure to imply a joint distribution of credit defaults. The copula that became widely used and whose parameter eventually became a quoted market price was a multivariate normal copula with a covariance (correlation) matrix given by

$$\begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & & \ddots & \\ \rho & \dots & \dots & 1 \end{pmatrix}.$$

Much like the versions of the Capital Asset Pricing Model we have seen already, the above model does two things: it provides a simplification of market relationships via market pricing and normative relationships, and produces an interpretable parameter.

Li's formula, however, is far more dangerous than Merton or Sharpe's. The model above, here told in generalities, but a rigorous treatment is not too much more involved, was used to estimate probabilities of joint defaults within pools of hundreds or even thousands of loans. The constant pairwise correlation is concerning, but the use of the normal distribution is even more so. Our previous analysis of the inability of the normal distribution to capture market extremes is apropos here as well.

The ease of implementation lead to this copula-based model here being used to mint huge numbers of triple-A rated bonds (made up from tiered levels of

pools of bonds). The pooled bonds were known as collateralized debt obligations, or CDOs. Concurrent with the acceptance of the modeling above, the CDO market grew from $275 billion in 2000 to $4.7 trillion in 2006.

Not only are correlations unstable (as we have already seen) and extreme events terribly likely, the CDO market (because of such great ratings by the ratings agencies) saw massive leverage. This was a recipe for disaster and culminated in the financial crisis of 2008. There were many people who could see this train-wreck coming far before it occurred, but in large part, the market did not. In effect, the market wasn't efficient at pricing pairwise correlations; or, even worse, systemic crashes.

Li has apparently been unavailable for comment since the crisis.

Even with the above cautionary tale, we maintain that the powerful capability of modeling joint and marginal distributions separately is incomparable. Again, our emphasis is on interpretation and empiricism rather than on normative modeling.

As such, we may propose uses for the copula approach along the lines of Meucci [24], suggesting to use a joint Student t copula with five degrees of freedom and flexible marginals, or various panic copulas which reflect the stylized features of asymmetry and increased correlation in a crisis we have noted previously. Such additions are not merely refinements to the approach presented in Li's paper, but well thought out and natural implementations of a flexible model in a market with a few known attributes. The copula modeling approach – much like the percentile modeling in VaR and CVaR interpretations of risk – allows the practitioner to articulate a wide variety of views. As such, its maligned history will likely be short.

3.3 Eigenvalues

We may further analyze the structure of the covariance matrix by studying its *eigenvalues and eigenvectors*. Recall that for a square matrix, $A \in \mathbb{R}^{N\times N}$, the scalar λ is an eigenvalue if

$$Av = \lambda v. \tag{3.20}$$

In this case, we say that the nonzero vector, v, is the eigenvector associated with λ. Notice that if v is an eigenvalue, then a scalar multiple, cv, satisfies

$$A(cv) = cAv = c\lambda v = \lambda(cv),$$

and hence we may assume without loss of generality that $||v|| = 1$.

Eigenvalues may be determined by considering that if

$$Av = \lambda v,$$

then

$$(A - \lambda I)v = 0$$

where I is the identity matrix. This implies, then, that $A - \lambda I$ is singular (i.e., it affords a nonzero solution to the above), and hence its determinant must be zero:

$$\det(A - \lambda I) = 0. \tag{3.21}$$

Equation (3.21) is known as the *characteristic equation*, and is a polynomial of degree N. We know that in $\mathbb{C}$ such an equation will surely have all of its roots. We are interested in the case where all of the eigenvalues are real, and positive definiteness (even semidefiniteness) is a sufficient condition for just such a result. We will denote that A is positive definite by

$$A \succ 0$$

and positive semidefinite by

$$A \succeq 0.$$

Theorem 3.3.1. The eigenvalues of a positive semidefinite real matrix, $A \succeq 0$, are real and nonnegative. If $A \succ 0$, then the eigenvalues are strictly positive.

Proof. Let $A \succeq 0$, and let v be a (nonzero) eigenvector with associated eigenvalue λ. We have immediately that

$$Av = \lambda v,$$

and, premultiplying both sides by v', we get

$$v'Av = \lambda v'v = \lambda.$$

Now, since $A \succeq 0$, the left hand side is nonnegative, and hence so is λ. The case where $A \succ 0$ follows identically, resulting in strictly positive λ.

We have already seen that the covariance matrix, Σ, of the real valued multivariate random variable X is positive semidefinite; viz., for $Y = w'X$,

$$Var(Y) = w'\Sigma w.$$

A necessary and sufficient condition for Σ, then, is to ensure that the variance of Y is nonzero; i.e., that it is not the case that there exists a linear combination

$$\sum_{i=1}^{N} w_i X_i$$

that is zero in the quotient space of random variables identifying constants. In this case, we say that $\{X_i\}_{i=1}^N$ is linearly independent. We have then that $\Sigma = Cov(X)$ is positive definite if and only if $\{X_i\}_{i=1}^N$ is linearly independent.

Notice that in our previous work on the multivariate normal and Student t distributions, an implicit assumption was that Σ was invertible. We now may formulate this condition based on the linear independence of the components of X.

In a financial context, we may interpret an eigenvector, e_i, as a vector of portfolio weights or positions. Consider,

$$\begin{aligned} Var(e_i'X) &= e_i'\Sigma e_i \\ &= \lambda_i e_i' e_i \\ &= \lambda_i. \end{aligned}$$

So that λ_i is exactly the variance of the portfolio with positions e_i. We call

$$e_i'X \tag{3.22}$$

the ith principal component of X, and will often refer to both e_i and $e_i'X$ as an *eigenportfolio*. The relative size of each eigenvalue is surprisingly meaningful and related to this observation and aids in problems of dimension reduction.

For a covariance matrix, $\Sigma \in \mathbb{R}^{N\times N}$ with eigenvalues $\{\lambda_i\}_{i=1}^N$ and associated eigenvectors $\{e_i\}_{i=1}^N$ with

$$\lambda_1 \geq \cdots \geq \lambda_N \geq 0$$

we have that the eigenvectors of distinct eigenvalues are orthogonal; viz.,

$$e_i' e_j = 0$$

if $\lambda_i \neq \lambda_j$. To prove this, we note that

$$\Sigma e_i = \lambda_i e_i$$

and

$$\Sigma e_j = \lambda_j e_j$$

and so

$$\begin{aligned} e_i'\Sigma e_j &= \lambda_j e_i' e_j \\ e_j'\Sigma e_i &= \lambda_i e_j' e_i. \end{aligned}$$

Now $e_i'\Sigma e_j = e_j'\Sigma e_i$ since Σ is symmetric. Therefore

$$\lambda_j e_i' e_j = \lambda_i e_j' e_i.$$

Similarly, $e_i' e_j = e_j' e_i$, giving

$$(\lambda_i - \lambda_j) e_i' e_j = 0.$$

Assuming that the eigenvalues of Σ are distinct, we may decompose the covariance matrix as

$$\Sigma = \sum_{i=1}^N \lambda_i e_i e_i'. \tag{3.23}$$

To see this, notice that

$$\begin{pmatrix} — & e_1 & — \\ & \vdots & \\ — & e_N & — \end{pmatrix} \begin{pmatrix} | & & | \\ e_1 & \cdots & e_N \\ | & & | \end{pmatrix} = \begin{pmatrix} e_1'e_1 & \cdots & e_1'e_N \\ \vdots & \vdots & \vdots \\ e_N'e_1 & & e_N'e_N \end{pmatrix} = I$$

by the orthogonality of eigenvectors of Σ. Now, for square matrices, A and B satisfying $AB = I$, we know (read: we leave to the reader) that $BA = I$. Hence

$$\begin{pmatrix} | & & | \\ e_1 & \cdots & e_N \\ | & & | \end{pmatrix} \begin{pmatrix} — & e_1 & — \\ & \vdots & \\ — & e_N & — \end{pmatrix} = I,$$

or

$$e_1e_1' + \cdots + e_Ne_N' = I.$$

Finally, since

$$\Sigma e_ie_i' = \lambda_i e_ie_i',$$

we have that

$$\begin{aligned} \Sigma &= \Sigma I \\ &= \Sigma \sum_i e_ie_i' \\ &= \sum_i \Sigma e_ie_i' \\ &= \sum_i \lambda_i e_ie_i' \end{aligned}$$

as desired.

As a result, utilizing the fact that the trace operator for a matrix is linear and the property that for square matrices A and B,

$$tr(AB) = tr(BA),$$

we may relate the sum of variances of X to the sum of the eigenvalues of Σ. In particular, we have

$$\begin{aligned} tr(\Sigma) &= tr\left(\sum_i \lambda_i e_ie_i'\right) \\ &= \sum_i \lambda_i tr\left(e_ie_i'\right) \\ &= \sum_i \lambda_i tr\left(e_i'e_i\right) \\ &= \sum_i \lambda_i, \end{aligned}$$

or

$$\sum_i \sigma_i^2 = \sum_i \lambda_i. \tag{3.24}$$

We call $\sum_i \lambda_i$ the *total variance* of Σ. In addition to relating the eigenvalues of Σ to the sum of variances, equation (3.24) also gives us a method for dimension reduction.

Example 3.3.1. Let X be an N-dimensional random vector representing the returns of N assets. For a threshold, τ, with

$$0 < \tau \leq 1$$

we may choose M eigenportfolios explaining $\tau\%$ of the total variance by choosing the smallest M satisfying

$$\frac{\sum_{i=1}^{M} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq \tau. \tag{3.25}$$

The related M eigenportfolios (or principal components), $e_i' X$, then comprise $\tau\%$ of the total variance. This is especially effective for high dimensional X such as when considering the composition of the S&P 500, for example.

In Figure 3.3, an estimated covariance matrix was calculated monthly for a cross-section of the 50 largest stocks at the time by market cap. The covariance was calculated using 121 trailing weeks of returns. The largest N eigenvalues were chosen according to (3.25), with $\tau = 80\%$. A smoothed approximation, looking at the mean N_t for the trailing 18 months is shown as well.

Throughout, no more than 18 eigenportfolios were needed to explain more than 80% of the total variance. This is a significant decrease from the original dimension of 50. Additionally, the figure shows that there was a steeped decrease in the number of eigenportfolios 'explaining the market' from the financial crisis through 2015 – only rebounding from the lows sometime in 2011. This effect is likely related to the various Quantitative Easing programs initiated by the Fed at the time.

We may also ask how much of the total variance is explained by the eigenportfolio related to the largest eigenvalue. Figure 3.3 shows the time variation of the explanatory power of this eigenportfolio. We see again the significant upswing after the Financial Crisis, achieving levels of market coordination not seen in the preceding twenty years. While the adage that in a crisis correlations go to one is evidenced here (read the explanatory power of the largest eigenvalue increases significantly), the figure is also instructive, showing that the market is dynamic and that heretofore unseen influences like the Fed's Qualitative Easing [12] can have significant and novel effects. However, we should note that these effects are interpretable – and perhaps even expected from the trained practitioner's eye – using the mathematical edifice already in place and established here.

Carrying this type of reasoning further, one may posit that if a necessary (but clearly not sufficient) condition for a bear market downturn is an uptick in

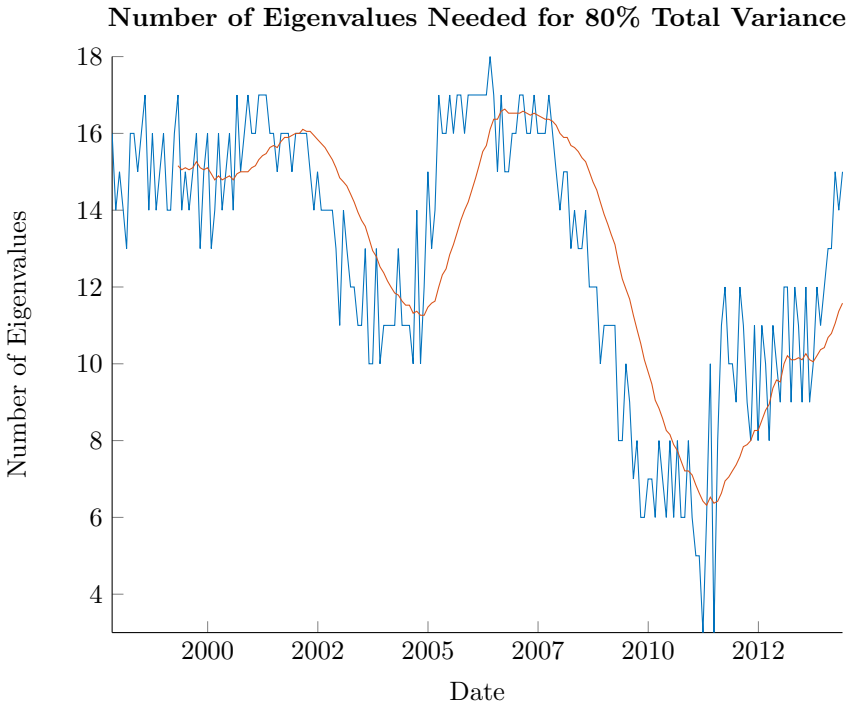

Figure 3.6: Plot of the number of eigenportfolios needed to account for 80% of total variance through time. Covariance is calculated using weekly returns over a trailing 121 week period with the largest 50 stocks by market cap evaluated monthly.

the explanatory power of the largest eigenvalue, a crisis post 2015 would require a break in the elevated levels seen since 2008. There is minor evidence of this in the figure as well.

Finally, we may look at the distribution of eigenvalues in a manner similar to our previous analysis of daily log returns for various stocks. Figure 3.3 shows the empirical density of the eigenvalues of the covariance matrix as before available on 12/31/2007. As with our discussion of the distribution of daily log returns, certain stylized features emerge.

Particularly, even with observations that are linearly independent, we see a peak of near-zero eigenvalues. In recent years, the tools of Random Matrix Theory (RMT) have been implemented in math finance to study this phenomenon. Authors like Bouchaud and Potters [4] present a methodology based on RMT to identify random, and hence noisy, eigenportfolios. Doing so seeks to modify the covariance (correlation) matrix to eliminate eigenportfolios with erroneously low contributions to risk. This effect is particularly important when considering mean-variance optimization.

In addition to a large bulk of eigenvalues clustering around zero, we also

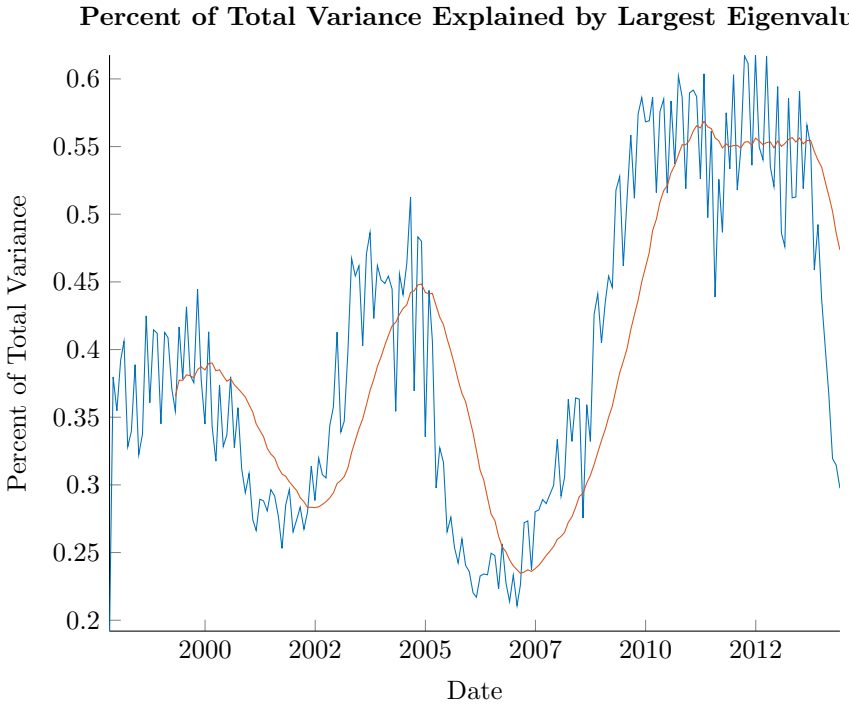

Figure 3.7: Plot of the percentage of total variance explained by the eigenportfolios associated with the largest eigenvalue through time. Covariance is calculated using weekly returns over a trailing 121 week period with the largest 50 stocks by market cap evaluated monthly.

note one very large eigenvalue – in this case, 2,745 times larger than the smallest eigenvalue, and 3.80 times larger than the second largest eigenvalue. This eigenvalue is documented as related to the market portfolio as in Avellaneda [1], but we cannot replicate these claims. However, we do note that the eigenportfolio for this eigenvalue very often has all positive entries (or, more accurately since eigenvectors are scalar independent, all entries share the same sign).

The covariance structure of equity returns, then, allows for a broad classification wherein the market is often driven by a dominant market portfolio, orders of magnitude larger than the smallest eigenvalue. This smallest eigenvalue, in turn, lies amongst a bulk of very near zero eigenvalues that may be classified in a technical sense as random noise. Further, the effects of the largest eigenvalue are interpretable in a dynamic sense, accounting for time specific features of the market as such.

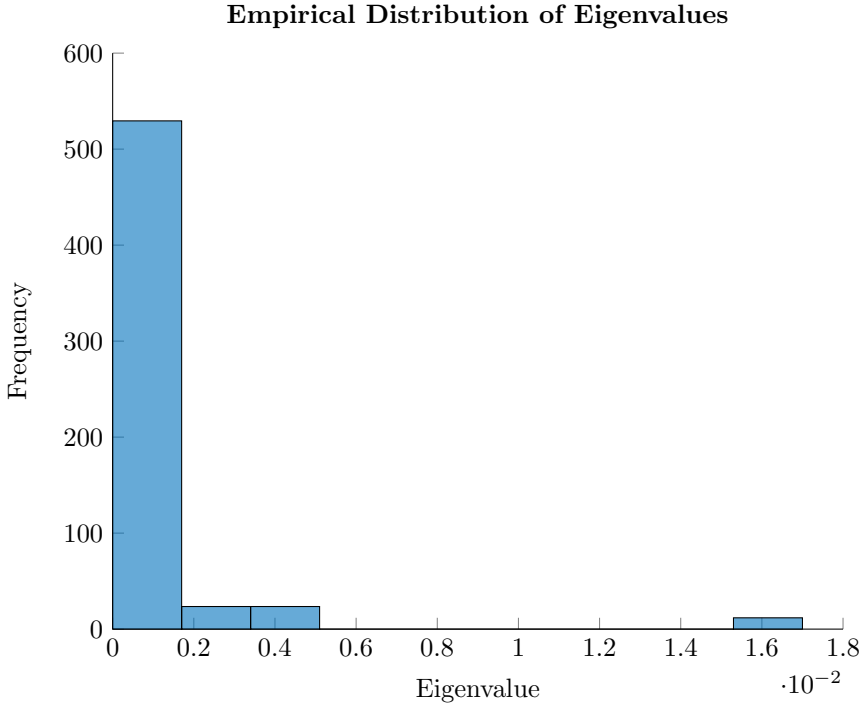

Figure 3.8: Empirical histogram of the eigenvalues of the covariance matrix available on 12/31/2007. Covariance is calculated using weekly returns over a trailing 121 week period with the largest 50 stocks by market cap on that date.

Exercises

1. Prove that for an inner product $(\cdot\,,\cdot)$, and for scalars$\{a_i\}$ and $\{b_j\}$, and $\{X_i\}$ and $\{Y_j\}$ in V, we have

$$\left(\sum_i a_i X_i, \sum_j b_j Y_j\right) = \sum_i a_i \sum_j b_j (X_i, Y_j).$$

2. Verify that $Cov(\cdot\,,\cdot)$ is symmetric and bilinear.

3. Prove that $||\cdot||^2 = (\cdot\,,\cdot)$ defines a norm. In particular, this gives that variance is a norm on the quotient space of random variables.

4. Prove for univariate random variables X and Y with means μ_X and μ_Y, respectively, that

$$Cov(X,Y) = \mathbb{E}\,(XY) - \mu_X \mu_Y.$$

5. Using notation as in (2.43), show that

$$\hat{\sigma}_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)$$

is an unbiased estimator of the covariance between X and Y.

6. For univariate random variables X and Y, and scalars, a, b, c, and d, if $Corr(X, Y) = \rho$, what is $Corr(a + bX, c + dY)$?

7. Verify that $f_U(x) = 1$ is the probability density function for uniformly distributed U on $[0, 1]$.

8. Use the change of variables theorem to determine the density for a uniformly distributed random variable on $[a, b]$.

9. Prove 3.18 for continuous F_i by looking at

$$\mathbb{P}(F_i(X_i) \leq u)$$

for $u \in [0, 1]$.

10. Using the IBM/S&P daily return data, construct a copula function simulation with $N = 1,000$ samples for the joint distribution of daily log returns where the joint distribution is Student t with five degrees of freedom with correlation matching the sample, and the marginals are also Student t with five degrees of freedom with means and volatilities matching the univariate sample means and standard deviations.

11. Using the cross-sectional and historical return data, plot the percentage of positive components of the eigenvector associated with the largest eigenvalue of the covariance through time; i.e., for eigenvector $e_t = (e_{1t}, \ldots, e_{Nt})$ at time t, plot

$$m_{it} = \frac{1}{N} \sum_{i=1}^{N} \delta_i(e_t)$$

where

$$\delta_i(e) = \begin{cases} 1 & \text{if } e_i > 0 \\ 0 & \text{otherwise.} \end{cases} .$$

Use the same methodology as in the chapter, choosing the largest 50 companies by market cap at each time, and using the full 121 weeks of returns available. Why can we assume without loss of generality that $m_{it} > 0.5$?

12. Prove that for square matrices A and B that if

$$AB = I$$

then

$$BA = I.$$

13. Consider the model

$$r = \alpha + \beta r_m + \epsilon$$

with r, r_m, and ϵ univariate random variables,

$$(r_m, \epsilon) = Cov(r_m, \epsilon) = 0,$$

and

$$\mathbb{E}(\epsilon) = 0.$$

Show that

$$\beta = \frac{Cov(r, r_m)}{Var(r_m)}$$

and

$$\alpha = \mathbb{E}(r) - \beta \mathbb{E}(r_m).$$

14. Prove that if the matrix A has eigenvalues $\{\lambda_i\}_{i=1}^N$, then A^k has eigenvalues $\{\lambda_i^k\}_{i=1}^N$.

15. Let $(\cdot, \cdot)_F$ be defined for matrices (of appropriate dimensions) by $(A, B)_F = tr(A'B)$.

    (a) Show that $(\cdot, \cdot)_F$ is an inner product.

    (b) For a covariance matrix, Σ, what is $\sqrt{(\Sigma, \Sigma)_F}$ in terms of the eigenvalues of Σ?

16. For the Euclidean vector norm $|| \cdot ||$,

$$||x|| = (x'x)^{\frac{1}{2}},$$

define the matrix norm $|| \cdot ||_2$ by

$$||A||_2 = \max_{||v||=1} ||Av||. \tag{3.26}$$

    (a) Show that

$$||A||_2 = \max_{||v||} \frac{||Av||}{||v||}.$$

    (b) Show that for any vector z,

$$||Az|| \leq ||A||_2 ||z||.$$

    (c) Show

$$||AB||_2 \leq ||A||_2 ||B||_2$$

    (d) Show that for a positive definite matrix A,

$$||A||_2 = \lambda_{max},$$

    where λ_{max} is the maximum eigenvalue of A.

Chapter 4

Ordinary Least Squares

Our work in Ordinary Least Squares (OLS) begins with the Capital Asset Pricing Model (CAPM). Fitting the parameters of the model to historical data is reduced to a calculus problem and an associated assumption about so-called idiosyncratic error terms. After establishing the methodology for calculating the CAPM parameters, we look at the implications of the model, giving testable hypotheses which we examine with market data. We find, as in Fama-French [10], for example, that these hypotheses are not borne out in the market, however.

The approach used to fit a CAPM model is justified through the establishment of the method of maximum likelihood, which, more generally justifies our approach in OLS. Our treatment of the subject is principally mathematical, relying in large part on the properties of projection operators.

We discuss and derive from three main assumptions (here identified as the Gauss-Markov assumptions and a sometimes-accompanying distributional assumption) the distributional properties of estimators of regression parameters. These distributions allow us to determine confidence intervals for our estimators as well as establish so-called null hypothesis tests [18, 6]. Our primary test statistics will be the t-test and F-test, which we build from first principles. The latter test statistic lends itself to model selection criteria, and we build upon this in our presentation of forward selection, backward elimination, and stepwise regression techniques.

We conclude with an introductory treatment on a generalization of ordinary least squares – aptly named Generalized Least Squares (GLS).

4.1 CAPM and Least Squares

The Capital Asset Pricing Model (CAPM) [22, 31] relates the time series of a given stock's returns, $\{r_t\}$, to the returns of the contemporaneous market, $\{m_t\}$, as

$$r_t - r_f = \beta(m_t - r_f) + \epsilon_t, \tag{4.1}$$

where r_f is the *risk-free rate*, the rate at which one may borrow with a probability of default on the bond being zero. Treasury bills are the obvious and most likely candidate here, the dynamics of the US debt notwithstanding.

We will not at present make any assumptions about the distribution of $\{\epsilon_t\}$. However, we will assume that the ϵ_t's are iid, and share the same distribution as some ϵ. Similarly, we will assume that the market returns, m_t are iid and distributed as some m. Finally, we assume as before that

$$\begin{aligned} Cov(m, \epsilon) &= 0 \\ Var(\epsilon) &= \sigma_\epsilon^2 \\ Var(m) &= \sigma_m^2. \end{aligned}$$

We consider the case where we have observations $\{r_t\}_{t=1}^T$ and $\{m_t\}_{t=1}^T$, drawn from r and m, respectively, and seek an estimate, $\hat{\beta}$ of β. We have already seen in (3.11) that

$$\beta = \frac{\rho\sigma_r}{\sigma_m} = \rho\frac{\sigma_r}{\sigma_m},$$

and we made mention in (3.14) that a method of moments estimator for β is

$$\hat{\beta} = \hat{\rho}\frac{\hat{\sigma}_r}{\hat{\sigma}_m}.$$

Here we show that this estimator coincides with what is called the *least squares* estimate for reasons that will become clear momentarily.

Let f be a function of the as-yet unknown β given by

$$f(\beta) = \sum_{t=1}^{N} \left(r_t - r_f - \beta(m_t - r_f)\right)^2. \tag{4.2}$$

Mathematically, finding the value which minimizes $f(\cdot)$ is an attractive and perhaps natural proposition. We know from calculus that $f(\cdot)$ is parabolic and hence has a single minimizer, which may be found where the derivative $f'(\cdot)$ is zero. We will now show that this minimizer is the method of moments estimator of β, why this link to least squares minimization exists, and that this minimizer is an unbiased estimator of β.

4.1.1 Minimizing $f(\cdot)$

The minimizer of $f(\cdot)$ above is exactly the method of moments estimator of β. The result follows from calculus and the assumption that the sample covariance between ϵ and m,

$$\hat{\sigma}_{\epsilon m} = \frac{1}{T-1}\sum_{t=1}^{T}(m_t - \bar{m})(\epsilon_t - \bar{\epsilon})$$

satisfies $\hat{\sigma}_{\epsilon,m} = 0$, with $\bar{m}$ and $\bar{\epsilon}$ the estimators for the mean we have seen previously. For notation, we denote

$$\begin{aligned} x_t &= m_t - r_f \\ y_t &= r_t - r_f. \end{aligned}$$

Notice that

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t = \bar{m} - r_f$$
$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t = \bar{r} - r_f,$$

and since variance and covariance are both shift independent,

$$\begin{aligned}
\hat{\sigma}_x^2 &= \hat{\sigma}_m^2 \\
\hat{\sigma}_y^2 &= \hat{\sigma}_r^2 \\
\hat{\sigma}_{xy}^2 &= \hat{\sigma}_{mr}^2 \\
\hat{\sigma}_{x\epsilon}^2 &= \hat{\sigma}_{m\epsilon}^2,
\end{aligned}$$

where the estimators of variance are as in (2.43).

In these new variables, $f(\cdot)$ becomes

$$f(\beta) = \sum_{t=1}^{T} (y_t - \beta x_t)^2 .$$

The minimum of $f(\cdot)$ occurs when the derivative is set to zero since the function is quadratic in β, hence we solve

$$\begin{aligned}
f'(\beta) &= 2\sum_{t=1}^{T} (y_t - \beta x_t)\cdot x_t = 0 \\
&= \sum_{t=1}^{T} \epsilon_t \cdot x_t = 0. \qquad (4.3)
\end{aligned}$$

Notice that for any sequences of random variables $\{u_t\}$ and $\{v_t\}$,

$$\sum_{t=1}^{T} u_t v_t = (T-1)\hat{\sigma}_{uv} + T\bar{u}\bar{v},$$

and

$$\sum_{t=1}^{T} u_t^2 = (T-1)\hat{\sigma}_u^2 + T\bar{u}^2.$$

These relationships are left to the reader.

Continuing, we have

$$\begin{aligned}
\sum_{t=1}^{T} (y_t - \beta x_t)\cdot x_t &= \sum_{t=1}^{T} y_t x_t - \beta \sum_{t=1}^{T} x_t^2 \\
&= (T-1)\hat{\sigma}_{xy} + T\bar{x}\bar{y} - \beta\left((T-1)\hat{\sigma}_x^2 + T\bar{x}^2\right) \\
&= (T-1)\left(\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2\right) + T\left(\bar{x}\bar{y} - \beta\bar{x}^2\right).
\end{aligned}$$

Our assumption that $\hat{\sigma}_{m\epsilon} = 0$, or equivalently, that $\hat{\sigma}_{x\epsilon} = 0$ gives

$$\sum_{t=1}^{T} x_t \epsilon_t - T\bar{x}\bar{\epsilon} = 0.$$

But since we know from (4.3) that $\sum_{t=1}^{T} x_t \epsilon_t = 0$, this gives that $\bar{x}\bar{\epsilon} = 0$ as well. Returning to our formulation above, we see then that

$$\begin{aligned}(T-1)\left(\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2\right) + T\left(\bar{x}\bar{y} - \beta\bar{x}^2\right) &= (T-1)\left(\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2\right) + T\bar{x}\left(\bar{y} - \beta\bar{x}\right) \\ &= (T-1)\left(\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2\right) + T\bar{x}\bar{\epsilon} \\ &= (T-1)\left(\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2\right).\end{aligned}$$

Solving $f'(\beta) = 0$ therefore amounts to solving

$$\hat{\sigma}_{xy} - \beta\hat{\sigma}_x^2 = 0,$$

which yields

$$\beta = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}$$

as expected.

It is critical to point out that we used an assumption that the error terms $\{\epsilon_t\}$ had zero correlation with the market returns $\{m_t\}$. Without this assumption, the estimate of β from this least squares approach keeps the terms eliminated in the final parts of our derivation; namely, we arrive at a solution

$$\beta = \frac{(T-1)\hat{\sigma}_{xy} + T\bar{x}\bar{y}}{(T-1)\hat{\sigma}_x^2 + T\bar{x}^2}.$$

Implicitly (and maybe naturally), we minimized the function $f(\cdot)$ with the *constraint* of idiosyncrasy of the error terms.

If we consider an alternative model,

$$r_t - r_f = \alpha + \beta(m_t - r_f) + \epsilon_t, \tag{4.4}$$

and look again at the least squares function, now in two variables,

$$f(\alpha, \beta) = \sum_{t=1}^{N} \left(r_t - r_f - (\alpha + \beta(m_t - r_f))\right)^2, \tag{4.5}$$

we may, by solving for where the gradient, ∇f, is zero,

$$\nabla f(\alpha, \beta) = \begin{pmatrix} \frac{\partial f}{\partial \alpha}(\alpha, \beta) \\ \frac{\partial f}{\partial \beta}(\alpha, \beta) \end{pmatrix} = 0.$$

we obtain a system of equations in α and β. For the former, we see –using the same notation as before– that

$$\frac{\partial f}{\partial \alpha}(\alpha, \beta) = -2\sum_{t=1}^{N} \left(r_t - r_f - (\alpha + \beta(m_t - r_f))\right).$$

Setting this equal to zero gives

$$\alpha = \bar{r}_t - r_f - \beta(\bar{m}_t - r_f).$$

We leave it to the reader to show that

$$\beta = \frac{\hat{\sigma}_{rm}}{\hat{\sigma}_m^2}.$$

We denote these estimates by $\hat{\alpha}$ and $\hat{\beta}$, respectively.

The intercept term in (4.4) has other implications. In particular for the least squares estimators $(\hat{\alpha}, \hat{\beta})$, the estimators of $\{\epsilon_t\}$, $\{\hat{\epsilon}_t\}$, satisfy (without explicit assumption)

$$\begin{aligned} \bar{\hat{\epsilon}} &= 0 \\ \hat{\sigma}_{m\epsilon} &= 0. \end{aligned}$$

Showing the first equation is straightforward:

$$\begin{aligned} \bar{\epsilon} &= \frac{1}{T}\sum_{t=1}^{T}\Big(r_t - r_f - \hat{\alpha} - \hat{\beta}(m_t - r_f)\Big) \\ &= \frac{1}{T}\sum_{t=1}^{T}\Big(r_t - r_f - \hat{\beta}(m_t - r_f)\Big) - \hat{\alpha} \\ &= \bar{r}_t - r_f - \hat{\beta}(\bar{m}_t - r_f) - \hat{\alpha} \end{aligned}$$

which is zero by the construction of $\hat{\alpha}$. The second equation follows from this and the partial derivative with respect to β of (4.5) being zero.

The Capital Asset Pricing Model [22, 31] is a normative model: the model in (4.1) omits an intercept term and defines the relationship of the idiosyncratic terms $\{\epsilon_t\}$ to the market through assumptions. This is not a model derived from observations and relationships in the market, but based on assumptions of how a market ought to perform. We prefer a somewhat nontraditional definition, relating a given security's adjusted returns to the market as in (4.4). The allowance for nonzero α has serious implications, though, which we address momentarily. The other major defining features of the model are obtained rather than assumed in this formulation, however. This intersection of parsimony and flexibility is hard to pass up.

We can formulate the relationship between random variables implied by (4.4), as

$$r - r_f = \alpha + \beta(m - r_f) + \epsilon.$$

Maintaining assumptions of idiosyncrasy of ϵ, which, follow from the least squares estimators, but are not properties of the model as such, and taking expectations on both sides, we get

$$\mathbb{E}(r - r_f) = \alpha + \beta(\mathbb{E}(m) - r_f),$$

or

$$\mathbb{E}(r) = r_f + \alpha + \beta(\mathbb{E}(m) - r_f).$$

This last line is intimately involved with the Efficient Market Hypothesis [28, 2]. In the case that we have a fully explanatory model, a nonzero α should be precluded in an efficient market. Any alternative allows for securities which give opportunities for risk free excess returns. Consider the case where $\alpha > 0$. A portfolio, Π, with returns r_Π, long one unit of the security with returns r and short β units of the market will have an expected return

$$\mathbb{E}(r_\Pi) = r_f + \alpha.$$

That is, the expected return of the portfolio will *exceed the risk free rate while having no linear relationship to the market.* Here we have used the fact that the β of a portfolio is the weighted sum of the βs of its positions. We forgo requiring a proof of this until we have established more optimization techniques as the current applications find a direct and amenable analogue there.

We say that such a portfolio provides an opportunity for *statistical arbitrage*: positive returns are expected in excess of those explained by selected factors. In our current model, the only factor (usually called a *risk factor*) is the market proper. A particular type of statistical arbitrage often claimed by academics and practitioners alike involves cross-sectional comparisons of fundamental or technical data; e.g., comparing ratios like earnings before interest and taxes over enterprise value at a given time for a set of stocks under consideration (we will have a more formal treatment of these terms in a following chapter). We will call these types of statistical arbitrage opportunities *anomalies*.

Empirically, the omission of α in (4.1) is a mistake. Equity anomalies abound. As such, we use (4.4) for the identification of anomalies rather than assume away their existence. Notice, however, that statistical arbitrage and anomalies are defined in relation to some model. The specification of that model, then, is critical in the assessment of perceived opportunities and statistical descriptions of any security or portfolio of securities. A consequence of the above is that one measure of a powerful descriptive model in the equity space is its ability to statistically identify an α of zero for many of the widely accepted and observed anomalies in the literature (as determined by models such as CAPM or its extensions).

Practitioners will often refer to the α of a given security, portfolio, strategy, or idea. These references necessarily presuppose some underlying model description – oftentimes the assumed model is CAPM. Similarly, β is also used as a descriptive characteristic for securities and portfolios alike. In practice, references to β as such are made to the sensitivity to the market; i.e., CAPM-β.

4.1.2 Interpretations and Implications

The Capital Asset Pricing Model (without intercept but with idiosyncratic assumptions) implies several relationships between returns and volatility, idiosyn-

cratic volatility, correlation to the market, and β. The model

$$r - r_f = \beta(m - r_f) + \epsilon$$

gives under the CAPM assumptions that

$$\begin{aligned}
\mathbb{E}(r) &= r_f + \beta(\mathbb{E}(m) - r_f) \\
&= r_f + \rho\frac{\sigma_r}{\sigma_m}(\mathbb{E}(m) - r_f) \\
&= r_f + \rho\sigma_r\frac{\mathbb{E}(m) - r_f}{\sigma_m},
\end{aligned}$$

where $\rho = Corr(r, m)$. The term $\frac{\mathbb{E}(m)-r_f}{\sigma_m}$ will make another appearance when we look at the so-called Sharpe Ratio. For now, however, we note that if the expected return of the market exceeds the risk free rate[1], $\mathbb{E}(r)$, is an increasing function of β, σ_r, and correlation to the market. Where in studying the latter two cases empirically, care must be taken to control for correlation and σ_r, respectively.

We may also solve for σ_r, obtaining a similar result,

$$\sigma_r = \sigma_m\frac{1}{\rho}\frac{\mathbb{E}(r) - r_f}{\mathbb{E}(m) - r_f}.$$

Here, as the expected return increases, leaving all else fixed, so does the volatility. Again, when testing this relationship empirically, a control for correlation must be in place.

The volatility of r may be decomposed into systemic and idiosyncratic components as

$$\begin{aligned}
Var(r - r_f) &= \beta(m - r_f) + \epsilon \\
\sigma_r^2 &= \beta^2\sigma_m^2 + \sigma_\epsilon^2.
\end{aligned}$$

As a result, controlling for β, *idiosyncratic risk*, σ_ϵ, is positively and monotonically related to higher returns.

We have already seen that systemic risk, σ_m, cannot be reduced while idiosyncratic risk may be diversified away. This seems to be an apparent drawback to the model: the identification of idiosyncratic risk and return is diluted away by diversification, leaving on the market. We will see an elegant proof that ameliorates this concern on the surface– namely that the model gives preferential treatment to the market over any other security or portfolio. This proof, however, will explicitly state that all investors should be holders only of a risk free asset and the market, in varying proportion determined by their risk preferences. We postpone this proof until we have established Merton's mean variance optimization problem.

[1] A very minor assumption in theory, almost by fiat, but one that may be violated using estimators over unfortunate time windows

In every case above risk – as proxied by σ_r – is directly proportional to expected returns. The implication of the CAPM model, then, and its interpretation within the framework of the Efficient Market Hypothesis, is that apparent excess returns are obtained by taking on risk.

This is not borne out in practice, however. In fact, the equity market prices risk perversely: returns to higher volatility names *underperform* low volatility names. This feature seems to generalize when one encounters the value, momentum, accrual, and fraud anomalies, for instance. In each of these cases, portfolios of *a priori* preferable names outperform less desirable names, *and they do so with more attractive risk profiles.*

In the next example we look at one of the implications of CAPM based on estimators observed in the equity markets over the last twenty years.

Example 4.1.1. Using the same cross sectional dataset we have used previously, spanning the period from 10/31/1997-5/31/2014 (200 months), we look at the average performance of volatility quartiles, recalculated monthly and held for one month. Our universe under consideration looks at the 1,000 largest US stocks by market cap each month – a liquid and tradeable set of securities.

Volatility	Q1	Q2	Q3	Q4
Average Performance (Ann.)	10.00%	10.41%	10.13%	9.44%
Volatility (Ann.)	12.07%	16.20%	19.58%	32.28%
Performance per Vol Point (Ann.)	0.83	0.64	0.52	0.29

From the table we see little to no evidence of compensation for realized volatility in the cross section. If anything a quick visual inspection of the annualized average performance indicates that the market slightly prefers less volatile stocks to more volatile. The decrease is not monotonic in annual performance, however.

Looking more closely, though, if we consider the annual performance per volatility point – something akin to $\frac{\mu}{\sigma}$ – we see a clear monotonic decrease as we go from lower to higher volatility names.

We have seen previously in our discussion of CAPM that the model implies a relationship between expected return and a product of correlation to the market and own-volatility. We did not control for correlation in this original analysis. This control can be approximated by performing so-called *double sorts.*

For each quartile of volatility, we may further subdivide into four quartiles by correlation. If equity returns follow the CAPM model, we would expect that within each volatility quartile, as we increase correlation (going from Quartile 1 to Quartile 4) we should see an increase in average annualized returns. And further, within each correlation quartile, an increase in volatility should correspond to an increase in expected return.

Volatility\Correlation	Q1	Q2	Q3	Q4
Q1	10.17%	10.30%	10.62%	8.94%
Q2	10.13%	11.32%	9.45%	10.71%
Q3	10.47%	9.31%	8.84%	11.91%
Q4	7.40%	9.06%	8.37%	12.89%

Visually, there does not appear to be evidence in the empirical table to support the hypothesis that, even when controlling for correlation, the market compensates investors for bearing risk in the equity markets – where, again, risk here is equated to realized own-volatility. We may conduct more rigorous statistical tests, but we postpone those slightly.

From the table above, there is evidence that over this period, *ex post* equity returns are negatively associated with *ex ante* volatility within each quartile of correlation, so that the control we established does not support the implications of the CAPM model.

Perhaps, though, we should focus again on some sort of risk-adjusted returns. We again take the ratio of the sample average return to the sample volatility and present the results in the table below.

Volatility\Correlation	Q1	Q2	Q3	Q4
Q1	0.86	0.89	0.85	0.59
Q2	0.70	0.72	0.57	0.53
Q3	0.58	0.50	0.42	0.49
Q4	0.25	0.27	0.24	0.36

The monotocity is more dramatic here. *But in the wrong direction of the hypotheses of CAPM.* In fact, we see that the most attractive performance per unit of risk is to be found in the lowest quartile of correlation within the lowest quartile of volatility.

One may conduct the same analysis as in the preceding example sorting by estimated β from (4.4). Based on the results here, one might expect that the empirical relationship between β and *ex post* returns is again negative. Interestingly, the double sort methodology in β and correlation does not produce the same results as the volatility/correlation double sorts. In particular, in the period under study, equity market returns show a preference for *low* β with *high* correlation.

At first glance, this seems to be counter-intuitive. However, we may interpret β as the sensitivity of an asset's return to movements in the market. Correlation, on the other hand, simply measures the linear relationship between the asset and market's returns. A high correlation indicates the asset's returns can in fact be explained by the market.

An alternative approach is to re-sort to (3.11) where we may rearrange terms to find

$$\rho = \beta \frac{\sigma_m}{\sigma_r},$$

so that a higher correlation indicates a lower own-volatility. These findings are left to the reader to verify.

4.2 Maximum Likelihood

In the preceding section we obtained estimates for α and β in CAPM models by finding critical points (minima) of quadratic functions. While we were

able to relate these values to method of moments estimators, we did not supply a justification for the usage of least squares. Here we establish one such justification.

Let $\{X_i\}_{i=1}^N$ be random variables with joint density $f_\theta(\cdot)$, where θ is some vector of parameters which determine the density. For example, if $f_\theta(\cdot) = \phi_{\mu,\Sigma}(\cdot)$, $\theta = (\mu, \Sigma)$. Or, in the case of a Student t distribution, θ may include the degrees of freedom. We have become familiar with the need for estimators, and the present case will be no exception. In particular, we identify an estimator $\hat{\theta}$.

Given observations $x_1, \ldots, x_N$, the *likelihood function* based on these observations is

$$L(\theta) = f_\theta(x_1, \ldots, x_N). \tag{4.6}$$

The *maximum likelihood estimator* of θ, $\hat{\theta}$, is the maximizer of

$$\max_\theta f_\theta(x_1, \ldots, x_N). \tag{4.7}$$

The general case is likely not applicable. A common implementation of maximum likelihood estimation assumes that $\{X_i\}_{i=1}^N$ are iid, reducing the joint density $f_\theta(\cdot)$ to a product of identical marginals,

$$f_\theta(x_1, \ldots, x_N) = \prod_{i=1}^N \tilde{f}_\theta(x_i). \tag{4.8}$$

Maximizing a product directly is generally difficult. Logarithms transform products into sums, preserve order, and are continuous on their domain. As such, the maximizer of

$$\max_\theta \log(f_\theta(x_1, \ldots, x_N)) \tag{4.9}$$

is exactly the maximizer of (4.7). In the case of iid $\{X_i\}_{i=1}^N$, (4.8) becomes

$$\log(f_\theta(x_1, \ldots, x_N)) = \sum_{i=1}^N \log(\tilde{f}_\theta(x_i)). \tag{4.10}$$

We call the logarithm of the likelihood function $\log(L(\theta))$ the log-likelihood function.

