Optimal Asset Allocation Strategies

Răzvan Ştefănescu*

June 26, 2016

Abstract

This study presents a survey on the current methodologies employed in the asset allocation industry. We review techniques from Modern Portfolio Theory and focus on the central role of the mean-variance optimization problem. The formalism of the Black-Litterman model is discussed with its Bayesian rigorous approach to combine the equilibrium returns predicted by the Capital Asset Pricing Model and investors' views. We also turned our attention to factor models to estimate asset returns and corresponding financial risks as an alternative to the traditional approaches. Linear, non-linear and machine learning regression techniques are made available for the construction of the factor models.

1 Introduction

Portfolio optimization refers to the procedure of selecting the proportions of different assets to be retained in a portfolio thus increasing its performance according to some criteria or metrics. With the work of Markowitz [29], the modern theory for portfolio optimization was born. The general approach is centered on the mean-variance optimization methodology for allocation of wealth to different investment alternatives having various risks. Its purpose is to derive portfolio weights that maximize the expected return for an a-priori selected level of risk (variance) or minimize the volatility for a specified expected return level.

The shift in paradigm proposed by Markowitz [29] was that the risk/return of single assets should not be considered individually, but rather in a portfolio setting, with efficient portfolios based on a minimal risk for a given return level or a maximal return for a given risk level. From mathematical point of view, these are optimization problems, with the former being a quadratic optimization problem with linear constraints, whereas the latter has a linear objective function with quadratic constraints. These optimal portfolios are searched in the so called opportunity set or feasible set and correspond to a specific curve known as the efficient frontier. Each point on this curve corresponds to a portfolio that has the smallest degree of risk for its level of expected return.

*email: rstefan@ncsu.edu

We will describe four important benchmark portfolio optimization problems inspired from the recent overview study on modern portfolio theory [15]. These are the mean-variance optimization, the minimum variance for a given expected return level, equal contributions to risk and maximum diversification. We discuss several optimization strategies to solve the mean-variance optimization problem with only equality constraints or both equality and inequality constraints. In the case only equality constraints are considered, we propose solutions based on the optimality conditions associated with the Lagrangian multiplier method and Augmented Lagrangian technique. For both equality and inequality constraints, Lagrangian multiplier method and quadratic penalty method are discussed. For the other three benchmark problems, we formulate the associated objective functions. Despite its elegance and rigourous mathematical foundation, Markowitz's approach suffers from a series of drawbacks, mostly due to the required estimation of the expected returns and their covariances.

The Black-Litterman model [3] seeks to overcome problems encountered by investors in applying modern portfolio theory in practice. The framework combines the equilibrium assumption that the asset allocation should be proportional to the market values with the particular views of the investor under a Bayesian analysis. More precisely, the model computes the implied risk premia and then deduces the optimized portfolio which agrees with the views of the portfolio manager. The Black-Litterman framework produces a single maximum Sharpe ratio (MSR) [43] optimal portfolio on the unconstrained mean-variance efficient frontier starting from an initial MSR optimal portfolio and active views. We present the entire formalism behind the Black-Litterman model which synthesize three different methods; i.e., portfolio optimization [29], the Capital Asset Pricing Model theory and introduction of constraints in asset allocation.

An alternative to estimate the expected assets' returns and their covariances has recently emerged and rely on factor models. This type of models exploits the observation that most return and risk characteristics for all asset classes can be well explained by particular building blocks, or factors. We describe various types of factors previously introduced in the literature specific to certain types of assets. Later we introduce the general form of factor models and describe how the uncertainties are propagated through these models. Factor models require much fewer observations for a reasonably accurate and stable estimate of covariance matrix and are less prone to false relationships than directly constructing the returns' covariances from historical data.

The mathematical notation utilized in this study uses bold letters to denote vectors. The remainder of the report is organized as follows. In Section 2, we introduce the readers to the Modern Portfolio Theory and present four important benchmark portfolio optimization problems. For the mean-variance optimization problem with equality and inequality constraints, we formulate various solutions based on the optimization theory. Section 3 presents the Black-Litterman formalism. In Section 4, we describe the recently introduced factor models and the associated strategy to construct covariances of the assets' returns. We draw conclusions in Section 5. Several linear, non-linear and machine learning regression techniques that can be potentially employed to construct factor models are summarized in the Appendix.

2 Modern Portfolio Theory

The tradeoff between risk and return represents a concern for every investor. In an attempt to increase the expected return on investment, an investor must be willing to accept greater risks. Modern portfolio theory studies how to model this tradeoff given a collection of N possible assets with returns $\boldsymbol{r} = \{r_i\}_{i=1,2,\cdots,N}$. The returns r_i are usually unknown and they are modeled as random variables following a normal distribution. Every normal distribution can be characterized by mean and variance

$$\mu_i = E[r_i]; \quad \sigma_i^2 = E[(r_i - \mu_i)^2].$$

The variance quantifies the spread of the variable r_i about its mean, so that larger values of σ_i indicate riskier assets. We can define correlations between pairs of returns since the returns are not in general independent

$$\rho_{i,j} = \frac{E[(r_i - \mu_i)(r_j - \mu_j)]}{\sigma_i \sigma_j}, \ i, j = 1, 2, \cdots, N.$$

The correlation measures the tendency of the return on assets i and j to shift in the same direction. An investor designs a portfolio by putting a fraction w_i of the available funds into asset i, for $i = 1, 2, \cdots, N$. Depending on the type of assets considered by the investor, different constraints have to be imposed on the weights w_i.

The return of the portfolio is given by

$$R = \sum_{i=1}^{N} w_i r_i.$$

To measure the desirability of the portfolio, we need to obtain measures of its expected return and variance. The expected return is simply

$$E(R) = E\left[\sum_{i=1}^{N} w_i r_i\right] = \sum_{i=1}^{N} w_i E[r_i] = \boldsymbol{w}^T \boldsymbol{\mu},$$

while the variance is given by

$$Var(R) = E[(R - E(R))^2] = \sum_{i=1}^{N}\sum_{j=1}^{N} w_i w_j \sigma_i \sigma_j \rho_{i,j} = \boldsymbol{w}^T \Sigma \boldsymbol{w},$$

where the $N \times N$ symmetric positive semidefinite covariance matrix Σ is defined by

$$\Sigma_{i,j} = \rho_{i,j} \sigma_i \sigma_j.$$

The opportunity set, also know as the feasible set, consists of all portfolios that could be formed from a series of N assets. This opportunity set can be plot in $[Var(R), E(R)]$ space. An efficient portfolio has either a higher return than any other portfolios in its risk class or

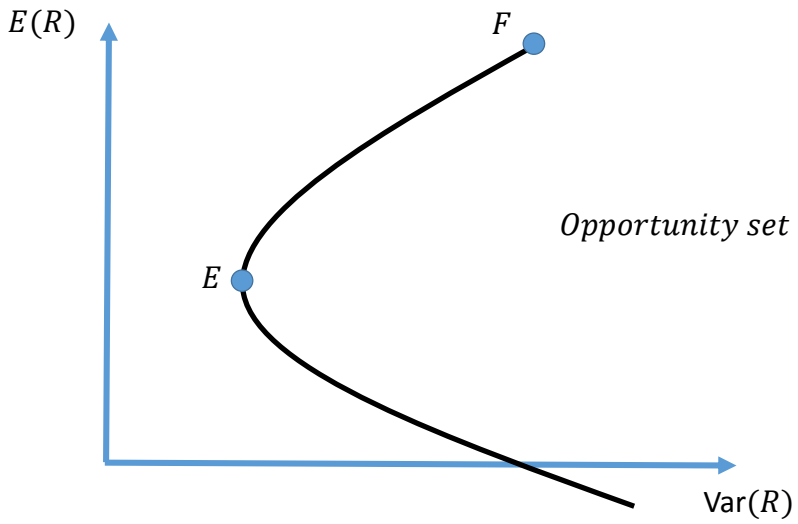

Figure 1: Efficient frontier and opportunity set

less risk than any other portfolio with the same return. The portfolios lying on the efficient frontier, described by the curve between points E and F in Figure 1, are all efficient.

The main difficulty in applying portfolio optimization techniques to real-life investments consists in the lack of information regarding the expected returns, variances, and correlations for the investments in question. The financial professionals often mix historical data with their own insights to produce estimates of these quantities.

Assuming the investor has a complete knowledge of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ of the assets' returns we formulate four different major portfolio optimization problems. A detailed characterization of the first two problems is available in Meucci [32, Chapters 6.3-6.5], whereas Choueifaty and Coignard [8] and Maillard et al. [28] analyze the latter frameworks. An overview of current state of-the-art for modern portfolio theory is given in Francis and Kim [15].

