Optimal Financial Portfolio Selection

George Derpanopoulos

May 8, 2018

Abstract

Modern Portfolio Theory (MPT) has been the canonical theoretical model of portfolio selection for over 60 years, yet it faces limited adoption among practitioners. This is because MPT's main inputs, assets' expected returns and covariances, are estimated with noise, while the solution to its optimization problem requires the inversion of an ill-conditioned matrix. As a result, MPT often produces unstable portfolios with extreme weights. This study reviews and evaluates several methods for altering MPT's inputs and optimization problem to produce more stable and diversified portfolios, without discarding MPT's intuitive assumptions and structure. These methods are: robust estimators and shrinkage estimators of expected returns and covariances, covariances based on statistical models of returns, sparse graphical models of inverse covariance matrices, filtered covariance matrices, portfolio optimization that incorporates uncertainty in expected returns, and portfolio optimization with penalties on weights' norm. To evaluate competing methods, I construct their respective portfolios using monthly data on 92 assets and 90 rolling training periods of 15-year length. Comparing these portfolios' out-of-sample performance across several financial metrics and rolling test periods of 6-month length, I find that most alternatives outperform the standard MPT approach. However, I also find that only the L1-norm penalized portfolio marginally outperforms the benchmark equal-weighted portfolio, and owes its good performance to limiting short-sales.

1 Introduction

Since the Nobel-prize-winning work of Markowitz (1952, 1959), modern portfolio theory (MPT) has attracted the attention of several disciplines: finance, statistics, economics, operations research, electrical engineering, and computer science, among others.[1] Despite its attractive normative and statistical properties, the MPT framework is far from universally adopted by portfolio managers, a phenomenon Michaud (1989) labeled the "Markowitz enigma". MPT's low adoption rate is owed to several limitations, some theoretical, others practical. Among the former, the largest limitation is the instability of optimal portfolio weights and the seemingly non-sensical investment decisions they produce. Among the latter, the largest practical limitation is that MPT produces portfolios that incur high transaction costs, especially for individual retail investors.[2] Crucially, the increasing dominance of passive over active investing, coupled with the advent of so-called "robo-advisors", has re-popularized MPT.[3] However, the framework's core shortcomings remain, thus creating a stronger call to repair MPT.

In this study, I explore several approaches to overcoming MPT's main limitations. Since MPT is composed of three main ingredients – estimates of assets' expected returns, estimates of covariance in those returns, and a constrained optimization problem – I present approaches that attempt to improve each of those ingredients. These approaches include robust estimators and shrinkage estimators of expected returns and covariances, covariances based on statistical models of returns, sparse graphical models of inverse covariance matrices, filtering covariance matrices based on their spectral properties, portfolio optimization that incorporates uncertainty in expected returns, and portfolio optimization with penalties on the norm of weights. Throughout, I attempt to present approaches from diverse methodological backgrounds in a unified framework. Additionally, I devote more attention to newer

[1] Throughout, I use "Modern Portfolio Theory", "Markowitz model", and "mean-variance portfolio optimization" interchangeably.

[2] I return to the limitations of MPT in Section 2.

[3] Wealthfront, the second largest robo-advisor (as measured by assets-under-management), champions the use of MPT in several research notes on its methodology (Wealthfront, 2018).

methods (e.g. sparse graphical modeling, penalized optimization), which have a stronger grounding in statistical theory and also present more opportunities for development.

To evaluate the above approaches, I conduct a comprehensive empirical comparison of their financial performance. Using real data on the monthly returns of 92 assets listed on US stock exchanges over the last 60 years, I construct the portfolios selected by each method in 90 rolling training periods of 15-year length, hold each portfolio for 6 months, and record its out-of-sample performance across a range of financial metrics—return, risk, risk-adjusted returns, diversification, gross exposure, exposure through short-sales, and turnover. In addition to the standard MPT portfolio and the alternative approaches mentioned above, I evaluate two benchmarks, the S&P500 index and the equal-weighted portfolio, resulting in 18 portfolios.

Overall, in line with the literature, I find that the standard MPT approach performs poorly, and most alternative methods beat it. On the contrary, but also in agreement with previous findings, I find the equal-weighted portfolio to be a tough benchmark. Indeed, only the L1-norm penalized portfolio beats the equal-weighted portfolio on risk-adjusted returns. This is most likely owed to the fact that the optimal value of the L1 penalty parameter calls for no short-selling. Moreover, the L1 portfolio is the top-performer across essentially all metrics, which confirms previous research that short-selling creates a net loss for portfolio performance—by inducing more risk, leverage, and instability than the additional returns generated. Following the L1 portfolio's performance is the L2-norm penalized portfolio and the portfolio constructed through filtered covariance matrices, while portfolios constructed through shrinkage estimators and model-based estimators also perform respectably.

The remainder of this study proceeds as follows. Section 2 reviews the standard MPT framework and its statistical limitations. Section 3 reviews the various approaches for dealing with those limitations. Section 4 presents an empirical evaluation of competing approaches and discusses its results. Section 5 summarizes and concludes.

2 MPT Framework

In this section, I present MPT's assumptions, optimization problem, and solution, as well as its statistical limitations.

2.1 Standard Approach

Assumptions MPT makes numerous assumptions about investors and the market, which I adopt throughout this study, unless otherwise noted. First, investors are identical, risk-averse, and rational.[4]. Second, investors minimize portfolio risk while maximizing portfolio expected returns, optimizing over just one period. Third, investors select portfolios solely on the basis of their expected returns and risk, where the latter is measured as the variance of portfolio returns.[5] Fourth, asset returns are stationary over time. Fifth, investors know the price of all assets considered for investment, and update their portfolio according to changes in asset prices immediately and costlessly. Sixth, asset prices are exogenous (no investor's choices affect asset prices). Seventh, all assets considered for investment are infinitely liquid, thus trades of any size can be made on those assets. Eighth, investors can take negative or "short" positions on assets.[6] Ninth, investors can borrow and lend without risk and at the same interest rate. Tenth, investors incur no transaction costs (e.g. taxes, brokerage fees, bid-ask spreads, foreign exchange commissions). Finally, investors allocate all of their budget to their portfolio (no savings).

[4] Formally, from a utility theory perspective, risk-aversion implies that the investor's expected utility is a concave function of her wealth.

[5] Some formulations of MPT assume that returns are distributed multivariate normal. Strictly speaking, this assumption is too strong, since mean-variance optimization of returns only requires that returns are fully described by their mean, variance, and covariance. However, though there are distributions other than the multivariate normal that satisfy this requirement, assuming a multivariate normal has practical advantages.

[6] A short-sale or short postion on an asset is a bet that its price will decrease. Mathematically, a short position on asset i is represented as a negative portfolio weight ($w_i < 0$). In practice, opening or initiating a short position on asset i at time t involves borrowing the asset from a broker and selling it at its current value $p_{i,t}$. Closing or covering a short position on i, say at $t+1$, involves selling i, receiving $p_{i,t+1}$, returning $p_{i,t}$ plus any fees (d) charged by the broker, and retaining $p_{i,t} - p_{i,t+1} - d$. Clearly, if i's price decreases by more than d during $[t, t+1]$, a short-sale is profitable. Note that an investor's maximum gain from shorting i is $p_{i,t} - d$, while her maximum loss is infinite. Due to this risk, many institutional investors are prohibited from directly engaging in short-selling (e.g. university endowment funds).

Notation Throughout, I use $i = 1, \ldots, N$ to denote assets and $t = 1, \ldots, T$ to denote time periods. For all i and t, asset prices are denoted by $p_{i,t} \in \mathbb{R}_+$, asset returns (between adjacent time periods) by $r_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}} \in \mathbb{R}$, (true) expected asset returns (during some time period) by $\mu_i = \mathbb{E}[r_i] \in \mathbb{R}$, (true) covariances of asset returns (during some time period) by $\sigma_{ij} \in \mathbb{R}$, (true) variances of asset returns (during some time period) by $\sigma_{ii} = \sigma_i^2 \in \mathbb{R}_+$, and portfolio weights (fraction of wealth allocated to each asset) by $w_i \in \mathbb{R}$. Column vectors are denoted by lower-case bold letters, such as that of assets' expected returns ($\mathbb{E}[\boldsymbol{r}] = \boldsymbol{\mu} \in \mathbb{R}^N$) and that of portfolio weights ($\boldsymbol{w} \in \mathbb{R}^N$), and matrices are denoted by upper-case bold letters, such as that of the true covariances of asset returns ($\mathbb{E}[(\boldsymbol{r} - \boldsymbol{\mu})(\boldsymbol{r} - \boldsymbol{\mu})^\intercal] = \boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$). Estimates of true parameters are denoted through the hat symbol, as with the sample estimate of the covariance matrix of asset returns ($\hat{\boldsymbol{\Sigma}}$).

Optimization Problem There are many formulations of the MPT optimization problem and the resulting optimal (Markowitz) portfolio. For this reason, I attempt to frame the optimization problem in a way that is transferable across the different approaches I will present. The Markowitz portfolio arises from an investor solving the MPT problem: selecting a set of weights that minimizes portfolio risk, subject to meeting a level of expected returns and allocating all of her wealth (normalized to 1)

$$\boldsymbol{w}_* = \underset{\boldsymbol{w}}{\operatorname{argmin}} \quad \boldsymbol{w}^\intercal \boldsymbol{\Sigma} \boldsymbol{w} \tag{1a}$$

$$\text{s.t.} \quad \boldsymbol{w}^\intercal \boldsymbol{\mu} = r_* \tag{1b}$$

$$\boldsymbol{w}^\intercal \mathbf{1}_N = 1 \tag{1c}$$

where r_* is the target level of returns and $\mathbf{1}_N$ is an N-length column vector.[7]

Risk-Free Borrowing/Lending Assume the investor can borrow and lend through selling and buying a risk-free asset (e.g. US Treasury bill), with fixed return, typically $r_f >$

[7]The more common formulation of the MPT optimization problem does not set a level of target returns and, instead, minimizes portfolio risk minus returns, or maximizes portfolio returns minus risk (DeMiguel, Garlappi and Uppal, 2009).

0.[8] Denoting the amount invested in the risk-free asset as w_f, 1c now becomes $\tilde{w}_f + \tilde{\boldsymbol{w}}^\intercal \mathbf{1}_N = 1$, which can be rearranged to $\tilde{\boldsymbol{w}}^\intercal \mathbf{1}_N = 1 - \tilde{w}_f$. This allows us to choose weights subject only to the target returns constraint (1b), since the allocation constraint (1c) can always be satisfied simply by lending or borrowing enough to force the total wealth allocated to 1. Furthermore, note that portfolio return is $r_p = \tilde{w}_f r_f + \tilde{\boldsymbol{w}}^\intercal \boldsymbol{r}$. Using the rearranged expression for $\tilde{w}_f$, portfolio return can be expressed as the sum of the return from the risk-free asset plus "excess return" from the N risky assets: $r_p = r_f + \tilde{\boldsymbol{w}}^\intercal(\boldsymbol{r} - \mathbf{1}_N r_f)$. Defining assets' excess return vector $\tilde{\boldsymbol{r}} = \boldsymbol{r} - \mathbf{1}_N r_f$, portfolio return becomes $r_p = r_f + \tilde{\boldsymbol{w}}^\intercal \tilde{\boldsymbol{r}}$, or $\tilde{r}_p = \tilde{\boldsymbol{w}}^\intercal \tilde{\boldsymbol{r}}$. Finally, the vector of assets' *expected* excess return is $\tilde{\boldsymbol{\mu}} = \mathbb{E}[\tilde{\boldsymbol{r}}] = \mathbb{E}[\boldsymbol{r}] - \mathbf{1}_N r_f = \boldsymbol{\mu} - \mathbf{1}_N r_f$. Substituting this into constraint 1b and dropping constraint 1c, we have the optimization problem

$$\tilde{\boldsymbol{w}}_* = \underset{\tilde{\boldsymbol{w}}}{\operatorname{argmin}} \quad \tilde{\boldsymbol{w}}^\intercal \Sigma \tilde{w} \tag{2a}$$

$$\text{s.t.} \quad \tilde{\boldsymbol{w}}^\intercal \tilde{\mu} = \tilde{r}_* \tag{2b}$$

where $\tilde{r}_* = r_* - r_f$. Notice that converting problem 1 into problem 2a only involves subtracting r_f from risky assets' expected return vector $\boldsymbol{\mu}$ to get $\tilde{\boldsymbol{\mu}}$; the covariance matrix $\boldsymbol{\Sigma}$ is the same as problem 1, as r_f is fixed and does not co-vary with the N risky assets.[9]

Optimal Portfolio To solve problem 2a, we set up the Lagrangian $L(\tilde{\boldsymbol{w}}, \lambda) = \tilde{\boldsymbol{w}}^\intercal \boldsymbol{\Sigma} \tilde{\boldsymbol{w}} - \lambda(\tilde{\boldsymbol{w}}^\intercal \tilde{\boldsymbol{\mu}} - \tilde{r}_*)$ and solve the FOC for $\tilde{\boldsymbol{w}}$ to get an expression for the Lagrange parameter $\lambda_* = \frac{2\tilde{\boldsymbol{\mu}}\boldsymbol{\Sigma}\tilde{\boldsymbol{w}}}{\tilde{\boldsymbol{\mu}}^\intercal \tilde{\boldsymbol{\mu}}}$. Substituting λ_* into the Lagrangian, we get a function $L(\tilde{\boldsymbol{w}})$ that we can solve for the portfolio weights vector. Assuming that $\boldsymbol{\Sigma}$ is strictly positive definite, this process

[8] In addition to simplifying derivations, the existence of a risk-free borrowing/lending rate is a realistic assumption, as most investors allocate part of their wealth to risk-free assets. r_f is usually the mean annual return of the 10-year US treasury bill over the period under study.

[9] Note that the portfolio weight for the risk-free asset does not enter the optimal weights vector $\tilde{\boldsymbol{w}}_*$ ($\tilde{w}_{*f} = 1 - \tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N$).

yields optimal weights

$$\tilde{\boldsymbol{w}}_* = \frac{\tilde{r}_*}{\tilde{\boldsymbol{\mu}}^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}} \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}} \tag{3}$$

Note four things in the above expression. First, $\tilde{\boldsymbol{w}}_*$ excludes the weight for the risk-free asset $\tilde{w}_{f*}$, and hence (3) gives the *absolute weights* for the N risky assets. To obtain $\tilde{w}_{f*}$ we rearrange $\tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N = 1 - \tilde{w}_{f*}$ to $\tilde{w}_{f*} = 1 - \tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N$. Clearly, if the sum of absolute weights is lower (greater) than 1, $\tilde{w}_{f*}$ is negative (positive) and investors must borrow (lend) to form their positions on risky assets. The second thing to note is that, to obtain the *relative weights* for risky assets $\tilde{\boldsymbol{w}}_*^R$, we must normalize the absolute weights by their sum, such that $\tilde{\boldsymbol{w}}_*^R = \frac{\tilde{\boldsymbol{w}}_*}{\tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N}$. Relative weights show how the wealth invested in risky assets is distributed among them. Third, note that the relative weights expression can be rewritten as

$$\tilde{\boldsymbol{w}}_*^R = \frac{\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}}{\mathbf{1}_N^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}} \tag{4}$$

whose only arguments are $\boldsymbol{\Sigma}^{-1}$ and $\tilde{\boldsymbol{\mu}}$ and is proportional to $\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}$. Relatedly, the fraction term in (3) is a scalar, which means that absolute weights are also proportional to $\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}$. These related points hint at the importance of correctly estimating the expected returns vector and inverted covariance matrix, a topic to which I return later in this section and in Section 3. The fourth point to note is that the excess return target $\tilde{r}_*$ does not enter $\tilde{\boldsymbol{w}}_*^R$, and thus does not affect the allocation of wealth *within risky assets* (relative weights). Irrespective of $\tilde{r}_*$, all optimal portfolios contain the same proportion of risky asset i vs asset j, and these proportions are determined by $\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}$. What $\tilde{r}_*$ does change is the allocation of wealth *between risk-free and risky assets* (absolute weights). Specifically, the higher the excess returns targeted, the smaller the weight placed on the risk-free asset (more borrowing) and the larger the absolute weights on the risky assets, but the latter always maintain the same proportions. In short, assets' expected return and covariance determine the *proportions* in which wealth is held among risky assets, while the excess return target determines *how* *much* wealth is held in risky assets.

