DIVERSIFICATION IN AN IVAR FRAMEWORK

Emiel Lemahieu - K19010410
Supervisor: Dr. Fotis Papailias

This dissertation is submitted in part fulfilment of the requirement for the degree of M.Sc. Finance Analytics

Diversification in an iVaR framework

Emiel Lemahieu*

September 12, 2020

Abstract

This paper addresses some of the oft-quoted shortcomings of standard portfolio construction frameworks. We first argue that because of (1) their risk perception, i.e. the formalization of the risk-reward tradeoff for the ranking of portfolios, and (2) because of their instability, they in practical applications often lack the properties of what we consider a rational portfolio. The proposed algorithms revolve around the construction of diversified portfolios in an iVaR framework. iVaR was introduced by the financial technology firm InvestSuite and embraces the human perception of risk in a portfolio: the frequency of drawdowns, their magnitude, and the time to recover from them. A common challenge in any portfolio construction framework is to make the allocation diversified, i.e. achieving low risk by spreading over many lowly correlated assets rather than over low risk assets in a concentrated way. We delve into the literature to find out what diversification means in the most tractable construction frameworks. We then try to define diversification in an iVaR framework. Next, we compare three distinct viewpoints on how to achieve an optimally diversified iVaR portfolio: maximizing diversification benefits, penalizing excessive concentrations, and recursive optimization on subuniverses or clusters of assets. We backtest these strategies and compare out-of-sample metrics such as risk-adjusted returns and diversification ratios with undiversified iVaR and standard construction frameworks using Hansen's bootstrapped model confidence set. We finally draw conclusions on which viewpoint proves most valuable for a more rational and robust approach to portfolio construction.

Keywords – Diversification, iVaR, Markowitz' Curse, Robust Optimization, Most-Diversified Portfolio, Clustering, Hierarchical Risk Parity, Hansen MCS test

*The author would like to express deep gratitude to Fotis Papailias, Emmanuel Wildiers, Laurent Sorber, Mathieu De Baets, Maartje M.E. Lemmens, Steven De Blieck, Sander Ducastel and Victor Hutse for their valuable comments.

Contents

I. Introduction

"*It's always better to be lucky than smart.*" (Robert C. Merton)

There is something inherently irrational about traditional portfolio construction frameworks. Rational decisions are defined as (1) the best decisions at every point in time, (2) based on all available information at that point. (1) Best can be defined as the investor's ability to at least rank his investment opportunities (*ordinal utility*), without the need for the exact differences in value between opportunities, and to pick the prevailing options. This ranking should be coherent, consistent and transitive. (2) All information means the investor is able to identify all investible assets and their attributes. Rationality w.r.t. portfolios thus implies [8]:

- the ability to observe all possible alternatives and their attributes.
- the ability to at least ordinally rank portfolios (i.e. not necessarily continuously 'score').
- the ability to choose the best portfolios in a way that is consistent and transitive. I.e. if we prefer portfolio A over B, and B over C, we do not pick C over A.

At least in practical applications traditional portfolio construction techniques typically lack many to all of these properties. The most obvious violation of rationality is that (robo)investors are unable to deal with the complexity of the world and therefore cannot possibly consider all the combinations of all investible assets in the world. Even for the small subset of assets in their horizon, they do not know all the *relevant* attributes. It would be naive to claim that machines would (ever) be able to observe all these attributes, but it is a simple fact that quantitative portfolio techniques typically reduce the number of features to be explored about an asset to the moments of its historical performance. As we will argue later on, this is wrong but often very useful nevertheless. However, it stays particularly surprising that historically, portfolio theory quickly revolved around limiting the number of inputs to the optimization algorithm, while finance and outperforming — or identifying the *best ranked* — portfolios is all about holding and exploiting asymmetric information. The only explanation in favor of these traditional views is market efficiency and *rational ignorance*, when the benefits of exploiting additional information do not outweigh the costs of collecting it. This is somehow how defenders of the efficient market hypothesis respond to Stiglitz' *informational paradox*[1] [6]. Even if this would be true, it is hard to believe that moments of the historical performance would suffice.

The second condition is more often violated than not as well by traditional methods, since the ranking implied by their utility function[2] is notoriously instable. Admittedly, information changes over time and thus portfolio preferences should be updated. However, traditional (variance-based) portfolio techniques change their implied rankings drastically for a small fluctuation in the input. This makes us suspicious about the quality of our initial ranking and leads us to suspect that issues of noisy estimation, misspecification or overfitting are at the root of this instability, rather than a signal

[1] 'If markets are efficient, why do banks and funds invest in multi-billion informational systems?'

[2] i.e. most notoriously quadratic utility in modern portfolio theory.

that the world has genuinely changed. Importantly, traditional approaches do not only suffer from noise-induced but also signal-induced instability [35] inherent to their risk-reward formalization. These limitations are then accounted for with techniques we are going to discuss later, but stability is often forced by constraints or naive shrinkage. Portfolios are then robust by correction, not construction. With the dawn of roboadvisors, automated investors should be rational by design, not correction. Although it is in traditional frameworks easy to solve for the best portfolio, this inherent instability makes us also often violate the third condition of consistency and transitivity over time, as completely opposed preferences can be implied over subsequent periods of time. Moreover, these issues oftentimes boil down to excessive concentrations in our portfolio. Stability and diversification, therefore, are inextricably linked.