2.1 Mean-variance optimization

Ideally, we would like to find a portfolio for which the expected return $\boldsymbol{w}^T\boldsymbol{\mu}$ is large while the variance $\boldsymbol{w}^T\Sigma\boldsymbol{w}$ is small. In the model proposed by Markowitz [30], these two aims are combined into a single objective function with the aid of a risk tolerance parameter denoted by δ. The value selected for the positive parameter δ depends on the preferences of the individual investor. More conservative investors are tempted to minimize the risk in their portfolio, and would choose a large value of δ to increase the weight of the variance measure in the cost function. More bold investors are prepared to take more risks to gain a higher expected return, so they would choose a smaller value of δ. In this subsection, we search the solution of the following problem

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{\mu} - \delta\boldsymbol{w}^T\Sigma\boldsymbol{w}, \text{ subject to } A\boldsymbol{w} = \boldsymbol{a}. \tag{2.1}$$

The solution of the mean-variance optimization problem is graphically represented by point E in Figure 1. The problem (2.1) can be written as a minimization problem; i.e,

$$\min_{\boldsymbol{w}} J(\boldsymbol{w}) = \delta\boldsymbol{w}^T\Sigma\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{\mu}, \text{ subject to } A\boldsymbol{w} = \boldsymbol{a}. \tag{2.2}$$

which is a quadratic programming problem with equality constraints.

To solve this problem, one can apply, for example, the Lagrangian multiplier method or the Augmented Lagrangian technique.

- Lagrangian multiplier method: The Lagrangian function is constructed

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = \delta\boldsymbol{w}^T\Sigma\boldsymbol{w} - \boldsymbol{w}^T\boldsymbol{\mu} - \boldsymbol{\lambda}^T(A\boldsymbol{w} - \boldsymbol{a}). \tag{2.3}$$

  The first-order necessary conditions for $\boldsymbol{w}^*$ to be a solution for (2.2) state that there is a vector $\boldsymbol{\lambda}^*$, such that the following system of equations is satisfied:

$$\begin{bmatrix} 2\delta\Sigma & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^* \\ \boldsymbol{\lambda}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{a} \end{bmatrix}.$$

  Direct solutions are based on factoring the full Karush-Kuhn-Tucker (KKT) system, Schur-Complement or null-space methods. Iterative solutions of the KKT system can be obtained by using a conjugate gradient method applied to the reduced system or employing a projected conjugate gradient method.

- The augmented Lagrangian method reduces the possibility of ill conditioning by introducing explicit Lagrange multipliers estimates into the function to be minimized. The augmented Lagrangian function is described below

$$L_A(\boldsymbol{w}, \boldsymbol{\lambda}; \alpha) = \delta\boldsymbol{w}^T\Sigma\boldsymbol{w} - \boldsymbol{w}^T\mu - \boldsymbol{\lambda}^T(A\boldsymbol{w} - \boldsymbol{a}) + \frac{\alpha}{2}(A\boldsymbol{w} - \boldsymbol{a})^T(A\boldsymbol{w} - \boldsymbol{a}). \tag{2.4}$$

  The augmented Lagrangian function (2.4) differs from the standard Lagrangian by the presence of the squared terms. An iterative algorithm [34, Alg.17.3, p.515] is usually employed to search for the minima of L_A for increasing values of α. At the $k-$th iteration, the optimality condition is given by

$$0 \approx \nabla_{\boldsymbol{w}} L_A(\boldsymbol{w}^k, \boldsymbol{\lambda}^k; \alpha^k) = \delta\Sigma\boldsymbol{w}^k - \mu - A^T\boldsymbol{\lambda}^k + \frac{\alpha_k}{2}A^T(A\boldsymbol{w}^k - \boldsymbol{a}) \tag{2.5}$$

  and the Lagrange multiplier vector $\boldsymbol{\lambda}^k$ is updated using formula

$$\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k - \alpha^k(A\boldsymbol{w}^k - \boldsymbol{a}). \tag{2.6}$$

Searching for an optimal portfolio by minimizing the cost function in (2.2) may lead to negative weights $\boldsymbol{w}$. Such solutions may not be realistic. Some large public portfolios and mutual funds are legally forbidden to use leverage. When negative weights are not allowed, additional inequality constraints must be added to the optimization problem (2.2) leading to

$$\min_{\boldsymbol{w}} J(\boldsymbol{w}) = \delta \boldsymbol{w}^T \Sigma \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\mu}, \text{ subject to } A\boldsymbol{w} = \boldsymbol{a}, \ \boldsymbol{w} \geq 0. \tag{2.7}$$

Such problem can be solved using inequality-constrained quadratic programming or quadratic penalty method.

- Quadratic Programming: First we start by defining the Lagrangian for this problem

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = \delta \boldsymbol{w}^T \Sigma \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\mu} - \boldsymbol{\lambda}^T (A\boldsymbol{w} - \boldsymbol{a}) - \boldsymbol{\beta}^T \boldsymbol{w}. \tag{2.8}$$

  Let us define the active set $\mathcal{A}(\boldsymbol{w}^*) = \{i |\ w_i^* = 0\}$ of indices of the constraints for which equality holds at optimum. The first order optimality conditions associated with this Lagrangian are given below

$$\begin{aligned} 2\delta \Sigma \boldsymbol{w}^* \boldsymbol{\mu} - A\boldsymbol{\lambda}^* - \bar{\boldsymbol{\beta}}^* &= 0 \\ A\boldsymbol{w}^* &= \boldsymbol{a}, \\ w_i^* &= 0, \ i \in \mathcal{A}(\boldsymbol{w}^*), \\ w_i &\geq 0, \ i \notin \mathcal{A}(\boldsymbol{w}^*), \\ \boldsymbol{\lambda}^* &\geq 0, \\ \bar{\beta}_i^* &\geq 0, \ i \in \mathcal{A}(\boldsymbol{w}^*), \end{aligned} \tag{2.9}$$

  where $\bar{\beta}_i^* = 0$, for $i \notin \mathcal{A}(\boldsymbol{w}^*)$.

  Solutions of the KKT system (2.9) can be obtained by applying active-set methods, interior point methods and gradient projection method.

- Quadratic penalty method: For the general constrained optimization problem (2.7), with equality and inequality constraints, we can define the quadratic penalty function

$$Q(\boldsymbol{w}; \alpha) = \delta \boldsymbol{w}^T \Sigma \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{\mu} + \frac{\alpha}{2} (A\boldsymbol{w} - \boldsymbol{a})^T (A\boldsymbol{w} - \boldsymbol{a}) + \frac{\alpha}{2} [\boldsymbol{w}]^{-T} [\boldsymbol{w}]^-, \tag{2.10}$$

  where $[\boldsymbol{w}]^- = \max(-\boldsymbol{w}, 0)$. A general framework for algorithms based on the quadratic penalty function is described in [34, Alg.17.1, p.501].

2.2 Minimum variance for a given expected return level

The investor is assumed to minimize the following utility function

$$J(\boldsymbol{w}) = \boldsymbol{w}^T \Sigma \boldsymbol{w}, \text{ subject to } \boldsymbol{w}^T \boldsymbol{\mu} = \mu_0, \ A\boldsymbol{w} = \boldsymbol{a}$$

In this problem we are looking for the portfolio with minimal variance satisfying linear constraints and having expected return equal to μ_0. The solution of the problem will identify

a portfolio on the efficient frontier. The optimal portfolio can be calculated using methods such as Lagrangian multiplier and augmented Lagrangian described in previous subsection dedicated to the mean-variance optimization method. Additional inequality constraints can be also incorporated.

2.3 Equal contributions to risk

The investor is assumed to maximize

$$J(\boldsymbol{w}) = -\sum_{i,j=1;i<j}^{N} (RC_i(\boldsymbol{w}) - RC_j(\boldsymbol{w}))^2,$$

where $RC_i(\boldsymbol{w}) = w_i(\Sigma \boldsymbol{w})_i$ represents the contribution of asset i to the portfolio variance. This expresses that the investor looks for a portfolio with similar risk contributions with respect to the Euclidean distance. Typically, constraints as in subsection 2.2 can be utilized.

2.4 Maximum diversification

The investor is assumed to maximize

$$J(\boldsymbol{w}) = \frac{\sum_{i=1}^{N} w_i \sqrt{\Sigma_{i,i}}}{\sqrt{\boldsymbol{w}^T \Sigma \boldsymbol{w}}},$$

where $\Sigma_{i,i}$ is the i-th element on the diagonal of Σ. This expresses that the investor maximizes the benefits of diversification. Typically, constraints as in subsection 2.2 can be utilized.

