Why estimation alone causes Markowitz portfolio selection to fail and what we might do about it

Why estimation alone causes Markowitz portfolio selection to fail and what we might do about it

Elmira Mynbayeva[a,1], John D. Lamb[a,*], Yuan Zhao[a]

[a]*University of Aberdeen Business School*
MacRobert Building
581 King Street
Aberdeen AB24 5UD

Abstract

Markowitz optimisation is well known to work poorly in practice, but it has not been clear why this happens. We show both theoretically and empirically that Markowitz optimisation is likely to fail badly, even with normally-distributed data, with no time series or correlation effects, and even with shrinkage estimators to reduce estimation risk. A core problem is that very often we cannot confidently distinguish between the mean returns of most assets. We develop a method, based on a sequentially rejective test procedure, to help remedy this problem by identifying subsets of assets indistinguishable in mean or variance. We test our method against naive Markowitz and compare it to other methods, including bootstrap aggregation, proposed to remedy the poor practical performance of Markowitz optimisation. We use out-of-sample and bootstrap tests on data from several market indices and hedge funds. We find our method is more robust than naive Markowitz and outperforms equally weighted portfolios but bootstrap aggregation works, as expected, better when we cannot distinguish among means. We also find evidence that covariance shrinkage improves performance.

Keywords:
Portfolio optimisation, Multivariate statistics, Homogeneous subsets, Estimation risk, Bootstrap aggregation

*Corresponding author
Email addresses: r02em16@abdn.ac.uk (Elmira Mynbayeva), J.D.Lamb@abdn.ac.uk (John D. Lamb), y.zhao@abdn.ac.uk (Yuan Zhao)

[1]Supported by JSC "Center for International Programs" – Bolashak

1. Introduction

It well known that Markowitz (1952) optimisation does not work well in practice. This is known as the 'Markowitz optimisation enigma' (Michaud, 1989). Many of the thousands of studies on Markowitz optimisation (Zhang et al., 2018) suggest reasons for this enigma, such as skewness, kurtosis, time-series effects, estimation error and estimation risk, and inappropriate optimisation problems, all of which undeniably have an effect. We show, first, that all that is required for Markowitz optimisation to fail, often spectacularly, is that we cannot distinguish with confidence between means or variances. Second, we develop methods (i) to identify when we can and cannot distinguish means and variances, and (ii) to modify Markowitz optimisation to work better with this information.

The following example should make this less abstract. Suppose we have just two assets. Standard Markowitz optimisation, with or without shrinkage (Section 1.1), assumes the difference between their sample means is known and fixed. In reality, the best we can do is estimate a distribution for it. If we are not confident that the means are different we should expect a minimum-variance portfolio to be at least as good as a Markowitz one. Similarly, if we are not confident the variances are different, assuming they are equal is unlikely to make Markowitz optimisation worse.

We generalise this idea. We develop a sequentially rejective multiple test procedure (Shaffer, 1986; Lamb and Tee, 2012) to identify subsets of assets that we are not confident differ in mean or variance. We call these homogeneous subsets. When comparing assets we use bootstrap methods so that we need not assume normal data. And we modify Markowitz optimisation to avoid distinguishing between means and variances within the homogeneous subsets.

The modified methods are simple. They use Markowitz assets when we can distinguish assets and variance minimisation or equally-weighted portfolio (*ewp* for short) when we cannot. This makes them insensitive to chance misestimation of means and variances. They can be combined with bootstrap aggregation (Frahm, 2015; Michaud and Michaud, 2007) or with robust optimisation methods (Fliege and Werner, 2014; Xidonas et al., 2020). We anticipate that they can also be combined with time-series methods and with shrinkage estimators, though both will need some further development.

1.1. Background

If we knew the means and covariance matrix of a set of asset returns exactly, Markowitz optimisation would select an optimal portfolio. But we never know

them exactly. So we must plug in estimates for them (Kan et al., 2007). Plug-in estimates work well in techniques such as regression and factor analysis. So it is natural to expect them to work for Markowitz optimisation too and attribute failure to complicating issues such as time-series effects, estimation risk and non-normality. We show failure is likely even in the absence of such issues.

Time-series effects are a complicating issue that can bias our estimates of means and variances. We remove them in Section 2 so that we can demonstrate the Markowitz optimisation enigma is due to inaccuracy rather than bias. And we ignore them in Sections 3 and 4 so that we can develop methods to deal with inaccuracy.

Inaccuracy in estimates is another complicating factor. We separate this inaccuracy into estimation error and estimation risk, though they are sometimes treated as if they were the same. Estimation error is uncertainty in estimates that, on average, are correct. Estimation risk is the expected value of the loss function (usually a mean-squared difference) between a population statistic such as the mean and an estimate of that statistic. Sample statistics do not minimise estimation risk when we apply them to a vector of as few as $n = 3$ assets (Stein, 1955). Estimation risk, increases with n and is worse when the range of statistics (e.g. means) is small. We wish to eliminate estimation risk, because it will lead to less than optimal portfolios. There are several ways to reduce estimation risk (Herold and Maurer, 2006). We use the shrinkage estimators for the mean (Jorion, 1986) and covariance matrix (Ledoit and Wolf, 2017), because they are the best developed. Another possibility is to restrict portfolio weights (Herold and Maurer, 2006). This is a cruder heuristic, and it may work well because it also reduces the effects of inaccuracy. The most extreme version of this heuristic is the ewp. It works surprisingly well in practice and we use it for comparison.

Non-normality is a complicating issue that affects accuracy. There are methods to handle skewness and kurtosis (Harvey et al., 2010; Kolm et al., 2014), coskewness and cokurtosis (Cerrato et al., 2017). Typically they modify the mean–variance framework. We do not. More complex frameworks may mask the causes of the Markowitz optimisation enigma: we show the enigma persists even in normal data. Nonetheless, non-normality affects how accurately we know the mean and variance of a portfolio and we seek ways to deal with it.

Inaccuracy in plug-in estimates creates two issues. First, we optimise over the data rather than the true means and covariance matrix. Section 2 shows that optimisation may be very sensitive to this misestimation. Second, even if we know the true means and covariances and have normal data, different portfolios (convex combinations of assets) differ widely in how accurately we may predict

their future means and variances. Then fuzzy or robust optimisation (Zhang et al., 2018) and the recent method of Meade et al. (2021) may help. We focus on the first problem and note both that what we do may reduce the second and that what we do may be combined with methods to deal with the second.

We note also an interesting method to deal with uncertainty in the mean and covariance estimates: the bootstrap aggregation method of Michaud and Michaud (2007, 1998). This method has limited theoretical justification. The problem is this. Even if bootstrap resampling accurately represents the range of possible true values of the means and covariance matrix, that does not guarantee that the resampled optimal portfolios accurately represent the range of future optimal portfolios or that their average is the best one. Frahm (2015) summarises well empirical studies showing mixed evidence for the performance of this method, but shows, under some weak assumptions, it will on average do no worse than the portfolio selection strategy to which it is applied. So we can, in principle, apply it to the methods we introduce and hope for further improvement. We test this.

2. Why estimation alone causes Markowitz portfolio selection to fail

We now demonstrate that there are circumstances where Markowitz optimisation must fail badly, even when there are no complicating factors such as non-normal data or time-series effects and when we use shrinkage estimators. We do this by considering the Markowitz-optimal solutions we should expect to find in cases where we know the true optimal solutions. Usually we assume here independently distributed assets with identical variances, because it is only then that we can easily derive results without simulation. We consider more realistic assumptions in Section 3.3, and Section 4 shows that assuming the means are close or identical is not very unrealistic, and that often variances are also close to identical.

Section 2.1 explains estimation risk and shrinkage estimators. Then Section 2.2 introduces the models we investigate. While the first model is unusual, it is the one we can use for analytic results in Scetion 2.3. The second model allows us to investigate, more plausibly, what happens when we try to minimise variance. Then Sections 2.4 and 2.5 use simulation to explore how Markowitz optimisation fails with and without shrinkage estimators.

2.1. Estimation risk and shrinkage estimators

Suppose we estimate a multivariate statistic $\theta = (\theta_1, \ldots, \theta_n)$ by a finite-sample estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$. We can define *estimation risk* by a risk function, usually

$$\mathcal{R}(\theta, \hat{\theta}) = \sum_{i=1}^{n} \mathbb{E}\left[\left(\theta_i - \hat{\theta}_i\right)^2\right].$$

Ideally we want $\mathcal{R}(\theta, \hat{\theta}) = 0$, in which case $\hat{\theta}$ is said to be *admissible*. But it has long been known (Stein, 1955; Fourdrinier et al., 2018) that even the sample mean is not admissible for $n \geq 3$. No estimator is known to be admissible in general and the best we can usually do is to reduce estimation risk as much as we can.

We distinguish estimation risk from *estimation error,* which is the error that arises because finite-sample estimators are not perfectly accurate. An admissible estimator will still have estimation error and estimation risk can be thought of as a multivariate form of bias. The problem with the usual definition of bias,

$$b(\theta, \hat{\theta}) = \mathbb{E}\left[\theta - \hat{\theta}\right] = \left(b(\theta_1, \hat{\theta}_1), \ldots, b(\theta_n, \hat{\theta}_n)\right),$$

is that it ignores the very real possibility that data from which θ_i is estimated might also contain information about θ_j.

To reduce estimation risk we use the following estimator (Jorion, 1985, 1986) for the (vector) mean of n assets observed over T periods:

$$\mu_J = \frac{\varphi}{\varphi + T}\mu_0 \mathbf{1} + \frac{T}{\varphi + T}\hat{\mu}, \tag{1}$$

where $\hat{\mu}$ is the vector of sample mean returns, μ_0 is the mean of $\hat{\mu}$ and $\mathbf{1}$ is a vector of ones of length n. The parameter φ is estimated as

$$\varphi = \frac{n + 2}{(\hat{\mu} - \mu_0 \mathbf{1})^\top \tilde{\Sigma}^{-1} (\hat{\mu} - \mu_0 \mathbf{1})},$$

where $\tilde{\Sigma}$ is an estimate of the covariance matrix of the asset returns, which estimates the minimum-variance portfolio.