This same lack of robustness brings us to another contradiction in modern-day finance. On the one hand, we are told that markets are extremely efficient and hard-to-beat. On the other hand, many asset management firms use plain-vanilla financial analysis techniques and simple factor regression models to screen assets and construct multi-billion portfolios. This anomaly epitomises why many asset managers fail to consistently beat a benchmark, and occasional outperformance is more likely to be a product of chance — being lucky — rather than skill — being smart.

The aim of this paper, and diversification in general, is to provide asset managers with more robust decision making. Let's say any investment decision is just a sample out of the total population of potential decisions, which decision would consistently outperform others? It's the very decision that (1) maximises the amount of relevant information incorporated in that decision and (2) optimizes how efficiently we use that information.

- Firstly, statistics 101 teaches us that more information — or a more efficient use of it — means lower variance, and lower variance in-sample (which determine our decisions), give us more robust results out-of-sample (which determine the results of our decisions).
- Secondly, the quintessential problem with human decision-making boils down to bias: we fail to cope with the complexity of the world and increasingly big data sets that fall beyond the grasp of standard econometrics. As a result, asset managers (and their models) make shortcuts, and these shortcuts lead to biased investments. The algorithms deployed in this paper are prone to bias and overfitting as well, but it is much easier to detect and correct for such biases with machines, rather than with humans.

This paper is a humble attempt to use information in a more efficient way and allow for more robust decision making. We do not claim the aim of a paper would be to construct a superior data set and resolve the first condition. Rather, we want to provide a framework that is likely to outperform traditional techniques on this platonic archetype of a data set. We should therefore be pragmatic and never forget that we are essentially trying to rank portfolios in more rational way, by trying to formulate the *right* problem and not applying some statistical trickery on the *wrong* problem. So the question comes down to: based on the information at hand, how can we compare portfolios and generate weights for the assets in our horizon that reflect the best ranked one? Following our discussion this means (1) *redefining the tradeoff or maximization* *of utility in terms that adhere to what investors really see as risk and reward* (Chapter I), and (2) *making the whole process more robust or less instable* (Chapter II).

The paper is structured as follows. This introduction I. will set the stage. After a brief history of traditional risk measures, we will define iVaR and compare it to classical measures. Next, we define diversification. What does it mean for a portfolio to be diversified? Where do diversification benefits stem from and how do we quantify them? Section II. answers these questions supporting its motivations on a brief literature review. This allows us to elaborate on the link between iVaR and these classical viewpoints. Section III. describes the data sets we used and then zooms in on the obtained return, risk and diversification characteristics of each of the portfolio construction frameworks for multi-period simulations. We draw conclusions on the usefulness of these alternative views on diversification and their implications for drawdown frameworks such as iVaR in section IV..

A. A brief history of traditional risk measures and their shortcomings

"*If you give a pilot an altimeter that is sometimes defective, he will crash the plane. Give him nothing, and he will look out of the window.*" (Nassim Nicholas Taleb)

Most of current-day portfolio optimizers and roboadvisors are based on Markowitz' mean-variance optimization framework. Former CUNY professor and Nobel laureate Harry Markowitz introduced a method [1] and algorithm [3] for asset allocation under uncertainty in the 1950s, based on variance (1) or volatility as a risk measure. The method laid the foundations for a vast field of literature, now referred to as Modern Portfolio Theory (MPT). Moreover, the mean-variance tradeoff is a theoretical cornerstone in other financial theories such as the Capital Asset Pricing Model (CAPM) and the Efficient Market Hypothesis (EMH), with well-known applications in practical and empirical finance[3].

$$\sigma_p^2 = \frac{1}{T}\sum_{t=1}^{T}(r_{p,t} - E(r_{p,t}))^2 \tag{1}$$

According to his theory, investors can compare all possible combinations or weights w of the assets in their horizon according to the expected return ($E(r_{p,t})$) and their risk (σ_p^2). One such combination dominates another if it offers a higher return for the same variance or lower variance for the same return. Since we assume means and (co)variances provide sufficient information for this tradeoff, we assume returns are normally distributed — the natural distribution that is fully described by its first two moments only. A result of using variance and the normal (or any elliptical distribution), is that the boundary set of all risk-reward tradeoffs is convex. The solution to the method is then to determine the tangent point between this boundary — also called the Efficient Frontier — and the lines composed of all combinations of a risk-free rate and risky assets on the frontier. This point will maximize the ratio expected return over variance, also known as the Sharpe ratio, which embodies the tradeoff we started from.