3 Black and Litterman model

The Black-Litterman model (BL) is an asset allocation model proposed in [3]. The model couples the market equilibrium with the investors views to generate a set of expected returns of the portfolio's assets subsequently used to define a mean-variance optimization problem whose solutions are the optimal portfolio weights. In the first step, BL model employs a Bayesian approach to construct a distribution for the expected assets' returns. Usually the Capital Asset Pricing Model (CAPM) [26] is utilized to design the normal equilibrium prior, whereas the investor's opinion is taken into account and forms a normal likelihood function. In consequence a posterior distribution for the expected assets' returns is obtained together with an updated covariance matrix. In the second step, the optimal portfolio's weights are computed as the solution of the unconstrained mean-variance optimization, with the mean and updated covariance matrix associated with the posterior distribution of the expected assets' returns and a fixed risk aversion parameter. A graphical representation of the model taken from [38] is depicted in Figure 2.

We follow the algorithm proposed in [22]. Again, we assume that there are N-assets in the market, which may include equities, bonds, currencies, and other assets. The returns

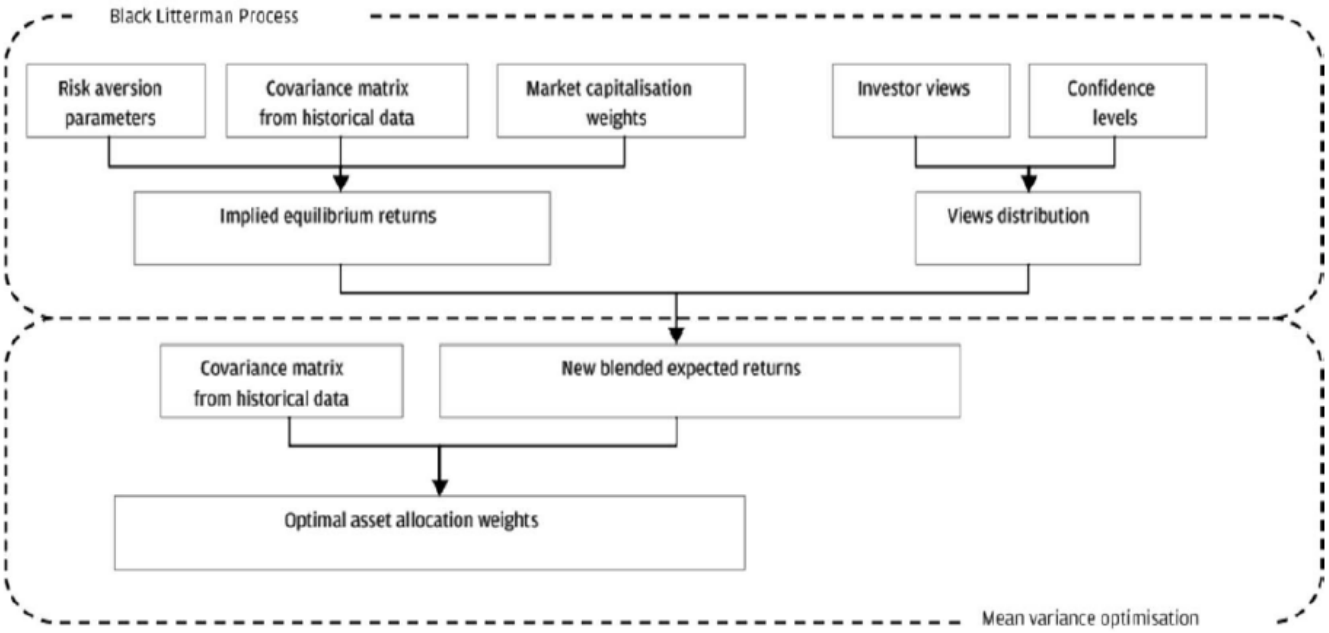

Figure 2: The Black-Litterman framework

of these assets are assumed to have a normal distribution with μ being the expected return and Σ the covariance matrix

$$\boldsymbol{r} \sim N(\boldsymbol{\mu}, \Sigma). \tag{3.1}$$

By denoting the equilibrium market portfolio's weighs $\boldsymbol{w}_{eq}$ and assuming the average risk tolerance of the world is represented by the risk aversion parameter δ, we can compute the equilibrium risk premiums, $\boldsymbol{\Pi}$ using reverse optimization

$$\boldsymbol{\Pi} = \delta \Sigma \boldsymbol{w}_{eq}. \tag{3.2}$$

The expected returns $\boldsymbol{\mu}$ are assumed to be random variables themselves, normally distributed with the equilibrium risk premiums as mean derived from the CAPM model; i.e.,

$$\boldsymbol{\mu} = \boldsymbol{\Pi} + \boldsymbol{\varepsilon}^{(e)}. \tag{3.3}$$

Here $\boldsymbol{\varepsilon}^{(e)}$ is a normally distributed random vector with zero mean and covariance matrix $\tau\Sigma$, where τ is a scalar indicating the uncertainty in the CAPM prior.

In addition to the CAPM prior, the investor also has a number of opinions regarding the market returns. An opinion is expressed as a statement regarding the expected return of the portfolio's assets represented as a normal distribution. Let K be the total number of the views, P be a $K \times N$ matrix whose rows are these portfolio's weights and Q be a K-vector of the expected returns of these portfolio' assets. That is

$$P^T = (\boldsymbol{p}_1,\ \boldsymbol{p}_2,\ \cdots,\ \boldsymbol{p}_k) \tag{3.4}$$

$$\boldsymbol{Q}^T = (q_1,\ q_2,\ \cdots,\ q_k) \tag{3.5}$$

Then the investor's views are expressed as

$$\boldsymbol{P}_\mu = \boldsymbol{Q} + \boldsymbol{\varepsilon}^{(v)} \tag{3.6}$$

where $\boldsymbol{\varepsilon}^{(v)}$ is an unobservable normally distributed random vector with zero mean and a diagonal covariance matrix $\Omega \in \mathbb{R}^{N\times N}$. It is also assumed that the $\boldsymbol{\varepsilon}^{(e)}$ and $\boldsymbol{\varepsilon}^{(v)}$ are independent

$$\begin{pmatrix} \boldsymbol{\varepsilon}^{(e)} \\ \boldsymbol{\varepsilon}^{(v)} \end{pmatrix} \sim N\left(0, \begin{bmatrix} \tau\Sigma & 0 \\ 0 & \Omega \end{bmatrix}\right).$$

Now, these views are combined with the CAPM prior using a Bayesian formulation. The result is that the expected returns are distributed as $N(\bar{\boldsymbol{\mu}}, \bar{M}^{-1})$, where the mean $\bar{\boldsymbol{\mu}}$ is given by

$$\bar{\boldsymbol{\mu}} = [(\tau\Sigma)^{-1} + P^T\Omega^{-1}P]^{-1}[(\tau\Sigma)^{-1}\boldsymbol{\Pi} + P^T\Omega^{-1}\boldsymbol{Q}], \tag{3.7}$$

and the covariance matrix $\bar{M}^{-1}$ is given by

$$\bar{M}^{-1} = [(\tau\Sigma)^{-1} + P^T\Omega^{-1}P]^{-1}. \tag{3.8}$$

Next, we compute the covariance matrix of the returns about the estimated mean $\bar{\Sigma} = \Sigma + \bar{M}^{-1}$ and we obtain the following distribution for the returns

$$\boldsymbol{r} \sim N(\bar{\boldsymbol{\mu}}, \bar{\Sigma}). \tag{3.9}$$

For an investor with the risk aversion parameter δ, the optimal portfolio can be constructed using the unconstrained mean-variance optimization method

$$\max_{\boldsymbol{w}} J(\boldsymbol{w}),\ \ J(\boldsymbol{w}) = \boldsymbol{w}^T * \bar{\boldsymbol{\mu}} - \frac{\delta}{2}\boldsymbol{w}^T\bar{\Sigma}\boldsymbol{w}. \tag{3.10}$$

The first order condition yields

$$\boldsymbol{w}^* = \frac{1}{\delta}\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}}. \tag{3.11}$$

Among the advantages of the BL model, Cheung [7] underlined the inclusion of the uncertainty through the use of Bayes's rule, the smoothness of the investor's view processing, and robust allocation due to shrinkage. Some practical aspects were identified by Cheung [6], such as the curse of dimensionality, confidence parameter setting, investor's view correlations, prior setting, factor-based portfolio construction, selection of optimization algorithm, risk model quality, and linearity and normality assumptions.

4 Factor models

Due to the recent market turbulence and stress, alternatives to traditional asset allocation methodologies emerged. One such alternative can be achieved using factor-based investing, which relies on the real data analysis that most assets' returns and associated risks can be well explained by certain building blocks or factors. According to Podkaminer [36], a factor-based investment strategy allows the investor to remix the factors into portfolios that are better diversified and more efficient than traditional portfolios. Moreover, factor based portfolios should be more advantageous for investors since their components can be directly linked to independent risks and individually rewarded by the market for their level of risk.

Factors can be defined as the basic building blocks of asset classes representing a source of common risk exposures across asset classes. Factors are the smallest systematic units that impact investment return and risk characteristics. Podkaminer [36] enumerated a series of factors such as inflation, GDP growth, currency, and convexity.