Optimal Return & Risk Since the optimal portfolio satisfies constraint 2b, by definition its return is $\tilde{r}_p = \tilde{r}_*$. In turn, the variance of the optimal portfolio is $\sigma_p^2 = \tilde{\boldsymbol{w}}_*^\intercal \boldsymbol{\Sigma} \tilde{\boldsymbol{w}}_*$, which implies

$$\sigma_p^2 = \frac{\tilde{r}_*^2}{\tilde{\boldsymbol{\mu}}^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}} \iff \sigma_p = (\sqrt{\tilde{\boldsymbol{\mu}}^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}})^{-1} \tilde{r}_* \tag{5}$$

Clearly, optimal portfolio risk is a linear function of target excess return $\tilde{r}_*$. This reveals a core intuition of the MPT framework: all else equal, optimal portfolios must take on higher risk if they target higher returns. Conversely, investors can lower their portfolio risk by targeting lower returns.[10]

Efficient Frontier More importantly, (5) gives the optimal combinations of excess return and risk that are feasible given assets' excess return and covariance, known as the *Efficient Frontier* (EF) (Markowitz, 1952). Because we assume risk-free borrowing and lending, the EF is a straight line given by (5). In my formulation, the EF spans $(\tilde{r}, \sigma)$ space, originates at $(0, 0)$, and has slope $(\sqrt{\tilde{\boldsymbol{\mu}}^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}})^{-1}$.[11] Combinations of excess return and risk below the EF are suboptimal, while combinations above it are unfeasible (given assets' excess return and covariance). Along the EF, if investors target the risk-free rate of return and seek no excess return ($r_* = r_f \Leftrightarrow \tilde{r}_* = 0$), their optimal portfolio consists of investing all wealth in the risk-free asset ($\tilde{w}_{f*} = 1 \Leftrightarrow \tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N = 0$) and takes on zero risk ($\sigma_p = 0$). The higher the excess return target, the further along the EF the optimal portfolio lies. For some excess return target $\tilde{r}_*'$, the optimal portfolio is the reverse of the zero-risk portfolio and consists of investing all wealth in the N risky assets ($\tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N = 1 \Leftrightarrow \tilde{w}_{f*} = 0$). Any excess return target less than $\tilde{r}_*'$ implies that investors lend a portion of their wealth at rate r_f. Conversely, if the

[10] A corollary is that a rise in the risk-free rate will, *ceteris paribus*, decrease portfolio risk, as investors allocate a smaller share of their wealth to risky assets. Mathematically, this is because an increase in r_f implies lower target excess return and thus lower optimal portfolio return.

[11] The standard presentation of the EF has expected returns on the vertical axis and risk on the horizontal axis. Since my optimization problem involves choosing a target portfolio return (1b, 2b), I place returns on the horizontal axis.

excess return target exceeds $\tilde{r}'_*$, investors must borrow at rate r_f to invest more than their total wealth in risky assets ($\tilde{\boldsymbol{w}}_*^\intercal \mathbf{1}_N > 1 \Leftrightarrow \tilde{w}_{f*} < 0$).[12] Once again, it is worth emphasizing that all portfolios along the EF contain the same proportions of risky assets, as determined by assets' returns and covariance.

Sharpe Ratio Another useful quantity in MPT is the *Sharpe Ratio* (SR), defined as the ratio of excess return to risk, the latter measured in standard deviation terms: $SR = \frac{\tilde{r}}{\sigma}$. Rearranging (5) we see that the SR of optimal portfolios is $SR_* = \frac{\tilde{r}_*}{\sigma_p} = \sqrt{\tilde{\boldsymbol{\mu}}^\intercal \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{\mu}}}$, the slope of the EF.[13] I return to SR in Section 4, since it is a common metric for evaluating portfolio performance.

2.2 Issues

Calculating (4) requires estimating assets' expected returns vector and (inverse) covariance matrix.[14] Traditionally, the sample (plug-in) estimators are used: $\hat{\boldsymbol{\mu}}_{sample} \equiv \boldsymbol{m}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{r}_t$ and $\hat{\boldsymbol{\Sigma}}_{sample} \equiv \boldsymbol{S}_T = \frac{1}{T}\sum_{t=1}^{T}[\boldsymbol{r}_t \boldsymbol{r}_t^\intercal] - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\intercal$, where $\boldsymbol{r}_t$ is the N-vector of assets' returns between $t-1$ and t. Asymptotically, these estimators are unbiased. In practice, there is a wealth of evidence that these estimators produce portfolios whose weights are very sensitive to changes in $\boldsymbol{r}_t$.

[12] In the presence of risk-free borrowing and lending the EF is also called the *best Capital Allocation Line*, because it is the line denoting optimal (best) linear combinations (allocations) of capital between the risk-free-portfolio and the risky-assets-only portfolio. Note that borrowing at rate r_f is equivalent to a short position on the risk-free asset. Portfolios that involve borrowing to finance investments on risky assets are called "leveraged".

[13] In more standard formulations of MPT, where return is measured on the vertical axis and risk-free lending/borrowing is unavailable, the EF is not a straight line, but rather the upper portion of a parabola (also known as the Markowitz bullet). In turn, SR_* is the slope of the CAL that originates at r_f and is tangent to the EF. In that formulation, the best CAL is that with the largest slope (largest SR), because it maximizes excess return (vertical distance, difference between portfolio return and r_f) for the risk incurred (horizontal distance).

[14] To simplify notation, henceforth I present portfolio optimization in terms of expected returns unadjusted for r_f (i.e. $\boldsymbol{r}$ and $\boldsymbol{\mu}$ instead of $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{\mu}}$). All issues presented also hold when adjusting for r_f.

2.2.1 Estimating Expected Returns

At least since Merton (1980), the consensus is that estimating $\boldsymbol{\mu}$ is very difficult, more so than estimating $\boldsymbol{\Sigma}^{-1}$. Even if a portion of assets' expected return depends on the market's expected return μ_m, as in most asset price models, and μ_m is constant over time, it would still take a very long time series to estimate $\boldsymbol{\mu}$ accurately (Merton, 1980, p. 326). Of course, the assumption of constant expected market return is unreasonable, and relaxing it makes estimating $\boldsymbol{\mu}$ even harder. Some solutions have been developed to deal with the estimation of $\boldsymbol{\mu}$, which I briefly review in Section 3. Unfortunately, none of these solutions are satisfactory, thus shifting researchers' attention to estimating $\boldsymbol{\Sigma}^{-1}$. For the same reason, researchers have accepted that most of MPT's benefits for portfolio performance come mostly from reducing risk, rather than increasing returns (Jorion, 1985). Indeed, Jagannathan and Ma (2003) claim that "estimation error in the sample mean is so large that nothing much is lost in ignoring the mean altogether" (p. 1652).

2.2.2 Estimating the Covariance/Precision Matrix

The standard approach employs the sample estimator of the covariance matrix and inverts it to obtain $\hat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{S}_T^{-1}$. (In Section 3, I review newer methods for estimating $\boldsymbol{\Sigma}^{-1}$ directly.) Using $\boldsymbol{S}_T^{-1}$ raises four issues. The first is quasi-normative: it assumes that historical values of covariances are useful estimates of their future values, a naive assumption according to some theories of asset prices (Elton and Gruber, 1973). The second obstacle in using $\boldsymbol{S}_T^{-1}$ is that it requires estimating $\frac{N(N+1)}{2}$ parameters, a computationally intensive task. The third issue the use of $\boldsymbol{S}_T^{-1}$ creates is that when $N > T$, $\boldsymbol{S}_T$ is not full rank and is thus non-invertible; this necessitates the use of a pseudo-inverse. Moreover, even when $\boldsymbol{S}_T$ is full rank, $\boldsymbol{S}_T^{-1}$ is a biased estimate of $\boldsymbol{\Sigma}^{-1}$ (Senneret et al., 2016). Finally, he largest limitation of using $\boldsymbol{S}_T^{-1}$ is that, in practice, it generates unstable and/or extreme portfolio weights, which are far from optimal and are not reasonable representations of investors' behavior.

Scholars take two approaches to interpreting the instability of portfolio weights constructured through $\boldsymbol{S}_T^{-1}$. One approach treats it as the result of estimation error from $\boldsymbol{S}_T$ passing on to $\boldsymbol{S}_T^{-1}$ and the portfolio weights. DeMiguel, Garlappi and Uppal (2009) find that for portfolios constructed through $\boldsymbol{S}_T^{-1}$ to outperform the naive $1/N$ portfolio, where wealth is allocated equally across the N risky assets, more than $6,000$ months (50 years) of data are required, even when N is only 50.[15] The authors attribute the Markowitz portfolio's underperformance to small estimation errors from $\boldsymbol{S}_T$ swinging portfolio weights to extreme values. Yet, under this approach, using $\boldsymbol{S}_T^{-1}$ *can* produce stable and sensible portfolio weights; if T is sufficiently large, given N, to remove substantial estimation error in $\boldsymbol{S}_T$. On the contrary, a different approach attributes the unstable weights that $\boldsymbol{S}_T^{-1}$ generates to ill-conditioning in the *true* (population) covariance matrix $\boldsymbol{\Sigma}$, owing to the inherently high correlation in asset returns. Brodie et al. (2009) argue that, even if $\boldsymbol{\Sigma}$ were known and $\boldsymbol{S}_T$ were redundant, inverting a large-N $\boldsymbol{\Sigma}$ would produce the same unstable weights that $\boldsymbol{S}_T^{-1}$ produces—even errors as small as those from rounding would be amplified enough to destabilize the inversion of $\boldsymbol{\Sigma}$. This means that, stable and optimal portfolio selection is impossible for large N, independent of T, due to the highly correlated nature of asset returns. Regardless of the cause of unstable weights in the standard MPT solution, it is clear that alternative approaches are needed to salvage MPT's contribution to portfolio selection.

2.2.3 Instability Example

Before proceeding, it is worth illustrating the instability of the standard MPT solution, using a simple example augmented from DeMiguel, Garlappi and Uppal (2009). Assume two assets with annual returns of the same mean and standard deviation, 0.08 and 0.20, respectively, and a correlation of 0.99. The standard MPT solution consists of relative weights $w_{*1} = w_{*2} = 0.5$. Now assume that asset 1's mean return is unknown and estimated at 0.09, instead of its true value of 0.08. The standard MPT solution now yields $w_{*1} = 6.35$ and

[15] In reality, investors face a much larger universe of assets (typically several thousands) and use a much shorter time window to optimize their portfolio (typically 60–120 months), thereby multiplying the estimation error in $\boldsymbol{S}_T$ and the suboptimality of the respective weights.

$w_{*2} = -5.35$. Note that we obtain a completely different solution only because estimation error shows that one asset's mean return is slightly superior. A similar result obtains if we instead decrease asset 1's risk to 0.175, keeping both assets' mean returns at 0.08, asset 2's risk at 0.20, and correlation at 0.99: now $w_{*1} = 4.04$ and $w_{*2} = -3.03$. This example demonstrates the sensitivity of the standard MPT solution to estimates of assets' expected return and covariance: the algorithm attempts to exploit the slightest difference in assets' returns and/or risk in optimizing the objective function. It is for this reason that Michaud (1989) labels the standard MPT solution an "estimation-error maximizer" (p. 33).

3 An Overview of Solutions

There are three broad approaches to improving MPT. The first involves reducing estimation error for expected returns, with robust estimators and shrinkage estimators being the two main tools to achieve that. The second approach for creating portfolios with stable and sensible weights is to reduce estimation error in the (inverse) covariance matrix. This is a more popular approach with more tools available to achieve it, including not just robust estimators and shrinkage estimators, but also covariance matrices based on statistical models, sparse graphical models, random matrix theory, and high-frequency data. The third approach for stabilizing portfolio optimization targets portfolio weights directly, through robust optimization techniques, explicit restrictions on the weights, or restrictions on the weights' norm. This section briefly reviews each approach.

3.1 Robust Estimation of Expected Returns

3.1.1 Robust Estimators

As mentioned in Section 2, estimating $\boldsymbol{\mu}$ is prone to more error than estimating $\boldsymbol{\Sigma}^{-1}$, forcing researchers to focus on the latter task. That said, researchers have explored two ways to improve $\hat{\boldsymbol{\mu}}$. First, instead of using the (sample) mean $\boldsymbol{m} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{r}_t$ to estimate (population) expected returns, we can use the (sample) truncated/trimmed mean or Winsorized mean (Martin, Clark and Green, 2010). The trimmed mean is calculated after *discarding* the $k\%$ most extreme values, while the Winsorized mean is calculated after *replacing* those values with the next $k\%$ most extreme values (e.g. 1st and 10th decile replaced by 2nd and 10th, respectively). Both estimators are more robust to deviations from distributional assumptions on returns (e.g. multivariate normality). However, if returns are non-stationary, neither estimator, nor other robust estimators like the M-estimator, can provide unbiased estimates of expected returns. Additionally, robust estimators require researchers to make subjective decisions about tuning parameters like k.

3.1.2 Shrinkage Estimators

To improve performance in the face of non-stationary returns and to automate the selection of tuning parameters, researchers have turned to *Shrinkage Estimators*. The shrinkage approach begins with the observation that, much like uncertainty in actual expected returns (asset risk), uncertainty in the *estimate* of expected returns (estimation risk) implies a loss of investor utility (Jorion, 1985). Thus, the optimization problem should minimize utility loss from selecting a portfolio based on sample estimates, instead of true values. In this line of thought, the solution is clearly not to estimate each asset's expected return individually, but to select an estimator that minimizes utility loss from aggregate parameter uncertainty.