Example 4.2.1. Let $\{X_i\}_{i=1}^N$ be univariate normal random variables, and assume that the X_i's are iid. We have that $\theta = (\mu, \sigma)$. The likelihood function given observations $\{x_i\}_{i=1}^N$ is

$$L(\mu, \sigma) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

or

$$L(\mu, \sigma) \propto \prod_{i=1}^N \frac{1}{\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

where $\propto$ denotes proportional to; viz., maximizing $L(\cdot)$ is equivalent to maximizing $c \cdot L(\cdot)$ for any constant c, so we may eliminate the normalization factor $\frac{1}{\sqrt{2\pi}}$. The log-likelihood function becomes

$$l_0(\mu, \sigma) = -\sum_{i=1}^{N} \left[\log(\sigma) + \frac{(x_i - \mu)^2}{2\sigma^2} \right].$$

Now maximizing $l_0(\cdot)$ is equivalent to minimizing

$$l(\mu, \sigma) = \sum_{i=1}^{N} \left[\log(\sigma) + \frac{(x_i - \mu)^2}{2\sigma^2} \right].$$

As before, the minimum will occur exactly when $\nabla l = 0$.

Looking at the $\frac{\partial l}{\partial \mu}$ in the gradient, we solve

$$-\frac{1}{\sigma^2} \sum_{i=1}^{N} [x_i - \mu] = 0,$$

giving the estimator we have seen previously:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

The maximum likelihood estimator of σ^2, however, is given by

$$\tilde{s}_N^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2,$$

a biased estimator of the variance.

Maximum likelihood also finds application in the model

$$y_t = \beta' x_t + \epsilon_t$$

where $\beta \in \mathbb{R}^p$, $x_t \in \mathbb{R}^p$ is nonrandom, and the $\{\epsilon_t\}$ are iid distributed as

$$\epsilon \sim N(0, \sigma^2).$$

In this case, the likelihood function in (β, σ^2) may be written as,

$$\begin{aligned} L(\beta, \sigma^2) &\propto \prod_{i=1}^{N} \frac{1}{\sigma} \exp\left(\frac{\epsilon_t^2}{2\sigma^2} \right) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma} \exp\left(\frac{(y_t - \beta' x_t)^2}{2\sigma^2} \right). \end{aligned}$$

The log-likelihood function is

$$l_0(\beta, \sigma^2) = -\sum_{i=1}^{N} \log(\sigma) + \frac{(y_t - \beta' x_t)^2}{2\sigma^2}$$

which is maximized exactly when the function

$$l(\beta, \sigma^2) = N \log(\sigma) + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_t - \beta' x_t)^2$$

is minimized. The minimum occurs when the gradient is identically zero,

$$\frac{\partial l}{\beta_i} = 0$$

for $i = 1, \ldots, N$. This coincides exactly with the minimizer of

$$f(\beta) = \sum_{i=1}^{N} (y_t - \beta' x_t)^2$$

so that minimizing the quadratic function $f(\cdot)$ coincides with the maximum likelihood estimator when residuals are independent and normally distributed. We will explicitly state when this assumption is necessary in our proofs that follow.

4.3 OLS

Consider the model

$$y_t = \beta' x_t + \epsilon_t \tag{4.11}$$

with $\beta \in \mathbb{R}^p$, $x_t \in \mathbb{R}^p$ nonrandom, $\{\epsilon_t\}$ univariate random variables. In particular, we assume that each y_t is a univariate random variable, and that the randomness is driven by ϵ_t. We may rewrite the above as

$$Y = X\beta + \epsilon \tag{4.12}$$

with

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{T1} & \ldots & x_{Tp} \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_T \end{pmatrix}.$$

Here, ϵ is now a random vector.

We call the estimator, $\hat{\beta}$, obtained from minimizing the quadratic function

$$\min_{\beta} ||Y - X\beta||^2 \tag{4.13}$$

the *ordinary least squares (OLS) estimate* of β. We have seen in the previous section that in the case that $\epsilon \sim N(0, \sigma^2 I)$, the OLS estimate and maximum likelihood estimate coincide.

The following assumptions are collectively, often referred to as the *Gauss-Markov assumptions*. We have

1. $X \in \mathbb{R}^{N \times p}$ is a nonrandom matrix with $N > p$ and full rank.
2. ϵ is a random vector with $\mathbb{E}(\epsilon) = 0$.
3. $Var(\epsilon_t) = \sigma^2$ for all t, and $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$; i.e., $Cov(\epsilon) = \sigma^2 I$.

In addition to the above, we will resort to a distributional assumption for ϵ when we develop distributional properties for $\hat{\beta}$. In particular, we will assume that $\epsilon \sim N(0, \sigma^2 I)$.

Theorem 4.3.1. Under the first two Gauss-Markov assumptions, the OLS estimate of β given by

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4.14}$$

is unbiased.

Proof. Since X is full rank, we know that $(X'X)^{-1}$ is invertible. Looking at the equation

$$Y = X\beta + \epsilon$$

then, we have

$$\begin{aligned} X'Y &= X'X\beta + X'\epsilon \\ (X'X)^{-1}X'Y &= \beta + (X'X)^{-1}X'\epsilon. \end{aligned}$$

Defining $\hat{\beta}$ as the term on the left hand side,

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

we see that

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}(\beta) + \mathbb{E}\left((X'X)^{-1}X'\epsilon\right) \\ &= \beta + (X'X)^{-1}X'\mathbb{E}(\epsilon) \\ &= \beta \end{aligned}$$

so that $\hat{\beta}$ is an unbiased estimator of β.

Notice in our previous theorem that the estimator $\hat{\beta}$ is a random variable since Y is a random vector driven by ϵ. As such, we may obtain a point estimate based on a specific set of observations, but we may also want to, as in our work with the estimator for the mean, understand the distribution of $\hat{\beta}$. Similarly, the ϵ values are unobservable. As such, we will be left with estimators of ϵ. Under the full Gauss-Markov assumptions with the distributional assumption on ϵ, we may satisfactorily address both of these curiosities. Before doing so, however,

we identify an important structural property of the matrix $X(X'X)^{-1}X'$ which will be used extensively and which, further, lends to more interpretability of the OLS estimators.

We define the matrix H as

$$H = X(X'X)^{-1}X'. \tag{4.15}$$

In the literature, H is often called the *hat matrix*. We have immediately that

$$\begin{aligned} HY &= X(X'X)^{-1}X'Y \\ &= X\hat{\beta}. \end{aligned}$$

That is, the effect of multiplying Y by H is to obtain our estimate, $X\hat{\beta}$ of $X\beta$. We call

$$\hat{Y} = X\hat{\beta} \tag{4.16}$$

and show in what follows that $\hat{Y}$ is the projection of Y onto the space spanned by the columns of X. We begin by showing that H is a *projection matrix*.

We say a matrix P is a projection matrix if it satisfies the following two conditions: $P^2 = P$ and $P' = P$. P is necessarily a square matrix. Projection matrices, P, have several useful properties.

Projection Matrix Property 1: The only eigenvalues of P a projection matrix are 0 and 1.

Proof. Let λ be an eigenvalue of P with associated eigenvalue, v. By the first condition of being a projection matrix, we have that

$$\begin{aligned} Pv &= \lambda v \\ P^2v &= P(\lambda v) = \lambda Pv = \lambda^2 v \end{aligned}$$

so that

$$\lambda v = \lambda^2 v$$

or $(\lambda^2 - \lambda)v = 0$. Since v is nonzero, we have that $\lambda(\lambda - 1) = 0$, proving the result.

The intuition behind the preceding result is that projection matrices (under some appropriate basis) fix some subspace of $\mathbb{R}^N$ and send to zero its orthogonal complement. This motivates the third property we will note. First, we look at $I - P$.

Projection Matrix Property 3: $I - P$ is also a projection.

Proof. We verify the properties of a projection matrix. We have

$$\begin{aligned} (I - P)^2 &= I - 2P + P^2 \\ &= I - P \end{aligned}$$

and

$$(I - P)' = I' - P' = I - P.$$

Returning to interpreting the eigenvalue property, we have:

Projection Matrix Property 3: Any vector x may be decomposed as $x = u + v$ where $(u, v) = 0$ and $u = Px$. In this case, $||x||^2 = ||u||^2 + ||v||^2$.

Proof. The statement of the property dictates the proof. Let $u = Px$. We verify the claims in turn. We have yet to define v, but note that if

$$x = u + v$$

then necessarily

$$v = x - u = x - Px = (I - P)x.$$

The inner product of u and v is now

$$\begin{aligned}(u, v) &= (Px, (I - P)x) \\ &= P(x, x)(I - P)' \\ &= ||x||^2 P(I - P) \\ &= 0\end{aligned}$$

Since $P(I - P) = P - P^2 = 0$. The norm of x, then becomes

$$\begin{aligned}||x||^2 &= (x, x) \\ &= (u + v, u + v) \\ &= (u, u) + 2(u, v) + (v, v) \\ &= ||u||^2 + ||v||^2.\end{aligned}$$

We call $u = Px$ in the above proof the *projection of x by P*.

Returning to the matrix obtained from our OLS estimate, the matrix H is a projection matrix since

$$\begin{aligned}H^2 &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &\quad \text{by associativity} \\ &= X(X'X)^{-1}X' \\ &= H.\end{aligned}$$

Now for matrices A and B, $(A')^{-1} = (A^{-1})'$, and $(AB)' = B'A'$ so that

$$\begin{aligned}H' &= \left(X(X'X)^{-1}X'\right)' \\ &= X\left((X'X)^{-1}\right)'X' \\ &= X\left((X'X)'\right)^{-1}X' \\ &= X\left((X'X)\right)^{-1}X' \\ &= H.\end{aligned}$$

As a result, $\hat{Y} = HY$ is the projection of Y by H. By construction, then, $\hat{Y}$ is the projcetion of Y onto the space spanned by the columns of X.

Defining the estimator of the residuals, ϵ by

$$\hat{\epsilon} = Y - \hat{Y}, \tag{4.17}$$

we have that $\hat{\epsilon} = (I - H)Y$, or that the estimated residuals are a projection of Y onto the orthogonal complement of the space spanned by the columns of X. Necessarily, then, $\hat{\epsilon}$ and $\hat{Y}$ are orthogonal; i.e., $(\hat{\epsilon}, \hat{Y}) = 0$. We also have the deconstruction $Y = \hat{Y} + \hat{\epsilon}$ with norm $||Y||^2 = ||\hat{Y}||^2 + ||\hat{\epsilon}||^2$.

We may also show that

$$\hat{\epsilon} = (I - H)\epsilon \tag{4.18}$$

but leave this to the reader.

Notice that under the Gauss-Markov assumptions, the covariance of the estimator of ϵ is not a diagonal matrix. In particular, we have

$$\begin{aligned} Cov(\hat{\epsilon}) &= Cov((I - H)Y) \\ &= (I - H)Cov(Y)(I - H)' \\ &= (I - H)Cov(\epsilon)(I - H) \\ &= \sigma^2(I - H)^2 \\ &= \sigma^2(I - H). \end{aligned}$$

And

$$Var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii}),$$

so that the variables, the so-called *jackknife residuals*,

$$\frac{\hat{\epsilon}_i}{\sigma\sqrt{1 - h_{ii}}} \tag{4.19}$$

have constant unit variance if the model is specified correctly. Notice, however, that σ is not known and must be estimated. We return to this issue in the exercises, but will note it below as well in determining some statistical properties for $\hat{\beta}$.

Under these same assumptions, we also have that the covariance of $\hat{\beta}$ is

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}. \tag{4.20}$$

This follows directly from $\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$, giving

$$\begin{aligned} Cov(\hat{\beta}) &= Cov(\beta + (X'X)^{-1}X'\epsilon) \\ &= (X'X)^{-1}X'Cov(\epsilon)\left((X'X)^{-1}X'\right)' \\ &= \sigma^2(X'X)^{-1}X'\left((X'X)^{-1}X'\right)' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

If we assume further that $\epsilon \sim N(0, \sigma^2 I)$, then from what has proceeded, $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$. In this case,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0, 1), \tag{4.21}$$

where $C = (X'X)^{-1}$.

Notice that in both of the preceding examples we have formulated our final result in terms of the unknown parameter σ. We will necessarily require an estimator s of σ. Having this estimator in hand, we will be in a position to look at the distribution of

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}, \tag{4.22}$$

which, perhaps surprisingly, will be the Student t distribution.

We begin by proving that s defined by

$$s^2 = \frac{1}{N-p}\sum_{t=1}^{N} \hat{\epsilon}_t^2 \tag{4.23}$$

is an unbiased estimator of σ, and that under the Gauss-Markov assumptions with added distributional requirement, $\epsilon \sim N(0, \sigma^2 I)$,

$$\frac{N-p}{\sigma^2} s^2 \sim \chi^2_{N-p} \tag{4.24}$$

where χ^2_ν is the Chi-Squared distribution with ν degrees of freedom defined as the sum of squared iid standard normal variables; viz., we say $W \sim \chi^2_\nu$ if

$$W = Z_1^2 + \cdots + Z_\nu^2 \tag{4.25}$$

with $Z_i \sim N(0,1)$ iid.

Proof. Let

$$\hat{\epsilon} = \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_N \end{pmatrix}.$$

be the OLS estimate of ϵ in (4.12). Since $\mathbb{E}(\hat{\epsilon}) = 0$, the covariance of $\hat{\epsilon}$ is exactly

$$Cov(\hat{\epsilon}) = \mathbb{E}(\hat{\epsilon}\hat{\epsilon}')$$

We have, implementing a technique seen previously, that

$$||\hat{\epsilon}||^2 = \hat{\epsilon}'\hat{\epsilon} = tr\left(\hat{\epsilon}'\hat{\epsilon}\right).$$

Now by linearity we have that for random matrices, A, $\mathbb{E}(tr(A)) = tr(\mathbb{E}(A))$ so that

$$\mathbb{E}\left(tr\left(\hat{\epsilon}'\hat{\epsilon}\right)\right) = tr\left(\mathbb{E}\left(\hat{\epsilon}'\hat{\epsilon}\right)\right).$$

And, as before,

$$tr\left(\mathbb{E}\left(\hat{\epsilon}'\hat{\epsilon}\right)\right) = tr\left(\mathbb{E}\left(\hat{\epsilon}\hat{\epsilon}'\right)\right).$$

So that

$$\begin{aligned} \mathbb{E}\left(||\hat{\epsilon}||^2\right) &= tr(Cov(\hat{\epsilon})) \\ &= tr(\sigma^2(I-H)). \end{aligned}$$

Since H is a projection into a p-dimensional space, there exists a basis such that

$$H = \begin{pmatrix} I_p & 0 \\ 0 & 0_{N-p} \end{pmatrix}.$$

This follows from the fact that the eigenvalues of H are zero or one, and that the rank of H is p. Clearly, under this basis, $tr(H) = p$. Finally, since $tr(\cdot)$ is independent of the choice of basis, $tr(H) = p$ under any basis. Similarly, $I - H$ will necessarily have trace

$$tr(I - H) = N - p.$$

Therefore,

$$\begin{aligned} \mathbb{E}\left(||\hat{\epsilon}||^2\right) &= \sigma^2 tr(I - H) \\ &= \sigma^2(N - p), \end{aligned}$$

or,

$$\begin{aligned} \mathbb{E}\left(\sum_{t=1}^{T} \hat{\epsilon}_t^2\right) &= \mathbb{E}\left(\sum_{t=1}^{T} (y_t - \beta' x_t)^2\right) \\ &= \sigma^2(N - p), \end{aligned}$$

giving that the *scaled sum of the squared residuals*,

$$s^2 = \frac{1}{N - p}\mathbb{E}\left(||\hat{\epsilon}||^2\right)$$

is an unbiased estimator of σ^2. This result is independent of the distributional assumption of ϵ; i.e., it used only the Gauss-Markov assumptions.

If we further make the distributional assumption that ϵ, $\epsilon \sim N(0, \sigma^2 I)$, we may show that

$$\frac{1}{\sigma^2}||\hat{\epsilon}||^2 \sim \chi^2_{N-p} \tag{4.26}$$

as in (4.24).

Let Q be the orthogonal change of basis matrix for the given matrix H such that

$$QHQ' = \begin{pmatrix} I_p & 0 \\ 0 & 0_{N-p} \end{pmatrix}.$$

Notice that this also implies that

$$Q(I - H)Q' = \begin{pmatrix} 0_p & 0 \\ 0 & I_{N-p} \end{pmatrix}.$$

For this Q, define

$$Z = \frac{1}{\sigma} Q\epsilon = \begin{pmatrix} Z_1 \\ \dots \\ Z_N \end{pmatrix}.$$

The key observation here is that $\hat{\epsilon} = (I - H)\epsilon$ and that the simplified basis allows for a more tractable approach.

Under our last assumption, Z is a multivariate normal random variable since a linear combination of jointly normal random variables is jointly normal. The mean and covariance of Z are given by

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}\left(\frac{1}{\sigma}Q\epsilon\right) \\ &= \frac{1}{\sigma}Q\mathbb{E}\left(\epsilon\right) \\ &= 0, \end{aligned}$$

by linearity of the expectation operator, and

$$\begin{aligned} Cov(Z) &= Cov\left(\frac{1}{\sigma}Q\epsilon\right) \\ &= \frac{1}{\sigma^2}QCov(\epsilon)Q' \\ &= \frac{\sigma^2}{\sigma^2}QIQ' \\ &= I \end{aligned}$$

since $QQ' = Q'Q = I$. Hence

$$Z \sim N(0, I).$$

This also shows that the $\{Z_t\}$ are iid. We may recover ϵ as

$$\epsilon = \sigma Q' Z.$$

Since Q is orthogonal, we have that

$$\begin{aligned} ||Q\hat{\epsilon}||^2 &= \hat{\epsilon}'Q'Q\hat{\epsilon} \\ &= \hat{\epsilon}'\hat{\epsilon} \\ &= ||\hat{\epsilon}||^2, \end{aligned}$$

giving

$$\begin{aligned} ||\hat{\epsilon}||^2 &= ||Q\hat{\epsilon}||^2 \\ &= ||Q(I-H)\epsilon||^2 \\ &= ||Q(I-H)Q'Q\epsilon||^2 \\ &= \sigma^2||Q(I-H)Q'Z||^2 \\ &= \sigma^2 Z' \begin{pmatrix} 0_p & 0 \\ 0 & I_{N-p} \end{pmatrix}^2 Z \\ &= \sigma^2 \sum_{t=p+1}^{N} Z_t^2. \end{aligned}$$

As a result, we clearly have that $\frac{1}{\sigma^2}||\hat{\epsilon}||^2 \sim \chi^2_{N-p}$ as desired.

We may now return to the ratio in (4.22),

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}.$$

We proved that under the full Gauss-Markov assumptions with additional distributional assumption for residuals,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1).$$

We use this fact along with the result just obtained to decompose (4.22) as

$$\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{c_{ii}}}}{\frac{s}{\sigma}}.$$

The result is a ratio of random variables with known distributions. In particular, in the numerator we have a standard normal random variable. The denominator is the square root of a χ^2 random variable with $N-p$ degrees of freedom scaled by $\frac{1}{N-p}$.

Remarkably, a random variable

$$T = \frac{Z}{\sqrt{\frac{1}{M}X}} \tag{4.27}$$

with $Z \sim N(0,1)$ and $X \sim \chi^2_M$ and Z and X independent is distributed as a standard Student t random variable with M degrees of freedom as seen in (2.13); i.e., $T \sim \mathrm{St}(0,1;M)$. We will abbreviate this notation as

$$T \sim t_M. \tag{4.28}$$

We do not prove the correspondence between our initial definition and that presented here, however.

The independence between $\hat{\beta}_i$ and s^2 has not been shown either. With the same assumptions as above we have that

$$\begin{aligned} \hat{\epsilon} &= (I-H)\epsilon \\ X(\hat{\beta}-\beta) &= H\epsilon, \end{aligned}$$

so that the covariance inner product between $\hat{\epsilon}$ and $X(\hat{\beta}-\beta)$ satisfies

$$\begin{aligned} (\hat{\epsilon}, X(\hat{\beta}-\beta)) &= ((I-H)\epsilon, H\epsilon) \\ &= (I-H)\sigma^2 IH \\ &= \sigma^2(H-H^2) = 0. \end{aligned}$$

Under the normality assumption for ϵ, this shows that $\hat{\epsilon}$ and $X(\hat{\beta} - \beta)$ are independent as well. Since we are also assuming that X is full rank,

$$\begin{aligned} 0 &= (\hat{\epsilon}, X\hat{\beta}) \left((X'X)^{-1} X' \right)' \\ &= (\hat{\epsilon}, \hat{\beta}) \end{aligned}$$

as well, so that $\hat{\epsilon}$ and $\hat{\beta}$ are independent. Therefore s^2 and the components of $\hat{\beta}$ are independent as well.

4.3.1 Hypothesis Testing, Confidence Intervals, and Prediction Intervals

With the distribution of $\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}$ now known (under the extended Gauss-Markov assumptions), we may construct symmetric *confidence intervals* for each β_i in turn. In particular, since t_{N-p} is symmetric about zero, a $100(1-\alpha)\%$ confidence interval for $\frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}}$ is given by

$$-t_{N-p;1-\alpha/2} \leq \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \leq t_{N-p;1-\alpha/2}, \tag{4.29}$$

or

$$\hat{\beta}_i - s\sqrt{c_{ii}} t_{N-p;1-\alpha/2} \leq \beta \leq \hat{\beta}_i + s\sqrt{c_{ii}} t_{N-p;1-\alpha/2}, \tag{4.30}$$

where $t_{N-p;1-\alpha/2}$ is the $1 - \alpha/2$ quantile of t_{N-p}. Clearly, with this choice,

$$\mathbb{P}\left(\left| \frac{\beta_i - \beta}{s\sqrt{c_{ii}}} \right| \leq t_{N-p;1-\alpha/2} \right) = 1 - \alpha.$$

We emphasize once more that this is a component-wise test.

A common issue in model development is determining whether a particular variable should be included in the model. As one approximation to the answer answer, we may approach the question by asking whether a particular β_i is likely to be zero. Necessarily this requires setting a specific confidence level. For example, for a fixed α, we may determine if the interval

$$\left(\hat{\beta}_i - s\sqrt{c_{ii}} t_{N-p;1-\alpha/2}, \hat{\beta}_i + s\sqrt{c_{ii}} t_{N-p;1-\alpha/2} \right)$$

contains zero. If it does, there is evidence that the ith variable may be omitted from the model (relative to this particular confidence level α).

The preceding reasoning may be formalized into so-called *hypothesis testing* wherein a *null hypothesis* is formulated – and typically denoted by H_0 – and statistical values are then measured. If the statistical values are extremely unlikely (relative to some threshold), we may reject the null hypothesis.

Example 4.3.1. We consider the null hypothesis

$$H_0 : \beta_i = 0.$$

Under the null hypothesis,

$$b_i = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}} \sim t_{N-p}.$$

The probability of observing a value at least as large as $|b_i|$ is exactly

$$p_0 = 2 \cdot (1 - St^{-1}_{0,1;N-p}(b_i)).$$

We *reject the null hypothesis at significance level* α if $p_0 < \alpha$. We call p_0 the *p-value* of the test statistic associated with H_0.

Notice that we reject the null hypothesis exactly when

$$|b_i| > t_{N-p;1-\alpha/2},$$

which coincides with our previous observation that the confidence interval for β_i would in fact contain zero.

Falsely rejecting the null hypothesis when it is in fact true is called a *type I error*. When the extended Gauss-Markov assumptions obtain, the probability of a type I error is equivalent to the confidence level α. It is not true that this holds generally.

Example 4.3.2. We return to examining *ex post* equity returns sorted by volatility quartiles. In our previous example we provided evidence that there is monotonicity in these *ex post* returns, but in a direction counter to that implied by CAPM. As such, we investigate constructing portfolios *long in low trailing volatility* and *short in high volatility* stocks.

At each month, we sort the same universe of 1,000 stocks considered previously by own volatility into quartiles, with quartile four holding the highest volatility names. At time t, let r_{1t} and r_{4t} be the average return over the next month $(t, t+1]$ for the lowest and highest quartile buckets, respectively. In addition, let σ_{1t} and σ_{4t} be the square root of the average trailing variance of the lowest and highest buckets.

We construct a portfolio with approximately equal long and short volatility by being long one unit of a portfolio of equal weighted names from quartile one, and short $\frac{\sigma_{1t}}{\sigma_{4t}}$ units of the equally weighted names from quartile four. The return over the month post construction is given by

$$r_t = r_{1t} - \frac{\sigma_{1t}}{\sigma_{4t}} r_{4t}.$$

Over the period in our backtest, the average monthly return for this portfolio is 55 bps, annualizing to approximately 6.58%.

Regressing on a market proxy, we find the CAPM α and β (ignoring the risk free rate) to be 0.60 and -0.0702, respectively. That is, the returns of the portfolio are generated without exposure to the market.

Further, we may consider the null hypothesis

$$H_0 : \alpha = 0.$$

A confidence interval for α at the confidence level 0.05 is found to be $[0.0013, 0.0106]$ so that we may reject the null hypothesis at the 5% level.

In addition to examining confidence intervals (and making inferences based on these intervals), we may also construct confidence intervals about the mean response from the linear models considered. In particular, and again, for the original OLS model (4.12), OLS (least squares) estimates of β, $\hat{\beta}$, and assuming the Gauss-Markov with distributional assumptions obtains, an unbiased estimate of the variance of an input, x_t, is given by

$$\begin{aligned} Var\left(\hat{y}_t\right) &= Var\left(\hat{\beta}' x_t\right) && (4.31)\\ &= x_t' Cov\left(\hat{\beta}\right) x_t && (4.32)\\ &= \sigma^2 x_t' \left(X'X\right)^{-1} x_t. && (4.33) \end{aligned}$$

Following the same procedure as above, we find that a $100 \cdot (1-\alpha)\%$ confidence interval for $\beta' x_t$ is

$$\hat{\beta}' x_t \pm \left(s\sqrt{x_t' C x_t}\right) t_{N-p;1-\alpha/2}, \tag{4.34}$$

where, continuing our convention, we denote $C = \left(X'X\right)^{-1}$. The complete verification of this result, following that of (4.29), is left as an exercise.

The confidence interval just exhibited is both an estimate of the confidence interval of the mean and dependent on $x = x_t$ being an observation in the derivation of $\hat{\beta}$. We may also consider the case of a new observation, or, equivalently, a prediction interval about an observation x as opposed to the mean response. This is in contrast with the previous result as we may no longer discard the variance of ϵ. We have,

$$\begin{aligned} Var\left(\hat{y}\right) &= Var\left(\hat{\beta}' x + \epsilon\right)\\ &= x' Cov\left(\hat{\beta}\right) x + \sigma^2\\ &= \sigma^2 x' \left(X'X\right)^{-1} x + \sigma^2, \end{aligned}$$

so that, in the usual manner, a $100 \cdot (1-\alpha)\%$ prediction interval for $\beta' x$ is

$$\hat{\beta}' x \pm \left(s\sqrt{1 + x_t' C x_t}\right) t_{N-p;1-\alpha/2}. \tag{4.35}$$

4.3.2 Submodel Testing

The ability to test the null hypothesis for the inclusion or exclusion of a particular variable is a powerful result. However, we may also want to construct null hypotheses about linear submodels of the full model specified in (4.12); i.e., we would like to consider the case that

$$\mathbb{E}(Y) \subseteq \mathcal{L}_0$$

for some (strict) subset $\mathcal{L}_0 \subset \mathcal{L}$, where $\mathcal{L}$ is the space spanned by the columns of X. Necessarily the dimension of $\mathcal{L}_0$ will be less than $dim(\mathcal{L})$. We will denote $dim(\mathcal{L}_0)$ by r in what follows.

Under the assumptions of the model, we have

$$\mathbb{E}(Y) = \mathbb{E}(X\beta + \epsilon) = X\beta, \tag{4.36}$$

so that clearly

$$\mathbb{E}(Y) \subseteq \mathcal{L}.$$

From our previous work, ordinary least squares projects Y onto $\mathcal{L}$ through the matrix H. For a fixed subspace $\mathcal{L}_0 \subset \mathcal{L}$, we may consider a similar projection of Y onto $\mathcal{L}_0$ by some matrix G, with

$$\hat{Y}_0 = GY. \tag{4.37}$$

Notice that since H projects into $\mathcal{L}$, $GH = G$ and $HG = G$. As a result, $Y - \hat{Y}_0$ and $\hat{Y} - \hat{Y}_0$ are orthogonal. First note that

$$\begin{aligned} \mathbb{E}(Y - \hat{Y}) &= \mathbb{E}((I - H)\epsilon) = 0 \\ \mathbb{E}(\hat{Y}_0 - \hat{Y}) &= \mathbb{E}((H - G)\epsilon) = 0. \end{aligned}$$

The covariance is therefore

$$\begin{aligned} Cov(Y - \hat{Y}, \hat{Y}_0 - \hat{Y}) &= Cov((I - H)\epsilon, (H - G)\epsilon) \\ &= (I - H)Cov(\epsilon)(H - G)' \\ &= \sigma^2(I - H)(H - G) \\ &= \sigma^2\left(H - H^2 - G + HG\right) \\ &= 0. \end{aligned}$$

Assuming joint normality of ϵ implies that $Y - \hat{Y}$ and $\hat{Y}_0 - \hat{Y}$ are independent. As a result, $||Y - \hat{Y}||^2$ and $||\hat{Y}_0 - \hat{Y}||^2$ are independent as well.

We next consider the null hypothesis

$$H_0 : \mathbb{E}(Y) \subseteq \mathcal{L}_0. \tag{4.38}$$

Under the null hypothesis, the expected value of Y lies in a space where we don't need all of the columns of X. In other words, some subset (or linear combination) of variables is sufficient to describe Y.

It also follows under H_0 and the Gauss-Markov assumptions that $\frac{1}{p-r}||\hat{Y} - \hat{Y}_0||^2$ is an unbiased estimator of σ^2, and further that under the extended Gauss-Markov assumptions,

$$\frac{1}{p-r}||\hat{Y} - \hat{Y}_0||^2 \sim \chi^2_{p-r} \tag{4.39}$$

The result follows considering the projection $H - G$ (rather than $I - H$ as in the preceding proof) and the change of basis matrix Q satisfying

$$Q(H - G)Q' = \begin{pmatrix} 0_{N-(p-r)} & 0 \\ 0 & I_{p-r} \end{pmatrix}.$$

The test statistic we will use to evaluate H_0 will be

$$F = \frac{N-p}{p-r}\frac{||\hat{Y}-\hat{Y}_0||^2}{||Y-\hat{Y}||^2} \tag{4.40}$$

We say (with clear motivation stemming from the results above) that a random variable constructed as the ratio of independent χ^2 random variables as in

$$F_{d_1,d_2} = \frac{\frac{1}{M}U}{\frac{1}{N}V} \tag{4.41}$$

follows the *F distribution with* (d_1, d_2) *degrees of freedom*, where d_1 and d_2 are the degrees of freedom for U and V, respectively. Since under our assumptions, $\frac{1}{\sigma^2}||Y-\hat{Y}||^2 \sim \chi^2_{N-p}$ and $\frac{1}{\sigma^2}||\hat{Y}-\hat{Y}_0||^2 \sim \chi^2_{p-r}$ and independence has already been shown,

$$\frac{\frac{1}{\sigma^2(p-r)}||\hat{Y}-\hat{Y}_0||^2}{\frac{1}{\sigma^2(N-p)}||Y-\hat{Y}||^2} = \frac{N-p}{p-r}\frac{||\hat{Y}-\hat{Y}_0||^2}{||Y-\hat{Y}||^2} \sim F_{p-r,N-p},$$

giving that the test statistic in (4.40) may be tested against a known distribution. Notice that by orthogonality we have

$$||Y-\hat{Y}_0||^2 = ||Y-\hat{Y}||^2 + ||\hat{Y}-\hat{Y}_0||^2, \tag{4.42}$$

or

$$||\hat{Y}-\hat{Y}_0||^2 = ||Y-\hat{Y}_0||^2 - ||Y-\hat{Y}||^2.$$

This gives an alternate formulation of the above as

$$\frac{N-p}{p-r}\frac{||\hat{Y}-\hat{Y}_0||^2}{||Y-\hat{Y}||^2} = \frac{N-p}{p-r}\frac{||Y-\hat{Y}_0||^2 - ||Y-\hat{Y}||^2}{||Y-\hat{Y}||^2}.$$

This refactoring is more than cosmetic as the ratio now only involves the terms $||Y-\hat{Y}_0^2||^2$ and $||Y-\hat{Y}^2||^2$, the sum of the squared residuals from the submodel and full model, respectively; i.e., we have

$$\frac{N-p}{p-r}\frac{||\hat{Y}-\hat{Y}_0||^2}{||Y-\hat{Y}||^2} = \frac{N-p}{p-r}\frac{RSS_{\mathcal{L}_0} - RSS_{\mathcal{L}}}{RSS_{\mathcal{L}}},$$

where RSS is defined as the *residual sum of squares* of a given model. In particular, the full model residual sum of squares is given by

$$RSS_{\mathcal{L}} = ||Y-X\hat{\beta}||^2 = \sum_{t=1}^{T}\left(y_t - \hat{\beta}' x_t\right)^2.$$

We call the null hypothesis test constructed above an *F test*. As before, we reject the null hypothesis, H_0 at confidence level α if $F > F_{p-r,N-p;1-\alpha}$, where as before, $F_{d_1,d_2,1-\alpha}$ is the $100\cdot(1-\alpha)$ percentile of the F distribution with d_1

and d_2 degrees of freedom. Notice that an F test is a one sided test, in contrast to the t test. An F test may be constructed whenever we have *nested models*, and gives a method for evaluating submodels; viz., if we do not reject the null hypothesis, then we may prefer the submodel giving rise to $\mathcal{L}_0$.

Example 4.3.3. Consider the model

$$\textit{Model 1 } y_t = \beta_1 x_{1t} + \cdots + \beta_p x_{Nt} + \epsilon_t$$

and assume there are $t = 1, \ldots, T$ observations and that the extended Gauss-Markov assumptions hold. Let

$$H_0 : \beta_1 = 2\beta_2.$$

Under H_0, we consider

$$\begin{aligned} \textit{Model 2 } y_t &= (2\beta_2)x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{Nt} + \epsilon_t \\ &= \beta_2(2x_{1t} + x_{2t}) + \cdots + \beta_p x_{Nt} + \epsilon_t. \end{aligned}$$

Obtaining the RSS of both models (denoted by RSS_1 and RSS_2, respectively) after this second formulation is now clear. As is the reframing of H_0 in terms of a subspace of $\mathcal{L} = span(X)$. An F test at confidence level α checks

$$(N-p)\frac{RSS_2 - RSS_1}{RSS_1}$$

against the value $F_{1,N-p;1-\alpha}$. If the test statistic is greater than $F_{1,N-p;1-\alpha}$, we reject the null hypothesis.

4.3.3 Variable Selection

In every case in OLS, a model with more input variables will fit better than one with less. However, additional regressors in the design matrix, $X \in \mathbb{R}^{N\times p}$, increases the chances of *collinearity* among the columns of X. In the worst case, this gives a singular $X'X$, but may also produce nearly singular matrices as well.

Let

$$X'X = Q\Lambda Q'$$

be the decomposition of $X'X$ into its eigenvalues; i.e., $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ and Q orthonormal. The inverse of $X'X$ is given

$$(X'X)^{-1} = Q\Lambda^{-1}Q'.$$

If one of the λ_i is close to zero we have that Λ^{-1} has a very large diagonal element. Therefore, in the case of collinearity in the columns of X, the covariance of $\hat{\beta}$ given by $\sigma^2 (X'X)^{-1}$, is potentially very large. In other words, in the case of collinearity, the OLS estimates are less certain.

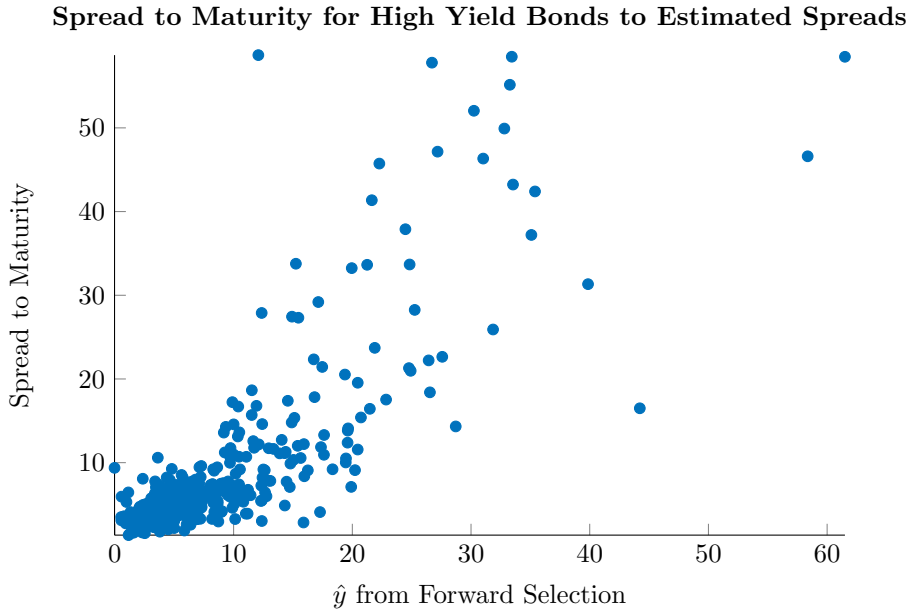

Figure 4.1: Spread to maturity for bonds in the Merrill High Yield Index on 2/28/2016 against estimated values from a forward-selected model chosen.

Hence some balance between the addition of variables for fit and estimation accuracy must necessarily be observed. In what follows we look at various stepwise selection procedures. In each case we let P be the maximum number of possible regressors. Notice that considering each possible model is computationally difficult due to the combinatoric nature of the problem. Instead we focus on building up a model one factor at a time, building a model by paring away factors from the P available, and a hybrid of these two methods.

Forward Selection

For a design matrix $\tilde{X} \in \mathbb{R}^{N \times p'}$, we define the *partial correlation* between $\{y_i\}_{i=1}^N$ and $\{w_i\}_{i=1}^N$ with respect to $\tilde{X}$ to be the correlation of the estimated residuals $\{\hat{\epsilon}_{yi}\}$ and $\{\hat{\epsilon}_{wi}\}$ given by

$$\hat{\epsilon}_{yi} = y_i - \hat{y}_i$$

and

$$\hat{\epsilon}_{wi} = w_i - \hat{w}_i$$

where $\hat{y}_\cdot$ and $\hat{w}_\cdot$ are the OLS regression estimates of $y_\cdot$ and $w_\cdot$, respectively, using the design matrix $\tilde{X}$.

In the forward selection algorithm, we initialize by calculating the correlation between $\{y_i\}_{i=1}^N$ and each column of X, $\{x_i j\}_{i=1}^N$, for $j = 1, \ldots, P$. This gives

a set of correlations $\{\hat{\rho}_j\}_{j=1}^{P}$. The first factor of the model is the regressor associated with the maximum correlation in this set.

Let $\tilde{X}_P$ be the design matrix with all regressors included, and define $\tilde{X}_{P-k}$ as the resulting matrix after removing the k regressors added to the model under the forward selection procedure. Similarly, let the design matrix of the model at this point be denoted X_k. To determine the next addition (should there be one), we look at the partial correlation between $\{y_i\}_{i=1}^{N}$ and each column of $\tilde{X}_{P-k}$ with respect to X_k.

A candidate model is then considered by adding the regressor with maximum partial correlation to the k factors already selected. Denote this new regressor by x^{k+1}. A partial F-test is then conducted at a fixed confidence level (usually 0.05 or 0.01) between the candidate model the model with k factors. If we fail to reject the null hypothesis

$$H_0 : \beta_{k+1} = 0,$$

the procedure terminates and the model is fixed with k factors and design matrix X_k. Otherwise, the design matrix is updated to X_{k+1} by adding x^{k+1} and another regressor is considered.

Example 4.3.4. We have primarily considered equity returns in our modeling thus far. Further, we have only considered technical data; i.e., data coming from market pricing. Here we look at explaining the cross section corporate bond spreads through a model using a mixture of technical and so-called fundamental data. We consider four possible regressors:

- *Debt to Market Cap*: defined as the ratio of the sum of long and short term debt as reported by a company on its balance sheet to the most recent market cap (stock price $\times$ common shares outstanding).
- *Enterprise Value*: The sum of
  - *Market Cap*: most recent stock price $\times$ the number of common shares outstanding
  - *Preferred Equity*: the value of preferred equity reported on the company's balance sheet. Preferred equity offers a slightly senior position to equity holders, often with fixed dividend payments similar to a fixed income instrument
  - *Total Debt*: the sum of long and short term debt held by the company as reported on their balance sheet
  - *Minority Interest*: the value of the company's subsidiaries owned by minority shareholders as reported on the balance sheet

  minus *Cash* and *Short Term Investments*.

  Enterprise value is a proxy for a total valuation of the firm. If another company were to take over, common and preferred shares, minority interest, and debt would all have to be considered. Short term investments would be liquidated as cash, and cash is, well, cash.

- *Equity Volatility*: Standard deviation of 121 weeks of equity returns.
- *CreditGrades Default Probability*: Risky debt instruments have some probability that they will no longer pay their contractual obligations in the future – else everywhere there would be one common rate for borrowers of the same tenor, all of whom are perfectly creditworthy. Upon default, the assets of the firm are distributed amongst bondholders after liquidation. This portion of value recovered in liquidation is the *recovery value*. CreditGrades extended work by Merton (the so-called *value of the firm*) to incorporate balance sheet items and volatility to approximate a probability of default at the firm level. We will see these ideas fully formed in a later chapter. Here we take it as a reasonable input to describe spreads.

The correlation between these items and the corporate bond spreads we consider is given in the following table.

	Spread	D/M	EV	Equity Vol	CreditGrades p
Spread	1	0.59	-0.07	0.77	0.71
D/M		1	0.02	0.45	0.55
EV			1	-0.11	-0.00
Equity Vol				1	0.87
Credit Grades p					1

Immediately apparent is the high correlation between Equity Volatility and spreads. And, even more, the relationship between the CreditGrades default probability and equity volatility. We see, too, a positive relationship between spreads and leverage (Debt to Market Cap) and a slightly negative and near zero relationship to size as proxied by Enterprise Value.

The environment at the time this data was pulled was distressed, with great concern for the viability of many names in the energy sector.

We note, too, that the relationship between risk and compensation stands in stark contrast to the results for equities seen above where we saw compensation for *lower risk*. This first experience in bond pricing shows a hint of variation that holds true more generally in the credit markets.

We proceed to build a regression model out of the factors under consideration using the forward selection procedure. Using the notation above, we have $P = 4$, and

$$\tilde{X}_P = \begin{pmatrix} | & | & | & | \\ \text{DM} & \text{EV} & \text{VOL} & \text{CG} \\ | & | & | & | \end{pmatrix} \in \mathbb{R}^{N \times P}$$

with $N = 413$. We normalize each variable, and fix a confidence level of α_c. Based on the correlation matrix above, we initialize with

$$X_1 = \begin{pmatrix} | \\ \text{VOL} \\ | \end{pmatrix}$$

which gives

$$\tilde{X}_{P-1} = \begin{pmatrix} | & | & | \\ \text{DM} & \text{EV} & \text{CG} \\ | & | & | \end{pmatrix} \in \mathbb{R}^{N \times P-1}$$

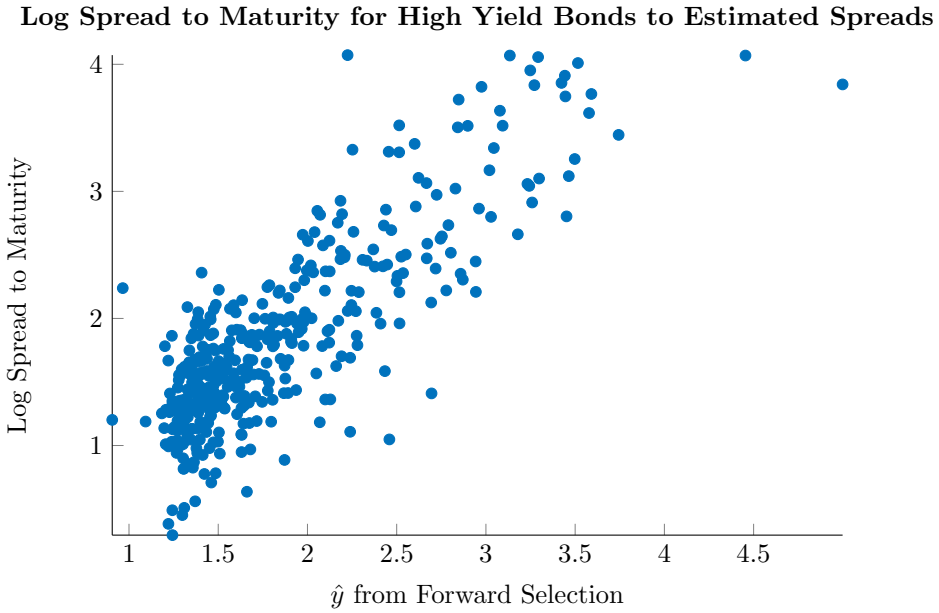

Figure 4.2: Log spread to maturity for bonds in the Merrill High Yield Index on 2/28/2016 against estimated values from a forward-selected model chosen.