The algorithm, the Critical Line Algorithm (CLA), is less well-known and solves the

[3]Such as widely adopted factor strategies, smart beta, and so forth.

quadratic optimization problem inherent to the method. In essence, it solves a sequence of standard mean-variance optimization problems, where we get the optimal portfolio for some critical level of risk aversion. Apart from pioneering a systematic methodology for assessing the risk-reward tradeoff of risky assets and this technical contribution, Markowitz' main insight was that investors should not assess risky investments as isolated opportunities, but rather investigate how adding risky assets to a portfolio might improve the overall risk-adjusted return of that portfolio (cf. next chapter). Therefore, his famous 1952 paper was actually the seminal paper on how to quantitatively assess diversification (rather than qualitatively, using e.g. constraints on asset classes or GICS[4] industries). This is exactly the aim of this paper.
MPT has many attractions: it is simple to understand and interpret results (using variance-based measures such as Sharpe ratios, etc.) and it can be linked to other financial theories and concepts. However, its limitations are also extensively documented in literature. Many of them refer to the simplistic assumptions that underlie mean-variance optimization. The most prevalent one is that we choose variance as our measure of risk, which means we assume returns are normally distributed. Therefore, we ignore many of the features of the real profit-and-loss distribution (PnL) of candidate portfolios and squeeze them into a normal straightjacket. (Conditional) value-at-risk and drawdown measures try to remediate this shortcoming. Moreover, variance or volatility has some other oft-quoted shortcomings, most importantly being symmetrical and sign-ignorant. The first limitation means that volatility punishes positive dispersion, i.e. surprise positive returns, to the same extent as surprise negative returns. The latter means that variance does not punish consistent negative returns.

The first answer to these issues was not proposed many years later, but only three months after the famous Markowitz 1952 paper, by Roy [2]. Indeed, Roy's theories were not developed as a response to Markowitz, but concurrently. Markowitz later acknowledged [11] that if Roy had published his work three months earlier and developed a theory using his mean-SV efficient sets, we would be talking about Roy's portfolio theory today instead of Markowitz' portfolio theory. SV stands for semivariance, the first of the so-called *downside risk measures*. Roy argued that investors consider risk asymmetrically as any deviation below some target return. Investors prefer safety of the invested principal first, and will set some minimum acceptable return. Roy called the minimum acceptable return the disaster level, and the portfolio technique Roy's *safety first* principle. The resulting risk measure is not variance (1), but a similar expression (2). We consider risk as variance below some target return R:

$$SV_t = \frac{1}{T} \sum_{t=1}^{T} \max\left(R - r_{p,t}, 0\right)^2 \tag{2}$$

We call SV_t the below-target semivariance. This should not be confused with a more specific measure, the below-mean semivariance SV_m, where $R = E(r_{p,t})$, which is highly adopted in literature:

$$SV_m = \frac{1}{T} \sum_{t=1}^{T} \max\left(E(r_{p,t}) - r_{p,t}, 0\right)^2 \tag{3}$$

(3) is closest in meaning to variance and can be seen als 'half-variance' for symmetric distributions. For portfolios with symmetric distributions, $\sigma_p^2 / SV_m = 2$, which is not

[4]Global Industry Classification Standard developed by Standard & Poor's, a commonly used classficiation to assess or guarantee diversification.

true for skewed portfolios. This gives SV_m the interesting interpretation as a measure of skew to compare Markowitz with Roy criteria when constructing portfolios. It also illustrates that variance can be seen as a special case of scaled semivariance where $R = E(r_{p,t})$.

Where SV is a generalization of σ, the work of Bawa and Fishburn [7] in the 1970s generalized semivariance as part of the family of lower partial moments (LPM). Compared to equation (2) and (3), there is nothing spectacular going on in equation (4) below which describes the (α, R)-LPM:

$$(\alpha, R) - LPM = \frac{1}{T}\sum_{t=1}^{T} \max\left(R - r_{p,t}, 0\right)^{\alpha} \tag{4}$$

What is spectacular, however, is that this generalization of $\alpha = 2$ to any α in $\mathbb{R}$ resolved many issues and academic discussions about downside risk measures, especially the implications for the utility of wealth. Markowitz assumed quadratic utility, which he borrowed from von Neumann and Morgenstern utility functions [19], and flow naturally out of his equations because of the use of variance as a risk measure. For Roy's SV it was not that obvious, and from a theoretical stance that is one of the reasons why the adoption of Roy's measure was much less than the famous (Nobel prize winning) MPT model. Roy once mockingly said [11]: "*A man who seeks financial advice will not be happy with the suggestion to maximize his quadratic utility.*" We can attest to that, but unfortunately the (financial) academic world often picks mathematical derivations and elegance before realism. Or in the words of Wassily Leontief [38]: *"Page after page of professional economic journals are filled with mathematical formulas leading the reader from sets of more or less plausible but entirely arbitrary assumptions to precisely stated but irrelevant theoretical conclusions.*"

Looking at (4), we see that SV corresponds to $\alpha = 2$, and $\alpha = 0$ corresponds to the probability that the return is lower than R. Therefore, the latter case is called the below target probability (BTP). The breakthrough of the LPM family is that α can be any real value, like 2 or 4, including fractional values like 2.11 or 4.32. Fishburn [7] proves that these different α values belong to different risk profiles for a mean-LPM rational investor, where low values are indicative of risk-seeking investors, and high values pertain to risk-averse investors. There is no reason to prove this here or delve into utility theory, it is only worth mentioning that because of LPM downside risk measures are fully 'compliant' with utility theory and classical risk measures such as variance and semivariance are special cases. Moreover, (conditional) value-at-risk measures can be categorized as belonging to the LPM family, as we discuss below.

