Chapter 2
Single-Period Problems

Risk is a part of God's game, alike for men and nations.
— Warren Buffett

In the single-period portfolio choice problem, the investor is assumed to make allocation decisions once and for all at the beginning of a given period (e.g. one quarter or one year), based on estimated prospects for the risk and return relationships of a universe of N investable assets over the horizon. Once made, the allocation decisions are not allowed to change until the end of the period; the impact of decisions arising in subsequent periods is not considered in this case, and for this reason, single-period problems lead to so-called *myopic* policies. Markowitz (1952) introduced the basic formulation, including expressions for the expected portfolio return and variance in terms of the portfolio weights and expected returns, variances and covariances of individual assets. He also introduced the *efficient frontier* and its depiction on the mean-variance plane. Since the original formulation uses the asset variances (and covariances) as the risk measure, the methodology is often called *mean–variance allocation*.

Despite their original conceptual simplicity, single-period problems are a large topic in which the optimization step is but one aspect. Just as important are the choice of utility function (§2.4/p. 11), risk measures (§2.5/p. 14), problem constraints (§2.6/p. 17) and forecasting models (§2.7/p. 23). Moreover, delicate issues related to the stability and econometrics of the obtained solutions need to be addressed for a successful implementation of the approach (§2.8/p. 29). This entails a rather involved methodology for single-period portfolio choice, which can be summarized by Fig. 2.1.

2.1 Basic Formulation

Let $\mathbf{R}_{t+1} \in \mathbb{R}^N$ be a vector of random *asset returns* between times t and $t+1$ (see §1.2/p. 5 for a summary of the time index conventions). Assume that the investor

N. Chapados, *Portfolio Choice Problems: An Introductory Survey of Single and Multiperiod Models*, SpringerBriefs in Electrical and Computer Engineering 3,
DOI 10.1007/978-1-4614-0577-1_2, © The Author 2011

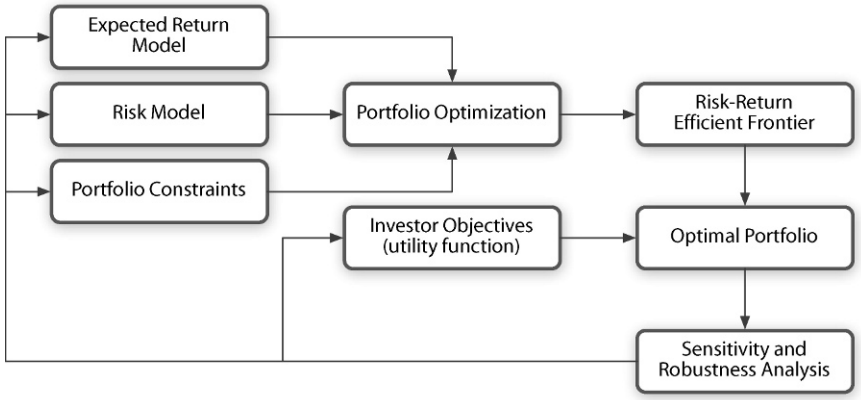

Fig. 2.1 Methodological steps surrounding the Markowitz single-period investment process; adapted from Exhibit 2.2 (p. 21) of Fabozzi *et al.* (2006).

makes, given the information available at time t, a forecast of the first two moments of the distribution of future returns,

$$\mu_{t+1|t} = \mathbb{E}_t\left[\mathbf{R}_{t+1}\right]$$
$$\Sigma_{t+1|t} = \mathrm{Cov}_t\left[\mathbf{R}_{t+1}\right],$$

where the $\mathbb{E}_t[\cdot]$ and $\mathrm{Cov}_t[\cdot]$ denote, respectively, the expectation and covariance matrix of a (vector) random variable conditioned on the information available at time t. For simplicity in this section, since single-period modeling does not explicitly consider the consequences of time, we drop the time subscripts on the above quantities, which we write simply as $\mathbf{R}$, μ and Σ. Likewise, the return on the risk-free asset during the period is denoted by R_f.

The investor allocates its capital among the N assets, forming a portfolio $\mathbf{w} \in \mathbb{R}^N$ where each element $\mathbf{w}_i$, the *weight* of asset i, represents the fraction of total capital held in the asset. The expected portfolio return and variance are given respectively by

$$\mu_P = \mathbf{w}'\mu \qquad \text{and} \qquad \sigma_P^2 = \mathbf{w}'\Sigma\,\mathbf{w}. \tag{2.1}$$

We shall make the following assumptions about the assets:

1. There are no "redundant" assets, i.e. no asset return can be obtained as a linear combination of the returns of other assets.
2. All assets are risky (have positive return variance), which implies, in conjunction with the above assumption, that the covariance matrix Σ is nonsingular. (The inclusion of a risk-free asset is treated in §2.4/p. 11.)

Definition 2 (Efficiency). A portfolio $\mathbf{w}$ is said to be **efficient** if it is the lowest-variance portfolio for a given level of expected return.

The portfolio choice problem seeks to directly find efficient portfolios by determining an "optimal" vector of asset weights. The minimum-variance formulation of the problem considers the expected portfolio variance as the measure of risk. It takes the form

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\Sigma\,\mathbf{w} \tag{2.2}$$

$$\text{subject to} \quad \mathbf{w}'\mu = \rho, \tag{2.3}$$

$$\mathbf{w}'\iota = 1. \tag{2.4}$$

The objective function, eq. (2.2), seeks the vector of weights which minimizes the total expected portfolio variance, subject to constraint (2.3) which requires a portfolio return of ρ (which can be viewed as the desired or target return), and constraint (2.4) which specifies that all capital must be invested. We consider other types of constraints — and their implication on the solution methods — in §2.6/p. 17.

2.2 Solution

Since all constraints are of equality type, problem (2.2) can be solved analytically by introducing Lagrange multipliers. The general solution is derived in §A.1/p. 71. To borrow notation from that section, we set

$$\mathbf{A} = \begin{pmatrix} \mu' \\ \iota' \end{pmatrix} \qquad\qquad \mathbf{b} = \begin{pmatrix} \rho \\ 1 \end{pmatrix},$$

and obtain the optimal weights $\mathbf{w}^*$ by substitution into eq. (A.10). Some algebraic manipulation yields the somewhat simplified but enlightening form, showing the optimal weights $\mathbf{w}^*$ as being linear in the desired return ρ, (Merton, 1972; Fabozzi *et al.*, 2007)

$$\mathbf{w}^* = \mathbf{g} + \mathbf{h}\rho, \tag{2.5}$$

where

$$\mathbf{g} = \frac{\Sigma^{-1}(c\iota - b\mu)}{d}, \qquad\qquad \mathbf{h} = \frac{\Sigma^{-1}(a\mu - b\iota)}{d},$$

and

$$a = \iota'\Sigma^{-1}\iota, \qquad b = \iota'\Sigma^{-1}\mu, \qquad c = \mu'\Sigma^{-1}\mu, \qquad d = ac - b^2.$$

Similarly, the globally minimum-variance portfolio (GMV) is obtained without imposing the expected-return constraint, yielding portfolio weights and variance respectively given by

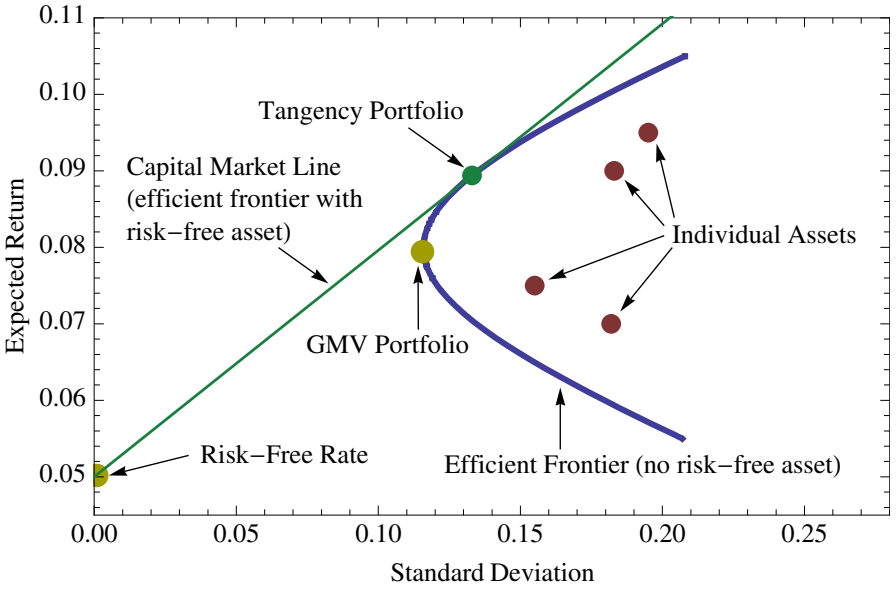

Fig. 2.2 Efficient frontier obtained from four assets specified in the text; the Global Minimum Variance (GMV) portfolio has a lower risk (as measured by the standard deviation of returns) than any individual asset, showing the benefits of diversification.

$$\mathbf{w}^*_{\text{GMV}} = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota} \qquad \text{and} \qquad \sigma^2_{\text{GMV}} = \frac{1}{\iota'\Sigma^{-1}\iota}. \tag{2.6}$$

The above solutions yield two important insights. First, as will be illustrated next, it reflects the benefits of diversification. Second, it highlights that ultimately, higher returns can only be obtained by taking on higher leverage — thence more risk — since the optimal weight vector is linear in the target return ρ.