The iterative algorithm then proceeds. For each column of $\tilde{X}_{P-1}$ – denoted by $X_{P-1,j}$ for $j = 1, \ldots, P-1$ – we regress

$$X_{P-1,j} = \alpha_{j,1} + X_1\beta + \epsilon_{j,1}$$

and obtain estimates $\hat{X}_{P-1,j}$ of $X_{P-1,j}$ and estimated residuals

$$\hat{\epsilon}_{j,1} \in \mathbb{R}^N$$

for $j = 1, \ldots, P-1$. We also regress $Y =$ *spread to maturity* on X_1 as

$$Y = \alpha_{Y,1} + X_1\beta + \epsilon_{Y,1}.$$

Again, we calculate estimated residuals of this model

$$\hat{\epsilon}_{Y,1} \in \mathbb{R}^N.$$

From these values, we calculate the partial correlation between each column of $\tilde{X}_{P-1}$ and Y with respect to X_1 as the correlations between $\hat{\epsilon}_{j,1}$ and $\hat{\epsilon}_{Y,1}$. We choose as a candidate addition to the model the variable with highest partial correlation. Perhaps not surprisingly the next addition is the CreditGrades default probability.

We obtain a candidate model

$$Y = \alpha_{Y,1'} + X_{1'}\beta + \epsilon_{Y,1'}.$$

where

$$X_{1'} = \begin{pmatrix} | & | \\ \text{VOL} & \text{CG} \\ | & | \end{pmatrix},$$

and conduct an F test with respect to the submodel consisting of X_1 and intercept. The test statistic, $f_{1'}$ is (ostensibly) distributed as $F_{1,N-3}$ If $f_{1'}$ is less than the critical value of $F_{1,N-3}$ determined by α_c, then the iteration is terminated and the model is X_1 with intercept. Otherwise, we add the factor to the model and update so that

$$X_2 = X_{1'}$$

and $\tilde{X}_{P-2}$ is obatined by removing the newest addition to the model. In this case,

$$\tilde{X}_{P-2} = \begin{pmatrix} | & | \\ \text{DM} & \text{EV} \\ | & | \end{pmatrix}.$$

The iteration continues until the stopping condition is met or there are no more possible additions.

Here, Debt to Market Cap is added next, followed by the candidate addition of Enterprise Value which fails to reject the null hypothesis that the submodel is sufficiently explanatory.

Figure 4.3.3 shows the result of plotting the Y (spreads) against $\hat{Y}$ as determined by the forward selection model. We expect to see, in a well-specified model, a linear relationship between the two with something close to homoscedasticity (uniform variance), allowing for the fact that estimated residual variances aren't constant as determined by (4.19).

Based on the dispersion seen for higher spread values, we are motivated to consider a transformation of the dependent variable, in this case a log transform. We ensure positivity of yields (screening for issues in data transmission from the pricing vendor), and then carry out the same procedure as above with respect to log spreads. This time, the forward selection procedure picks out every available regressor. Figure 4.3.4 appears more tightly clustered around some linear relationship.

Backward Elimination

In the backwards elimination procedure, a confidence level α is fixed and the initial design matrix consists of the full model with all P regressors. Denote this initial design matrix by X_0. The partial F-statistic for each regressor is then calculated, giving $\{F_j\}$, for $j = 1, \ldots, P$. The minimum, F_-, of these F-statistics is compared to $F_{1,N-P;1-\alpha}$. If $F_- > F_{1,N-P;1-\alpha}$, the procedure terminates. Otherwise, we cannot reject that null hypothesis that the regressor associated with F_- should be omitted. The design matrix X_0 is updated to X_1 with this factor omitted. The procedure terminates with design matrix X_k if the associated F_- satisfies $F_- > F_{1,N-(P-k);1-\alpha}$.

Stepwise Regression

The two previous procedures may be utilized in tandem. For example, a forward selection procedure may be followed by backward elimination.

4.4 Generalized Least Squares

Throughout we have assumed that in

$$Y = X\beta + \epsilon$$

we had $Cov(\epsilon) = \sigma^2 I$. Here, we generalize to the case where $Cov(\epsilon) = V$ for some matrix $V \in \mathbb{R}^{N \times N}$. Notice that this is equivalent – since all randomness in our model is derived from ϵ – to $Cov(Y) = V$.

The general case is handled directly with the tools already in place. Throughout, we assume $V \succ 0$, and that the eigenvalues of V are unique. This implies that there exists an orthonormal change of basis matrix Q satisfying

$$QVQ' = \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix}$$

for diagonal matrix Λ. In what follows, we are guided by the goal of linearly transforming $Y = X\beta + \epsilon$ via $AY = AX\beta + A\epsilon$ so that

$$\begin{aligned} Cov(AY) &= Cov(A\epsilon) \\ &= ACov(\epsilon)A' \\ &= AVA' \end{aligned}$$

is a diagonal matrix. We will leverage the decomposition above. In essence, we seek a square root of V which we can invert.

For the matrix Λ above, define $\Lambda^{\frac{1}{2}}$ as

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_N} \end{pmatrix}.$$

Clearly in this case we have that the inverse of Λ^{frac12}, which we denote by $\Lambda^{-\frac{1}{2}}$. is given by

$$\Lambda^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\lambda_N}} \end{pmatrix}.$$

Define P by

$$P = Q\Lambda^{\frac{1}{2}}Q' \tag{4.43}$$

for the same orthonormal change of basis matrix, Q, used for V. We have that

$$\begin{aligned} P^2 &= Q\Lambda^{\frac{1}{2}}Q'Q\Lambda^{\frac{1}{2}}Q' \\ &= Q\Lambda Q' \\ &= V. \end{aligned}$$

In addition, notice that $P = P'$. We may also show that P is positive definite. Lastly,

$$P^{-1} = Q\Lambda^{-\frac{1}{2}}Q'.$$

Looking back to our original regression equation, we have, then, that

$$P^{-1}Y = P^{-1}X\beta + P^{-1}\epsilon.$$

Defining

$$Y^* = P^{-1}Y \tag{4.44}$$
$$X^* = P^{-1}X \tag{4.45}$$
$$\epsilon^* = P^{-1}\epsilon, \tag{4.46}$$

the modified regression equation

$$Y^* = X^*\beta + \epsilon^* \tag{4.47}$$

satisfies

$$\begin{aligned} Cov(Y^*) &= Cov(\epsilon^*) \\ &= P^{-1}Cov(\epsilon)P^{-1} \\ &= P^{-1}VP^{-1} \\ &= P^{-1}(P^2)P^{-1} \\ &= I. \end{aligned}$$

We have then, that if $\mathbb{E}(\epsilon) = 0$, (4.47) satisfies the Gauss-Markov assumptions (with $\sigma^2 = 1$). Further, if the distributional assumption, $\epsilon \sim N(0, V)$ obtains, then $\epsilon^* \sim N(0, I)$.

The projection matrix for (4.47) is

$$H^* = P^{-1}X\left(X'V^{-1}X\right)^{-1}X'P^{-1}. \tag{4.48}$$

The proof is left to the reader. As before, we have

$$H^*Y^* = X^*\hat{\beta}$$

with

$$\begin{aligned} \hat{\beta} &= \left(X^{*'}X\right)^{-1}X^{*'}Y^* \\ &= \left(X'V^{-1}X\right)^{-1}X'V^{-1}Y. \end{aligned} \tag{4.49}$$

Under the distributional assumption, we also have

$$\begin{aligned} \hat{\beta} &\sim N\left(\beta, \left(X^{*'}X^*\right)^{-1}\right) \\ &\sim N\left(\beta, \left(X'V^{-1}X\right)^{-1}\right). \end{aligned} \tag{4.50}$$

4.5 Collinearity and the Condition Number

Throughout this chapter we have required that the design matrix, X, be full rank, resulting in a nonsingular, $X'X$. We have also seen how the uncertainty in OLS parameter estimates is affected by this design matrix, specifically as the inverse, $(X'X)^{-1}$ is critical to defining the variance of these estimates. In this section we examine some features of the OLS estimate through the lens of linear algebra. In particular, we ask what impact small changes in Y might have in the estimate, $\hat{\beta}$.

Our approach will be via the matrix norm defined in (3.26) and a related measure called the *condition number* of a matrix. Consider the linear system of equations

$$Ax = b,$$

and assume for the moment that A is invertible.

For nonsingular matrices, we define the condition number, here fixing the matrix norm as $||\cdot||_2$ as in (3.26), but noting that any norm may be used, as

$$\kappa(A) = ||A||_2||A^{-1}||_2. \tag{4.51}$$

Looking back at the system of equations, we have, for some perturbation in b, Δb, that there is some new solution $x + \Delta x$; viz.,

$$A(x + \Delta x) = b + \Delta b.$$

We would like to understand how relative changes in b, $\frac{||\Delta b||}{||b||}$, translate to relative changes in x, $\frac{||\Delta x||}{||x||}$. Looking at the original system, we have

$$||Ax|| = ||b|| \leq ||A||_2||x||,$$

and, similarly, the solution to the perturbed system must satisfy $A\Delta x = \Delta b$, so that

$$||\Delta x|| = ||A^{-1}\Delta b|| \leq ||A^{-1}||_2||\Delta b||.$$

Putting these two inequalities together, we have

$$\begin{aligned} \frac{||\Delta x||}{||x||} &\leq ||A||_2||A^{-1}||_2\frac{||\Delta b||}{||b||} \\ &= \kappa(A)\frac{||\Delta b||}{||b||}. \end{aligned} \tag{4.52}$$

Interpreting this result, we see that relative changes in b scaled by the *condition number* of A provide an upper bound for the resulting relative change in x. This bound may be made tighter, and in doing so, one may produce examples where the inequality gives an equality; i.e., there is no lower bound. This is slightly beyond the scope of what we provide here, however.

Even so, (4.52) remains insightful. For, all things being equal, we might prefer a smaller condition number than a greater one. It is somewhat direct to

show that in the case of a nonsingular, A, we have $\kappa(A) \in [1, \infty)$, with $\kappa(A) = 1$ indicating that in some basis, A is a scalar multiple of the identity.

We leave as an exercise to show that for nonsingular matrices,

$$||A^{-1}||_2 = \frac{1}{\min_{||v||=1} ||Av||},$$

thus implying that if a matrix is 'close to singular', then the condition number tends to infinity. This may be understood directly by resorting to an eigenvalue interpretation, where singularity is identified with zero valued eigenvalues. In our present work, we have then, that near-multicollinearity in the columns of A results in a large condition number, and solutions very sensitive to inputs.

In the case of ordinary least squares, the familiar setup

$$X\hat{\beta} = \hat{Y}$$

implies that

$$||\hat{Y}|| \leq ||X||_2 ||\hat{\beta}||,$$

where as usual, $\hat{\beta} = (X'X)^{-1}X'Y$. In the case that $\hat{Y}$ is perturbed by $\Delta\hat{Y}$, we have as above that

$$||\Delta\hat{\beta}|| \leq ||X^{\dagger}||_2 ||\Delta\hat{Y}||,$$

where $X^{\dagger} = (X'X)^{-1}X'$. The matrix, $X^{\dagger}$, in the present case is called the *Moore-Penrose pseudoinverse*. The above gives that

$$\begin{aligned} \frac{||\Delta\hat{\beta}||}{||\hat{\beta}||} &\leq ||X||_2 ||X^{\dagger}||_2 \frac{||\Delta\hat{Y}||}{||\hat{Y}||} \\ &= \kappa(X) \frac{||\Delta\hat{Y}||}{||\hat{Y}||}, \end{aligned} \tag{4.53}$$

where we have defined the condition number, $\kappa(X)$, for non-square, but full rank matrices via $X^{\dagger}$. Notice that these definitions coincide for invertible (square) matrices, and that if we allow for non-singularity and rank-deficient matrices to have an infinite condition number, for matrices generally.

The condition number of the design matrix X, then, is a direct measure of the sensitivity of the parameter estimate, $\hat{\beta}$. We say that a matrix is *ill-conditioned* if its condition number is very large, and *well conditioned* otherwise. In the context of collinearity and ordinary least squares, then, we see that variable selection is critical, and a preference for orthogonality over similarity generally obtains. Conversely, and in practice, identifying that a particular design matrix is ill-conditioned signals to the practitioner that their choice of predictors is in need of culling as one or more inputs are near some vector in the span of the remaining variables.

Exercises

1. Show that for any sequences of random variables $\{u_t\}$ and $\{v_t\}$,

$$\sum_{t=1}^{T} u_t v_t = (T-1)\hat{\sigma}_{uv} + T\bar{u}\bar{v},$$

and

$$\sum_{t=1}^{T} u_t^2 = (T-1)\hat{\sigma}_u^2 + T\bar{u}^2.$$

2. Prove that the least squares estimate of β for the model

$$r_t - r_f = \alpha + \beta(m_t - r_f) + \epsilon_t$$

is exactly

$$\beta = \frac{\hat{\sigma}_{rm}}{\hat{\sigma}_m^2}.$$

3. Using the model in the preceding problem, show that $\hat{\sigma}_{m\epsilon} = 0$.

4. Let $\{X_i\}_{i=1}^N$ be univariate normal random variables, and assume that the X_i's are iid. Given observations $\{x_i\}_{i=1}^N$, show that the maximum likelihood estimator of variance is given by

$$\tilde{s}_N^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

What is $\mathbb{E}(\tilde{s}_N^2)$?

5. For $i = 1, 2$, let $\{r_{i,t}\}_{t=1}^{N_i}$ be iid with $Var(r_i) = \sigma^2$. Let

$$\hat{M}_i = \frac{1}{N_i}\sum_{t=1}^{N_i} r_{i,t}$$

be an estimator of $\mu_i = \mathbb{E}(r_i)$ for $i = 1, 2$, and assume $\hat{M}_1$ and $\hat{M}_2$ are independent.

   (a) What is $\sigma_{12}^2 = Var(\hat{M}_1 - \hat{M}_2)$?

   (b) Assume further that $r_i \sim N(\mu_i, \sigma^2)$, and let s_i^2 be the unbiased estimator of variance shown previously. Let

$$s^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}.$$

      i. What is $\mathbb{E}(s^2)$?

      ii. How is s^2 distributed?

iii. Show

$$\frac{\frac{(\hat{M}_1-\mu_1)-(\hat{M}_2-\mu_2)}{\sigma_{12}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{(\hat{M}_1-\mu_1)-(\hat{M}_2-\mu_2)}{s\sqrt{\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}.$$

iv. How is this last ratio distributed?

v. Construct a 95% confidence interval for $\mu_1 - \mu_2$.

vi. Write down the confidence interval you would use to check the null hypothesis at confidence level $\alpha = 5\%$ that the means are the same; i.e.,

$$H_0 : \mu_1 = \mu_2.$$

6. Suppose $\hat{\beta}_1$ and $\hat{\beta}_2$ each minimize

$$\sum_{t=1}^{T} (y_{i,t} - \beta_i x_t)^2$$

for $i = 1$ and 2, respectively, where $y_\cdot \in \mathbb{R}$ and $x_\cdot \in \mathbb{R}$

$$f(\beta_1, \beta_2) = \sum_{t=1}^{T} (w_1 y_{1,t} + w_2 y_{2,t} - w_1\beta_1 x_t - w_2\beta_2 x_t)^2$$

Explain why the CAPM β of two portfolios is the weighted sum of the β's.

7. Prove that (4.18):

$$\hat{\epsilon} = (I - H)\epsilon.$$

8. Let H and G be projections. Denote the space spanned by the columns of a matrix, A, by $span(A)$, and let $\mathcal{L} = span(H)$ and $\mathcal{L}_0 = span(G)$. If $\mathcal{L}_0 \subseteq \mathcal{L}$, show that $H - G$ is a projection.

9. Prove (4.39) under the extended Gauss-Markov assumptions.

10. Prove (4.42).

11. Verify the results regarding the cross-sectional performance of β. Expand the findings by performing a double sort with β and then correlation. In each case, report annualized returns, annualized volatility, and the ratio of the two for each quinitle.

12. Using the same data as in the long-short volatility portfolio analysis, produce a portfolio which is β neutral at construction using even weight decile portfolios sorted on *ex ante* CAPM β (found using historical returns and including an intercept term, but ignoring the risk free rate). Report the average monthly return to your portfolio, its α and β, and test the null hypothesis that $\alpha = 0$ at the 0.05 confidence level.

13. Prove directly that P in (4.43) is positive definite by showing that for a given vector $w \not\equiv 0$, $w'Pw > 0$.

14. The jackknifed residuals shown in (4.19) used the population variance of residuals, σ. We define the internally studentized residuals as

$$\tilde{\epsilon}_i = \frac{\hat{\epsilon}_i}{s\sqrt{1-h_{ii}}} \tag{4.54}$$

with s the unbiased estimator of σ given by (4.23). Using the results of the chapter, how is $\tilde{\epsilon}_i$ distributed? How might you use these results to detect and remove outliers from a regression?

15. Verify (4.34) following the proof of (4.29).

16. Explain qualitatively the result of (4.35) and outline potential uses.

17. Prove (4.48).

18. Show that for a nonsingular matrix, A,

$$||A^{-1}||_2 = \frac{1}{\min_{||v||=1} ||Av||}.$$

19. Prove that the $||\cdot||_2$ is invariant under a change of basis. Conclude that the condition number of a sqaure matrix with real eigenvalues is $\frac{\max_i |\lambda_i|}{\min_i |\lambda_i|}$, where i ranges over all eigenvalues of the matrix, and the convention of infinite condition number for nonsingular matrices is used.

20. Use a backward elimination procedure for the spread data discussed in this chapter and the four factors provided. Report the variables chosen at each stage along with associated test statistics and p values. Repeat the process for log spreads.

21. Use a forward selection followed by backward elimination procedure for the spread data discussed in this chapter and the four factors provided. Report the variables chosen at each stage along with associated test statistics and p values. Repeat the process for log spreads.

22. The file `fama_french.txt` contains monthly performance numbers for various portfolios: the market excess return (MARKET), a long-short portfolio on size (SMB), and a long-short value portfolio (HML). For each portfolio:

    (a) Fit the regression model

$$r_t = \mu_1 \mathbf{1}_{\{t<\tau\}} + \mu_2 \mathbf{1}_{\{t\geq\tau\}} + \epsilon_t$$

    with τ = 12/31/1969 (196912).

    (b) Test the null hypothesis that $\mu_1 = \mu_2$ at the 95% confidence level.

Chapter 5

Math Preliminaries

In this chapter, we review some of the basic mathematical results needed in subsequent chapters. With an eye towards establishing the machinery of Taylor series, we present definitions and results for the gradient, hessian, and Jacobian of multivariate functions. Next we look at the growth rate categorizations established via big O and little o notation. This culminates in establishing several formulations of Taylor's theorem.

The following section deals with convex functions, of primary importance in our work in optimization. By using various results from Taylor's theorem, we prove three equivalencies depending on continuity assumptions.

5.1 Multivariate Calculus

In what follows, we will classify functions by their continuity properties. We say $f \in \mathbb{C}^N$ if f is N-times differentiable with a continuous Nth derivative. In this vein, continuous functions will be denoted by the class $\mathbb{C}^0$. Notice that if $f \in \mathbb{C}^N$, then $f \in \mathbb{C}^{N-k}$ for $k = 0, \ldots, N$.

5.1.1 The Gradient and Hessian

The gradient is the direct analogue of the first derivative in univariate calculus. For a function

$$f : \mathbb{R}^N \to \mathbb{R},$$

the *gradient* of f is denoted by ∇f and defined by

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_N} \end{pmatrix}. \tag{5.1}$$

where

$$\left.\frac{\partial f}{\partial x_i}\right|_x = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h} \tag{5.2}$$

and $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N$ is the vector with a 1 in the ith component and zeros everywhere else.

As a result, the gradient is a function

$$\nabla f : \mathbb{R}^N \to \mathbb{R}^N.$$

We define a *stationary point* of f as any point x^* such that $\nabla f(x^*) = 0$, and will often use notation $\nabla f^* = \nabla f(x^*)$ when the context is clear.

Certain useful properties of the gradient include that the gradient points in the direction of greatest increase for f (and as a corollary, $-\nabla f$ points in the direction of greatest decrease), and that $\nabla f(x)$ is tangent to the level curve of f at x.

We prove this first statement below.

Proof. To show that the gradient points in the direction of greatest increase, we must first define the *directional derivative*. For a unit vector u, the directional derivative at a point x is given by

$$D_u f(x) = (\nabla f(x), u). \tag{5.3}$$

The directional derivative measures the rate of change in the direction u. Notice that by Cauchy-Schwarz,

$$D_u f(x) = cos\theta ||\nabla f(x)|| \cdot ||u||$$

where $cos\theta$ is defined as the angle between $\nabla f(x)$ and u. As an immediate result, to maximize $D_u f(x)$, then, we require that $cos\theta = 1$, or that the angle between $\nabla f(x)$ and u must be 0. That is, the direction of greatest increase points in the direction of the gradient.

Similarly, to find the direction which decreases f the most, we must choose θ so that $cos\theta = -1$, giving that the direction of greatest decrease as $-\nabla f(x)$. □

We define a *descent direction* for the function f at a point x as any direction u satisfying $D_u f(x) < 0$. That is, a descent direction satisfies

$$u' \nabla f(x) < 0. \tag{5.4}$$

From the above, it is clear that $-\nabla f(x)$ is a descent direction. Further, we have that for any $H \succ 0$, then $-H\nabla f(x)$ is a descent direction as well.

In the sequel, we will be concerned with finding minima of functions. Based on the preceding result, the astute reader may see an immediate application of the gradient.

The *hessian* of f is the direct analogue of the second derivative in univariate calculus, and is denoted by $\nabla^2 f$. It is defined by the matrix of mixed partials

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_N \partial x_1} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_N} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{pmatrix}. \tag{5.5}$$

If $f \in \mathbb{C}^2$, then $\nabla^2 f$ is symmetric. Notice that $\nabla^2 f(x)$ is a function

$$\nabla^2 f : \mathbb{R}^N \to \mathbb{R}^{N \times N}.$$

For univariate functions, the second derivative gives an indication of whether the function is 'concave up' or 'concave down,' according to whether the second derivative is positive or negative, respectively. While we cite these terms out of familiarity, the preferred technical terms will be shown to be 'convex' and 'concave.' In any event, the hessian provides similar color on the shape of the function f, with a positive definite hessian indicating that the function is locally convex. We will formally define these terms and derive this result in a subsequent section.

5.1.2 The Jacobian

For a function, F,

$$F : \mathbb{R}^N \to \mathbb{R}^M$$

given by

$$F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{pmatrix}$$

with $f_i(x) : \mathbb{R}^N \to \mathbb{R}$ for each $i = 1, \ldots, N$, the *Jacobian* of F is denoted ∇F, and defined by

$$\nabla F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \cdots & \frac{\partial f_M}{\partial x_N} \end{pmatrix}. \tag{5.6}$$

The Jacobian of F may be written in terms of the gradients of f_i:

$$\nabla F = \begin{pmatrix} - & \nabla f_1' & - \\ & \vdots & \\ - & \nabla f_M' & - \end{pmatrix}. \tag{5.7}$$

Notice that the Jacobian is a function

$$\nabla F : \mathbb{R}^N \to \mathbb{R}^{M \times N},$$

and that for $f : \mathbb{R}^N \to \mathbb{R}$

$$\nabla(\nabla f) = \nabla^2 f, \tag{5.8}$$

or 'the Jacobian of the gradient is the hessian.'

The preceding observation highlights the intersection of notation. The apparent misuse is perhaps excusable when one considers that both the gradient and the Jacobian are the best linear approximations of their respective functions. For example, the Jacobian of

$$F(x) = Ax$$

for a constant matrix $A \in \mathbb{R}^{M\times N}$ is $\nabla F = A$. We derive this formally for familiarity with the concepts involved in the next proof.

Proof. For matrix

$$A = \begin{pmatrix} a_{11} & \dots & a_{1N} \\ & \ddots & \\ a_{M1} & \dots & a_{MN} \end{pmatrix},$$

we have

$$Ax = \begin{pmatrix} \sum_{i=1}^{N} a_{1i}x_i \\ \vdots \\ \sum_{i=1}^{N} a_{Mi}x_i \end{pmatrix}.$$

Define $A_j(x) = \sum_{i=1}^{N} a_{ji}x_i$ for $j = 1, \dots, M$. The gradient of each A_j is given by

$$\nabla A_j = \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jN} \end{pmatrix},$$

so that the Jacobian of F is

$$\nabla F = \begin{pmatrix} - & \nabla A_1' & - \\ & \vdots & \\ - & \nabla A_M' & - \end{pmatrix}.$$

Observing, finally, that ∇A_j is exactly the transpose of the jth row of A, ∇F is exactly A.

□

The product rule of univariate calculus has a direct analogue as well. Let F and G both be functions from $\mathbb{R}^N \to \mathbb{R}^M$,

$$F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_M(x) \end{pmatrix} \qquad G(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_M(x) \end{pmatrix}.$$

Then we have

$$\nabla(F'G) = \nabla F'G + \nabla G'F. \tag{5.9}$$

Note that on the left hand side we see the gradient operator and on the right the Jacobian.

Proof. We have that

$$\begin{aligned}\nabla(F'G) &= \nabla\left(\sum_{i=1}^{N} f_i g_i\right) \\ &= \begin{pmatrix} \sum_{i=1}^{M} \frac{\partial f_i}{\partial x_1} g_i \\ \vdots \\ \sum_{i=1}^{M} \frac{\partial f_i}{\partial x_N} g_i \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{M} \frac{\partial g_i}{\partial x_1} f_i \\ \vdots \\ \sum_{i=1}^{M} \frac{\partial g_i}{\partial x_N} f_i \end{pmatrix}.\end{aligned}$$

Observing from (5.6) that

$$\nabla F' = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \cdots & \frac{\partial f_M}{\partial x_1} \\ \vdots & & & \vdots \\ \frac{\partial f_1}{\partial x_N} & \cdots & \cdots & \frac{\partial f_M}{\partial x_N} \end{pmatrix}.$$

we see that

$$\nabla F'G = \begin{pmatrix} \sum_{i=1}^{M} \frac{\partial f_i}{\partial x_1} g_i \\ \vdots \\ \sum_{i=1}^{M} \frac{\partial f_i}{\partial x_N} g_i \end{pmatrix},$$

and similarly $\nabla G'F$ is equal to the second summand above, proving the result. □

We say that $q : \mathbb{R}^N \to \mathbb{R}$ is quadratic if q may be written as

$$q(x) = x'Gx + g'x + c \tag{5.10}$$

for constants G, g, and c. The gradient of q requires us to know $\nabla(x'Gx)$. This follows immediately from the above results, since

$$\begin{aligned}\nabla(x'Gx) &= \nabla(x)'Gx + \nabla(Gx)'x \\ &= Gx + G'x \\ &= (G + G')x.\end{aligned}$$

This gives that

$$\nabla q = (G + G')x + g. \tag{5.11}$$

In the case that G is symmetric, this reduces to the familiar $\nabla q = 2Gx + g$. The hessian of q is the Jacobian of (5.11), which, since ∇q is linear, is

$$\nabla^2 q = G + G'. \tag{5.12}$$

5.1.3 Big O and Little o

Big O and little o notation are used to categorize real valued functions with respect to their growth rates – everywhere, at a single point, or in the limit as

their dependent variable tends to infinity. We consider real valued functions, f and g,

$$f : \mathbb{R}^N \to \mathbb{R}$$
$$g : \mathbb{R}^N \to \mathbb{R}.$$

We say that f *is big O of* g as x approaches a if for a given δ there exists an M such that

$$||f(x)|| \leq M||g(x)|| \tag{5.13}$$

for all x satisfying $||x - a|| \leq \delta$.

If the limit $\lim_{x\to a} f(x)$ exists, and

$$\lim_{x\to a} \frac{||f(x)||}{||g(x)||} = c < \infty$$

then f is big O of g. The notation is often abused, stating $f = O(g)$ to indicate f is big O of g. Of course, since $O(g)$ is a set, it may be more appropriate to write $f \in O(g)$, but the previous notation is ubiquitous.

Example 5.1.1. It is easy to show that

- $\pi = O(1)$ everywhere
- $10x^2 + 3x + 17 = O(x^2)$ everywhere
- $\sin x = O(1)$ as $x \to 0$
- $\sin x = O(x)$ as $x \to 0$
- $\sin x \neq O(x^2)$ as $x \to 0$

Several useful properties obtain as well, including:

1. $O(f) \cdot O(g) = O(f \cdot g)$
2. $O(f) + O(g) = O(|f| + |g|)$
3. $f + O(g) = O(|f| + |g|)$
4. $O(c \cdot f) = O(f)$ for $c \in \mathbb{R}$, $c \neq 0$
5. if $\lim_{x\to a} |g(x)| = C_g < \infty$, then $O(f \cdot g) = O(f)$

To give a flavor of how to work with the notation, we prove the first property, leaving several others to the reader in the exercises.

Proof. Let f_0 be $O(f)$, g_0 be $O(g)$, and h_0 be $O(f \cdot g)$. Then we know that there exist constants, M_f, M_g, and M_h such that

$$\begin{aligned} ||f_0|| &\leq M_f||f|| \\ ||g_0|| &\leq M_g||g|| \\ ||h_0|| &\leq M_h||f \cdot g||. \end{aligned}$$

To prove that $O(f) \cdot O(g) = O(f \cdot g)$, we need to show that there exists an M_h^* such that $||h_0|| \leq M_h^* ||f|| \cdot ||g||$ and an M_{fg}^* such that $||f_0|| \cdot ||g_0|| \leq M_{fg}^* ||f \cdot g||$. Each claim is straightforward. We have

$$\begin{aligned} ||h_0|| &\leq M_h ||f \cdot g|| \\ &= M_h ||f|| \cdot ||g||, \end{aligned}$$

so we may set $M_h^* = M_h$. On the other hand,

$$\begin{aligned} ||f_0 \cdot g_0|| &= ||f_0|| \cdot ||g_0|| \\ &\leq M_f ||f|| \cdot M_g ||g|| \\ &= M_f M_g ||f \cdot g||, \end{aligned}$$

and we may set $M_{fg}^* = M_f M_g$. □

We say that f *is little o of* g as x approaches a if

$$\lim_{x \to a} \frac{||f(x)||}{||g(x)||} = 0. \tag{5.14}$$

And, again, the notation will be stretched so that we may write $f = o(g)$ to indicate f is little o of g. Clearly, little o notation is a stronger statement about the relative growth rates of f and g than big O. As such, if $f = o(g)$, then $f = O(g)$. We also have

1. $c \cdot o(f) = o(f)$ for $c \in \mathbb{R}$
2. $o(f) \cdot O(g) = o(f \cdot g)$

These properties are left as exercises for the reader.

It is important to note that if $f = o(g)$ at a, then based on the definition, for any $\epsilon > 0$, there exists a δ such that

$$||f(x)|| \leq \epsilon ||g(x)||$$

for all $||x - a|| \leq \delta$.

5.1.4 Taylor Series

The tool we will use to study optimization of smooth functions is Taylor's Theorem. At times we will have to use various formulations, each a result of continuity assumptions of $f : \mathbb{R}^N \to \mathbb{R}$. Recall the first and second order Taylor approximations of univariate functions,

$$f(x + \delta) \approx f(x) + \delta f'(x) \tag{5.15}$$

$$f(x + \delta) \approx f(x) + \delta f'(x) + \frac{1}{2} \delta^2 f''(x). \tag{5.16}$$

In the multivariate case, $f : \mathbb{R}^N \to \mathbb{R}$, we have:

1. For f differentiable,

$$f(x) = f(x_0) + \nabla f(x_0)'(x - x_0) + o(||x - x_o||). \tag{5.17}$$

Notice that we may write the above as

$$f(x + \delta) = f(x) + \nabla f(x)'\delta + o(||\delta||).$$

An alternate formulation with this same continuity assumption is

$$f(x + \delta) = f(x) + \nabla f(x + t\delta)'\delta \tag{5.18}$$

for some $t \in (0, 1)$.

This second formulation in t will be extremely useful in many of the proofs that follow.

2. For f twice differentiable

$$f(x) = f(x_0)+\nabla f(x_0)'(x-x_0)+\frac{1}{2}(x-x_0)'\nabla^2 f(x_0)(x-x_0)+o(||x-x_o||^2). \tag{5.19}$$

And again, we may rewrite this in terms of a perturbation, δ, as

$$f(x + \delta) = f(x) + \nabla f(x)'\delta + \frac{1}{2}\delta'\nabla^2 f(x)\delta + o(||\delta||^2),$$

and have a formulation with strict equality as

$$f(x + \delta) = f(x) + \nabla f(x)'\delta + \delta'\nabla^2 f(x + t\delta)'\delta \tag{5.20}$$

for some $t \in (0, 1)$.

3. For f twice differentiable with continuous second derivative; i.e., $f \in \mathbb{C}^2$,

$$f(x) = f(x_0) + \nabla f(x_0)'(x - x_0) + O(||x - x_o||^2), \tag{5.21}$$

or, as before

$$f(x + \delta) = f(x) + \nabla f(x)'\delta + O(||\delta||^2).$$

We also have the integral formulation

$$\nabla f(x + \delta) = \nabla f(x) + \int_0^1 \nabla^2 f(x + t\delta)\delta \, dt \tag{5.22}$$

For each of the above formulations in x_0, we identify them as *Taylor expansions of f about x_0*.

Finally, for $F : \mathbb{R}^N \to \mathbb{R}^M$ and F differentiable, we also have, for ∇F indicating the Jacobian,

$$F(x) = F(x_0) + \nabla F(x_0)(x - x_0) + o(||x - x_0||) \tag{5.23}$$

and

$$F(x+\delta) = F(x) + \nabla F(x+t\delta)\delta \tag{5.24}$$

for some $t \in (0,1)$.

This equation also gives an expansion of the gradient of f for f twice differentiable. Namely,

$$\nabla f(x) = \nabla f(x_0) + \nabla^2 f(x_0)(x-x_0) + o(||x-x_0||) \tag{5.25}$$

since the Jacobian of the gradient is the hessian.

5.2 Convex Functions

Convex functions occupy a particularly attractive position in optimization. As we shall see, many appealing properties such as uniqueness of global optima are readily obtained when the function under consideration is convex. In this section we establish several equivalent formulations of convexity, each with an accompanying continuity assumption. We begin with the definition for a *convex set*.

We say that a set, K, of points in $\mathbb{R}^N$ is convex if for any x_0 and x_1 in K, x_θ, defined as

$$x_\theta = (1-\theta)x_0 + \theta x_1, \tag{5.26}$$

lies in K as well, for $\theta \in [0,1]$. Put another way, a set is convex if every line segment between two points in K is completely contained in K as well. One may show that, more generally, if K is convex, then

$$\sum_{i=1}^{N} \theta_i x_i \in K \tag{5.27}$$

when each $x_i \in K$ and $\sum_i \theta_i = 1$ with $\theta_i > 0$ for $i = 1, \ldots, N$. This exercise is left to the reader. We say in the preceding case that $\sum_i \theta_i x_i$ is a *convex combination* of $\{x_i\}_{i=1}^N$.

A function, $f : \mathbb{R}^N \to R$ is convex if

$$f(x_\theta) \leq (1-\theta)f(x_0) + \theta f(x_1). \tag{5.28}$$

In the work below, we will further reduce this notation to simply

$$f_\theta \leq (1-\theta)f_0 + \theta f_1.$$

The definition simply states that the value of the function as x ranges from x_0 to x_1 lies below the line segment connecting x_0 and x_1.

We next prove the following equivalencies:

1. For f differentiable, f is convex if and only if

$$\nabla f_0'(x_1 - x_0) \leq f_1 - f_0. \tag{5.29}$$

Or, equivalently,

$$f_0 + \nabla f_0'(x_1 - x_0) \leq f_1$$

giving that the f always lies above its linear approximation.

2. For $f \in \mathbb{C}^2$, f is convex if and only if

$$\nabla^2 f \succeq 0. \tag{5.30}$$

In this case the function resembles (at least in low enough dimensions) an upward facing bowl.

We begin by proving (5.29).

Proof. Suppose f is differentiable and

$$\nabla f_0'(x_1 - x_0) \leq f_1 - f_0.$$

Then, using the notation in 5.26 and following 5.28, we have that for any $\theta \in (0, 1)$, both

$$\nabla f_\theta'(x_1 - x_\theta) \quad \leq \quad f_1 - f_\theta$$

and

$$\nabla f_\theta'(x_0 - x_\theta) \quad \leq \quad f_0 - f_\theta.$$

Multiplying this first inequality in f_θ by θ and the second by $1 - \theta$ gives

$$\begin{aligned} \theta \nabla f_\theta'(x_1 - x_\theta) &\leq \theta(f_1 - f_\theta) \\ \nabla f_\theta'(\theta x_1 - \theta x_\theta) &\leq \theta f_1 - \theta f_\theta \end{aligned}$$

and

$$\nabla f_\theta'((1-\theta)x_0 - (1-\theta)x_\theta) \quad \leq \quad (1-\theta)f_0 - (1-\theta)f_\theta.$$

Adding these two resulting inequalities gives

$$\begin{aligned} \nabla f_\theta(\theta x_1 + (1-\theta)x_0 - (\theta + 1 - \theta)x_\theta) &\leq \theta f_1 + (1-\theta)f_0 - (\theta + 1 - \theta)f_\theta \\ \nabla f_\theta(x_\theta - x_\theta) &\leq \theta f_1 + (1-\theta)f_0 - f_\theta \\ 0 &\leq \theta f_1 + (1-\theta)f_0 - f_\theta \end{aligned}$$

so that

$$f_\theta \leq (1-\theta)f_0 + \theta f_1$$

as desired.

To prove the converse, we assume that f is convex and establish (5.29). Since f is convex, we know that

$$f_\theta \leq (1-\theta)f_0 + \theta f_1,$$

and so

$$f_\theta - f_0 \leq \theta(f_1 - f_0),$$

giving

$$\frac{f_\theta - f_0}{\theta} \leq f_1 - f_0.$$

This final formulation gives an indication of how to introduce the gradient. Considering x_θ as a point on the line segment connecting fixed x_0 and x_1, we expand f in a Taylor Series as in (5.17) about x_0 as

$$f(x_0 + \theta(x_1 - x_0)) = f(x_0) + \nabla f(x_0)'(\theta(x_1 - x_0)) + o(\theta||x_1 - x_0||)$$

or

$$f_\theta = f_0 + \nabla f_0'(\theta(x_1 - x_0)) + o(\theta).$$

Rearranging terms, we have

$$\frac{f_\theta - f_0}{\theta} = \nabla f_0'(x_1 - x_0) + \frac{o(\theta)}{\theta},$$

and taking the limit as $\theta \downarrow 0$ (taking the limit from the right since $\theta \in (0, 1)$, we see

$$\begin{aligned} \lim_{\theta \downarrow 0} \frac{f_\theta - f_0}{\theta} &= \nabla f_0'(x_1 - x_0) + \lim_{\theta \downarrow 0} \frac{o(\theta)}{\theta} \\ &= \nabla f_0'(x_1 - x_0). \end{aligned}$$

We are left to determine a bound for $\lim_{\theta \downarrow 0} \frac{f_\theta - f_0}{\theta}$. But it is clear from our previous observation that

$$\lim_{\theta \downarrow 0} \frac{f_\theta - f_0}{\theta} \leq \lim_{\theta \downarrow 0} (f_1 - f_0) = f_1 - f_0,$$

so that

$$\nabla f_0'(x_1 - x_0) \leq f_1 - f_0$$

completing the proof. □

Next, we prove (5.30).

Proof. Suppose f is twice differentiable and that

$$\nabla^2 f(x) \succeq 0$$

for all x. Expanding f about x_0 as before and using 5.20, we have

$$f_1 = f_0 + \nabla f_0'(x_1 - x_0) + \frac{1}{2}(x_1 - x_0)'\nabla^2 f(x_0 + t(x_1 - x_0))(x_1 - x_0)$$

for some $t \in (0, 1)$. By the positive semidefiniteness of the hessian, the second summand is nonnegative, and hence

$$\begin{aligned} f_1 &\geq f_0 + \nabla f_0'(x_1 - x_0) \\ f_1 - f_0 &\geq \nabla f_0'(x_1 - x_0) \end{aligned}$$

giving that f is convex by (5.29).

We next assume that f is convex and twice differentiable. Let $x_1 = x_0 + \alpha s$ for $s \neq 0$ and α a positive scalar, and expand the gradient of f about x_0 as in (5.25),

$$\nabla f_1 = \nabla f_0 + \nabla^2 f_0(\alpha s) + o(\alpha). \tag{5.31}$$

From (5.29), we know that both

$$\begin{aligned} \nabla f_0'(x_1 - x_0) &\leq f_1 - f_0 \\ \nabla f_1'(x_0 - x_1) &\leq f_0 - f_1 \end{aligned}$$

giving

$$\nabla f_0'(x_1 - x_0) \leq f_1 - f_0 \leq \nabla f_1'(x_1 - x_0).$$

Replacing x_1 by $x_0 + \alpha s$, we see that

$$\nabla f_0'(\alpha s) \leq f_1 - f_0 \leq \nabla f_1'(\alpha s). \tag{5.32}$$

Now, premultiplying (5.31) by αs, we see that

$$(\alpha s)' \nabla f_1 = (\alpha s)' \nabla f_0 + (\alpha s)' \nabla^2 f_0(\alpha s) + o(\alpha^2).$$

Combining this result with the preceding set of inequalities, we get

$$\begin{aligned} \nabla f_0'(\alpha s) \leq f_1 - f_0 &\leq (\alpha s)' \nabla f_0 + (\alpha s)' \nabla^2 f_0(\alpha s) + o(\alpha^2) \\ 0 \leq f_1 - f_0 &\leq \alpha^2 s' \nabla^2 f_0 s + o(\alpha^2). \end{aligned}$$

Dividing through by α^2 and taking the limit as $\alpha \downarrow 0$, we get

$$\begin{aligned} 0 &\leq s' \nabla^2 f_0 s + \lim_{\alpha \downarrow 0} \frac{o(\alpha^2)}{\alpha^2} \\ 0 &\leq s' \nabla^2 f_0 s, \end{aligned}$$

proving the result since x_0 and s were arbitrarily chosen. □

Exercises

1. Prove $O(f) + O(g) = O\left(|f| + |g|\right)$.
2. Prove that if f is $o(g)$, then f is $O(g)$.
3. Prove $o(f) \cdot O(g) = o(f \cdot g)$.
4. Ledoit and Wolf [19] consider various biased estimators of the mean and covariance. In the following problems we look at some of their preliminary results.

   (a) Let $X \in \mathbb{R}^N$ be a multivariate random variable with mean μ and covariance Σ. Let $\hat{\mu}$ be the unbiased estimator of the mean,

   $$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} \mu_t$$

   and $f \in \mathbb{R}^N$ a constant. A shrinkage estimator of the mean is given by

   $$(1-\alpha)\hat{\mu} + \alpha f$$

   for some α. To determine α, Ledoit and Wolf consider the expected loss function

   $$R(\alpha) = \mathbb{E}\left(||(1-\alpha)\hat{\mu} + \alpha f - \mu||^2\right).$$

   i. Show that $R(\alpha)$ is minimized when

   $$\alpha^* = \frac{\mathbb{E}\left(||\hat{\mu} - \mu||^2\right)}{\mathbb{E}\left(||\hat{\mu} - \mu||^2\right) + ||f - \mu||^2}$$

   ii. Show that

   $$\mathbb{E}\left(||\hat{\mu} - \mu||^2\right) = \frac{1}{T} tr(\Sigma).$$

   Use the expectation of quadratic forms rule: for Y a random vector, and A a matrix, $\mathbb{E}(Y'AY) = tr(A \cdot S) + m'Am$, where $\mathbb{E}(Y) = m$, $Cov(Y) = S$.

   iii. Show that, as a result,

   $$\alpha^* = \frac{(N/T)\bar{\sigma}^2}{(N/T)\bar{\sigma}^2 + ||f - \mu||^2}$$

   where $\bar{\sigma}^2 = \frac{1}{N} tr(\Sigma)$.

   (b) We have defined the condition number of a positive definite matrix, $A \in \mathbb{R}^{N \times N}$, as the ratio of the maximum and minimum eigenvalues of A. Consider

   $$\Sigma_s = (1-\alpha)\hat{\Sigma} + \alpha F.$$

Ledoit and Wolf [20] suggest another shrinkage method (i.e., a biased estimator of the covariance matrix, Σ) as a convex combination of the sample covariance and a matrix F of the form

$$F = \sigma^2 I.$$

i. Show that the condition number of Σ_s is

$$k(\alpha) = \kappa(\Sigma_s) = \frac{(1-\alpha)\bar{\lambda} + \alpha\sigma^2}{(1-\alpha)\underline{\lambda} + \alpha\sigma^2}$$

where $\bar{\lambda}$ and $\underline{\lambda}$ are the maximum and minimum eigenvalues of $\hat{\Sigma}$ respectively.

ii. Where is $k(\alpha)$ increasing or decreasing on $[0, 1]$? Where does it attain its maximum? Its minimum?