Value-at-risk (VaR) was introduced by JPMorgan's RiskMetrics in the 1990s and looks at the left tail of the PnL. VaR was basically an answer to JPMorgan's CEO request to give him, every day at 4pm, a back-of-the-envelope estimate of how much money his desks could lose over the next 24 hours with a certain very low probability. This is a very intuitive definition of risk: how much money we could lose because of certain extremely infrequent but negative events (with a certain low probability p or confidence level $cl = 1 - p$), on a certain asset or portfolio, over a certain time horizon. VaR (5) corresponds to a cutoff point deep in the left tail of our PnL distribution. More intuitively:

$$p = 1 - cl = Pr(r < VaR_{1-cl}(F)) = F(VaR_{1-cl}(F)) \tag{5}$$

where p is the probability that the return is worse (smaller) than VaR[5]. Taking a close look at (5) explains the link with LPM. VaR can be seen as the inverse of LPM of order 0, and is just a different way of looking at the same thing. Where the BTP gives you p when we assess a disaster level of return R, VaR gives you R when we assess a very low probability p. BTP and VaR bring you to the same viewpoint on top of the same hill, just the tracks to take you there are different. This is explained in more detail in Appendix 1.

VaR ignited a broad field of literature on how to determine this distribution (e.g. parametrically, non-parametrically, Monte Carlo, etc.), evaluate it over time, and incorporate this risk management practice into portfolio management. Moreover, VaR adoption by regulators challenged banks with the huge data management task of aggregating their risk data in such a way that it was possible to calculate VaR at portfolio, desk and bank level. VaR's main attraction is that it is very simple to understand, and (under certain assumptions) calculate and backtest. Moreover, it is expressed in a measure that is intuitively close to our sense of risk: Pounds.

However, VaR has some major limitations. First and foremost, VaR does not look beyond the cutoff, i.e. what happens deeper in the tail. Therefore, although we can set this probability very low, we are ignorant of those scenarios that are even less probable but still possible. The Global Financial Crisis (GFC) of 2008 and the unprecedented volatility caused by the sudden worldwide outbreak of the Coronavirus underscore the importance of considering all worst-case scenarios or Black Swans, not just one scenario. Another consequence of considering only one cutoff point is that some diversification benefits tend to get lost in a VaR framework. Individual assets with very high potential losses that are less likely than the cutoff point can go unnoticed by VaR. When we combine many of these in a portfolio, the joint probability function will tell us that the odds of one such an asset going to ruin is higher than the individual probability. If now the potential loss is measured by VaR, the risk of a portfolio is higher than the sum of the risks of individual assets. In mathematical terms, VaR lacks subadditivity. By definition of diversification, a portfolio of assets should never be more risky than holding similar assets individually. This is a main point of the next chapter. That is why regulators have adopted Expected Shortfall, the subadditive expected tail loss, as a new measure of market risk in the FRTB (Fundamental Review of the Trading Book) in the revised Basel III framework, often referred to as Basel IV.

Expected shortfall (ES) or Conditional Value-at-Risk (CVaR) is very similar in nature to VaR, but it answers a slightly different question: how much money do we expect to lose on a certain asset or portfolio, with a certain probability p (or confidence level $cl = 1 - p$) over a certain time horizon, given that things are at least as bad as VaR suggests? So if things turn south, how bad can it get? This is simply the probability-weighted average loss beyond some cutoff point (VaR):

[5]Or, equivalently, that the loss is higher, since we can define VaR as a positive number drawn from the LnP or loss-and-profit distribution.

$$CVaR_{1-cl}(F) = E(r|r < VaR_{1-cl}(F)) \tag{6}$$

Again take a close look at equation (6): it is conceptually intimitely related to the LPM of order 1, where we start from specifying a probability *cl* instead of a disaster return R. Hence, ES or CVaR is the inverse concept of a lower-partial moment of order 1 where R = VaR. Note that LPM would take the mean excess loss $R - r$, while CVaR just takes the average loss in excess of VaR, but essentially the two things measure the same risk. This is again explained in more detail in Appendix 1.

CVaR is preferred over VaR since it is subadditive, and since it has other desired properties which makes it a so-called coherent risk measure (see below). In brief, we can say that it inherits the attractions of VaR, and adds this desirable property. However, its main limitation is its perception of risk. CVaR is an extremely useful measure for the determination of a bank's, or even your personal, capital adequacy. How levered a bank's business model can be, or how much exposure you can bear with taking a bet, both depends on the expected size of potential severe losses. However, for portfolio optimization purposes, CVaR misses a crucial point. It by construction neglects smaller losses and their persistence over time, as ES is essentially a *static* measure. For investors, smaller losses matter as well, especially if they are autocorrelated or persistent over time (the '*dynamics*' of losses). This should be captured by a risk measure used in a portfolio construction framework, in contrast to traditional risk management and scenario analyses. Small but persistent losses will, as much as more severe singletons, influence the efficiency of the allocation. In summary, CVaR has desirable properties but it does not capture (1) all relevant losses to rational investors, nor (2) the time dimension.

Finally, we can introduce drawdown risk measures. Concepts such as peak-to-valley drawdowns[6], recovery periods or time-under-water [7], Calmar ratios [8], Sterling ratios [9] and Burke ratios [10] are commonly used to assess fund performance ex-post. Moreover, drawdowns naturally influence fund management by regulations and the behaviour of client accounts. For instance, [18] suggests for US CTA portfolios that (1) regulators may issue warnings at 15% drawdowns and shut down accounts automatically at 20%, (2) clients, on the other hand, will very unlikely tolerate 50% drawdowns or (even small) drawdowns that last for longer than a year.