To illustrate these solutions, consider a four-asset problem specified as

$$\mu = \begin{bmatrix} 0.095 \\ 0.070 \\ 0.090 \\ 0.075 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 0.0380 & 0.0085 & 0.0089 & 0.0066 \\ 0.0085 & 0.0331 & 0.0156 & 0.0039 \\ 0.0089 & 0.0156 & 0.0334 & 0.0070 \\ 0.0066 & 0.0039 & 0.0070 & 0.0240 \end{bmatrix}.$$

The efficient frontier for this example is plotted in Fig. 2.2, under the label "Efficient Frontier (no risk-free asset)".

2.3 Risk-Free Asset, Tangency Portfolio, Separation

When one of the assets can be considered risk-free (i.e. a return variance of zero and necessarily an identically zero covariance with all other assets), the above formulation cannot be used directly since the covariance matrix Σ would not be invertible. In this context, it can be shown that all efficient portfolios are formed by a linear combination of the risk-free asset and the *tangency portfolio* located on the risky-assets efficient frontier. These portfolios are located on what is known as the Capital Market Line (CML). These concepts, for a risk-free rate of 5%, are depicted on Fig. 2.2.

As derived in §A.2/p. 72, the risky-asset proportions of the tangency portfolio, given a risk-free rate R_f, are obtained as

$$\mathbf{w}^{\text{TGP}} = \frac{\Sigma^{-1}(\mu - R_f)}{\iota'\Sigma^{-1}(\mu - R_f)}.$$

A central consequence of the efficiency of all portfolios along the CML is that it is optimal for all investors (who share a common view about μ and Σ) to hold the *tangency portfolio in some proportion*. Investors only differ in their exposure to it, or alternatively, in how they allocate their holdings between the risk-free and tangency portfolio. This result was originally established by Tobin (1958) (see also Merton (1990, ch. 2)) and is an example of *separation* or *mutual fund* theorems.[1]

In the presence of a risk-free asset, portfolio optimization problems can be formulated without insisting on the "sum-to-one" constraint (2.4), since the unallocated fraction of capital, $1 - \mathbf{w}'\iota$, can be invested in the risk-free asset (or assumed to be borrowable at the risk-free rate in the case of a negative fraction).

Geometrically, from Fig. 2.2, the tangency portfolio can also be seen to maximize the *Sharpe ratio* (Sharpe 1966, 1994), defined as the expected portfolio excess return (over the risk-free rate R_f) per unit of portfolio return standard deviation,

$$\text{SR} \triangleq \frac{\mu_P - R_f}{\sigma_P},$$

with μ_P and σ_P given by eq. (2.1). A formal derivation of the relationship between the Sharpe ratio and the tangency portfolio appears in §A.2/p. 72.

2.4 Utility Maximization

Problem (2.2) does not specify what the "appropriate" level of target return ρ should be; this question should be decided by the investor and is a direct function of the risk

[1] This result also serves as a foundation for the celebrated Capital Asset Pricing Model (CAPM), which assumes, among other things, that all investors do share common views about μ and Σ, and examines equilibrium consequences; see §2.7.1/p. 23.

s/he is *willing* and *able* to bear. Markowitz (1959) introduces a formulation wherein the investor's expected utility is directly maximized. He considered the following quadratic form, written in terms of the portfolio return R_P,

$$U_\lambda(R_P) = R_P - \frac{\lambda}{2} R_P^2,$$

where λ is represents the investor's *risk aversion*, and in this context quantifies how the investor is willing to trade each incremental unit of expected return against a corresponding increase in variance of return.[2]

A rational decision maker would seek to maximize its *expected utility*, which is computed as

$$\begin{aligned}\mathbb{E}[U_\lambda(R_P)] &= \mu_P - \frac{\lambda}{2}\sigma_P^2 \\ &= \mathbf{w}'\mu - \frac{\lambda}{2}\mathbf{w}'\Sigma\,\mathbf{w},\end{aligned}$$

where $\mathbf{w}$ is, as above, the weight given on each asset within the portfolio and μ_P and σ_P^2 are respectively the mean and variance of the portfolio return distribution, given by eq. (2.1). The expected quadratic utility maximization problem is then written as

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathbf{w}'\mu - \frac{\lambda}{2}\mathbf{w}'\Sigma\,\mathbf{w} \tag{2.7}$$

$$\text{subject to} \quad \mathbf{w}'\iota = 1. \tag{2.8}$$

When no further constraint is imposed, an analytical solution for $\mathbf{w}^*$ is easily found by introducing Lagrange multipliers, similarly to the solution for problem (2.2).[3]

Proposition 1. *The unconstrained minimum-variance portfolio* (2.2)–(2.3) *and maximum quadratic utility* (2.7) *formulations are equivalent.*

Proof. The equality constraint (2.3) is incorporated in the minimum-variance objective (2.2) through an unconstrained Lagrange multiplier $\nu \in \mathbb{R}$, yielding the problem

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}'\Sigma\,\mathbf{w} - \nu(\mathbf{w}'\mu - \rho),$$

with first-order conditions for optimality given by $\Sigma\,\mathbf{w} - \nu\mu = 0$, yielding optimal solution

$$\mathbf{w}^* = \nu\Sigma^{-1}\mu, \tag{2.9}$$

where ν is found by substitution as $\nu = \frac{\rho}{\mu'\Sigma^{-1}\mu}$.

[2] Many formulations of utility theory focus on the utility of *terminal wealth*, instead of the portfolio return; Markowitz explicitly considers the latter (e.g. Markowitz 1959, p. 208), and this convention is almost universally followed in mean-variance problems. An alternative formulation of quadratic utility in terms of terminal wealth would slightly change the resulting equations.

[3] See, e.g. Chapados (2000) for a derivation.

Consider, on the other hand, the first-order optimality conditions of problem (2.7), $\mu - \lambda \Sigma \mathbf{w} = 0$, yielding optimal solution

$$\mathbf{w}^* = \frac{1}{\lambda} \Sigma^{-1} \mu. \tag{2.10}$$

Comparing eq. (2.9) and (2.10), it suffices to take $\lambda = \mu' \Sigma^{-1} \mu / \rho$ to obtain the equivalence. □

This result confirms that in order to target a higher expected portfolio return ρ, the investor must exhibit a lower risk aversion.

Obviously, quadratic utility is but one of a number of utility functions that have been proposed to model the behavior of economic agents. The more general problem is easily written in terms of expected utility maximization,

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \int_{\mathbf{R}} U(\mathbf{w}'\mathbf{R}) \, \mathrm{d}P(\mathbf{R}), \tag{2.11}$$

subject to the budget constraint (2.8), where $U(\cdot)$ is a utility function and $P(\mathbf{R})$ is the next-period return distribution. In particular, Mossin (1968) proves that constant relative risk aversion (CRRA) functions[4] are the only ones permitted if constant asset proportions are to be optimal, i.e. the investment in the risky asset does not depend on the level of initial wealth. Merton (1969) establishes the same result in a continuous-time setting. Moreover, Campbell and Viceira (2002) strongly argue in favor of CRRA utilities on the basis of the long-run observed behavior of the economy. However, for a large number of utility functions and "reasonable" return distributions, several studies (Levy and Markowitz, 1979; Kallberg and Ziemba, 1983) have established that single-period optimal portfolios under quadratic utility are very close to those obtained under alternative utilities.

A special case of some interest is the logarithmic utility, defined as $U(R) = \log(1+R)$. This utility function is maximized by considering a Taylor series expansion of $1+R$ around $R = 0$,

$$\log(1+R) = R - \frac{R^2}{2} + O(R^3).$$

For relatively small returns, this is seen to be equivalent to the maximization of quadratic utility, problem (2.7), with $\lambda = 1$. The optimal weights under this utility function are given precisely by the tangency portfolio for a risk-free rate of zero (which also maximizes the Sharpe Ratio, see §A.2/p. 72). This property led some authors to confer a special aura to the logarithmic utility as being somehow "better",

[4] For a utility function $U(W)$, the Arrow–Pratt measure of relative risk aversion (Arrow, 1965; Pratt, 1964) is defined as

$$\mathrm{RRA}(W) = -\frac{W U''(W)}{U'(W)}.$$

A CRRA utility function is one for which $\mathrm{RRA}(W)$ is a constant independent of W. Such functions are sometimes said to exhibit *iso-elastic marginal utility*.

a point discussed, and found to be fallacious in a multiperiod setting, by Merton and Samuelson (1974). We return to the logarithmic utility in §3.4/p. 49.

Some utility functions have been proposed to incorporate parameter estimation uncertainty, the subject of *robust optimization*, which is covered in §2.8.5/p. 34.