5. Find the gradient and Hessian function for $f(x, y) = 100(y-x^2)^2+(1-x)^2$. Show that for the the local minimizer $x^* = (1, 1)$, the gradient vanishes and the Hessian is positive definite.

6. Prove that if Σ is a covariance matrix, the function

$$f(x) = x'\Sigma x$$

is convex.

7. Prove that the intersection of finitely many convex sets is convex.

8. Prove (5.27).

Chapter 6

Unconstrained Optimization

Much of the work we have seen in regression may be framed as examples of unconstrained optimization. That is, the problems obtained solutions by minimizing some loss function; viz., the L^2 norm, or sum of squares. In that setting, we both identified solutions such as the $\hat{\beta}$ of an OLS regression as well as discussed such a solution's distributional properties. Here, we focus on the former task, even while revisiting examples in OLS and a related problem in index replication.

The work begins by outlining conditions for a point, x^*, to be a solution of a minimization problem, $\min_x f(x)$. We then identify useful necessary and sufficient conditions that x^* must obtain to be optimal when working with a smooth f [5]. We also establish some attractive properties that convex f have in the current context.

Finally, we discuss Newton's method, an algorithm useful in identifying stationary points. As we shall see, under some assumptions, we may prove convergence of the method to a stationary point when $f \in \mathbb{C}^2$. The algorithm provides a template for future work in constrained optimization as well.

6.1 Preliminaries

For a function

$$f : \mathbb{R}^N \to \mathbb{R},$$

we say that x^* is a *global minimizer* if

$$f(x*) \leq f(x) \tag{6.1}$$

for all $x \in \mathbb{R}^N$. We say that x^* is a *strict global minimizer* if the inequality above is strict for all $x \neq x^*$. Finally, x^* is a *local minimizer* if there exists a

neighborhood, $\mathcal{N}(x^*)$, about x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{N}(x)$. The strict qualifier may be applied here as well.

Our notation will often be abbreviated as

$$\begin{aligned} f^* &= f(x^*) \\ \nabla f^* &= \nabla f(x^*) \\ \nabla^2 f^* &= \nabla^2 f(x^*). \end{aligned}$$

In this section we study *unconstrained minimization problems*; i.e., problems of the form

$$\min_x f(x). \tag{6.2}$$

In this formulation, f is called the *objective function*, and the solution – should one exist – is a global minimizer of f.

6.1.1 Necessary and Sufficient Conditions

We will prove various necessary and sufficient conditions that obtain at ∇f^* and $\nabla^2 f^*$, referred to as first and second order conditions, respectively.

Recall that a *necessary condition* for x^* to be a minimizer is a condition that x^* must satisfy if it is to be a minimizer, while a *sufficient condition* for x^* to be a minimizer is a condition that implies x^* is a minimizer of f.

We proceed with proofs of a first order necessary condition, a second order necessary condition, and a second order sufficient condition.

First Order Necessary Condition Let x^* be a local minimizer of $f \in \mathbb{C}^1$ near x^*, then $\nabla f^* = 0$. That is, x^* is a stationary point.

Proof. Assume by contradiction that $\nabla f^* \neq 0$. Then we may find a δ satisfying

$$\delta' \nabla f^* < 0.$$

(For example, choose $\delta = -\nabla f^*$.) By continuity, there exists a τ such that

$$\delta' \nabla f(x^* + \hat{t}\delta) < 0.$$

for all $\hat{t} \in [0, \tau]$.

A Taylor expansion of f about x^* gives

$$f(x^* + \hat{t}\delta) = f^* + \hat{t}\delta' \nabla f(x^* + t \cdot \hat{t}\delta)$$

for some $t \in (0, 1)$. Now since $t \cdot \hat{t}$ remains in $[0, \tau]$, $\delta' \nabla f(x^* + t \cdot \hat{t}\delta) < 0$, and

$$f(x^* + \hat{t}\delta) < f^*,$$

contradicting the fact that x^* is a local minimizer of f. Hence $\nabla f^* = 0$. □

Second Order Necessary Condition Let x^* be a local minimizer of $f \in \mathbb{C}^2$ near x^*, then $\nabla^2 f^* \succeq 0$.

The proof proceeds just as before, utilizing the fact that we now know that x^* is a stationary point as well.

Proof. Assume by contradiction that $\nabla^2 f^*$ is not positive semidefinite. Then we may find a δ such that

$$\delta' \nabla^2 f^* \delta < 0.$$

Since f is in $\mathbb{C}^2$ near x^*, we have by continuity that there exists a τ such that

$$\delta' \nabla^2 f(x^* + \hat{t}\delta)\delta < 0$$

for all $\hat{t} \in [0, \tau]$.

Taking a second order expansion of f about x^* gives

$$f(x^* + \hat{t}\delta) = f^* + \hat{t}\delta' \nabla f^* + \frac{1}{2}\hat{t}^2 \delta' \nabla^2 f(x^* + t \cdot \hat{t}\delta)\delta$$

for some $t \in (0, 1)$. Since x^* is a stationary point, this reduces to

$$f(x^* + \hat{t}\delta) = f^* + \frac{1}{2}\hat{t}^2 \delta' \nabla^2 f(x^* + t \cdot \hat{t}\delta)\delta.$$

Finally, $t \cdot \hat{t} \in [0, \tau]$ so that $\delta' \nabla^2 f(x^* + t \cdot \hat{t}\delta)\delta < 0$, and

$$f(x^* + \hat{t}\delta) < f^*,$$

contradicting the fact that x^* is a local minimizer of f. Hence $\nabla^2 f^* \succeq 0$. □

Second Order Sufficient Condition Let x^* be a stationary point, and assume $\nabla^2 f^* \succ 0$. If $f \in \mathbb{C}^2$ in a neighborhood of x^*, then x^* is a local minimizer.

Proof. By continuity, we have that there exists a τ such that

$$\nabla^2 f(x^* + \hat{t}\delta) \succ 0$$

for all $\hat{t} \in [0, \tau]$. Expanding f about x^*, there exists a $t \in (0, 1)$ so that

$$\begin{aligned} f(x^* + \hat{t}\delta) &= f^* + \hat{t}\delta' \nabla f^* + \frac{1}{2}\hat{t}^2 \delta' \nabla^2 f(x^* + t \cdot \hat{t}\delta)\delta \\ f(x^* + \hat{t}\delta) &= f^* + \frac{1}{2}\hat{t}^2 \delta' \nabla^2 f(x^* + t \cdot \hat{t}\delta)\delta \end{aligned}$$

since x^* is a stationary point.

As before, $t \cdot \hat{t} \in [0, \tau]$, so that $\nabla^2 f(x^* + t \cdot \hat{t}\delta) \succ 0$, and hence

$$f(x^* + \hat{t}\delta) \geq f^*$$

as desired. □

Several examples of finding extrema are immediately available to us having identified the above necessary and sufficient conditions. For example, we may frame OLS and GLS in terms of an optimization framework. Further, we introduce a version of so-called index tracking.

Example 6.1.1. We have already considered the OLS objective function (e.g., (4.2), (4.13))

$$f(\beta) = ||Y - X\beta||^2$$

for $Y \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N \times p}$. We may rewrite f as

$$\begin{aligned} f(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= \beta' X' X\beta - 2\beta' X' Y + Y' Y \end{aligned}$$

showing explicitly that f is a quadratic function of β. Since $X'X$ is symmetric, we have

$$\nabla f(\beta) = 2X' X\beta - 2X' Y.$$

Setting the gradient equal to zero to find β^* gives

$$\beta^* = (X' X)^{-1} X' Y,$$

an answer seen several times in the preceding work in statistics. Now, since the hessian of f is

$$\nabla^2 f(\beta) = 2X' X$$

we have that β^* is a minimizer by virtue of the positive semidefiniteness of $X'X$ and global continuity of f.

Notice that in the OLS-as-optimziation presentation here, no discussion of the random component ϵ was considered. In fact, no statistical properties of the estimator β^* are discernible here whatsoever. The reduction to (i.e., justification for) a quadratic minimization problem required an argument about the distributional properties of ϵ. The fact that the OLS $\hat{\beta}$ was a projection is also nowhere to be seen in this casting.

Example 6.1.2. We may also consider the generalized least squares case as an optimization problem. In the GLS case, the objective function becomes

$$f(\beta) = ||Y - X\beta||^2_{V^{-1}}$$

where $V = Cov(\epsilon)$ and $||x||^2_A = (x, x)_A = x' Ax$ is defined as the inner product with respect to a positive definite matrix A. Here the expansion of f yields

$$\begin{aligned} f(\beta) &= (Y - X\beta)' V^{-1} (Y - X\beta) \\ &= \beta' X' V^{-1} X\beta - 2\beta' X' V^{-1} Y + Y' V^{-1} Y \end{aligned}$$

with gradient

$$\nabla f(\beta) = 2X' V^{-1} X\beta - 2X' V^{-1} Y.$$

which gives stationary point

$$\beta^* = \left(X'V^{-1}X\right)^{-1} X'V^{-1}Y$$

as in (4.49). Again, f is $\mathbb{C}^2$ everywhere with Hessian $X'V^{-1}X$ which is positive semidefinite when X is full rank and V is a covariance matrix so the above result is sufficient for optimality of f.

As before, the justification of this particular objective function is lacking in the discussion along with distributional properties of the estimators.

In our next example, we look at an index tracking model. That is, given a target security's returns through time, $\{s_t\}_{t=1}^T$, we identify optimal weights for a basket of tradeable securities, $\{r_{i,t}\}_{t=1}^T$, for $i = 1, \ldots, N$.

Example 6.1.3. We define

$$r_t = \begin{pmatrix} r_{1,t} \\ \vdots \\ r_{N,t} \end{pmatrix}.$$

We consider the objective function in weights, $w \in \mathbb{R}^N$,

$$f(w) = \sum_{t=1}^{T} \left(s_t - r_t' w\right)^2$$

which may be rewritten as in ordinary least squares as

$$f(w) = ||S - Rw||^2 \tag{6.3}$$

for

$$S = \begin{pmatrix} s_1 \\ \vdots \\ s_T \end{pmatrix}.$$

and $R_{ij} = r_{j,i}$; i.e., the jth row of R consists of the N secruties returns at time j.

We proceed as before, yielding gradient

$$\nabla f(w) = 2R'Rw - 2R'S$$

and optimal solution

$$w^* = (R'R)^{-1}R'S, \tag{6.4}$$

or the OLS solution in another guise.

We may apply the above to a particular set of data. Using the same cross-sectional data as in previous studies, we build index tracking portfolios from weekly return data of the 50 largest companies available each of the 200 months from 10/31/1997 to 5/31/2014. In particular, $T = 117$ weeks and $N = 50$ in

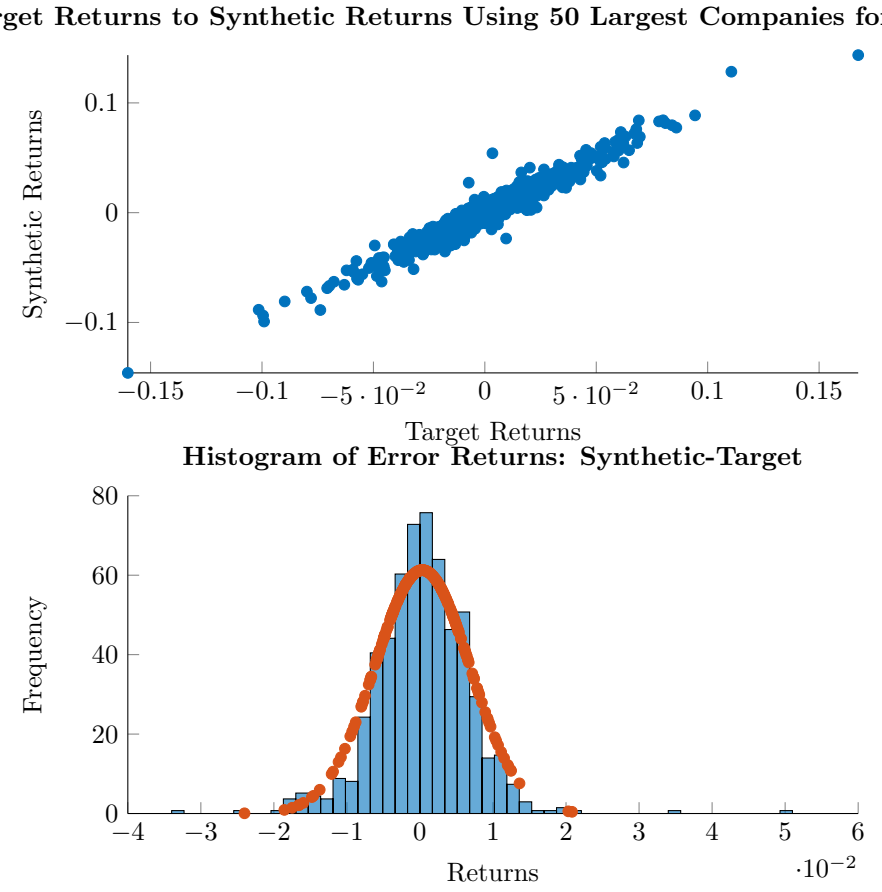

Figure 6.1: Summary statistics for the out of sample weekly return performance of an index tracking methodology obtained from an unconstrained quadratic optimization problem.

the above notation. We then use the optimal weights for each time period given by (6.4) to evaluate performance over the next four weeks.

Summary statistics are presented visually in Figure 6.1.1. The 95% confidence interval for the CAPM α of the synthetic asset's weekly returns is $[-0.0001, 0.0008]$, while the CAPM β confidence interval is $[0.9538, 0.9866]$, significantly different from 1.0. We cannot reject the null hypothesis that the synthetic return's α is different from zero – the rare case where this is not a disappointing statement to make.

Finally, across the roughly 800 weeks studied, the average tracking error was 3 bps (0.03%). This annualizes to 1.67% over nearly 17 years. These are appealing characteristics by and large. However, the procedure is not without its faults.

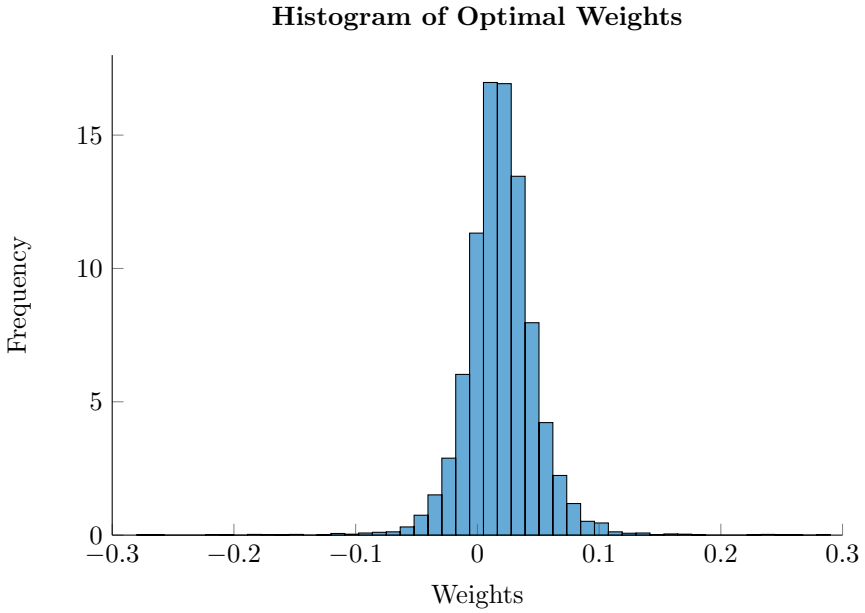

Figure 6.2: Histogram of all optimal weights found applying the index tracking procedure over 200 distinct months. Large positive and negative weights are apparent as well as a common occurrence of negative weights, generally.

In particular, the optimization is prone to large and negative values of w_i^*, as can be seen in Figure 6.1.3. Approximately 20% of all of the weights found are negative, and position sizing as large as $\pm 28\%$ was seen. For a market tracking portfolio, one may argue that the appearance of negative positions is spurious. In practice, the added difficulty (and cost) of shorting securities makes this more than just an academic observation. Additionally, the near symmetry for outsized positioning bodes poorly for the method, indicating potentially spurious results.

The observations noted here will be seen again when we give mean-variance optimization a formal treatment. Our observations are not novel by any means, however, and are well noted in the current literature. We will see that a parsimonious solution relies on understanding the eigenvalues of the covariance of the returns involved. As a preview: the model results here are sensitive to near-zero eigenvalues. Solutions abound, however. In particular, an immediate and obvious fix is to *constrain* the possible weight sizing available to the model. Another is to ameliorate the eigenvalue just identified. This solution leverages the field of random matrix theory and is a burgeoning area of interest in the field.

6.1.2 Convex Functions

Convex functions have particularly attractive properties with respect to their extrema. In particular, local minimizers are also global minimizers, and the first order necessary condition for optimality above is also sufficient. We prove each of these claims here.

Let x^* be a local minimizer of f, convex. Then x^* is a global minimizer of f.

Proof. Assume by contradiction that there exists a z such that $f(z) < f^*$. Then by convexity, for $x_\theta = \theta z + (1-\theta)x^*$,

$$f_\theta \leq \theta f(z) + (1-\theta)f^* < f^*.$$

Now, any neighborhood of x^* will contain x_θ for sufficiently small θ, though, contradicting the fact that x^* is a local minimizer. Therefore x^* must be a global minimizer. □

Next, suppose that f is convex and differentiable at x^*, then if x^* is a stationary point, x^* is a global minimizer of f.

Proof. We again proceed by contradiction, assuming there exists a z such that $f(z) < f^*$. By (5.29) we have

$$\nabla f^{*'}(z-x) \leq f(z) - f^*$$

implying since x^* is a stationary point that

$$f^* \leq f(z),$$

a contradiction. □

6.2 Newton's Method

Based on our work above, to find minima of $f : \mathbb{R}^N \to \mathbb{R}$, we are motivated to identify stationary points. Our prototype algorithm for doing so will be Newton's Method, an iterative procedure developed to find the roots of functions.

For a general $F : \mathbb{R}^N \to \mathbb{R}^N$, Newton's Method utilizes the linear approximation of F given by

$$F_l(x+\delta) = F(x) + \nabla F(x)\delta$$

and solves where this linear approximation, F_l is zero; viz.,

$$\delta = -(\nabla F(x))^{-1} F(x).$$

Immediately we see that we must have invertibility of the Jacobian for this particular implementation of the method to be well defined. Iterations continue, updating x with this particular δ.

Given that we are particularly interested in the roots of the gradient function, $\nabla f : \mathbb{R}^N \to \mathbb{R}^N$, we may rewrite the above in terms of the gradient and hessian as

$$\delta = -(\nabla^2 f(x))^{-1} \nabla f(x). \tag{6.5}$$

This δ is often referred to as the *Newton step* or *Newton direction*. We may also, for reasons that will become clear in later exposition, call this a *full* Newton step.

We use superscript notation to indicate iterates within an algorithm; e.g., x^k as the kth iterate, δ^k as the kth Newton step, and $\nabla f^k = \nabla f(x^k)$, and so on. The algorithm is outlined as follows, initializing with some x^0 and having small and large threshold parameters ϵ and K, respectively:

Algorithm 1 Newton's Method

Initialize x^0
while $||\nabla f^k|| > \epsilon$ and $k < K$ **do**
  $\delta^k = -(\nabla^2 f^k)^{-1} \nabla f^k$
  $x^{k+1} \leftarrow x^k + \delta^k$

Of course, Algorithm 1 presupposes some kind of stopping condition in the norm of the gradient is possible; viz., there is a tacit assumption that we will be able to find an iterate k such that $||\nabla f^k||$ is in fact smaller than some prescribed ϵ. And, of course, in this case we would expect some proximity to a stationary point. The maximum iteration parameter, K, is just realistic coding.

Our next theorem addresses the as yet aspirational view of finding a point x^k near a stationary point. We outline several assumptions first.

Theorem 6.2.1. For $f \in \mathbb{C}^2$ in a neighborhood of x^* and with hessian satisfying both $\nabla^2 f^* \succ 0$ and $||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq M||x_1 - x_2||$ for any x_1 and x_2 in that neighborhood, then for x^k sufficiently close to x^*, Newton's method converges at second order and is well defined at each iterate.

In the above, quadratic convergence of a sequence $\{x^k\}$ to x^* is defined by the satisfaction of

$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq C \tag{6.6}$$

for some constant, C. Of course, this implies that $||x^{k+1} - x^*|| = O\left(||x^k - x^*||^2\right)$.

Proof. Since

$$\nabla^2 f^* \succ 0$$

and $f \in \mathbb{C}^2$ in a neighborhood of x^*, there exists a neighborhood about x^* such that the hessian remains positive definite. In addition, since

$$||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq M||x_1 - x_2||$$

in a neighborhood of x^* the inverse of the hessian is bounded in that neighborhood.

These two neighborhoods may be taken jointly to identify a neighborhood, $\mathcal{N}(x^*)$, such that

$$\begin{aligned} \nabla^2 f(x) &\succ 0 \\ ||\nabla^2 f(x)^{-1}|| &< M' \end{aligned}$$

for all $x \in \mathcal{N}(x^*)$ and some M'.

Suppose x^k lies in $\mathcal{N}(x^*)$ for some iterate k. We may define h^k by

$$x^* = x^k + h^k.$$

A Taylor expansion of the gradient of f about x^k gives

$$\nabla f(x^k + h^k) = \nabla f^k + \nabla^2 f^k h^k + O(||h^k||^2).$$

But since $\nabla f(x^k + h^k) = \nabla f^* = 0$, we get

$$0 = \nabla f^k + \nabla^2 f^k h^k + O(||h^k||^2).$$

Multiplying through by $(\nabla^2 f^k)^{-1}$ gives

$$\begin{aligned} -h^k &= (\nabla^2 f^k)^{-1} \nabla f^k + O(||h^k||^2) \\ \delta^k - h^k &= O(||h^k||^2) \end{aligned}$$

where the big O term remains the same since the norm of $(\nabla^2 f^k)^{-1}$ is bounded on $\mathcal{N}(x^*)$.

With a little bit of algebraic manipulation, we see that

$$\begin{aligned} x^k + h^k &= x^{k+1} + h^{k+1} \\ x^k + h^k &= x^k + \delta^k + h^{k+1} \\ -\delta^k + h^k &= h^{k+1}. \end{aligned}$$

So that

$$h^{k+1} = O(||h^k||^2). \tag{6.7}$$

If we can show that x^{k+1} remains in $\mathcal{N}(x^*)$, the proof is complete as we already have second order convergence via (6.7).

Equation (6.7) gives that there exists a constant depending on k, M_k, such that

$$||h^{k+1}|| \leq M_k ||h^k||^2.$$

But by the integral formulation of the Taylor expansion given in (5.22) coupled with the derivation above, we may uniformly bound all iterates by some $\tilde{M}$ by the continuity of $\nabla^2 f$ as

$$||h^{k+1}|| \leq \tilde{M} ||h^k||^2.$$

This exercise is left to the reader. For this $\tilde{M}$, assume that there is some k_0 such that

$$||h^{k_0}|| \leq \frac{\alpha}{\tilde{M}}$$

for some $\alpha \in (0,1)$ sufficiently small to remain in $\mathcal{N}(x^*)$. We have then,

$$\begin{aligned} ||h^{k_0+1}|| &\leq \tilde{M}||h^{k_0}||^2 \\ &\leq \tilde{M}\left(\frac{\alpha}{\tilde{M}}\right)||h^{k_0}|| \\ &\leq \alpha||h^{k_0}|| \end{aligned}$$

so that x^{k_0+1} remains in $\mathcal{N}(x^*)$ and all of the previous assumptions hold as well. Finally,

$$\begin{aligned} ||h^{k_0+2}|| &\leq \tilde{M}||h^{k_0+1}||^2 \\ &\leq \tilde{M}\frac{\alpha}{\tilde{M}}\alpha||h^{k_0}|| \\ &\leq \alpha^2||h^{k_0}|| \end{aligned}$$

and in general, $||h^{k_0+N}|| \leq \alpha^N||h^{k_0}||$, so that $lim_{k\to\infty}||h^k|| = 0$. □

Exercises

1. Consider

$$\begin{aligned} f(\beta) &= ||Y - X\beta||^2_{V^{-1}} \\ &= \beta' X' V^{-1} X\beta - 2\beta' X' V^{-1} Y + Y' V^{-1} Y. \end{aligned}$$

   (a) Show that the hessian of f is exactly $\nabla^2 f(\beta) = 2X'V^{-1}X$.

   (b) Prove that $X'V^{-1}X$ is positive semidefinite when X is full rank and V is a covariance matrix.

2. We have seen the definition of a norm for a square matrix A given by

$$||A||_2 = \max_{||v||=1} ||Av||$$

   and proved in an exercise that

$$||A||_2 = \max_{||v||} \frac{||Av||}{||v||}.$$

   Suppose A is invertible. What is $||A^{-1}||$? Prove your claim.

3. Suppose $(\hat{\beta}^1, \ldots, \hat{\beta}^N)$ are the OLS estimates of

$$Y^i = X\beta^i + \epsilon^i$$

   for assets $i = 1, \ldots, N$. Let $Y = (Y^1\, Y^2\, \cdots\, Y^N) \in \mathbb{R}^{M\times N}$, where $Y^i \in \mathbb{R}^M$. We have seen that OLS regression is the solution to a minimization of norm.

   (a) For some fixed w, find the solution to

$$\min_{\beta} ||Yw - X\beta||^2$$

   in terms of w, X, and Y.

   (b) Conclude that the multifactor β of a portfolio is the weighted sum of asset βs.

4. For this problem, you will need to write code for the multivariate Newton's method. Consider the function

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2$$

   (a) Using the multivariate version of Newton's method with initial point $(4, 4)$, find a stationary point of f.

   (b) Plot your iterates (x_i, y_i). If possible, include the level curves of f in your plot.

   (c) Plot $|\nabla f(x_i, y_i)|$ for each of your iterates.

5. In Merton's structural model, we have for r the risk free rate, F the debt barrier, and E and σ_E the given equity value and equity volatility, respectively, that

$$V_t\Phi(d_1) - E - Fe^{-r(T-t)}\Phi(d_2) \quad = \quad 0 \tag{6.8}$$

and

$$\sigma_V V - E\Phi^{-1}(d_1)\sigma_E = 0 \tag{6.9}$$

where

$$d_1 = \frac{\ln\left(\frac{V}{F}\right) + \left(r + \frac{\sigma_V^2}{2}\right)(T-t)}{\sigma_V\sqrt{T-t}},$$
$$d_2 = d_1 - \sigma_V\sqrt{T-t},$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal density, with derivative denoted by the probability density function $\phi(\cdot)$. For fixed E and σ_E, let

$$M(V, \sigma_V) \quad = \quad \begin{pmatrix} V_t\Phi(d_1) - E - Fe^{-r(T-t)}\Phi(d_2) \\ \sigma_V V - E\Phi^{-1}(d_1)\sigma_E \end{pmatrix}.$$

   (a) Find the Jacobian of M, ∇M.

   (b) Write a Newton's Method algorithm using your Jacobian function above to find (V, σ_V) when $(E, F, \sigma_E, r, T, t) = (1, 0.5, 0.20, 0.0025, 1, 0)$.

6. Consider the shrinkage estimator problem where

$$R(\alpha) = \alpha F + (1-\alpha)S - \Sigma$$

where $F = (f_{ij})$ is the shrinkage target, $S = (s_{ij})$ is the sample covariance, and $\Sigma = (\sigma_{ij})$ is the covariance. Let

$$L(\alpha) = ||R(\alpha)||^2 = tr(R(\alpha)'R(\alpha))$$

   (a) Show that $\mathbb{E}(L(\alpha))$ is minimized at

$$\alpha^* = \frac{\sum_{i,j} Var(s_{ij}) - Cov(f_{ij}, s_{ij})}{\sum_{i,j} \mathbb{E}(f_{ij} - \sigma_{ij})^2 + Var(s_{ij}) - 2Cov(f_{ij}, s_{ij})}$$

   (b) Show that you may write the denominator above as

$$\sum_{i,j} Var(f_{ij} - s_{ij}) + (\mathbb{E}(f_{ij}) - \sigma_{ij})^2$$

Chapter 7

Constrained Optimization

In this chapter, our work in unconstrained optimization is expanded to the case of identifying optima under some set of constraints for the decision variable x. As in the previous chapter, we begin with some preliminary discussion and establish necessary and sufficient conditions for a point x^* to be a solution for a class of objective functions. Whereas before this was a fairly straightforward endeavor, here we feel the need to illustrate the problem at hand with a few examples for clarity. After this motivation, we lay out the so-called Karush-Kuhn-Tucker (KKT) conditions for constrained optimization and provide a derivation of the first order necessary conditions. In so doing, we also develop some understanding of the associated Lagrange multipliers of the problem.

As was the case for unconstrained optimization, we again emphasize the attractive properties of convex functions.

With the basics in hand, we look at two broad classes of constrained optimization problems: Quadratic Programming with Linear Constraints (QPLC) and Linear Programming with Linear Constraints (LPLC). In each case, we establish standard forms of the problem and give some motivation as to how they might solved algorithmically by utilizing the KKT conditions of the particular problem. In the case of LPLC, we also lightly develop the concept of the dual.

We conclude with examples in QPLC and LPLC, ranging from a portfolio optimization problem identifying the market portfolio as the minimum variance portfolio with $\beta = 1$ to a regression technique called quantile regression.

7.1 Preliminaries

For functions $f : \mathbb{R}^N \to \mathbb{R}$ and $c_i : \mathbb{R}^N \to \mathbb{R}$,*constrained optimization problems* are given by

$$\begin{aligned} \min_x \quad & f(x) \\ & c_i(x) = 0 \ i \in \mathcal{E} \\ & c_i(x) \geq 0 \ i \in \mathcal{I}, \end{aligned} \tag{7.1}$$

where $\mathcal{E}$ and $\mathcal{I}$ denote sets of equality and inequality indices, respectively. We say that a point $\tilde{x}$ is a *feasible point*, or simply *feasible*, if it satisfies $c_i(\tilde{x}) = 0$ for $i \in \mathcal{E}$ and $c_i(\tilde{x}) \geq 0$ for $i \in \mathcal{I}$. The *feasible region* is the set of all feasible points, denoted

$$\chi = \{x \,|\, x \text{ is a feasible point}\}. \tag{7.2}$$

The functions $c_i(\cdot)$ are called the equality and inequality constraints based on membership in $\mathcal{E}$ and $\mathcal{I}$, respectively. Oftentimes, in the literature, the equation set in (7.1) is stated simply as

$$\min_{x \in \chi} f(x)$$

with the feasible set, χ defined by some set of functions $\{c_i(\cdot)\}_{i \in \mathcal{I} \cup \mathcal{E}}$.

To obtain solutions to (7.1), we have to develop a different set of criteria than in the unconstrained case; viz., χ may not contain any stationary points of $f(\cdot)$. Our main tool will be the so-called Karush-Kuhn-Tucker (KKT) conditions, which we motivate in several examples that follow. We will consider in turn the case of a single equality constraint and then a single inequality constraint. These examples, are not intended to be fully rigorous, but rather to illustrate the core ideas of the necessity of the KKT conditions using the math previously presented before tackling the proof of the same.

7.1.1 The Case of One Equality Constraint

Consider

$$\begin{aligned} \min_{x} \quad & f(x) \\ & c_1(x) = 0, \end{aligned}$$

where clearly, $\mathcal{E} = \{1\}$ and $\mathcal{I} = \{\varnothing\}$, the empty set. Assume that both f and $c_1 \in \mathbb{C}^1$ so that each function and its gradient are continuous. Then we have by (5.17) that

$$c_1(x + \delta) = c_1(x) + \delta' \nabla c_1(x) + o(||\delta||). \tag{7.3}$$

For a feasible x, then, up to first order, a move in the direction δ will stay feasible if

$$c_1(x) + \delta' \nabla c_1(x) = 0.$$

But since x is feasible, $c_1(x) = 0$, so this reduces to requiring

$$\delta' \nabla c_1(x) = 0. \tag{7.4}$$

For δ to be a descent direction as in (5.4), we require

$$\delta' \nabla f(x) < 0. \tag{7.5}$$

In the case that x is not a stationary point and $\nabla f(x) \neq \lambda \nabla c_1(x)$ for some constant λ, we may show there exists a δ simultaneously satisfying (7.4) and

(7.5). That is, a direction that both maintains feasibility (up to first order) and is a descent direction; viz.,

$$\delta = -\nabla f + \frac{(\nabla f' \nabla c_1) \cdot \nabla c_1}{||\nabla c_1||^2} \tag{7.6}$$

For this δ, we see that

$$\begin{aligned} \delta' \nabla c_1(x) &= -\nabla f' \nabla c_1 + \frac{(\nabla f' \nabla c_1) \cdot \nabla c_1' \nabla c_1}{||\nabla c_1||^2} \\ &= -\nabla f' \nabla c_1 + \nabla f' \nabla c_1 \\ &= 0, \end{aligned}$$

so that δ maintains first order feasibility. Continuing in the same fashion,

$$\begin{aligned} \delta' \nabla f(x) &= -\nabla f' \nabla f + \frac{(\nabla f' \nabla c_1) \cdot \nabla c_1' \nabla f}{||\nabla c_1||^2} \\ &= -||\nabla f||^2 + \frac{|\nabla f' \nabla c_1|^2}{||\nabla c_1||^2}. \end{aligned}$$

By Cauchy-Schwarz, we have that $|\nabla f' \nabla c_1|^2 = \beta ||\nabla f||^2 ||\nabla c_1||^2$, for some $\beta \in [0, 1]$. But since by assumption, $\nabla f(x) \neq \lambda \nabla c_1(x)$, the range for β is reduced to $\beta \in [0, 1)$. Hence

$$\begin{aligned} \delta' \nabla f(x) &= -||\nabla f||^2 + \frac{\beta ||\nabla f||^2 ||\nabla c_1||^2}{||\nabla c_1||^2} \\ &= -||\nabla f||^2 + \beta ||\nabla f||^2 \\ &= (\beta - 1)||\nabla f||^2 \\ &< 0. \end{aligned}$$

Giving that this direction is also a descent direction.

In conclusion, so long as $\nabla f(x) \neq \lambda \nabla c_1(x)$, any non-stationary x may be perturbed by δ to obtain $x + \delta$ that is both feasible and a descent direction. We are motivated to consider an equation such as

$$\mathcal{L}(x, \lambda) = f(x) - \lambda c_1(x)$$

and require, for a necessary condition of optimality, that

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f^* - \lambda^* \nabla c_1^* = 0.$$

We will see that this condition is indeed the case, generalizing to many constraints.

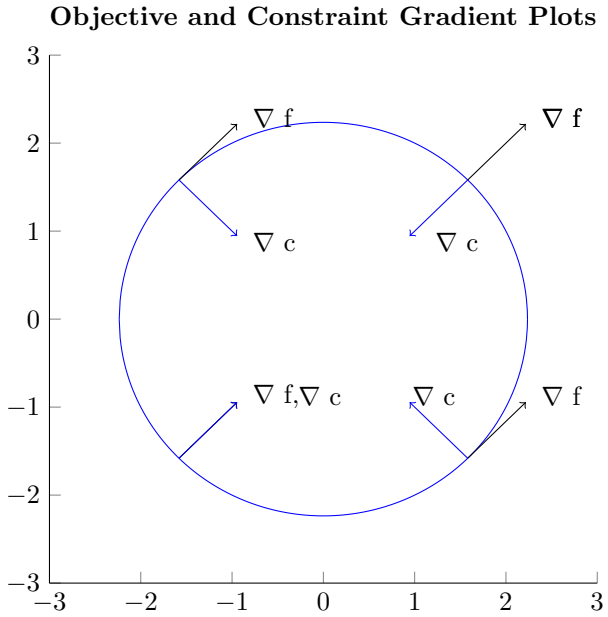

Figure 7.1: Plot of the gradient of the objective function and constraint function for minimizing $2x_1 + 2x_2$ on the circle of radius $\sqrt{5}$.

Example 7.1.1. Consider the following minimization problem constrained to a circle of radius $\sqrt{5}$:

$$\begin{aligned} \min_x \quad & f(x) = 2x_1 + 2x_2 \\ & c(x) = 5 - x_1^2 - x_2^2 = 0 \end{aligned}$$

Notice that (by careful construction) this problem exhibits symmetry in the two variables in question. As such, the problem reduces to

$$\begin{aligned} \min_x \quad & f(x) = 4x \\ & c(x) = 5 - 2x^2 = 0 \end{aligned}$$

where now $x \in \mathbb{R}$. A solution is clearly determined from the constraints, which determine that $x = \pm\sqrt{\frac{5}{2}}$. This gives four possible points where the minimum may occur, and we find that $x^* = -\sqrt{\frac{5}{2}}$ achieves the minimum value of f on the circle.

Now, the gradient of the constraint function is given by

$$\nabla c(x) = \begin{pmatrix} -2x_1 \\ -2x_2 \end{pmatrix}.$$

Plotting the gradient of f, $\nabla f \equiv (2,2)'$, we see in Figure 7.1.1, where the magnitude of the vectors have each been normalized, that at the minimum, x^*, the gradients of f and c are parallel. Elsewhere at the maximum ($-x^*$), we have that the angle between ∇f and ∇c is π.

7.1.2 The Case of One Inequality Constraint

We next look at the case of one inequality constraint,

$$\begin{aligned} \min_x \quad & f(x) \\ & c_1(x) \geq 0, \end{aligned}$$

with, $\mathcal{E} = \{\varnothing\}$ and $\mathcal{I} = \{1\}$. Again, assuming that f and $c_1 \in \mathbb{C}^1$, we have in a manner similar to the above that first order requirements for a feasible descent direction (for nonstationary point, x) are

$$\begin{aligned} c_1(x) + \delta' \nabla c_1(x) &\geq 0 \\ \delta' \nabla f(x) &< 0. \end{aligned}$$

We consider two cases: x on the interior of the disc and x on the boundary.

When on the interior, if not at a stationary point, we may find a γ small enough so that $\delta = -\gamma \nabla f(x)$ remains feasible. Clearly, as well, δ is a descent direction.

On the boundary, $c_1(x) = 0$, so we require

$$\begin{aligned} \delta' \nabla c_1(x) &\geq 0 \\ \delta' \nabla f(x) &< 0. \end{aligned}$$

By Cauchy-Schwarz, we have then that

$$\begin{aligned} \delta' \nabla c_1 &= \cos\theta_1 ||\delta|| \cdot ||\nabla c_1||, \, \theta_1 \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \\ \delta' \nabla f &= \cos\theta_2 ||\delta|| \cdot ||\nabla f||, \, \theta_2 \in \left(\frac{\pi}{2}, \frac{3\pi}{2}\right). \end{aligned}$$

The set of feasible descent directions is determined by a cone constructed from the intersection of the regions determined above. One may convince themselves from this that if ∇f and ∇c_1 are not parallel, there will always be a cone of feasible directions reducing f.

We again see motivation to consider $\mathcal{L}(x, \lambda) = f(x) - \lambda c_1(x)$ and the necessity of requiring $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ for x^* a minimizer of f on the feasible region. In this case, we have an additional requirement that $\lambda^* \geq 0$, with positivity of λ^* if x^* is on the boundary. On the interior, optimality occurs at a stationary point, and hence $\lambda^* = 0$ in this simple example.

These requirements are called *strict complementarity*. In particular, the product, $\lambda^* c_1^* = 0$, but either $\lambda^* > 0$ and $c_1 = 0$ or $\lambda^* = 0$ and $c_1 > 0$. The pair are never simultaneously zero unless f has a stationary point on the boundary.

7.1.3 First and Second Order Conditions

As in the unconstrained case, we obtain first and second order necessary and sufficient conditions for optimality for the constrained problem (7.1).

Before proceeding, we say that x is a *regular point* of (7.1) if

$$\{\nabla c_i(x) \,|\, i \in \mathcal{A}(x)\} \tag{7.7}$$

is linearly independent, where $\mathcal{A}(x)$ is the *active set* at x given by

$$\mathcal{A}(x) = \mathcal{E} \cup \{i \in \mathcal{I} \,|\, c_i(x) = 0\}, \tag{7.8}$$

or the set of indices where equality obtains in the constraint set. We further say that an inequality constraint, c_i is *active* at x if $c_i(x) = 0$.

Finally, for a feasible point x, and motivated by the linear approximations used in the examples above, we define the *set of linearized feasible directions*, $\mathcal{F}(x)$ by

$$\mathcal{F}(x) = \left\{ d \;\middle|\; \begin{array}{ll} d'\nabla c_i(x) = 0, & i \in \mathcal{E} \\ d'\nabla c_i(x) \geq 0, & i \in \mathcal{A}(x) \cap \mathcal{I} \end{array} \right\} \tag{7.9}$$

With these definitions, the necessary and sufficient Karush-Kuhn-Tucker conditions for constrained optimization problems given by (7.1) may be stated. In each case that follows, we define the *Lagrangian function* by

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x). \tag{7.10}$$

Each component, λ_i, of the vector λ (whose size is equal to the total number of constraints) is called a *Lagrange multiplier*.

First Order Necessary Condition For x^* a regular point and local solution to the constrained optimization problem (7.1) with f and each c_i in $\mathbb{C}^1$, there exists a λ^* such that the following conditions hold:

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \lambda^*) &= 0 && \text{(7.11a)} \\ x^* &\text{ is feasible} && \text{(7.11b)} \\ \lambda_i^* &\geq 0, \text{ for all } i \in \mathcal{I} && \text{(7.11c)} \\ \lambda_i^* c_i(x^*) &= 0, \text{ for all } i \in \mathcal{E} \cup \mathcal{I}. && \text{(7.11d)} \end{aligned}$$

Collectively the equations in (7.11) are known as the Karush-Kuhn-Tucker conditions, or KKT, while the final conditions are called complementarity conditions.

Second Order Necessary Condition To state the second order necessary condition here, we must first define the *critical cone*. For $\mathcal{F}(x^*)$ the set of feasible linear directions as in (7.9) and (x^*, λ^*) a pair satisfying the KKT conditions (7.11), the *critical cone*, $\mathcal{C}(x^*, \lambda^*)$, is defined by

$$\mathcal{C}(x^*, \lambda^*) = \{d \in \mathcal{F}(x^*) \,|\, d'\nabla c_i^* = 0,\, i \in \mathcal{A}(x^*) \cap \mathcal{I}, \text{ with } \lambda_i^* > 0\}. \tag{7.12}$$

By complementarity, we see that the critical cone contains the set of directions that maintain linear feasibility and also retain up to first order the same active constraint set. Further it follows from complementarity of KKT and the definition here that if $d \in \mathcal{C}(x^*, \lambda^*)$, then

$$\lambda_i^* d' \nabla c_i^* = 0$$

for all $i \in \mathcal{E} \cup \mathcal{I}$.

We have that if $d \in \mathcal{C}(x^*, \lambda^*)$,

$$d' \nabla f^* = \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* d' \nabla c_i^* = 0. \tag{7.13}$$

As a result, the critical cone contains the set of all feasible linear directions for which the first order necessary conditions are not clear with regards to an increase or decrease in f.

For x^* a regular point and local solution to the constrained optimization problem (7.1) with f and each c_i in $\mathbb{C}^2$, and λ^* a solution to (7.11), then

$$d' \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) d \geq 0 \tag{7.14}$$

for all $d \in \mathcal{C}(x^*, \lambda^*)$.

Second Order Sufficient Condition For f and each c_i in $\mathbb{C}^2$, if x^* is feasible and (x^*, λ^*) satisfy the KKT conditions and

$$d' \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) d > 0 \tag{7.15}$$

for all $d \in \mathcal{C}(x^*, \lambda^*)$, then x^* is a strict local minimizer of (7.1).

7.1.4 Mathematical Derivation for First Order Conditions

In much of the preceding section, we gave motivating examples for the first order necessary conditions encapsulated in the KKT conditions just cited. Here we give a formal treatment using the inverse mapping theorem.

As before, we begin by considering the case of having only equality constraints. Define the map $F : \mathbb{R}^N \to \mathbb{R}^M$ by

$$F(x) = \begin{pmatrix} c_1(x) \\ \vdots \\ c_m(x) \end{pmatrix}$$

where $\mathcal{E} = \{1, \ldots, m\}$. If x^* is optimal, and the Jacobian of F is locally invertible at x^* (equivalently, x^* is a regular point), then there exists a smooth map $G : \mathbb{R}^N \to \mathbb{R}^N$ such that G satisfies

$$x^* = G(z^*)$$

for some z^*, with $\nabla G(z^*)$ invertible. Further, the coordinates z may be chosen such that the composite constraint functions, $\tilde{c}_i(z) = c_i(G(z))$ satisfy

$$\tilde{c}_i(z) = z_i.$$

In z, then, the original constraints become simply $z_i = 0$ for $i \in \mathcal{E}$.

As a result, the composite objective function, $\tilde{f}(z) = f(G(z))$ must satisfy the first order necessary condition

$$\frac{\partial \tilde{f}}{\partial z_i}(z^*) = 0$$

for $i \in \{m+1, \ldots, N\}$.