Notice that equation (1) is a weighted sum of the mean of sample means (left) and the sample mean (right). Thus it shrinks the sample mean estimate towards the mean of sample means, but does not make two means identical unless $\phi = 0$, which is unlikely. In general, standard estimators tend to underestimate the smallest, and overestimate the largest, statistic, and so estimators that reduce this misestimation are often called *shrinkage estimators.*

Jorion (1986) uses the sample covariance matrix $\hat{\Sigma}$ to estimate $\tilde{\Sigma}$. The justification for this choice is that the sample covariance matrix is more stable than the sample mean vectors (Merton, 1980). However, high-dimensional covariance matrices can be singular or inversion may enhance estimation error. So we also estimate $\tilde{\Sigma}$ using the linear and nonlinear shrinkage estimators of Ledoit and Wolf (2004) and Ledoit and Wolf (2017).

The linear shrinkage estimator shrinks the sample covariance matrix towards the identity matrix. Ledoit and Wolf (2004) give the details, which are easy to implement in `R`. The nonlinear estimator works by shrinking the estimates of the sample covariance eigenvalues. The method, outlined in Ledoit and Wolf (2017), is complex, but is implemented in the `nlshrink` `R` package.

Notice that estimation risk affects not just mean and variance, but any statistic we might wish to estimate, including those used to generate the box and ellipsoidal uncertainties used in robust optimisation (Fliege and Werner, 2014; Xidonas et al., 2020). So while robust optimisation deals well with estimation error, we still need to reduce estimation risk to use it more effectively. This creates two problems. First, good shrinkage estimators are only known for a few statistics such as mean and variance. Second, no current shrinkage estimators make population statistic estimates identical when it may be best to assume that they are identical—see Section 2.3.

2.2. Markowitz optimisation problems

We consider a range of Markowitz optimisation problems. Suppose we have assets $\mathbf{a} = (a_1, \ldots, a_n)^\top$. Write $\hat{\mu}$ for the sample mean return, a vector of length n, and write $\hat{\Sigma}$ for the $n \times n$ sample covariance matrix. Then we wish to choose a vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$ of portfolio weights: that is, w_i is the proportion invested in a_i and $\mathbf{w}^\top\mathbf{a}$ is the *portfolio* or *virtual asset.* We assume we neither leave some amount uninvested nor use short-selling. So

$$\mathbf{w}^\top\mathbf{1} = 1 \quad \text{and} \quad \mathbf{w} \geq \mathbf{0}, \tag{2}$$

where $\mathbf{1}$ is the vector of length n all of whose entries are 1.

We consider the following general Markowitz optimisation problem.

$$\max_{\mathbf{w}} \quad \hat{\mu}^\top\mathbf{w} - R\mathbf{w}^\top\hat{\Sigma}\mathbf{w}, \quad \text{subject to constraints (2)}, \tag{3}$$

where $R \geq 0$, the *coefficient of risk tolerance* determines our attitude to risk. Choosing $R = 0$ and $R = 10000$, we get the portfolio selection strategies *max* and *min,* which seek to maximise return or minimise risk. We also consider two

other Markowitz optimisation problems. The first, which we call *min-c,* seeks the lowest variance portfolio with a minimum required return b:

$$\min_{\mathbf{w}} \quad \mathbf{w}^\top \hat{\Sigma} \mathbf{w}, \quad \text{subject to } \hat{\mu}^\top \mathbf{w} \geq b \text{ and constraints (2).} \tag{4}$$

The second, which we call *max-c,* seeks the highest mean portfolio with a maximum variance b:

$$\max_{\mathbf{w}} \quad \hat{\mu}^\top \mathbf{w} \quad \text{subject to } \mathbf{w}^\top \hat{\Sigma} \mathbf{w} \leq b \text{ and constraints (2).} \tag{5}$$

Setting $w_i = 1/n$ for $i = 1, \ldots, n$ gives an equally-weighted portfolio, satisfying constraints (2). The ewp is a popular asset allocation strategy (Benartzi and Thaler, 2001) often used as a benchmark to compare other strategies (Hwang et al., 2018; DeMiguel et al., 2009).

2.3. Theoretical considerations

We now consider some of cases where we can show why and how badly Markowitz optimisation should fail. We assume here that we have independent normally distributed assets with variance $\sigma^2 = 4$ observed over $T = 100$ time periods.

Suppose first that the n assets have identical mean return. We choose a value $\mu = 2$, because that makes it easier to show some effects in Figure 1 (left). But the value does not affect the analysis. Then, writing $\overline{X}_k$ for the random variable describing the mean return of the kth asset, we have $\overline{X}_k \sim N(\mu, \sigma^2/T)$. And any reasonable Markowitz optimisation should give the same solution: an ewp with mean return μ and variance σ^2/n^2.

Suppose we try to maximise mean with no constraint on variance other than that it is minimum subject to mean being maximised. Since $\sigma^2/T > 0$, $\mathbb{P}(\overline{X}_j = \overline{X}_k) = 0$ $(j, k = 1, \ldots, n,\ j \neq k)$. So, with probability 1 we will choose a portfolio with exactly one asset. The expected value of the observed variance of this portfolio will be σ^2. And we can estimate the expected value of the mean of this portfolio using order statistics.

An *order statistic* (David, 2008) $\overline{X}_{(k)}$ $(k = 1, \ldots, n)$ is the kth smallest of a set of n observations from the same distribution. And it is well known that if the distribution function is F then

$$\mathbb{E}\left[\overline{X}_k\right] = k\binom{n}{k} \int_0^1 F^{-1}(u) u^{k-1} (1-u)^{n-k} \, \mathrm{d}u. \tag{6}$$

In practice, this integral usually needs to be evaluated numerically and we do so using the QAG method of the GNU Scientific Library (Galassi et al., 2009). The expected value of the mean of the portfolio is $\mathbb{E}[\overline{X}_n]$ and we show this value as the solid line in Figure 1 (left) for increasing values of n. The dotted line shows the true mean return of 2. We may conclude that if optimisation is dominated by mean maximisation, we should expect a portfolio with few assets and substantially overestimated future returns for that portfolio.

We can use equation (6) to estimate the expected value kth smallest observed asset standard deviation S_k, because the asset variances have sampling distribution $\sigma^2/(T-1)\,\chi^2_{T-1}$. We plot $S_{(1)}$ and $S_{(n)}$ as dashed lines on Figure 1 (left) to show how the range of standard deviations increases with n. However, the optimal solution of a minimum variance portfolio depends on the whole covariance matrix. Although we know this follows a Wishart distribution (Eaton, 2007), we do not know of a way to use this to calculate the expected value of the minimum variance portfolio for $n > 2$. (For $n = 2$ we can use formulae given by (Nadarajah and Kotz, 2008)). We can, however, reasonably conclude that optimisation based on sample data will choose unequal weights, give a portfolio that does not minimise variance and underestimate the true minimum variance of σ^2/n.

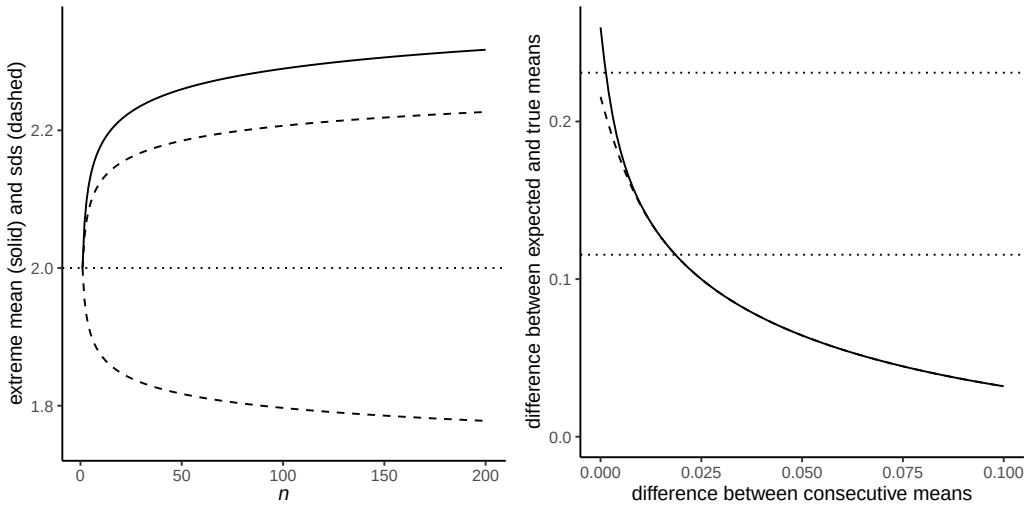

Figure 1 Charts showing effects of using sample rather than true statistics

We have assumed so far all assets have the same mean return. In practice, they may be similar but are identical with probability 0. So we may reasonably ask how well Markowitz optimisation will do if the means are different. We can generalise equation (6) to answer this, though will only elaborate on the estimate of $\overline{X}_{(n)}$,

because that is what gets chosen in mean maximisation. Suppose the means are given by a vector $\mu = (\mu_1, \ldots, \mu_n)$ and the distributions of the asset returns by $F_1, \ldots, F_k$. Then

$$F_{(n)}(x;\mu) = \mathbb{P}\left(\overline{X}_n < x;\mu\right) = \prod_{j=1}^{n} F_j(x;\mu_j).$$

We differentiate this to get the density function

$$f_{(n)}(x;\mu) = \sum_{i=1}^{n} f_i(x;\mu_i) \prod_{j=1}^{n} \delta_{ij} F_j(x;\mu_j),$$

where $f_i(x;\mu_i)$ is the density of the mean of the ith asset and δ_{ij} is the Kronecker delta. Then

$$\mathbb{E}\left[\overline{X}_{(n)}\right] = \int_{-\infty}^{\infty} \sum_{i=1}^{n} f_i(x;\mu_i) \prod_{j=1}^{n} \delta_{ij} F_j(x;\mu_j)\, \mathrm{d}x.$$

As before, this integral must be evaluated numerically when the $\overline{X}_k$ are normally distributed. Figure 1 (right) shows for $n = 20$ (dashed) and $n = 50$ (solid) that the amount that $X_{(n)}$ overestimates the μ_n when when $\mu_1 \leq \cdots \leq \mu_n$ are equally spaced with the horizontal axis showing the difference between consecutive values of μ_k. The dashed lines are at 1 and 2 standard deviations. Unless the means are more dispersed than we usually see in practice, we expect mean maximisation to noticeably overestimate the expected mean return with some risk of not even choosing the asset with true maximum return.