2.5 Risk Measures

The exposition so far assumes that the investor considers the variance of the portfolio return distribution to be an adequate measure of risk. This measure has the major shortcoming that it considers positive return surprises to be as equally unpleasant as negative return surprises, a property that would surely be dismissed by most real-world investors! A number of alternative measures have been proposed throughout the years that attempt to quantify *portfolio downside risk*, starting with Markowitz's original treatment of the semivariance. This section briefly reviews the most significant possibilities. Nawrocki (1999) surveys the field more extensively.

2.5.1 Semivariance

Semivariance was originally considered by Markowitz (1959, Chapter 9) as a simple measure of downside risk. Whereas the variance is a symmetrical measure, semivariance only considers movements that fall below the mean; as such, its value depends on the *skewness* (third moment) of the distribution. For a scalar random variable X with mean μ, semivariance is defined as

$$\sigma^2_{\min} = \mathbb{E}\left[\min\left[X-\mu,0\right]^2\right].$$

This measure can be used instead of portfolio variance in Problem (2.2). Although there is no closed-form solution to the mean-semivariance problem, Jin *et al.* (2006) establish the existence of the one-period mean-semivariance efficient frontier and review the literature examining its applications. Furthermore, Estrada (2007) provides an approximation to the semivariance that lends itself well to analytical solutions and reports good results on a number of problems.

2.5.2 Roy's Safety First

The Roy (1952) "safety-first" criterion puts portfolio risk in a more concrete setting than Markowitz' consideration of the second moment of returns. As Roy argued, the investor first decides on a minimum acceptable return that would ensure the preservation of a desired portion of his capital; he then proceeds with portfolio opti-

mization by minimizing the probability of experiencing a return below the "disaster level". Let R_0 be the investor's minimum acceptable return and consider the problem

$$\begin{aligned} &\text{minimize} && P(R_P \leq R_0) \\ &\text{subject to} && \mathbf{w}'\iota = 1 \qquad \text{(budget)}. \end{aligned}$$

Since the return distribution probability is not known precisely, this minimization may appear unfeasible. However, by Chebyshev's inequality, we have

$$P(R_P \leq R_0) \leq \frac{\sigma_P^2}{(\mu_P - R_0)^2},$$

which, taking square roots, yields the approximate problem

$$\min_{\mathbf{w}} \frac{\sigma_P}{\mu_P - R_0}$$

subject to the budget constraint. If the R_0 is the risk-free rate, this problem is equivalent to maximizing the Sharpe ratio (Sharpe, 1966).

2.5.3 Value-at-Risk

Value-at-Risk (VaR) was developed by JP Morgan in the early 1990's and made popular in a widely-circulated technical document (RiskMetrics, 1996) and associated software product. Intuitively, the level-α VaR (e.g. $\alpha = 95\%$) of a portfolio over a certain time horizon h is the portfolio return R_P such that the fraction α of returns will be better than R_P over the horizon. More formally, the level-α VaR of a portfolio is defined as the $1-\alpha$-percentile of the portfolio return distribution,

$$\text{VaR}_\alpha(R_P) = -\inf_R \{R : P(R_P \geq R) \geq \alpha\},$$

where all returns are computed over horizon h. (The minus sign in the definition serves to make the risk measure positive.) The location of the VaR of an hypothetical asset return distribution, and its relationship to the CVaR (treated next) is shown in Fig. 2.3.

Value-at-Risk is regarded as a more plausible measure of portfolio risk than the variance since it accounts (in theory) for skewness and kurtosis in the return distribution.[5] In addition to its origins in risk management, it has received wide attention in a portfolio choice context where the VaR simply substitutes for the variance as the risk measure (Alexander and Baptista, 2002; Mittnik *et al.*, 2003; Chow and Kritzman, 2002; Chapados, 2000).

[5] In practice, it is common to compute the VaR under a normal approximation due to its analytical tractability, which of course disregards higher-order moments in the underlying true distribution.

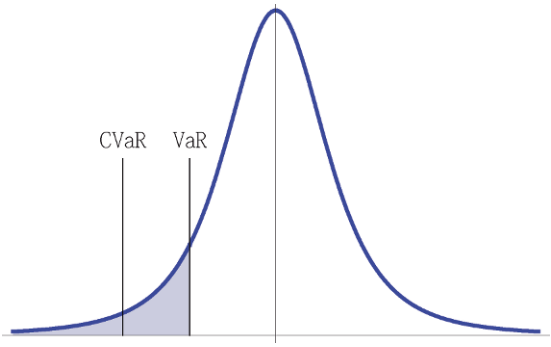

Fig. 2.3 90% Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR) for a Student $t(3)$ distribution. For fat-tailed distributions, the CVaR point can represent an expected loss much more significant than the VaR.

2.5.4 Conditional Value-at-Risk

In spite of its wide use, the VaR, as a measure of risk, suffers from a major defect: its lack of *subadditivity* (Artzner *et al.*, 1999). For a risk measure ρ applied to portfolios P_1 and P_2, subadditivity is satisfied if

$$\rho(P_1 + P_2) \leq \rho(P_1) + \rho(P_2),$$

which is a statement of the benefits of diversification—the risk of a diversified portfolio cannot be more than the risk of any of its constituents. That the VaR does not satisfy this property can lead to a number of counterintuitive results, particularly for firm-wide risk management, where it can appear that a more diversified portfolio exhibits a higher risk (Rau-Bredow, 2004).

A closely related measure that does satisfy subadditivity is the *conditional value at risk* (CVaR)—also called *expected shortfall* or *expected tail loss*—defined as the expected return conditional on observing a return lower than the VaR:

$$\text{CVaR}_\alpha(R_P) = \mathbb{E}[R_P \,|\, R_P < \text{VaR}_\alpha(R_P)],$$

where, as for the VaR, the returns are computed over a given time horizon h. In Fig. 2.3, this corresponds to an expectation taken within the shaded area. In a portfolio context, the CVaR has been studied by Krokhmal *et al.* (2002) and Consigli (2004).

2.5.5 Other Measures

In the past few years, there has been an explosion of alternative risk measures based on the modeling of tail phenomena (e.g. Malevergne and Sornette 2005a). Although it is not our focus to describe them in depth, Rachev *et al.* (2005) provide a good survey of the relevant literature, especially of measures related to portfolio selection.

Farinelli *et al.* (2006) provide computational portfolio allocation results comparing eleven alternative performance measure ratios.

2.6 Additional Constraints

Portfolio optimization problems, regardless of the form of the objective function or type of risk measure, are often solved with a number of constraints that attempt to capture *a priori* knowledge that the analyst possesses on what should be "good" solutions, embody investment objectives of the fund, or comply with regulatory requirements. It should be noted that with most of these constraints, Problem (2.2) can no longer be solved analytically but must instead be tackled with quadratic programming (Luenberger and Ye, 2007; Bertsekas, 2000) or mixed-integer programming (Wolsey and Nemhauser, 1999). Constraints also play a *regularization* role that can serve to mitigate sampling variance and estimation error in the mean return and risk forecasts; this is covered in §2.8/p. 29.

Some of the more common constraints are as follows. More comprehensive treatments appear in Fabozzi *et al.* (2006) and Qian *et al.* (2007). In line with the first reference, the rest of this section makes use of the following notation: we denote the current holdings of an investor by $\mathbf{w}_0$, the target holdings to be invested over the next period (i.e. the variables resulting from optimization) by $\mathbf{w}$, and their difference (the traded amount in each asset) by $\mathbf{x} = \mathbf{w} - \mathbf{w}_0$.[6] Furthermore, let $\mathbf{p}_0$ be the current price vector of the assets, and W_0 the current total portfolio value. The amount to be invested in asset i is given by $W_0\mathbf{w}_i$ and the number of shares[7] held is $n_i = W_0\mathbf{w}_i/\mathbf{p}_{0i}$.

2.6.1 No Short-Sales Constraint

This corresponds to the requirement that all portfolio weights be non-negative, namely

$$\mathbf{w}_i \geq 0, \quad \text{for all } i,$$

[6] The absolute traded amount, $|\mathbf{x}| = |\mathbf{w} - \mathbf{w}_0|$, shall be of significance, especially when considering transaction costs. The usual way of incorporating a term of this kind in a mathematical program is to introduce two variables,

$$\mathbf{x}^+ = \mathbf{w} - \mathbf{w}_0 \qquad \text{and} \qquad \mathbf{x}^- = \mathbf{w}_0 - \mathbf{w}$$

along with the constraints

$$\mathbf{x}^+ \geq \mathbf{0} \qquad \text{and} \qquad \mathbf{x}^- \geq \mathbf{0}$$

and use the sum $\mathbf{x}^+ + \mathbf{x}^-$ whenever $|\mathbf{x}|$ appears.