Defining

$$\lambda_i^* = \frac{\partial \tilde{f}}{\partial z_i}(z^*)$$

for all indexes, then, we must have

$$\frac{\partial \tilde{f}}{\partial z_i}(z^*) - \lambda_j^* \frac{\partial \tilde{c}_j}{\partial z_i}(z^*) = 0$$

over all indexes i, and for any $j \in \mathcal{E}$ since $\tilde{c}_j(z) \equiv z_j$. This may be succinctly written as

$$\nabla \tilde{f}(z^*) - \sum_{i \in \mathcal{E}} \lambda_i^* \nabla \tilde{c}_i(z^*) = 0. \tag{7.16}$$

Finally, since $x^* = G(z^*)$, we may write the above in the original coordinates as

$$\nabla f(x^*) - \sum_{i \in \mathcal{E}} \lambda_i^* \nabla c_i(x^*) = 0. \tag{7.17}$$

Hence we have shown that if x^* is a regular point, a necessary condition for x^* to be optimal is that there exists a λ^* such that the gradient of the Lagrangian for the problem (as previously defined) vanishes for both gradients in x and λ at (x^*, λ^*); viz.,

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \tag{7.18}$$

$$\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = 0, \tag{7.19}$$

the second equation being a concise way to denote the feasibility of x^*. Notice that this is exactly the first order necessary condition previously stated.

The generalization to inequality constraints is handled similarly. As before, we may (smoothly) change to an easier coordinate system, but now accounting for the active set at x^*, $\mathcal{A}(x^*)$, rather than just the equality set, $\mathcal{E}$. Again, so long as the active constraints are linearly independent at x^*, there exists a G as before satisfying

$$\tilde{c}_i(z) = c(G(x)) = z_i$$

for $i \in \mathcal{A}(x^*)$.

In this coordinate system, we seek to minimize $\tilde{f}(z) = f(G(z))$ subject to

$$\begin{aligned} z_i &= 0,\ i \in \mathcal{E} \\ z_i &\geq 0,\ i \in \mathcal{A}(x^*) \cap \mathcal{I}, \end{aligned}$$

with $z_i^* = 0$ for $i \in \mathcal{A}(x^*)$ by construction. The first order necessary conditions of optimality of z^* become, then,

$$\begin{aligned} \frac{\partial \tilde{f}}{\partial z_i}(z^*) &= 0,\ i \notin \mathcal{A}(x^*) \\ \frac{\partial \tilde{f}}{\partial z_i}(z^*) &\geq 0,\ i \in \mathcal{A}(x^*) \cap \mathcal{I}. \end{aligned}$$

Defining

$$\begin{aligned} \lambda_i^* &= \frac{\partial \tilde{f}}{\partial z_i}(z^*),\ i \in \mathcal{A}(x^*) \\ \lambda_i^* &= 0,\ i \notin \mathcal{A}(x^*), \end{aligned} \tag{7.20}$$

we have that

$$\frac{\partial \tilde{f}}{\partial z_i}(z^*) - \lambda_j^* \frac{\partial \tilde{c}_j}{\partial z_i}(z^*) = 0.$$

This may be seen by considering the various partitions of the index, i. In any partition, however, we note that the term $\lambda_j \frac{\partial \tilde{c}_j}{\partial z_i}(z^*)$ is only nonzero when $j = i$ by our choice of coordinate system and definition of $\lambda_\cdot$ above.

For $i \notin \mathcal{E} \cup \mathcal{I}$, $\frac{\partial \tilde{f}}{\partial z_i}(z^*)$ vanishes and there are no j's that can equal i in $\lambda_j^* \frac{\partial \tilde{c}_j}{\partial z_i}$ on this set of indexes. Hence these terms are all zero as well.

When $i \in \mathcal{A}(x^*)$, we have that

$$\lambda_i^* \frac{\partial \tilde{c}_i}{\partial z_i}(z^*) = \frac{\partial \tilde{f}}{\partial z_i}(z^*)$$

and hence the sum in question is zero.

Finally, those indexes in the inequality set which are inactive must have both vanishing $\frac{\partial \tilde{f}}{\partial z_i}(z^*)$ and λ_i^*.

As before, we may write these results concisely as

$$\nabla \tilde{f}(z^*) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla \tilde{c}_i(z^*) = 0, \tag{7.21}$$

or, in the original coordinate system

$$\nabla f(x^*) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla c_i(x^*) = 0. \tag{7.22}$$

The result may be summarized by saying that in the case of both equality and inequality constraints, a necessary condition for a feasible regular point, x^*, to be optimal is that there exists λ^* satisfying

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$$

with those λ_i^* associated with inequality constraints being positive when inactive and zero when active; that is, strict complementarity holds across all constraints. As before, this result coincides with our previous summary of the KKT conditions.

Finally, we notice that (7.20) indicates an interpretation of the Lagrange multipliers satisfying the KKT conditions. From the equations there is some sense that λ_i^* is equivalent to the sensitivity of the objective function to small changes in the ith constraint; viz., for the inactive constraints, we would expect the impact to be zero since the constraints are nonbinding.

While this is clear in the alternative coordinate system produced, we may also prove this result with a more elementary derivation. Focusing on the standard optimization problem (7.1), we let x^* and λ^* satisfy the associated first order KKT conditions. Next we consider the case of changing the jth constraint in the active set at x^* from $c_j(x) = 0$ to $c_j(x) = \delta$.

The result of this change would be to obtain a new optimal solution for the corresponding problem which we may write as $x^* + \Delta x^*$. The approximate change in objective function values is given by

$$f(x^* + \Delta x^*) - f(x^*) \approx \nabla f(x^*)' \Delta x^*,$$

which we may make exact with notation as in (5.17), for example. By the first order KKT conditions, we have that

$$\nabla f(x^*)' \Delta x^* = \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*)' \Delta x^*.$$

We may also write a linear approximation of the constraint functions as

$$c_i(x^* + \Delta x^*) - c_i(x^*) \approx \nabla c_i(x^*)' \Delta x^*,$$

for each i in the active set. For each $i \neq j$, however, we have that the left hand side of this equation is zero. For $i = j$, we are left with

$$\delta = \nabla c_j(x^*)' \Delta x^*.$$

This gives immediately that

$$\nabla f(x^*)' \Delta x^* = \lambda_j^* \delta,$$

and hence

$$\frac{f(x^* + \Delta x^*) - f(x^*)}{\delta} \approx \lambda_j^*.$$

Taking the limit as $\delta \to 0$ confirms the result.

We note that, smoothness requirements for both f and $c.$ are established based upon the Taylor approximations used. In particular, the above proof required that these functions simply be differentiable at x^*.

7.2 Convex Functions and KKT

The first order KKT conditions on (x^*, λ^*) are sufficient in the case that f is convex and each c_i is concave (i.e., $-c_i$ is convex). In other words, the satisfaction of the KKT conditions in this case ensures a solution to the constrained problem.

Before proceeding, we note that the set

$$K = \{x | c_i(x) \geq k_i\} \tag{7.23}$$

is convex when each c_i is concave and leave the proof to the reader. With respect to the constrained optimization problem we are considering, this implies that the feasible set is convex.

Assume next that (x^*, λ^*) satisfy the first order KKT conditions with convex f and concave c_i. For any point, x_0, we have that

$$f(x_0) \geq f(x_0) - \sum_i \lambda_i^* c_i(x_0). \tag{7.24}$$

And by convexity of f,

$$f(x_0) \geq f(x^*) + \nabla f(x^*)'(x_0 - x^*). \tag{7.25}$$

Similarly, by concavity of $c.$,

$$-c_i(x_0) \geq -c_i(x^*) - \nabla c_i(x^*)'(x_0 - x^*). \tag{7.26}$$

By (7.24), (7.25), and (7.26), then,

$$\begin{aligned}
f(x_0) &\geq f(x_0) - \sum_i \lambda_i^* c_i(x_0) \\
&\geq f(x^*) + \nabla f(x^*)'(x_0 - x^*) - \sum_i \lambda_i^* c_i(x_0) \\
&\geq f(x^*) + \nabla f(x^*)'(x_0 - x^*) - \sum_i (c_i(x^*) + \nabla c_i(x^*)'(x_0 - x^*)) \\
&\geq f(x^*) + \left(\nabla f(x^*) - \sum_i \lambda_i^* \nabla c_i(x_0) \right)' (x_0 - x^*) - \sum_i \lambda_i^* c_i(x^*).
\end{aligned}$$

At x^*, the KKT conditions give that the gradient of the Lagrangian is zero so that

$$\nabla f(x^*) - \sum_i \lambda_i^* \nabla c_i(x^*) = 0$$

and further, by strict complementarity,

$$\sum_i \lambda_i^* c_i(x^*) = 0.$$

These observations in concert with the last line of the inequality above yield that $f(x_0) \geq f(x^*)$ and show that $f(x^*)$ is the minimum value obtained on the constrained set.

7.3 Quadratic Programming with Linear Constraints

Of particular importance will be the case of a quadratic objective function with linear constraints. Here we consider the so-called Quadratic Problem with Linear Constraints (QPLC) given by

$$\begin{aligned} \min_x \quad & q(x) = \frac{1}{2}x'Qx - r'x \\ & Ax = b. \end{aligned} \tag{7.27}$$

The general case of linear inequality constraints is also a QPLC.

We assume that $A \in \mathbb{R}^{M\times N}$ is full rank and, for now, that Q is simply symmetric. We will tighten our assumptions on Q below. Relating this new formulation to the notation above, we may write

$$A = \begin{pmatrix} — & a_1 & — \\ — & a_2 & — \\ & \vdots & \\ — & a_M & — \end{pmatrix}$$

as

$$c_i(x) = a_i \cdot x - b_i.$$

The gradient of q is found directly to be

$$\nabla q(x) = Qx - r$$

since Q is symmetric, and the gradient of each c_i is exactly a_i'. As a result, we may state the KKT condition on the gradient of the Lagrangian as

$$Qx^* - r - \sum_i \lambda_i^* a_i' = 0. \tag{7.28}$$

Looking at the final sum, we see that

$$\begin{aligned} \sum_i \lambda_i^* a_i' &= \lambda_1^* \begin{pmatrix} | \\ a_1 \\ | \end{pmatrix} + \cdots + \lambda_M^* \begin{pmatrix} | \\ a_M \\ | \end{pmatrix} \\ &= A'\lambda. \end{aligned}$$

As a result, (7.28) may be rewritten as

$$Qx^* - r - A'\lambda^* = 0. \tag{7.29}$$

The only remaining condition, feasibility, is given by $Ax^* = b$.

In all, then, the KKT conditions for this problem may be summarized as a system of linear equations

$$\begin{pmatrix} Q & -A' \\ A & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} r \\ b \end{pmatrix}. \tag{7.30}$$

In the specific case of $Q \succ 0$, we know from our previous work that the system must have a unique solution. Here we prove a uniqueness result for a less strict condition on Q.

For K the matrix given in (7.30) and Q symmetric and satisfying $Z'QZ \succ 0$ for Z the matrix made up of the columns of the basis of the null space of A, $null(A)$, there exists a unique solution to (7.27).

Proof. We prove that K is nonsingular. Suppose $(v,\, w)'$ satisfies

$$K \begin{pmatrix} \mathrm{v} \\ \mathrm{w} \end{pmatrix} = 0$$

for some v and w not identically zero. Then we must have $Av = 0$ and $v \in null(A)$. Now, the inner product of $(v,\, w)'$ with respect to K gives

$$\begin{aligned} (v'\, w') \begin{pmatrix} Q & -A' \\ A & 0 \end{pmatrix} \begin{pmatrix} \mathrm{v} \\ \mathrm{w} \end{pmatrix} &= (v'\, w') \begin{pmatrix} Qv - A'w \\ Av \end{pmatrix} \\ &= v'Qv - v'Aw + w'Av \\ &= 0. \end{aligned}$$

Since v is in the null space of A, this reduces to

$$v'Qv = 0.$$

Continuing, we may write $v = Zu$ for some u since Z spans $null(A)$. This gives, then, that

$$u'Z'QZu = 0,$$

contradicting the positive definiteness of $Z'QZ$. Therefore $u \equiv 0$ and so is v.

Next we show that w must be zero as well. From the first set of equations in the system, we have that

$$Qv - A'w = 0,$$

which reduces to $A'w = 0$ as $v \equiv 0$. Finally, since A has full rank, $A'w = 0$ implies w must be zero as well.

We conclude that K is nonsingular as desired.

□

While it is interesting that the process shown above yields the solution to the constrained QP problem is exactly that of a linear system of equations in the extended variable set including the Lagrange multipliers, it is also a sketch that leads to a more general class of algorithms called interior point methods. We give a first indication of such an implementation in the exercises; viz., how might the above be generalized to non-quadratic objective functions and how might an iteration scheme be developed?

7.4 Linear Programming with Linear Constraints

Another indispensable framework is the Linear Program with Linear Constraints (LPLC), or simply, Linear Programming, the standard form of which is given by

$$\begin{aligned} \min_x \quad & c'x \\ & Ax = b \\ & x \geq 0. \end{aligned} \tag{7.31}$$

The formulation in (7.31) is more flexible than it might first appear. For example, inequality constraints such as

$$Cx \geq d$$

may be described with the use of so-called slack variables, s, as

$$\begin{aligned} Cx - s &= d \\ s &\geq 0. \end{aligned}$$

The nonnegativity constraint on x is also less restrictive than might otherwise be supposed. The general case of

$$\begin{aligned} \min_x \quad & c'x \\ & Ax = b \\ & Cx \geq d \end{aligned} \tag{7.32}$$

may be written in the standard form by writing x as a difference of its positive and negative parts,

$$x = x_+ - x_-$$

with each of x_+ and x_- being nonnegative. The objective function in this case becomes

$$\begin{pmatrix} c \\ -c \\ 0 \end{pmatrix}' \begin{pmatrix} x_+ \\ x_- \\ s \end{pmatrix}$$

where we have anticipated the need for slack variables. Completing the formulation of (7.32) in standard form is left as an exercise.

Without loss of generality, then, we may focus on problems of the type (7.31). The Lagrangian for this problem is

$$\mathcal{L}(x, \lambda, \eta) = c'x - \lambda'(Ax - b) - \eta'x,$$

and the KKT conditions are

$$\nabla_x \mathcal{L}(x^*, \lambda^*, \eta^*) = c - A'\lambda^* - \eta^* = 0 \tag{7.33a}$$

$$Ax^* = b \tag{7.33b}$$

$$x^* \geq 0 \tag{7.33c}$$

$$\eta^* \geq 0 \tag{7.33d}$$

$$\eta^{*\prime} x^* = 0. \tag{7.33e}$$

Notice that the complementarity condition has been replaced by $\eta^{*\prime} x^* = 0$, but that these conditions are equivalent due to the non-negativity of x^* and η^*.

We next show directly that the KKT conditions in (7.33) are sufficient for an optimal solution of (7.32). For any feasible $\tilde{x}$, we have that the value in the objective function is

$$\begin{aligned} c'\tilde{x} &= (A'\lambda^* + \eta^*)'\tilde{x} \\ &= \lambda^{*\prime} A\tilde{x} + \eta^{*\prime}\tilde{x} \\ &= \lambda^{*\prime} b + \eta^{*\prime}\tilde{x} \\ &\geq \lambda^{*\prime} b. \end{aligned}$$

Now the objective function at x^* is related to this final linear equation in the Lagrange multiplier, λ^* vis a vis the same steps as above, but noting that $\eta^{*\prime}\tilde{x^*} = 0$. By doing so, we see that

$$c'x^* = \lambda^{*\prime} b. \tag{7.34}$$

Putting this into the last line of the above, we conclude that $c'\tilde{x} \geq c'x^*$ and the KKT conditions are sufficient for x^* to be a solution to the original problem.

The relationship in (7.34) is significant in its own right. In fact, with this equation and an analysis of the KKT conditions in (7.33), we may construct the so-called dual problem and note the Strong Duality Theorem of linear programming. In what follows, we will refer to our original linear programming problem as the primal.

Suppose that (x^*, λ^*, η^*) is a solution to the first order KKT conditions in (7.33). The condition

$$c - A'\lambda^* - \eta^* = 0$$

is indicative of a linear inequality in λ. With this intuition along with (7.34), we write the dual problem of (7.31) as

$$\begin{aligned} \max_{\lambda} \quad & b'\lambda \\ & A'\lambda \leq c. \end{aligned} \tag{7.35}$$

Verifying that the KKT conditions of (7.35) coincide with those of the primal confirms this choice. We have

$$\begin{aligned} \nabla_\lambda \tilde{\mathcal{L}}(x^*, \lambda^*, \eta^*) &= -b + Ax^* = 0 \\ A'\lambda^* &\leq c \\ x^* &\geq 0 \\ x_i^*(a_i x^* - c) &= 0. \end{aligned}$$

Substituting $\eta^* = c - A'\lambda^*$ and rearranging some terms gives

$$\begin{aligned} Ax^* &= b \\ \eta^* &\geq 0 \\ x^* &\geq 0 \\ \eta^{*\prime} x^* &= 0, \end{aligned}$$

which is exactly the KKT conditions determined for the original problem. We have already shown, too, that the objective function values at optimal solutions are equal for both the primal and dual. The variables (λ, η) are often referred to as the dual variables for (7.31).

We collect these results in the following theorem, the second portion of which requires some further development, but is readily available from the above.

Theorem 7.4.1 (Strong Duality)**.**

If the primal problem (7.31) affords a finite solution, then so does the dual (7.35), and their objective functions are equal. Conversely, if the dual has a finite solution so does the primal.

If the primal (dual) is unbounded, then the dual (primal) is infeasible.

As a last point of discussion in this introductory exposition of Linear Programming, and in a manner similar to that shown for Quadratic Programming with Linear Constraints above, we motivate an algorithmic technique for solving (7.31). These algorithms are referred to as primal-dual interior point methods, and as we shall soon see, what follows is mainly a restatement of results we have already obtained. The updating techinque we indicate should also feel familiar.

We begin by writing the primal KKT conditions in a functional form by first defining

$$F(x, \lambda, \eta) = \begin{pmatrix} A'\lambda + \eta - c \\ Ax - b \\ \text{XSe} \end{pmatrix} \tag{7.36}$$

where $X = diag(x_1, \ldots, x_N)$ and $S = diag(\eta_1, \ldots, \eta_N)$ and e is a vector of ones. We may concisely state the conditions now as

$$\begin{aligned} F &= 0 \\ \eta' x &= 0. \end{aligned}$$

Interior point primal-dual methods generate a sequence (x^k, λ^k, η^k) satisfying $\eta^k > 0$ and $x^k > 0$ (hence the qualifier 'interior') and approximately solving $F^k = 0$.

Given (x^k, λ^k, η^k), we would like to construct an update $(\Delta x, \Delta\lambda, \Delta\eta)$ such that $F^{k+1} = 0$. As we have seen with other nonlinear functions, especially when considering Newton's method, we may approximate the nearly linear F by way of its Jacobian. We have

$$\nabla F(x, \lambda, \eta) = \begin{pmatrix} 0 & A' & I \\ A & 0 & 0 \\ S & 0 & X \end{pmatrix}. \tag{7.37}$$

To solve F^{k+1} up to first order, we see that the updates $(\Delta x, \Delta\lambda, \Delta\eta)$ must satisfy

$$\nabla F^k \begin{pmatrix} x^{k+1} \\ \lambda^{k+1} \\ \eta^{k+1} \end{pmatrix} = \begin{pmatrix} c \\ b \\ 0 \end{pmatrix}.$$

Or,

$$\nabla F^k \begin{pmatrix} x^k \\ \lambda^k \\ \eta^k \end{pmatrix} + \nabla F^k \begin{pmatrix} \Delta x \\ \Delta\lambda \\ \Delta\eta \end{pmatrix} = \begin{pmatrix} c \\ b \\ 0 \end{pmatrix}.$$

Which gives

$$\nabla F^k \begin{pmatrix} \Delta x \\ \Delta\lambda \\ \Delta\eta \end{pmatrix} = \begin{pmatrix} c - A'\lambda^k - \eta^k \\ b - Ax^k \\ -S^k x^k - X^k \eta^k \end{pmatrix},$$

with updates now available by solving the resulting linear system.

As usual, the choice of how much of a Newton step to take is a critical refinement of the process, but a fairly accurate representation of the algorithm is obtained from the above and determining a choice of update length α so that

$$(x^{k+1}, \lambda^{k+1}, \eta^{k+1}) = (x^k, \lambda^k, \eta^k) + \alpha(\Delta x, \Delta\lambda, \Delta\eta). \tag{7.38}$$

We have omitted some important details, however. In particular, we have not dealt with the question of remaining interior; viz., we have not handled $x^{k+1} = 0$ and $s^{k+1} = 0$ in our treatment. The full algorithm (along with its convergence behavior) is beyond the scope of the text, unfortunately, but the interested reader should be well equipped to engage such material from this point.

7.5 Constrained Optimization Examples

As in the unconstrained case, several examples become accessible having identified necessary and sufficient conditions for optimality. Here we establish constrained optimization problems with analytic solutions identifying the eigenvalues of a positive definite matrix, constructing prediction intervals for generalized least squares problems, relating minimum variance portfolios with $\beta = 1$ to the market portfolio, and establishing quantile regression as a Linear Programming problem.

Example 7.5.1. The eigenvalues of a covariance matrix, Σ may be determined via constrained optimization. Consider

$$\begin{aligned} \min_x \quad & x'\Sigma x \\ & ||x||^2 = 1. \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(x, \lambda) = x'\Sigma x - \lambda\left(||x||^2 - 1\right)$$

which has gradient

$$\nabla_x \mathcal{L}(x, \lambda) = 2\Sigma x - 2\lambda x.$$

We require by (7.11) that $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$, giving

$$\Sigma x^* = \lambda^* x^*.$$

Hence x^* is an eigenvector with associated eigenvalue λ^*. Let this first eigenvector, eigenvalue pair be denoted (v_1, γ_1).

The next smallest eigenvalue may be found in a similar fashion. Namely by solving

$$\begin{aligned} \min_x \quad & x'\Sigma x \\ & ||x||^2 = 1 \\ & v_1'x = 0. \end{aligned}$$

That is, find the eigenvector with smallest eigenvalue perpendicular to v_1. The first portion of this sentence needs to be validated; i.e., we are not sure that the solution here is in fact an eigenvector. We proceed with the Lagrangian as before, with

$$\mathcal{L}(x, \lambda) = x'\Sigma x - \lambda_1 \left(||x||^2 - 1 \right) - \lambda_2 (v_1'x).$$

The gradient is then

$$\nabla_x \mathcal{L}(x, \lambda) = 2\Sigma x - 2\lambda_1 x - \lambda_2 v_1,$$

and a pair, (x^*, λ^*) satisfying KKT gives $\nabla_x \mathcal{L}(x, \lambda) = 0$. Premultiplying the gradient of the Lagrangian by v_1 and dividing through by 2, we have

$$v_1'\Sigma x^* - \lambda_1^* v_1' x^* - \frac{1}{2}\lambda_2^* ||v_1||^2 = 0.$$

By feasibility of the current problem, $v_1'x^* = 0$. Feasibility in the preceding problem gave $||v_1||^2 = 1$ as well. The first order condition on $\mathcal{L}$ gives that $\lambda_2^* = 0$:

$$\begin{aligned} v_1'\Sigma x^* - \frac{1}{2}\lambda_2^* &= 0 \\ \gamma_1 v_1' x^* - \frac{1}{2}\lambda_2^* &= 0 \\ \frac{1}{2}\lambda_2^* &= 0 \end{aligned}$$

Hence we have

$$\Sigma x^* - \lambda_1^* x^* = 0$$

as before, and x^* is an eigenvector with associated eigenvalue λ_1^*. Necessarily, $\lambda_1^* \geq \gamma_1$.

The procedure may be continued to find all of the eigenvalues of Σ.

We next return to an issue not presented in our original treatment of Generalized Least Squares. Namely, given a new observation, X_{N+1}, what is the best predictor $\hat{y}_{N+1}$?

Example 7.5.2. Consider the GLS assumptions as before,

$$Y = X\beta + \epsilon$$

with $Cov(\epsilon) = V$ for covariance matrix $V \in \mathbb{R}^{N\times N}$, and estimated $\hat{\beta}$ given by (4.49):

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}Y.$$

Let X_{N+1} be a new observation, and let

$$\hat{y}_{N+1} = w'Y$$

be a linear predictor for some w. To ensure this estimator is unbiased, we require

$$\mathbb{E}\left(\hat{y}_{N+1} - y_{N+1}\right) = 0,$$

which gives

$$\begin{aligned}\mathbb{E}\left(w'Y - y_{N+1}\right) &= \mathbb{E}\left(w'(X\beta+\epsilon) - (X_{N+1}\beta + \epsilon_{N+1})\right)\\ &= \left(w'X - X_{N+1}\right)\beta = 0.\end{aligned}$$

Since this is true for every w that gives an unbiased estimator, we require $w'X - X_{N+1} = 0$. Last, we note that the prediction error may be written

$$w'\epsilon - \epsilon_{N+1}.$$

Two issues arise: the first is that w is not uniquely determined; the second is that ϵ_{N+1} and ϵ are correlated under the generalized least squares assumption. Resolving the second issue is simply a matter of making an assumption about the covariance between ϵ_{N+1} and ϵ. We assume

$$Cov(\epsilon_{N+1}, \epsilon) = s \in \mathbb{R}^N.$$

Resolving the non-uniqueness of w may be handled by solving an optimization problem. The objective function we consider is the variance of the prediction error given by

$$\begin{aligned}Var(w'\epsilon - \epsilon_{N+1}) &= Var(w'\epsilon) - 2Cov(\epsilon_{N+1}, w'\epsilon) + Var(\epsilon_{N+1})\\ &= w'Vw - 2w's + \sigma^2\end{aligned}$$

where σ^2 is the variance of ϵ_{N+1}.

Noting that we may rescale the objective function and omit the constant σ^2, the constrained optimization problem we consider is

$$\begin{aligned}\min_w \quad & \frac{1}{2}w'Vw - w's\\ & X'w = X'_{N+1}\end{aligned}$$

whose Lagrangian is

$$\mathcal{L}(w,\lambda) = \frac{1}{2}w'Vw - w's - \lambda'(X'w - X'_{N+1})$$

with gradient in w

$$\nabla_w \mathcal{L}(w,\lambda) = Vw - s - X\lambda$$

which, as before, we set equal to zero.

Coupling this statement about the gradient of $\mathcal{L}$ with the constraint $X'w = X'_{N+1}$ gives a system of equations as in (7.30). Here we have

$$\begin{pmatrix} V & -X \\ X' & 0 \end{pmatrix} \begin{pmatrix} w \\ \lambda \end{pmatrix} = \begin{pmatrix} s \\ X'_{N+1} \end{pmatrix}.$$

Next, we note a result using something called a Schur complement that identifies the inverse of a partitioned matrix. Let

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

with A invertible. Then M has inverse

$$\begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix} \tag{7.39}$$

where

$$(M/A) = D - CA^{-1}B \tag{7.40}$$

is called the *Schur complement* of M with respect to A. We leave the verification to the reader.

Applying this result to the system of equations at hand gives that

$$w^* = \left(V^{-1} - V^{-1}X(X'V^{-1}X)X'V^{-1}\right)s + \left(V^{-1}X(X'V^{-1}X)^{-1}\right)X'_{N+1}$$

minimizes the prediction error variance and yields an unbiased estimator.

For this w^* we have

$$\begin{aligned} \hat{y}_{N+1} &= {w^*}'Y \\ &= s'\left(V^{-1} - V^{-1}X(X'V^{-1}X)X'V^{-1}\right)Y + X_{N+1}\left((X'V^{-1}X)^{-1}X'V^{-1}\right)Y. \end{aligned}$$

Recalling that

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1}X'V^{-1}Y,$$

this reduces rather nicely to

$$\begin{aligned} s'V^{-1}Y - V^{-1}X\hat{\beta} + X_{N+1}\hat{\beta} \\ X_{N+1}\hat{\beta} + s'V^{-1}(Y - X\hat{\beta}). \end{aligned}$$

Notice that V and s are not specified by the data necessarily. As a matter of practice, these parameters are usually modeled with a small set of parameters themselves.

As a final example using a quadratic objective function, we consider the problem of determining a portfolio with minimum variance whose CAPM β is one. We shall explore this topic in much further depth in subsequent work when we introduce mean-variance optimization more thoroughly.

Example 7.5.3. We know that for a vector of portfolio weights, $w \in \mathbb{R}^N$ and random return vector $r \in \mathbb{R}^N$, the variance of the portfolio is given by

$$Var(w'r) = w'\Sigma w$$

where $\Sigma = Cov(r)$.

We may identify the portfolio with smallest variance whose CAPM β is one by solving

$$\begin{aligned} \min_w \quad & \frac{1}{2} w'\Sigma w \\ & \beta' w = 1 \end{aligned}$$

where $\beta \in \mathbb{R}^N$ is the vector of individual asset β's; i.e., β_i is the CAPM β of asset i. We leave to the reader to prove that the portfolio β is in fact the weighted sum of asset β's.

The solution to the optimization problem is readily obtained by considering the Lagrangian,

$$\mathcal{L}(w, \lambda) = \frac{1}{2} w'\Sigma w - \lambda(\beta' w - 1)$$

with gradient in w

$$\nabla_w \mathcal{L}(w, \lambda) = \Sigma w - \lambda \beta.$$

The optimal w^* satisfies $\nabla_w \mathcal{L}^* = 0$, giving

$$w^* = \lambda \Sigma^{-1} \beta.$$

For w^* to be feasible, we require $\beta' w^* = 1$, so that

$$\begin{aligned} \lambda^* \beta' \Sigma^{-1} \beta &= 1 \\ \lambda^* &= \left(\beta' \Sigma^{-1} \beta\right)^{-1}. \end{aligned}$$

Putting this all together,

$$w^* = \frac{\Sigma^{-1} \beta}{\beta' \Sigma^{-1} \beta}.$$

The weight w^* may be identified further. Let the weight used to construct the market returns be w_m and suppose that the returns in question, r, contain all components of the market index.

We know that for each asset, r_i, β_i is given by

$$\beta_i = \frac{Cov(r_i, r_m)}{Var(r_m)},$$

The covariance between any particular asset return, r_i, and the market return $r_m = w_m' r$ is given, using e_i as the vector with a one in the ith component and zeros elsewhere, by

$$\begin{aligned} Cov(r_i, r_m) &= Cov(e_i' r, w_m' r) \\ &= e_i' \Sigma w_m \\ &= \Sigma_i w_m \end{aligned}$$

where, necessarily, Σ_i is the ith row of the covariance matrix, Σ. The variance of the market return is clearly given by $w_m' \Sigma w_m$. As a result,

$$\beta = \frac{\Sigma w_m}{w_m' \Sigma w_m}.$$

Returning to w^*, we have

$$\begin{aligned} w^* &= \frac{\Sigma^{-1}\beta}{\beta' \Sigma^{-1}\beta} \\ &= \frac{\Sigma^{-1}\Sigma w_m}{w_m' \Sigma \Sigma^{-1} \Sigma w_m} \frac{1}{w_m' \Sigma w_m} (w_m' \Sigma w_m)^2 \\ &= w_m. \end{aligned}$$

So that the minimum variance portfolio with unit β is the market portfolio.

Finally, we develop a linear programming problem related to regression. Here, we introduce the tilting function, $\rho_\tau(\cdot)$, parameterized in τ. We will see that this function is related to identifying the sample τ quantile of a distribution and will use this fact to develop what is called quantile regression. In both the case of quantile value estimation and quantile regression, we will show that the problem may be written as a linear programming problem in standard form.

Example 7.5.4. Given the sample $\{y_t\}_{t=1}^N$, the sample τ quantile may be obtained by solving

$$\min_q \sum_{t=1}^{N} \rho_\tau(y_t - q) \tag{7.41}$$

where ρ_τ is the tilting function defined by

$$\rho_\tau(x) = \tau \max(x, 0) + (1 - \tau) \max(-x, 0). \tag{7.42}$$

Equivalently,

$$\rho_\tau(x) = \begin{cases} \tau x & \text{if } x \geq 0 \\ (\tau - 1)x & \text{if } x < 0. \end{cases}$$

Notice that ρ_τ takes nonnegative values.

We justify that the sample quantile, $\hat{q}_\tau$, may be obtained via ρ_τ, through a proof using the population quantile, $q_\tau = F^{-1}(\tau)$, where F is the cumulative distribution function of the random variable Y.

For any real valued u, we have that

$$\begin{aligned}
\mathbb{E}\left(\rho_\tau(Y-u)\right) &= \int_{-\infty}^{\infty} \rho_\tau(y-u)dF(y) \\
&= \int_{-\infty}^{u} (\tau-1)(y-u)dF(y) + \int_{u}^{\infty} \tau y dF(y) \\
&= (\tau-1)\int_{-\infty}^{u} y dF(y) \\
&\quad -(\tau-1)u\int_{-\infty}^{u} dF(y) \\
&\quad +\tau\int_{u}^{\infty} y dF(x) \\
&\quad -\tau u\int_{u}^{\infty} dF(x).
\end{aligned}$$

This expectation may be written as a function of u parameterized by τ as $L_\tau(u)$. We are interested in minimizing this expected loss function and so take the first derivative in u. Doing so, we see that

$$L'_\tau(u) = (1-\tau)\int_{-\infty}^{u} dF(y) - \tau\int_{u}^{\infty} dF(y),$$

which may further be simplified as

$$L'_\tau(u) = (1-\tau)F(u) - \tau(1-F(u)) = F(u) - \tau,$$

and hence the function attains a minimum at $u = F^{-1}(\tau)$ as originally claimed.

Having established the quantile property of ρ_τ, we next write (7.41) as an LP problem. For

$$\rho_\tau(y_t - u) = \tau(y_t-u)_+ + (1-\tau)(u-y_t)_+,$$

we construct variables, $z_{t,+}$ and $z_{t,-}$ such that each variable is nonnegative, and in addition,

$$\begin{aligned}
z_{t,+} &\geq y_t - u \\
z_{t,-} &\geq u - y_t.
\end{aligned}$$

In these new variables, we may write (7.41) as

$$\begin{aligned}
\min_{u,z_+,z_-} \quad & \sum_{t=1}^{N} \tau z_{t,+} + (1-\tau)z_{t,-} \\
& z_+ + u \geq Y \\
& z_- - u \geq -Y \\
& z_+ \geq 0 \\
& z_- \geq 0,
\end{aligned} \tag{7.43}$$

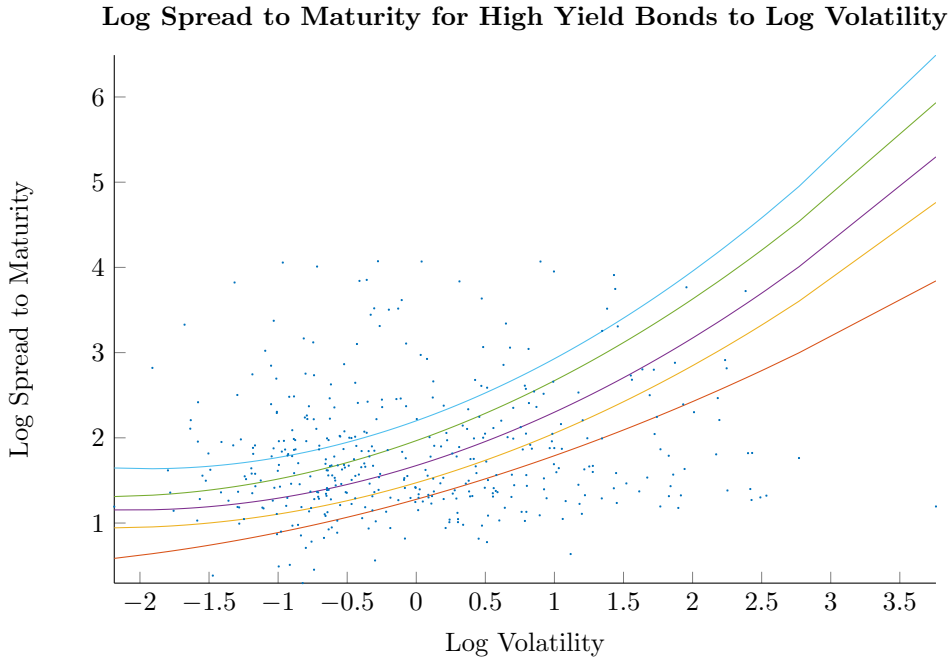

Figure 7.2: Plot of various quantile regression levels when regressing log spreads on log realized equity volatility and its square, including an intercept term. Quantile levels range from (bottom to top): 10%, 25%, 50%,75%, and 90%.

where $Y = (y_1, \ldots, y_N)'$. From here, the additional variables and constraints needed to put the problem in standard form should be clear. We note, too, that these particular constraints may be refined a bit more by simply requiring requiring to $z_+ - z_- = y - u$. We have left the framework above for clarity of exposition, but encourage the interested reader to reframe the problem in this manner.

Based on the above, we define quantile regression as

$$\min_{\beta} \sum_{t=1}^{N} \rho_\tau(y_t - \beta' x_t), \tag{7.44}$$

in a manner similar to (4.11). The only difference from previous work in terms of the optimization problem that is implied being the change of loss function from the L^2 norm to $\rho_\tau(\cdot)$. As this new problem is linear in β, we may write (7.44) as an LP problem in standard form as in (7.41). We note that an intercept term is required.

An example of the results from a single input variable are give in Figure 7.5.4. Here we have used log volatility and its square to estimate quantile regressors

for various quantiles. The boundary curves represent the 0.10 and 0.90 quantile regressions (from bottom to top). Notice that the shape of the regressions for these values differ slightly from what might be expected in a confidence interval curve from ordinary least squares as given by (4.34).

Example 7.5.5. Using the same framework as the previous example, we next consider the Huber loss based regression,

$$\min_{\beta} \sum_{t=1}^{N} \rho_C(y_t - \beta' x_t) \tag{7.45}$$

for some set of observations $\{x_t\}_{t=1}^N$, and with $\rho_C(\cdot)$ the Huber loss with parameter C, defined by

$$\rho_C(x) = \begin{cases} |x|^2 & \text{for } |x| \leq C \\ 2C|x| - C^2 & \text{for } |x| > C. \end{cases} \tag{7.46}$$

With a bit of examination, it becomes clear that the Huber loss considers square losses (L^2) inside a band of width $2C$, and linear (L^1) outside this band, with some care taken to assure continuity at the boundary. We proceed by formulating (7.45) as a constrained quadratic programming problem.

To do so, we again introduce auxiliary variables. In the present case, we seek variables, z_t and u_t, such that $|y_t - \beta' x_t| = z_t + u_t$ and z_t takes up that portion of the absolute value up to C and u_t the remainder. This designation is actually sufficient to to proceed. Consider

$$\begin{aligned} \min_{z,u,\beta} \quad & ||z||^2 + 2C1'u \\ & y_t - \beta' x_t \leq z_t + u_t \text{ for } 1 \leq t \leq N \\ & -z_t - u_t \leq y_t - \beta' x_t \text{ for } 1 \leq t \leq N \\ & 0 \leq z \leq C \\ & 0 \leq u \end{aligned} \tag{7.47}$$

Noting the identification,

$$|y_t - \beta' x_t| = \min(|y_t - \beta' x_t|, C) + \max(|y_t - \beta' x_t| - C, 0)$$

and that at the solution to (7.47), we will have equality in the first two conditions gives

$$|y_t - \beta' x_t| = z_t + u_t,$$

and the correspondence between (7.45) and (7.47) follows.

Clearly as $C \to \infty$, the above loss function tends towards the standard least squares problem considered extensively already. For intermediate, C, however, we see that the Huber loss based regression reduces the overall impact (and hence, the fitting to) outliers in $\{y_t\}_t$ as determined by C. The question of how to determine such C is left to a subsequent chapter.

Exercises

1. We say that a set C is a cone if for $\alpha > 0$, $\alpha c \in C$ for every $c \in C$. Prove that $\mathcal{F}(x)$ as in (7.9) is a cone.

2. Prove that if x^* is a regular point, with (x^*, λ^*) satisfying (7.11), then λ^* is unique.

3. Let (x^*, λ^*) satisfy (7.11). Show that $\mathcal{L}(x^*, \lambda^*) = f^*$.

4. Prove that the set
$$K = \{x | c_i(x) \geq k_i\}$$
is convex when each c_i is concave and leave the proof to the reader.

5. Verify (7.39), that a partitioned matrix
$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$
with A invertible has inverse
$$\begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix}$$
where
$$(M/A) = D - CA^{-1}B.$$

6. Rigorously prove that each optimal Lagrange multiplier in the active set is exactly the sensitivity of the objective function value at the optimal solution when a small change in the constraint value is made. That is, modify the proof shown in the text by using little-o notation as in (5.17).

7. What is the impact of a nonzero Lagrange multiplier when the objective function is multiplied by a scalar $c > 0$? What is the impact of multiplying a single constraint function, $c_i(x)$ by c (leaving the objective function fixed)? (This result coupled with the previous exercise shows that while Lagrange multipliers may be viewed as sensitivities, they are also scale dependent; i.e., they are not necessarily unitless.)

8. Let $f(x) \in \mathbb{C}^2$.

   (a) Using a degree two Taylor expansion, write down a quadratic approximation $f_q(x + \delta) = h + g'\delta + \frac{1}{2}\delta' G\delta$ in terms of $f(x)$, the gradient $\nabla f(x)$, and the Hessian, $\nabla^2 f(x)$.

   (b) Let $q(\delta) = f_q(x + \delta)$. Assuming the matrix G you found above is always positive definite, find an equation for δ^*, the minimizer of $q(\delta)$.

   (c) How does your δ^* above relate to the Newton step we derived in Newton's method?

(d) Next, we generalize the work we did to obtain what we called the K-matrix in a quadratic programming problem with equality constraints. Consider

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

i. Let x^k be a feasible point. Define $x^{k+1} = x^k + \delta$ for some δ. Write down a condition that must hold for δ for x^{k+1} to remain feasible.

ii. Let $q_k(\delta) = f_q(x^k + \delta)$, where $f_q(x^k + \delta)$ is the approximation of $f(x^k + \delta)$ using a second degree Taylor expansion as earlier in the exam. Consider the approximate problem

$$\begin{aligned} \min_{\delta} \quad & q_k(\delta) \\ \text{subject to} \quad & \tilde{A}\delta = \tilde{b} \end{aligned}$$

where $\tilde{A}$ and $\tilde{b}$ are the conditions needed to keep x^{k+1} feasible as determined above. Write down the Lagrangian for this new problem in terms of $f(x^k)$, the gradient $\nabla f(x^k)$, the Hessian, $\nabla^2 f(x^k)$, $\tilde{A}$, and $\tilde{b}$.

iii. Write down the KKT conditions for the new problem in δ.

iv. Combine the above conditions into a matrix equation in a similar manner as we did for the equality constrained quadratic programming problem. You are looking for a matrix K and a vector κ such that

$$K \begin{pmatrix} \delta \\ \lambda \end{pmatrix} = \kappa$$

where λ is the vector of Lagrange multipliers.

9. Write (7.32) in standard form.

10. Write (7.44) in standard form.

11. Complete the proof of the Strong Duality Theorem by showing that if the primal (dual) is unbounded, then the dual (primal) is infeasible.

12. Let $f(x) = \sum_{i=1}^{N} x_i \ln x_i$ for $x \in \mathbb{R}^N$, and $x \geq 0$.

(a) What is $\nabla f(x)$?

(b) Consider the problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & \tfrac{1}{N} \textstyle\sum_{i=1}^{N} x_i = \mu \end{aligned}$$

i. Write down the KKT conditions for this problem.

ii. Using your KKT equations, solve for x_i in terms of λ, the Lagrange multiplier used in the KKT condition.

13. Convince yourself that the condition that x be a regular point coincides with the Jacobian of

$$F(x) = \begin{pmatrix} c_1(x) \\ \vdots \\ c_m(x) \end{pmatrix}$$

being invertible for $\mathcal{A}(x) = \{1, \ldots, m\}$.

Chapter 8

Mean-Variance Optimization

We arrive finally at mean-variance optimization, a hallmark of modern portfolio theory, having established the statistical and optimization framework needed for a fairly full treatment. In this chapter, we will develop the standard framing for the problem and identify some of its most salient features such as the efficient frontier in mean-variance space. We will also prove a theorem – which, if true, would greatly reduce portfolio management complexity – called the mutual fund separation theorem. We will also establish the capital market line of mean-variance optimization and in so doing we will connect mean-variance optimization to the Capital Asset Pricing Model directly.

From the standard problem, we look at several variants, including Sharpe Ratio optimization, portfolio updates in a mean-variance optimal setting, and proving equivalency of maximizing returns under a quadratic constraint.

8.1 The Standard Problem

Merton's mean-variance optimization [22] equates investor risk with portfolio variance. That is, for stochastic returns, $r \in \mathbb{R}^N$, and fixed portfolio weights, $w \in \mathbb{R}^N$, the variance of the portfolio is given by $Var(w'r) = w'\Sigma w$ with $\Sigma = Cov(r)$, and this quantity completely determines the risk an investor is taking. In its original formulation, minimizing variance is paired with a requirement of some expected return. Letting $\mu = \mathbb{E}(r)$, the linearity of expectation gives that the expected portfolio return is $\mu' w$.

For the present treatment, we will assume that r is linearly independent as defined previously; that is, Σ is positive definite.