We note three consequences of these observations. First, the bias in the estimate of the optimum mean is an increasing function of n. Second, since shrinkage estimators for the mean such as equation (1) do not shrink to a global mean, they will not change the optimal portfolio but should reduce the bias in the estimated portfolio mean. Third, Section 4.1 shows that it is common not to be able to distinguish mean asset returns. Then max should do little better than choose a single asset randomly and the method of Michaud and Michaud (2007) should choose a portfolio very close to the ewp.

We might hope to consider the min or min-c strategies in a similar way. However, we do not even know how to solve them numerically. So we now consider simulation.

2.4. Simulation with equal means and variances

We continue considering the case of n independently distributed assets, each with distribution $N(\mu, \sigma^2)$, for which the max, min and min-c strategies should all give the same solution as ewp. To use simulation, we must choose values for n and the number of time periods T to sample over, besides those of μ and σ^2. We choose $T = 300$ because that is typical of the number of months for which we can get real data. We simulate two cases, $n = 20$ and $n = 50$, so that we can see the likely effects of increasing the number of assets used for optimisation. We choose $\mu = 1$ and $\sigma^2 = 4$ and note that the optimal coefficients are independent of this choice. We do not test the min-c strategy here, because it only makes sense when the means are different. For every strategy the (true) optimal portfolio has mean 1 and standard deviation 0.447 ($n = 20$) or 0.283 ($n = 50$).

We choose $T = 300$ and $\sigma^2 = 4$ because these are of the order of magnitude of the parameters available in real data, such as in Section 4. Notice that we do not immediately need $T = 300$, because we could simulate samples from $N(\mu, \sigma^2/T)$. But we need $T > 1$ for bootstrap aggregation at then end of this subsection and to generate and simulate the effects of homogeneous subsets in Section 3.

We simulate returns for n assets and record the optimal portfolio weights given by the max and min Markowitz strategies applied to the simulated data. We then compute the population and sample (from simulated data) portfolio mean and portfolio standard deviation given the optimal portfolio weights from each strategy and also given the ewp portfolio weights.

Table 1 Summary results for simulated data

	$n = 20$			$n = 50$		
	ewp	max	min	ewp	max	min
μ	1	1	1	1	1	1
$\overline{x}$	1	1.215	0.998	1	1.259	0.998
σ	0.447	2	0.461	0.283	2	0.308
$\overline{s}$	0.447	1.989	0.429	0.283	2	0.259

Table 1 summarises the average over 100 simulations of the population portfolio mean μ, sample portfolio mean $\overline{x}$, population portfolio standard deviation σ, and sample portfolio standard deviation $\overline{s}$.

Markowitz optimisation fails in every case, even with normal data and no time series effects. The max strategy sample mean overestimates the true mean. The min strategy sample standard deviations underestimate the true minimum standard

deviations, and their optimal portfolios would on average have standard deviation noticeably higher than the known minimum. The results also get worse as we go from $n = 20$ to $n = 50$.

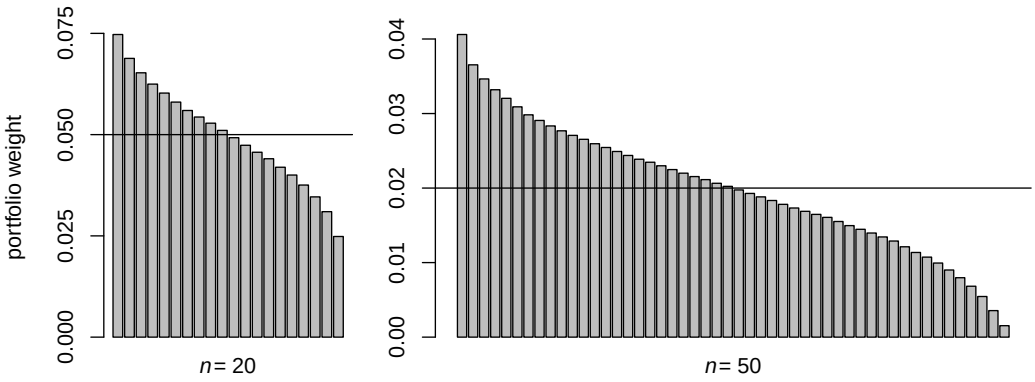

Figure 2 Average sorted min portfolio weights for simulated data

In every case max gives a portfolio with exactly one asset, as Section 2.3 predicts. Figure 2 shows the average sorted portfolio weights for min. The horizontal lines shows the weights we should get if Markowitz found the optimal (ewp) solutions. Clearly what we should get and what we do get are very different: Markowitz fails to diversify enough when we have maximised the opportunity for diversification.

We note briefly how a bootstrap aggregation method, like that of Michaud and Michaud (2007), performs on the simulated data. We expect it to perform better than Markowitz optimisation, because the data satisfies the assumptions of Frahm (2015). We use the bootstrap data generating process of Lamb and Tee (2012), which we can summarise as follows. We let r_{ti} be the return of asset i in time period $t \in \{1, \ldots, T\}$ and for some $\tau \leq T$ we write R^{in} for the $\tau \times n$ matrix with (t, i)th entry r_{ti}. Then we generate B replications of R^{in} as follows.

for $b = 1, \ldots, B$:
  for $t = 1, \ldots, \tau$:
    select $u(t, b)$ uniformly at random from $\{1, \ldots, \tau\}$
let R^{in}_b be the $\tau \times n$ matrix whose (t, i)th entry is $r_{u(t,b),i}$

We use $B = 2000$ (a common choice) bootstrap replications. We use $\tau = T$ here and report the We use the same data generating process later, both when generating homogeneous subsets in Section 3.1 and for bootstrap tests in 4.3, where R^{in} is in-sample data and we set $\tau = T/2$. performance of the average of the average of the

B vectors of portfolio weights. We summarise the results in Table 2. The bootstrap aggregation method performs markedly better. We also find the portfolio weights are much closer to ewp.

Table 2 Summary results for simulated data: bootstrap aggregation

	$n = 20$			$n = 50$		
	ewp	max	min	ewp	max	min
$\overline{x}$	1	1.003	1.002	1	1.005	1.005
$\overline{s}$	0.447	0.438	0.434	0.283	0.298	0.295

Notice that we sample τ rows of R^{in} randomly with replacement to get each R_b^{in} so that we preserve the correlation between assets. This is not necessary here, but it is helpful later, when we use the same data generating process to generate homogeneous subsets in Section 3.1 and for bootstrap tests in 4.3. There we set $\tau = T/2$ so that R^{in} is the matrix of returns of an in-sample subset of the data. We do not consider here correlation of returns time, but note that can be done using, for example, the maximum-entropy bootstrap method of Vinod (2004).

2.5. Simulation with shrinkage estimators

Sections 2.3 and 2.4 show that we cannot attribute the failure of Markowitz optimisation to non-normality or time series effects. But both real data and simulated data are subject to estimation risk (Jorion, 1986). Section 2.3 indicates that shrinkage estimators (Jorion, 1986; Ledoit and Wolf, 2004, 2017) will not change the portfolio the max strategy selects but may reduce the bias in the portfolio mean. We now investigate how much these shrinkage estimators improve portfolio selection.

We simulate data exactly as in Section 2.4 to test the effect of shrinkage estimators. We test either without shrinkage estimators for the mean or with Jorion (1986) mean shrinkage (J). And we test three possibilities for the covariance matrix: the sample covariance ($\hat{\Sigma}$), or the linear (LW) or nonlinear (NL) shrinkage covariance matrices of Ledoit and Wolf.

Table 3 summarises the results of the tests. Each row summarises the result of 100 tests with different random numbers. The first two columns shows the number of assets simulated and the optimisation strategy. As in section 2.4 we omit min-c, because all assets have (true) mean 1. The remaining columns show the shrinkage estimators used, the average sample portfolio mean and standard

Table 3 Summary simulation results with shrinkage estimators

n	Strategy	Shrinkage	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$
	ewp		1.000	0.447	20	0.05	0.05
	max	J, NL	1.211	3.929	1.03	0	0.99
20		J, $\hat{\Sigma}$	1.002	0.434	19.99	0.026	0.075
	min	J, LW	0.998	0.446	20	0.049	0.051
		J, NL	1	0.447	20	0.048	0.052
	ewp		1	0.283	50	0.02	0.02
	max	J, NL	1.260	1.985	1.02	0	0.993
50		J, $\hat{\Sigma}$	1.005	0.257	43.16	0.001	0.04
	min	J, LW	0.998	0.283	50	0.02	0.02
		J, NL	0.996	0.283	50	0.02	0.02

deviation, the average number of assets with non-negligible portfolio weights (m), and the average smallest ($w_{\min}$) and largest ($w_{\max}$) portfolio weights.