We mention here various types of factors previously employed in the literature:

- factors that characterize forecast economic views: interest rates, slope of the yield curve, corporate bond spreads, equity returns, changes in volatility, commodity returns, changes in liquidity.
- fundamental factors: momentum, value and size capturing features such as industry mem- bership, country membership, valuation ratios, and technical indicators. These type of factors were employed in Fama-French [13, 14] or Carhart models [4].
- macroeconomic factors [36]: low/high growth, low/high inflation, flight to quality, surprises in inflation, surprises in GNP.
- statistical factors based on principal components analysis (PCA) or singular value decomposition (SVD)

One possible factor classification based on the type of exposure among various categories is shown in Figure 3. Previous studies showed that each type of asset class is sensitive to specific factors. For example, U.S. equity and U.S. corporate bonds share some common risk exposures according to [36], as seen in Figure 4.

Several factor premiums are documented in the commodity market and we mention the momentum [10], carry [10, 20] low-volatility [16] and value [1]. Currency returns depend on various risk factors [31] such as carry factor, value, momentum [35] or trend. Documented factors for bond markets comprise low-risk [16, 24, 25], momentum [24], value [1, 9], size [24] and carry [37]. For global equities, several factors including carry [37], momentum and value [1] were discovered. Strategy/product, fund size, vintage, industry focus and geographic exposure were identified as factors applicable to private equities [27]. Various factors have been recognized for real estate and we enumerate real estate exposure, vintage year exposure, property sector exposure and leverage [11].

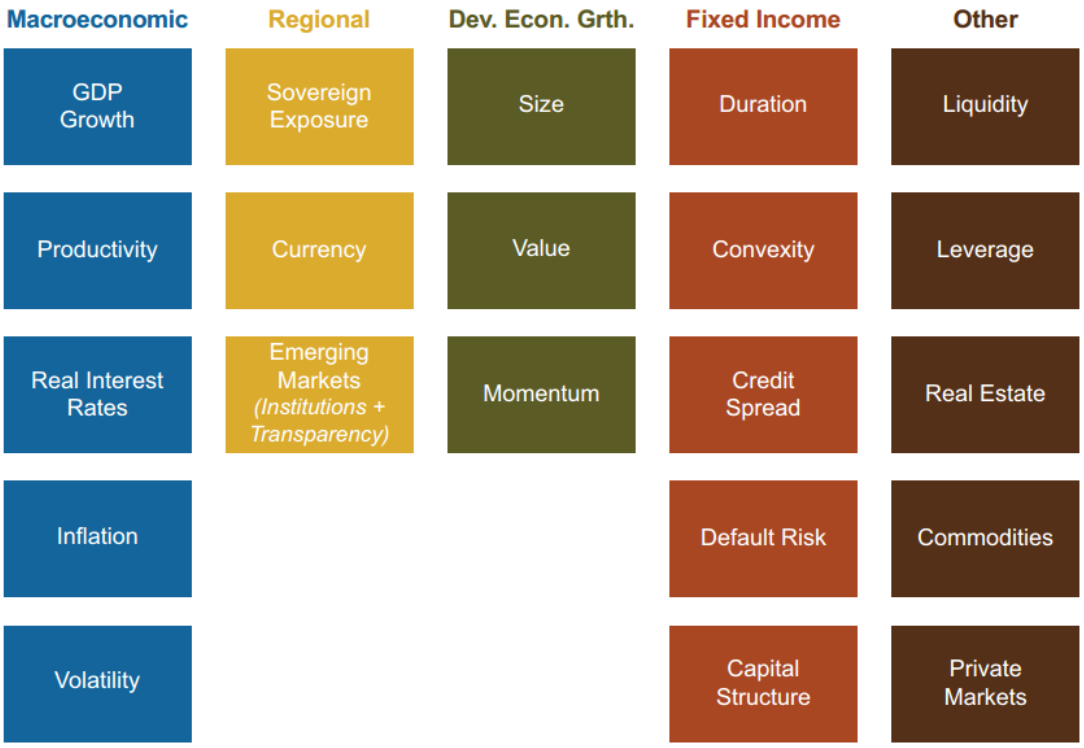

Figure 3: Sampling of factor and potential groupings. Figure taken from Podkaminer [36].

4.1 Motivation

In order to construct optimal portfolios, the investor needs to measure the volatilities, covariance matrix and correlations of the underlying assets' returns using historical data. However, this process is prone to great deal of false relationships and requires a large quantity of data depending on the considered number of assets.

For example, it is possible for the prices of two assets or stocks to move in the same direction but have a negative correlation measure, and viceversa. Figure 5 taken from Glantz and Kissell [19] is relevant. The price evolutions of two assets are depicted over 24 periods in top panel. Whereas the prices move in the same positive direction, the correlation between these two assets is negative. This is explained due to the excess terms being on the opposite sides of the price trend as seen in the bottom panel of Figure 5. Glantz and Kissell [19] concluded that relying on historical data alone, has a very high chance of providing incorrect price movement signals.

Moreover, if a portfolio consists of N-assets, the covariance matrix will be $N \times N$, with $N(N+1)/2$ unique covariance terms. To estimate these total parameters, a large number of observations is required to ensure that the number of degrees of freedom is at least positive. The number of unique elements in a covariance matrix is depicted in Figure 6(A) as a function of various number of assets. Figure 6(B) illustrates the number of days of data required for statistically significant estimation of covariance terms with various numbers of observations per each entry. Therefore, for a 500-asset portfolio (the size of the market index), a number

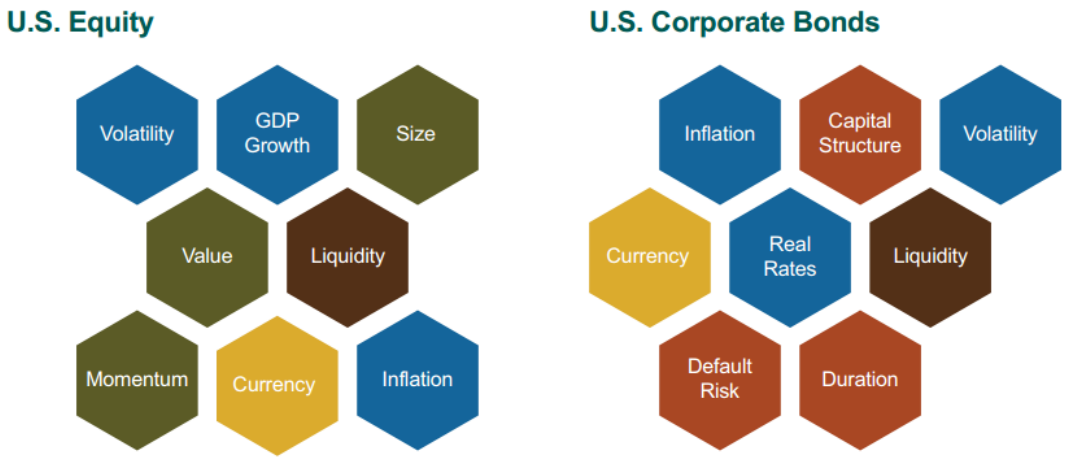

Figure 4: Common factor exposure across asset classes. Figure taken from Podkaminer [36].

of 5010 days of observations is required to achieve statistically meaningful results.

As underlined in Glantz and Kissell [19], factor models require much fewer observations for a reasonably accurate and stable estimate of covariance matrix and use a set of common explanatory factors across all assets, allowing for rigourous comparisons between factors.

4.2 General Formulation

A factor model has the following form

$$\mathbf{r}_i = \boldsymbol{\alpha}_i + F\boldsymbol{b}_i + \boldsymbol{e}_i, \tag{4.1}$$

$\mathbf{r}_i = \begin{bmatrix} r_{1i} \\ r_{2i} \\ \vdots \\ r_{ni} \end{bmatrix}$, $F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1k} \\ f_{21} & f_{22} & \cdots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nk} \end{bmatrix}$, $\boldsymbol{b}_k = \begin{bmatrix} b_{1i} \\ b_{2i} \\ \vdots \\ b_{ki} \end{bmatrix}$, $\boldsymbol{e}_i = \begin{bmatrix} e_{1i} \\ e_{2i} \\ \vdots \\ e_{ni} \end{bmatrix}$. The vector of asset returns for asset $i-$th is denoted by $\boldsymbol{r}_i$, whereas its entry r_{ti} is the associated return of asset $i-th$ in period t. The vector $\boldsymbol{\alpha}_i$ is constant, F is the matrix of factors' returns with f_{ti} denoting the factor $i-$th in period t, $\boldsymbol{b}_i$ denotes the vector of risk exposures with b_{ji} describing the risk sensitivity of asset i-th to factor j. Finally, $\boldsymbol{e}_i$ is the vector of errors with each entry e_{ti} characterizing the error of assets i-th in period t. Here we assumed n total number of time periods, N total number of stocks and k factors.