Jorion (1986) suggests using a Bayes–Stein estimator that shrinks each asset's sample mean m_i to the grand mean $\bar{m} = \frac{1}{N}\sum_i^N m_i$. The optimal amount of shrinkage is determined by the data; namely, N, T, $\bar{m}$, $\boldsymbol{m}_T$, and $\boldsymbol{S}_T^{-1}$.[16] Simulations show that the shrinkage estimator substantially reduces estimation risk, defined as the average loss of utility across repeated

[16] In particular, Jorion (1986) uses an *empirical* Bayes–Stein (shrinkage) estimator, to avoid using theory-driven priors in determining the optimal amount of shrinkage θ_*. In particular, $\theta_* = \frac{N+2}{(N+2)+(\boldsymbol{m}_T - m_0\boldsymbol{1}_N)^\intercal T \boldsymbol{S}_T^{-1}(\boldsymbol{m}_T - m_0\boldsymbol{1}_N)}$, where $m_0 = \frac{\boldsymbol{1}_N^\intercal \boldsymbol{S}_T^{-1}}{\boldsymbol{1}_N^\intercal \boldsymbol{S}_T^{-1}\boldsymbol{1}_N}\boldsymbol{m}_T$. This yields the shrunken estimated expected returns vector $\hat{\boldsymbol{\mu}}_{shrink} = (1-\theta_*)\boldsymbol{m}_T + \theta_*\boldsymbol{1}_N m_0$. As an alternative to tuning θ using data, Ruppert and Matteson (2015) suggest a bootstrap approach (pp. 482–484).

samples, and outperforms portfolios constructed using $\boldsymbol{m}_T$ (Jorion, 1986).[17] The author concludes that the standard MPT approach vastly overestimates potential portfolio gains from overweighting assets with high (estimated) expected returns, and argues that most potential MPT gains come from reducing portfolio risk. This points to the importance of correctly estimating the (inverted) covariance matrix.

3.2 Robust estimation of (inverse) covariance matrix

3.2.1 Robust Estimators

As mentioned above, there are several alternatives to $\boldsymbol{S}_T^{-1}$ for estimating the (inverse) covariance matrix. One approach is to use so-called *Robust Estimators* of $\boldsymbol{\Sigma}^{-1}$. Much like those for $\boldsymbol{\mu}$, these estimators are less sensitive to extreme observations of assets' co-movement. DeMiguel and Nogales (2009) employ M- and S-estimators, which exhibit lower estimation error than $\boldsymbol{S}_T$ when the sample distribution of the underlying variables belongs to the neighborhood of the assumed distribution (e.g. multivariate normal).[18] Pfaff (2016) (Ch. 11) reviews several alternatives, like the minimum volume ellipsoid estimator, the minimum covariance determinant estimator, and the orthogonalized Gnanadesikan–Ketterning (OGK) estimator. Note that all these are estimators of $\boldsymbol{\Sigma}$ and not its inverse, thus requiring us to invert the estimated matrices in order to use them as inputs to portfolio selection.

3.2.2 Model-Based: Single-Index

Statistical models of asset returns and covariances have been a popular tool in empirical finance since the Nobel prize-winning work of Sharpe (1963, 1964). The latter developed

[17] Merton (1980) also uses a Bayesian approach to augment the estimator of expected returns, but focuses on market return and does not rely on a shrinkage estimator. Specifically, he uses prior information from the Capital Asset Pricing Model (CAPM) to disaggregate expected market return into an expected risk-free rate and an expected excess return rate. By allowing the latter to vary over time, Merton accounts for non-stationarity of market returns.

[18] DeMiguel and Nogales (2009) go a step further than just using a robust estimator of $\boldsymbol{S}_T$, and integrate robust estimation of portfolio inputs with portfolio optimization, an approach they label "robust portfolio estimation". Specifically, the authors solve a single non-linear optimization problem, selecting weights that minimize an M- or S-estimate of portfolio risk.

factor models; linear models of asset returns that include anywhere from one to dozens of predictors (factors). The simplest factor model is the *Single-Index Model* (SIM), or one-factor model or market model, which regresses stocks' excess return on the excess return of the market index (a capitalization-weighted portfolio of all stocks) (Sharpe, 1963).[19] That is, the SIM specifies the excess return for stock i as $\tilde{r}_{i,t} = \alpha_i + \beta_i \tilde{r}_{m,t} + \epsilon_{i,t}$, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where $\tilde{r}_m$ is excess return on the market index, α_i and β_i are parameters to be estimated, and $\epsilon_{i,t}$ is random error, with $\mathbb{E}[\epsilon_i] = 0$, $\text{Var}[\epsilon_i] = \sigma^2_{\epsilon_i}$, $\text{Cov}[\epsilon_i, \epsilon_j] = 0 \; \forall j \neq i$, and $\text{Cov}[\tilde{r}_m, \epsilon_i] = 0$.

Due to the theoretical structure that the SIM imposes on stock returns, it produces a simple covariance matrix. Note that $\text{Var}[\tilde{r}_i] = \sigma^2_i = \beta^2_i \sigma^2_m + \sigma^2_{\epsilon_i}$ and $\text{Cov}[\tilde{r}_i, \tilde{r}_j] = \sigma_{ij} = \beta_i \beta_j \sigma^2_m$, where $\sigma^2_m = \text{Var}[\tilde{r}_m]$. Thus, the covariance matrix is $\boldsymbol{\Sigma}_{SIM} = \boldsymbol{\beta}\boldsymbol{\beta}^{\intercal}\sigma^2_m + \boldsymbol{\Sigma}_\epsilon$, where $\boldsymbol{\beta}$ is the length-N vector of β's from the SIM regressions and $\boldsymbol{\Sigma}_\epsilon$ is the regression errors' N-by-N diagonal covariance matrix. In turn, the precision matrix becomes $\boldsymbol{\Sigma}^{-1}_{SIM} = \boldsymbol{\Sigma}^{-1}_\epsilon - \frac{\boldsymbol{\Sigma}^{-1}_\epsilon \boldsymbol{\beta}\boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}_\epsilon}{\frac{1}{\sigma^2_m} + \boldsymbol{\beta}^{\intercal}\boldsymbol{\Sigma}^{-1}_\epsilon \boldsymbol{\beta}}$. Of course, the true SIM parameters are unknown, thus we replace $\boldsymbol{\beta}$, σ^2_m, and $\boldsymbol{\Sigma}_\epsilon$ with their sample/regression estimates. A finite sample bias-correction is also required, thus the estimated covariance matrix becomes $\hat{\boldsymbol{\Sigma}}_{SIM} = \hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\intercal}\hat{\sigma}^2_m + \frac{T-1}{T-2}\hat{\boldsymbol{\Sigma}}_\epsilon$, where $\hat{\boldsymbol{\beta}}$ contains the OLS estimates of the slope parameter, $\hat{\sigma}^2_m$ is the bias-corrected sample estimate of the market index's variance, and $\hat{\boldsymbol{\Sigma}}_\epsilon$ diagonal contains the OLS estimates of the residual variances $\hat{\sigma}^2_{\epsilon_i}$.[20] (Similar corrections apply to the inverse covariance matrix estimator.)

In addition to providing a model-based view of asset returns – a property that some find normatively attractive – the SIM has three other advantages over the standard MPT approach. First, SIM requires estimating only $2N+1$ parameters to construct the covariance matrix (N $\hat{\beta}_i$'s, N $\hat{\sigma}_{\epsilon_i}$'s, and $\hat{\sigma}^2_m$), versus the $\frac{N(N+1)}{2}$ parameters required by the standard approach (exceeds $2N+1$ for $N \geq 4$). Second, whenever we add a new asset to our sample, we only need to estimate its β and σ_ϵ to update our precision matrix, versus estimating the

[19] In practice, the market index is typically proxied by the S&P 500 or some similar index.

[20] This expression rests on the usual OLS property that the estimators of slope and residual variance ($\hat{\beta}_i$ and $\hat{\sigma}_{\epsilon_i}$) are independent.

asset's variance and covariance with every other asset in our sample ($N + 1$ parameters) and inverting our new covariance matrix to update our precision matrix. Finally, the SIM approach only requires $T > 2$ observations to estimate β and σ_ϵ for every asset (as well as σ_m^2) and thus to estimate the precision matrix. On the contrary, the standard approach requires $T > N$.

That said, these are merely theoretical requirements for statistical identification. In practice, as noted in Section 2, the standard approach requires $T \gg N$ to estimate $\boldsymbol{\Sigma}$ with reasonably low error, while fitting SIM regressions with few degrees of freedom result in very noise estimates of β's and σ_ϵ's, thereby adding noise to $\hat{\boldsymbol{\Sigma}}_{SIM}^{-1}$. Moreover, in the portfolio selection stage, both the standard and SIM approach to portfolio selection are vulnerable to estimation error from using the sample mean returns vector $\boldsymbol{m}_T$ (SIM is also exposed to estimation error from $\hat{\sigma}_m^2$).[21] Nevertheless, there is ample evidence that the SIM approach produces portfolios that are much less sensitive to estimation error than the standard approach and, therefore, perform better across a range of risk and return metrics (Senneret et al., 2016).

3.2.3 Model-Based: Adjusting β's

A popular technique for reducing noise in $\hat{\boldsymbol{\Sigma}}_{SIM}^{-1}$ is to use *Adjusted* β estimates. Blume (1975) suggests a three-step approach: first, fit SIMs for N stocks using returns from two adjacent periods of equal length (call the β estimates from these periods $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\beta}}_{t-1}$); second, regress $\hat{\boldsymbol{\beta}}_t$ on $\hat{\boldsymbol{\beta}}_{t-1}$ and obtain coefficients $\hat{\gamma}_0$ and $\hat{\gamma}_1$ capturing the temporal relationship between β's; third, insert $\hat{\boldsymbol{\beta}}_t$ in the beta-adjustment equation to predict the next period's β's ($\hat{\boldsymbol{\beta}}_{t+1} = \hat{\gamma}_0 + \hat{\gamma}_1\hat{\boldsymbol{\beta}}_t$). In practice, adjusted β's are higher (lower) for stocks with lower (higher) unadjusted β's. Though not an explicit goal of Blume (1975), adjusted β's can be combined with the usual ingredients ($\hat{\sigma}_m^2$ and $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1}$) to produce a beta-adjusted $\hat{\boldsymbol{\Sigma}}_{SIM}^{-1}$ for

[21] In practice, many researchers use a ranking approach with a cut-off rule to select the optimal portfolio, which requires the same parameters as the standard approach, but does not require inverting $\hat{\boldsymbol{\Sigma}}_{SIM}$ (Elton et al., 2009, Ch. 9). This approach yields the same portfolio as the standard approach ($\boldsymbol{w}_* = \hat{\boldsymbol{\Sigma}}_{SIM}^{-1}\boldsymbol{m}_T$), while simplifying computation furthermore.

portfolio selection. One limitation of this approach is that adjusted $\hat{\beta}$'s drift upwards over time: if the average $\hat{\beta}$ increases between $t-1$ and t, then predicting $\hat{\boldsymbol{\beta}}_{t+1}$ through $\hat{\gamma}_0 + \hat{\gamma}_1 \hat{\boldsymbol{\beta}}_t$ will invariably produce a larger estimate (Elton et al., 2009).

An alternative, Bayesian technique for adjusting β's is provided by Vasicek (1973): to shrink individual stocks' β estimates towards the average β estimate in proportion to each estimate's precision. In particular, Vasicek (1973) suggests the formula $\hat{\beta}_{i,t+1} = \frac{\hat{\sigma}^2_{\beta_{i,t}}}{\hat{\sigma}^2_{\bar{\beta}_t} + \hat{\sigma}^2_{\beta_{i,t}}} \bar{\beta}_t + \frac{\hat{\sigma}^2_{\bar{\beta}_t}}{\hat{\sigma}^2_{\bar{\beta}_t} + \hat{\sigma}^2_{\beta_{i,t}}} \hat{\beta}_{i,t} \ \ \forall i, t$, where $\hat{\beta}_{i,t}$ is the unadjusted β for stock i in period t, $\hat{\sigma}_{\beta_{i,t}}$ is its standard error, $\bar{\beta}_t$ is the mean unadjusted β across N stocks for period t, and $\hat{\sigma}_{\bar{\beta}_t}$ is its standard error. As the formula suggests, the more noisily a stock's β_i is estimated, the more it is shrunk towards $\bar{\beta}$. Unfortunately, this also induces a downward drift in the average adjusted β over time. Because β_i's are shrunk towards $\bar{\beta}$ in proportion to their standard error, larger β_i's are invariably shrunk by a larger percentage of their distance from $\bar{\beta}$ than small β_i's are raised. Thus, the average (adjusted) $\hat{\beta}_{i,t+1}$ will invariably be lower than the average (unadjusted) $\hat{\beta}_{i,t}$.

Despite their biases, though, both Blume's and Vasicek's adjustments result in more accurate forecasts of future β's versus the unadjusted historical β's (Klemkosky and Martin, 1975). Moreover, both types of adjusted β's produce covariance matrices that are significantly better forecasts of future periods' covariance matrices than the sample (historical) covariance matrix (Elton, Gruber and Urich, 1978). In short, in addition to simplifying computation for optimal portfolio selection, the SIM – with or without β-adjustment – also reduces estimation error in a key input of MPT, the inverse covariance matrix.

3.2.4 Model-Based: Multi-Index

Recalling that the SIM only includes one factor, the market return, it is possible to add more factors to account for non-market-related covariance in asset returns. *Multi-index models* (MIMs) are motivated by the observation that even after removing asset covariance owed to the market index, substantial covariance remains among stocks belonging to the same industry (King, 1966). MIMs come in several forms, but all assume that stocks' excess return is a linear function of the excess return on the respective industry index, which itself is a linear function of the excess return on the market index.

In the formulation of Cohen and Pogue (1967), the excess return of stock i as $\tilde{r}_{i,t} = \alpha_i + \beta_i \tilde{I}_{j,t} + \epsilon_{i,t}$, where $\tilde{I}_{j,t}$ is the excess return of industry j (all other variables are interpreted as in the SIM and the SIM's assumptions about $\epsilon_{i,t}$ carry over). In turn, the excess return on industries is modeled as $\tilde{I}_{j,t} = a_j + b_j \tilde{r}_{m,t} + e_{j,t}$, where a_j and b_j are parameters to be estimated and $e_{j,t}$ are errors with $\mathbb{E}[e_j] = 0$, $\text{Var}[e_j] = \sigma^2_{e_j}$, $\text{Cov}[e_j, e_{j'}] = 0 \quad \forall j' \neq j$, and $\text{Cov}[\epsilon_i, e_j] = 0$.[22] As in the SIM, MIMs produce a model-based variance-covariance structure: $\sigma_i^2 = \beta_i^2(b_j^2\sigma_m^2 + \sigma^2_{e_j}) + \sigma^2_{\epsilon_i}$, $\sigma_{ik} = \beta_i\beta_k(b_j^2\sigma_m^2 + \sigma^2_{e_j})$ for i and k are in the same industry j, and $\sigma_{ik} = \beta_i\beta_k b_j b_l \sigma_m^2$ for i in industry j and k in industry l. The resulting covariance matrix is $\boldsymbol{\Sigma}_{MIM} = \boldsymbol{\beta}\boldsymbol{\beta}^\intercal \odot (\boldsymbol{b}\boldsymbol{b}^\intercal \sigma_m^2 + \boldsymbol{\Sigma}_e) + \boldsymbol{\Sigma}_\epsilon$, where $\odot$ is the element-wise product, $\boldsymbol{b}$ is the length-N vector of b_j's from the industry regressions, $\boldsymbol{\Sigma}_e$ is the block-diagonal N-by-N covariance matrix of errors from the industry regressions, and other parameters are defined as in the SIM.[23] Once again, true values are unknown, so we substitute them with regression/sample/historical estimates to obtain the covariance matrix.[24]

In terms of portfolio performance, MIMs have produced relatively weak results: though they estimate the historical covariance matrix more accurately than the SIM, MIMs do not forecast future covariance matrices more accurately and, consequently, produce portfolios

[22]MIMs can be adapted to include many, uncorrelated industry indexes on the right-hand-side (the indexes can be orthogonalized through a technique like principal components analysis). Though this is useful for stocks of companies that trascend industry boundaries (e.g. Tesla, Inc. is viewed as an auto manufacturer, battery producer, and software company), for most stocks the addition of industry indexes might add more noise than information to the regressions (Elton et al., 2009, p. 159).