[7] Assuming stocks as the assets.

thereby prohibiting selling assets short. Regulatory constraints placed on mutual fund managers often mandate such a constraint. Markowitz' original formulation of the portfolio choice problem included those constraints as an integral part of his solution method, and many introductory treatments of the theory[8] include them by default, despite the impossibility of deriving an analytical solution for the optimal portfolio weights in their presence.[9]

2.6.2 Turnover and Transaction Costs Constraints

For large institutional portfolios, transaction costs can represent a sizable portion of total operational costs, especially for funds that take an *active management* (Grinold and Kahn, 2000) approach as opposed to a passive index-tracking objective. As such, we may incorporate constraints that attempt to minimize the relative or dollar turnover on individual assets, respectively

$$|\mathbf{x}_i| \leq U_i \qquad \text{and} \qquad W_0|\mathbf{x}_i| \leq \tilde{U}_i,$$

or the complete portfolio

$$\sum_i |\mathbf{x}_i| \leq U_P.$$

It is also possible to directly incorporate *transaction costs* into the objective function as a term to be minimized. In its simplest form, a transaction cost model simply imposes a proportional cost on the absolute value of traded quantities,

$$\text{prop cost}_i = W_0 \chi_i |\mathbf{x}_i|,$$

and the total portfolio cost given by

$$\text{prop cost}_P = W_0 \sum_i \chi_i |\mathbf{x}_i|, \tag{2.12}$$

where χ_i is the proportional cost of trading asset i. Assuming all χ_i and W_0 are non-negative, prop cost_P is nonnegative and hence the imposition of transaction costs penalizes portfolio performance. To understand their consequence on realized returns, let $\mathbf{p}_1$ be the asset prices at the end of the investment period and consider the relative return on asset i,

$$r_i = \frac{\mathbf{p}_{1i} - \mathbf{p}_{0i}}{\mathbf{p}_{0i}}.$$

Transaction costs affect portfolio return as

[8] E.g. Bodie *et al.* (2004).

[9] Non-negativity constraints can be seen as the "great divide" in optimization between analytical and non-analytical solutions; in the case of portfolio optimization, the latter require, as mentioned above, solution by quadratic programming.

$$\tilde{r}_i = \frac{\mathbf{p}_{1i} - \mathbf{p}_{0i} - W_0 \chi_i |\mathbf{x}_i|}{\mathbf{p}_{0i}}$$
$$= r_i - \frac{W_0}{\mathbf{p}_{0i}} \chi_i |\mathbf{x}_i| \quad = \quad r_i - n_i \chi_i |\mathbf{x}_i| \quad = \quad r_i - \tilde{\chi}_i |\mathbf{x}_i|,$$

where it is obvious that they adjust the portfolio relative return by a term proportional to the traded amount. Their effect can then directly be incorporated into the objective function for the quadratic utility maximization formulation, yielding the problem

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \mathbf{w}'\mu - \tilde{\chi}' |\mathbf{w} - \mathbf{w}_0| - \lambda \mathbf{w}' \Sigma \mathbf{w} \tag{2.13}$$

$$\text{subject to} \quad \mathbf{w}'\iota = 1. \tag{2.14}$$

The proportional costs structure is, however, only a starting point. As pointed out by Kissell and Glantz (2003), the totality of trading costs can be broken down according to an elaborate taxonomy that includes *explicit* (measurable) costs as well as more insidious *implicit* ones. Without delving into an intricate description, we can summarize them as follows:

Explicit Costs They include *fixed costs*, in the form of commissions (as outlined above) and fees (custodial fees, transfer fees). They also include *variable costs*, in the form of bid–ask spread (the difference between the price at which one can buy versus sell) and taxes.[10]

Implicit Costs They include *delay cost* (time between which a decision is made—for instance, by an allocation committee—and the actual trade is brought to the market), *price movement risk* (effect of underlying trends affecting the asset to be traded), *market impact costs* (deviation of the transaction price from the market price that would have prevailed had the trade not occurred), *timing risk* (cost attributable to general market volatility), *opportunity cost* (cost of not trading or not completing a trade).

Some of the implicit costs may not be costs at all but the source of trading profits depending on market conditions. A study by Wagner and Edwards (1998) shows that the price impact of a liquidity-demanding trade[11] averages -103 basis points[12] on a set of some 700,000 trades by more than 50 management firms in 1996, whereas the price impact of a liquidity-supplying trade generated *profits* of $+36$ basis points. In a liquidity-neutral market, the average price impact was -23 basis points. The effects of other implicit costs can likewise be decomposed according to market conditions.

There is a vast literature on transaction costs models, including how realistic nonlinear models of costs can be incorporated in asset-allocation models. This literature is well reviewed by Fabozzi *et al.* (2006, ch. 3).

---

[10] The proportional costs structure introduced previously can be seen as an adequate model of bid–ask spread, the most significant explicit cost for an institutional investor.

[11] For example, a "buy" trade executed when there are significantly more buyers than sellers.

[12] A *basis point* (bp) is one hundredth of one percent, i.e. $100\,\text{bp} = 1\%$.

2.6.3 Maximum Holdings Constraint

To ensure that the portfolio is not overly concentrated in a single asset, we can impose a constraint of the form

$$\mathbf{L} \leq \mathbf{w} \leq \mathbf{U},$$

where $\mathbf{L}$ and $\mathbf{U}$ are vectors specifying, respectively, the allowable lower and upper bounds for each asset. Likewise, we can ensure a sector $\mathscr{S} = \{i_1, i_2, \ldots, i_n\}$ (a set of asset indices) is not unduly weighted in the portfolio by imposing

$$L_{\mathscr{S}} \leq \sum_{i \in \mathscr{S}} \mathbf{w}_i \leq U_{\mathscr{S}},$$

with $L_{\mathscr{S}}$ and $U_{\mathscr{S}}$ denoting, respectively, the minimum and maximum exposure to the sector.

2.6.4 Maximum Tracking Error and Factor Exposure Constraint

The performance of portfolio managers is often compared to that of a *benchmark* such as the S&P 500 (Grinold and Kahn, 2000). Depending on the fund's style, the manager may seek to replicate the benchmark as closely as possible (using, for instance, a smaller number of assets than the benchmark), or to provide additional performance (the so-called "alpha") at the expense of taking on *active risk*, namely, deviating from the benchmark. This risk is quantified by the *tracking error*, defined next. Assume that the benchmark's and fund's investable universe are the same and that the (random) asset returns are given by $\mathbf{R}$. Let $\mathbf{w}_b$ denote the benchmark weights, $\mathbf{w}$ the decision variables, and R_B and R_P denote, respectively, the benchmark and portfolio returns,

$$R_B = \mathbf{w}_B' \mathbf{R} \qquad \text{and} \qquad R_P = \mathbf{w}' \mathbf{R}.$$

The tracking error is simply the variance of the return difference between the benchmark and the invested portfolio,[13]

$$\begin{aligned} \mathrm{TE}_P &= \mathrm{Var}[R_P - R_B] \\ &= \mathrm{Var}[\mathbf{w}_B' \mathbf{R} - \mathbf{w}' \mathbf{R}] \\ &= (\mathbf{w}_B - \mathbf{w})' \Sigma (\mathbf{w}_B - \mathbf{w}), \end{aligned}$$

with Σ the asset return covariance matrix. A quadratic tracking error constraint of the form

[13] More accurately, *tracking error* is usually reserved for the square-root of this variance, but for notational simplicity, we shall omit the square-roots in this overview.

$$(\mathbf{w}_B - \mathbf{w})' \Sigma (\mathbf{w}_B - \mathbf{w}) \leq \sigma_{\text{TE}}^2$$

can then be imposed to limit active risk. Note that this does not limit *total risk*, which would require additional constraints (Jorion, 2003).

In an analogous manner, one can restrict exposure to specific risk factors. Suppose that we posit the following decomposition for *explaining* the return of asset i as a linear combination of factors (additional background on factor models is given in §2.7/p. 23),

$$R_i = \alpha_i + \sum_{j=1}^{M} \beta_{i,j} F_j + \varepsilon_i,$$

where F_j is the random "return" associated with factor j[14] during the period, and $\beta_{i,j}$ is the exposure of asset i to factor j. This is written more succinctly as

$$\mathbf{R} = \alpha + \mathbf{B}\mathbf{F} + \varepsilon$$

with $\mathbf{B}$ and $\mathbf{F}$ respectively the matrix of factor exposures and the vector of one-period factor returns. This yields a portfolio return, given asset weights $\mathbf{w}$, of

$$R_P = \mathbf{w}'\alpha + \mathbf{w}'\mathbf{B}\mathbf{F} + \mathbf{w}'\varepsilon.$$

The exposure of the portfolio to factor j is given by $\sum_i \mathbf{w}_i \beta_{i,j}$. Bound or equality constraints may be placed on this exposure; for example, to ensure an *ex ante* neutral exposure to factor j one may impose

$$\sum_i \mathbf{w}_i \beta_{i,j} = 0.$$

Such constraints are commonly used in so-called "long-short equity" hedge funds, which are designed to be neutral to overall market fluctuations.[15]

2.6.5 Transaction Size, Cardinality and Round Lot Constraints

The following class of constraints is of a combinatorial nature and necessitates solution by *mixed integer programming* methods (Wolsey and Nemhauser, 1999). For convenience, we define the vector δ of binary indicator variables

$$\delta_i = \begin{cases} 1, & \text{if } \mathbf{w}_i \neq 0, \\ 0, & \text{if } \mathbf{w}_i = 0, \end{cases} \qquad i = 1, \ldots, N,$$

[14] For stocks, examples of likely factors would be the return on a broad market index, the return difference between growth and value stocks, and the return difference between large- and small-capitalization stocks; see §2.7/p. 23.