Parameters $\boldsymbol{b}_i$ of the model are obtained following a linear ordinary least squares (OLS) regression analysis described in the Appendix. Various weighting schemes can be applied depending on the investor views. For example, the more recent observations may have a

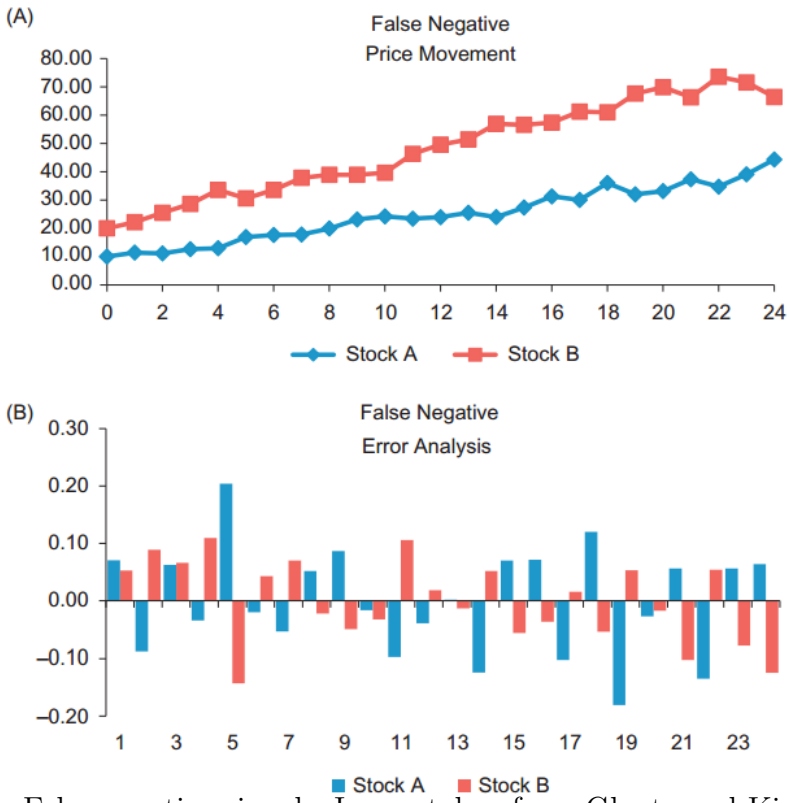

Figure 5: False negative signal - Image taken from Glantz and Kissell [19].

higher weight in the regression analysis. Several properties are required for a statistically correct regression analysis. First, the error vector $\boldsymbol{e}_i$ is normally distributed with mean 0 and some diagonal covariance matrix. Each factor f_{ti} is normally distributed $N(\bar{f}_{ti}, \sigma^2_{f_{ti}})$. The error terms e_{ti} and each factor f_{ti} are independent. The factors $f_{ti},\ i = 1, 2, .., k$ are mutually independent. Moreover, the error terms are mutually independent for all lagged time periods and across all stocks.

Often, real-world data lead to dependent factors which violates one of the above regression property. For these scenarios, regression can be employed to remove the correlations.

4.3 Estimating returns covariances using factor models

A general factor model with N assets can be written as

$$R = \alpha + F\beta + e, \tag{4.2}$$

where

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1N} \\ r_{21} & r_{22} & \cdots & f_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & f_{nN} \end{bmatrix}, \quad \alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1N} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nN} \end{bmatrix}, \quad F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1k} \\ f_{21} & f_{22} & \cdots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nk} \end{bmatrix},$$

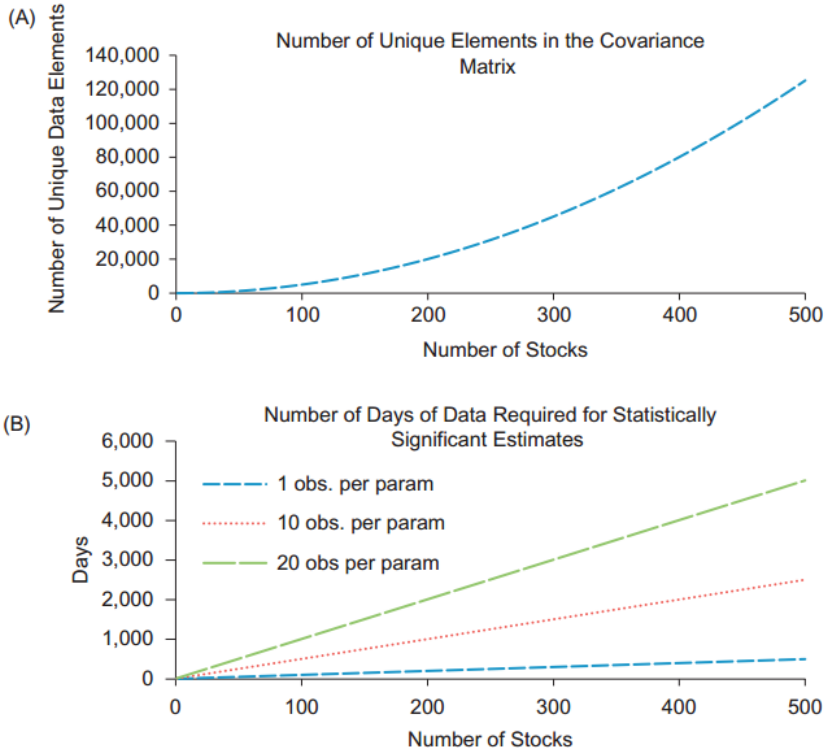

Figure 6: Data requirements for the construction of statistically significant covariances. Figure taken from Glantz and Kissell [19].

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1N} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kN} \end{bmatrix}, \quad e = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1N} \\ e_{21} & e_{22} & \cdots & e_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nN} \end{bmatrix}.$$

The formulas for the expected value of returns and excess returns, are given by

$$\begin{aligned} E[R] &= \alpha + \bar{F}\beta \\ R - E[R] &= (F - \bar{F})\beta + e. \end{aligned} \tag{4.3}$$

The analytic formulation of the covariance matrix C of the returns is computed as

$$\begin{aligned} C &= E\Big[(R - E[R])^T (R - E[R])\Big] = E\Big[((F - \bar{F})\beta + e)^T ((F - \bar{F})\beta + e)\Big] \\ &= E\Big[\beta^T (F - \bar{F})^T (F - \bar{F}) + 2\beta^T (F - \bar{F})^T e + e^T e\Big] \\ &= \beta^T E\Big[(F - \bar{F})^T (F - \bar{F})\Big]\beta + 2\beta^T E\Big[(F - \bar{F})^T e\Big] + E\Big[e^t e\Big]. \end{aligned} \tag{4.4}$$

Since factors returns and errors are independent than

$$\beta^T E\Big[(F - \bar{F})^T e] = 0.$$

Moreover, with the errors assumed independent accross all lagged time period and all assets, we obtain

$$E\Big[e^T e\Big] = \begin{bmatrix} \sigma_{e_1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{e_2}^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_{e_N}^2 \end{bmatrix} = \Lambda.$$

Finally

$$E\Big[(F-\bar{F})^T(F-\bar{F})\Big] = \begin{bmatrix} \sigma_{\boldsymbol{f}_1}^2 & cov(\boldsymbol{f}_1,\boldsymbol{f}_2) & \cdots & cov(\boldsymbol{f}_1,\boldsymbol{f}_k) \\ cov(\boldsymbol{f}_2,\boldsymbol{f}_1) & \sigma_{\boldsymbol{f}_2}^2 & \cdots & cov(\boldsymbol{f}_2,\boldsymbol{f}_k) \\ \cdots & \cdots & \cdots & \cdots \\ cov(\boldsymbol{f}_k,\boldsymbol{f}_1) & cov(\boldsymbol{f}_k,\boldsymbol{f}_2) & \cdots & cov(\boldsymbol{f}_k,\boldsymbol{f}_k) \end{bmatrix},$$

where $\boldsymbol{f}_i$ is the $i-$th column of matrix F.

The factor covariance matrix $cov(F)$ is a diagonal matrix in majority of the practical situations. Techniques to remove such correlations are available and rely on regression techniques. In consequence, the returns' covariance matrix derived from the factor model has the following expression

$$C = \beta^T cov(F)\beta + \Lambda. \tag{4.5}$$

4.4 Different types of factor models

Glantz and Kissell [19] classified the factor models into four categories depending on their applicability and type of employed factors. These classes include index models, macroeconomic models, cross-sectional or fundamental data models and statistical factor models.

Except for statistical factor models, the investor has to specify a-priori either a set of explanatory factors or a set of company-specific factor loadings. Index models and macroeconomic factor models are designed starting from a list of specific factors, while fundamental data models require factor loadings inputs and estimated explanatory factors.