[23]Before expressing the covariance matrix in this form, we must group stocks by industry and order them to reflect that grouping in all vectors/matrices. Note that in $\boldsymbol{b}$ all elements corresponding to stocks in the same industry have the same value (e.g. b_1 for industry 1, b_2 for industry 2, etc.). Similarly, within each block in $\boldsymbol{\Sigma}_e$ all elements have the same value (i.e. σ_{e_1} for block/industry 1, σ_{e_2} for block/industry 2, etc.), but blocks do not have to be of the same size (e.g. 5 stocks might belong in industry 1, 7 stocks in industry 2, etc.).

[24]Note that the MIM requires $2N + 2J + 1$ inputs, where J is the number of industries, versus the $2N + 1$ parameters of the SIM. I omit the full covariance and precision matrix expressions using sample estimates to conserve space. Moreover, more so than in the SIM, most practicioners resort to the ranking and cut-off rule approach to select optimal portfolios, instead of the standard approach ($\boldsymbol{w}_* = \hat{\boldsymbol{\Sigma}}^{-1}_{MIM}\boldsymbol{m}_T$) (Elton, Gruber and Padberg, 1977).

with inferior performance (Elton et al., 2009, pp. 159–162). Moreover, Chan, Karceski and Lakonishok (1999) find models with no more than three indexes outperform those with additional ones. In short, adding complexity to the SIM seems to import more noise than information, not least because it is hard to attribute covariance among stocks to factors other than the market.

3.2.5 Model-Based: Averaging Models

In contrast to index models (single or multi-index) stand models that try to reduce the noise inherent in stocks' covariance by averaging the latter. The simplest model is the *Constant Correlation Model* (CCM), which assumes that all stocks have the same correlation, equal to the sample (historical) mean correlation ($\hat{\rho}_T = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{k=i+1}^{N} \hat{\rho}_{ik}$) (Elton and Gruber, 1973). The resulting covariance is $\sigma_{ik} = \hat{\rho}_T \sigma_i \sigma_k \;\; \forall i, k$, thus the covariance matrix is $\mathbf{\Sigma}_{CCM} = \mathbf{P} \odot \boldsymbol{\sigma}\boldsymbol{\sigma}^{\intercal}$, where $\mathbf{P}$ is a N-by-N matrix with 1's on its diagonal and $\hat{\rho}_T$ as its off-diagonal elements and $\boldsymbol{\sigma}$ is the N-vector of assets' standard deviation. To form the optimal portfolio, the CCM has to be combined with the ranking and cut-off rule approach (Elton, Gruber and Padberg, 1976). Crucially, the CCM implicitly assumes that the historical correlation matrix only contains information about the the *average* correlation for future periods, but no information about pairwise correlations' deviation from that average. Clearly, this is a strong assumption that can be relaxed.

One way to do so is by allowing average pairwise correlations to differ by industry and also between industries, an approach known as the *Multi-Group Model* (MGM). In particular, the MGM assumes that all stocks within the same industry have the same correlation, estimated as the mean sample pairwise correlation between stocks in that industry ($\hat{\rho}_{jj} = \frac{2}{N_j(N_j-1)} \sum_{i=1}^{N_j-1} \sum_{k=i+1}^{N_j} \hat{\rho}_{ik}$, $j = 1, \ldots, J$, where N_j is the number of stocks in industry j), while stocks in different industries have a correlation equal to the mean sample pairwise correlation between stocks in those industries ($\hat{\rho}_{jl} = \frac{\sum_i^{N_j} \sum_k^{N_l} \hat{\rho}_{ik}}{N_j(N_l-1)}$, $j \neq l$). Again, the MGM is usually solved through a ranking and cut-off rule approach (I omit the complicated expression of the full covariance matrix, $\mathbf{\Sigma}_{MGM}$) (Elton, Gruber and Padberg, 1977).

In terms of performance, Elton et al. (2009, pp. 162–163) summarize the literature's findings: that both $\mathbf{\Sigma}_{CCM}$ and $\mathbf{\Sigma}_{MGM}$ are more accurate forecasts of future covariance matrices than $\boldsymbol{S}_T$, $\mathbf{\Sigma}_{MIM}$, and, more importantly, $\mathbf{\Sigma}_{SIM}$. Relatedly, both averaging models produce portfolios with superior performance than competing models, across a range of metrics. However, when comparing CCM to MGM, the authors find that relative forecasting accuracy and portfolio performance vary by the time period used to evaluate them. In any case, among model-based covariance matrices, it seems that averaging models are dominant.

3.2.6 Shrinkage Estimators

Another approach, which carries over from the robust estimation of $\boldsymbol{\mu}$ but is often combined with model-based estimates of $\mathbf{\Sigma}$, is *Shrinkage*. Again, the rationale behind shrinkage is to reduce the estimator's variance – and the utility loss stemming from estimation error – by shrinking it towards an (invariant) constant. This approach becomes more suitable (i) the less the bias introduced by the shrinkage target matrix, (ii) the larger the noise in the data, and (iii) the larger N/T. The first two conditions also apply to shrinking $\boldsymbol{\mu}$; the third one is much more relevant for shrinking $\mathbf{\Sigma}$, where each new asset adds N new terms to $\mathbf{\Sigma}$ that we have to (noisily) estimate. In general, the idea is to estimate $\hat{\mathbf{\Sigma}}_{shrink} = (1-\delta_*)\boldsymbol{S}_T + \delta_* \hat{\mathbf{\Sigma}}_{target}$, where δ_* is determined by a data-driven algorithm and $\hat{\mathbf{\Sigma}}_{target}$ is chosen through some prior belief about asset return covariance. The literature has explored several target matrices, most stemming from the factor and index models reviewed above. Crucially, due to the nature of these models, $\hat{\mathbf{\Sigma}}_{target}$ is usually estimated with much less error than $\boldsymbol{S}_T$, but factor models are arguably mis-specified and hence $\hat{\mathbf{\Sigma}}_{target}$ is biased. As such, shrinkage attempts to strike a balance between the biased but relatively precisely estimated $\hat{\mathbf{\Sigma}}_{target}$ and the unbiased but noisily estimated $\boldsymbol{S}_T$.

In terms of shrinkage target, Ledoit and Wolf (2003*b*) suggest $\hat{\mathbf{\Sigma}}_{SIM}$, while Ledoit and Wolf (2003*a*) recommend $\hat{\mathbf{\Sigma}}_{CCM}$ (both studies derive formulas for selecting δ_*). Crucially, Ledoit and Wolf (2003*a*) find that shrinkage towards the CCM covariance matrix outperforms both the standard and CCM approach, and Disatnik and Benninga (2007) find that simpler shrinkage targets produce portfolios that perform best. However, newer work has focused on non-linear shrinkage estimators (see Senneret et al. (2016, pp. 3–4) for a review). Note that most studies focus on shrinking $\mathbf{\Sigma}$, then inverting it for portfolio selection. Recently, researchers have considered directly shrinking $\mathbf{\Sigma}^{-1}$ (Kourtis, Dotsis and Markellos, 2012; DeMiguel, Martin-Utrera and Nogales, 2013).

3.2.7 Sparse Graphical Models

Another model-based approach for estimating the asset covariance matrix involves *Sparse Graphical Models* (SGMs).[25] SGMs model random variables (stock returns) as nodes and the dependencies between them as edges; the ensuing graph gives us a visual representation of the joint distribution of stock returns. Like shrinkage estimators, from a Bayesian perspective, SGMs implicitly use prior information on the covariance matrix's structure. In the case of stocks, this prior information is usually employed to impose sparsity on the covariance or precision matrix. (A sparse matrix is one with few non-zero off-diagonal elements.) Sparsity is motivated from both a theoretical and computational perspective.

Theoretically, the preference for sparsity depends on whether our prior information relates to the covariance or precision matrix—a sparse covariance matrix will yield a dense covariance matrix, and vice-versa. A sparse $\mathbf{\Sigma}$ implies that most stocks' returns are independent, an assumption that might only be valid for stocks in different industries and/or markets. Thus, a more reasonable structure for $\mathbf{\Sigma}$ is block-sparsity: zero covariance for stocks in different industries/markets (grouped as blocks in the matrix) and non-zero covariance for stocks in the same industry/market. A sparse $\mathbf{\Sigma}^{-1}$, on the other hand, implies that most stocks are *conditionally* independent. One way to justify this prior is by assuming that stock returns are

[25] Also referred to as Probabilistic Graphical Models. When edges between nodes are undirected, as in the case of modeling the asset covariance matrix, graphical models are also known as Markov Graphs, Random Networks, or Random Fields.

linearly related, such that knowing a subset of stocks makes remaining stocks (conditionally) independent. (For both $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ matrix, more (conditional) independencies between stocks result in a sparser matrix.) Another theoretical motivation for sparsity, which applies to both $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$, is a preference for parsimony. *À la* Occam's Razor, investors might seek the least complex and most compressed representation of $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-1}$, thereby penalizing matrices with many non-zero entries (more in Section 3.3.3).

In terms of computation, sparse matrices are easier to store and manipulate, as they have fewer non-zero entries. More importantly for the purposes of portfolio selection, for both $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$, sparsity can reduce estimation error. Because small estimated non-zero off-diagonal elements might be owed to noise, we might want to simply nullify them and instead focus on estimating large elements.[26] The key to reducing estimation error through sparse graphical models is deciding how many and which elements of $\mathbf{\Sigma}$ or $\mathbf{\Sigma}^{-1}$ to set to zero. To make this decision in a data-driven way, researchers turn to algorithms that balance model accuracy and complexity.

The state-of-the-art algorithm for sparse estimation of covariance/precision matrixes is the *Graphical LASSO* (GLASSO) (Friedman, Hastie and Tibshirani, 2008). The model assumes that variables (stock returns) are distributed Multivariate Normal and maximizes the log-likelihood of the covariance/precision matrix, penalized by the ℓ_1-norm ($||\boldsymbol{X}||_1 = \sum_i^N \sum_j^N |X_{ij}|$) and adjusted by a tuning parameter τ, to be selected from the data.[27] That is, GLASSO's estimated covariance matrix is $\hat{\mathbf{\Sigma}}_{GLASSO} = \text{argmin}_{\mathbf{\Sigma}} \log(\det[\mathbf{\Sigma}]) - \text{Tr}(\boldsymbol{S}_T\mathbf{\Sigma}^{-1}) - \tau||\mathbf{\Sigma}||_1$. However, this results in a non-convex problem, which is very difficult to solve, hence there is little work on sparse estimation of $\mathbf{\Sigma}$ for portfolio selection (Fan, Liao and Mincheva, 2011, 2013). On the contrary, more studies have applied GLASSO to estimating $\mathbf{\Sigma}^{-1}$ (Awoye, 2016; Goto and Xu, 2015; Senneret et al., 2016). Defining $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, GLASSO's

[26] This is known as a *screening effect/rule*, and a similar rationale underlies penalized regressions that induce sparsity in the regression coefficient vector, like the LASSO and its variants (Tibshirani, Wainwright and Hastie, 2015).

[27] Note that ℓ_1-norm penalization, in fact, is used to approximate ℓ_0-norm penalization, which directly penalizes the number of non-zero elements but produces a non-convex optimization problem.

estimated precision matrix is $\hat{\mathbf{\Theta}}_{GLASSO} = \text{argmin}_{\mathbf{\Theta}} \log(\det[\mathbf{\Theta}]) - \text{Tr}(\boldsymbol{S}_T\mathbf{\Theta}) - \tau||\mathbf{\Theta}||_1$. This problem is typically solved through the Pathwise Coordinate Descent algorithm (Bien and Tibshirani, 2011). In terms of performance, Awoye (2016) finds that portfolios selected by $\hat{\mathbf{\Theta}}_{GLASSO}$ outperform those selected through the SIM, shrinkage, and RMT (see next paragraph) approaches across a range of metrics, N/T ratios, and datasets on asset returns.[28] Similar findings for GLASSO are reported in Senneret et al. (2016) versus several factor and shrinkage models.

3.2.8 Random Matrix Theory

A different model-based approach to covariance matrix estimation stems from econophysics and is known as *Random Matrix Theory* (RMT). Based on the spectral properties of the asset covariance matrix, RMT's application to MPT begins with the observation that $\mathbf{\Sigma}$'s smallest and largest eigenvalues reflect minimum and maximum portfolio risk, respectively, and that $\mathbf{\Sigma}$'s eigenvectors determine optimal portfolio weights.[29] Crucially, Laloux et al. (1999) show that the smallest eigenvalue is particularly sensitive to estimation error; thus, using the latter results in portfolios with significantly underestimated risk. Relatedly, the authors note that the density of eigenvalues and the structure of eigenvectors of the sample correlation matrix $\boldsymbol{C}$ applied to standard portfolios of stocks (e.g. S&P 500) is statistically indistinguishable from that of a purely random matrix. For this reason, Laloux et al. (2000) develop an approach for filtering $\boldsymbol{C}$ to distinguish between eigenvalues that contain real information from those containing noise.[30] Specifically, the authors compare its eigenvalues to those of a purely random "null hypothesis" matrix $\boldsymbol{M}$ (dimensions $T \times N$). The latter follow a Marchenko-Pastur distribution with density $f(\lambda) = \frac{T}{2N\pi\lambda\sigma^2}\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{\min})}$,

[28] Note that, due to the sparsity induced via the ℓ_1 penalty, GLASSO is able to estimate $\mathbf{\Theta}$ even when $N < T$.

[29] Senneret et al. (2016) note that RMT implicitly extends factor models, by accounting for the fact that the actual number of factors that determine asset returns and covariance is unknown—for this reason, RMT applied to MPT is also known as the "latent factors approach" (p. 9).