For max we omit the cases where we use $\hat{\Sigma}$ or LW: their results are negligibly different from NL, as we expect. Table 1 shows the case where we use a sample mean instead of J. It is striking that mean shrinkage produces no improvement in the max strategy.

For min we omit the cases where we tested the sample mean instead of J. Again, as expected, the results are negligibly different. Also as expected, the combination J, $\hat{\Sigma}$ gives very similar results to those of Table 1 and Figure 2. However, the results for the combinations J, LW and J, NL are remarkably similar and very close to the true optimal ewp. At least in this case shrinkage has little effect on mean maximisation but is very effective in variance minimisation.

The rows labelled ideal, naive and J, NL of Table 6 show results comparable to those of Table 3, further supporting our observation that variance shrinkage is more effective than mean shrinkage.

3. How we might deal with the failure of Markowitz optimisation in practice

3.1. Homogeneous subsets

Multiple comparison procedures (Holm, 1979) help us identify homogeneous subsets. If we have a single comparison (e.g. between two assets) we can use a hypothesis test and test the null hypothesis that the assets are identical in some

statistic at a preselected significance level α. Suppose that we wish to make multiple comparisons with hypothesis tests $H_1, \ldots, H_m$ (usually we have $m = \binom{n}{2}$). We can write H_0 for the null hypothesis that all of $H_1, \ldots, H_m$ are true. Then we wish to estimate the probability α of rejecting at least one of $H_1, \ldots, H_m$ at significance level α' given that H_0 is true. We wish to choose α, the *familywise signicance level* and estimate α_t, the *experimentwise significance level*. We can use either the well-known Bonferroni (left) correction or the Šidák (1967) (right) equation with $t = m$:

$$\alpha_t \approx \alpha/t \approx 1 - (1 - \alpha)^{1/t}. \tag{7}$$

The Bonferroni correction gives slightly smaller values of α_t.

So far we have shown how to make one comparison. Suppose $p_1, \ldots, p_n$ are the p-values of $H_1, \ldots, H_n$ at significance level α'. Then, using the order-statistic notation of Section 2.3, we write $p_1, \ldots, p_n$ in increasing order as $p_{(1)}, \ldots, p_{(n)}$ with $H_{(i)}$ the null hypothesis matching $p_{(i)}$. Our comparison is to reject H_0 (and so $H_{(1)}$) if $p_{(1)} > \alpha_t$. But we wish to identify all the hypotheses we can reject at familywise significance level α. We can do this using the procedure of Shaffer (1986), which provides proofs and further details.

Suppose we have already rejected $H_{(1)}, \ldots, H_{(k-1)}$. Initially we put $t = m$, because we could suppose all of $H_1, \ldots, H_m$ were true. But once we reject some of them, not all the remaining hyotheses may be true. For example, if we are comparing the means μ_1, μ_2 and μ_3 of three assets and reject the hypothesis that $\mu_1 = \mu_2$, then we may have either $\mu_1 = \mu_3$ or $\mu_2 = \mu_3$ but not both. Shaffer (1986) deals with this by setting $t = t(k)$ ($k > 1$) to be the maximum number of null hypotheses that may be true given that $k - 1$ are false and shows how to calculate $t(k)$. If $t(k) = 0$ there are no further possible null hypotheses. Otherwise we keep rejecting null hypotheses while $p_{(k)} < \alpha_{t(k)}$. This gives us the multiple comparisons.

We go from multiple comparisons to homogeneous subsets by defining a strict partial order $\prec$ over the set $\theta_1, \ldots, \theta_n$ of statistics we are comparing. This order must satisfy (i) $\theta_i \not\prec \theta_i$ (asymmetry), (ii) if $\theta_i \prec \theta_j$ then $\theta_j \not\prec \theta_i$ (irreflexivity) and (iii) if $\theta_i \prec \theta_k$ and $\theta_k \prec \theta_j$ then $\theta_i \prec \theta_k$ (transitivity). We define $\prec$ by $\theta_i \prec \theta_j$ if either $\theta_i < \theta_j$ and we have rejected the hypothesis that $\theta_i = \theta_j$ or $\theta_i \prec \theta_j$ is implied by transitivity. We need the transitivity rule because we have not proved that it is not necessary and we conjecture that it is, in principle, necessary and suggest that this might be proved as a straightforward generalisation of Arrow's paradox Arrow (1950). In practice we have not needed the transitivity rule.

Once we have defined $\prec$ we define $\theta_i \sim \theta_j$ if neither $\theta_i \prec \theta_j$ nor $\theta_j \prec \theta_i$: that is if we have not rejected the null hypothesis that $\theta_i = \theta_j$ either explicitly or through transitivity. An antichain is a subset A of $\{1, \ldots, n\}$ such that $\theta_i \sim \theta_j$ for all $i, j \in A$. And a subset of $\{1, \ldots, n\}$ is homogeneous in θ if it is a maximal antichain: that is if it is an antichain and not a proper subset of any other antichain.

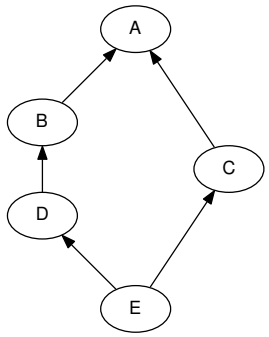

Figure 3 Hasse diagram showing choice of homogeneous subsets

Notice that the definition of homogeneous subsets guarantees that every asset must be in a homogeneous subset (we always have $\theta_i \sim \theta_i$), but not that it be in only one. That is, the homogeneous subsets do not necessarily partition the set of assets. Figure 3 illustrates how an asset can be in more than one homogeneous subset. It shows a Hasse diagram over a set of assets A–E: if $\theta_i \prec \theta_j$ ($i, j \in \{A, B, C, D, E\}$) then i is lower than j in the diagram and there is a path from i to j. The homogeneous subsets are $\{A\}$, $\{B, C\}$, $\{D, C\}$ and $\{E\}$. These do not partition the assets because C is contained in two of them.

The method we develop in Section 3.2 needs homogeneous subsets that partition the set of assets. In the example, we can obtain a partition by choosing to put C into only one homogeneous subset. This gives two possible solutions: $\{A\}, \{B, C\}, \{D\}, \{E\}$ and $\{A\}, \{B\}, \{C, D\}, \{E\}$. In general we obtain a partition of the set of assets into homogeneous subsets by repeatedly adding adding to the partial order a relation $\theta_i \prec \theta_j$ (and any that follow by transitivity) with smallest p-value satisfying (i)–(iii) of the following proposition.

Proposition 1. *Suppose we have homogeneous subsets that do not partition the set* $\{1, \ldots, n\}$. *Then there exits* $i, j, k \in \{1, \ldots, n\}$ *such that (i)* $\theta_i \sim \theta_j$, *(ii)* $\theta_i \sim \theta_k$ *and (iii) either* $\theta_j \prec \theta_k$ *or* $\theta_k \prec \theta_j$.

Proof. Choose I maximal such that $I = A \cap B$ for some pair A and B of distinct homogeneous subsets. Then $I \neq \emptyset$, because the homogeneous subsets do not

partition $\{1, \ldots, n\}$, and $\theta_i \sim \theta_{i'}$ for $i, i' \in I$. Since A and B are distinct maximal antichains, there exist $j \in A \setminus B$ and $k \in B \setminus A$. Since $I \subset A$, $\theta_j \sim \theta_i$ for $i \in I$; and since $I \subset B$, $\theta_k \sim \theta_i$ for $i \in I$: that is, (i) and (ii) hold. If (iii) does not hold, then $\theta_j \sim \theta_k$ and $I' = I \cup \{j, k\}$ satisfies $\theta_i \sim \theta_{i'}$ for $i \in I'$, contradicting the maximality of I. □

Notice that Proposition 1 guarantees we can find a relation to add to the partial order and that repeating the process must give homogeneous subsets that partition $\{1, \ldots, n\}$, because each step either finds a partition or adds one of finitely many remaining relations. While it is not essential that we choose a relation satisfying (i)–(iii), it helps us avoid creating more homogeneous subsets than necessary.

In the example, the process will add whichever of $D \to C$ or $C \to B$ has smallest p-value corresponding to its hypothesis test.

Typically multiple comparison procedures use t-tests to compare pairs of means (assume normality) and assume the tests are independent. These assumptions do not hold for assets. So, as Lamb and Tee (2012) suggest, we use bias-corrected accelerated bootstrap tests (Efron and Tibshirani, 1998). If asset i has unknown mean μ_i and variance σ_i these allow us to construct subsets homogeneous in μ_i, σ_i or $\mu_i - R\sigma_i^2$. The last of these is sensible in the more general model (3).

We use the bootstrap data generating process described at the end of Section 2.4 to generate replications $R_1, \ldots, R_B$ of the asset returns. And for each estimator (e.g. mean) θ of interest, we estimate the empirical distribution of θ_i ($i = 1, \ldots, n$) as $\theta_{i,1}, \ldots, \theta_{i,B}$. Thus we get paired estimates $\theta_{i,b}, \theta_{j,b}$ for each pair i, j in $\{1, \ldots, n\}$ and so can compare paired differences to get a more powerful test that does not ignore the correlation between asset returns. We use $B = 2000$, a typical value for the bootstrap. The the choice of B has less impact on the selection of homogeneous subsets than the value T, which depends on the available data. If T is too small, we cannot use the bootstrap at all, while the number of homogeneous subsets tends to n as $T \to \infty$: the more data we have the more confident we are that observed differences are real.