[15] For a factor-neutral constraint to make sense, the exposures $\beta_{i,j}$ must be standardized to have a mean of zero across assets.

where each element specifies whether a position is being taken in the corresponding asset.

A first class of combinatorial constraints aims at eliminating positions that are too small; such positions are often the result of a traditional unconstrained mean–variance optimization. The manager can require

$$|\mathbf{w}_i| \geq \delta_i \mathbf{L}_{\mathbf{w}_i},$$

where $\mathbf{L}_{\mathbf{w}_i}$ is the minimum (relative) position size allowed for asset i. Likewise a limit can be set on portfolio trades

$$|\mathbf{x}_i| \geq \delta_i \mathbf{L}_{\mathbf{x}_i},$$

with $\mathbf{L}_{\mathbf{x}_i}$ the minimum allowed trade size for asset i.

Next, *cardinality* constraints can be useful in problems that seek to replicate a benchmark using a smaller number of assets than the original universe. This may take the form of

$$\delta' \iota \leq K$$

where K is the maximum number of allowable assets. The impact of cardinality constraints on the shape of the efficient frontier is studied by Chang *et al.* (2000).

Finally, *round lot* constraints account for the fact that market-traded instruments are not infinitely divisible (contrarily to idealizations of finance theory)—it is common for stocks to be traded in multiples of 100 shares or more. If the lot size for asset i is given by the constant κ_i and the desired number of lots by η_i (an integer decision variable), we can enforce

$$W_0 \mathbf{w}_i = \kappa_i \eta_i \mathbf{p}_{0i}, \qquad \eta_i \in \mathbb{Z}.$$

In general, when imposing round lots, the budget constraint, $\sum_i \mathbf{w}_i = 1$, may no longer be satisfiable; in this case, one may settle for an approximate budget constraint, expressed as

$$\begin{aligned} \frac{1}{W_0} \sum_i \kappa_i \eta_i \mathbf{p}_{0i} + \xi^+ - \xi^- &= 1, \\ \xi^+, \xi^- &\geq 0, \\ \eta_i &\in \mathbb{Z}, \end{aligned}$$

where ξ^+ and ξ^- are "slack variables" to be minimized (by incorporating them in the objective function). Formulations of this type are analyzed by Kellerer *et al.* (2000).

2.7 Forecasting Models

Markowitz's method of portfolio construction is silent on how the required expected next-period asset returns and covariances are to be obtained. This section reviews the most commonly-used approaches in practice, starting with *factor models* and their uses in covariance modeling and expected return forecasts. We then briefly cover other expected-return forecasting approaches for equities, mostly based on *dividend discount models* and accounting ratios. Finally, extensive experience with mean-variance criteria suggest that they are extremely sensitive to parameter estimation error—very small changes in the forecasts can yield enormous changes in "optimal" portfolio weights, leading to doubt about the validity of the portfolios and possible considerable rebalancing costs when the decisions are implemented. This naturally paves the way for robust estimation methods and Bayesian approaches; we cover some of the methods that have been suggested to counter portfolio instability.

2.7.1 Factor Models

Factor models seek to explain the *cross-section*[16] of asset returns by a simple affine relationship, where the return of asset i over the period t is decomposed into the return of more elemental *factor returns* $F_{j,t}$,

$$R_{i,t} = \alpha_i + \sum_{j=1}^{M} \beta_{i,j} F_{j,t} + \varepsilon_{i,t}, \tag{2.15}$$

where α_i is a regression constant, $\beta_{i,j}$ are the factor exposures, and $\varepsilon_{i,t}$ is a zero-mean random unexplained component uncorrelated with factor returns.[17]

The grandfather of factor models is the Capital Asset Pricing Model (CAPM) of Sharpe (1964), Lintner (1965) and Mossin (1966); this model is generally derived from equilibrium considerations as a *positive theory* of collective investor behavior,[18] but we shall merely regard it as a simple one-factor model. It expresses the expected excess return[19] on asset i as a linear function of the return of the overall market portfolio, R_M,

$$\mathbb{E}[R_i - R_f] = \beta_i \mathbb{E}[R_M - R_f],$$

[16] As opposed to the time-series characteristics.

[17] It should be noted that what this literature refers to as *factors* almost exclusively consist of observable variables, what would simply be called covariates, explanatory or input variables in a more traditional statistical context. Latent factors are always referred to as such.

[18] In other words, it seeks to establish what consequences would arise if every investor behaved according to a set of hypotheses that include Markowitz's rules for portfolio choice among others.

[19] The return earned over the risk-free rate.

where, under the CAPM assumptions, α_i is identically zero.[20]

It has long been understood, at least since Merton (1973), that there exists the possibility that additional sources of *priced risk*, on top of the market portfolio, could impact expected asset returns. Generalizations of the CAPM are obtained in the context of the Arbitrage Pricing Theory (APT) of Ross (1976).[21] Assume that asset returns are distributed according to the factor structure of eq. (2.15), along with

$$\begin{aligned} \mathbb{E}[\varepsilon_i] = \mathbb{E}[F_k] &= 0 \\ \mathbb{E}[\varepsilon_i \varepsilon_j] = \mathbb{E}[\varepsilon_i F_j] = \mathbb{E}[F_i F_j] &= 0, \quad i \neq j \\ \mathbb{E}[\varepsilon_i^2] &= \sigma^2 < \infty. \end{aligned}$$

In this context, in the absence of arbitrage and under some technical conditions, Ross showed that the excess return on asset i is given by

$$\mathbb{E}[R_i - R_f] = \sum_{j=1}^{K} \beta_{i,j} \mathbb{E}[F_j - R_f].$$

Under the APT, each factor represents a priced systematic risk (a risk for which investors are seeking compensation), and the factor exposures $\beta_{i,j}$ quantify the *market price* of those risks (how much the investor is compensated in expected return for taking on a unit of risk).

Ross remains silent on how factors should be chosen. In addition to the CAPM market portfolio factor, several *pricing anomalies* have been documented in the 1980's and early 1990's suggesting additional factors, including long-run price reversal (De Bondt and Thaler, 1985), short-run price momentum (Jegadeesh and Titman, 1993), and a variety of effects due to firm size (market equity, ME, the stock price times the number of shares), earnings to price ratio (E/P), cash-flow to price ratio (C/P), book value to market value (BE/ME), and past sales growth (Banz, 1981; Basu, 1983; Rosenberg *et al.*, 1985; Lakonishok *et al.*, 1994). These results built up to an influential series of papers by Fama and French (1992; 1993; 1995; 1996), who show that the following two additional factors summarize well a number of empirical findings:

High-Minus-Low (HML) The difference between the return on a portfolio of high-book-to-market stocks and the return on a portfolio of low-book-to-market stocks.[22]

[20] Starting from the late-1960's, a huge literature has emerged aiming at testing the validity of the CAPM; see Campbell *et al.* (1997) for an overview.

[21] Technically, the CAPM is derived from equilibrium considerations whereas the APT is derived from a more fundamental "absence of arbitrage" principle; these minutiæ make little difference from a statistical estimation standpoint.

[22] The precise definition is slightly technical and appears in Fama and French (1996).

Small-Minus-Big (SMB) The difference between the return on a portfolio of small-capitalization stocks and the return on a portfolio of large-capitalization stocks.

Put together, Fama and French argue that a model of the form

$$\mathbb{E}[R_i - R_f] = \beta_i \mathbb{E}[R_M - R_f] + s_i \mathbb{E}[\text{SMB}] + h_i \mathbb{E}[\text{HML}]$$

can account for a large fraction of the cross-section of returns, and obtain time-series regression R^2 in the 0.90–0.95 range. The only factor significantly unaccounted for is the short-run price momentum, which is empirically analyzed by Carhart (1997).

Since the late 1990's, several large commercial factor models have become available, the best known of which is perhaps Barra's fundamental multifactor risk model for United States equities (Barra, 1998), which includes 13 risk indices and 55 industry groups.

2.7.2 Factor Models in Covariance Matrix Estimation

The estimation of covariance matrices for portfolios of many assets is a hard problem. As an illustration, consider the Russell 1000 index, whose sample covariance matrix

$$\hat{\Sigma} = \frac{1}{T-1} \sum_{t=1}^{T} (\mathbf{R}_t - \hat{\mu})(\mathbf{R}_t - \hat{\mu})'$$

contains 500,500 distinct entries;[23] an analysis with the tools of *random matrix theory* shows that for such large matrices, only a few eigenvalues of the sample covariance matrix carry information, the rest being the result of noise (Laloux *et al.*, 1999; Malevergne and Sornette, 2005b). This observation gave rise to a number of schemes to add structure to the estimator, often relying on *shrinkage methods* that attempt to find an optimal compromise between a restricted and unrestricted estimators (§2.8/p. 29).