[30] A key goal of filtering $\boldsymbol{C}$ is to eliminate very small eigenvalues, which are most likely products of estimation error. This is because the corresponding portfolios display artificially low risk – or even none – with positive expected returns, thereby "tricking" MPT optimization into selecting them.

where σ^2 is the variance of $\boldsymbol{M}$ and $\lambda_{min}^{max} = \sigma^2(1+\frac{N}{T}\pm 2\sqrt{\frac{N}{T}})$. All $\lambda > \lambda_{max}$ in $\boldsymbol{C}$ are retained, while $\lambda < \lambda_{max}$ are deleted or replaced with the average λ above λ_{max}. In short, RMT treats deviations from $\boldsymbol{M}$ as hints of valuable information in the respective eigenvalues of $\boldsymbol{C}$, and uses the corresponding eigenvectors to select the optimal portfolio (after converting $\boldsymbol{C}$ to a covariance matrix). However, El Karoui (2008) shows that, unless T/N is very large, identifying the significant eigenvectors of $\boldsymbol{\Sigma}$ through RMT is very difficult. Nevertheless, in terms of performance, Pantaleo et al. (2011) find that RMT estimators produce portfolios with comparable performance to shrinkage estimators.

3.2.9 High-Frequency Data

An entirely different approach involves using *High-Frequency Data* on assets' returns to estimate their covariance. The rationale behind this approach is intuitive: more granular data increases T/N, thereby potentially decreasing error in $\boldsymbol{S}_T$. The downside of this approach is that, due to the noisiness of asset returns, higher-dimension data can add more noise than information to covariance estimation. Additionally, as we increase the data's frequency, returns tend to deviate more from the normal distribution, thereby burdening the estimation of expected returns. Perhaps for this reason, Jagannathan and Ma (2003) find that portfolios constructed through covariance estimation of daily versus monthly returns do not differ much. Similarly, Liu (2009) compares portfolios constructed by estimating $\boldsymbol{S}_T$ on data of different frequencies and finds that, if at least 12 months of data are used in the optimization, portfolios' performance is statistically indistinguishable. For these reasons, high-frequency data has not received much attention from researchers.

3.3 Robust Selection of Portfolio Weights

A third approach to repairing the standard MPT framework targets the optimization stage, instead of its inputs $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. There are three ways to alter the optimization stage: incorporate uncertainty in the optimization, restrict portfolio weights, or use penalized optimization.

Note that all of these methods rely – to differing extents – on the standard sample estimates of portfolio inputs, $\boldsymbol{m}_T$ and $\boldsymbol{S}_T$.

3.3.1 Robust Optimization

Some researchers argue that, instead of optimizing a function based on noisy estimates of portfolio inputs $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{-1}$, we should directly incorporate parameter uncertainty and estimation error in the optimization problem. In other words, we should use an optimization technique that produces portfolio weights that are robust to (minor) changes in portfolio inputs—an approach labeled *Robust Optimization* (see Pfaff (2016, Ch. 10) for a review). Goldfarb and Iyengar (2003) focus on the expected returns vector – whose uncertainty is more problematic for portfolio selection – and use *uncertainty sets* to bound the perturbations in parameter inputs and their effect on portfolio weights. Specifically, the authors assume observed asset returns are determined by the multi-factor model $\boldsymbol{r} = \boldsymbol{\mu} + \boldsymbol{V}^\intercal \boldsymbol{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{f} \in \mathbb{R}^M$ is the vector of random returns of the $M < N$ factors that drive asset returns, $\boldsymbol{V} \in \mathbb{R}^{M \times N}$ is the factor loading matrix, and other terms have their usual meaning. Crucially, $\boldsymbol{\mu}$, $\boldsymbol{V}$, and the covariance matrices of $\boldsymbol{f}$ and $\boldsymbol{\epsilon}$, $\boldsymbol{F}$ and $\boldsymbol{\Sigma}_\epsilon$, respectively, are unknown but bounded within well-defined sets.[31] Using this model, Goldfarb and Iyengar (2003) show that the uncertainty sets for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are parameterized by the data $\boldsymbol{r}$ and $\boldsymbol{f}$, the estimators used to recover the model's parameters $\boldsymbol{\mu}$, $\boldsymbol{V}$, $\boldsymbol{F}$, and $\boldsymbol{\Sigma}_\epsilon$, and a parameter ω that tunes the confidence level. This allows the researcher to make probabilistic guarantees about the resulting portfolios' performance. Moreover, an attractive feature of this optimization problem is that, computationally, its difficulty is comparable to that of the quadratic programming required by the standard MPT problem (1). Tütüncü and Koenig (2004) extend this approach to produce a min-max solution to portfolio selection. Namely, the authors select weights that, for a target level of returns, minimize portfolio risk assuming the maximum possible value of risk within the uncertainty set. However, this problem requires a more

[31] Note that the model implies $\boldsymbol{\Sigma} = \boldsymbol{V}^\intercal \boldsymbol{F} \boldsymbol{V} + \boldsymbol{\Sigma}_\epsilon$.

complicated optimization algorithm.

3.3.2 Restricting Weights

A different way to alter the portfolio optimization stage is to restrict portfolio weights. Mathematically, this is equivalent to adding one or more constraints to the standard MPT optimization problem (1). Using sample estimates of portfolio inputs, Jagannathan and Ma (2003) find that restricting short-sales (i.e. forcing all weights to be non-negative) decreases optimal portfolio risk. Moreover, optimal No-Short-Sales (NSS) portfolios with $\boldsymbol{\Sigma}$ estimated through $\boldsymbol{S}_T$ perform comparably to unrestricted portfolios constructed with $\boldsymbol{\Sigma}$ estimated through factor models, shrinkage estimators, or high-frequency data. The reason short-sales constraints help was illustrated in the example of Section 2.2.3: short-selling multiplies portfolio instability owed to estimation error. That is, short-selling produces portfolios with large negative weights on assets with negative returns and/or high correlations with other assets, even when these attractive properties are the product of noise.

Interestingly, Jagannathan and Ma (2003) demonstrate that the NSS portfolio can also be obtained by shrinking the covariance matrix. Because assets with high (low) correlation with other assets, all else equal, receive large negative (positive) weights in unrestricted optimization, forcing all weights to be non-negative is analogous to shrinking (inflating) the rows in $\boldsymbol{\Sigma}$ corresponding to high- (low-)correlation assets. In particular, the authors show that restricting short-sales is equivalent to unrestricted portfolio optimization with $\boldsymbol{\Sigma}_{NSS} = \boldsymbol{S}_T + (\boldsymbol{\delta}\mathbf{1}_N^\intercal + \mathbf{1}_N\boldsymbol{\delta}^\intercal) - (\boldsymbol{\lambda}\mathbf{1}_N^\intercal + \mathbf{1}_N\boldsymbol{\lambda}^\intercal)$, where $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are length-N vectors of the Lagrange multipliers from non-negativity constraints and maximum-weight constraints, respectively, added to the standard MPT problem (1).[32] This correspondence between short-sale constraints and covariance matrix shrinkage is another way to understand why restricting weights can reduce estimation error and stabilize portfolio optimization. Despite the benefits of short-sale constraints for the standard approach, DeMiguel et al. (2009) find that they do

[32] Note that the δ's determine how much to shrink the rows in $\boldsymbol{S}_T$ corresponding to high-covariance assets and the λ's determine how much to inflate the rows corresponding to low-covariance assets.

not improve the performance of all optimization approaches: restricting short-sales when $\mathbf{\Sigma}$ is estimated through factor models or shrinkage estimators results in inferior portfolios. In other words, combining portfolio optimization techniques does not always result in a whole greater than the sum of its parts.

DeMiguel, Garlappi and Uppal (2009), in turn, suggest an even more radical restriction on portfolio weights: setting all weights equal to $1/N$, i.e. using an equally weighted portfolio. Surprisingly, using several different datasets and a range of performance metrics, they find that the $1/N$ allocation matches or beats 14 different portfolio selection methods, including short-sale constraints. In theory, some of these methods can outperform the equally weighted portfolio, but only if T exceeds N by dozens of orders of magnitude—an unrealistic data requirement. The authors attribute the failure of competing methods to beat the naive $1/N$ allocation to the insurmountable nature of estimation error in portfolio inputs. Even though these methods lessen the optimization's dependence on $\boldsymbol{S}_T$ and $\boldsymbol{m}_T$, estimation error is inherently so large that optimal diversification rarely beats naive diversification. The authors also offer a general reason for constraining weights, through the $1/N$ rule or other rules, over alternative approaches: weight constraints can be derived from asset-pricing theory and result in interpretable allocations.[33] The same cannot be said for constraints on assets' moments.

3.3.3 Penalized Optimization

A related approach to constraining weights is *Penalized Optimization* (also known as regularized optimization). The latter involves minimizing the sum of portfolio risk and a penalty term that is a function of portfolio weights. This approach is similar to weight restrictions in that both approaches constrain/shrink/restrict weights. However, penalized optimization constrains the *total* size (norm) of weights, while weight restrictions constrain *each* weight.

[33] Behr, Guettler and Miebs (2013) develop a method for flexibly bounding each portfolio weight, and demonstrate that their approach nests several others. Namely, they choose bounds that minimize the *ex ante* mean squared error of covariance matrix entries.

Different penalties on weights induce different properties in the resulting portfolio. The most popular penalty is the ℓ_1-norm, which induces sparsity in the portfolio weight vector, much like it does in the covariance matrix for sparse graphical models (Section 3.2.7).[34] In practice, *Sparse Optimal Portfolios* involve active positions on relatively few assets and short positions of moderate magnitude. The first application of sparse portfolio optimization is Brodie et al. (2009). I briefly review their approach.

We begin by noting that $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{r} - \boldsymbol{\mu})(\boldsymbol{r} - \boldsymbol{\mu})^\intercal]$ can be rewritten as $\mathbb{E}[\boldsymbol{r}_t\boldsymbol{r}_t^\intercal] - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$. Recalling that the objective function in Problem 1 is $\boldsymbol{w}^\intercal\boldsymbol{\Sigma}\boldsymbol{w}$ and the target returns constraint (1b) is $\boldsymbol{w}^\intercal\boldsymbol{\mu} = r_*$, after some manipulation the objective function can be rewritten as $\mathbb{E}[(r_* - \boldsymbol{w}^\intercal\boldsymbol{r}_t)^2]$. (Recall that there is also the allocation constraint (1c), $\boldsymbol{w}^\intercal\mathbf{1}_N = 1$.) Replacing expectation with sample average and $\boldsymbol{\mu}$ with its sample estimate $\boldsymbol{m}_T$, the objective function becomes $\frac{1}{T}\sum_t^T(r_* - \boldsymbol{w}^\intercal\boldsymbol{r}_t)^2$ and the target returns constraint becomes $\boldsymbol{w}^\intercal\boldsymbol{m}_T = r_*$. (The allocation constraint remains the same.) Notice that this problem is equivalent to a multivariate regression problem with panel data (i.e. given data, choose coefficients that minimize sum of squared deviations from response), with r_* as the response vector $\boldsymbol{y}$ ($y_t = r_* \; \forall t$), $\boldsymbol{w}$ as the coefficient vector $\boldsymbol{\beta}$, and $\boldsymbol{r}_t$ as the predictor vector $\boldsymbol{x}_t$. Adding, an ℓ_1 penalty to the objective function ($||\boldsymbol{w}||_1 = \sum_i^N |w_i|$), plus a tuning parameter τ, the sparse portfolio optimization problem becomes

$$\boldsymbol{w}_*^{[\tau]} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \quad \sum_t^T(r_* - \boldsymbol{w}^\intercal\boldsymbol{r}_t)^2 + \tau||\boldsymbol{w}||_1 \tag{6a}$$

$$\text{s.t.} \qquad \boldsymbol{w}^\intercal\boldsymbol{m}_T = r_* \tag{6b}$$

$$\boldsymbol{w}^\intercal\mathbf{1}_N = 1 \tag{6c}$$

where $1/T$ has been absorbed in τ. In other words, Brodie et al. (2009) reformulate the standard MPT problem (1) as a *LASSO* regression problem with two added constraints (Tibshirani, 1996).

[34]Sparse optimization methods have been popularized by the machine learning literature during the last 20 years, but have only appeared in the portfolio selection literature during the last decade.

To solve this problem, the authors use a variant of the homotopy / *Least-Angle Regression* (LARS) algorithm (Efron et al., 2004). The latter applies to unconstrained ℓ_1-penalized regression, thus the authors augment LARS to account for the two constraints above. In particular, given assets' mean returns and the target returns level, they first prescribe the affine subspace that satisfies the target returns and allocation constraints: $\{\mathcal{H} : \boldsymbol{w} \in \mathbb{R}^N \;\mid\; \boldsymbol{w}^\intercal \boldsymbol{m}_T = r_* \;\cap\; \boldsymbol{w}^\intercal \mathbf{1}_N = 1\}$. Then, they solve the problem $\boldsymbol{w}_*^{[\tau]} = \text{argmin}_{\boldsymbol{w}\in\mathcal{H}} \sum_t^T (r_* - \boldsymbol{w}^\intercal \boldsymbol{r}_t)^2 + \tau||\boldsymbol{w}||_1$. Starting from very large values, they gradually decrease τ and solve $\boldsymbol{w}_*^{[\tau]}$ (a linear system) only at breakpoints of the solution path—points where the slope changes. The locus of solutions $\boldsymbol{w}_*^{[\tau]}$ moves through $\mathbb{R}^N$ on a piecewise affine path, which is stored (see online appendix in Brodie et al. (2009) for more details). Along that locus, another quantity that is stored is the *active set*: the number of non-zero weights. Typically, as τ increases, the active set shrinks and the portfolio becomes sparser.[35]

The approach of Brodie et al. (2009) produces four interesting insights into penalized portfolio optimization. First, it is easy to show that the ℓ_1-norm only penalizes *negative* weights.[36] In other words, sparse optimization limits shorting. Relatedly, because the ℓ_1 penalty constrains negative weights' norm, it limits *total* shorting, unlike imposing *individual constraints* on weights as in Jagannathan and Ma (2003). Limits on total shorting are closer to actual investor behavior, which often takes the form of rules like the 130/30 portfolio (long positions equal to 130% of wealth, short positions equal to 30% of wealth). A second interesting result from Brodie et al. (2009) is that, if we select an extremely high value of τ, we obtain a very sparse portfolio consisting only of (weakly) positive weights. In fact, that portfolio is the same as the one obtained by solving the standard MPT problem (1)

[35] Note that penalized portfolio optimization still relies on the usual portfolio inputs: Brodie et al. (2009) use the sample estimator $\boldsymbol{m}_T$ for expected returns, and other authors that have reformulated the penalized portfolio optimization to include the covariane matrix (e.g. DeMiguel et al. (2009)) use $\boldsymbol{S}_T$ to estimate it. Fortunately, Henriques and Ortega (2014) find that the perfromance of ℓ_1-penalized optimization is robust to using different estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

[36] To see this, note that the allocation constraint implies $\sum_{w_i\geq 0} w_i + \sum_{w_i<0} w_i = 1 \Leftrightarrow \sum_{w_i\geq 0} w_i = 1 - \sum_{w_i<0} w_i$. This allows us to rewrite the penalty term $\tau \sum_i^N |w_i|$ in the objective function as $\tau(\sum_{w_i<0}|w_i| + \sum_{w_i\geq 0} w_i) = \tau(\sum_{w_i<0}|w_i| + 1 - \sum_{w_i<0} w_i) = \tau(\sum_{w_i<0}|w_i| - \sum_{w_i<0} w_i) + \tau = 2\tau \sum_{w_i<0}|w_i| + \tau$, where the latter τ does not affect optimization.

with a short-sales restriction, i.e. the NSS portfolio of Jagannathan and Ma (2003).[37] The equivalence between short-sales restrictions and very high ℓ_1-penalization underscores the destabilizing effect of short-selling on portfolio selection. A third noteworthy feature of the ℓ_1-norm approach is that *how many* assets appear in the active set is determined by τ, but *which* assets appear in the active set is implicitly determined by $\boldsymbol{\Sigma}$ (or its sample counterpart $\boldsymbol{S}_T$).[38] Once again, this emphasizes the importance of assets' covariance matrix for MPT. Finally, and relatedly, the stabilizing effect of the norm penalty on portfolio selection (see next paragraph) is due to its effect on the covariance matrix (this also holds for other norm penalties I review below). Namely, by appropriately constraining weights' size, we reduce the optimization's sensitivity to colinearities in asset returns. This is analogous to reducing estimation error for $\boldsymbol{\Sigma}^{-1}$, one of two sources of portfolio instability.