We use equation (7) (right) in the multiple comparison procedure, though this has little effect on the choice of homogeneous subsets. The procedure is, however, sensitive to the choice of experimentwise significance level α. Since our purpose is to test rather than calibrate a procedure we do not investigate this further here and set $\alpha = 0.05$, a common choice.

Note that estimation risk does does not affect pairwise comparisons (Stein, 1955) and the multiple comparison procedure is designed not to reject differences due to the risk that shrinkage is designed to reduce. That is, we can think of

homogeneous subsets as an alternative to shrinkage.

3.2. An optimisation heuristic with homogeneous subsets

Suppose we wish to use model (3) or (4) to optimise a portfolio over assets $A = \{a_1, \ldots, a_n\}$. We cannot assume the asset means and covariance matrix are known accurately enough for Markowitz optimisation to work. So we propose Heuristic 1 instead of pure Markowitz optimisation.

(a) Choose a familywise significance level α and statistic θ for homogeneous subsets and find subsets of $H_1, \ldots, H_k$ homogeneous in θ
(b) For $i = 1, \ldots, k$, for the assets $\{a_j : \ a_j \in H_i\}$, let $\overline{a}_i$ be either
  (ewp) the equally-weighted portfolio
  (min) the minimum-variance portfolio
(c) Solve model (3) or (4) over virtual assets $\overline{a}_1, \ldots, \overline{a}_k$

Heuristic 1 Optimisation with homogeneous subsets

Note that when the assets are independent and identically distributed, heuristic 1 reduces to choosing the ewp. And if the means and standard deviations are known exactly it reduces to Markowitz optimisation. In both cases it selects the portfolio we know to be optimal.

We have ignored here time dependency in asset returns. We noted in Section 2.4 that we could substitute the bootstap method we use with a time-series bootstrap method. Early tests, using the maximum-entropy bootstrap method of Vinod (2004) combined with ARIMA and GARCH suggest that when we account for time-series effects, we sometimes get a small increase in the number of homogeneous subsets for a data set. This is what we should expect.

3.3. Application of homogeneous subsets to simulated data

We test Heuristic 1 on simulated data to check that it performs as expected when we know the optimal solution to a problem. As in Sections 2.4 and 2.5 we test sets of $n = 20$ and $n = 50$ assets with independent normally distributed returns and generate data sets of $T = 300$ periods. We choose familywise significance $\alpha = 0.05$ in step (a) and test here the ewp variant in step (b).

We consider first the case where all means are 1 and standard deviations 0.5. In this case the optimal portfolio for all strategies is the ewp with portfolio mean 1 and standard deviation 0.112 ($n = 20$) or 0.071 ($n = 50$). We compute subsets homogeneous in mean for max and in standard deviation for min.

Table 4 Summary simulation results with homogeneous subsets

n	Strategy	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$
	ewp	1.000	0.112	20	0.05	0.05
20	max	1.001	0.113	19.65	0.042	0.511
	min	1.000	0.112	20	0.050	0.052
	ewp	1	0.071	50	0.02	0.02
50	max	1.003	0.173	37.25	0.009	0.24
	min	0.999	0.071	50	0.019	0.021

Table 4 summarises the results of the tests. It shows averages over 100 simulations of the portfolio mean and standard deviation, number of assets with non-negligible weights and smallest and largest portfolio weight. The min strategy performs well, as it did with shrinkage estimators (Table 3) but not with standard Markowitz optimisation (Table 1). The improvement in max is, however, striking. It has excellent performance for $n = 20$ and good for $n = 50$.

What is not obvious in the table is that the results for max are mixed. When $n = 50$, 25 of the tests optimised over homogeneous subsets of one or two assets and the rest optimised over homogeneous subsets of 46–50 assets. In this extreme case we can improve performance by increasing the number of bootstrap replications or the experimentwise significance used to generate the homogeneous subsets.

In practice we expect some variation in asset mean values. So we test this case. Specifically, we test cases where there are 4 or 10 assets with each of the mean values in $\{1, \ldots, 5\}$. We simulate $T = 300$ normally distributed independent random variates, all with standard deviation 0.05 as before. We test the max and min strategies and also a new strategy cmax-24, which is model (3) with $R = 24$. The ewp is no longer optimal, but we can compute the true optimal portfolio in each case. We label it as ideal in Table 5.

We compare min, max and cmax-24 also using optimisation with the shrinkage estimators of Section 2.5. Table 5 summarises the results averaged over 100 simulations. Column HS is the number of homogeneous subsets found. As before we use subsets homogeneous in mean for max and standard deviation for min. For cmax-24 we use subsets homogeneous in

$$\text{mean} - 24 \times \text{variance},$$

because this is what model (3) minimises.

We find that using homogeneous subsets works very well for max and min

Table 5 Results for simulated data with inhomogeneous means

n	Strategy	Method	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS
		HS	5.000	0.256	3.95	0.000	0.254	5.2
	max	J, NL	5.029	0.501	1.15	0.000	1.000	
		ideal	5	0.25	4	0	0.25	5
		HS	3.000	0.114	20	0.050	0.052	1.1
20	min	J, NL	3.000	0.113	20	0.048	0.051	
		ideal	3	0.112	20	0.05	0.05	1
		HS	4.314	0.168	10.25	0.000	0.115	2.6
	cmax-24	J, NL	4.660	0.187	10.2	0.000	0.170	
		ideal	4.667	0.186	8	0	0.167	5
		HS	4.999	0.161	9.85	0.000	0.102	5.4
	max	J, NL	5.043	0.500	1	0.000	1.000	
		ideal	5	0.158	10	0	0.1	5
		HS	3.000	0.071	50	0.019	0.025	1.7
50	min	J, NL	2.999	0.071	50	0.019	0.021	
		ideal	3	0.071	50	0.02	0.02	1
		HS	4.503	0.118	18.35	0.000	0.056	3.0
	cmax-24	J, NL	4.914	0.146	20	0.000	0.095	
		ideal	4.917	0.146	20	0	0.092	5

strategies, as in Table 4. We also find the shrinkage estimators work well for min but not max, as they did in Table 3. By contrast, shrinkage estimators work well for cmax-24.

These results suggest that using homogeneous subsets for means and a shrinkage estimator for variances should work well. Since mean maximisation is unusual, we test this with strategies min-c and max-c (problems (4) and (5)). We use bounds 4 for min-c and 0.625 for max-c. Table 6 summarises the results in the same format as Table 5. Where the method is labelled standard, we use Markowitz optimisation without shrinkage or homogeneous subsets. Although shrinkage estimators work well for min-c, it is notable that the combination of homogeneous subsets for means and a shrinkage estimator for variances works well in all cases, giving lower standard deviation when $n = 50$ and better estimates of the optimal coefficients when $n = 20$ or 50.

Table 6 Further results for simulated data with inhomogeneous means

n	Strategy	Method	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS
20	min-c	standard	4.000	0.134	17.1	0.000	0.117	
		HS, NL	4.000	0.137	17.65	0.000	0.100	5.28
		J, NL	4.000	0.137	17.43	0.000	0.101	
		ideal	4	0.137	16	0.000	0.1	5
	max-c	standard	4.994	0.250	4.68	0.000	0.314	
		HS, NL	4.996	0.249	5.57	0.000	0.250	5.2
		J, NL	4.997	0.250	4.77	0.000	0.274	
		ideal	5	0.25	4	0	0.25	5
50	min-c	standard	4.000	0.802	40.1	0.000	0.058	
		HS, NL	4.000	0.087	42.58	0.000	0.040	6.08
		J, NL	4.000	0.087	42.73	0.000	0.041	
		ideal	4	0.087	40	0	0.04	5
	max-c	standard	5.029	0.250	6.04	0.000	0.371	
		HS, NL	5.002	0.162	9.56	0.000	0.106	6.32
		J, NL	5.028	0.25	6.15	0.000	0.361	
		ideal	5	0.158	10	0	0.1	5

4. Empirical results

We now investigate the performance of the homogeneous subsets methods on real data. We consider the max-c, min-c and cmax strategies used in Section 3.3. We no longer know what the optimal portfolios should be; so we use an out-of-sample test (DeMiguel et al. (2009), Hwang et al. (2018)) to compare different strategies combined with our method. That is, we construct portfolios using the first half of each sample of asset returns and compare how well the portfolios perform on the second half. As before, we consider returns of samples of $n = 20$ and $n = 50$ assets.

4.1. Data and construction of homogeneous subsets

The data we use are monthly percentage returns of random samples of 20 or 50 stocks from the S&P 500, Nikkei 225, FTSE 100, and DAX market indices, obtained from Datastream (2019) None of these sets gave us more than one homogeneous subset of means. So we also use random samples of 20 and 50 monthly percentage returns of hedge funds obtained from Refinitiv (2018).

Table 7 summarises the data: n is the number of stocks or funds and T is the number of monthly returns available. We want stock data for as long as possible and have comparable data for 300 months for the first three indices. The DAX index comprises 30 major German companies. Since only 18 of them have data for 300 months and we need at least 20 we select the 20 oldest assets.

Table 7 Summary of data

Data set	Mean	Sd	Skewness	Kurtosis	n	T	Time span
FTSE	0.837	8.151	0.131	0.151	61	300	30/05/1994 – 30/04/2019
S&P 500	1.207	9.143	0.226	0.796	326	300	02/06/1994 – 02/05/2019
DAX	0.923	9.033	0.085	−0.186	20	280	01/02/1996 – 01/05/2019
Nikkei 225	0.577	9.805	0.370	−0.294	184	300	01/06/1994 – 01/05/2019
HF	0.531	3.711	−0.386	1.580	375	168	31/01/2005 – 31/12/2019

We construct homogeneous subsets using Heuristic 1(a) with $\alpha = 0.05$ and θ either sample mean or sample standard deviation using samples of $n = 20$ and (where available) $n = 50$ assets from each of the five data sets. Table 8 summarises the homogeneous subsets obtained from the first half of each sample. The first column indicates the data set used. The remaining columns show the number of subsets obtained that are homogeneous in mean or in standard deviation (sd).