We immediately arrive at Merton's original constrained optimization prob-

lem:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w' \Sigma w \\ & \mu' w = \mu^* \\ & 1' w = 1. \end{aligned} \tag{8.1}$$

Plainly, we see that we are minimizing risk (read variance) subject to an investor-specified required return, μ^*, being achieved in expectation, and portfolio weights summing to one. Looking at the form of the objective and constraints, the optimization problem is classified as a *quadraitc programming problem with linear constraints.* From our work in the previous chapter, we know that (8.1) has a unique solution obtained by looking at the Karush-Kuhn-Tucker conditions (7.11), which are both necessary and sufficient.

In particular, the vanishing gradient of the Lagrangian and two feasibility constraints give a system of equations in w, λ_1, and λ_2.

$$\Sigma w - \lambda_1 \mu - \lambda_2 1 = 0 \tag{8.2}$$

$$\mu' w = \mu^* \tag{8.3}$$

$$1' w = 1. \tag{8.4}$$

A system of equations in just λ_1 and λ_2 may be obtained by premultiplying (8.2) by μ' and $1'$:

$$\begin{aligned} b\lambda_1 + a\lambda_2 &= \mu^* \\ a\lambda_1 + c\lambda_2 &= 1, \end{aligned} \tag{8.5}$$

for choices of a, b, and c which we leave to the reader. The resulting matrix equation is simply

$$\begin{pmatrix} b & a \\ a & c \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \mu^* \\ 1 \end{pmatrix}, \tag{8.6}$$

which gives

$$\begin{aligned} \lambda_1^* &= \frac{c\mu^* - a}{d} \\ \lambda_2^* &= \frac{b - a\mu^*}{d}. \end{aligned}$$

where $d = bc - a^2$ is the determinant of (8.6). Due to the positive-definiteness of Σ^{-1}, d is positive (proof left to the reader).

From here, the optimal weights, w^*, are given by inverting Σ in (8.2) and using the analytic solutions for λ_1^* and λ_2^*,

$$w^* = \frac{c\mu^* - a}{d} \Sigma^{-1} \mu + \frac{b - a\mu^*}{d} \Sigma^{-1} 1. \tag{8.7}$$

The variance of the portfolio is calculated by premultiplying (8.2) by $w^{*\prime}$. This gives

$$\begin{aligned} Var(w^{*\prime} r) &= \lambda_1^* \mu^* + \lambda_2^* \\ &= \frac{c\mu^{*2} - 2a\mu^* + b}{d}. \end{aligned} \tag{8.8}$$

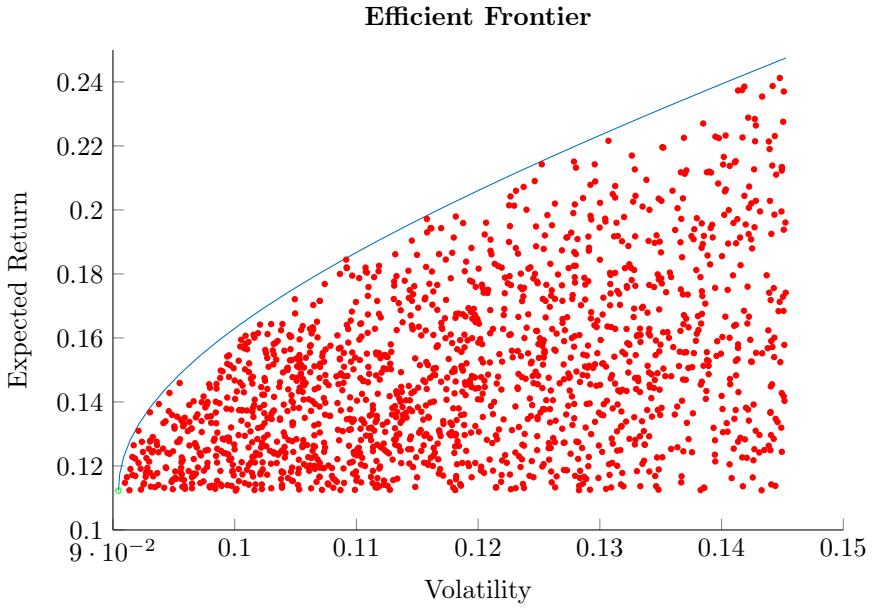

Figure 8.1: Plot of a Mean-Variance efficient frontier using a sample covariance matrix and mean return from trailing return data. A random sampling of portfolios is shown bounded by the efficient frontier, and the global mean-variance optimal portfolio is shown on the far left edge in green.

We have, finally, that the variance of the optimal portfolio of (8.1) is a quadratic function of μ^*, completely determined by Σ^{-1} and μ. We now arrive at our first comment about the variety of specifications that the problem allows. That is, (8.1) provides a curve of solutions, parameterized by μ^*, tracing out a parabola in volatility-return space. As of yet, we do not have a preference or a taxonomy for points on this curve, however. Looking at (8.8) gives us a preliminary insight. Namely, we may identify the minimum variance portfolio by taking the derivative with respect to μ^*, giving

$$\frac{d\sigma_{w^*}^2}{d\mu^*} = \frac{2c\mu^* - 2a}{d},$$

where $\sigma_{w^*}^2$ denotes the variance of the portfolio with weights w^*.

Doing so, we see that the variance function is minimized when

$$\mu^* = \frac{a}{c}. \tag{8.9}$$

We will denote this mean by $\underline{\mu}$. We call the portfolio with minimum feasible variance the *global minimum variance optimal portfolio*, or GMVO, and as

we shall see subsequently, the performance of these portfolios in our sample period is surprising relative to what the theory prescribes. Recalling our previous results on the variance anomaly, however, will give the astute reader some foreshadowing.

We may plot various portfolios in volatility-return space as in Figure 8.1. Notice that the curve of solutions parameterized by μ^* – the so-called *efficient frontier* – provides a boundary for possible mean-variance optimal portfolios in volatility-return space, and, as (8.8) implies, the boundary is quadratic in expected returns.

Under the assumptions of the model (namely that risk is variance and reward is expected return), investors should, for a given level of risk tolerance (read variance), require as much reward (read expected return) as possible. Geometrically, this means that investors will draw a line vertically until they hit the efficient frontier for any specified level of risk, going up until they maximize return. Similarly, for a given level of required return, investors will draw a line horizontally until they arrive at the minimal variance portfolio with that same expected return.

From (8.8) we have that, for a given level of expected return, μ^*,

$$\sigma = \sqrt{\frac{c\mu^{*2} - 2a\mu^* + b}{d}}.$$

Writing μ^* as a function of σ, we have

$$\begin{aligned} d\sigma^2 &= c\mu^{*2} - 2a\mu^* + b \\ 0 &= c\mu^{*2} - 2a\mu^* + (b - d\sigma^2), \end{aligned}$$

which by the quadratic equation gives

$$\begin{aligned} \mu^* &= \frac{a}{c} \pm \frac{\sqrt{a^2 - bc + dc\sigma^2}}{c} \\ &= \frac{a}{c} \pm \frac{\sqrt{d(c\sigma^2 - 1)}}{c} \end{aligned}$$

since $d = bc - a^2$. Now, since $\underline{\mu} = \frac{a}{c}$ is the return to the minimum variance portfolio, we have that the efficient frontier is given by

$$\mu^* = \underline{\mu} + \frac{1}{c}\sqrt{d(c\sigma^2 - 1)}. \tag{8.10}$$

8.1.1 Mutual Fund Separation Theorem

The theory also implies the so-called *mutual fund separation theorem*. At a high level, we have reduced every tradable position and portfolio to a point in mean-variance space[1]. That is, a pair of numbers completely describes any portfolio

[1]We will move (and indeed already have moved) interchangeably between the labels of mean-variance and volatility-return space. The former is identified from the original optimization problem, while the latter is useful in understanding the geometry of the space of solutions. Related, it is common in practice to refer to volatility-return space as *risk-return space*.

in the model. Given this reduction of dimensions in utility space, one may ask whether we may describe portfolios themselves with fewer parameters; i.e., is there a subset of portfolios which can yield any mean-variance combination an investor desires? The answer is yes. And perhaps not surprisingly you only need two.

For a point on the efficient frontier, we may rewrite (8.7) as

$$w^* = \frac{1}{d}\left(b\Sigma^{-1}1 - a\Sigma^{-1}\mu\right) + \frac{1}{d}\left(c\Sigma^{-1}\mu - a\Sigma^{-1}1\right)\mu^*,$$

or, simply

$$w^* = \alpha + \gamma\mu^*. \tag{8.11}$$

Notice that in this formulation, the only free parameter in determining a position on the efficient frontier (perhaps as expected) is the expected return, μ^*.

Next, suppose we have two points on the efficient frontier, (σ_1, μ_1) and (σ_2, μ_2), or, equivalently, two portfolios, Π_1 and Π_2, on the efficient frontier given by

$$\begin{aligned} w_1 &= \alpha + \gamma\mu_1 \\ w_2 &= \alpha + \gamma\mu_2. \end{aligned}$$

Consider a third portfolio, w_3, lying on the efficient frontier with expected return, μ_3. By the linearity of expectation, any portfolio constructed as a linear combination of Π_1 and Π_2 satisfying

$$a_1\mu_1 + a_2\mu_2 = \mu_3$$

will have expected return μ_3. By (8.11), to lie on the efficient frontier, we must also have

$$a_1 w_1 + a_2 w_2 = \alpha + \gamma\mu_3.$$

This implies, based on the parameterization of w_1 and w_2, that

$$a_1 + a_2 = 1.$$

A system of equations to identify w_3 as a linear combination of Π_1 and Π_2 emerges as

$$\begin{pmatrix} \mu_1 & \mu_2 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \mu_3 \\ 1 \end{pmatrix}, \tag{8.12}$$

which yields solution

$$\begin{aligned} a_1 &= \frac{\mu_2 - \mu_3}{\mu_2 - \mu_1} \\ a_2 &= \frac{\mu_3 - \mu_1}{\mu_2 - \mu_1}, \end{aligned}$$

or

$$w_3 = \frac{\mu_2 - \mu_3}{\mu_2 - \mu_1}w_1 + \frac{\mu_3 - \mu_1}{\mu_2 - \mu_1}w_2. \tag{8.13}$$

This is farily remarkable.

Under the assumptions of the model, then, any efficient portfolio may be constructed from just two other portfolios on the efficient frontier. The theorem gets its name by assuming that Π_1 and Π_2 are mutual funds through which all investors can express their mean-variance preferences. Inherent in the usefulness of this theory, however, is the stability of the covariance – an issue we have previously studied and in so doing have observed significant variation, especially in the market impact of dominant eigenvalues. That is, the theorem is interesting for its mathematical implications, but we do not have empirical evidence to pursue its application in practice.

A more thorough discussion of the application of the theory follows at the end of the chapter. First, we present another remarkable result, culminating in what is called the *Capital Market Line*.

8.1.2 The Capital Market Line

So far, we have assumed that all assets were risky. The introduction of a risk-free asset (read zero variance) has dramatic implications. In particular, we will see that in volatility-return space, the efficient frontier may be reduced to a line. Further, the previous mutual fund separation theorem may be formulated between the risk-free asset and a portfolio whose importance we will determine shortly. First, we outline a mean-variance optimization problem when one asset is risk-free.

As before, let μ and Σ be the expectation and covariance, respectively, of $r \in \mathbb{R}^N$, the stochastic vector of returns of N linearly independent risky assets. If we consider fully allocated portfolios in these N assets and the risk-free asset, we have, for $1'w = w^*$, r_f the return for the risk-free asset, and μ^* the expected return of the portfolio of these $N+1$ assets,

$$\begin{aligned} w^{*\prime}\mu + (1 - e'w^*)r_f &= \mu^* \\ w^{*\prime}(\mu - r_f) &= \mu^* - r_f. \end{aligned}$$

As before, we may set up a minimum variance portfolio optimization problem as

$$\begin{aligned} \min_w \quad & \frac{1}{2} w'\Sigma w \\ & (\mu - r_f)'w = \mu^* - r_f. \end{aligned} \tag{8.14}$$

Notice that we have not included the full allocation constraint since it is implied in the expected return requirement.

The Lagrangian of (8.14) is

$$\mathcal{L}(w, \lambda) = \frac{1}{2} w'\Sigma w - \lambda \left((\mu - r_f)'w - (\mu^* - r_f) \right), \tag{8.15}$$

which has gradient in w of

$$\nabla_w \mathcal{L}(w, \lambda) = \Sigma w - \lambda(\mu - r_f). \tag{8.16}$$

The gradient vanishes at

$$w^* = \lambda \Sigma^{-1}(\mu - r_f),$$

and λ^* satisfying the constraints of (8.14) is found by solving

$$w^{*\prime}(\mu - r_f) = \mu^* - r_f.$$

One may show, and the exercise is left to the reader, that the optimal risky weights are given by

$$w^* = (cr_f^2 - 2ar_f + b)^{-1}\Sigma^{-1}(\mu - r_f)(\mu^* - r_f) \tag{8.17}$$

for the same a, b, and c as in the preceding section.

The variance of the optimal portfolio in risky and risk-free assets for expected return μ^* is found to be

$$\sigma^2 = \frac{(\mu^* - r_f)^2}{cr_f^2 - 2ar_f + b}. \tag{8.18}$$

Assuming that $\underline{\mu} > r_f$, we have that

$$\mu^* = r_f + \sigma\sqrt{cr_f^2 - 2ar_f + b}. \tag{8.19}$$

Notice that, in contrast to (8.10), we now have a *linear* relationship between volatility and expected return. One of two cases may obtain: the line determined by (8.19) may dominate the efficient frontier, being tangent; or, this line is at times above and others below the efficient frontier. But, by construction, the answer is the former (so long as $\underline{\mu} > r_f$) since (8.14) has as a subset of feasible portfolios those considered in the original formulation (8.1). That is, for the same μ^*, the objective function of (8.14) is no bigger than that of (8.1) at the optimal solution. Therefore there exists a single portfolio on both the efficient frontier and the line specified by (8.19). We call this portfolio the *tangency portfolio* and the line connecting the risk-free asset and the tangency portfolio in volatility-return space the *Capital Market Line*.

Next we consider another mutual fund separation theorem in light of the Capital Market Line. In this iteration, we fix the two 'mutual funds' as the risk-free asset and the tangency portfolio, with weights for the latter denoted by w_T.

For a third portfolio lying on the efficient frontier with weights w_3, we show that its expected return, $\mu_3 = w_3'\mu$ may be obtained from a combination of the risk-free asset and the tangency portfolio, with the sum of the weights between these two being fully specified and summing to one. Based on the above, this gives that this particular combination lies on the Capital Market Line, reduces volatility as compared to the portfolio with weights w_3, and maintains the expected return μ_3.

Letting μ_T be the expected return of the tangency portfolio, we have by (8.17) that

$$w_T = \gamma_0(\mu_T - r_f)$$

with $\gamma_0 = (cr_f^2 - 2ar_f + b)^{-1}\Sigma^{-1}(\mu - r_f)$. This gives

$$\mu_T = r_f + (1'\gamma_0)^{-1}$$

since w_T is on the efficient frontier with sum of weights equal to one.

To relate the expected return of our third portfolio to the risk-free asset and the tangency portfolio, we require

$$a_0 r_f + a_1 \mu_T = \mu_3.$$

Combining this with the relationship in μ_T, we have

$$\begin{aligned} a_0 r_f + a_1(r_f + (1'\gamma_0)^{-1}) &= \mu_3 \\ (a_0 + a_1) r_f + a_1 (1'\gamma_0)^{-1} &= \mu_3. \end{aligned}$$

Requiring $a_0 + a_1 = 1$, this further reduces to

$$r_f + a_1(1'\gamma_0)^{-1} = \mu_3$$

so that

$$a_1 = (1'\gamma_0)(\mu_3 - r_f)$$

and

$$a_0 = 1 - (1'\gamma_0)(\mu_3 - r_f).$$

We conclude that the portfolio with weights a_0 and a_1 in the risk-free asset and tangency portfolio results in all of the features set out at the beginning of this construction. Namely, we have reduced volatility while maintaining expected returns, and the solution is to simply identify a mixture of two portfolios[2].

This is yet another truly remarkable result: in volatility-return space, simple combinations of the risk free asset and the tangency portfolio dominate all other portfolios – even those on the efficient frontier. This is clearly seen in Figure 8.1.2. The punchline of all of this work is that if all investors regard risk and reward purely in terms of mean and variance, then all analysis boils down to determining just what proportion of the tangency portfolio to choose[3].

We have previously noted caution with respect to reducing equities to their respective time series of returns; the current results extend this caution further. We postpone these thoughts for a bit longer, however.

The tangency portfolio may be identified by an equilibrium argument: namely, if all investors have the same utility, then the tangency portfolio must in fact be the market portfolio of the Capital Asset Pricing Model (4.1). Such an argument

[2]Of course, we have not yet outlined a procedure for identifying the weights of the tangency portfolio, focusing so far only on its defining characteristic.

[3]One additional assumption is that all investors must have the same allocation constraints; viz., the constraints to the optimization problem are identical for all investors. This of course does not obtain in practice.

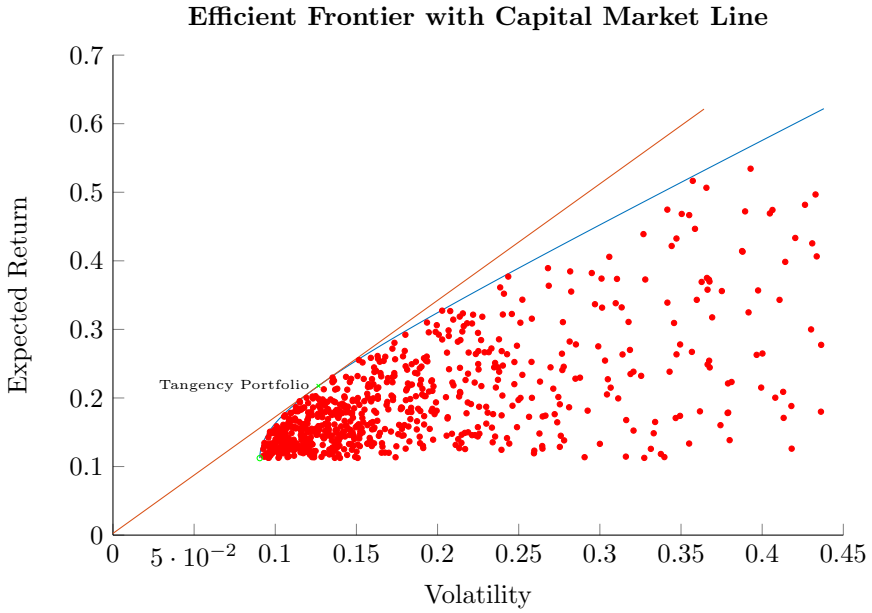

Figure 8.2: Plot of a Mean-Variance efficient frontier and Capital Market Line using a sample covariance matrix and mean return from trailing return data. A random sampling of portfolios is shown bounded by the efficient frontier, and both the global mean-variance optimal portfolio and tangent portfolio are shown. The Capital Market Line dominates the risky-asset frontier.

is a deviation from the approach of this book, however. Instead, we may prove that the tangency portfolio is the very same market portfolio by first assuming that the CAPM model holds and then making the weaker assumption that the market portfolio and tangency portfolio have the same expected return. These assumptions and the previously considered optimization problem identifying the minimum variance portfolio with $\beta = 1$ in (7.41) may be used to show that the market portfolio and tangency portfolio are one and the same.

8.1.3 A Geometric Interpretation and CAPM

As noted above, we may consider any portfolio as a point in volatility-return space, (σ, μ). The mutual fund separation theorem that gave rise to the Capital Market Line may be seen as a maximization of a particular slope. That is, we identified the Capital Market Line as the line emanating from the risk-free asset and going through the tangency portfolio. The slope of the Capital Market Line

$$s_T = \frac{\mu_T - r_f}{\sigma_T},$$

with (σ_T, μ_T) being the mean return and volatility of the tangency portfolio, then, is the maximum slope possible since it is tangent to the efficient frontier.

The slope for any other choice (σ_0, μ_0),

$$s_0 = \frac{\mu_0 - r_f}{\sigma_0}, \tag{8.20}$$

is called the *Sharpe Ratio* for that portfolio. The results above give a clear indication of its importance in the theory; viz., the Capital Market Line focuses on the change in expected return per change in volatility. We have already seen this ratio when considering the low volatility anomaly, however. There we focused on return-per-vol-point to identify attractiveness across deciles of *ex ante* volatility[4]. Here we continue with the geometric focus.

Motivated by these observations, we would like to, for any portfolio, Π, identify $\frac{\partial \mathbb{E}(r_\Pi)}{\partial \sigma_\Pi}$. Our previous encounters with CAPM and now the Capital Market Line give some indication of a path forward on this question. In particular, we will work in volatility-return space, and consider convex combinations of Π and the tangency portfolio. Some calculus will yield that $\frac{\partial \mathbb{E}(r_\Pi)}{\partial \sigma_\Pi}$ is exactly determined by the Capital Asset Pricing Model.

Proof. Let Π be identified by (σ_0, μ_0) in volatility-return space and let the tangency portfolio be given by Π_T, with volatility, expected return pair (σ_T, μ_T). Let the stochastic return of Π and Π_T be given by r_Π and r_T, respectively. A portfolio consisting of a convex combination of Π and Π_T is given by

$$\Pi_\theta = \theta\Pi + (1-\theta)\Pi_T,$$

with stochastic returns

$$r_\theta = \theta r_\Pi + (1-\theta) r_T.$$

Clearly at $\theta = 1$, $r_\theta = r_\Pi$ and at $\theta = 0$ we have the return of the tangency portfolio. Similarly

$$\left.\frac{\partial \mathbb{E}(r_\theta)}{\partial \sigma_\theta}\right|_{\theta=0} = \frac{\partial \mathbb{E}(r_T)}{\partial \sigma_T}. \tag{8.21}$$

This final partial is known based upon our work above, however, as this coincides with the slope of the Capital Market Line; i.e.,

$$\frac{\partial \mathbb{E}(r_T)}{\partial \sigma_T} = \frac{\mu_T - r_f}{\sigma_T}. \tag{8.22}$$

We proceed in identifying $\frac{\partial \mathbb{E}(r_\theta)}{\partial \sigma_\theta}$ by making use of the chain rule. In particular, we have that

$$\frac{\partial \mathbb{E}(r_\theta)}{\partial \sigma_\theta} = \frac{\partial \mathbb{E}(r_\theta)}{\partial \theta}\frac{\partial \theta}{\partial \sigma_\theta} = \frac{\partial \mathbb{E}(r_\theta)}{\partial \theta}\left(\frac{\partial \sigma_\theta}{\partial \theta}\right)^{-1}.$$

[4] A minor quibble may be made. In the previous analysis, we omitted the risk free rate, r_f. This ratio of return (as opposed to excess return) to volatility is oftentimes referred to as the *information ratio*.

That is, we are left to calculate

$$\frac{\partial \mathbb{E}(r_\theta)}{\partial \theta}$$

and

$$\frac{\partial \sigma_\theta}{\partial \theta},$$

specifically at $\theta = 0$.

By the linearity of expectation,

$$\mathbb{E}(r_\theta) \quad = \quad \theta\mu_0 + (1-\theta)\mu_T,$$

and so,

$$\left.\frac{\partial \mathbb{E}(r_\theta)}{\partial \theta}\right|_{\theta=0} = \mu_0 - \mu_T. \tag{8.23}$$

The volatility of Π_θ is similarly derived. We begin by finding the variance,

$$\begin{aligned} Var(r_\theta) &= Var(\theta r + (1-\theta)r_T) \\ &= \theta^2\sigma_0^2 + 2\theta(1-\theta)\sigma_{0,T} + (1-\theta)^2\sigma_T^2, \end{aligned}$$

where $\sigma_{0,T} = Cov(r_\Pi, r_T)$. One may show from here that

$$\frac{\partial \sigma_\theta}{\partial \theta} = \frac{\theta\sigma_0^2 + (1-2\theta)\sigma_{0,T} - (1-\theta)\sigma_T^2}{\sigma_\theta} \tag{8.24}$$

so that, at $\theta = 0$, we get

$$\left.\frac{\partial \sigma_\theta}{\partial \theta}\right|_{\theta=0} = \frac{\sigma_{0,T} - \sigma_T^2}{\sigma_T}. \tag{8.25}$$

Finally, we have that

$$\left.\frac{\partial \mathbb{E}(r_\theta)}{\partial \sigma_\theta}\right|_{\theta=0} = (\mu_0 - \mu_T) \cdot \left(\frac{\sigma_T}{\sigma_{0,T} - \sigma_T^2}\right).$$

And, as already mentioned above, the left hand side is the slope of the Capital Market Line, giving an equation relating μ_0 and the covariance of $\sigma_{0,T}$ to the tangency portfolio in volatility-return space:

$$\frac{\mu_T - r_f}{\sigma_T} = (\mu_0 - \mu_T) \cdot \left(\frac{\sigma_T}{\sigma_{0,T} - \sigma_T^2}\right). \tag{8.26}$$

Rearranging terms in the above yields the very recognizable relationship

$$\mu_0 - r_f = \frac{\sigma_{0,T}}{\sigma_T^2}(\mu_T - r_f), \tag{8.27}$$

which is just the expected value of the Capital Asset Pricing Model

$$r_\Pi - r_f = \beta_\Pi(r_T - r_f) + \epsilon$$

with the tangency portfolio taking the place of the market portfolio and under the very mild assumption that ϵ is idiosyncratic with zero expectation. Notice that, in particular, we have not assumed any distributional assumptions for the idiosyncratic ϵ. □

In the case that the CAPM model holds with these reduced assumptions on ϵ and the market portfolio is put in place of the tangency portfolio as prescribed, then we must have, considering the tangency portfolio as Π,

$$\mu_T - r_f = \beta_T(\mu_m - r_f).$$

Now, if $\mu_T = \mu_m$, then clearly $\beta_T = 1$. Based on our previous results identifying the market portfolio with the minimum variance, fully allocated portfolio with $\beta = 1$, this implies the market portfolio and tangency portfolio must be the same if the optimization under consideration is performed over all assets in the market.

In the simplest case, then, we may determine the tangency portfolio directly. However, two immediate variations come to mind. First, we may consider optimizations over some subset of securities. Second, we may extend (8.1) or (8.14) to include general linear constraints. In this case, much of the above work still obtains – in particular, a tangency portfolio exists under the same conditions as above – but the identification of the tangency portfolio and market portfolio is restricted to the fairly reduced example just presented. In addition, we will require some new machinery to determine the tangency portfolio from an optimization problem.

We consider the case of identifying the tangency portfolio with general constraints next.

8.2 Maximizing the Sharpe Ratio

Based on the preceding section, the importance of the Sharpe Ratio (8.20) is evident. While the original analysis only included an expected return constraint in the derivation, we are now interested in the general case

$$\begin{aligned} \max_w \quad & \frac{w'\mu - r_f}{\sqrt{w'\Sigma w}} \\ & Aw = b \\ & Cw \geq d. \end{aligned} \tag{8.28}$$

Additionally, for reasons discussed previously, we will require that there exists a feasible solution with $\mu' w > r_f$.

The above formulation does not fit cleanly into any of the optimization problems we have yet considered. To obtain a solution, we will need to modify the objective function through the introduction of some auxiliary variables. We will show that this new problem is equivalent to the original and its solution

will be apparent. For ease of exposition, we first consider the simpler problem

$$\begin{aligned} \max_{w} \quad & \frac{w'\mu - r_f}{\sqrt{w'\Sigma w}} \\ & 1'w = 1. \end{aligned} \tag{8.29}$$

Again, we assume that there is a feasible solution that outperforms the risk-free asset in expected returns.

We will show that we may rewrite (8.29) as

$$\begin{aligned} \min_{(y,\kappa)} \quad & y'\Sigma y \\ & (\mu - r_f)'y = 1 \\ & (y,\kappa) \in \chi^+ \end{aligned} \tag{8.30}$$

where y is seen to have the same dimension as w, and the new variable, κ is a real valued scalar defined through the sets

$$\chi = \{w : 1'w = 1\}$$

and

$$\chi^+ = \left\{(y,\kappa) : \kappa > 0, \frac{y}{\kappa} \in \chi\right\}. \tag{8.31}$$

In particular, if (y^*, κ^*) solves (8.30), then $w^* = \frac{y^*}{\kappa^*} \in \chi$ and solves (8.29).

Proof. The condition that there exist w in the feasible set such that $(\mu - r_f)'w > 0$ implies that for κ defined as $\kappa = ((\mu - r_f)'w)^{-1}$, may be constrained to be positive in the feasible set. With κ so defined, let $y = \kappa w$ and note that the objective function in (8.29) may be written in these new variables as

$$\begin{aligned} \frac{w'\mu - r_f}{\sqrt{w'\Sigma w}} &= \frac{1}{\kappa}\left(\sqrt{w'\Sigma w}\right)^{-1} \\ &= \frac{1}{\kappa}\left(\sqrt{\frac{1}{\kappa^2}y'\Sigma y}\right)^{-1} \\ &= \left(\sqrt{y'\Sigma y}\right)^{-1}. \end{aligned}$$

This gives immediately that the original objective function maximization is equivalent to minimizing $\sqrt{y'\Sigma y}$ over both κ and y (taking into account the feasible set) and hence equivalent to minimizing $y'\Sigma y$ over these same variables as in (8.30). We are left then to similarly rewrite the constraints of the original problem.

Before addressing this issue, we must first write y without a dependence on w. This is fairly immediate, however, as

$$\begin{aligned} y &= \kappa w \\ &= ((\mu - r_f)'w)^{-1} w \end{aligned}$$

implies that $(\mu - r_f)'y = 1$. This constraint, as seen in (8.30), then, ensures the relationship we defined above. With this final piece in place, the requirement $(y, \kappa) \in \chi^+$ clearly ensures the feasibility sets of the two problems coincide. Hence the two optimization problems are equivalent. □

One minor quibble may be made that we have not rewritten the original problem in a more familiar form (read with linear constraints). This is indeed the case. Instead we have constructed a so-called quadratic programming problem with a conic constraint, where the details of this distinction are left as an exercise for the reader. It is rather immediate based on the above, though, to see that we may replace the constraint $(y, \kappa) \in \chi^+$ with just the single constraint $\kappa > 0$, giving a fairly simple linearly constrained quadratic programming problem.

The case of general constraints considered in (8.28) is handled similarly. As before, we define auxiliary variables κ and y and cones, χ and χ^+ and assume there is a feasible solution that exceeds the risk free rate in expected returns. But for the definition of χ^+, the problem equivalent to (8.28) is identical to the last formulation. Namely, we have

$$\begin{aligned} \min_{(y,\kappa)} \quad & y'\Sigma y \\ & (\mu - r_f)'y = 1 \\ & (y, \kappa) \in \chi^+ \end{aligned} \tag{8.32}$$

with, now,

$$\chi = \{w : Aw = b,\, Cw \geq d\}$$

and

$$\chi^+ = \left\{(y, \kappa) : \kappa > 0, \frac{y}{\kappa} \in \chi\right\}. \tag{8.33}$$

Further, if (y^*, κ^*) solves (8.32), then $w^* = \frac{y^*}{\kappa^*} \in \chi$ and solves (8.28). The proof is left to the reader.

Again we have a tidy description of the constraints in (8.32), but note that these do not coincide with our usual presentation. This may be remedied immediately by rewriting $(y, \kappa) \in \chi^+$ as the trio of linear constraints

$$\begin{aligned} \kappa &> 0 \\ Ay - b\kappa &= 0 \\ Cy - d\kappa &\geq 0. \end{aligned} \tag{8.34}$$

8.3 Portfolio Updates

We have already noted the importance of the global mean-variance optimal portfolio and tangency portfolio above. Here we discuss applying the general procedure of mean-variance optimization to updating a portfolio. While doing so, we will also briefly mention some common types of constraints seen in practice.

Oftentimes when considering the generalization of (8.1),

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w' \Sigma w \\ & Aw = b \\ & Cw \geq d, \end{aligned} \tag{8.35}$$

or the general Sharpe optimization problem given in (8.28), it is necessary to work from an already existing portfolio. That is, rather than identifying the optimal weight, w^*, one must find an optimal update, Δ^*, based upon initial weights w_0.

Focusing on (8.35), we may write the problem in terms of an update to some initial w_0 as

$$\begin{aligned} \min_{\Delta} \quad & \frac{1}{2} (w_0 + \Delta)' \Sigma (w_0 + \Delta) \\ & A(w_0 + \Delta) = b \\ & C(w_0 + \Delta) \geq d, \end{aligned} \tag{8.36}$$

which of course may be rewritten as

$$\begin{aligned} \min_{\Delta} \quad & \frac{1}{2} \Delta' \Sigma \Delta + w_0' \Sigma \Delta \\ & A\Delta = b - Aw_0 \\ & C\Delta \geq d - Cw_0. \end{aligned} \tag{8.37}$$

The formulation (8.37) is then a quadratic programming problem in the decision variable Δ. A similar procedure may be applied to (8.28), and this is left as an exercise.

It is common to place gross turnover constraints on a portfolio update, or, slightly more sophisticated, to place linear transaction cost constraints on the problem. In either case, we are left with having to work with the absolute value of turnover. This entails having to keep track of the sign of each Δ_i; viz., a gross turnover constraint may be written as

$$\sum_i |\Delta_i| \leq \tau$$

for some τ.

We may partition Δ into a positive and negative part as in our development of linear programming as

$$\Delta = \Delta_+ - \Delta_- \tag{8.38}$$

where $\Delta_+ \geq 0$ and $\Delta_- \geq 0$. In this case, the turnover constraint above becomes

$$\sum_i \Delta_{+,i} + \Delta_{-,i} \leq \tau. \tag{8.39}$$

In the presence of turnover constraints, (8.37) is now optimized over a vector of dimension $3N$ when $\Delta \in \mathbb{R}^N$. How exactly this comes about is illustrated below.

We begin by letting

$$\tilde{\Delta} = \begin{pmatrix} \Delta \\ \Delta_+ \\ \Delta_- \end{pmatrix},$$

and writing the relationships given in (8.38) and (8.39) as an equality and inequality constraint, respectively. For (8.38), we have

$$[I \ -I \ I]\,\tilde{\Delta} = 0,$$

where I is the $N \times N$ identity matrix. Gross turnover is handled similarly by writing

$$[0 \ -1' \ -1']\,\tilde{\Delta} \geq -\tau.$$

Notice that we need not include Δ in our modified problem since the positive and negative parts as constructed will contain the same information. However, we retain the variable for potential ease of exposition.

The original problem may now be written in these new auxiliary variables:

$$\begin{aligned} \min_{\tilde{\Delta}} \quad & \frac{1}{2}\tilde{\Delta}'\tilde{\Sigma}\tilde{\Delta} + \tilde{w_0}'\tilde{\Sigma}\tilde{\Delta} \\ & [A \ 0 \ 0]\,\tilde{\Delta} = b - Aw_0 \\ & [C \ 0 \ 0]\,\tilde{\Delta} \geq d - Cw_0 \\ & [I \ -I \ -I]\,\tilde{\Delta} = 0 \\ & [0 \ -1' \ -1']\,\tilde{\Delta} \geq -\tau. \end{aligned} \tag{8.40}$$

The determination of $\tilde{w_0}$ and $\tilde{\Sigma}$ are left to the reader.

It should be noted that the move to an updating framework from a total portfolio optimization increases the likelihood of the constraint set being infeasible when gross turnover constraints are introduced. This is perhaps more clearly understood from thinking about the problem qualitatively; viz., if the incoming portfolio is not feasible, and a small number of trades are allowed, then it isn't possible to trade to feasibility. One remedy to this type of issue is to increase τ in the above until a feasible solution is obtained. This isn't always possible, however, as turnover constraints may be more rigid than others in live trading. Alternatively, b and d may be relaxed in turn, but this requires bespoke knowledge of the mandates driving the constraints being present in the first place.

Another way to handle the potential infeasibility of (8.40) based on a particular w_0 and turnover budget τ is to introduce so-called slack variables. For example, restricting a new variable γ to be nonnegative, we may modify

$$[Z \ -1' \ -1']\,\tilde{\Delta} \geq -\tau$$

to

$$[Z \ -1' \ -1']\,\tilde{\Delta} \geq -\tau - \gamma.$$

The new variable must be included in the objective function as well as, e.g.,

$$\min_{\tilde{\Delta},\,\gamma} \frac{1}{2}\tilde{\Delta}'\tilde{\Sigma}\tilde{\Delta} + \tilde{w_0}'\tilde{\Sigma}\tilde{\Delta} + C\gamma$$

for some scalar, C. This technique generalizes to the case of multiple slack variables where the change to the objective function is $C\sum_j \gamma_j$ in this case and equality constraints are modified to inequality constraints based on enforcing conditions as $|c_i(\Delta)| \leq \gamma_i$.

8.4 Equivalence of Maximizing Returns

We have so far looked at the problem of minimizing portfolio variance with a set of linear constraints, with particular focus on an expected return constraint. Here we establish the equivalence between this original problem and maximizing returns while constraining portfolio variance. While this is a slightly mundane equivalence (based on the insights already made via the efficient frontier work), it is also one commonly needed in practice as a portfolio manager may have to adhere to mandated risk controls; i.e., a portfolio manager will be interested in identifying the best portfolio opportunity given a maximum bound on risk.

In this spirit, we will prove an equivalence between

$$\begin{aligned} \min_w \quad & \frac{1}{2}w'\Sigma w \\ & Aw = b \\ & \mu' w \geq \mu^* \\ & Cw \geq d \end{aligned} \tag{8.41}$$

and

$$\begin{aligned} \min_w \quad & -\mu' w \\ & Aw = b \\ & \frac{1}{2}w'\Sigma w \leq \nu^* \\ & Cw \geq d \end{aligned} \tag{8.42}$$

where we have explicitly broken out the return requirement in (8.41) and modified the objective and similar constraint to obtain (8.42). The parameter ν^* has not yet been specified, but the interested reader will likely guess what this value must be. Note, too, that while we have not as of yet worked with quadratic constraints, we will still utilize the Karush-Kuhn-Tucker conditions (7.11), which are again both necessary and sufficient in this case.

We next prove that for w^* a solution to (8.41), with $\frac{1}{2}w^{*\prime}\Sigma w^* = \nu^*$, then w^* solves (8.42) with the same ν^*.

Proof. The Lagrangian of (8.41) is given by

$$\mathcal{L}(w, \delta, \lambda, \eta) = \frac{1}{2}w'\Sigma w - \delta(\mu' w - \mu^*) - \lambda'(Aw - b) - \eta'(Cw - d).$$

We know that at the optimal solution there exists Lagrange multipliers satisfying the KKT conditions

$$\begin{aligned}
\nabla_w \mathcal{L}(w^*, \delta^*, \lambda^*, \eta^*) &= 0 \\
Aw^* &= b \\
Cw^* &\geq d \\
\delta^*(\mu' w^* - \mu^*) &= 0 \\
\eta_i^*(C_i w^* - d_i) &= 0.
\end{aligned}$$

We also have a Lagrangian for (8.42)

$$\tilde{\mathcal{L}}(\tilde{w}, \tilde{\delta}, \tilde{\lambda}, \tilde{\eta}) = -\mu' \tilde{w} - \tilde{\delta}\left(\frac{1}{2}\tilde{w}'\Sigma\tilde{w} - \nu^*\right) - \tilde{\lambda}'(A\tilde{w} - b) - \tilde{\eta}'(C\tilde{w} - d)$$

and associated KKT conditions at the optimal solution

$$\begin{aligned}
\nabla_w \tilde{\mathcal{L}}(\tilde{w}^*, \delta^*, \lambda^*, \eta^*) &= 0 \\
A\tilde{w}^* &= b \\
C\tilde{w}^* &\geq d \\
\tilde{\delta}^*\left(\frac{1}{2}\tilde{w}'\Sigma\tilde{w} - \nu^*\right) &= 0 \\
\tilde{\eta}_i^*(C_i\tilde{w}^* - d_i) &= 0.
\end{aligned}$$

To confirm the equivalence we are looking for, we verify that

$$\begin{aligned}
\tilde{w}^* &= w^* \\
\tilde{\delta}^* &= \frac{1}{\delta^*} \\
\tilde{\lambda}^* &= \frac{\lambda^*}{\delta^*} \\
\tilde{\eta}^* &= \frac{\eta^*}{\delta^*}
\end{aligned}$$

satisfy the KKT conditions for (8.42) when $(w^*, \delta^*, \lambda^*, \eta^*)$ satisfy the KKT conditions for (8.41).

Since feasibility in the constraints outside of the mean return or variance are identical between the two problems, we are left to confirm

$$\begin{aligned}
\nabla_w \tilde{\mathcal{L}}(\tilde{w}^*, \delta^*, \lambda^*, \eta^*) &= 0 \\
\tilde{\delta}^*\left(\frac{1}{2}\tilde{w}'\Sigma\tilde{w} - \nu^*\right) &= 0 \\
\tilde{\eta}_i^*(C_i\tilde{w}^* - d_i) &= 0.
\end{aligned}$$

We leave the verification of the gradient vanishing as a problem in the exercises. Looking at the remaining conditions, we know that

$$\eta_i^*(C_i w^* - d_i) = 0$$

for each row of C, and hence it is immediate that

$$\tilde{\eta}_i^*(C_i\tilde{w}^* - d_i) = 0$$

by the definition of $\tilde{\eta}^*$ and $\tilde{w}$. Further, strict complementarity holds. Next, since

$$\begin{aligned} \frac{1}{2}w^{*\prime}\Sigma w^* &= \nu^* \\ \tilde{\delta}^*\left(\frac{1}{2}\tilde{w}'\Sigma\tilde{w} - \nu^*\right) &= 0, \end{aligned}$$

again with strict complementarity.

Therefore the two problems as constructed have identical solutions, and refer to the same point on the efficient frontier. □

A third formulation may be considered, motivated by the Lagrangian of the first problem. Here, a parameterization of the efficient frontier is explicit in γ:

$$\begin{aligned} \min_w \quad & \frac{1}{2}w'\Sigma w - \gamma\mu' w \\ & Aw = b \\ & Cw \geq d. \end{aligned} \tag{8.43}$$

The parameter γ is sometimes referred to as an appetite for risk, with the case of $\gamma = 0$ coinciding with the global mean-variance optimal portfolio.

8.5 Pitfalls and a Word of Caution

The unconstrained mean-variance problem encountered at the beginning of the chapter is rife with issues in practice, so much so that it the procedure itself has at times been referred to as an 'error maximizer' when the sample covariance matrix is used. Some indication that this might be the case was already seen, but in a different context; namely, when the distribution of eigenvalues was considered as in Figure 3.3, we noted that the bulk of the eigenvalues (by count) were near zero, with a handful very large eigenvalues explaining the majority of the variance.

In the present context, this phenomenon leads to stark results. Consider that without loss of generality, we may consider any of the preceding mean-variance optimization problems in an alternative basis. Changing basis to that of the eigenvectors of Σ yields

$$w'\Sigma w = (Qw)'\Lambda(Qw)$$

for some change of basis matrix satisfying $Q\Sigma Q' = \Lambda = diag(\lambda_1, \ldots, \lambda_N)$. But, we've seen that there are several eigenvectors with eigenvalues near zero. Assuming that the $\lambda_1 < \lambda_2 < \cdots < \lambda_N$, then the optimizer will maximize the return of the eigenportfolios e_1, e_2, etc., if their expected returns are nonzero

since there it perceives that these eigenportfolios have little to no risk. If no exposure constraints are enforced (viz., $|w| \leq \delta$), then the optimizer will view the total allocation constraint in our original problem as a net constraint.

The result of all of this is that when using the sample covariance matrix without informative constraints (especially maximum exposure constraints), the distribution of optimal weights tends toward heavy concentrations in both longs and shorts. This is literally counter to every bit of intuition presented so far that has led to this point.

Luckily, several remedies are available. One in particular has already been alluded to: incorporating constraints greatly mitigates the concentrations issues noted above. On the one hand, this has been seen as simply codifying preferences rather mean-variance optimization as such; i.e., the idea that constraints may resolve issues with the initial problem may be seen solely as the power of the *a priori* intuition of the practitioner imposing those constraints rather than as a boon to the process as such. On the other hand, we will later show that we may view the addition of constraints as a modification of the covariance matrix.

We will also consider alternative modifications of the covariance. Our focus will range from factor models, to so-called shrinkage estimators, to results with origins in random matrix theory.

Before we get to these advances in approach, however, we are left with some intuition that (8.1) is not suitable for practical implementation without due consideration of the input covariance matrix.

Exercises

1. How can you be sure that each point on the efficient frontier is uniquely represented by a single portfolio?

2. Determine a, b, and c in (8.5).

3. Establish (8.8).

4. Use Cauchy-Schwarz to show that

$$\left(\mu'\Sigma^{-1}\mu\right)\left(1'\Sigma^{-1}1\right) - \left(\mu'\Sigma^{-1}1\right)^2 > 0.$$

   Use this result to show that the determinant of (8.6) is positive.

5. Prove (8.17).

6. Verify (8.18).

7. Why must we assume $\underline{\mu} > r_f$ preceding (8.19) to get the result noted? What happens when $\underline{\mu} = r_f$? When $\underline{\mu} < r_f$?

8. Assume that both the CAPM model holds and that the expected return of the market is the same as that of the tangency portfolio. Prove that the tangency portfolio is the market portfolio.