In addition to the interesting points revealed above, sparse optimal portfolios display four advantages. First and foremost, they reduce the negative effect of estimation error on portfolio selection. Fan, Zhang and Yu (2012) prove that, for a wide range of values of the penalty term, the sparse optimal portfolio is robust to estimation error in portfolio inputs. Specifically, the authors show that estimation risk is bounded by a quadratic function of the ℓ_1-norm of portfolio weights. Thus, constraining weights is equivalent to constraining estimation risk (Li, 2015). Moreover, Fan, Zhang and Yu (2012) demonstrate that the true and empirical sparse optimal portfolio yield roughly the same utility, and theoretical and empirical risk are approximately equal. A second benefit of sparse portfolios is that their small number of active positions implies low brokerage and other transaction costs (e.g. trading commissions, bid-ask spreads), since investors have to execute fewer trades.[39]

[37] To see why the max-τ portfolio is the NSS portfolio, first note that the ℓ_1 penalty on weights is equivalent to adding a constraint that the sum of weights' absolute value must be lower than some value $\bar{w}$. The minimum $\bar{w}$ that is consistent with the allocation constraint is 1; a lower $\bar{w}$ would violate the allocation constraint, as weights cannot sum to 1 if absolute weights sum to less than 1. Given that, note that, if any weight is negative, it is impossible for weights to sum to 1 (allocation constraint) *and* absolute weights to sum to 1 (min-$\bar{w}$ constraint). In other words, the max-τ portfolio can only have weakly positive weights, i.e. is the NSS portfolio.

[38] Though $\boldsymbol{\Sigma}$ (nor $\boldsymbol{S}_T$) appear in the sparse optimization problem, note that the objective function penalizes assets' relative deviations from the target returns level.

[39] For institutional investors, transaction costs are less significant, as they benefit from economies of scale and

Relatedly, sparse portfolios are very attractive to investors facing liquidity constraints (e.g. margin requirements by their broker), and hence seek to minimize leveraged positions, short or long. A third benefit of sparse portfolios is that they tend to be more stable; changes in assets' expected return and covariance result in smaller changes for sparse portfolios versus non-sparse ones. This minimizes the time spent opening, rebalancing, and closing positions on assets. The fourth advantage of sparse portfolios is that they reflect a cognitive preference for parsimony, which is embedded in many domains of rational human behavior (Gabaix, 2014).

However, ℓ_1-penalized portfolios are not without limitations. Significant criticisms of using the ℓ_1 penalty is that it often results in under-diversified portfolios with extreme weights (Yen, 2015), that it produces biased estimates of large weights (Fastrich, Paterlini and Winker, 2015), and that produces intertemporally unstable portfolios when assets are highly correlated (De Mol, 2016).[40] Though these criticisms have some theoretical basis, they lose their importance if ℓ_1-penalized portfolios still perform better than competing methods' portfolios—a claim that must be assessed empirically (more below). A less significant criticism of ℓ_1-penalization is that, even though the ℓ_1-norm is the only penalty that induces sparsity while being a convex combination of weights, it nonetheless creates a tough optimization problem without a closed-form solution (De Mol, 2016). Similarly, the non-linear shrinkage induced by the ℓ_1-norm obscures determination of the efficient frontier; thus, for a given parameter, we have to trace the efficient frontier by solving the optimization problem point-by-point for each value of target returns (De Mol, 2016).

Partly due to the limitations of the ℓ_1-norm, DeMiguel et al. (2009) develop a broader approach to penalized optimization, ***A**-norm penalization*. To do this, the authors supplement the standard MPT problem (1) with the additional constraint $\boldsymbol{w}^\intercal \boldsymbol{A} \boldsymbol{w} \leq \delta$, where $\boldsymbol{A}$ is a $N \times N$ positive-definite matrix and $\delta > 0$. Their approach reveals some interesting nuances

market power.

[40] The last criticism is due to the LASSO's well-known instability in selecting active coefficients (weights) when predictors (asset returns) are highly correlated. In the context of portfolio selection, this instability manifests itself as high portfolio turnover. This matters only insofar as turnover is a criterion for evaluating portfolios.

of penalized portfolio optimization. First, the authors show that the ℓ_1-norm imposes an upper limit on total shorting, equal to $\frac{\delta-1}{2}$. Note that a larger δ (weaker penalty) allows more shorting, while $\delta = \infty$ yields the standard (short-sale unconstrained) optimal portfolio. Second, A-norm penalized portfolios are equivalent to the optimal portfolios obtained by shrinking $\boldsymbol{S}_T$ towards $\boldsymbol{A}$. For example, solving the $\boldsymbol{A}$-norm penalized optimization with $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{SIM}$ results in the same portfolio as solving the standard optimization with $\boldsymbol{S}_T$ shrunk towards the single-index model's matrix; that is, the shrinkage estimator suggested by Ledoit and Wolf (2003*b*) (Section 3.2.6). Similarly, setting $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{CCM}$ yields the same portfolio as shrinking $\boldsymbol{S}_T$ towards the constant correlation model's matrix; i.e. the approach of Ledoit and Wolf (2003*a*) (Section 3.2.6). Once again, these examples underline the connection between different portfolio optimization methods and shrinking (elements of) the covariance matrix. Third, the special case of $\boldsymbol{A} = \boldsymbol{I}_N$ results in ℓ_2-norm penalization. More importantly, reformulating the portfolio problem from an optimization with the constraint $\boldsymbol{w}^\intercal \boldsymbol{I}_N \boldsymbol{w} \leq \delta$ to a ℓ_2-penalized regression problem, as Brodie et al. (2009) do for the ℓ_1-norm, results in a *Ridge Regression* problem (with a tuning parameter and two added constraints) (Hoerl and Kennard, 1970). Though it does not induce sparsity, Ridge regression is more stable than LASSO, and it also admits a closed-form solution (De Mol, 2016). Fourth, DeMiguel et al. (2009) give interesting Bayesian interpretations to the use of various norms in portfolio optimization. For example, if investors' prior is that portfolio weights are i.i.d. following the Double Exponential, then their *maximum a posteriori* estimate are the weights produced by ℓ_1-penalized optimization.[41] Finally, the authors provide a useful algorithm for tuning the penalty parameter (δ in their formulation): k-fold cross validation (k-CV) to choose the δ that minimizes mean portfolio risk across samples. However, in addition to being computationally intensive, Yen (2015) argues that k-CV produces unstable sequences of the penalty parameter, which can weaken portfolio performance to the point of nullifying gains from

[41] The $\boldsymbol{A}$-norm portfolio, in turn, produces weights equal to the mode of the posterior resulting from a Multivariate Normal prior with covariance matrix $\boldsymbol{A}$. Similarly, the weights of the ℓ_2-norm portfolio are analogous to the mode of the posterior resulting from a standard Multivariate Normal prior (with covariance matrix $\boldsymbol{I}_N$).

penalization.

Several other studies have contributed to the penalized portfolio optimization approach. Carrasco and Noumon (2011) give a formal treatment of k-CV and other algorithms for tuning the penalty parameter for various norms. Fernandes, Rocha and Souza (2012) incorporate information on assets' industry and use a penalty that induces sparsity while accounting for within-industry correlations. Fastrich, Paterlini and Winker (2015) use a non-convex penalty in a model similar to the Weighted LASSO, which maintains sparsity while reducing LASSO's bias when estimating large weights.[42] The authors' approach gives a different penalty to each weight, overweighing assets that improve portfolio performance more than the ℓ_1 penalty does (and vice-versa for assets that weaken portfolio performance). Finally, Yen (2015) introduces a weighted norm approach that combines the ℓ_1 and squared ℓ_2 penalties. The author's rationale is that the ℓ_1-norm creates sparse and stable weights, while the ℓ_2-norm mitigates the under-diversification and extreme weights often produced by the ℓ_1-norm.

Overall, penalized optimization methods perform very favorably compared to other methods. Across a range of datasets and performance metrics, Brodie et al. (2009) find that the ℓ_1-penalized portfolio outperforms the NSS and $1/N$ portfolio. An interesting pattern they find is that portfolio complexity (number of active weights) has a curvilinear effect on the portfolio's Sharpe Ratio: both too few and too many active weights reduce SR. Similarly, DeMiguel et al. (2009) find that most norm-constrained portfolios have a higher SR than benchmark portfolios like NSS and $1/N$, have a similar SR to those of more robust methods like covariance shrinkage, but have lower turnover than the latter. Despite the encouraging findings of the above researchers, it is worth assessing the performance of penalized optimization – and other portfolio selection methods – through a comprehensive comparison.

[42] Yen (2015) notes that, though they burden computation, non-convex penalties improve portfolio performance more than convex penalties when T/N is small.

4 Evaluating Competing Approaches

4.1 Data

To compare the performance of the approaches reviewed in Section 3, I use real data on asset returns, instead of synthetic data. Though the latter provide more statistical power and flexibility in evaluating competing methods' performance, they are usually paired with estimation criteria like (Root) Mean Squared Error. Crucially, these estimation criteria assume that there is a ground-truth model, which is not true for asset returns (Senneret et al., 2016, p. 6). As such, I evaluate competing methods by constructing their respective portfolios using real data and comparing their out-of-sample (OOS) performance across a range of financial metrics (see Section 4.3 for more).

I use a dataset employed by several other studies, the Fama and French 100 Portfolios (FF100) (Brodie et al., 2009; Goto and Xu, 2015; Senneret et al., 2016). FF100 includes the monthly returns of 100 portfolios formed using all equities listed on US stock exchanges.[43] Note that, because I use this dataset, the portfolios I will form will be *portfolios of portfolios of stocks*. That said, this comes at no cost: apart from the Multi-Index Model and Multi-Group Model, none of the approaches reviewed in Section 3 require their inputs to be individual stocks. I subset the last 60 years of data (2/1958–2/2018), which forces me to drop 8 portfolios that have missing returns. Thus, my resulting dataset has a ratio $T/N \approx 8$, with $T = 720$ and $N = 92$.

[43] The FF100 portfolios are constructed in the following manner. At the end of every June, all US-listed stocks are divided into two sets of deciles: one based on their market value (size) and another based on the ratio of their book value to market value (book-to-market). Stocks' size is determined through the closing price of the last trading day of June, and book-to-market is determined using the book value disclosed in the fourth-quarter earnings report of the preceding year and the market value based on the last trading day of the preceding December. The 100 portfolios are formed through the intersection of the 10 size portfolios and the 10 book-to-market porfolios. These portfolios are held until the next June, at which point the process is repeated to reflect annual changes in size and book-to-market. See http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/det_100_port_sz.html for more details.

4.2 Optimization & Trading Strategy

I follow the literature in adopting a rolling-window strategy to construct and evaluate competing portfolios. I use training windows with a length of about 15 years ($T_{train} = 184$). This gives me a T/N ratio of roughly 2 for constructing the portfolios, sufficient to train even the more "greedy" algorithms. My test windows have a length of 6 months ($T_{test} = 6$), giving rise to 90 test periods. Shorter test windows result in more test periods, but amplify within-test period noise from stock market volatility and burden computation (more test periods require more training periods). Note that the test window's length is the amount of time that the investor will hold the respective portfolio before re-solving the optimization problem using the next training window. In other words, I assume that investors rebalance their portfolios every 6 months, which is realistic for retail investors.

To illustrate my approach in more detail, consider method k, which estimates assets' covariance through $\hat{\boldsymbol{\Sigma}}_k$ and assets' expected returns through $\hat{\boldsymbol{\mu}}_k$. First, for each training period $t = 1, \ldots, 90$, I estimate $\hat{\boldsymbol{\Sigma}}_{k,t}$ and $\hat{\boldsymbol{\mu}}_{k,t}$, and I obtain optimal relative weights $\boldsymbol{w}_{kt} = \frac{\hat{\boldsymbol{\Sigma}}_{k,t}^{-1}\hat{\boldsymbol{\mu}}_{k,t}}{\mathbf{1}_N^{\intercal}\hat{\boldsymbol{\Sigma}}_{k,t}^{-1}\hat{\boldsymbol{\mu}}_{k,t}}$ (I ignore excess returns over r_f for simplicity). Then, for each test period and for each evaluation metric, I use realized returns to calculate the respective portfolio's OOS performance. For example, for training period 1 (2/1958 – 5/1973) I obtain $\boldsymbol{w}_{k1}$, and I use realized moments in test period 1 (6/1973 – 11/1973), $\boldsymbol{m}_1$ and $\boldsymbol{S}_1$, to estimate k's OOS expected return and risk in that period, $\hat{\mu}_{k,1} = \boldsymbol{w}_{k,1}^{\intercal}\boldsymbol{m}_1$ and $\hat{\sigma}_{k,1} = \sqrt{\boldsymbol{w}_{k,1}^{\intercal}\boldsymbol{S}_1\boldsymbol{w}_{k,1}}$ (and similarly for the other performance metrics in the next section). For period 2, I shift the training window to incorporate period 1's test window and repeat the same process. This process is applied to all the portfolio selection methods evaluated, with minor adjustments made for methods that require tuning of parameters (e.g. shrinkage).