In general, drawdowns are the differences ξ between a portfolio level V and its running maximum or high-water mark. Drawdown measures can be based on the worst-case plunge or maximum drawdown MDD, average drawdown ADD or conditional drawdown $\alpha - CDD$ (average of α% worst drawdowns, see Appendix 3). The time to recover from drawdowns or reach a new high-water mark is called the recovery period or time-under-water.

$$AD(V) = (\xi_1, ..., \xi_T), \xi_t = \max_{s<t}(V_s) - V_t \tag{7}$$

$$MDD(V) = \max_{s<T}(\xi_s) \tag{8}$$

[6] MDD from peak to trough.
[7] Time from one to the next high-water mark.
[8] Expected return over MDD ratio.
[9] Expected return over ADD ratio.
[10] Expected return over root-mean-squared drawdowns.

$$ADD(V) = \frac{1}{T} \sum_{t=1}^{T} \xi_t \tag{9}$$

The maximum drawdown can be seen as a *dynamic* generalization of VaR or worst-case *static* loss, while conditional drawdown can be seen as a *dynamic* generalization of CVaR or expected *static* loss. Effectively, we are just replacing the VaR functional by an AD functional [18]. This is explained in more detail in Appendix 3.

Figure 31 below summarizes our brief history so far.

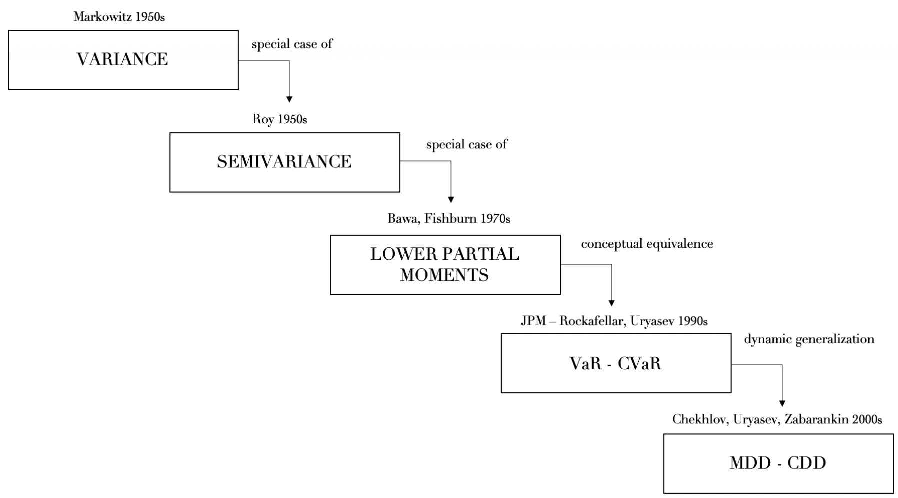


Figure 1: Overview risk measures

Given that regulators stress the concept, and that it naturally influences client behaviour, it is obvious that one could optimize for drawdowns in a portfolio construction framework. However, it was not until Chekhlov, Uryasev and Zabarankin [18] published a paper on portfolio construction with drawdown risk measures that it was properly considered a measure used to generate mean-drawdown efficient sets and optimize portfolios *ex-ante*. Conceptually, this might be surprising, but the reasons are twofold. Firstly, solving for optimal portfolios with drawdown measures is mathematically an order of magnitude more difficult than mean-variance optimization, although conceptually equally easy to grasp. Secondly, when we do not merely utilize a measure as ex-post evaluation but as part of an optimization algorithm, the risk measure has to have some desirable mathematical properties[11]. These required desirable properties did not reach wide academic consensus but after the work of Artzner and Delbaen in 1999, in what they call coherent risk measures [10]. The 6 axioms of good risk measures are listed and explained in Appendix 2. [18] proved that drawdown measures are coherent and use them in optimization. Appendix 3 provides the intuition behind their proofs. This is essential, as it makes sure concepts such as diversification hold in a construction framework. That is why we will delve into this again in the next

[11] This in order to be sure that we do not get any impermissible behaviour, for instance on different asset classes or environments (e.g. the VaR example on options).

chapter. Chekhlov et al. further elaborated models that generate sample V paths and optimize returns under drawdown constraints (MDD, ADD and $\alpha - CDD$). However, they did not incorporate drawdowns in their objective as such. Essentially, their optimization algorithm — like others [18][16][15][13] — is a maximum return objective under a maximum drawdown constraint. This gives us a set of mean-drawdown efficient portfolios, but does not pinpoint the optimal portfolio[12].

In summary of this section, we can say downside risk measures are an important improvement over symmetric ones. Further, the widely adopted VaR and CVaR measures can be examined within the wider family of LPM models. Finally, we said that drawdown measures are dynamic generalizations of (C)VaR measures and provided an argument for their use in optimization. They are dynamic, intuitive and conceptually easy to grasp. However, the reason why they have not yet been widely adopted, and Markowitz optimization is still omnipresent in e.g. the roboadvisory industry, is that it is an order of magnitude more difficult to solve for than variance. Including drawdown measures in your optimization problem requires the translation of your problem into a more complex LP program such as in Chekhlov et al. Even doing so, we argued this is more optimizing *with* drawdowns rather than *for* drawdowns. Drawdowns are not truly used as risk measure, but rather as constraints[13]. This is where portfolio construction with drawdowns still misses a valuable opportunity.