In the case of statistical factor models, the investor does not make any prior assumptions regarding the list of explanatory factors. These factors together with the sensitivities of the returns with respect to these factors are both computed inside this framework.

Index models Depending on the number of indexes used as factors, these models are known as single-index and multi-index models. The single-index model employs only a single major market index such as the SP500. In the case of multi-index models, additional information based on the asset's sector index and asset's industry index are usually incorporated.

Macroeconomic models A macroeconomic model searches for relationships between assets' returns and a collection of macroeconomic variables such as inflation, GDP, industrial production, bond yields, etc. An advantage of using macroeconomic data as underlying factors is that such variables are easily measured and have solid economic meaning. However, the dynamic of a macroeconomic model may not sufficiently capture the correlation

structure of price movement across assets especially when different economic regimes are considered. Chen et al. [5] identified four macroeconomic factors with significant explanatory power with asset return. These are unanticipated changes in inflation, unanticipated changes in industrial production, unanticipated changes in the yield between high-grade and low-grade corporate bonds and unanticipated changes in the yield between long-term government bonds and t-bills. They are still valid for the current market.

Cross-sectional models Cross-sectional models approximate asset returns using a set of input data particular to each company, instead of the common factors across all the assets. The particular data consists of company characteristics and balance sheet information. More precisely, it includes trading activity metrics such as average daily trading volume, price momentum, size, etc. This type of data is also named fundamental data, explaining the alternative terminology of fundamental model used to characterize this particular factor models. Fama and French [12] found that market returns, company size (market capitalization), and book-to-market ratio could be used as factors due to their high correlations with the assets' returns. Whereas it seems naturally to incorporate fundamental data into multi-factor models, this method has some limitations. For example, it is difficult to access the company-specific data for a large variety of assets.

Statistical factor models For this category of models, the explanatory factors are directly derived from historical data. In this way, any preconceived bias is removed from the model which is advantageous. The major disadvantage is the lack of any real-world meaning of the factors thus making the analysis of the portfolio managers difficult to understand. Various methods can be used to derive statistical factor models. Among them we mention the eigendecomposition based on factorization of the sample covariance matrix, the singular value decomposition relying on the factorization of assets' returns matrix and factor analysis based on a maximum likelihood estimate of the correlations across stocks.

5 Conclusions

In this study, we have reviewed recent methodologies for optimal asset allocation problem from Modern Portfolio Theory to the Black-Litterman model and factor models. These methods were rigourously analysed together with their negative and positive implementation aspects. We underline that the highest difficulty in generating optimal portfolios is caused by the lack of accuracy in the prediction of the expected assets' returns and their covariances. The Black-Litterman model tries to alleviate this issue by constructing posterior distributions of the returns based on priors obtained from the CAPM model and likelihood functions derived from investors views of the market. Factor models recently emerged in the financial industry due to the continued market turbulence and asset price uncertainty and present great potential to increase the accuracy of the predicted assets' returns.

Several academics and practitioners studies presented irrefutable empirical evidences of non-normality in asset returns [42]. This suggests that non-linear factor models would be

more reliable than their linear counterparts at least for some specific market regimes. In consequence, non-linear regression techniques should be employed and we advocate for the use of neural network techniques also described in the Appendix. We also consider that factor models can be greatly improved using regime switching and sensitivity analysis as described in [23]. Better forecasted returns will greatly contribute to decreasing the uncertainty in the current optimal portfolio strategies leading to better asset allocation management.

References

[1] Clifford S Asness, Tobias J Moskowitz, and Lasse Heje Pedersen. Value and momentum everywhere. The Journal of Finance, 68(3):929–985, 2013.

[2] Christopher M Bishop. Pattern recognition. Machine Learning, 128, 2006.

[3] Fischer Black and Robert B Litterman. Asset allocation: combining investor views with market equilibrium. The Journal of Fixed Income, 1(2):7–18, 1991.

[4] Mark M Carhart. On persistence in mutual fund performance. The Journal of Finance, 52(1):57–82, 1997.

[5] Nai-Fu Chen, Richard Roll, and Stephen A Ross. Economic forces and the stock market. Journal of business, pages 383–403, 1986.

[6] Wing Cheung. The Black–Litterman model explained. Journal of Asset Management, 11(4):229–243, 2010.

[7] Wing Cheung. The augmented Black–Litterman model: a ranking-free approach to factor-based portfolio construction and beyond. Quantitative Finance, 13(2):301–316, 2013.

[8] Y. Choueifaty and Y. Coignard. Toward maximum diversification. The Journal of Portfolio Management, 35(1):40–51, 2008.

[9] Maria Correia, Scott Richardson, and İrem Tuna. Value investing in credit markets. Review of Accounting Studies, 17(3):572–609, 2012.

[10] Claude B Erb and Campbell R Harvey. The strategic and tactical value of commodity futures. Financial Analysts Journal, 62(2):69–97, 2006.

[11] Real estate portfolio risk management and monitoring. Paul Mouchakkaa. Technical report, 2012. URL `http://www.morganstanley.com/realestate/Observatory.pdf`.

[12] Eugene F Fama and Kenneth R French. The cross-section of expected stock returns. the Journal of Finance, 47(2):427–465, 1992.

[13] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1):3–56, 1993.

[14] Eugene F. Fama and Kenneth R. French. A five-factor asset pricing model. Technical report, SSRN, 2014. URL http://ssrn.com/abstract=2287202.

[15] Jack Clark Francis and Dongcheol Kim. Modern Portfolio Theory: foundations, analysis, and new developments, volume 795. John Wiley & Sons, 2013.

[16] Andrea Frazzini and Lasse Heje Pedersen. Betting against beta. Journal of Financial Economics, 111(1):1–25, 2014.

[17] G David Garson. Testing statistical assumptions. Asheboro, NC: Statistical Associates Publishing, 2012.

[18] A George, F Seber, and CJ Wild. Nonlinear regression, 2003.

[19] Morton Glantz and Robert Kissell. Multi-Asset Risk Modeling: Techniques for a Global Economy in an Electronic and Algorithmic Trading Era. Academic Press, 2013.

[20] Gary Gorton and K Geert Rouwenhorst. Facts and fantasies about commodity futures. Financial Analysts Journal, 62(2):47–68, 2006.

[21] Trevor J.. Hastie, Robert John Tibshirani, and Jerome H Friedman. The elements of statistical learning: data mining, inference, and prediction. Springer, 2011.

[22] Guangliang He and Robert Litterman. The intuition behind Black-Litterman model portfolios. Available at SSRN 334304, 2002.

[23] Cristian Homescu. Better investing through factors, regimes and sensitivity analysis. Regimes and Sensitivity Analysis (January 25, 2015), 2015.

[24] Patrick Houweling and Jeroen Van Zundert. Factor investing in the corporate bond market. Available at SSRN 2516322, 2015.

[25] Antti Ilmanen, Rory Byrne, Heinz Gunasekera, and Robert Minikin. Which Risks Have Been Best Rewarded? The Journal of Portfolio Management, 30(2):53–57, 2004.

[26] Michael C Jensen, Fischer Black, and Myron S Scholes. The capital asset pricing model: Some empirical tests. 1972.

[27] SP Kothari, Konstantin Danilov, and Gitanjali Meher Swamy. Generating superior performance in private equity: A new investment methodology. 2013.

[28] S. Maillard, T. Roncalli, and J Teiletche. The properties of equally weighted risk contribution portfolios. The Journal of Portfolio Management, 36:60–70, 2010.

[29] Harry Markowitz. Portfolio selection. The Journal of Finance, 7(1):77–91, 1952.

[30] Harry M Markowitz, G Peter Todd, and William F Sharpe. Mean-variance analysis in portfolio choice and capital markets, volume 66. John Wiley & Sons, 2000.

[31] Lukas Menkhoff, Lucio Sarno, Maik Schmeling, and Andreas Schrimpf. Currency momentum strategies. Journal of Financial Economics, 106(3):660–684, 2012.

[32] Attilio Meucci. Risk and asset allocation. Springer Science & Business Media, 2009.

[33] Kevin P Murphy. Machine learning: a probabilistic perspective. MIT press, 2012.

[34] Jorge Nocedal and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.

[35] John Okunev and Derek White. Do momentum-based strategies still work in foreign currency markets? Journal of Financial and Quantitative Analysis, 38(02):425–447, 2003.

[36] Eugene L Podkaminer. Risk factors as building blocks for portfolio diversification: The chemistry of asset allocation. Investment Risk and Performance Newsletter, 2013(1), 2013.

[37] L. Pedersen R. Koijen, T. Moskowitz and E. Vrugt. Carry. Technical report, 2013. URL `http://www.nber.org/papers/w19325`.

[38] Prasad Ramani. Evolution of portfolio construction methodologies. Technical report, 2012. URL `www.garp.org/media/1110362/evolutionofportfolioconstructionmethodologies_prasadramani_121212.pdf`.