4.3 Performance Metrics

In addition to the two quantities mentioned above, expected return and risk, I use several other financial performance metrics to compare portfolios. (All metrics reflect OOS performance, also known as *ex post* performance in the literature.) The third metric, *Sharpe Ratio* (SR), is the ratio of expected (excess) return to risk. Recall that it was introduced in Section 2 as the slope of the efficient frontier. The fourth metric is the *Inverse Herfindahl-Hirschman Index* (HH^{-1}), defined as the inverse of the sum of squared portfolio weight weights: $\mathrm{HH}_t^{-1} = (\sum_i^N w_{i,t}^2)^{-1}$, $t = 1, \ldots, T$. Since HH is a (well-known) measure of concentration, IHH acts as a measure of diversification. (Note that $\mathrm{IHH}_{max} = N$.) Moreover, higher IHH usually leads to more stable portfolios over time, thereby limiting transaction costs (Senneret et al., 2016). The fifth metric I use is *Gross Exposure* (GE), defined as the sum of absolute portfolio weights: $\mathrm{GE}_t = \sum_i^N |w_{i,t}|$.[44] Since the portfolio allocation constraint requires that $\sum_i^N w_i = 1$, GE is the value of all positions in the portfolio, short and long, and GE> 1 implies a leveraged portfolio. In other words, if the investor's capital equals \$100K and her portfolio's GE equals 20, it means that her exposure is \$2M and her leverage is \$1.9M. Even with costless borrowing, limiting gross exposure is desirable because, all else equal, it lowers portfolio risk (Fan, Zhang and Yu, 2012). The sixth metric is I use is the *fraction of gross exposure owed to short positions* (Short), defined as: $\mathrm{Short}_t = \sum_{i:\, w_i<0} |w_{i,t}|/\mathrm{GE}_t$. Though this is not a standard performance metric, it identifies the portfolio's dependence on shorting. And since shorting is not available to some institutional investors, Short reveals the extent to which each method is open to all investors. Finally, I calculate *Turnover* (TO), defined as the change in portfolio weights between holding periods: $\mathrm{TO}_t = \sum_i^N |w_{i,t} - w_{i,t-1}|$. Despite not including transaction costs in the optimization problem, lower turnover decreases transaction costs and, all else equal, increases investor utility–especially for retail investors.

4.4 Methods & Computation

I compare the performance of 16 different methods reviewed in Section 3, in addition to the standard MPT portfolio, the $1/N$ portfolio, and the market index, resulting in 19 portfolios. Before presenting my results, I outline some implementation details. Every method other

[44] Note that HH_t^{-1} is equivalent to the portfolio's inverse squared ℓ_2-norm, while GE is equivalent to its ℓ_1-norm.

than the standard approach and benchmarks involves replacing *either* the sample estimator of $\boldsymbol{\Sigma}$ *or* the sample estimator of $\boldsymbol{\mu}$ with an alternative estimator. Thus, if a particular method involves estimating $\boldsymbol{\Sigma}$ with $\hat{\boldsymbol{\Sigma}}_k$, then it is implied that $\boldsymbol{\mu}$ is estimated by $\boldsymbol{m}$, and the vice-versa. In addition, aside from GLASSO, which directly estimates $\boldsymbol{\Sigma}^{-1}$, all other methods require inverting $\boldsymbol{S}$ or $\hat{\boldsymbol{\Sigma}}_k$ with standard numerical to compute the optimal portfolio. For all methods, I compute the optimal relative weights (Equation 4), which do not depend on the target excess return constraint (2b). Relatedly, all of the performance metrics I present refer to portfolios fully described by their relative weights and with no risk-free lending/borrowing (i.e. tangency portfolios). To compute those portfolios, though, I must assume a risk-free rate, so I set $r_f = 0.52$.[45] All computations were done in `R`, using base functions where possible.

For the *Standard MPT Portfolio* (Standard), I use the sample estimators $\boldsymbol{m}$ and $\boldsymbol{S}$. For the market index, I use the returns on the S&P500.[46] Among robust estimators of expected returns (Section 3.1.1), I use the *Trimmed Mean* with $k = 0.1$ and the *Winsorized Mean* with $k = 0.2$. Among robust estimators of the covariance matrix (Section 3.2.1), I use the *Minimum Covariance Determinant* (MCD) and *Minimum Volume Ellipsoid* (MVE) estimators, both implemented using the `rrcov` package.[47] Among model-based estimators, I use the *Single-Index Model* (SIM) with S&P500 as the index and the *Constant Correlation Model* (CCM). I also use *Vasicek-adjusted betas* and *Blume-adjusted betas* to obtain corrected SIM covariance matrices. Again, note that, because the assets in my data are portfolios and not individual stocks, I cannot assign them to groups/industries and apply the Multi-Index Model or Multi-Group Model. In terms of shrinkage estimators, I use *Shrinkage to the CCM* and *Shrinkage to the SIM* to estimate the covariance matrix (Section 3.2.6) and *Shrinkage to*

[45] Because my data is monthly, $r_f = 0.52$ corresponds to 6.23% non-compounded annual return (fixed income), equal to the average yield of 10-year Treasury notes during the period studied. Data for this calculation was downloaded from https://finance.yahoo.com/quote/%5ETNX/history.

[46] Data was downloaded from https://finance.yahoo.com/quote/%5EGSPC/history.

[47] I also experiment with the Orthogonalized Gnanadesikan-Ketterning estimator, M-estimators, S-estmators, and other functions included in `rrcov` and related packages, but find that MCD and MVE consistently outperform other robust estimators of covariance.

the Grand Mean ($\bar{m}$) to estimate expected returns (Section 3.1.2). All shrinkage models are implemented through the `tawny` package, which computes the optimal shrinkage coefficient. In terms of *Sparse Graphical Models*, GLASSO (Section 3.2.7) is implemented through the `huge` package, which tunes the regularization parameter τ through a Rotation Information Criterion approach (Zhao et al., 2012). Instead of selecting τ through CV or subsampling, this approach directly estimates the optimal level of sparsity using random rotations of the precision matrix. *Random Matrix Theory* (RMT) estimators are implemented using the `covmat` package. The latter fits a Marchenko-Pastur density curve to the eigenvalues of the sample correlation matrix, then replaces eigenvalues below a cutoff with the average eigenvalue above the cutoff and converts the filtered correlation matrix to a covariance matrix.[48] *L1-* and *L2-Penalized Portfolios* are constructed using the `cccp` package, and the regularization parameter τ is tuned using T-fold CV, as described in DeMiguel et al. (2009, fn. 13).[49] In particular, I use a grid of 10 equally-spaced values for τ and select the value that minimizes OOS portfolio variance across the $T = 184$ folds. For each value of τ, this involves first solving Problem 6 for $t = 1, \ldots, T$ to obtain weight vectors $\boldsymbol{w}_t^{[\tau]}$, then calculating OOS portfolio returns on the withheld periods $\hat{r}_{t[\tau]} = \boldsymbol{w}_{t[\tau]}^{\intercal} \boldsymbol{r}_t$, and finally calculating OOS variance across the T folds $\hat{\sigma}^2_{[\tau]} = \frac{1}{T-1} \sum_t^T (\hat{r}_{t[\tau]} - \bar{r}_{[\tau]})^2$, where $\bar{r}_{[\tau]} = \frac{1}{T} \sum_t^T r_{t[\tau]}$.[50] The optimal τ minimizes $\hat{\sigma}^2_{[\tau]}$. Note that a new τ has to be computed for each of the 90 rolling training windows, and that this tuning algorithm is computationally very heavy.

4.5 Results

In Table 1 I follow the literature and report means and standard errors for all my performance metrics. The latter are calculated across the 90 rolling test periods, each of 6-month length. As such, the numbers I report should be interpreted substantively within the context of a 6-month portfolio holding period. For example, the mean risk reported is the

[48] I also try deleting eigenvalues below the cutoff, and I obtain very similar portfolios.

[49] I also consulted some Matlab code in McKenzie (2017) and Saucedo (2014).

[50] Following Brodie et al. (2009), I set the target return r_* in each training period equal to the average return achieved by the 1/N portfolio in that period.

standard deviation in returns investors will experience from holding the respective portfolio in a typical period, not the standard deviation in their returns across all holding periods (45 years).[51] Note that using the latter approach does not change the relative ordering of methods' performance. A visual comparison of different methods' average return and risk is shown in Figure 4.5, where points to the north-west of a point denoting a method represent superior methods.

Table 1 shows significant variation in the performance of the various methods. In terms or return, the best-performers are factor models: the SIM covariance matrix with Blume- and Vasicek-adjusted betas, as well as Shrinkage-SIM, which is a mixture between the standard and SIM approach. Curiously, the CCM covariance matrix and its shrinkage counterpart deliver highly negative average returns—the only methods to do so. Overall, the model-based methods easily outperform the standard approach across all metrics.

However, when factoring-in risk, model-based methods lose their edge, and the newer methods of RMT and penalized optimization dominate. Indeed, the L1-penalized portfolio carries even lower risk than the equal-weighted 1/N portfolio, a particularly tough benchmark to beat on risk (DeMiguel, Garlappi and Uppal, 2009). That said, the L1 portfolio is riskier than simply buying and holding the S&P500 market index, a very popular approach for retail investors. In line with previous findings, the standard portfolio is relatively risky, but, interestingly, portfolios that use robust estimators of expected returns or covariance are even riskier.

Combining return and risk in considering the Sharpe Ratio, we see that the newer methods (last 3 in table) dominate again. Closely following them are GLASSO and a couple of shrinkage estimators. Moreover, all of these methods beat the standard approach and the market in terms of SR, but only L1-penalization with SR$=0.44$ beats the 1/N portfolio. Substantively, the L1 portfolio's SR implies that for every 10 percentage points of risk the investor takes on, she is rewarded with 4.4 percentage points of monthly returns. Again, we

[51] In other words, the observations used to calculate means and standard errors are the holding periods (90), not the months across all holding periods (540).

see that portfolios using robust moment estimators perform poorly on SR, mostly due to their large risk.

Moving on to the less standard performance metrics – the inverse Herfindhal index, gross exposure, gross exposure owed to short positions, and turnover – the clear winner is the L1 portfolio. Crucially, the optimal penalty parameter selected through CV completely restricts short-sales. By doing so, the L1 portfolio results in high diversification (IHH= 85.6 while $\text{IHH}_{max} = 92$), no excess exposure or borrowing (GE=1), no short positions, and near-zero turnover ($TO = 0.04$). As such, the L1 portfolio almost matches the ultra-diversified and ultra-stable market and 1/N portfolios, while delivering higher risk-adjusted returns (SR). On the contrary, mostly because they allow short-sales, other methods produce not just significantly larger exposure, but also much less diversification and higher turnover than the L1, 1/N, and market portfolios. Nevertheless, apart from robust estimators, almost all methods still beat the standard approach on IHH, GE, Short, and TO.

Overall, Table 1 adds to the literature's consensus on the standard approach's shortcomings (Michaud, 1989), the benefits of constraining short-sales (Jagannathan and Ma, 2003), and the difficulty of beating the naive 1/N portfolio DeMiguel, Garlappi and Uppal (2009). Moreover, the L1 portfolio's dominance and, to a lesser extent, the respectable performance of the L2, RMT, and GLASSO portfolios emphasize the potential of new methods for portfolio optimization. That said, given the large standard errors for almost all point estimates – especially for return, risk, and SR – we cannot make *statistically confident* statements about the relative performance of different methods. Though the noise in estimates of portfolio performance is largely owed to the inherent volatility of asset returns, it does prevent us from making strictly statistical inferences about portfolio performance. In short, without estimates of uncertainty better tailored to asset returns, we can only discriminate between portfolios on a financial basis, not a statistical one.

Table 1: Mean and SE of Performance Metrics

Approach	Method	Return	Risk	SR	IHH	GE	Short	Turnover
Benchmarks	S&P500	0.7	3.96	0.32	NA	1	0	0
		(1.74)	(1.92)	(0.59)	(NA)	(0)	(0)	(0)
	1/N	1.23	4.63	0.42	92	1	0	0
		(2.06)	(2.25)	(0.61)	(0)	(0)	(0)	(0)
	Standard	1.17	12.86	0.29	0.08	49.03	0.48	28.28
		(11.24)	(18.31)	(0.51)	(0.05)	(75.41)	(0.01)	(66.53)
Robust $\boldsymbol{\mu}$	Trim	0.31	17.47	0.26	0.08	62.95	0.48	45.63
		(20.43)	(43.4)	(0.48)	(0.05)	(150.56)	(0.01)	(155.19)
	Winsor	1.22	14.07	0.23	0.07	50.22	0.48	34.06
		(10.13)	(20.91)	(0.49)	(0.05)	(80.98)	(0.01)	(84.38)
Robust $\boldsymbol{\Sigma}$	MCD	2.79	15.01	0.2	0.05	57.37	0.49	70.29
		(13.4)	(18.42)	(0.54)	(0.03)	(82.4)	(0.01)	(103.85)
	MVE	1.81	14.92	0.19	0.04	55.89	0.49	66.48
		(6.19)	(14.63)	(0.48)	(0.03)	(48.66)	(0.01)	(56.59)
Model-Based	SIM	2.71	10.83	0.32	0.55	**11.63**	**0.45**	1.98
		(6.36)	(7.26)	(0.66)	(0.31)	(3.85)	(0.01)	(0.72)
	Vasicek	**2.84**	11.73	0.3	0.46	12.69	0.46	**2.2**
		(6.92)	(7.89)	(0.65)	(0.27)	(4.3)	(0.01)	(0.81)
	Blume	**2.92**	11.3	0.32	0.48	**12.25**	0.46	**2.09**
		(6.56)	(7.73)	(0.63)	(0.27)	(4.03)	(0.01)	(0.65)
	CCM	-3.75	27.07	0.34	**0.95**	29.16	**0.44**	24.54
		(82.24)	(94.01)	(0.69)	(0.78)	(88.02)	(0.03)	(95.64)
Shrinkage	Shrink-CCM	-3.78	26.34	0.36	**0.56**	30.77	0.46	25.29
		(82.23)	(94.14)	(0.65)	(0.41)	(87.67)	(0.02)	(95.48)
	Shrink-$\bar{m}$	1.11	11.32	0.32	0.09	42.24	0.48	20.82
		(10.16)	(13.53)	(0.55)	(0.06)	(50.01)	(0.01)	(37.71)
	Shrink-SIM	**2.91**	13.49	0.18	0.05	51.25	0.49	48.78
		(16.06)	(16.78)	(0.53)	(0.04)	(57.36)	(0.01)	(84.15)
Graphical	GLASSO	2.63	9.6	0.36	0.16	20.66	0.47	5.25
		(5.57)	(5.39)	(0.71)	(0.09)	(6.79)	(0.01)	(2.16)
Spectral	RMT	1.41	**5.51**	**0.38**	0.47	14.41	0.46	3.7
		(2.99)	(3.5)	(0.6)	(0.33)	(7.36)	(0.02)	(2.93)
Penalized	L1	1.26	**4.61**	**0.44**	**85.59**	**1**	**0**	**0.04**
		(2.07)	(2.28)	(0.62)	(3.44)	(0)	(0)	(0.02)
	L2	1.55	**5.23**	**0.38**	0.44	14.03	0.46	3.48
		(2.95)	(3.09)	(0.69)	(0.31)	(6.5)	(0.02)	(2.51)

NOTES: Means and standard errors of monthly OOS performance metrics across 90 rolling test periods, each with length 6 months. Bold numbers identify the three best-performing portfolios in each metric—excluding benchmark portfolios. Underlying portfolios constructed using monthly returns data for the period 2/1958 – 2/2018 on $N = 92$ portfolios from FF100, with rolling training periods of length $T = 184$. See Section 4.2 for details on portfolio optimization/trading strategy and Section 3 for a review of methods/portfolios. Mean SR is calculated across test periods, not by dividing mean return by mean risk. See Section 4.3 for more details on metrics.