B. iVaR: minimizing stress for investors

"*The proof of the pudding is in the backtesting.*" (Emmanuel Wildiers)

So following our discussion in the previous section, what do we desire from our risk measure? It should take into account (1) the magnitude of losses, (2) the frequency of losses and (3) the time to recover from them. We want to design our risk measure in such a way that it is natural to investors. What is risk? It is more than a likelihood of losing money. In terms of the life of an asset manager, it is the amount of worried phone calls (s)he gets from clients (frequency), the severity of the call (magnitude) and the duration of the period these calls accumulate (duration of the loss). These elements constitute financial risk, or financial stress, from a human-centred perspective. Say we want to provide investors with a ride that is as smooth as possible, or as stress-free as possible, how can we measure both losses and the time to make up for them?

If we would try to account for both the size and frequency of *losses*, we need a downside risk measure and in particular a drawdown measure, since we also want to add *time* as a dimension. Minimizing these dimensions corresponds to penalizing deviations of our portfolio from *monotonic growth*. However, we argued optimizing for this is not as trivial as assessing some ex-post average drawdown of a timeseries, but requires solving a complex combinatorial problem for w. The proposed measure in this paper is the *expected average residual from monotonic growth*, from here called iVaR. iVaR is an integrated measure of risk, in the sense that it calculates the integral of drawdowns over time for many generated trajectories (see Fig. 2). iVaR is the average drawdown or

[12] Although one could opt for the respective maximum Calmar and Sterling portfolios.

[13] For instance, their rationale could be used by a roboadvisory app to simply maximize client returns while restricting maximum drawdown to 25% because otherwise clients would withdraw their money and delete the app. It is clear that this would be a useful application and drawdowns have advantages over traditional measures, but this does not formalize our perception of risk any better.

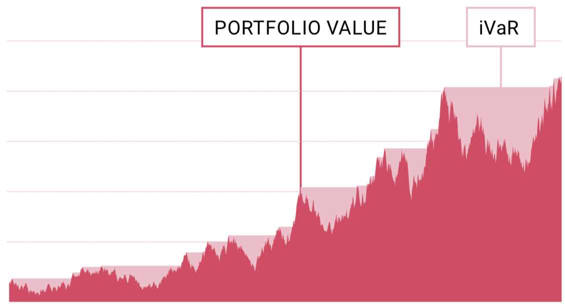


Figure 2: iVaR as an average expected drawdown (%)

time-weighted average distance between many simulated portfolios and their running maxima. From these simulated trajectories, we want the portfolio that provides us the smoothest ride, i.e. first we want to minimize iVaR.

Figure 2 illustrates iVaR graphically: the time-weighted average drawdown is the area under the imaginal horizontal line that establishes the running maximum and the actual portfolio value. Minimizing the integral of these areas by reconfiguring the weights of our portfolio corresponds to pushing our portfolio towards monotonic growth, or 'minimizing stress for investors'.

Mathematically, the closest closed-form expression to iVaR is (9), which we could use *ex-post* to assess an 'individual' iVaR of a specific portfolio. However, our goal was to translate this intuition into an ex-ante optimization problem. As such, we cannot assess iVaR in 'isolation'[14], but we can build on (9) in designing diversification measures.

In summary, as a natural extension to our previous discussion on risk measures, iVaR calculates the residuals w.r.t. monotonic growth and minimizes them explicitly in the objective of the optimization problem. This requires going from the relatively simple '$\max_w Return(w)$, $s.t. Drawdown(w) < D$' program we discussed before, to a more complex '$\min_w iVaR(w)$' problem[15]. As we argued, this is an order of magnitude more difficult and computationally intensive than mean-variance optimization, but provided current innovations in convex optimization and modern-day processing capacity this optimization opportunity has now become feasible. The goal of this paper is to optimize for iVaR, to provide investors with a smooth ride, while at the same time making sure that we do not invest in low-iVaR assets in a concentrated way. The easiest way to guarantee this is by adding additional constraints in terms of geography, sectors, and so forth, i.e. simply adding '*s.t.*'s to the algo. This is ad hoc, often arbitrary and almost always suboptimal. In the next chapters we will therefore measure and monitor for diversification using different viewpoints, ultimately tweaking the model's objective in our pursuit of optimal diversification.

[14] I.e. iVaR should be assessed at portfolio level after solving the MILP problem (thus be derived from the optimal objective value) and is conceptually less meaningful on individual trajectories. We do not provide the more technical definition of the problem here to keep the flow of the introduction, but it is provided in the next chapter.

[15] And its natural extensions (cf. next chapter).

II. What is diversification?

"*My ventures are not in one bottom trusted, nor to one place; nor is my whole estate. Upon the fortune of this present year: Therefore, my merchandise makes me not sad.*" (William Shakespeare in The Merchant of Venice.)