[39] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[40] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[41] David Ruppert. Statistics and finance: an introduction. Springer Science & Business Media, 2004.

[42] Bernd Scherer. Portfolio construction and risk budgeting. Risk Books, 2002.

[43] William F Sharpe. The sharpe ratio. The Journal of Portfolio Management, 21(1): 49–58, 1994.

Appendix - Regression Techniques

Linear Regression

Regression diagnostics: testing the assumptions of linear regression

Here we provide a general overview of the principal assumptions of linear regression models, how to diagnose and fix them. For an exhaustive list of statistical assumptions, we refer to

[17]. Linear regression makes several key assumptions. These are linearity of the relationship between dependent and independent variables, multicollinearity of the predictors, normal iid residuals and homoscedasticity.

Violating the linearity will most likely lead to large prediction errors especially when one extrapolates outside the range of the sample data. *How to diagnose*: nonlinearity can be noticed in a plot of observed versus predicted values or a plot of residuals versus predicted values. The points should be symmetrically distributed around a diagonal line in the former plot or around a horizontal line in the latter plot, with a constant variance. For multiple regression models, nonlinearity is revealed by noticeable patterns in the plots of the residuals versus individual independent variables. *How to fix*: one may consider applying a nonlinear transformation to the dependent and/or independent variables. For example, the log transformation can be considered an option if the data are strictly positive. In case a log transformation is applied to the dependent variable only, this is equivalent to assuming that it behaves exponentially as a function of the independent variables. Other possibility to consider is adding another regressor that is a nonlinear function of one of the other variables. For example, if the graph of residuals versus forecasted variables suggests a parabolic curve, then it makes sense to use the square of the current predictors as new independent variables. Finding missing entirely different independent variables can also address the linearity issue.

Linear regression also assumes that there is little or no multicollinearity in the data. Lack of multicollinearity in the predictors leads to a non-unique solution of the linear regression problem. This can be induced by having two or more perfectly correlated predictor variables. It can be also noticed when there is not enough observations available compared to the number of parameters to be estimated. *How to diagnose*: one has to compute the matrix of Pearson's bivariate correlation among all independent variables and check for correlation coefficients equal with 1. For such cases, the duplicated predictors need to be removed. One can also employ a tolerance statistic to quantify the influence of one independent variable on all others independent variables. The tolerance formula is $T = 1 - R^2$, where R^2 is the coefficient of determination. For $T < 0.01$, the multicollinearity in the data is certain. The variance inflation factor and condition index can be also used to diagnose multicollinearity. *How to fix:* In case multicollinearity is found in the data, centering the data around the mean may solve the problem. Other alternatives can be considered such as factor analysis or simply rotating the factors to assure independence.

Independent observations are assumed by most statistical procedures including linear regression. *How to diagnose:* one can check the plots of the residuals versus independent variables. The residuals should be randomly and symmetrically distributed around zero under all conditions. In particular, there should be no correlation between consecutive errors. *How to fix:* the linearity assumption has to be checked together with possible omitted variables.

Breaking the normality assumption of the error creates problems for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts. For significantly non-normal error distributions, the confidence intervals may be too wide or too narrow. *How to diagnose:* there are several methods such as Kolmogorov-Smirnov test, Shapiro-Wilk test, Jarque-Bera test, and Anderson-Darling test. One can also

analyze a Q-Q plot, compute the skewness and kurtosis, or plot the histogram or the kernel density estimation. *How to fix*: this can be explained by significant non-normal distributions of the dependent and independent variables or violations of the linearity assumption. For such scenarios, a nonlinear transformation of variables might cure both problems. Another possibility is that there are two or more subsets of the data having different statistical properties and building separate models should fix the problem.

The violation of the constant variance (homoscedasticity) makes difficult to measure the true standard deviation of the forecast errors, resulting in confidence intervals that are too wide or too narrow. Heteroscedasticity may also have the effect of giving too much weight to a small subset of the data when estimating coefficients. The presence of outliers may cause such behaviour. *How to diagnose*: by simply plotting the residuals versus predicted values one can identify the lack of constant variance. *How to fix:* In case the dependent variable is strictly positive and for residual-versus-predicted plots showing a proportionality between the size of the errors and the size of the predictions, a log transformation applied to the dependent variable may fix the issue. Normalizing the data may also alleviate the problem. Heteroscedasticity can also be a consequence of an important violation of the linearity or independence assumptions.

General formulation

Without reducing the generality, we can rewrite the factor model (4.2) as below

$$R = F\beta_0 + e, \tag{5.1}$$

and recall that the errors are unbiased and independent and identically distributed with variances σ_{ij}^2

$$E[e_{ij}] = 0 \text{ and } var(e_{ij}) = \sigma_{ij}^2,\ cov(e_{ij}, e_{kl}) = 0.$$

To construct an estimator $\hat{\beta}$ for the risk sensitivity matrix β_0, we search for β that minimizes the ordinary least square functional

$$\min_{\beta} J(\beta) = (R - F\beta)^T (R - F\beta).$$

The solution to this problem is to set the gradient ∇J equal to 0 and solving for β. Specifically, one sets

$$\nabla_\beta J(\beta) = 2F^T(R - F\beta) = 0$$

to obtain the least-square estimate

$$\hat{\beta} = (F^T F)^{-1} F^T R. \tag{5.2}$$

Overfitting occurs when a model captures idiosyncrasies of the input data, rather than generalizing. One way to avoid overfitting is to perform regularization. Introducing a penalty term for the size of the weight we obtain

Ridge regularization regression problem

$$\min_{\beta} = (R - F\beta)^T (R - F\beta) + \frac{\lambda}{2}\|\beta\|_{FRO}^2,$$

and

Lasso regularization regression problem

$$\min_{\beta} = (R - F\beta)^T(R - F\beta) + \frac{\lambda}{2}\|\beta\|_1,$$

where $\|\cdot\|^2_{FRO}$ is the Frobenius norm and $\|\cdot\|_1$ is the matrix $1-$norm.

The constraint for ridge regression is a disk, whereas for lasso is a rhomboid. Both methods find the first point where the elliptical contours of the least squares error function hit the constraint region. Since the rhomboid has corners, the iterate have several components equal with zero. One can generalize ridge and lasso regression viewing them as Bayes estimates (see book).

Variance of $\hat{\beta}$

In order to asses the sampling properties of $\hat{\beta}$ defined in (5.2), we assume that the observations are uncorrelated and have constant variance σ^2, and that the predictors are fixed (non random). By simply rearranging matrices to vectors in equation (5.1), we obtain the vector notation

$$\mathbf{R} = F\boldsymbol{\beta}_0 + \mathbf{e}, \tag{5.3}$$

where $\mathbf{R}$, $\boldsymbol{\beta}_0$ and $\mathbf{e}$ are vectors.

The variance-covariance matrix of the least squares parameters estimates is easily derived

$$var(\hat{\boldsymbol{\beta}}) = (F^T F)^{-1}\sigma^2. \tag{5.4}$$

To test the hypothesis that a particular coefficient $\beta_j = 0$, we form the standardized coefficient or Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}, \tag{5.5}$$

where v_j is the jth diagonal element of $F^T F^{-1}$. In the above equation, we also used the sample standard deviance $\hat{\sigma}$. Under the null hypothesis that $\beta_j = 0$, z_j is distributed as $t-$student distribution with $M - p - 1$ degrees of freedom. Here M is the size of the observations vector $\mathbf{R}$ and p is the size of the vector $\hat{\boldsymbol{\beta}}$. A large value of z_j will lead to rejection of the null hypothesis. If $\hat{\sigma}$ is replaced by a known value σ, than z_j would have a standard normal distribution.

Model Selection

The Akaike information criterion (AIC) and Bayesian information criterion (BIC) are model selection criteria based on the log-likelihood and can be used to select among the independent variables.

According to Ruppert [41], AIC is defined as

$$-2log(L) + 2 \cdot p, \tag{5.6}$$

where L is the likelihood evaluated at the maximum likelihood. The Bayesian information criterion is defined as

$$-2log(L) + log(M) \cdot p, \tag{5.7}$$

where M is the length of the observation series and p is the number of considered independent variables. The best model according to either criterion is the model that minimizes that criterion.

For normal likelihood functions, $-2log(L)$ is equal with the sum of the squared residuals (SSQ) over the constant variance of the residuals σ^2. According to [21], the AIC information criterion is defined as

$$AIC = \frac{SSQ}{M} + 2\frac{p}{M}\hat{\sigma}^2 \tag{5.8}$$

whereas BIC is formulated as

$$BIC = \frac{SSQ}{\sigma^2} + log(M)\frac{p}{M}\sigma^2. \tag{5.9}$$

For model selection purposes, there is no clear choice between AIC and BIC. However, BIC has an important asymptotic property. Giving a family of models, including the true model, the probability that BIC will select the correct model approaches one as the sample size $M \to \infty$. This is not the case for AIC, which tends to select models which are too complex for $M \to \infty$. On the other hand, for finite number of observations, BIC often select models that are too simple, since it heavily penalizes the model complexity.