An obvious application of factor models is to the estimation of covariance matrices. This approach can be traced back to a suggestion by Sharpe (1963), and relies on the factor decomposition of eq. (2.15). Assume that firm-specific residual returns, ε_i, are uncorrelated for two different firms,

$$\mathbb{E}[\varepsilon_i \varepsilon_j] = \begin{cases} 0, & i \neq j, \\ \sigma_i^2, & i = j. \end{cases}$$

The covariance between returns R_i and R_j is obtained from eq. (2.15) as

[23] Obtained as $1000 \times 1001/2$.

$$\begin{aligned}\mathrm{Cov}[R_i, R_j] &= \sum_{k=1}^{M} \mathrm{Cov}[\beta_{i,k} F_i, \beta_{j,k} F_j] + \mathrm{Cov}[\varepsilon_i, \varepsilon_j] \\ &= \sum_{k=1}^{M} \beta_{i,k}\beta_{j,k}\,\mathrm{Cov}[F_i, F_j] + \delta_{i,j}\sigma_i^2,\end{aligned}$$

where $\delta_{i,j}$ is the Kronecker delta. This expression illustrates that under a factor model of returns, the covariance between arbitrary assets depends only on the *covariance matrix between the individual factors*, which (for the small number of factors used in practice) is a much more tractable quantity to estimate with statistical reliability. Current methods for covariance modeling are reviewed by Fabozzi *et al.* (2006) and Qian *et al.* (2007).

2.7.3 Factor Models in Expected Return Estimation

Forecasting expected asset returns is recognized as notoriously difficult — so much so that this apparent unforecastability gave rise to the Efficient Market Hypothesis (EMH) and a famous proof that prices should fluctuate randomly (Cootner, 1964; Samuelson, 1965; Fama, 1970). Empirically, it is often observed that the simplest predictors, a constant based on the historical average return or even the constant *zero*,[24] perform the best out of sample. More recently, with advances in computing power and improvements in the quality and quantity of available data, mounting evidence has started to accumulate in favor of some *very small* forecastability (Lo and MacKinlay, 1999), possibly arising from market imperfections. However, exploiting any residual forecastability, especially when accounting for trading costs, remains of the utmost challenge.

Factor models can provide some direction in this respect and are generally used by relating the returns at time t with the observed factors at the same time, and then positing a dynamical model for making forecasts of the factors themselves. It is common to utilize a Vector Autoregressive (VAR) model for establishing the dynamics (Hamilton, 1994), yielding an overall forecasting model specified as

$$\begin{aligned}\mathbf{R}_t &= \alpha + \beta' \mathbf{F}_t + \varepsilon_{\mathbf{R},t} \\ \mathbf{F}_{t+1} &= \mathbf{a} + \mathbf{B}\mathbf{F}_t + \varepsilon_{\mathbf{F},t+1},\end{aligned}$$

where $\mathbf{a}$ is a vector and $\mathbf{B}$ is a matrix of first-order autoregression factors.

An example that has received wide attention is the forecastability of stock returns by the dividend yield.[25] Brandt (2004) estimates the following parameters for the

[24] Which is surprisingly effective in the case of daily stock returns.

[25] The first evidence is presented in Campbell and Shiller (1988) and Fama and French (1988); Campbell (1991) presents an interesting decomposition of stock returns wherein he shows that unexpected stock returns must be associated with changes in expected future dividends or expected future returns, and attributes a third of the variance in U.S. unexpected returns over the 1927–88

quarterly returns of the value-weighted CRSP[26] index

$$\begin{bmatrix} r^e_{t+1} \\ d_{t+1} - p_{t+1} \end{bmatrix} = \begin{bmatrix} 0.2049 \\ (0.0839) \\ -0.1694 \\ (0.0845) \end{bmatrix} + \begin{bmatrix} 0.0568 \\ (0.0249) \\ 0.9514 \\ (0.0251) \end{bmatrix} (d_t - p_t) + \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix} \tag{2.16}$$

$$\begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.0062 & -0.0060 \\ -0.0060 & 0.0063 \end{bmatrix} \right),$$

where r^e_t denotes the log excess return of the index and $d_t - p_t$ is the log dividend yield, computed from the log of the trailing-twelve-month sum of monthly dividends d_t and the current index level p_t.[27] In parenthesis are the Newey and West (1987) standard errors. These results serve to illustrate that whatever forecastability remains, although statistically significant over a long sample, remains low.

2.7.4 Other Expected Return Forecasting Models

A different angle on forecasting models for equities is provided by the *fundamental analysis* of a firm's fair value. The starting point in this line of study is the *dividend discount model* (DDM), introduced by Williams (1938), stating that the price of one share of stock should be given by the sum of discounted future dividend payments,

$$P_t = \mathbb{E}_t \left[\sum_{\tau=1}^{\infty} \frac{D_{t+\tau}}{(1+R_{t+\tau})^{\tau}} \right], \tag{2.17}$$

where D_t is the dividend to be paid in (future) period t and R_t are discount rates.[28] It should be noted that the discount rate is generally higher than the prevailing risk-free rate and reflects the market's expectations on the prospects of future dividend payments; a greater risk on the dividend stream entails a higher discount rate. In other words, it can be viewed as the rate of return that investors *require* for bearing the risk of holding the equity. Consider a simplification wherein we keep the discount factor constant (i.e. not time-varying, but still unknown) with value R and assume a constant growth rate g for dividends,[29] $D_{t+1} = D_t(1+g) = D_1(1+g)^{t-1}$,

---

period to the first component, a third to the second, and the final third to their covariance. For use of the dividend yield in an asset allocation context, see e.g. Kandel and Stambaugh (1996) and Brennan *et al.* (1997).

[26] Center for Research in Security Prices, based at the University of Chicago; www.crsp.com.

[27] The estimation period in this example is from April 1952 to December 1996, and the results are fairly stable across different estimation periods.

[28] This model can be adapted to a similar *free cash flow* relationship for stocks that do not pay dividends.

[29] This hypothesis is valid, for instance, under the scenario where a business grows its earnings at a constant rate and maintains the same dividend payout ratio.

which allows to write

$$P_t = \mathbb{E}_t\left[\sum_{\tau=1}^{\infty} \frac{D_{t+\tau+1}(1+g)^{\tau-1}}{(1+R)^{\tau}}\right] = \mathbb{E}_t\left[\frac{D_{t+1}}{R-g}\right].$$

This is referred to as the Gordon (1962) growth model. Now assuming that price P_t is observed on the market and that R independent of D_{t+1} (the latter is generally a quite well ascertained quantity), the expected implied discount rate—thence the implied expected return on the security—can be solved for as

$$\mathbb{E}_t[R] = \frac{\mathbb{E}_t[D_{t+1}]}{P_t} + g.$$

Unfortunately, this model is very sensitive to inaccuracies in its inputs, and for this reason, so-called *residual income valuation* models (RIM) have been proposed that exploit the fundamental accounting *clean surplus relationship* linking the balance sheet and income statement

$$B_t = B_{t-1} + E_t - D_t, \tag{2.18}$$

where B_t is the firm's book value per share at time t and E_t the earnings per share generated during period t. This states that the period-to-period variation in the firm's value is given by increases resulting from the period activities (net earnings) minus payments to shareholders (dividends) (Edwards and Bell, 1961; Ohlson, 1995). Define the "abnormal" earnings, assuming a constant discount factor R, as

$$E_t^a \triangleq E_t - R B_{t-1};$$

in this context, R can be interpreted as the required return on equity expected at the start of each period. This relationship, in conjunction with eq. (2.18), allows to write the dividends for period t as

$$D_t = E_t^a - B_t + (1+R) B_{t-1}.$$

Substituting in eq. (2.17), we obtain

$$\begin{aligned}
P_t &= \mathbb{E}_t\left[\frac{D_{t+1}}{1+R} + \frac{D_{t+2}}{(1+R)^2} + \cdots\right] \\
&= \mathbb{E}_t\left[\frac{E_{t+1}^a - B_{t+1} + (1+R)B_t}{1+R} + \frac{E_{t+2}^a - B_{t+2} + (1+R)B_{t+1}}{(1+R)^2} + \cdots\right] \\
&= B_t + \mathbb{E}_t\left[\sum_{\tau=1}^{\infty} \frac{E_{t+\tau}^a}{(1+R)^{\tau}}\right] \\
&= B_t + \mathbb{E}_t\left[\sum_{\tau=1}^{\infty} \frac{E_{t+\tau} - R B_{t+\tau-1}}{(1+R)^{\tau}}\right].
\end{aligned}$$

Under some assumptions, Philips (2003) derives the following expression for the expected returns

$$\mathbb{E}_t[R] = \frac{\mathbb{E}_t[E_{t+1}] - gB_t}{P_t} + g,$$

where P_t and B_t are readily available and E_{t+1} is often estimated by analysts that follow a stock.[30] The growth rate g can conservatively be taken as the growth of nominal GDP.[31] Claus and Thomas (2001) find relationships based on residual income valuations to be much less sensitive to errors than the Gordon model.