Table 8 Homogeneous subsets

	Mean	Sd	Mean	Sd
	$n = 20$		$n = 50$	
FTSE	1	4	1	4
S&P	1	4	1	4
DAX	1	4	–	–
Nikkei	1	3	1	6
HF	2	6	3	7

Figure 4 show the means and standard deviations of the stocks used to create homogeneous subsets using different symbols for different homogeneous subsets. We show only subsets homogeneous in standard deviation because in every case

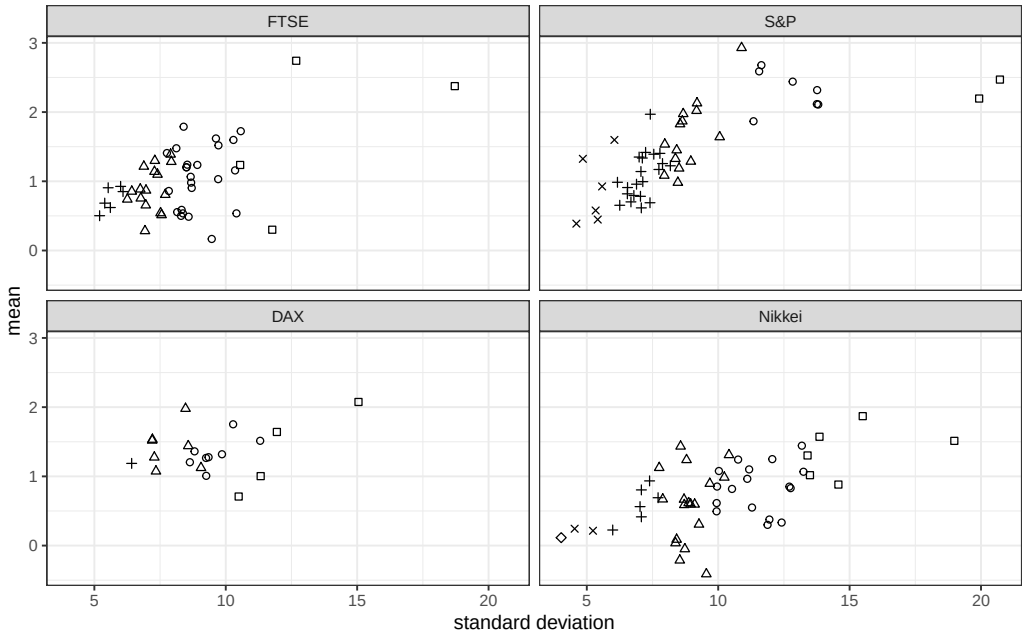

Figure 4 Homogeneous standard-deviation subsets for four asset classes

there was only one subset homogeneous in mean. We do not have 50 DAX stocks and so Figure 4 shows 20 assets for DAX and 50 for the other three asset classes.

Figure 5 shows the means and standard deviations of the sample of 50 hedge funds. The charts on the left use different symbols for the subsets homogeneous in mean. The charts on the right use different symbols for subsets homogeneous in standard deviation.

4.2. Empirical tests

We now test Heuristic 1 on the data of Section 4.1. We consider three optimisation strategies, min-c, max-c and cmax (problems (4), (5) and (3)), because we expect these to behave differently. The first two need a bound. For max-c and min-c we choose the bound to be the mean and the variance of the data. We need need a parameter R for cmax and choose $R = \hat{\mu}^\top \overline{\mathbf{w}}/(\overline{\mathbf{w}}^\top \hat{\Sigma} \overline{\mathbf{w}})$, where $\overline{\mathbf{w}}$ is the ewp. Then the objective of problem (3) for the whole data set will be 0 for the ewp and larger values of the objective indicate better performing portfolios.

We test Heuristic 1 with $\alpha = 0.05$ in step (1) and with both ewp and min in step (b) for samples of 20 and 50 assets from each of the five asset classes of Section 4.1. As before, we compare with the ewp and naive Markowitz optimisation.

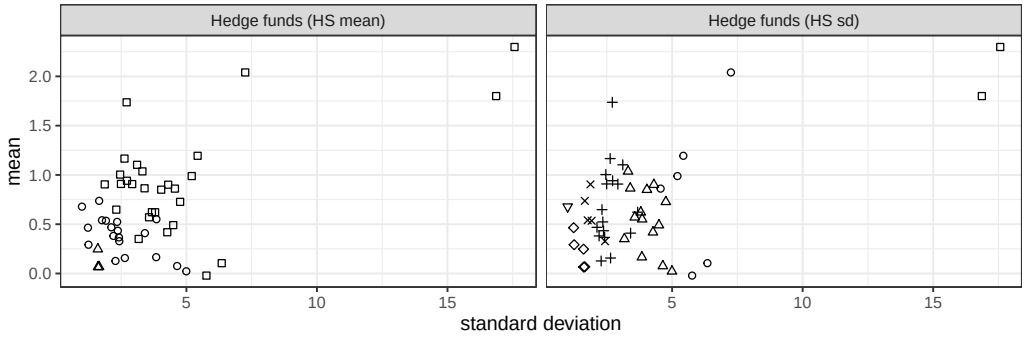

Figure 5 Homogeneous subsets for portfolios of hedge funds

We also compare the effects of using shrinkage estimators for mean or covariance. As in Section 3.3 we find using homogeneous subsets of standard deviations is no more effective than using shrinkage estimators for covariance. And when we use subsets homogeneous subsets in mean together with ewp in step (b) of Heuristic 1 we get an equally-weighted portfolio of all, or in the case of hedge funds the best homogeneous subset of, the assets. So we report only the min case. Similarly, the results for samples of 20 and 50 assets similar and so we report only those for samples of 50 assets.

Tables 9–11 summarise the results of tests for 50 assets. HF is the hedge funds. We write naive for Markowitz optimisation without shrinkage or homogeneous subsets, J or HS if we use Jorion or subsets homogeneous in mean and NL or LW for nonlinear or linear covariance shrinkage. We always try to use NL, but occasionally this makes the covariance matrix ill-conditioned and then we use LW.

Table 9 shows the results for the min-c strategy. The asterisks indicate where the bound on mean is met. The bounds for FTSE and HF are 0.837 and 0.532. All other methods perform better out-of-sample than ewp. There is no clear difference in performance among the other methods, though naive Markowitz gives less good diversification. And the minimum-variance portfolio satisfies the bound on mean when we use Jorion shrinkage or homogeneous subsets. So, in these cases the last two methods give the same portfolio.

Table 10 shows the results for the min-c strategy. Covariance shrinkage negligibly influences the results and using homogeneous subsets gives results close to ewp, with strong diversification but weaker performance out-of-sample.

Table 11 shows the results for the cmax strategy. The values of R are 0.0429 (FTSE), 0.0545 (S&P), 0.0166 (Nikkei) and 0.1506 (HF). The obj columns of the

Table 9 Results for the portfolios of $n = 50$ assets: min-c

		In-sample		Out-of-sample				
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$
FTSE	ewp	1.017	4.408	0.664	4.903	50	0.02	0.02
	naive	0.933	3.034	0.405	3.813	19	0.008	0.127
	J/LW	0.989	2.963	0.439	3.759	22	0.002	0.109
	HS/LW	1.017	2.963	0.439	3.759	22	0.002	0.109
S&P	ewp	1.437	3.926	0.791	5.039	50	0.02	0.02
	naive	1.207*	2.701	0.657	4.178	22	0.001	0.212
	J/NL	1.297	2.716	0.604	4.013	22	0.003	0.133
	HS/NL	1.437	2.716	0.604	4.013	22	0.003	0.133
Nikkei	ewp	0.74	5.801	0.421	6.314	50	0.02	0.02
	naive	0.577*	3.102	0.195	5.453	15	0.003	0.255
	J/NL	0.641	3.119	0.204	5.082	17	0.006	0.258
	HS/NL	0.74	3.119	0.204	5.082	17	0.006	0.258
HF	ewp	0.673	2.329	0.441	1.262	50	0.02	0.02
	naive	0.711	0.701	0.466	0.778	11	0.003	0.362
	J/NL	0.68	0.776	0.387	0.791	17	0.001	0.215
	HS/NL	0.594	0.776	0.387	0.791	17	0.001	0.215

Table 10 Results for the portfolios of $n = 50$ assets: max-c

		In-sample		Out-of-sample				
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$
	ewp	1.017	4.408	0.664	4.903	50	0.02	0.02
	naive	2.371	8.43 *	1.197	5.695	4	0.058	0.547
FTSE	J/LW	1.5	8.43 *	1.224	5.832	4	0.037	0.569
	HS	1.017	4.402	0.663	4.91	49	0.014	0.023
	HS/LW	1.017	4.197	0.664	4.937	50	0.013	0.022
	ewp	1.437	3.926	0.791	5.039	50	0.02	0.02
	naive	2.88	9.595*	1.462	8.734	3	0.05	0.825
S&P	J/NL	1.978	9.595*	1.47	9.167	2	0.112	0.888
	HS	1.437	3.867	0.787	5.006	50	0.016	0.025
	HS/NL	1.437	3.781	0.789	5.03	50	0.018	0.022
	ewp	0.74	5.801	0.421	6.314	50	0.02	0.02
	naive	1.671	10.063*	0.819	8.994	4	0.031	0.495
Nikkei	J/NL	1.065	10.063*	0.886	9.181	4	0.02	0.547
	HS	0.74	5.714	0.418	6.265	50	0.018	0.025
	HS/NL	0.74	5.672	0.422	6.309	50	0.018	0.021
	ewp	0.673	2.329	0.441	1.262	50	0.02	0.02
	naive	1.925	4.752*	0.915	2.426	3	0.095	0.46
HF	J/NL	1.613	4.752*	0.905	2.46	3	0.092	0.471
	HS	0.93	2.482	0.44	1.343	28	0.025	0.041
	HS/NL	0.93	2.44	0.439	1.34	28	0.025	0.04