9. Prove (8.24) and (8.25).

10. Show that χ^+ as defined by (8.31) is a cone. That is, for every $u \in \chi^+$, and positive scalar, α, $\alpha u \in \chi^+$. Show further that χ^+ is a convex cone by showing that for any u and v in χ^+, then $\alpha u + \gamma v \in \chi^+$ for positive scalars, α and γ. How can you use this information to show that (8.30) has a unique global minimum?

11. Prove that if $(y, \kappa)\chi^+$ as in (8.33), with y and κ defined as in the case of simple constraints, then the feasible sets of (8.32) and (8.28) are the equivalent. Conclude that the two optimization problems are equivalent.

12. Prove that the constraints given in (8.34) are equivalent to χ^+ coupled with the requirements on κ for the general Sharpe optimization problem. Rewrite these constraints in matrix notation.

13. Write a portfolio updating problem similar to that given in (8.37) based upon the Sharpe optimization problem (8.28).

14. Determine $\tilde{w_0}$ and $\tilde{\Sigma}$ for the portfolio update probelm with turnover constraints.

15. Assuming turnover costs are linear in Δ, how would you modify (8.40) to enforce a turnover cost budget of τ_0 rather than having a gross turnover constraint?

16. Focusing on (8.35), write the specific constraints needed to construct a portfolio with gross notional of 2 ($\sum i|w_i| = 2$), and net notional of zero ($\sum_i w_i = 0$). This will require a similar approach taken in the case of turnover constraints, but without the introduction of Δ; i.e., this is not an update, but a full portfolio optimization.

17. Again using the framework of (8.35), write the constraints needed for a portfolio optimization problem ensuring that net exposure to the value anomaly measured by top and bottom quintile is positive, no short (long) exposure exists in the top (bottom) quintile of value, gross exposure of the portfolio is 2, and net exposure is between -0.2 and 0.2. For notation, let ν be the vector of value quintiles; i.e., ν is a vector with values in $\{1, 2, 3, 4, 5\}$ of the same dimension as the decision variable of the problem, with values of 5 indicating 'high' value.

18. Generalize (8.40) to robustly handle the case of infeasibility by introducing slack variables in the problem where appropriate. Why must the constant C noted in this discussion be used? How might one choose the size of C?

19. In the proof of the equivalence between minimizing variance and maximizing return, the solution depended on dividing by δ^*. How can we be sure that δ^* is not zero?

20. Prove that the gradient of the Lagrangian vanishes for the choices of weights and Lagrange multipliers used for (8.42). This completes the needed verifications for the equivalence proof.

21. Construct a 3×3 covariance matrix and an expected return vector such that your constructed covariance matrix has a near-zero eigenvalue with positive expected return for the associated eigenportfolio. Solve (8.1) using a solver and determine the weight of the eigenportfolio of the near-zero eigenvalue.

Chapter 9

Coherent Measures of Risk

Alternatives to the variance-as-risk measure are ubiquitous, and, as indicated from our previous observations such as the inverse relationship between variance-as-risk and *ex-post* returns, warranted. In this chapter we identify the requirements for so-called *coherent measures of risk*. Following the approach of previous chapters, we do not take these conditions as necessary in any sense, but, rather as a taxonomy of what one might articulate as a reasonable risk measure. For example, at the outset of the text we compared unlevered and levered portfolios and noted that any reasonable measure of risk should account for this intuitive approach. This, among other similar features, is indeed the case for coherent measures of risk.

After establishing the standard conditions, we give several examples of potential measures of risk, some of which may fail one or another requirement of being a coherent measure of risk. A few particularly popular risk measures will be formulated as linear programming problems.

9.1 Definition and Examples

We focus specifically on coherent measures of risk from the frame of portfolio positions. Specifically, for a fixed universe of N securities, we define a portfolio, Π, as the positions in these assets,

$$\Pi = \begin{pmatrix} p_1 \\ \vdots \\ p_N \end{pmatrix}.$$

In this formulation, we treat the p_i as nonrandom. It is common to refer to both the portfolio as Π as well as its vector of positions.

The performance of a portfolio is subject to some stochastic element. We let

$$r = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix}$$

denote the performance of the N assets in our universe, with r_i corresponding to the ith asset. We may regard r as either normalized performance (read percent change) or changes in value measured in the currency of Π; in the former case, we often assume that the p_i are normalized as weights as well, while in the latter, p_i will denote the position in security i, irrespective of other positions taken. Where necessary, we will specify which of these interpretations for r is being used. In contrast to our work in mean-variance optimization, however, we will prefer using positions and changes in security values over returns and normalized positions in the present work and will assume this is the case unless otherwise specified. Of course, there is also an implicit assumption of a unit of time over which r is evaluated. The profit or loss of the portfolio Π over a given period is simply, $\Pi' r$, as we have seen previously. Finally, to be consistent with conditions that follow, we will fix the sign of r so that a positive value indicates a loss for a long (positive) position.

Before stating the mathematical formulation of what makes a coherent measure of risk, we first give some qualitative motivation. First, we would expect that any reasonable risk measure should respect leverage. That is, a levered portfolio should multiply the risk of its un-levered counterpart. Next, we would want that the addition of a risk-free asset to a portfolio should not add to the risk of the portfolio. This condition will actually be strengthened in what follows. Third, based on all that has preceded, we have some preference for diversification, and we would prefer a risk measure that reflects this. Finally, if one portfolio almost surely exceeds another under all scenarios, r, then we should like to say that the first is less risky than the second.

With these features established, we say

$$\rho : \Pi \quad \rightarrow \quad \mathbb{R} \tag{9.1}$$

is a coherent measure of risk if $\rho(\cdot)$ satisfies the following conditions:

Positive Homogeneity For λ a positive scalar, $\rho(\lambda \cdot \Pi) = \lambda \rho(\Pi)$.

Translation Invariance For p_f a position in the risk-free asset with profit, r_f, $\rho(\Pi + p_f) = \rho(\Pi) - p_f \cdot r_f$.

Monotonicity If $\Pi_1' r < \Pi_2' r$ for all instances of r, then $\rho(\Pi_1) < \rho(\Pi_2)$.

Subadditivity $\rho(\Pi_1 + \Pi_2) \leq \rho(\Pi_1) + \rho(\Pi_2)$.

We have as archetypical examples of risk in portfolio variance and volatility. It should be readily apparent that these examples will not satisfy all of the conditions above. Translation Invariance is immediately an issue as we know

variance is translation invariant in the traditional sense; viz., adding a nonrandom scalar to a random variable does not affect the variance of the random variable. Furthermore, Positive Homogeneity gives an indication that volatility may be better suited to the remaining conditions than variance. Continuing this cursory discussion, it may also be apparent to some readers that coherent measures of risk will focus on values of the distribution of loss (which must be denoted as positive in the above framing) rather than the moments of these distributions.

We proceed by analyzing various risk measures according to the conditions just set out.

9.1.1 Volatility

If we define $\rho(\Pi)$ by the volatility of the portfolio loss $\Pi' r$, $\nu(\Pi) = \sqrt{Var(\Pi' r)}$, or in terms of the inner product and associated norm defined previously,

$$\nu(\Pi) = \sqrt{(\Pi,\, \Pi)_\Sigma} = ||\Pi||_\Sigma,$$

for Σ the covariance of r, many of the conditions for a coherent measure of risk are immediately evident.

To check Positive Homogeneity, we simply calculate, for $\lambda > 0$,

$$\begin{aligned} \nu(\lambda\Pi) &= ||\lambda\Pi||_\Sigma \\ &= \lambda||\Pi||_\Sigma \\ &= \lambda\nu(\Pi). \end{aligned}$$

As alluded to earlier, volatility does not adhere to the Translation Invariance condition of a coherent measure of risk since the inner product defining volatility is translation invariant with respect to nonrandom shifts; viz.,

$$\begin{aligned} \nu(\Pi + p_f) &= ||\Pi + p_f||_\Sigma \\ &= ||\Pi||_\Sigma \\ &= \nu(\Pi). \end{aligned}$$

This same reasoning precludes Monotonicity from obtaining for volatility. Consider two identical portfolios but for an additional long position in the risk-free asset in the first portfolio. Clearly, in terms of the positions of the portfolios (and not weights), the first portfolio's losses will always be less than the seconds since the risk-free asset will create a parallel shift in these losses. However, as just shown the risks of these portfolios are equal. More generally, we notice that since coherent measures of risk are location dependent, while volatility and variance are not, this condition will fail in more general cases as well.

Lastly, to verify that volatility is Subadditive, we begin with the square of

$\nu(\cdot)$ for ease of exposition,

$$\begin{aligned}
\nu(\Pi_1 + \Pi_2)^2 &= ||\Pi_1 + \Pi_2||_\Sigma^2 \\
&= ||\Pi_1||_\Sigma^2 + ||\Pi_2||_\Sigma^2 + 2(\Pi_1,\, \Pi_2)_\Sigma \\
&\leq ||\Pi_1||_\Sigma^2 + ||\Pi_2||_\Sigma^2 + 2\,|(\Pi_1,\, \Pi_2)_\Sigma| \\
&\leq ||\Pi_1||_\Sigma^2 + ||\Pi_2||_\Sigma^2 + 2||\Pi_1||_\Sigma \cdot ||\Pi_2||_\Sigma \\
&= \left(||\Pi_1||_\Sigma + ||\Pi_2||_\Sigma\right)^2,
\end{aligned}$$

where the second inequality is obtained from Cauchy-Schwarz. Taking square roots, we have

$$\nu(\Pi_1 + \Pi_2) \leq ||\Pi_1||_\Sigma + ||\Pi_2||_\Sigma = \nu(\Pi_1) + \nu(\Pi_2).$$

The above construction gives some guidance in identifying why variance is not Subadditive in the general case. This is left as an exercise.

In analyzing volatility above, we noted that a coherent measure of risk must be location dependent. One rather immediate candidate, consistent with previous work, might be to define a risk measure based on both expectation and volatility as

$$\rho_\gamma(\Pi) = \mathbb{E}(\Pi' r) + \gamma \cdot ||\Pi||_\Sigma \tag{9.2}$$

for some fixed $\gamma > 0$. We leave it to the reader to show that Positive Homogeneity, Translation Invariance, and Subadditivity hold for ρ_γ. Monotonicity fails in the general case, however, since we have not specified r sufficiently to connect the rank ordering of portfolio performance to the values given by ρ_γ.

This, too, gives some indication of a way to construct a coherent measure of risk from this work. Namely, in the case that r has an elliptical distribution, ρ_γ is directly related to the percentiles of the loss distribution defined by $\Pi' r$. As a simplified example, let $r \sim N(\mu, \Sigma)$. We have that the cumulative density function for portfolio losses, Y, is then given by the standard normal cumulative density function, Φ, by

$$F_\Pi(y) = \Phi\left(\frac{y - \Pi'\mu}{||\Pi||_\Sigma}\right)$$

where, again, $F_\Pi(y) = \mathbb{P}(Y < y)$. The definition of ρ_γ, then, implies that γ exactly determines a probability; viz., for fixed γ,

$$\begin{aligned}
F_\Pi(\mathbb{E}(\Pi' r) + \gamma \cdot ||\Pi||_\Sigma) &= \Phi\left(\frac{\mathbb{E}(\Pi' r) + \gamma \cdot ||\Pi||_\Sigma - \Pi'\mu}{||\Pi||_\Sigma}\right) \\
&= \Phi(\gamma).
\end{aligned}$$

That is, γ determines the probability that portfolio losses are bounded above by $\rho_\gamma(\Pi)$; viz., $\mathbb{P}(Y < \mathbb{E}(\Pi' r) + \gamma \cdot ||\Pi||_\Sigma)$.

We may then reformulate the problem in terms of these percentiles. For example, if we want to look at the 95^{th} percentile of portfolio losses, we fix γ as $\gamma = \Phi^{-1}(0.95)$. For this γ, $\rho_\gamma(\Pi)$ is the value such that losses will only exceed $\rho_\gamma(\Pi)$ 5% of the time.

Returning to the question of constructing a coherent measure of risk via volatility, we have, finally, that for normally distributed, r, ρ_γ is a coherent measure of risk since we may now verify that it satisfies Monotonicity. This follows directly from the work we just established since if $\Pi_1' r < \Pi_2' r$ for all instances of r, then, necessarily, the percentiles of the loss distribution for Π_1 are all less than those of Π_2. Consequently, $\rho_\gamma(\Pi_1) < \rho_\gamma(\Pi_2)$.

The generalization to elliptical distributions follows this same pattern, and the exercise is left to the reader.

A further generalization of the above leads to the concept of Value-at-Risk, which we formally define and analyze in the next example.

9.1.2 Value-at-Risk

Given a percentile, β, the β Value-at-Risk, or β-VaR, is the smallest value, α_β, such that the probability of losses exceeding α_β is $1 - \beta$. In the continuous example given above, this may be stated more concisely since α_β is simply the β percentile of losses.[1]

It is useful to formalize some underlying concepts. If r has probability density function $f(\cdot)$, then the probability that portfolio losses of the portfolio, Π, do not exceed some specified value, α, is given by the integral

$$\Psi(\Pi, \alpha) = \int_{\Pi' r < \alpha} f(r) dr. \tag{9.3}$$

The β-VaR is determined from Ψ as

$$\alpha_\beta(\Pi) = \min \{\alpha \in \mathbb{R} \,|\, \Psi(\Pi, \alpha) \geq \beta\}, \tag{9.4}$$

for some $\beta \in (0, 1)$. It is common to discuss β as a percent; e.g., we often say 95% VaR for $\beta = 0.95$. In spite of our statement that we will consider continuous portfolio losses in our treatment, we note that the definition given in (9.4) accounts for the non-continuous case by assigning $\alpha_\beta(\Pi)$ to the leftmost point in the nonempty interval consisting of values, α such that $\Psi(\Pi, \alpha) \geq \beta$.

The β-VaR of a portfolio may be approximated by sampling according to the distribution of r. For such samples, $\{r_k\}_{i=1}^K$, the integral defining (9.3) is approximated as

$$\Psi(\Pi, \alpha) \approx \frac{1}{K} \sum_{k=1}^{K} \delta(\Pi' r_k < \alpha) \tag{9.5}$$

$\delta(\Pi' r_k < \alpha) = 1$ if $\Pi' r_k < \alpha$ and 0 otherwise. Notice that the density function is replaced by $\frac{1}{K}$ since r_k are sampled according to the density of r.

We have already established that β-VaR is a coherent measure of risk when r has an elliptical distribution. However, this is not true in the general case.

[1] It is somewhat unfortunate to reuse the variables, α and β in this way, but it is common in the literature.

Positive Homogeneity of β-VaR follows directly from the definitions. For $\lambda > 0$,

$$\begin{aligned}\Psi(\lambda\Pi, \alpha) &= \int_{\lambda\Pi' r < \alpha} f(r)dr \\ &= \int_{\Pi' r < \frac{\alpha}{\lambda}} f(r)dr \\ &= \Psi\left(\Pi, \frac{\alpha}{\lambda}\right).\end{aligned}$$

So that

$$\begin{aligned}\alpha_\beta(\lambda\Pi) &= \min\{\alpha \in \mathbb{R} \,|\, \Psi(\lambda\Pi, \alpha) \geq \beta\} \\ &= \min\left\{\alpha \in \mathbb{R} \,|\, \Psi\left(\Pi, \frac{\alpha}{\lambda}\right) \geq \beta\right\} \\ &= \min\left\{\lambda\frac{\alpha}{\lambda} \in \mathbb{R} \,|\, \Psi\left(\Pi, \frac{\alpha}{\lambda}\right) \geq \beta\right\} \\ &= \lambda \min\left\{\frac{\alpha}{\lambda} \in \mathbb{R} \,|\, \Psi\left(\Pi, \frac{\alpha}{\lambda}\right) \geq \beta\right\} \\ &= \lambda\alpha_\beta(\Pi).\end{aligned}$$

Translation Invariance follows similarly since for a risk free position, p_f,

$$\begin{aligned}\Psi(\Pi + p_f, \alpha) &= \int_{\Pi' r - p_f r_f < \alpha} f(r)dr \\ &= \int_{\Pi' r < \alpha + p_f r_f} f(r)dr \\ &= \Psi\left(\Pi, \alpha + p_f r_f\right).\end{aligned}$$

The remainder of this verification is left as an exercise.

Monotonicity is similarly obtained. For $\Pi_1' r < \Pi_2' r$ for all instances of r, then

$$\begin{aligned}\alpha_\beta(\Pi_1) &= \min\{\alpha \in \mathbb{R} \,|\, \Psi(\Pi_1, \alpha) \geq \beta\} \\ &\leq \min\{\alpha \in \mathbb{R} \,|\, \Psi(\Pi_2, \alpha) \geq \beta\} \\ &= \alpha_\beta(\Pi_2).\end{aligned}$$

The techniques to verify these conditions cannot be used to establish Subadditivity. And, in fact, β-VaR is not subadditive in the general case. Consider a portfolio with a single position in each of two defaultable bonds, B_1 and B_2, with recoveries of \$40 and \$60, respectively. In the case of default, the value of each bond is given by its recovery value. Let the current values of B_1 and B_2 be \$100 and \$105, respectively, and assume each bond has a 3% chance of defaulting over the next year. Assume further that defaults are independent, and if no default occurs, the each bond's price increases by \$1.

We begin by looking at the portfolio holding both bonds and consider the loss if a default occurs under the three combinations of defaults (B_1 defaults alone, B_2 defaults alone, or both default) and their probabilities,

$$\begin{aligned}&(.03 \cdot 0.97)((100-40)+(105-106))+\\&(.97 \cdot 0.03)((100-101)+(105-60))+\\&\quad(.03 \cdot 0.03)((100-40)+(105-60)),\end{aligned}$$

which is a loss of \$8.24 with a probability of 5.91%. As a result, the 95% VaR of the portfolio is \$8.24. For each of the bonds considered in isolation, though, the 95% VaR is -\$1, indicating that diversification *increases* risk.

One feature of β-VaR that may be gleaned from the final example on its lack of general subadditivity is that β-VaR is not concerned with losses in the right portion of the loss distribution past $\alpha_\beta(\Pi)$; viz., in the case of defaultable bonds, the default losses were not accounted for when considering the 95% VaR. One solution to account for such tail behavior is to compute the average loss conditioned on exceeding $\alpha_\beta(\Pi)$. This is exactly the definition of conditional value at risk, or CVaR, which we consider in our next example.

9.1.3 Conditional Value-at-Risk

The β-CVaR of the portfolio Π, which we will denote by $\phi_\beta(\Pi)$, is the average of losses exceeding $\alpha_\beta(\Pi)$, the β-VaR. Formally, we have

$$\phi_\beta(\Pi) = \frac{1}{1-\beta}\int_{\Pi' r \geq \alpha_\beta(\Pi)} \Pi' r f(r) dr, \tag{9.6}$$

where, as usual, $f(\cdot)$ is the probability density function for r.

We leave the verification that β-CVaR is a coherent measure of risk as an exercise. To show Monotonicity and Subadditvity, it is useful to write $\phi_\beta(\Pi)$ as the average of all $\tilde{\beta}$-VaR values of the portfolio Π as $\tilde{\beta}$ ranges from β to 1. That is,

$$\phi_\beta(\Pi) = \frac{1}{1-\beta}\int_\beta^1 \alpha_{\tilde{\beta}}(\Pi) d\tilde{\beta}. \tag{9.7}$$

To prove this, we note that we may write $\phi_\beta(\Pi)$ with respect to the distribution of portfolio losses directly,

$$\phi_\beta(\Pi) = \frac{1}{1-\beta}\int_{\alpha_\beta(\Pi)}^\infty y f_{\mathcal{L}}(y) dy, \tag{9.8}$$

where $f_{\mathcal{L}}(\cdot)$ is the probability density function for portfolio losses. If we let $\tilde{\beta} = F_{\mathcal{L}}(y)$, with $F_{\mathcal{L}}(\cdot)$ the cumulative distribution function, then $d\tilde{\beta} = f_{\mathcal{L}}(y)dy$. By noticing that $F_{\mathcal{L}}(y)$ is exactly $\Psi(\Pi, y)$, we have, using leftpoint values, $F^{-1}(\tilde{\beta}) = y = \alpha_{\tilde{\beta}}(\Pi)$. Completing the change of variables in (9.8) gives (9.7) as desired.

Notice that (9.7) gives that β-CVaR is always at least as large as β-VaR. As such, if a particular risk tolerance is defined by β-VaR, matching this risk

tolerance with a β-CVaR target will always suffice. Further, as the latter is a coherent measure of risk, it may be more desirable to focus on β-CVaR directly.

Another useful formula for β-CVaR is given by

$$\phi_\beta(\Pi) = \min_\alpha \left(\alpha + \frac{1}{1-\beta} \int_{r \in \mathbb{R}^N} [\Pi' r - \alpha]_+ f(r) dr \right), \tag{9.9}$$

where, as usual, $[x]_+ = \max(0, x)$. Defining

$$G_\beta(\Pi, \alpha) = \alpha + \frac{1}{1-\beta} \int_{r \in \mathbb{R}^N} [\Pi' r - \alpha]_+ f(r) dr,$$

it may be shown that for fixed Π [29, 30],

$$\frac{\partial G_\beta}{\partial \alpha} = 1 + \frac{1}{1-\beta}(\Psi(\Pi, \alpha) - 1). \tag{9.10}$$

Taking a second partial with respect to α shows that G_β is also convex in α. As a result, the function is minimized in α when $\frac{\partial G_\beta}{\partial \alpha} = 0$; in other words, when $\Psi(\Pi, \alpha) = \beta$. We have established earlier that this is satisfied exactly when $\alpha = \alpha_\beta(\Pi)$. Hence, we are left to verify that

$$\phi_\beta(\Pi) = \alpha_\beta(\Pi) + \frac{1}{1-\beta} \int_{r \in \mathbb{R}^N} [\Pi' r - \alpha_\beta(\Pi)]_+ f(r) dr.$$

We have

$$\begin{aligned} \int_{r \in \mathbb{R}^N} [\Pi' r - \alpha_\beta(\Pi)]_+ f(r) dr &= \int_{\Pi' r \geq \alpha_\beta(\Pi)} (\Pi' r - \alpha_\beta(\Pi)) f(r) dr \\ &= \int_{\Pi' r \geq \alpha_\beta(\Pi)} \Pi' r f(r) dr - (1-\beta)\alpha_\beta(\Pi) \\ &= (1-\beta)\phi_\beta(\Pi) - (1-\beta)\alpha_\beta(\Pi), \end{aligned}$$

so that multiplying through by $(1-\beta)^{-1}$ confirms the result.

Finally, we note, but do not prove here, that $G_\beta(\Pi, \alpha)$ is convex in the joint variables of Π and α [27].

We are now in a position, again, to provide an approximation to β-CVaR according to sampling from the distribution of r. Using the same notation as before, we have

$$G_\beta(\Pi, \alpha) = \alpha + \frac{1}{1-\beta} \frac{1}{K} \sum_{k=1}^{K} [\Pi' r_k - \alpha]_+, \tag{9.11}$$

and $\phi_\beta(\Pi)$ may thus be approximated by minimizing (9.11) over α. We note that if β is close to 1, it may be beneficial (for numerical stability) to minimize $(1-\beta)G_\beta(\Pi, \alpha)$. The question of how to handle the $[\cdot]_+$ function in a minimization setting will be addressed subsequently in the chapter.

Both β-VaR and β-CVaR are quantile based risk measures, relying on the distribution of portfolio losses. One feature that they do not retain as a result is the serial ordering of portfolio losses; viz.; like all risk measures considered so far, they disregard whether large portfolio losses may cluster or not, a taxonomy of returns we have noted previously. Common risk measures which are order-dependent are constructed from portfolio drawdowns. We examine some of these examples next.

9.1.4 Drawdown Measures

An especially common portfolio evaluation metric in hedge funds is to calculate the maximum drawdown over various time windows. Using the same notation as we have throughout, we now add some serial component to our portfolio loss variable, r, which we denote, as usual, by r_t. We define the relative portfolio value at time, τ, of a portfolio, Π, by

$$P(\Pi, k) = -\Pi' \sum_{t=1}^{k} r_t, \tag{9.12}$$

where the minus sign accounts for our convention of $r.$ being a loss and discrete time steps are chosen. In the case of a continuous process, we may write

$$P(\Pi, \tau) = -\Pi' \int_1^{\tau} r_t dt.$$

In both discrete and continuous settings, the drawdown function at time τ is given by

$$D(\Pi, \tau) = \max_{1 \leq t \leq \tau} P(\Pi, t) - P(\Pi, \tau). \tag{9.13}$$

We will specifically write $D(\Pi, k)$ when using a discrete sample. The maximum drawdown over a period $t \in [1, T]$ is then

$$M(\Pi) = \max_{0 \leq t \leq T} D(\Pi, t), \tag{9.14}$$

and the period average drawdown, or simply average drawdown, over the same time window is

$$A(\Pi) = \frac{1}{T} \sum_{k=1}^{T} D(\Pi, k). \tag{9.15}$$

As we have written it, the average drawdown is drawdown per unit time denoted by the discrete steps in $[1, T]$. In the case of a continuous process, we may write the average drawdown as

$$A(\Pi) = \frac{1}{T} \int_{t=0}^{T} D(\Pi, t) dt.$$

Neither maximum drawdown nor average drawdown are coherent measures of risk as each fails Translation Invariance and Subadditivity. Monotonicity may

be shown if, for instance, the definition is extended to multiple periods; viz., $\Pi_1' r_t < \Pi_2' r_t$ for every t. These verifications are left as exercises.

Analogues of both β-VaR and β-CVaR may be constructed from the maximum drawdown distribution. For example, following Chekhlov, Uryasev, and Zabarankin [7], we may define β-Drawdown-at-Risk (β-DaR) and β-Conditional Drawdown at Risk (β-CDaR) as

$$\delta_\beta(\Pi) = \min\left\{\delta \in \mathbb{R} \,|\, \mathbb{P}(D(\Pi, t) > \delta) \geq \beta \text{ for } t \in [0, T]\right\}, \tag{9.16}$$

and

$$\Delta_\beta(\Pi) = \frac{1}{1-\beta}\frac{1}{T}\int_{D(\Pi,t)\geq\delta_\beta(\Pi)} D(\Pi, t)dt. \tag{9.17}$$

Following our previous work, one may show that (9.17) may be also be written as

$$\Delta_\beta(\Pi) = \frac{1}{1-\beta}\frac{1}{T}\int_\beta^1 \delta_\beta(\Pi)d\beta. \tag{9.18}$$

Similarly, it is helpful to write (9.17) as a minimization problem in $\delta_\beta(\Pi)$, and the resulting equation should be familiar:

$$\Delta_\beta(\Pi) = \min_\delta \left(\delta + \frac{1}{1-\beta}\frac{1}{T}\int_0^T (D(\Pi, t) - \delta)_+ dt\right). \tag{9.19}$$

Of particular utility is being able to approximate the objective function in (9.19), which may be accomplished by writing

$$D_\beta(\Pi, \delta) \quad = \quad \delta + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K}[\max_{0\leq j\leq k}(P(\Pi, j) - P(\Pi, k)) - \delta]_+ \tag{9.20}$$

where r_t has been sampled according to the distribution of losses for the securities in question. We will return to this problem when we define drawdown constraints in an optimization setting. Before proceeding, however, we note that there is some difference between the sampling properties used to approximate β-CVaR and β-CDaR exhibited here. Namely, in the former case, sampling was according to the distribution of losses, irrespective of serial ordering. As such, as K grew, we might expect that the value obtained converged to the population β-CVaR. (This is the case.) In the above, however, K represents the number of time steps in the path of losses which we have used to define portfolio value. As such, to converge to (9.19), one would have to average (9.20) over several paths.

We state, but do not prove, that β-CDaR satisfies all conditions of being a coherent measure of risk but Translation Invariance, where $\Delta_\beta(\Pi + p_f) = \Delta_\beta(\Pi)$. These verifications for the continuous case are left as exercises.

9.2 Implementation as Linear Constraints

Several of the new risk measures introduced in this chapter rely on an unspecified distribution for the loss variable, r. That is, while we might associate variance

(or volatility) as a risk measure associated with and justified by assuming r has a normal distribution via arguments based on CAPM, we have no theoretical underpinning to determine such a distribution for, say, β-CVaR generally. This is true even while we showed an equivalence of β-CVaR to volatility as a risk measure when the joint distribution of losses is assumed normal. While on the one hand this distribution relaxation allows greater flexibility, on the other, we have yet to specify how we might implement any of these new risk measures in practice.

In this section we outline procedures to include any of β-CVaR, average drawdowns, maximum drawdowns, and β-CDaR as linear constraints in identifying some optimal portfolio, Π^*. We do not include constraints in β-VaR or β-DaR, as these are outside the scope of the text.[2] Throughout, we will assume a baseline portfolio optimization problem of the type seen in (7.32); that is,

$$\begin{aligned} \min_{\Pi} \quad & \Pi'\hat{r} \\ & A\Pi = b \\ & C\Pi \geq d. \end{aligned} \tag{9.21}$$

We will then identify auxiliary variables which we must add to the problem for each risk measure considered. Afterwards, we discuss methods to simulate specified distributions for r; largely collecting previous results from the text as opposed to introducing any new concept in this regard.

Given this choice of arrangement, we begin by first assuming some sampling technique from the loss variable, r, is possible, and then discuss a method for such sampling.

9.2.1 β-CVaR Constraints

Assume that $\{r_k\}_{k=1}^K$ is sampled according to the distribution of r, and β is fixed between 0 and 1. The discretization of the objective function used to determine β-CVaR given by (9.11),

$$G_\beta(\Pi, \alpha) = \alpha + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K}[\Pi' r_k - \alpha]_+,$$

may be rewritten using auxiliary variables as

$$\begin{aligned} \alpha + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K} z_k & \\ z_k \geq 0 & \\ z_k \geq \Pi' r_k - \alpha. & \end{aligned} \tag{9.22}$$

[2] This may be surprising given that we will be able to, for instance, constrain β-CVaR, and by doing so, obtain the β-VaR of the portfolio. However, the integration of β-VaR used to obtain β-CVaR is the key feature that distinguishes the two, as it leads to smoothness as well as Subadditivity.

Rockafellar and Uryasev [27] establish that when minimization of (9.21) is carried out in $\Pi \times \alpha \times z$, with $\alpha + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K} z_k$ bounded above by some risk tolerance as an additional constraint, the result is that α^* will be the approximate β-VaR (approximate due to discretization), and the approximate β-CVaR is represented and bound by the constraint as well.

Putting this all together, we have that the β-CVaR constrained problem becomes

$$\begin{aligned} \min_{(\Pi,\, \alpha,\, z)} \quad & \Pi'\hat{r} \\ A\Pi &= b \\ C\Pi &\geq d \\ \alpha + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K} z_k &\leq \omega \\ z &\geq 0 \\ z_k + -\Pi' r_k + \alpha &\geq 0. \end{aligned} \tag{9.23}$$

We leave it as an exercise to rewrite this problem in matrix form. The triplet (Π^*, α^*, z^*) solving (9.23) yields the β-VaR for the portfolio, Π^*, in α^*. Of course, the β-CVaR of the portfolio is $\alpha^* + \frac{1}{1-\beta}\frac{1}{K}\sum_{k=1}^{K} z_k^*$, and this provides an upper bound for β-VaR, as noted previously. As a result, while we do not give a method of how to constrain β-VaR here, we may conservatively constrain a portfolio to not exceed some β-VaR limit by using that same limit in the above.

We note that (9.23) is convex in (Π, α); a result which may proved directly by resorting to the KKT conditions of the problem. Consequently, we may also, for instance, create an analogous mean-β-CVaR curve as previously considered in mean-variance optimization. Finally, and as seen previously, an alternate formulation of (9.23) with β-CVaR objective should be apparent.

9.2.2 Drawdown Constraints

We next assume that $\{r_t\}_{t=1}^{T}$ is sampled according to the distribution of r, noting that some serial relationships may be introduced; i.e., we do not assume necessarily that these observations are iid, and the ordering in t cannot be disregarded.

The maximum drawdown for this sample, for fixed Π, is given by

$$\begin{aligned} M(\Pi) &= \max_{1\leq k\leq T} D(\Pi, k) \\ &= \max_{1\leq k\leq T} \left(\max_{1\leq j\leq k} \left(P(\Pi, j)\right) - P(\Pi, k) \right) \\ &= \max_{1\leq k\leq T} \left(\max_{1\leq j\leq k} \left(-\Pi' \sum_{t=1}^{j} r_t \right) + \Pi' \sum_{t=1}^{k} r_t \right) \\ &= \max_{1\leq k\leq T} \left(\max_{1\leq j\leq k} \left(-\Pi' R_j\right) + \Pi' R_k \right) \end{aligned}$$

where we have introduced the variable $R_j = \sum_{t=0}^{j} r_t$ as a cumulative uncompounded loss.

As in the previous subsection, we may rewrite $M(\Pi)$ using auxiliary variables, this time as

$$\begin{aligned} z_k &\geq -\Pi' R_k \\ z_k &\geq z_{k-1} \end{aligned} \tag{9.24}$$

for $k = 1, \ldots, T$, and $z_0 = 0$.

The addition of a maximum drawdown constraint to (9.21) becomes, then,

$$\begin{aligned} \min_{(\Pi,\ z)} \Pi' \hat{r} & \\ A\Pi &= b \\ C\Pi &\geq d \\ z_k + \Pi' R_k &\leq \omega \\ z_k + \Pi' R_k &\geq 0 \\ z_k - z_{k-1} &\geq 0, \end{aligned} \tag{9.25}$$

where, again, $z_0 = 0$, and k ranges from 1 to T. Notice that $z_k + \Pi' R_k$ stands in for $D(\Pi, k)$. In (9.25), each $D(\Pi, k)$ is constrained by some upper bound, ensuring the maximum drawdown is likewise bounded. The formulation of average drawdown constraints is similarly handled and the exercise is left to the reader.

These same auxiliary variables may be used to formulate a β-CDaR constraint. The approach should recognizable from the β-CVaR case. Here, we leverage (9.20)

$$D_\beta(\Pi, \delta) \quad = \quad \delta + \frac{1}{1-\beta} \frac{1}{T} \sum_{k=1}^{T} \left[\max_{1 \leq j \leq k} (-\Pi' R_j) + \Pi' R_k - \delta \right]_+$$

which now becomes

$$\begin{aligned} \delta + \frac{1}{1-\beta} \frac{1}{T} \sum_{k=1}^{T} u_k & \\ u_k &\geq 0 \\ u_k &\geq z_k + \Pi' R_k - \alpha \\ z_k &\geq -\Pi' R_k \\ z_k &\geq z_{k-1}, \end{aligned} \tag{9.26}$$

with, again, $z_0 = 0$, and for $k = 1, \ldots, T$. A β-CDaR constrained version of

(9.21) may now be given as

$$\begin{aligned} \min_{(\Pi,\,\delta,\,u,\,z)} \quad & \Pi'\hat{r} \\ A\Pi &= b \\ C\Pi &\geq d \\ \delta + \frac{1}{1-\beta}\frac{1}{T}\sum_{k=1}^{T} u_k &\leq \omega \\ u_k &\geq 0 \\ u_k - z_k - \Pi' R_k + \alpha &\geq 0 \\ z_k + \Pi' R_k &\geq 0 \\ z_k - z_{k-1} &\geq 0. \end{aligned} \tag{9.27}$$

Checklov, Uryasev, and Zabarankin [7] establish, just as with β-CVaR, that the optimization with β-CDaR constraints results in δ^* being the β-DaR and $D_\beta(\Pi^*, \delta^*)$ the β-CDaR of the portfolio, Π^*.

In both examples presented here, we have considered a single path $\{r_t\}_{t=1}^T$. While it is perhaps reasonable to use, say, a historical set of losses for such a path, generally speaking this is myopic with respect to the distribution of losses; viz., the maximum drawdown of one sample path may not be indicative of an expected maximum drawdown. The above work may be easily modified to account for multiple sample paths, and we leave this as an exercise for the reader. In addition to this being an exercise, we suggest that in practice this method is preferred; i.e., defining a random process for r_t from which many sample paths may be used and incorporated in, for example, (9.27).

9.2.3 Sampling from r

One of the standard tools for the sampling problem required for β-CVaR is the theory of copulas, first seen in (3.16). Recall that the copula framework allows for the flexibility of specifying marginal distributions as well as joint distributions.

Based on these results, we may construct joint distributions, $F(\cdot)$ and marginals, $\{F_i(\cdot)\}$ based on *a priori* views. For example, in equity returns, we have frequently emphasized a preference for using a Student t distribution with five degrees of freedom for both marginals as well as the joint distribution of returns as this allows for excess kurtosis. In the present case, we note that such a choice reduces the β-CVaR optimization constraint to a variance constraint as the distribution is elliptical. Even so, two reasons may justify the slight complexity added over using a normal distribution.

First, if a portfolio is managed to a β-CVaR level, we have already seen ample evidence that normal distributions are inadequate in capturing tail behavior and will thus underestimate tail risk. So that while percentiles of the loss distribution map identically to portfolio variance in both the Student t and

normally distributed case, the former gives a better measure of expected tail losses.

Second, it is not uncommon to consider portfolios with both equities and securities based on these equities; viz., equity options. In this case, changes in option security prices are nonlinear in changes in the respective underlying equity prices. As such, an elliptical distribution of returns for equities does not translate to an elliptical distribution for the portfolio loss random variable. This highlights a key differentiator between a β-CVaR optimization procedure and a mean-variance one. The next example considers this with some more specificity.

Example 9.2.1. Suppose that a portfolio is to be constructed with positions in some fixed set of equities, p_E and options written on those equities, p_O. For our present purposes, it is sufficient to note that it is common to estimate changes in option prices via Taylor expansions in the underlying stock price; viz.,

$$\Delta v_{O,j}(\Delta v_{E,j}) \approx \frac{\partial v_{O,j}}{\partial v_{E,j}} \Delta v_{E,j} + \frac{1}{2} \frac{\partial^2 v_{O,j}}{\partial v_{E,j}^2} \Delta v_{E,j}^2, \tag{9.28}$$

where $v_{E,j}$ and $v_{O,j}$ denote the current price of the j^{th} equity and option, respectively.[3]

If equity returns, ζ are sampled from $St_{\mu,\Sigma;\nu}(\cdot)$, with $\nu = 5$, and μ and Σ matching the sample moments of the historical returns, then for sample ζ_i, the associated loss variable for the j^{th} equity is exactly

$$r_{i,E,j} = -\Delta v_{i,E,j} = -\zeta_{i,j} \cdot v_{E,j}.$$

The loss variable for the associated option in this same sample is then

$$r_{i,O,j} = -\Delta v_{i,O,j} = \frac{\partial v_{O,j}}{\partial v_{E,j}} r_{i,E,j} - \frac{1}{2} \frac{\partial^2 v_{O,j}}{\partial v_{E,j}^2} r_{i,E,j}^2.$$

A full accounting of loss under the simulation driven by ζ_i may be achieved in the single vector,

$$r_i = \begin{pmatrix} r_{i,E} \\ r_{i,O} \end{pmatrix}$$

From here, it should be clear how to construct a β-CVaR constrained portfolio. Notice that significant information would be lost if one were to focus solely on variance as a risk proxy as the distribution of portfolio losses is not ellipsoidal.

We note that we have not covered the needed theory to introduce a serial relationship in r_t. That is, all of the above assumes iid samples from the distribution of portfolio losses. To more carefully treat the drawdown constraints considered in this chapter, it is necessary to introduce some time series dynamics into the processes involved.

[3]The interested reader may look at the Black-Scholes formula which gives a closed form solution for so-called European call and put options, which give the buyer the right (but not the obligation) to buy or sell a stock, respectively, for a specified price on a specified future date.

Exercises

1. Show that the expectation operator is a coherent measure of risk. Discuss why this might not be an appropriate risk measure, and note that as a result the conditions for a coherent measure of risk may only be construed as necessary, but not sufficient, for risk management.

2. Prove that variance is not Subadditive.

3. Prove that $\rho_\gamma(\cdot)$, as defined by (9.2), satisfies Positive Homogeneity, Translation Invariance, and Subadditivity.

4. Prove that if r is elliptically distributed (not just normally distributed), then $\rho_\gamma(\cdot)$ is a coherent measure of risk and coincides with Value-at-Risk for some β. Expressly determine this β.

   (a) Suppose that $r = B \cdot f + \epsilon$ for some vector of stochastic factors, $f \in \mathbb{R}^M$ and idiosnycratic component, $\epsilon \in \mathbb{R}^N$. If both f and ϵ are elliptically distributed, prove that ρ_γ is a coherent measure of risk when profit and loss are given by r.

5. Show that a convex combination of coherent measures of risk is itself a coherent measure of risk.

6. Suppose $\rho(\cdot)$ satisfies all conditions of a coherent measure of risk but Translation Invariance, but that $\rho(\Pi + p_f) = \rho(\Pi)$. Prove that $\tilde{\rho}(\Pi) = \mathbb{E}(\Pi) + \gamma\rho(\Pi)$ is a coherent measure of risk for any $\gamma > 0$.

7. Finish the verification that β-VaR satisfies Translation Invariance.

8. Prove that β-CVaR is a coherent measure of risk.

9. Show that G_β as given in (9.11) is convex in α for fixed Π.

10. Prove that both maximum drawdown and average drawdown as defined by (9.14) and (9.15), respectively, satisfy Positive Homogeneity and Monotonicity. Give an examples of maximum drawdown failing Translation Invariance and Subadditivity.

11. Prove that in the continuous case, β-CDaR satisfies all conditions of being a coherent measure of risk but Translation Invariance, where $\Delta_\beta(\Pi+p_f) = \Delta_\beta(\Pi)$. You may assume $\delta_\beta(\Pi)$ satisfies the Monotonicity condition and that (9.18) holds. How might you modify a risk measure based on β-CDaR to make it a coherent measure of risk?

12. Rewrite (9.23) in matrix form.

    (a) How would you modify the problem to incorporate an initial portfolio, Π_0, and turnover constraints?

13. How might you tell if the β-CVaR constraint in (9.23) is binding? If it is not binding, how might this change the interpretation of α^*?

14. Write the average drawdown analogue of (9.25).

15. What are the limiting cases for β-CDaR as $\beta \downarrow 0$ and $\beta \uparrow 1$?

16. Focusing on the maximum drawdown constrained problem (9.25), suppose M sample paths are drawn rather than just one. Write an optimization problem constraining the average maximum drawdown over these M sample paths.

17. Following the methodology in Chapter 2, select the fifty largest stocks in the cross-sectional and historical return data on the last date available. Using the full 121 weeks of returns available, simulate 5,000 samples from a Student t distribution using the sample mean and covariance and five degrees of freedom as parameters. Carry out a β-CVaR minimization with $\beta = 0.95$. Include a no short sale constraint, maximum position of 5% constraint, and gross notional of 1. Bound the β-CVaR by the β-VaR from the sample assuming an evenly weighted portfolio. For the optimal portfolio:

    (a) What are the β-VaR and β-CVaR values from the optimizer?

    (b) How do these compare to the β-VaR and β-CVaR values from the sample?

    (c) How does the β-VaR compare to analytic approximation given by using $\rho_\gamma(\cdot)$?

Chapter 10

Covariance Modifications

Just as we may alter our particular risk measure, we may also focus directly on the covariance matrix. Of course, as a driver for the elliptical distributions we have seen, this may be an effort which has multiple avenues of impact; e.g., directly in the objective function in a mean-variance setting, or perhaps as the driver of returns in a mean-CVaR setting. In any event, the importance and centrality of the covariance in much of what we have covered should be clear. Furthermore, the difficulties surrounding the sample covariance matrix, in terms of estimation, variation through time, and as an input in an objective function of an optimization problem have been made clear.

In this chapter, we consider several modifications to the covariance matrix; all of which are biased estimators of the covariance, and rely on varying amounts of predefined structure. First, we will formally outline the so-called *factor model* approach. Here, structure is explicit and full. A similar modification relies on a combination of the sample and some factor model based estimator. Broadly, these are *shrinkage estimators*, and while the literature is very robust, we will focus on a small subset of these results. Finally, as we have seen that constrained mean-variance optimization generally outperforms its unconstrained counterpart, we present a modification to the covariance unifying these observations via a *constrained maximum likelihood* correspondence.

10.1 Factor Models

In the standard factor model framework, we write

$$r_t = \sum_{k=1}^{K} \beta_k f_{k,t} + \epsilon_t, \tag{10.1}$$

with $f_k t$ the k^{th} *common factor* as observed at time t, r_t the individual stock return being examined (sometimes with a difference of risk-free rate taken out), β_k the *factor loading* on the k^{th} common factor for this individual stock, and

ϵ_t the idiosyncratic error under the model. When several stocks (or, more generally, securities) are considered a subscript of i is included as $r_{t,i}$, $\beta_{k,i}$, and $\epsilon_{t,i}$ to indicate reference to the i^{th} stock.

We have already encountered such models by way of CAPM. Further, the relation to the standard OLS framework of (4.11) and (4.12) is immediate, the slight change of notation to $f.$ notwithstanding. Of course, this interpretation makes a tacit assumption that the model is fit using a time series. More on this below.