What is diversification? It sounds like an easy question since everyone has an intuitive sense of what *diversified* means. A diversified wardrobe, for instance, means one has a variety of clothes, covering the risk of extremely cold or warm weather. When we look for all-weather investments, how do we define diversified? Common phrases such as *"Don't put all your eggs in one basket"* and the Shakespeare quote epitomize our intuition: we spread out the risk of our portfolio over different investments such that our exposure does not stem from one source. Our asset allocation becomes a risk allocation, and next to an analysis of return contribution, we also want to know the risk contributions of our positions.

In a variance framework, these contributions are relatively simple to determine. It's a variance adjudication problem, much like all standard regression problems. However, the relationship between our need for diversification and the reliability of our covariance estimates is a bit convoluted. We will delve into this later in this chapter. For (conditional) VaR and drawdown measures, this is less trivial to determine. Therefore, we first need to answer some more philosophical questions about diversification.

The fact that we want to spread out our investment in different investments, implies that some assets are more similar than others. Hence, similarity and dissimilarity between assets, and how we measure them, is at the heart of quantifying diversification.

We should imagine our economy as an interconnected network of agents that essentially establish contracts between each other. These agents can be evaluated at the level of individuals or firms. Sometimes these contracts become more valuable (frequency and monetary value implied by the contract increases) and sometimes they become less valuable, or even default on their obligations. An economy is an interconnected system of these obligations, and it works like a watch. Sometimes one piece fails, and most other parts of the watch break down as well. This is the source of the risk, and some parts of the watch are more affected than others. The goal of diversification is to systematically identify these relationships and make sure not all of our positions are exposed to the same predictable source of risk. With predictable, we do not mean that we know when it is going to break down, but we know from the past there is a proven link between source X and asset Y. When we evaluate the investible assets in our horizon and calculate correlations, we assume this source is a latent source and we can observe it through its common effect on the time series of their returns. We are building a graph similar to the interconnected network of economic relationships, albeit an oversimplification. What we do is we see every company as a node, connected with every other company through an edge. A graph with 500 companies of the SPY, would generate 124,750 such relationships. Every edge has many attributes about that relationship: the nature of the relationship, the activity (if any) between the firms, whether they are in a similar industry, whether they are located in the same country, and so and so forth. These are all pieces of information that can help us in spreading risk over many sources. However, portfolio optimization literature tends to reduce this information to a single number: correlation. Correlation (Pearson's rho) is a linear measure of association that can be estimated from the return series and tells us the following. The value for every edge ranges between -1 and 1, will be 1 if they are identical in exposure to the source, 0 if they have no shared source of risk and -1 if they have exactly the same source but opposite exposure to it. However, correlation is just one way — and in the case of iVaR not the ideal one — to look at similarity or dissimilarity, as we will discuss later on.

The remainder of this chapter hinges considerably on this concept of the graph and how the most tractable frameworks see this graph. With correlation, most of it has been said, but we will delve deeper into the issues caused by this graph for high-dimensional settings (the Curse of Dimensionality and Markowitz' Curse of Diversification) and how literature tried to resolve it (using shrinkage in robust optimization, or hierarchical relationships discovered with clustering). In terms of frameworks, we will look at modern portfolio theory, maximum deconcentration, minimum variance, inverse variance, maximum decorrelation, equal risk contribution, most-diversified and hierarchical risk parity allocations. Next, we make the link with iVaR in subsection B.

A. Different views on diversification

a. Markowitz' modern portfolio theory (MPT)

"*A man who seeks financial advice will not be happy with the suggestion to maximize his quadratic utility.*" (Arthur Roy)

We again start with the father of portfolio theory, Harry Markowitz. Recall that the optimal portfolio according to the mean-variance criterion maximizes the Sharpe ratio (SR):

$$SR = \frac{E(r_p - r_f)}{\sigma_p} \tag{10}$$

This ratio represents the expected value of the excess return of the portfolio p over the risk-free rate f, divided by the risk or standard deviation of the portfolio. Therefore, MPT gives us the maximum Sharpe portfolio. The objective passed to the quadratic solver can also be written as[16]:

$$\max_{w}(\mathbf{w}^T\mathbf{r} - \frac{\lambda}{2}\sqrt{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}) \tag{11}$$

where $\mathbf{w}$ is the vector of weights we try to determine, $\mathbf{r}$ the vector of expected returns, Σ the variance-covariance matrix and λ a tuning parameter. In theoretical terms, we maximize quadratic utility that is linearly proportional with expected return and quadratically and negatively proportional with the risk, where λ represents the risk aversion of the investor.

Now what is the impact of diversification in such a framework and where do diversification benefits stem from? The power of diversification in a mean-variance framework is aptly illustrated in the next set of simple equations derived by Markowitz in the

[16]The max SR portfolio is then obtained by solving recursively for MV efficient portfolios, where the optimal portfolio corresponds to a critical level of risk aversion $\hat{\lambda}$. This is illustrated in Appendix 8.