Table 11 Results for the portfolios of $n = 50$ assets: cmax

		In-sample		Out-of-sample				
Assets	Method	$\overline{x}$	obj	$\overline{x}$	obj	m	$w_{\min}$	$w_{\max}$
FTSE	ewp	1.017	0.183	0.664	−0.368	50	0.02	0.02
	naive	1.559	0.902	0.699	0.083	10	0.045	0.172
	J/LW	1.091	0.665	0.524	−0.087	19	0.003	0.11
	HS/LW	1.017	0.64	0.439	−0.167	22	0.002	0.109
S&P	ewp	1.437	0.597	0.791	−0.592	50	0.02	0.02
	naive	1.844	1.103	0.826	−0.37	17	0.011	0.21
	J/NL	1.427	0.959	0.701	−0.253	25	0.01	0.157
	HS/NL	1.437	1.035	0.604	−0.273	22	0.003	0.133
Nikkei	ewp	0.74	0.183	0.421	−0.239	50	0.02	0.02
	naive	1.251	0.762	0.343	−0.456	12	0.009	0.269
	J/NL	0.769	0.544	0.235	−0.263	19	0.007	0.158
	HS/NL	0.74	0.579	0.204	−0.224	17	0.006	0.258
HF	ewp	0.673	−0.144	0.441	0.201	50	0.02	0.02
	naive	1.34	1.023	0.858	0.586	7	0.04	0.447
	J/NL	1.112	0.836	0.718	0.496	8	0.01	0.331
	HS/NL	0.93	0.776	0.334	0.139	13	0.002	0.201

table show the optimal value of the objective function of the optimisation problem. We chose R to give a value of 0 for ewp over the complete set of assets, in-sample and out-of-sample. So the larger the value of obj, the better the performance of the portfolio. We note that all other methods do better than ewp, none is better than others out-of-sample in all cases and homogeneous subsets and covariance shrinkage both improve diversification.

4.3. Bootstrap tests

We now consider bootstrap replications of the tests of Section 4.2. We have two reasons to do this. First, it gives more robust conclusions. Second, it allows some evaluation of the bootstrap method of Michaud and Michaud (2007).

We need a data generation process for bootstrap replication. We write R^{in} and R^{out} for the matrices whose (t, i)th entries are $r_{t,i}$ and $r_{T/2+t,i}$ ($t = 1, \ldots, T/2$, $i = 1, \ldots, n$. Then we use the method described at the end of Section 2.4 with $\tau = T/2$, which is always an integer in our data, to generate $B = 2000$ replications of the in-sample data (R^{in}). We compute optimal portfolios from R^{in}_b ($b = 1, \ldots, B$) by various methods and compute the in-sample and out-of-sample means and standard deviations from R^{in}_b and R^{out}. We test samples of $n = 20$ and 50 assets but, as in Section 4.2, report results only for 50, because the results are not very different. We also omit the ewp results. For each test we show two rows showing lower and upper 95% bootstrap percentile confidence bounds (Efron and Tibshirani, 1998) for each column.

Table 12 summarises the results for min-c and should be compared with Table 9. In some cases the virtual assets of Heuristic 1 step (c) had mean greater than the optimisation bound and so there was no feasible solution. Column val indicates the number of cases where a feasible solution was found. In this case it is striking that there is very little difference in the out-of-sample confidence intervals for naive Markowitz and Markowitz with covariance shrinkage and homogeneous subsets.

Table 13 summarises the results for max-c and should be compared with Table 10. This time all the resampled in-sample data sets have feasible solutions. Standard Markowitz performs much better in-sample, as we should expect. But the combination of homogeneous subsets with covariance shrinkage gives narrower out-of-sample confidence intervals and better diversification in each case.

We end this Section by investigating how a bootstrap aggregation (Breiman, 1996; Frahm, 2015) method can be combined and compared with the method we develop. This method is simple and similar to that of Michaud and Michaud (2007). It applies an optimisation strategy to B bootstrap resampled data sets

Table 12 Confidence intervals for portfolios of $n = 50$ assets: min-c

		In-sample		Out-of-sample						
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS	val
FTSE	naive	0.837*	2.427	0.304	3.598	13	0.001	0.115		2000
		1.478	3.264	0.638	4.238	22	0.02	0.237		
	HS/NL	0.285	2.424	0.296	3.604	13	0.001	0.115	1	1507
		1.691	3.223	0.642	4.238	22	0.019	0.247	2	
S&P	naive	1.207*	2.111	0.54	3.78	13	0.001	0.135		2000
		1.551	2.938	0.732	4.653	22	0.015	0.282		
	HS/NL	0.754	2.09	0.532	3.763	13	0.001	0.137	1	1679
		1.947	2.829	0.694	4.597	21	0.015	0.287	3	
Nikkei	naive	0.577*	2.524	0.098	4.921	9	0.001	0.213		2000
		0.875	3.626	0.3	6.374	17	0.024	0.528		
	HS/NL	−0.165	2.452	0.088	4.98	9	0.001	0.228	1	1344
		1.621	3.262	0.292	6.362	16	0.022	0.549	2	
HF	naive	0.557	0.533	0.356	0.638	8	0.001	0.18		2000
		0.899	0.731	0.525	1.166	14	0.035	0.559		
	HS/NL	0.187	0.534	0.328	0.641	8	0.001	0.171	2	1846
		1.154	0.833	0.524	1.22	14	0.036	0.55	4	

Table 13 Confidence intervals for portfolios of $n = 50$ assets: max-c

		In-sample		Out-of-sample					
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS
FTSE	naive	1.817	7.384	0.073	4.736	1	0.004	0.324	
		4.033	8.43 *	1.652	13.15	6	1	1	
	HS/NL	0.35	3.762	0.332	4.713	2	0.004	0.023	1
		1.84	6.913	0.911	6.57	50	0.359	0.733	2
S&P	naive	2.494	8.923	0.384	5.149	1	0.003	0.322	
		4.942	9.595*	1.645	9.761	6	1	1	
	HS/NL	0.921	3.452	0.77	4.918	6	0.009	0.024	1
		2.352	5.927	1.03	6.21	50	0.074	0.291	3
Nikkei	naive	1.418	8.306	−0.151	7.033	1	0.004	0.314	
		3.959	10.063*	1.038	12.201	7	1	1	
	HS/NL	−0.083	5.05	0.328	6.132	2	0.007	0.024	1
		1.838	8.38	0.501	8.187	50	0.26	0.825	2
HF	naive	1.499	2.533	−0.046	2.227	1	0.006	0.35	
		3.17	4.752*	1.643	3.802	5	1	1	
	HS/NL	0.345	1.984	0.382	1.172	9	0.015	0.033	2
		1.758	3.41	0.642	1.945	36	0.065	0.163	4

Table 14 Bootstrap aggregation combined with other methods: min-c

		In-sample		Out-of-sample					
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS
FTSE	ewp	1.017	4.408	0.664	4.903	50	0.02	0.02	
	naive	0.96	3.068	0.46	3.765	33	0.002	0.108	
	HS/NL	0.936	3.066	0.456	3.761	33	0.001	0.107	1.91
S&P	ewp	1.437	3.926	0.791	5.039	50	0.02	0.02	
	naive	1.161	2.715	0.63	4.103	38	0.001	0.177	
	HS/NL	1.091	2.686	0.608	4.057	37	0.001	0.165	1.99
Nikkei	ewp	0.74	5.801	0.421	6.314	50	0.02	0.02	
	naive	0.481	3.034	0.192	5.44	30	0.001	0.297	
	HS/NL	0.403	2.976	0.184	5.532	25	0.001	0.339	1.89
HF	ewp	0.673	2.329	0.441	1.262	50	0.02	0.02	
	naive	0.703	0.719	0.444	0.766	20	0.001	0.332	
	HS/NL	0.714	0.726	0.437	0.784	23	0.002	0.309	3.12

Table 15 Bootstrap aggregation combined with other methods: max-c

		In-sample		Out-of-sample					
Assets	Method	$\overline{x}$	$\overline{s}$	$\overline{x}$	$\overline{s}$	m	$w_{\min}$	$w_{\max}$	HS
	ewp	1.017	4.408	0.664	4.903	50	0.02	0.02	
FTSE	naive	1.944	6.233	0.991	5.445	36	0.001	0.295	
	HS/NL	1.043	4.417	0.667	4.934	50	0.017	0.024	1.91
	ewp	1.437	3.926	0.791	5.039	50	0.02	0.02	
S&P	naive	2.447	6.476	1.289	6.001	26	0.001	0.215	
	HS/NL	1.549	4.136	0.836	5.231	50	0.011	0.025	1.99
	ewp	0.74	5.801	0.421	6.314	50	0.02	0.02	
Nikkei	naive	1.287	7.025	0.447	7.587	35	0.001	0.123	
	HS/NL	0.75	5.718	0.423	6.302	50	0.014	0.024	1.89
	ewp	0.673	2.329	0.441	1.262	50	0.02	0.02	
HF	naive	1.697	3.745	0.836	1.935	12	0.003	0.386	
	HS/NL	0.966	2.535	0.473	1.366	41	0.002	0.053	3.12

to generate B resampled optimal portfolios. Then it uses the average of these portfolios as the estimate of the optimal portfolio.