The topic of expected return forecasts is much richer than this brief overview can provide. In particular, we must omit treatment of a sizable literature on the information regarding the implied probability distribution of returns that option markets provide (e.g. Pan and Poteshman 2006; Aït-Sahalia and Brandt 2007). A review of several recently-proposed methodologies for forecasting expected returns appears in Satchell (2007).

2.8 Forecast Stability and Econometric Issues

A longstanding critique of Markowitz's mean-variance method of portfolio choice stems from the often-observed erratic nature of the optimal weights: unless expected returns are "perfectly matched" to the covariance matrix, it is frequent to arrive at *corner solutions* wherein a small number of assets get allocated most of the weight, with problem constraints strongly governing the obtained solution. It almost appears as if the theory's foundational goal of *efficient diversification of investment*[32] somehow gets lost along the way. Moreover, the obtained solutions tend to be unstable, both cross-sectionally (small changes to the forecasts have a large impact on the weights) and over time (optimal portfolios often change drastically from one period to the next, leading to important costs due to turnover).

Michaud (1989) argues that extreme and unstable portfolio weights are inherent to mean-variance optimizers due to forecast estimation error: by virtue of mere statistical fluctuation, large positive (negative) weights are assigned to assets that have large positive (negative) estimation error in expected return and/or large negative (positive) error in variance. This arises because in the classical mean-variance

[30] Analyst forecasts of earnings have themselves long been subject to investigation, including the early work of Crichfield *et al.* (1978) and Givoly and Lakonishok (1984), who generally find forecasts to improve as the earnings publication date approaches. More recently, Friesen and Weller (2006) consider a Bayesian framework in which analysts constantly revise their forecasts based on newly-revised information; in this context, the authors report strong evidence of biases, including overconfidence and cognitive dissonance biases.

[31] For firms whose capital structure consists of a mixture of equity and debt, this is indeed a very conservative assumption. The growth rate of nominal GDP would normally characterize the return on the firm's *assets*. In contrast, the return on *equity*—the quantity represented by g—would be magnified by the firm's financial leverage, i.e. its use of debt.

[32] The subtitle in Markowitz's 1959 treatment of the subject.

paradigm, forecasts are totally disconnected from optimization: the former are "plugged into" the latter (hence the name *plug-in estimates*), and in a sense the optimizer "does not know" that the forecasts are but point estimates that also have an associated standard error. This led Michaud to his bon mot that *mean-variance optimizers act as statistical error maximizers.*

Michaud (1989), Jobson and Korkie (1980), Best and Grauer (1991) and Chopra and Ziemba (1993) study the impact of estimation uncertainty, where it is often observed to be much larger than that of asset risk itself. In particular, the plug-in estimates are found to be extremely unreliable, their performance dropping rapidly as the number of assets increases. This led to a variety of approaches to "robustify" the optimal portfolios, including shrinkage estimators, Bayesian approaches, resampling methods and robust optimization, summarized next. It should be noted that the practitioner's little-told secret of imposing optimization constraints, such as those reviewed in §2.6/p. 17, already serves to stabilize the portfolio by truncating extreme weights, and was confirmed by Frost and Savarino (1988) to generally improve performance. In this context, constraints can be interpreted as providing a *post hoc* regularization of the estimator, a point elaborated upon by Jagannathan and Ma (2003).

A very complete review of the literature on the econometrics of portfolio choice appears in Brandt (2004).

2.8.1 Shrinkage Estimators

It is known since Stein (1956) that biased estimators often have better finite-sample properties (lower sample variance) than unbiased ones.[33] In particular, consider estimating the mean of an N-dimensional ($N \geq 3$) multivariate normal distribution with *known covariance matrix* Σ, subject to the quadratic loss function

$$L(\hat{\mu}, \mu) = (\hat{\mu} - \mu)' \Sigma^{-1} (\hat{\mu} - \mu),$$

where μ is the true mean. In this context, the usual sample mean $\hat{\mu}$ is not the best estimator (James and Stein, 1961). The James-Stein *shrinkage estimator*

$$\hat{\mu}_{JS} = (1 - w)\hat{\mu} + w\,\mu_0\,\iota, \qquad 0 < w < 1,$$

exhibits a lower quadratic loss, where μ_0 is an arbitrary "common" constant and is called the shrinkage target. The optimal trade-off between bias and variance is achieved by

$$w^* = \min\left(1, \frac{(N-2)/T}{(\hat{\mu} - \mu_0 \iota)' \Sigma^{-1} (\hat{\mu} - \mu_0 \iota)}\right).$$

[33] This *bias–variance trade-off* is related to the notion of capacity control which is studied in depth in machine learning; see, e.g. Bishop (2006) and Hastie *et al.* (2009) for textbook treatments.

More generally, shrinkage methods involve the combination of an unstructured estimator (with a large number of degrees of freedom and likely high sample variance) and a highly structured one (with a small number or even zero degrees of freedom). Jobson and Ratti (1979) and Jorion (1986) have studied them in a portfolio context, demonstrating that their benefits carries to the estimation of expected returns and obtain good performance of the resulting portfolios. Similarly, Frost and Savarino (1986) and Ledoit and Wolf (2004) apply them to the estimation of covariance matrices. Brandt (2004) suggests applying shrinkage estimation directly to the optimal portfolio weights, where the shrinkage target can be some *ex ante* reasonable weights such as $1/N$ or those of a benchmark portfolio.

2.8.2 Bayesian Approaches

In contrast to the "plug-in" approaches presented previously which sought to obtain the single best estimates of the next-period return mean and variance, a Bayesian or decision-theoretic approach would explicitly carry the estimation uncertainty to the optimization. Consider an explicit parametrization of the next-period return distribution, $P(\mathbf{R} \mid \theta)$, in terms of a parameter vector θ, allowing us to rewrite the expected utility maximization, eq. (2.11), as

$$\mathbf{w}^*(\theta) = \arg\max_{\mathbf{w}} \int_{\mathbf{R}} U(\mathbf{w}'\mathbf{R}) \, \mathrm{d}P(\mathbf{R} \mid \theta).$$

A Bayesian investor would not commit to a single choice of parameter vector θ, but would instead consider the posterior distribution of parameters, given by Bayes' rule as

$$P(\theta \mid \mathscr{D}) = \frac{P(\mathscr{D} \mid \theta) P_0(\theta)}{P(\mathscr{D})},$$

where $\mathscr{D}$ is some data (obviously only known up to before the start of the forecast period) and $P_0(\theta)$ is a (subjective) prior distribution on parameter values. The investor's subjective distribution of asset returns, given the data, is obtained by marginalizing out the parameters,

$$P(\mathbf{R} \mid \mathscr{D}) = \int_{\theta} P(\mathbf{R} \mid \theta) \, \mathrm{d}P(\theta \mid \mathscr{D}),$$

yielding to reformulating the expected utility maximization problem for finding optimal portfolio weights as

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \int_{\theta} \left[\int_{\mathbf{R}} U(\mathbf{w}'\mathbf{R}) \, \mathrm{d}P(\mathbf{R} \mid \theta) \right] \mathrm{d}P(\theta \mid \mathscr{D}).$$

This approach to portfolio choice was pioneered as early as the 1960's by Zellner and Chetty (1965) and further studied by Klein and Bawa (1976) and Brown (1978).

More recently, the notion of a "learning investor" was revisited in the context of the increasing evidence on the (mild) predictability of returns in works by Kandel and Stambaugh (1996) and Barberis (2000); see §3.5/p. 55.

2.8.3 The Black-Litterman Model

A different path to Bayesian estimation relies on the implications of an underlying economic equilibrium model, which can serve to provide the "prior" in a portfolio choice context. This is embodied in the Black and Litterman (1992) model, widely used by practitioners. Our presentation of this model draws from Fabozzi *et al.* (2006).

Consider the expected-return relationship for asset i given by the CAPM (§2.7.1/p. 23),

$$\Pi_i = \mathbb{E}[R_i - R_f] = \beta_i \mathbb{E}[R_M - R_f], \tag{2.19}$$

where β_i is obtained as a regression coefficient,

$$\beta_i = \frac{\mathrm{Cov}[R_i, R_M]}{\sigma_M^2},$$

with σ_M^2 the variance of the market portfolio. We shall denote by $\mathbf{w}_M$ the weights of the market portfolio, such that its return can be written as

$$R_M = \sum_{j=1}^{N} \mathbf{w}_{M,j} R_j.$$

Then eq. (2.19) can be rewritten as

$$\begin{aligned}
\Pi_i &= \beta_i \mathbb{E}[R_M - R_f] \\
&= \frac{\mathrm{Cov}[R_i, R_M]}{\sigma_M^2} \mathbb{E}[R_M - R_f] \\
&= \frac{\mathrm{Cov}[R_i, \sum_{j=1}^{N} \mathbf{w}_{M,j} R_j]}{\sigma_M^2} \mathbb{E}[R_M - R_f] \\
&= \frac{\mathbb{E}[R_M - R_f]}{\sigma_M^2} \sum_{j=1}^{N} \mathbf{w}_{M,j} \mathrm{Cov}[R_i, R_j],
\end{aligned}$$

or in matrix form,

$$\Pi = \delta \Sigma \mathbf{w}_M \qquad \text{with } \delta = \frac{\mathbb{E}[R_M - R_f]}{\sigma_M^2}.$$

Although the true expected asset returns μ are unknown, we can posit that the equilibrium model provides a sensible approximation in the form of

$$\Pi = \mu + \varepsilon_\Pi, \qquad \varepsilon_\Pi \sim N(0, \tau\Sigma), \tag{2.20}$$

where $\tau \ll 1$ is a small constant.[34] We can view ε_Π as a "confidence interval" in which the true expected returns are approximated by the equilibrium model: a small τ implies a high confidence in the equilibrium estimates and vice versa.