In addition to the single factor model of CAPM [22, 31], another very well-regarded factor model is the Fama-French extension to CAPM [10]. This model resolves the apparent outperformance after controlling for market exposure using CAPM of small market cap stocks to large and higher value stocks to lower as measured by the cross-section of book to price. This is achieved, consistent with our approach so far, by accounting for these exposures via the regression

$$r_t - r_f = \beta_m(m_t - r_f) + \beta_h h_t + \beta_v v_t + \epsilon_t. \tag{10.2}$$

The particular method of construction of the time series for the common factors, h_t and v_t, may be seen in the original paper. Broadly, they are constructed via cross-sectional rank ordered portfolios, controlling for remaining factor exposures; viz., h_t is determined as the average return (in the cross-section) of 'small high value' and 'big high value' minus the average return (again, in the cross-section) of 'small low value' and 'big low value', where 'small' and 'big' are with respect to the median market capitalization of the cross-section at that particular time.

In every case, we assume that the mean and covariance of the common factors is constant over time. Writing $\mathbf{f}_t = (f_{1,t}, \ldots, f_{K,t})'$, this implies

$$\mathbb{E}(\mathbf{f}_t) = \mu_f \tag{10.3}$$

$$Cov(\mathbf{f}_t) = \Omega_f \tag{10.4}$$

for all t. We will also make a distinction when the β's of (10.1) are fit using observations obtained from a time series for any given stock, and when the Gauss-Markov assumptions hold for each of these regressions (not necessarily with the distributional assumption). As an example of the distinction when fitting a time series regression, we will assume that $Cov(f_{k,t}, \epsilon_{i,t}) = 0$ for all k, t, and i.

In addition to these usual assumptions, we also assume that the idiosyncratic components for stock i and j are uncorrelated; i.e.,

$$Cov(\epsilon_{i,t}, \epsilon_{j,t}) = \begin{cases} \sigma_i^2 & i = j \\ 0 & i \neq j \end{cases} \tag{10.5}$$

where a lack of time dependence is implicit in the definition.

Notice that if a constant factor is included, the generalization of α as in the CAPM model with intercept, (4.4), is immediate.

10.1.1 Time Series Models: Observed Common Factors

For stocks $i = 1, \ldots, N$, each of which is fit to (10.1) using a time series, $t = 1, \ldots, T$, the model now specifies (omitting a particular time subscript),

$$r = \mathbf{B}\mathbf{f} + \epsilon, \tag{10.6}$$

where r is the vector of returns for the cross-section, $(r_1, \ldots, r_N)$, and $B \in \mathbb{R}^{NxK}$ is the matrix of factor loadings,

$$\mathbf{B} = \begin{pmatrix} — & \beta_1 & — \\ & \vdots & \\ — & \beta_N & — \end{pmatrix}, \tag{10.7}$$

with $\beta_i = (\beta_{1,i}, \ldots, \beta_{K,i})'$ for each i.

Under the assumptions of the model, we have that the covariance of the cross-section is simply

$$Cov(r) = \mathbf{B}\Omega_f\mathbf{B}' + \mathbf{D}, \tag{10.8}$$

where $D = diag(\sigma_1^2, \ldots, \sigma_N^2)$. Further, the OLS estimates of β_i and residual variance, σ_i^2 given by $\hat{\beta}_i$ and s_i^2, respectively, for each i, yield the unbiased estimator of the covariance

$$\hat{Cov}(r) = \hat{\mathbf{B}}\hat{\Omega}_f\hat{\mathbf{B}}' + \hat{\mathbf{D}}, \tag{10.9}$$

where in addition to inputting estimates, the unbiased estimator of the factor covariance, $\hat{\Omega}_f$ is also used. In general, the number of common factors is far less than the number of securities, $K << N$, remedying in small part the issue of insufficient observations in time normally encountered when estimating the sample (the so-called large N, small T problem).

Portfolio variance using (10.9) may be obtained directly, and the exercise is left to the reader. Similarly, if one common factor is simply a vector of ones, the model allows for an interpretation of excess return after controlling for factor exposure.

In the case that every stock has the same range of observed returns, a single multivariate regression may be performed. To see this, let

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & & r_{1,N} \\ \vdots & \ldots & \vdots \\ r_{T,1} & & r_{T,N}, \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} f_{1,1} & & f_{1,K} \\ \vdots & \ldots & \vdots \\ f_{T,1} & & f_{T,K}, \end{pmatrix} \tag{10.10}$$

and let $\mathbf{B}$ be as before. Finally, define $\mathbf{E}$ based on the residuals $\epsilon.$ similarly. Then the simultaneous system of equations given by (10.11) for every stock i is

$$\mathbf{R} = \mathbf{F}\mathbf{B} + \mathbf{E}. \tag{10.11}$$

One may verify that the usual approach taken in the OLS setting – namely, a modification of (4.14) – solves the matrix equation under the Gauss-Markov assumptions.

10.1.2 Cross-Sectional Models: Observed Factor Loadings

The model may also be fit via the cross-section at a specific time. In particular, using the same variables, but adapting to the use in the cross-section, we may write, now assuming that $\mathbf{B}$ is given rather than $\mathbf{f}.$,

$$r_t = \mathbf{B}\mathbf{f}_t + \epsilon_t. \tag{10.12}$$

In this model, the common factors are unobserved and estimated from security features at a given time. Of note is that the residual covariance is no longer homoscedastic. In particular, we must assume (10.5) holds. If residual variances, σ_i^2, are assumed known, then, using the GLS formula (4.49), we have

$$\hat{\mathbf{f}}_t = \left(\mathbf{B}'\mathbf{D}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{D}^{-1}r_t. \tag{10.13}$$

The OLS estimate will necessarily be biased.

Given $\left\{\hat{\mathbf{f}}_t\right\}_{t=1}^{T}$, an estimate of the covariance of unobserved common factors is given by the sample,

$$\hat{\Omega}_f = \frac{1}{T-1}\sum_{t=1}^{T}\left(\hat{\mathbf{f}}_t - \bar{\mathbf{f}}\right)\left(\hat{\mathbf{f}}_t - \bar{\mathbf{f}}\right)', \tag{10.14}$$

where $\bar{\mathbf{f}}$ is the sample mean,

$$\bar{\mathbf{f}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{f}}_t. \tag{10.15}$$

For the cross-sectional model with unobserved common factors, the estimate of the covariance is again given by (10.9), but with (10.14) as input for $\hat{\Omega}_f$.

The astute reader may notice that the residual variances used in the formulation of (10.13) are not directly observable, but nonetheless needed for this model. These variances may be estimated from a simple OLS regression to obtain a time series of residuals $\{\hat{\epsilon}_t\}_{t=1}^{T}$, and sample variances may be obtained from the time series for each security in the cross-section. The diagonal matrix of sample variances $\hat{\mathbf{D}}$ may then be used in (10.13).

Finally, one may show that the common factors given in (10.13) may be interpreted as the portfolio return where the portfolio weights are the solution to a minimum variance problem (using $\mathbf{D}$ as covariance) subject to each portfolio factor β being one. The exercise, again, is left to the reader.

10.1.3 Statistical Factor Models: Principal Component Analysis

In the previous subsections, either common factors or factor loadings were observed. In either case, the choice of factors was determined *a priori*. Presently we consider a factor model determined from the time series of returns directly. Our previous results give an indication of an attractive set of factors; namely, the eigenportfolios of (3.22) which we include again here for ease.

For $\hat{\Sigma} \in \mathbb{R}^{N \times N}$, the sample covariance matrix of the observed time series, $\mathbf{R}$, given in (10.10), we denote the eigenvalues and eigenportfolios or $\hat{\Sigma}$ as

$$\lambda_1 \geq \cdots \geq \lambda_N \geq 0$$

and

$$e_1, \ldots, e_N,$$

respectively. Using the same dimension reduction technique exhibited in (3.25), we may choose K eigenportfolios explaining some prescribed fraction, τ, of the total variance as $\{e_k\}_{k=1}^{K}$. The statistical common factor for this choice is then

$$\mathbf{f}_T = \begin{pmatrix} e_1' r_T \\ \vdots \\ e_K' r_T \end{pmatrix}, \tag{10.16}$$

with $r_T = (r_{T,1}, \ldots, r_{T,N})'$, and notation reflecting the time window dependence of the definition.

Exactly as in the time series factor model approach given previously, factor loadings per security may be found via (10.1) using these K statistical factors, and the resulting factor model covariance is identical to that given in the time series with common factors.

One feature of the factors used in this statistical approach is that f_k and f_j are orthogonal by construction, with a key benefit being that attribution of returns is non-overlapping. Of course, this feature may be achieved with any set of chosen factors using stepwise regressions. That exercise is left to the reader.

10.2 Shrinkage Estimators

In the previous section, methods for developing well structured alternatives to the sample covariance were exhibited. An alternative stance would be to retain some information content in the sample covariance itself; viz., using a convex combination of the sample and, say, a time series model with observed common factors. Ledoit and Wolf [19] provide an elegant solution to just such an approach.

For a sample covariance matrix, $\hat{\Sigma} \in \mathbb{R}^{N \times N}$, based on returns $\mathbf{r}$ as in (10.6) with T observations and a structured covariance matrix alternative estimate (as, for example, in the preceding sections), $\hat{\Omega}$, the *shrinkage estimator for the covariance* is defined by

$$\Sigma_s = (1 - \alpha)\hat{\Sigma} + \alpha\hat{\Omega}. \tag{10.17}$$

As Ledoit and Wolf note, determing $\alpha \in [0, 1)$ is the technically challenging part. We will follow their approach here.

Throughout, we will use continue to use notation for the entries of a given covariance matrix with row-and-column subscripts; viz., the ij^{th} entry of $\hat{\Sigma}$ ($\hat{\Omega}$) will be denoted s_{ij} (ω_{ij}), and the i^{th} diagonal element will be denoted as s_i^2 ($\hat{\omega_i}^2$).

To begin, we define the Frobenius norm on an $N \times N$ matrix, A, with entries a_{ij}, as,

$$\|A\|_F^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}^2. \tag{10.18}$$

One may show (and the exercise is left to the reader) that

$$\|A\|_F^2 = tr(A^2) = \sum_{i=1}^{N} \lambda_i^2, \tag{10.19}$$

where $\{\lambda_i\}_{i=1}^N$ are the eigenvalues of A. If we assume that the true covariance of returns is denoted by Σ, we may define a loss function in α by

$$L(\alpha) = \|(1-\alpha)\hat{\Sigma} + \alpha\hat{\Omega} - \Sigma\|_F^2. \tag{10.20}$$

That is, we would like to minimize the distance in Frobenius norm between the shrinkage estimator, Σ_s, as a function of α, and Σ. This loss function is a random variable, and so we focus on the expectation,

$$\mathbb{E}(L(\alpha)) = R(\alpha) = \mathbb{E}\left(\|(1-\alpha)\hat{\Sigma} + \alpha\hat{\Omega} - \Sigma\|_F^2\right). \tag{10.21}$$

As in a previous exercise, we may see that

$$\begin{aligned} R(\alpha) &= \sum_{i=1}^{N}\sum_{j=1}^{N}(1-\alpha)^2 Var(s_{ij}) + \alpha^2 Var(\hat{\omega}_{ij}) + 2\alpha(1-\alpha)Cov(\hat{\omega}_{ij}, s_{ij}) \\ &\quad +\alpha^2(\omega_{ij} - \sigma_{ij})^2, \end{aligned}$$

where $\mathbb{E}(\hat{\Omega}) = \Omega$, with entries ω_{ij} and similarly with Σ and σ_{ij}.

By taking a first deriviate of $R(\alpha)$ and solving for the stationary point, the optimal α is found to be

$$\alpha^* = \frac{\sum_{i=1}^N \sum_{j=1}^N Var(s_{ij}) - Cov(\hat{\omega}_{ij}, s_{ij})}{\sum_{i=1}^N \sum_{j=1}^N Cov(\hat{\omega}_{ij}, s_{ij}) + (\omega_{ij} - \sigma_{ij})^2}. \tag{10.22}$$

This, of course, is elegant but not ready for application as Ω and Σ – and thus ω_{ij} and $\sigma_i j$ – are not known. A consistent estimator for α^* is needed, and, indeed, Ledoit and Wolf establish that

$$\hat{\alpha}^* = \frac{1}{T}\frac{p-q}{c}, \tag{10.23}$$

for p and c defined as

$$\begin{aligned} p &= \frac{1}{T}\sum_{t=1}^{T}\left((r_{i,t} - \hat{\mu}_i)(r_{j,t} - \hat{\mu}_j) - s_{ij}\right)^2, \\ c &= \sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{\omega}_{ij} - \hat{\sigma}_{ij})^2. \end{aligned}$$

These values are independent of the choice of structured covariance matrix, Ω. The calculation of q, being a consistent estimator for the term $\sum_{i=1}^{N}\sum_{j=1}^{N} Cov(\hat{\omega}_{ij}, s_{ij})$, is dependent on this choice, however. In addition, it is unlikely, but possible, that α^* may not be in $[0,1]$. In these cases, a simple truncation is needed in practice; that is, $\alpha^* = \max\left(0, \min\left(\frac{1}{T}\frac{p-q}{c}, 1\right)\right)$.

We consider two structured covariance matrices in turn: first a CAPM-based single factor model; and, second, a multiple of the identity.

10.3 Constant Correlation Target

In balancing structure and data, an exemplar of a shrinkage target focusing on simplicity is that of a constant pairwise correlation matrix; viz., $\omega_{ij} = \bar{\rho}\sigma_i\sigma_j$ if $i \neq j$, otherwise $\omega_{ii} = \sigma_i^2$, using the notation in the present section. Here, $\bar{\rho}$ is simply the average

$$\bar{\rho} = \frac{2}{(N-1)N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\rho_{ij},$$

where ρ_{ij} is the usual correlation statistic between returns r_i and r_j. The sample of this average is given as one might expect as

$$\hat{\rho} = \frac{2}{(N-1)N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij},$$

which results in population and sample covariance matrix targets,

$$\mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & & \bar{\rho}\sigma_1\sigma_N \\ \vdots & \cdots & \vdots \\ \bar{\rho}\sigma_1\sigma_N & & \sigma_N^2 \end{pmatrix}, \quad \hat{\mathbf{\Omega}} = \begin{pmatrix} s_1^2 & & \hat{\rho}s_1 s_N \\ \vdots & \cdots & \vdots \\ \hat{\rho}s_1 s_N & & s_N^2 \end{pmatrix},$$

respectively.

Under these assumptions, Ledoit and Wolf [20] show that

$$\begin{aligned} q &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left((r_{i,t}-\hat{\mu}_i)^2 - s_i^2\right)^2 \\ &\quad + \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\frac{\bar{\rho}}{2}\left(\frac{s_j}{s_i}\zeta_{ii,ij} + \frac{s_i}{s_j}\zeta_{jj,ij}\right), \end{aligned}$$

where

$$\begin{aligned} \zeta_{ii,ij} &= \sum_{t=1}^{T}\left((r_{i,t}-\hat{\mu}_i)^2 - s_i^2\right)\left((r_{i,t}-\hat{\mu}_i)(r_{j,t}-\hat{\mu}_j) - s_{ij}\right) \\ \zeta_{jj,ij} &= \sum_{t=1}^{T}\left((r_{j,t}-\hat{\mu}_j)^2 - s_j^2\right)\left((r_{i,t}-\hat{\mu}_i)(r_{j,t}-\hat{\mu}_j) - s_{ij}\right). \end{aligned}$$

10.4 Shrinking to CAPM

For the oft-revisited CAPM framework,

$$r_{i,t} - r_f = \alpha_i + \beta_i(m_t - r_f) + \epsilon_{i,t},$$

the structured covariance matrix is of the familiar form

$$\Omega = \sigma_m^2 \beta\beta' + \mathbf{D},$$

with estimator,

$$\hat{\Omega} = \sigma_m^2 \hat{\beta}\hat{\beta}' + \hat{\mathbf{D}},$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_N)'$ (and similarly for $\hat{\beta}$).

In this case, and again, Ledoit and Wolf [20] prove

$$q = \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{s_{jm}}{s_m^2}\zeta_{im,ij} + \frac{s_{im}}{s_m^2}\zeta_{jm,ij} - \frac{s_{im}s_{jm}}{s_m^2}\zeta_{m,ij}$$

where

$$\begin{aligned}
\zeta_{im,ij} &= \sum_{t=1}^{T} \left((r_{i,t} - \hat{\mu}_i)(m_t - \hat{\mu}_m) - \hat{\rho}_{im}s_i s_m\right)\left((r_{i,t} - \hat{\mu}_i)(r_{j,t} - \hat{\mu}_j) - s_{ij}\right) \\
\zeta_{jm,ij} &= \sum_{t=1}^{T} \left((r_{j,t} - \hat{\mu}_j)(m_t - \hat{\mu}_m) - \hat{\rho}_{jm}s_j s_m\right)\left((r_{i,t} - \hat{\mu}_i)(r_{j,t} - \hat{\mu}_j) - s_{ij}\right) \\
\zeta_{m,ij} &= \sum_{t=1}^{T} \left((m_t - \hat{\mu}_m)^2 - s_m^2\right)^2,
\end{aligned}$$

with $\hat{\rho}_{km}$ the sample correlation between return $r_{k\cdot}$ and the market return, m_t, and s_m^2 the sample market variance.

10.4.1 Some Comments on Shrinkage

Typically within this text, derivations of results take precedence, while in the establishment of the two example shrinkage targets above, results are shown without proof. These derivations, while direct, are outside the scope of this text; in particular, we have not covered distributional statments (including variance and covariance) for the general entries of a covariance matrix. However, the usefulness of the shrinkage estimator approach cannot be understated. Using historical return data, one may show that general mean-variance optimization problems result in more attractive risk-adjusted returns than, say, using the sample covariance matrix alone with its known inadequacies or any of several standard factor model based approaches.

The interested reader is urged to consult the original reference papers and pursue the statistical background necessary for a full treatment. Further, a generalization of the results shown here is possible, particularly in the identification of q above when Ω has a factor model structure as in (10.8), but this extenstion is left to the reader.

10.5 Constraints as Modifications to the Covariance

Given any covariance matrix as input to a portfolio optimization problem, we have already seen that investors often impose constraints to forcibly yield portfolios that reflect *a priori* position and exposure requirements; viz., no shortsales, diversification, and particular style or sector exposures. Jagannathan and Ma identify the role of some example constraints in acting like a shrinkage estimator for the sample covariance [14]. In a word, they establish why *imposing the wrong constraints helps*.

In this section, we extend the work of Jagannathan and Ma to the case of general constraints, with attention paid to the application of the method as a modifier in the sense of Black and Litterman [3]; i.e., as a way to express investor views via constraints to obtain modified means and covariances (either jointly or individually). In contrast to Black-Litterman, however, we allow for the expression of views as both excess expected returns as well as desired exposures.

We begin by considering the usual constrained mean-variance optimization problem, with input sample covariance matrix, S, and arbitrary return vector m:

$$\begin{array}{ll} \min\limits_{w} & \frac{1}{2} w' S w \\ \text{s.t.} & A w = b \\ & C w \leq d \\ & m' w = \mu^*, \end{array} \tag{10.24}$$

and assume that no matter the choice of covariance matrix, S, we have $S \succeq 0$. The equality and inequality constraints – A and C, respectively – reflect the desired properties of the target portfolio; viz., β-neutral, fully allocated, long or short constraints, or maximum exposure requirements, as we have seen in previous chapters.

The by-now familiar Lagrangian of (10.24) is given by

$$\mathcal{L}(w, \lambda, \eta) = \frac{1}{2} w' S w + \lambda' (A w - b) + \eta' (C w - d) + \lambda_0 (m' w - \mu^*)$$

with gradient

$$\nabla_w \mathcal{L} = S w + A' \lambda + C' \eta + \lambda_0 m$$

and Karesh-Kuhn-Tucker (KKT) conditions (both necessary and sufficient in

this case) for the optimal solution is given by

$$\begin{aligned} \nabla_w \mathcal{L}\left(\omega^*, \lambda^*, \eta^*\right) &= 0 \\ A\omega^* &= b \\ C\omega^* &\leq d \\ m'\omega^* &= \mu^* \\ \eta^* &\geq 0 \\ \eta_i^* \left(C_i\omega^* - d_i\right) &= 0 \end{aligned} \tag{10.25}$$

where C_i is the ith row of the matrix C.

The approach identifies a covariance matrix, $\tilde{\Sigma}$, based on S, such that the solution of (10.24) and the solution to the minimally constrained mean-constrained problem

$$\begin{aligned} \min_w \quad & \frac{1}{2} w'\tilde{\Sigma} w \\ \text{s.t.} \quad & \tilde{\mu}' w = \tilde{\mu}^* \end{aligned} \tag{10.26}$$

are the same.

The unique Lagrange multipliers of (10.25), $(\lambda^*, \eta^*, \lambda_0^*)$, and solution, ω^* of (10.24) will remain fixed in the notation that follows. Further, it is also be helpful to define

$$K = A'\lambda^* + C'\eta^* \tag{10.27}$$

$$\kappa = \lambda^{*\prime} b + \eta^{*\prime} d, \tag{10.28}$$

and notice $\omega^{*\prime} K = \kappa$.

Finally, define

$$\Delta = \frac{1}{\mu^*}(Km' + mK) - \frac{2\kappa}{\mu^{*2}} M \tag{10.29}$$

with $M = mm'$.

The following propositions hold:

Proposition 10.5.1. If ω^* is the solution to (10.24), ω^* is also a solution of (10.26) with $\tilde{\mu} = m$ (and necessarily $\tilde{\mu}^* = \mu^*$) and

$$\tilde{\Sigma} = S + \Delta \tag{10.30}$$

with Δ as in (10.29).

Proposition 10.5.2. If $\tilde{\Sigma}$ given by (10.30) is invertible, then we may identify $\tilde{\Sigma}$ as the solution to a constrained maximum likelihood estimation problem with constraints motivated by (10.24) when returns are assumed to be iid normal and m and S are the sample mean and covariance, respectively.

Proposition 10.5.3. If $S \succeq 0$, then so is $\tilde{\Sigma}$, and if $S \succ 0$, then $\tilde{\Sigma}$ is positive definite on the feasible set defined by (10.24). Since $\tilde{\Sigma}$ is symmetric, it is also therefore a covariance matrix.

The above results may be interpreted as a shrinkage estimator on the input covariance, S.

We next prove the above propositions by first establishing some preliminary maximum likelihood results. Assuming for the proof that returns are jointly normal, $r \sim N(\mu, \Sigma)$, and iid, the log likelihood function is

$$l(\mu, \Sigma) \propto -\frac{T}{2} ln|\Sigma| - \frac{1}{2} \sum_{t=1}^{T} (r_t - \mu)' \Sigma^{-1} (r_t - \mu).$$

This may be written in $\Lambda = \Sigma^{-1}$ when Σ is invertible:

$$l_0(\mu, \Lambda) = -l(\mu, \Sigma^{-1}). \tag{10.31}$$

The optimization of the log likelihood function may be informed by the constraints of both the (so-called) unconstrained and constrained problem above. We may formulate the constraints in (10.24) in Λ via the relationship

$$\omega^* = \tilde{\mu}^* (\tilde{\mu}' \tilde{\Sigma}^{-1} \tilde{\mu})^{-1} \tilde{\Sigma}^{-1} \tilde{\mu} \tag{10.32}$$

as

$$\begin{aligned} \tilde{\mu}^* A \Lambda \mu - (\mu' \Lambda \mu) b &= 0 \\ \tilde{\mu}^* C \Lambda \mu - (\mu' \Lambda \mu) d &\leq 0. \end{aligned}$$

We arrive, finally, at a constrained maximum likelihood problem

$$\begin{aligned} \min_{\mu, \Lambda} \quad & l_0(\mu, \Lambda) \\ \text{s.t.} \quad & \tilde{\mu}^* A \Lambda \mu - (\mu' \Lambda \mu) b = 0 \\ & \tilde{\mu}^* C \Lambda \mu - (\mu' \Lambda \mu) d \leq 0. \end{aligned} \tag{10.33}$$

For ease of exposition, we give the partials of l_0 in each of μ and Λ here:

$$\frac{\partial l_0}{\partial \mu} = T \Lambda \mu - T \Lambda \hat{\mu} \tag{10.34}$$

$$\begin{aligned} \frac{\partial l_0}{\partial \Lambda} = & T((-\Lambda^{-1} + \hat{\Sigma} + (\mu - \hat{\mu})(\mu - \hat{\mu})')) \\ & -\frac{T}{2} diag(-\Lambda^{-1} + \hat{\Sigma} + (\mu - \hat{\mu})(\mu - \hat{\mu})'). \end{aligned} \tag{10.35}$$

The Lagrangian of (10.33) is

$$\begin{aligned} \mathcal{L}(\mu, \Lambda, \xi, \delta) = & \; l_0(\mu, \Lambda) + \xi' \left(\tilde{\mu}^* A \Lambda \mu - (\mu' \Lambda \mu) b \right) \\ & + \delta' \left(\tilde{\mu}^* C \Lambda \mu - (\mu' \Lambda \mu) d \right), \end{aligned} \tag{10.36}$$

with partials in μ and Λ

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mu} &= T\Lambda\mu - T\Lambda\hat{\mu} + \tilde{\mu}^*\Lambda(A'\xi + C'\delta) \\
&\quad -2(\xi' b + \delta' d)\Lambda\mu \qquad (10.37) \\
\frac{\partial \mathcal{L}}{\partial \Lambda} &= T((-\Lambda^{-1} + \hat{\Sigma} + (\mu - \hat{\mu})(\mu - \hat{\mu})')) \\
&\quad +\tilde{\mu}^* (\xi' A\mu' + \mu A'\xi + \delta' C\mu' + \mu C'\delta) \\
&\quad -2\mu\mu'(\xi' b + \delta' d) \\
&\quad -\frac{T}{2} diag(-\Lambda^{-1} + \hat{\Sigma} + (\mu - \hat{\mu})(\mu - \hat{\mu})') \\
&\quad -\frac{\tilde{\mu}^*}{2} diag\left(\xi' A\mu' + \mu A'\xi + \delta' C\mu' + \mu C'\delta\right) \\
&\quad +diag(\mu\mu'(\xi' b + \delta' d)) \qquad (10.38)
\end{aligned}$$

Begin by setting $\tilde{\mu}^* = \mu^*$, and consider, for $l_m(\Lambda) = l_0(m, \Lambda)$,

$$\begin{aligned}
\min_{\Lambda} \quad & l_m(\Lambda) \\
\text{s.t.} \quad & \mu^* A\Lambda m - (m'\Lambda m)b = 0 \\
& \mu^* C\Lambda m - (m'\Lambda m)d \leq 0.
\end{aligned} \qquad (10.39)$$

The Lagrangian is obtained exactly as in (10.36), replacing μ with m. Letting

$$\Delta(\xi, \delta) = \frac{\mu^*}{T}\left(\xi' Am' + mA'\xi + \delta' Cm' + mC'\delta\right) - \frac{2}{T} M(\xi' b + \delta' d)$$

with

$$M = mm'.$$

We have that the KKT conditions for this problem are, for $\Omega^* = \Lambda^{*-1}$,

$$\begin{aligned}
\Omega^* &= \hat{\Sigma} + \Delta(\xi^*, \delta^*) + (m - \hat{\mu})(m - \hat{\mu})' \\
\Lambda^* & \quad \text{feasible in (10.39)} \\
\delta^* &\geq 0 \\
\delta_i^* \left(\mu^* C_i \Lambda m - (m'\Lambda m) d_i\right) &= 0.
\end{aligned}$$

Next consider

$$\frac{1}{T}\xi^* = \frac{1}{\mu^{*2}}\lambda^*, \quad \tfrac{1}{T}\delta^* = \tfrac{1}{\mu^{*2}}\eta^*$$

with (λ^*, η^*) the optimal Lagrange multipliers from (10.24), where for now we assume that the sample mean and covariance have been used as inputs in the constrained mean-variance optimization problem; that is, $S = \hat{\Sigma}$ and $m = \hat{\mu}$ in (10.24). We next define

$$\tilde{\Sigma} = \hat{\Sigma} + \Delta\left(\frac{T}{\mu^{*2}}\lambda^*, \frac{T}{\mu^{*2}}\eta^*\right).$$

Notice that $\Delta\left(\frac{T}{\mu^{*2}}\lambda^*, \frac{T}{\mu^{*2}}\eta^*\right)$ is the same Δ in (10.29). In terms of K and κ from (10.27) and (10.28), we have

$$\tilde{\Sigma} = \hat{\Sigma} + \frac{1}{\mu^*}(Km' + mK') - \frac{2\kappa}{\mu^{*2}}M,$$

giving

$$\begin{aligned}
\tilde{\Sigma}\omega^* &= \hat{\Sigma}\omega^* + \frac{1}{\mu^*}(Km' + mK')\omega^* - \frac{2\kappa}{\mu^{*2}}M\omega^* \\
&= -K - \lambda_0^* m + \frac{1}{\mu^*}(K\mu^* + m\kappa) - \frac{2\kappa}{\mu^*}m \\
&= \left(-\lambda_0^* - \frac{\kappa}{\mu^*}\right)m.
\end{aligned}$$

Hence ω^* satisfies the funcional form required for the mean-constrained minimum variance problem in $\tilde{\Sigma}$. Since we know that $m'\omega^* = \mu^*$, we conclude that ω^* is a solution to (10.26) proving Proposition 10.5.1 when $S = \hat{\Sigma}$ and $m = \hat{\mu}$ (so that $\Omega^* = \tilde{\Sigma}$ in that case as well).

Now, if $\tilde{\Sigma}$ is nonsingular, the preceding result implies that

$$\omega^* = \mu^*(\tilde{\mu}'\tilde{\Sigma}^{-1}\tilde{\mu})^{-1}\mu^*$$

as in (10.32). With this relationship, verifying the feasibility of $\Lambda^* = \tilde{\Sigma}^{-1}$ in the KKT conditions for the constrained maximum likelihood problem is straightforward. Similarly, since $\delta^* = \frac{1}{\mu^{*2}}\eta^*$, the nonnegativity and complementarity conditions are clear, verifying Proposition 10.5.2

In the case of general S and m, the same construction obtains using the Lagrange multipliers from (10.24). In particular, for

$$\tilde{\Sigma} = S + \frac{1}{\mu^*}(Km' + mK') - \frac{2\kappa}{\mu^{*2}}M$$

the same results follow except that the modified covariance no longer coincides with that from the constrained maximum likelihood problem; i.e., Proposition 10.5.1 holds for general S and m. Notice, too, that the $(m-\hat{\mu})(m-\hat{\mu})'$ term in the modified covariance is necessarily omitted – rather than being simply zero in the case of $m = \hat{\mu}$ – as m is biased in the general case.

We are left to verify $\tilde{\Sigma}$ is a covariance matrix and do so in the general case of m and S as inputs to the original problem. Since symmetry is immediate, we are only left with the question of definiteness. We have, for arbitrary nonzero

w, and $S \succeq 0$,

$$\begin{aligned}
w'\tilde{\Sigma}w &= w'(S+\Delta)w \\
&= w'Sw + \frac{2}{\mu^*}w'Km'w - \frac{2\kappa}{\mu^{*2}}w'Mw \\
&= w'Sw + \frac{2}{\mu^*}w'(-S\omega^* - \lambda_0^* m)m'w - \frac{2\kappa}{\mu^{*2}}w'Mw \\
&\quad \text{(by first-order KKT)} \\
&= w'Sw - \frac{2}{\mu^*}w'S\omega^* m'w - \frac{2\lambda_0^*}{\mu^*}w'Mw - \frac{2\kappa}{\mu^{*2}}w'Mw \\
&\geq w'Sw - \frac{2}{\mu^*}|w'S\omega^*| \cdot |m'w| - 2\left(\frac{\lambda_0^*}{\mu^*} + \frac{\kappa}{\mu^{*2}}\right)w'Mw \\
&\geq w'Sw - \frac{2}{\mu^*}|w'Sw|^{1/2}|\omega^{*\prime}S\omega^*|^{1/2}| \cdot |m'w| - 2\left(\frac{\lambda_0^*}{\mu^*} + \frac{\kappa}{\mu^{*2}}\right)w'Mw \\
&\quad \text{(by Cauchy Schwarz)} \\
&= \left((w'Sw)^{1/2} - \frac{|m'w|}{\mu^*}(\omega^{*\prime}S\omega^*)^{1/2}\right)^2 \\
&\quad -\frac{|m'w|^2}{\mu^{*2}}(\omega^{*\prime}S\omega^*) - 2\left(\frac{\lambda_0^*}{\mu^*} + \frac{\kappa}{\mu^{*2}}\right)w'Mw.
\end{aligned}$$

Now, since

$$\omega^{*\prime}S\omega^* = -\lambda_0^*\mu^* - \kappa,$$

we have

$$\begin{aligned}
\frac{|m'w|^2}{\mu^{*2}}(\omega^{*\prime}S\omega^*) &= \frac{|m'w|^2}{\mu^{*2}}(-\lambda_0^*\mu^* - \kappa) \\
&= -\left(\frac{\lambda_0^*}{\mu^*} + \frac{\kappa}{\mu^{*2}}\right)w'Mw,
\end{aligned}$$

So that

$$\begin{aligned}
w'\tilde{\Sigma}w &\geq \left((w'Sw)^{1/2} - \frac{|m'w|}{\mu^*}(\omega^{*\prime}S\omega^*)^{1/2}\right)^2 - \left(\frac{\lambda_0^*}{\mu^*} + \frac{\kappa}{\mu^{*2}}\right)w'Mw \\
&\geq \left((w'Sw)^{1/2} - \frac{|m'w|}{\mu^*}(\omega^{*\prime}S\omega^*)^{1/2}\right)^2 + \frac{1}{\mu^{*2}}(\omega^{*\prime}S\omega^*)w'Mw.
\end{aligned}$$

Finally, then,

$$w'\tilde{\Sigma}w \geq 0,$$

as desired, with strict inequality on the original feasible set when $\hat{\Sigma} \succ 0$, proving Proposition 10.5.3.

Notice also that for this choice of $\tilde{\Sigma}$, we have that the variance of the optimal portfolio in (10.26) is fixed; viz.,

$$\omega^{*\prime}\tilde{\Sigma}\omega^* = \omega^{*\prime}S\omega^*. \tag{10.40}$$

That is, the shrinkage estimator keeps the calculated variance of the optimal portfolio the same.

10.5.1 Value Constraint Modification

Consider the inclusion of a value constraint in portfolio construction. Denote the set of stocks in the upper decile of EBIT/EV (the value anomaly seen in previous chapters) by $\mathcal{E}$.

Let $c \in \mathbb{R}^N$ be defined by

$$c_i = \begin{cases} 1 & \text{if stock } i \text{ is in } \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{10.41}$$

so that a requirement that $\nu > 0$ percent of portfolio holdings are in the top decile of EBIT/EV may be written as

$$c'w \geq \nu.$$

The resulting statistical arbitrage constrained mean-variance optimization problem is given by

$$\begin{aligned} \min_w \quad & \frac{1}{2} w'Sw \\ & -c'w \leq -\nu \\ & m'w \geq \mu^*. \end{aligned} \tag{10.42}$$

The modified covariance matrix obtained via (10.30) for this constraint is

$$\begin{aligned} \tilde{\Sigma} &= S + \Delta \\ &= S - \frac{\eta^*}{\mu^*}(cm' + mc') + \frac{2\eta^*\nu}{\mu^{*2}} M \end{aligned}$$

Assuming the constraint is binding and hence $\eta^* > 0$, the modified input variance of stock i becomes

$$\tilde{\sigma}_i^2 = s_i^2 - 2\frac{\eta^*}{\mu^*} c_i m_i + \frac{2\eta^*\nu}{\mu^{*2}} m_i^2$$

The second order correction arising from the M term results in an increase of each stock's input variance, proportional to that stock's input squared mean. For stocks in $\mathcal{E}$, however, variance is reduced by $2\frac{\eta^*}{\mu^*}|m_i|$ if the stock's expected return is positive; and increased by the same factor otherwise. Stocks not in the upper decile do not have this linear correction in m_i. In the final analysis, the first-order term results in a model preference for high EBIT/EV names with positive expectation, while the second order term punishes *ex ante* large expectations.

In addition, it may be shown using market data that those portfolios constructed using $\tilde{\Sigma}$ in the minimally constrained portfolio optimization problem have statistically significant exposure to the value factor that was used in its

Covariance	α		β_m		β_v	
S	0.0026	0.0044	0.3982	0.2216	0.1096*	0.1662**
Σ_{LW}	0.0025	0.0044	0.3781	0.1754	0.1119*	0.1864**
$S_{\mathcal{F}_{1,\mathcal{E}}}$	0.0037	0.0057	0.1886	0.1974	0.3822**	0.3185**
$\Sigma_{LW,\mathcal{F}_{1,\mathcal{E}}}$	0.0035	0.0054	0.1462	0.1439	0.4099**	0.3411**

Table 10.1: Exposures of *ex post* returns using the covaraiances listed under the feasible set descriptions given in each panel, rebalancing monthly from October 1997 to July 2015. For each column, data is presented in order for $N = 100$ and $N = 1,000$. Significance of t tests are at 1% and 5% levels: t tests for unmodified estimators are against the null hypothesis of $\beta_v = 0$ (exposure to value) and indicated using $(*)$; t tests for modified estimators are one sided tests of greater exposure against the relative unmodified estimator and indicated using $(\star)$. Statistical significance is only reported for value exposure.

construction. That is, in the balance between structure and data, the method outlined in this section provides a tool to tailor exactly what exposures might be desired in the structured alternative. This may be seen in Table 10.1, where both sample and optimal shrinkage covariance matrices are modified using (10.30) under the constraint sets,

$$\begin{aligned} \mathcal{F}_1 &= \{w \,|\, 1'w = 1,\, w \geq 0\} \\ \mathcal{F}_{1,\mathcal{E}} &= \mathcal{F}_0 \cap \mathcal{E}_\nu. \end{aligned}$$

The data used to construct this table considers $N = 100$ and $N = 1,000$ stocks over 200 monthly dates from 10/31/1997 through 5/31/2014, and is comprised of the 1,000 largest domestic stocks on NYSE and AMXE by market capitalization each month, with a requirement that the share price was greater than \$5 at the close on each month end date. Sample covariance matrices, S, are constructed using 121 weeks of trailing returns at each monthly rebalance, and shrinkage estimators Σ_{LW} are constructed as in the previous section with a constant correlation target; modifications following the value-constraint modification established in (10.42) are denoted by $S_{\mathcal{F}_{1,\mathcal{E}}}$ and $\Sigma_{LW,\mathcal{F}_{1,\mathcal{E}}}$, respectively.

From the table it may be seen that modifying each of the sample, S, and the shrinkage estimator using constant correlation target, Σ_{LW}, results in an *ex post* increase in exposure to value (measured by a regression on a proxy value factor) from 0.1096 andd 0.1119 to .3822 and.4099, respectively, in the case of $N = 100$ stocks. Similar results obtain in the case of $N = 1,000$ stocks where the sample covariance matrix is underdetermined. These increases are statistically significant at the 1% level in each case.

That is, the empirical exercise bolsters the theoretical claim established by the derivations in (10.5.2).

Exercises

1. For a portfolio with weights, w, what is the variance of $w'r$ if the covariance of r is estimated using (10.9)?

2. Prove that the usual approach from OLS solves the matrix equation given in (10.11). What is the unbiased estimate of the residual covariance?

3. Explain why (10.13) holds.

4. Rigorously determine $\hat{\mathbf{D}}$ for use in the cross-sectional model using GLS estimates in (10.13) using the method suggested in the text.

5. The common factors from the cross-sectional model given by (10.13) may be interpreted as factor returns of a constrained quadratic program. Consider

$$\begin{aligned} \min_{w} \quad & \frac{1}{2} w' \mathbf{D} w \\ & \mathbf{B}' w = 1. \end{aligned} \tag{10.43}$$

   (a) Interpret the constrained optimization problem above.

   (b) Solve for w^*.

   (c) Show that $\hat{\mathbf{f}}_t = w^{*'} r_t$ using the definition given in (10.13).

   When the additional constraint $1'w = 1$ is included, the resulting portfolio is called the *factor mimicking portfolio*.

6. Prove that the statistical factors given in (10.16) are orthogonal.

7. Given a set of common factors, $\{f_k\}_{k=1}^K$, provide a constructive method for creating a set of pairwise orthogonal common factors $\{\tilde{f}_k\}_{k=1}^K$. What concerns might you have for your approach, specifically related to time sensitivity?

8. Show that, for a matrix, $A \in \mathbb{R}^{N \times N}$, with eigenvalues $\{\lambda_i\}_{i=1}^N$, the Frobenius norm satisfies:

   (a) $\|A\|_F^2 = tr(A^2)$.

   (b) $\|A\|_F^2 = \sum_{i=1}^N \lambda_i^2$.

9. Find $R'(\alpha)$ and $R''(\alpha)$ for $R(\alpha)$ defined in (10.21).

   (a) Establish (10.22).

   (b) Explain why α^* is a unique minima by examing $R''(\alpha)$.

   (c) Following the methodology in Chapter 2, select the fifty largest stocks in the cross-sectional and historical return data on the last date available. Use the full 121 weeks of returns available to:

i. Shrinking to a constant correlation target, identify the shrinkage target, shrinkage intensity, and shrinkage estimator as established in this chapter.

ii. Compare the distribution of eigenvalues from the sample covariance matrix and your identified shrinkage target. Discuss any observations you may have.

10. Prove (10.32).

11. Prove Proposition (10.5.1).

Bibliography

[1] Marco Avellaneda. Hierarchical pca and applications to portfolio management. *SSRN Electronic Journal*, 2019.

[2] L. Bachelier. Theorie de la speculation. *Annales Scientifiques de l'Ecole Normale Superieure*, 17:21–86, 1900.

[3] F. Black and R. Litterman. Asset allocation: Combining investor views with market equilibrium. *The Journal of Fixed Income*, 1:7–18, 1991.

[4] Jean-Phillipe Bouchard and Marc Potters. *Financial applications of random matrix theory: a short review.* The Oxford Handbook of Random Matrix Theory, 2015.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[6] G. Casella and R.L. Berger. *Statistical Inference.* Duxbury Press, Pacific Grove, 2nd edition, 2002.

[7] A. Chekhlov, S. Uryasev, and M. Zabarankin. Portfolio optimization with drawdown constraints. pages 209–228. World Scientific Publishing Co. Pte. Ltd., 2004.

[8] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001.

[9] Eugene F Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1992.

[10] Eugene F. Fama and Kenneth French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–466, 1992.

[11] Yahoo Finance. International business machines corporation (ibm). `https://finance.yahoo.com/quote/IBM/`, 2022.

[12] Simon Gilchrist and Egon Zakrajsek. The impact of the federal reserve's large-scale asset purchase programs on corporate credit risk. *Journal of Money, Credit and Banking*, 45(2):29–57, 2013.

[13] William Sealy Gosset. The probable error of a mean. *Biometrika*, 6 (1):1–25, 1908.

[14] R. Jagannathan and T. Ma. Risk reduction in large portfolios: Why imposing the wrong constraint helps. *The Journal of Finance*, 58:1651–1683, 2003.

[15] Jonathan Keehner. How a goldman hedge fund shrank a third in a week. *Financial Times*, Aug14, 2007.

[16] Amir E. Khandani and Andrew W. Lo. What happened to the quants in august 2007? *Journal of Investment Management*, 5:29–78, 2007.

[17] Amir E. Khandani and Andrew W. Lo. What happened to the quants in august 2007?: Evidence from factors and transactions data. *Journal of Financial Markets*, 14:1–46, 2011.

[18] Tze Leung Lai and Haipeng Xing. *Statistical Models and Methods for Financial Markets*. Springer New York, NY, 2008.

[19] O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, 10(5):603–621, 2003.

[20] O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. *Journal of Portfolio Management*, 30(4):110–119, 2004.

[21] David Li. On default correlation: A copula function approach. *Journal of Fixed Income*, 9(4):43–54, 2000.

[22] R.C. Merton. An intertemporal capital asset pricing model. *Econometrica*, 41(5):867–887, 1973.

[23] Attilio Meucci. *Risk and Asset Allocation*, volume 1st ed. Springer, 2005.

[24] Attilio Meucci. A new breed of copulas for risk and portfolio management. *Risk*, 24(9):122–126, 2011.

[25] Federal Reserve Bank of St. Louis. S and p dow jones indices llc, s and p 500 [sp500], retrieved from fred, federal reserve bank of st. louis. `https://fred.stlouisfed.org/series/SP500`, 2022.

[26] Andrew Redleaf and Richard Vigilante. *Panic: The Betrayal of Capitalism by Wall Street and Washington*. Richard Vigilante Books, 2010.

[27] R. Terrell Rockafellar and Stan Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–42, 2000.

[28] Paul A. Samuelson. Proof that properly anticipated prices fluctuate randomly. In *The World Scientific Handbook of Futures Markets, World Scientific Handbook in Financial Economics Series, vol. 5,*, pages 25–38. World Scientific, 2015.

[29] A. Shapiro and Y. Wardi. Nondifferentiability of the steady-state function in discrete event dynamic systems. *Automatic Control, IEEE Transactions on*, 39:1707 – 1711, 09 1994.

[30] A. Shapiro and Y. Wardi. Convergence analysis of stochastic algorithms. *Mathematics of Operations Research*, 21:615–628, 08 1996.

[31] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

[32] A Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris,*, 8:229–231, 1959.