Since we have already generated $B = 2000$ resampled in-sample data sets for each of the strategies and methods we test, we can use these for bootstrap aggregation. We construct an optimal portfolio estimate $\mathbf{w}$ as the average of the 2000 resampled in-sample optimal portfolios. Then we calculate the mean and standard deviation of the returns in-sample and out-of-sample using $\mathbf{w}$. When we use a method that includes homogeneous subsets and covariance shrinkage we recalculate the shrinkage covariances for each resampled data set rather than resample from a data set with shrinkage covariances. We currently have no way to generate such a data set.

Table 14 summarises the performance of the average optimal portfolio with a min-c strategy and 50 assets. Columns m, $w_{\min}$ and $w_{\max}$ show the number of non-negligible weights, the smallest non-negligible weight and the largest weight in the average portfolio. Column HS shows that average number of homogeneous subsets found in 2000 resamples. The unsurprising result is that the average (bootstrap aggregated) portfolio out-of-sample performance is about the same when we use naive Markowitz and when we use homogeneous subsets and covariance shrinkage. Both achieve lower standard deviation than the ewp.

The results of Table 15 are more interesting. It summarises the performance of the average optimal portfolio with a max-c strategy and 50 assets. Both naive

Markowitz optimisation and homogeneous subsets with covariance shrinkage give higher out-of-sample mean than the ewp. But the naive average portfolio gives a higher mean in each case without exceeding the bound on standard deviation.

5. Discussion and conclusions

Perhaps the greatest challenge in portfolio optimisation based on historic data is that we are trying to optimise over a small amount of information in the presence of a large amount of noise. The noise includes uncertainty in the true values of statistics such as mean and variance that we wish to optimise over. And it is complicated by the fact that real assets have time-series effects and are not normally distributed.

When statistics works it is because the average effect of noise is zero. So, it is tempting to attribute the failure of Markowitz portfolio optimisation to time-series effects, asymmetry of distributions or estimation risk. We have demonstrated that Markowitz portfolio optimisation can fail badly even in the absence of such issues. Rather than averaging out the effects of noise, quadratic or quadratically constrained optimisation can select on noise alone and be an extreme case of fitting the data better than the population. We have also shown that, while covariance shrinkage can be helpful, Markowitz optimisation is too sensitive to small differences in mean values for the currently-known mean shrinkage estimators to prevent it from choosing a portfolio that is far from the true optimum.

If we hope to use optimisation to control the level of risk in a portfolio we must find better ways to identify when differences, especially in mean return, are more plausibly due to information than noise. We introduce a method of homogeneous subsets that can help. In essence, it allows us to cluster assets into subsets so that assets within a subset are plausibly indistinguishable on some statistic such as mean return, while assets in different subsets are plausibly distinguishable.

We find homogeneous subsets in means to be the most informative. There are two reasons for this. First, Markowitz optimisation is more sensitive to the effects of noise in the mean than in variances. Second, homogeneity of variance is complicated by the fact that when two assets have indistinguishable variance, their covariance still matters: a convex combination of the assets may still reduce variance. Our tests suggest that it is better to use covariance shrinkage which deals with both variances and covariances to limit the effects of noise on covariance.

When we construct homogeneous subsets from historic asset data we find new and informative results. First, while all asset sets vary in standard deviation, our selections from the FTSE, Standard and Poor, DAX and Nikkei all plausibly have

indistinguishable mean returns, at least over an in-sample period. We had to look at hedge-fund returns, which use a range of investment strategies, to find distinguishable means. This limits how much improvement we may expect to find by using the method (Heuristic 1) we develop to deal with homogeneous subsets.

When we test Heuristic 1 on historic asset returns we find it is more robust than naive Markowitz optimisation, but still allows us to select a portfolio with lower risk or higher return than the ewp. It works best on mean maximisation, though a bootstrap aggregation method, similar to that of Michaud and Michaud (2007), appears to perform better and is also computationally less expensive.

It is tempting to conclude that the empirical evidence favours bootstrap aggregation as a generic portfolio selection strategy. However, we should note that bootstrap aggregation should work best when, as we found, differences in mean value are mostly due to noise (Frahm, 2015). Heuristic 1 has clearer theoretical justification and may perform better if we can adjust it to deal with complicating issues such as time-series effects, which can be done by replacing the bootstrap with time-series bootstrap methods. The selection of homogeneous subsets depends on the choice of an experimentwise significance level α. We chose $\alpha = 0.05$ and a less conservative choice might allow us a better balance between eliminating noise and preserving information. More generally, we might be able to develop our methods from one where we make a binary decision about whether a subset is homogeneous or not to one where we optimise based on some estimate of how likely it is assets are homogeneous in some statistic.

All of these require further research. But while there is no clear evidence (see the discussion in Frahm (2015)) in favour of any particular portfolio selection strategy, identifying plausibly homogeneous subsets may be useful in helping choose which strategy or combination of strategies to use.

References

Arrow, K.J., 1950. A difficulty in the concept of social welfare. Journal of Political Economy 58, 328–346.

Benartzi, S., Thaler, R.H., 2001. Naive diversification strategies in defined contribution saving plans. American Economic Review 91, 79–98.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Cerrato, M., Crosby, J., Kim, M., Zhao, Y., 2017. Relation between higher order

comoments and dependence structure of equity portfolio. Journal of Empirical Finance 40, 101–120.

Datastream, 2019. Refinitiv Datastream. URL: https://solutions.refinitiv.com/datastream-macroeconomic-analysis. Subscription Service; Accessed: May 2019.

David, H.A., 2008. Order Statistics. Wiley-Interscience.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy? Review of Financial Studies 22, 1915–1953.

Eaton, M.L., 2007. The Wishart Distribution, in: Multivariate Statistics. Institute of Mathematical Statistics Lecture Notes - Monograph Series. chapter 8, pp. 302–333.

Efron, B., Tibshirani, R., 1998. An Introduction to the Bootstrap. Chapman and Hall/CRC.

Fliege, J., Werner, R., 2014. Robust multiobjective optimization & applications in portfolio optimization. European Journal of Operational Research 234, 422–433.

Fourdrinier, D., Strawderman, W.E., Wells, M.T., 2018. Shrinkage Estimation. Springer.

Frahm, G., 2015. A theoretical foundation of portfolio resampling. Theory and Decision 79, 107–132.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F., 2009. GNU Scientific Library Reference Manual. Network Theory. URL: https://gsl.gnu.org.

Harvey, C.R., Liechty, J.C., Liechty, M.W., Peter, M., 2010. Portfolio selection with higher moments. Quantitative Finance 10, 469–485.

Herold, U., Maurer, R., 2006. Portfolio Choice and Estimation Risk: A Comparison Of Bayesian To Heuristic Approaches. ASTIN Bulletin 36, 135–160.

Holm, S., 1979. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6, 65–70.

Hwang, I., Xu, S., In, F., 2018. Naive versus optimal diversification: Tail risk and performance. European Journal of Operational Research 265, 372–388.

Jorion, P., 1985. International Portfolio Diversification with Estimation Risk. Journal of Business 58, 259–278.

Jorion, P., 1986. Bayes-Stein Estimation for Portfolio Analysis. Journal of Financial and Quantitative Analysis 21, 279–292.

Kan, R., Zhou, G., Raymond, K., Guofu, Z., 2007. Optimal Portfolio Choice with Parameter Uncertainty. Journal of Financial & Quantitative Analysis 42, 621–659.

Kolm, P.N., Tütüncü, R., Fabozzi, F.J., 2014. 60 Years of portfolio optimization: Practical challenges and current trends. European Journal of Operational Research 234, 356–371.

Lamb, J.D., Tee, K.H., 2012. Resampling DEA estimates of investment fund performance. European Journal of Operational Research 223, 834–841.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365–411.

Ledoit, O., Wolf, M., 2017. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. Review of Financial Studies 30, 4349–4388.

Markowitz, H., 1952. Portfolio selection. The Journal of Finance 65, 1581–1611.

Meade, N., Beasley, J.E., Adcock, C.J., 2021. Quantitative portfolio selection: using density forecasting to find consistent portfolios. European Journal of Operational Research 288, 1053–1067.

Merton, R.C., 1980. On estimating the expected return on the market. Journal of Financial Economics 8, 323–361.

Michaud, R.O., 1989. The Markowitz Optimization Enigma: Is 'Optimized' Optimal? Financial Analysts Journal 45, 31–42.

Michaud, R.O., Michaud, R., 2007. Estimation Error and Portfolio Optimization: A Resampling Solution. The Journal Of Investment Management 45(1), 31–42.

Michaud, R.O., Michaud, R.O., 1998. Efficient Asset Management. Oxford University Press, Inc. Published, Boston.

Nadarajah, S., Kotz, S., 2008. Exact distribution of the max/min of two Gaussian random variables. IEEE Transactions on very large scale integration 16, 210–212.

Refinitiv, 2018. Lipper TASS database. URL: https://www.refinitiv.com/en/%0Apolicies/third-party-provider-terms/tass-database. Subscription Service; Accessed: April 2018.

Shaffer, J.P., 1986. Modified Sequentially Rejective Multiple Test Procedures. Journal of the American Statistical Association 81, 826–831.

Šidák, Z., 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62, 626–633.

Stein, C., 1955. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, in: Proceedings of the 3rd Berkeley Symposium on Probability and Statistic, pp. 197–208.

Vinod, H.D., 2004. Ranking mutual funds using unconventional utility theory and stochastic dominance. Journal of Empirical Finance 11, 353–377.

Xidonas, P., Steuer, R., Hassapis, C., 2020. Robust portfolio optimization: a categorized bibliographic review. Annals of Operations Research 292, 533–552.

Zhang, Y., Li, X., Guo, S., 2018. Portfolio selection problems with Markowitz's mean–variance framework: a review of literature. Fuzzy Optimization and Decision Making 17, 125–158.