Now suppose that the investor holds particular *views* on some assets or combinations of assets; examples are "the expected return of asset i will be x percent", or "asset j will outperform asset k by z percent". Each view has an attached *confidence* reflecting how strongly the investor believes them. We can formally express the K views as a vector $\mathbf{q} \in \mathbb{R}^K$,

$$\mathbf{q} = \mathbf{P}\mu + \varepsilon_\mathbf{q}, \qquad \varepsilon_\mathbf{q} \sim N(0, \Omega), \tag{2.21}$$

where $\mathbf{P}$ is a $K \times N$ matrix of view combinations and Ω is a $K \times K$ matrix of view confidences. For example, in a universe of $N = 3$ assets, the investor may believe that

- Asset 1 will have a return of 1.5%.
- Asset 3 will outperform asset 2 by 4%.

This yields the following form for the views

$$\begin{bmatrix} 1.5\% \\ 4\% \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{\mathbf{q},1} \\ \varepsilon_{\mathbf{q},2} \end{bmatrix},$$

for some view confidence matrix Ω, which is commonly diagonal. Both eq. (2.20) and (2.21) are expressed in terms of the unknown expected returns μ. The Black-Litterman model uses the *mixed estimator* of Theil and Goldberger (1961) to combine the information from two data sources—here the equilibrium model and the investor views—into a single posterior estimator. Start by "stacking" the two equations as follows,

$$\mathbf{y} = \mathbf{X}\mu + \varepsilon, \qquad \varepsilon \sim N(0, \mathbf{V})$$

where

$$\mathbf{y} = \begin{bmatrix} \Pi \\ \mathbf{q} \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} \mathbf{I}_N \\ \mathbf{P} \end{bmatrix}, \qquad \mathbf{V} = \begin{bmatrix} \tau\Sigma & \\ & \Omega \end{bmatrix}.$$

We can rely on a standard generalized least squares (GLS) estimator (Greene, 2007) to arrive at the *Black-Litterman* estimator for expected returns,

[34] Values in the neighborhood of 0.1–0.3 often give satisfactory results for U.S. equities.

$$
\begin{aligned}
\hat{\mu}_{\mathrm{BL}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
&= \left(\left[\mathbf{I}_N \mathbf{P}'\right] \begin{bmatrix} (\tau\Sigma)^{-1} & \\ & \Omega^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_N \\ \mathbf{P} \end{bmatrix} \right)^{-1} \left[\mathbf{I}_N \mathbf{P}'\right] \begin{bmatrix} (\tau\Sigma)^{-1} & \\ & \Omega^{-1} \end{bmatrix} \begin{bmatrix} \Pi \\ \mathbf{q} \end{bmatrix} \\
&= \left(\left[\mathbf{I}_N \mathbf{P}'\right] \begin{bmatrix} (\tau\Sigma)^{-1} \\ \Omega^{-1}\mathbf{P} \end{bmatrix} \right)^{-1} \left[\mathbf{I}_N \mathbf{P}'\right] \begin{bmatrix} (\tau\Sigma)^{-1}\Pi \\ \Omega^{-1}\mathbf{q} \end{bmatrix} \\
&= \left[(\tau\Sigma)^{-1} + \mathbf{P}'\Omega^{-1}\mathbf{P}\right]^{-1} \left[(\tau\Sigma)^{-1}\Pi + \mathbf{P}'\Omega^{-1}\mathbf{q}\right].
\end{aligned}
$$

This estimator is then used with the standard mean-variance problem formulation, e.g. eq. (2.2) or eq. (2.7). Practical experience with this model, documenting the much greater stability of the resulting portfolio weights than would otherwise be obtained, is related in Bevan and Winkelmann (1998), Litterman (2003), and Fabozzi *et al.* (2006).

2.8.4 Portfolio Resampling

The Black-Litterman estimator still operates before portfolio optimization takes place; its benefits can be traced to a reduced "impedance mismatch" between the expected return estimator and the associated covariance matrix. In contrast, portfolio resampling techniques (Michaud, 1998; Scherer, 2002) attempt to make direct use of the forecast distribution of returns by repeatedly drawing a large number of (*expected-return* , *covariance-matrix*) pairs, and for each computing an efficient frontier, namely a set of (*portfolio-return* , *portfolio-risk*) pairs, over some reasonable risk range. Then those efficient frontiers are averaged over all drawings, and the resulting frontier used to make an allocation decision. Markowitz and Usmen (2003) compare this approach to one similar to the Bayesian approach of p. 31 and observe a good performance of the resampling approach.

A practical limitation to the approach is with respect to *portfolio constraints*: in general, there is no guarantee that the averaged portfolio weights (after resampling) will obey the inequality constraints set in the original optimization problem. Also, due to the high number of optimization steps it requires, it is computationally expensive.

2.8.5 Robust Portfolio Allocation

In recent years, several reformulations of the mean-variance problem have received wide attention that attempt to incorporate estimation uncertainty within the optimization step—not "before", as for the Black-Litterman model, or "around" as for portfolio resampling. They are collectively known as *robust optimization* tech-

niques, and are related to minimax estimators in decision theory.[35] Robust methods in mathematical programming were introduced by Ben-Tal and Nemirovski (1999) and further studied in a portfolio choice context by Goldfarb and Iyengar (2003) and Tütüncü and Koenig (2004) among others. Fabozzi *et al.* (2007) provides a good survey of the current literature.

The starting point of these approaches is to consider the *uncertainty set* of the model parameters (the next-period expected returns and their covariances for a portfolio problem) and to ask: "what is the worst-case realization of model parameters that can arise?", and from there to maximize the utility of this worst-case outcome. Consider the simplest type of uncertainty region given in the form of "box" intervals

$$\mathscr{U} = \{(\mu, \Sigma) : \mu_L < \mu < \mu_U, \Sigma_L < \Sigma < \Sigma_U, \Sigma \succ 0\},$$

where in this context the $<$ operator should be interpreted elementwise for both vectors and matrices.

The robust portfolio problem with quadratic utility is expressed as

$$\max_{\mathbf{w}} \left\{ \min_{(\mu,\Sigma)\in\mathscr{U}} \mu'\mathbf{w} - \lambda \mathbf{w}'\Sigma\,\mathbf{w} \right\}$$

which for the above form of the uncertainty region separates out as

$$\max_{\mathbf{w}} \left\{ \min_{\mu\in\mathscr{U}^{\mu}} \mu'\mathbf{w} + \max_{\Sigma\in\mathscr{U}^{\Sigma}} \lambda \mathbf{w}'\Sigma\,\mathbf{w} \right\}.$$

This can be expressed as a saddle-point problem and solved in polynomial time (Halldórsson and Tütüncü, 2003). Simpler results can be obtained by considering other types of uncertainty sets; for instance, when only uncertainty in expected returns is considered, the box constraints reduce to a quadratic program of nearly the same complexity as the original mean-variance problem; similarly, an ellipsoidal constraint set yields a second-order cone program (SOCP), which is efficiently solved by interior-point methods (Boyd and Vandenberghe, 2004). More recently, Bertsimas and Pachamanova (2008) studied a number of robust optimization approaches to the multiperiod portfolio problem (see next section) in the presence of transaction costs; in particular, they advocate linear formulations that yield significant computational savings.

2.8.6 Portfolio Robustness: a Synthesis?

In light of the large variety of proposed methods for improving the performance of mean-variance allocation, one may wonder if a particular method turns out to be "best". To the author's knowledge, a systematic comparison between all of the

[35] Robust optimization should not be confused with *robust estimation* in statistics, devoted to establishing the properties of outlier-resistant estimators.

approaches presented in this section has yet to be published. However, an element of insight has recently been provided by DeMiguel *et al.* (2009), who compare 14 different models on a number of datasets (including U.S. and world equity markets) on three criteria: the out-of-sample Sharpe ratio, certainty equivalent return (from the perspective of a mean-variance investor) and portfolio turnover. On these measures, it is found that *none of the "sophisticated" models consistently beat the naïve* $1/N$ *benchmark* (uniform portfolio weights), *out of sample*. These results suggest that, for the models considered, estimation error still largely dominates any gains obtained from "optimal" diversification.

http://www.springer.com/978-1-4614-0576-4

Portfolio Choice Problems
An Introductory Survey of Single and Multiperiod Models
Chapados, N.
2011, X, 96 p. 8 illus., Softcover
ISBN: 978-1-4614-0576-4