Nonlinear Regression

The relationship between the underlying factors and assets' return can be modeled using a non-linear model

$$R = f(\beta_0) + e.$$

The ordinary least squares estimate is obtained by minimizing

$$\min_{\beta} J(\beta) = (R - f(\beta))^T (R - f(\beta)).$$

The difficulty is that analytic expressions for these minimizers generally can not be obtained for nonlinearly parameterized problems. Instead, the solutions must be obtained by minimizing the least square objective function. Several results analogous to the linear theory can be found in [18].

Supervised machine learning techniques for regression models

Neural networks and Gaussian processes are used to build a probabilistic model $\phi : \mathbf{z} \mapsto \hat{y}$, where ϕ is a transformation function that learns through the input features $\mathbf{z}$ to estimate the output $y \in \mathbb{R}$, [33]. In the context of factor models, $\mathbf{z}$ can be seen as an underlying factor and y is an asset return. Extension to multiple factors is straightforward. In what follows we briefly review the Gaussian process and Neural network techniques.

Gaussian process kernel method

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution [39]. A Gaussian process is fully described by its mean and covariance functions

$$\phi(\mathbf{z}) \sim \text{gp}\left(m(\mathbf{z}), \mathbf{K}\right),$$

where $m(\mathbf{z}) = \mathbb{E}\left[\phi(\mathbf{z})\right]$ [39], and $\mathbf{K}$ is the covariance matrix with entries

$$K_{i,j} = \mathbb{E}\left[\left(\phi(\mathbf{z}^i) - m(\mathbf{z}^i)\right)\left(\phi(\mathbf{z}^j) - m(\mathbf{z}^j)\right)\right]$$

The commonly used squared-exponential-covariance Gaussian kernel [39] is introduced below

$$k : \mathbb{R}^r \times \mathbb{R}^r \mapsto \mathbb{R},\ k(\mathbf{z}^i, \mathbf{z}^j) = \sigma_\phi^2 \exp\left(-\frac{\|\mathbf{z}^i - \mathbf{z}^j\|}{2\,\hbar^2}\right) + \sigma_n^2\,\delta_{i,j}, \tag{5.10}$$

and $K_{ij} = k(\mathbf{z}^i, \mathbf{z}^j)$, where $\mathbf{z}^i$ and $\mathbf{z}^j$ are the pairs of data points in training or test samples, δ is the Kronecker delta symbol and $\|\cdot\|$ is some appropriate norm. This model has three hyper-parameters. The length-scale $\hbar$ governs the correlation among data points. The signal variance $\sigma_\phi^2 \in \mathbb{R}$ and the noise variance $\sigma_n^2 \in \mathbb{R}$ govern the precision of variance and noise, respectively.

Consider a set of training data points $\mathbf{Z} = [\mathbf{z}^1\ \mathbf{z}^2 \cdots \mathbf{z}^n] \in \mathbb{R}^{r\times n}$ and the corresponding noisy observations $\mathbf{y} = [y^1\ y^2 \cdots y^n] \in \mathbb{R}^{1\times n}$,

$$y^i = \phi(\mathbf{z}^i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}\left(0, \sigma_n^2\right), \quad i = 1, \ldots, n. \tag{5.11}$$

Consider also the set of test points $\mathbf{Z}^* = [\mathbf{z}^{1*}\ \mathbf{z}^{2*} \cdots \mathbf{z}^{m*}] \in \mathbb{R}^{r\times m}$ and the predictions $\hat{\mathbf{y}} = [\hat{y}^1\ \hat{y}^2 \cdots \hat{y}^m] \in \mathbb{R}^{1\times m}$,

$$\hat{y}^i = \phi\left(\mathbf{z}^{i*}\right), \quad i = 1, \ldots, m. \tag{5.12}$$

For a Gaussian prior the joint distribution of training outputs $\mathbf{y}$ and test outputs $\hat{\mathbf{y}}$ is

$$\begin{bmatrix}\mathbf{y}^T \\ \hat{\mathbf{y}}^\mathbf{T}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}(\mathbf{Z})^T \\ \mathbf{m}(\mathbf{Z}^*)^T\end{bmatrix},\ \begin{bmatrix}\mathbf{K} & \mathbf{K}^* \\ \mathbf{K}^{*T} & \mathbf{K}^{**}\end{bmatrix}\right), \tag{5.13}$$

where

$$\mathbf{m}(\mathbf{Z}) = [m(\mathbf{z}^1)\ m(\mathbf{z}^2) \cdot m(\mathbf{z}^n)] \in \mathbb{R}^{1\times n},\ \mathbf{m}(\mathbf{Z}^*) = [m(\mathbf{z}^{1*})\ m(\mathbf{z}^{2*}) \cdot m(\mathbf{z}^{m*})] \in \mathbb{R}^{1\times m},$$

$$\mathbf{K}^* = (K^*_{ij})_{i=1,..,n;\ j=1,..,m} = k(\mathbf{z}^i, \mathbf{z}^{j*}) \text{ and } \mathbf{K}^{**} = (K^{**}_{ij})_{i=1,..,m;\ j=1,..,m} = k(\mathbf{z}^{i*}, \mathbf{z}^{j*}).$$

The predictive distribution represents the posterior after observing the data [2] and is given by:

$$p\left(\hat{\mathbf{y}}|\mathbf{Z}, \mathbf{y}, \mathbf{Z}^*\right) \sim \mathcal{N}\left(\mathbf{K}^{*T}\mathbf{K}^{-1}\mathbf{y}\,,\, \mathbf{K}^{**} - \mathbf{K}^{*T}\mathbf{K}^{-1}K^*\right). \tag{5.14}$$

The prediction of Gaussian process will depend on the choice of the mean and covariance functions, and on their hyperparameters $\hbar$, σ^2_ϕ and σ^2 which can be inferred from the data by minimizing the marginal negative log-likelihood function $\boldsymbol{\theta} = \arg\min L(\boldsymbol{\theta})$, where

$$L(\boldsymbol{\theta}) = -\log\, p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta}) = \frac{1}{2}\log\det(\mathbf{K}) + \frac{1}{2}(\mathbf{y} - \mathbf{m}(\mathbf{Z}))\,\mathbf{K}^{-1}\,(\mathbf{y} - \mathbf{m}(\mathbf{Z}))^T + \frac{n}{2}\,\log(2\pi)\,.$$

Artificial Neural networks

Artificial Neural networks (ANNs) detect the pattern of data by discovering the input–output relationships. ANNs consist of neurons and connections between the neurons (weights). Neurons are organized in layers, where at least three layers of neurons (an input layer, a hidden layer, and an output layer) are required for construction of a neural network. The input layer distributes input signals or factors $\mathbf{z} = [z_1\ z_2 \cdots z_r]$ to the hidden layer. For a neural network with L hidden layers and m^ℓ neurons in each hidden layer, let $\hat{\mathbf{y}}^\ell = [\hat{y}^\ell_1\ \hat{y}^\ell_2 \cdots]$ be the vector of outputs from layer ℓ, $\mathbf{b}^\ell = [b^\ell_1\ b^\ell_2 \cdots]$ the biases at layer ℓ, and $\mathbf{w}^\ell_j = [w^\ell_{j1} w^\ell_{j2} \cdots]$ the weights connecting the neuron j to the input. The vectors $\hat{\mathbf{y}}^\ell$ and $\mathbf{w}^\ell_j$ share the same dimension which varies along the layers depending on the number of input features, neurons and outputs. Then the feed-forward operation is:

$$\begin{aligned} x^{\ell+1}_j &= \mathbf{w}^{\ell+1^T}_j \hat{\mathbf{y}}^\ell + \mathbf{b}^{\ell+1}_j, \quad \hat{\mathbf{y}}^0 = \mathbf{z}, \quad j = 1, \cdots m^\ell \\ \hat{y}^{\ell+1}_j &= \varphi\left(\mathbf{x}^{\ell+1}\right), \quad \ell = 0, 1, \cdots, L-1. \end{aligned}$$

The differentiable function φ is the transfer function and can be log-sigmoid, hyperbolic tangent sigmoid, or linear transfer function.

The training process of ANN adjusts the weights and the biases in order to reproduce the desired outputs when fed the given inputs. The training process via the back propagation algorithm [40] uses a gradient descent method to modify weights and thresholds such that the error between the desired output and the output signal of the network is minimized. In supervised learning the network is provided with samples from which it discovers the relations of inputs and outputs. The output of the network is compared with the desired output, and the error is back-propagated through the network and the weights will be adjusted. This process is repeated during several iterations, until the network output is close to the desired output.