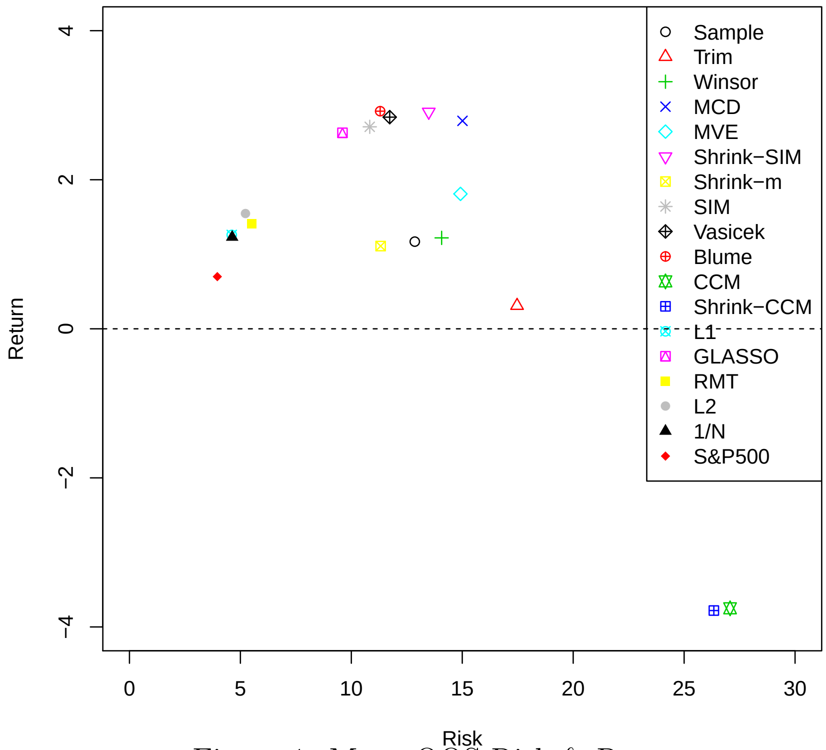

Figure 1: Mean OOS Risk & Return

NOTES: Mean out-of-sample excess monthly return and risk over risk-free rate from following respective portfolio selection method. Averaged across 89 6-month holding periods (6/1973 – 11/2017).

5 Conclusion

This study reviewed the canonical model of optimal financial portfolio selection, the Markowitz model, exposed its limitations, introduced several competing approaches for addressing those limitations, and compared these approaches' performance using a comprehensive empirical exercise. Throughout, I gave more attention to newer approaches, while attempting to bridge the gap between the diverse methodological backgrounds of the researchers and practitioners studying optimal portfolio selection. Overall, I find some promise in the performance of newer techniques, like penalizing weights' norm, filtering the covariance matrix through random matrix theory, and estimating the precision matrix through sparse graphical models. These methods outperform the standard Markowitz model and market index across a range of metrics, while the best-performer, the L1-penalized portfolio, also beats the demanding

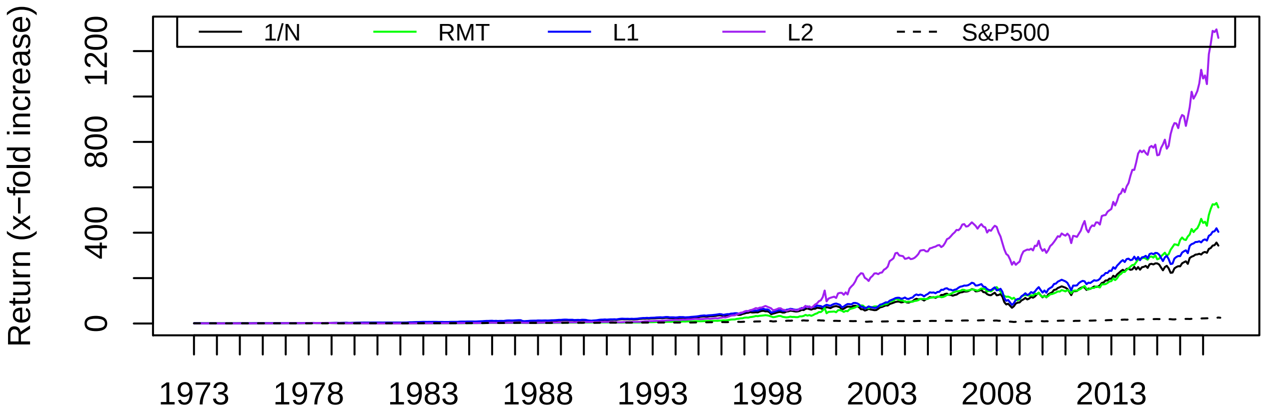

Figure 2: Realized Cumulative Returns

NOTES: Cumulative excess returns over risk-free rate from following respective portfolio selection method. Plotted for 89 6-month holding periods (6/1973 – 11/2017).

benchmark set by the equally weighted portfolio.

There at two natural extensions of this study. First, as noted in the previous section, the estimation of uncertainty in portfolio performance, which would facilitate statistical inference in comparing approaches. Literature on this topic is relatively nascent, thus I leave this advance to future work (Ledoit and Wolf, 2008; DeMiguel et al., 2009). The second obvious way to supplenent my comparison of different methods is to use additional data sources. Potential parameters of the data one could vary are the market assets are traded on (eg US- vs Japan-listed stocks), the number of assets considered (N), the length of the rolling training windows (T) and, by extension, the ratio of the previous two parameters (T/N), and the length of the rolling test/holding periods and, by extension, the number of rolling training and test periods.

References

Awoye, Oluwatoyin Abimbola. 2016. Markowitz Minimum Variance Portfolio Optimization using New Machine Learning Methods PhD thesis (UCL) University College London.

Behr, Patrick, Andre Guettler and Felix Miebs. 2013. “On portfolio optimization: Imposing the right constraints.” *Journal of Banking & Finance* 37(4):1232–1242.

Bien, Jacob and Robert J Tibshirani. 2011. “Sparse estimation of a covariance matrix.” *Biometrika* 98(4):807–820.

Blume, Marshall E. 1975. “Betas and their regression tendencies.” *The Journal of Finance* 30(3):785–795.

Brodie, Joshua, Ingrid Daubechies, Christine De Mol, Domenico Giannone and Ignace Loris. 2009. “Sparse and stable Markowitz portfolios.” *Proceedings of the National Academy of Sciences* 106(30):12267–12272.

Carrasco, Marine and Nérée Noumon. 2011. Optimal portfolio selection using regularization. Technical report Working Paper.

Chan, Louis KC, Jason Karceski and Josef Lakonishok. 1999. “On portfolio optimization: Forecasting covariances and choosing the risk model.” *The Review of Financial Studies* 12(5):937–974.

Cohen, Kalman J and Jerry A Pogue. 1967. “An empirical evaluation of alternative portfolio-selection models.” *The Journal of Business* 40(2):166–193.

De Mol, Christine. 2016. “Sparse Markowitz Portfolios.” *Financial Signal Processing and Machine Learning* pp. 11–22.

DeMiguel, Victor, Alberto Martin-Utrera and Francisco J Nogales. 2013. “Size matters: Optimal calibration of shrinkage estimators for portfolio selection.” *Journal of Banking & Finance* 37(8):3018–3034.

DeMiguel, Victor and Francisco J Nogales. 2009. “Portfolio selection with robust estimation.” *Operations Research* 57(3):560–577.

DeMiguel, Victor, Lorenzo Garlappi, Francisco J Nogales and Raman Uppal. 2009. “A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms.” *Management Science* 55(5):798–812.

DeMiguel, Victor, Lorenzo Garlappi and Raman Uppal. 2009. “Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?” *The Review of Financial Studies* 22(5).

Disatnik, David J and Simon Benninga. 2007. “Shrinking the covariance matrix.” *Journal of Portfolio Management* 33(4):55.

Efron, Bradley, Trevor Hastie, Iain Johnstone, Robert Tibshirani et al. 2004. "Least angle regression." *The Annals of statistics* 32(2):407–499.

El Karoui, Noureddine. 2008. "Spectrum estimation for large dimensional covariance matrices using random matrix theory." *The Annals of Statistics* 36(6):2757–2790.

Elton, Edwin J and Martin J Gruber. 1973. "Estimating the dependence structure of share prices: implications for portfolio selection." *The Journal of Finance* 28(5):1203–1232.

Elton, Edwin J, Martin J Gruber and Manfred W Padberg. 1976. "Simple criteria for optimal portfolio selection." *The Journal of Finance* 31(5):1341–1357.

Elton, Edwin J, Martin J Gruber and Manfred W Padberg. 1977. "Simple rules for optimal portfolio selection: the multi group case." *Journal of Financial and Quantitative Analysis* 12(3):329–345.

Elton, Edwin J, Martin J Gruber, Stephen J Brown and William N Goetzmann. 2009. *Modern portfolio theory and investment analysis*. John Wiley & Sons.

Elton, Edwin J, Martin J Gruber and Thomas J Urich. 1978. "Are betas best?" *The Journal of Finance* 33(5):1375–1384.

Fan, Jianqing, Jingjin Zhang and Ke Yu. 2012. "Vast portfolio selection with gross-exposure constraints." *Journal of the American Statistical Association* 107(498):592–606.

Fan, Jianqing, Yuan Liao and Martina Mincheva. 2011. "High dimensional covariance matrix estimation in approximate factor models." *Annals of statistics* 39(6):3320.

Fan, Jianqing, Yuan Liao and Martina Mincheva. 2013. "Large covariance estimation by thresholding principal orthogonal complements." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(4):603–680.

Fastrich, Björn, Sandra Paterlini and Peter Winker. 2015. "Constructing optimal sparse portfolios using regularization methods." *Computational Management Science* 12(3):417–434.

Fernandes, Marcelo, Guilherme Rocha and Thiago Souza. 2012. "Regularized minimum-variance portfolios using asset group information.".

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2008. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9(3):432–441.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129(4).

Goldfarb, Donald and Garud Iyengar. 2003. "Robust portfolio selection problems." *Mathematics of operations research* 28(1):1–38.

Goto, Shingo and Yan Xu. 2015. "Improving mean variance optimization through sparse hedging restrictions." *Journal of Financial and Quantitative Analysis* 50(6):1415–1441.

Henriques, Julie and Juan-Pablo Ortega. 2014. "Construction, management, and performance of sparse Markowitz portfolios." *Studies in Nonlinear Dynamics & Econometrics* 18(4):383–402.

Hoerl, Arthur E and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12(1):55–67.

Jagannathan, Ravi and Tongshu Ma. 2003. "Risk reduction in large portfolios: Why imposing the wrong constraints helps." *The Journal of Finance* 58(4):1651–1683.

Jorion, Philippe. 1985. "International portfolio diversification with estimation risk." *Journal of Business* pp. 259–278.

Jorion, Philippe. 1986. "Bayes-Stein estimation for portfolio analysis." *Journal of Financial and Quantitative Analysis* 21(3):279–292.

King, Benjamin F. 1966. "Market and industry factors in stock price behavior." *The Journal of Business* 39(1):139–190.

Klemkosky, Robert C and John D Martin. 1975. "The adjustment of beta forecasts." *The Journal of Finance* 30(4):1123–1128.

Kourtis, Apostolos, George Dotsis and Raphael N Markellos. 2012. "Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix." *Journal of Banking & Finance* 36(9):2522–2531.

Laloux, Laurent, Pierre Cizeau, Jean-Philippe Bouchaud and Marc Potters. 1999. "Noise dressing of financial correlation matrices." *Physical Review Letters* 83(7):1467.

Laloux, Laurent, Pierre Cizeau, Marc Potters and Jean-Philippe Bouchaud. 2000. "Random matrix theory and financial correlations." *International Journal of Theoretical and Applied Finance* 3(03):391–397.

Ledoit, Oliver and Michael Wolf. 2008. "Robust performance hypothesis testing with the Sharpe ratio." *Journal of Empirical Finance* 15(5):850–859.

Ledoit, Olivier and Michael Wolf. 2003*a*. "Honey, I shrunk the sample covariance matrix.".

Ledoit, Olivier and Michael Wolf. 2003*b*. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *Journal of empirical finance* 10(5):603–621.

Li, Jiahan. 2015. "Sparse and stable portfolio selection with parameter uncertainty." *Journal of Business & Economic Statistics* 33(3):381–392.

Liu, Qianqiu. 2009. "On portfolio optimization: How and when do we benefit from high-frequency data?" *Journal of Applied Econometrics* 24(4):560–582.

Markowitz, Harry M. 1952. "Portfolio selection." *The Journal of Finance* 7(1):77–91.

Markowitz, Harry M. 1959. *Portfolio Selection: Efficient Diversification of Investments.* John Wiley & Sons.

Martin, R Douglas, Andrew Clark and Christopher G Green. 2010. Robust portfolio construction. In *Handbook of Portfolio Construction.* Springer pp. 337–380.

McKenzie, Daniel. 2017. "Optimal Sparse Markowitz Portfolios." *Working Paper* . http://danielmckenzie.github.io/WriteUp.pdf.

Merton, Robert C. 1980. "On estimating the expected return on the market: An exploratory investigation." *Journal of financial economics* 8(4):323–361.

Michaud, Richard O. 1989. "The Markowitz optimization enigma: Is "optimized" optimal?" *Financial Analysts Journal* 45(1):31–42.

Pantaleo, Ester, Michele Tumminello, Fabrizio Lillo and Rosario N Mantegna. 2011. "When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators." *Quantitative Finance* 11(7):1067–1080.

Pfaff, Bernhard. 2016. *Financial risk modelling and portfolio optimization with R.* John Wiley & Sons.

Ruppert, David and David S Matteson. 2015. *Statistics and Data Analysis for Financial Engineering: with R examples.* Springer.

Saucedo, Alejandro. 2014. Sparse and Group Regression Models in Portfolio Optimization PhD thesis University of Southampton.

Senneret, Marc, Yannick Malevergne, Patrice Abry, Gerald Perrin and Laurent Jaffres. 2016. "Covariance versus precision matrix estimation for efficient asset allocation." *IEEE Journal of Selected Topics in Signal Processing* 10(6):982–993.

Sharpe, William F. 1963. "A simplified model for portfolio analysis." *Management science* 9(2):277–293.

Sharpe, William F. 1964. "Capital asset prices: A theory of market equilibrium under conditions of risk." *The journal of finance* 19(3):425–442.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tibshirani, Robert, Martin Wainwright and Trevor Hastie. 2015. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC.

Tütüncü, Reha H and M Koenig. 2004. "Robust asset allocation." *Annals of Operations Research* 132(1-4):157–187.

Vasicek, Oldrich A. 1973. "A Note on Using Cross-sectional Information in Bayesian Estimation of Security Betas." *The Journal of Finance* 28(5):1233–1239.

Wealthfront. 2018. "Wealthfront Investment Methodology White Paper.".
URL: *https://research.wealthfront.com/whitepapers/investment-methodology/*

Yen, Yu-Min. 2015. "Sparse weighted-norm minimum variance portfolios." *Review of Finance* 20(3):1259–1287.

Zhao, Tuo, Han Liu, Kathryn Roeder, John Lafferty and Larry Wasserman. 2012. "The huge package for high-dimensional undirected graph estimation in R." *Journal of Machine Learning Research* 13(Apr):1059–1062.