1950s. If we define portfolio variance as:

$$\sigma_p^2 = \sum_i \sum_j w_i w_j cov(r_i, r_j) \tag{12}$$

The risk of the portfolio will decrease if we add low (ideally negative) elements to our covariance matrix, which will increase the Sharpe ratio of the portfolio. When we rewrite (12) using the average risk of our individual positions (13) and their average covariance (14), we find the following very interesting expression (15).

$$\bar{\sigma}^2 = \frac{1}{n} \sum_i \sigma_i^2 \tag{13}$$

$$c\bar{o}v = \frac{1}{n(n-1)} \sum_i \sum_j cov(r_i, r_j) \tag{14}$$

$$\sigma_p^2 = \frac{1}{n}\bar{\sigma}^2 + \frac{n-1}{n} c\bar{o}v \tag{15}$$

The first term of equation (15) tells us that by 'naive' diversification, i.e. simply adding random instruments to our portfolio therefore increasing n, we can drastically reduce portfolio variance. This explains why some empirical evidence [28] suggests that on average a portfolio of only 12 stocks exploits almost 90% of the diversification benefits [17]. Moreover, after adding 40 stocks to the portfolio the additional reduction in variance of naive diversification becomes negligible (our first term $\frac{1}{n}\bar{\sigma}^2$ is close to zero and remaining variance is due to $c\bar{o}v$).

The second term of (15) explains the power of 'smart' diversification, i.e. picking those instruments that are lowly (ideally negatively) correlated with each other. This requires managerial skill and boils down to picking instruments over different asset classes, geographies and industries.

This type of insight generated by MPT proves the common adage that models can be wrong, but still useful. MPT relies on simple assumptions (cf. introduction), but gives us insight in the relationship between expected variability in a portfolio — albeit a naive measure of variability —, the number of assets in our portfolio[18] and how we should pick them. It is very nice in theory, but breaks down in practice.

This brings us to our graph and the link with covariance misspecification. For every edge, we need to estimate a correlation measure. Consider the SPY, even if we can estimate one such value with relatively small error, the combined error of 124,750 such estimates can be enormous. This is the intuitive explanation for covariance misspecification and Markowitz' Curse of Diversification.

The more statistical explanation is that it is a Curse of Dimensionality. If we consider a universe of N assets, we need to determine N(N-1)/2 correlations. This is the relationship between 500 nodes and 124,750 edges. The Curse of Dimensionality generally refers to issues with statistical estimation where the number of parameters to be estimated scales badly with the number of units N. This leads to an ill-posed problem: an explosion in variance in the estimated coefficients and occasionally our

[17] The reduction in variance by adding random instruments, compared to a fully concentrated portfolio.

[18] For instance, the size of a reasonably diversified portfolio (be it in variance terms) is a crucial parameter for any fund.

solver tries to solve an underidentified set of equations. The latter would lead to a perfectly overfitted family of solutions, rather than a unique solution. In our case, the number of correlations scales quadratically with N. Given a fixed sample T for every N, we only have TxN observations, while we have N(N-1)/2 correlations. As a result, we need to collect at least $T > (N-1)/2$ to avoid underidentification, which is not a trivial job given T is restricted in time and N should be large (cf. our definition of rationality). Even in less extreme cases we typically consume degrees of freedom to such a large extent that the variance in the estimates of our covariance matrix is impractically large.

Secondly, the higher the average estimated covariance, the higher these errors. This means that the more correlated the assets in our scope are on average, or the higher our actual need for 'smart' diversification, the larger the errors in the solution will be. From a more statistical viewpoint, this is because the estimated covariance matrix $\mathbf{\Sigma}$ is ill-conditioned, i.e. it has a high condition number which is the ratio of the maximum and minimum eigenvalues[19]. This inevitably leads to large errors in our eventual solution, as we need to invert $\mathbf{\Sigma}$ according to CLA. In more practical terms, the higher the average correlation in our universe, the more uncertainty in $\mathbf{\Sigma}$, the more uncertainty or less stable the solution $\mathbf{w}$ will be. This unfortunate fact about MPT and CLA is known as Markowitz Curse of Diversification [33]. Statistically this implies high standard deviations around the estimated weights, and large changes in weights when the input data changes a little bit. In practical terms, this means instability, high concentration and high turnover when rebalancing your portfolio.

A commonly used technique to reduce the uncertainty in $\mathbf{\Sigma}$ is shrinkage and robust optimization. The term shrinkage refers to a broad class of penalized estimations. Penalized essentially means we both try to minimize the deviation of our estimation from its true value[20], as well as the number of parameters to do so. A penalty forces our algorithm to estimate on average smaller covariances, but with smaller variance in the estimate. The most popular alternatives are Tinkhonov or Ridge estimation, which uses an L2-norm penalty[21], and the Least Absolute Shrinkage and Selection Operator (LASSO) which uses an L1-norm penalty[22]. In the jargon of robust optimization, these are called the quadratic and box uncertainty terms in our objective function [23]. Both of them make our estimate biased — the estimated value has a different expectation than the true value — but reduce variance — as the tuning parameter of the penalty somehow ends up in a denominator in the variance formula. There is no need to prove this here. What is important, however, is that L2-like penalties can never shrink correlations to exactly zero[23]. This effectively means cutting edges in the graph for assets that are very lowly correlated and only introduce additional variance in the estimation. With LASSO, however, we can perform *covariance selection*, and shrinkage gets a more visual interpretation.

Figure 3 and 4 illustrate the idea for the constituents of the S&P500 and BEL20 (20

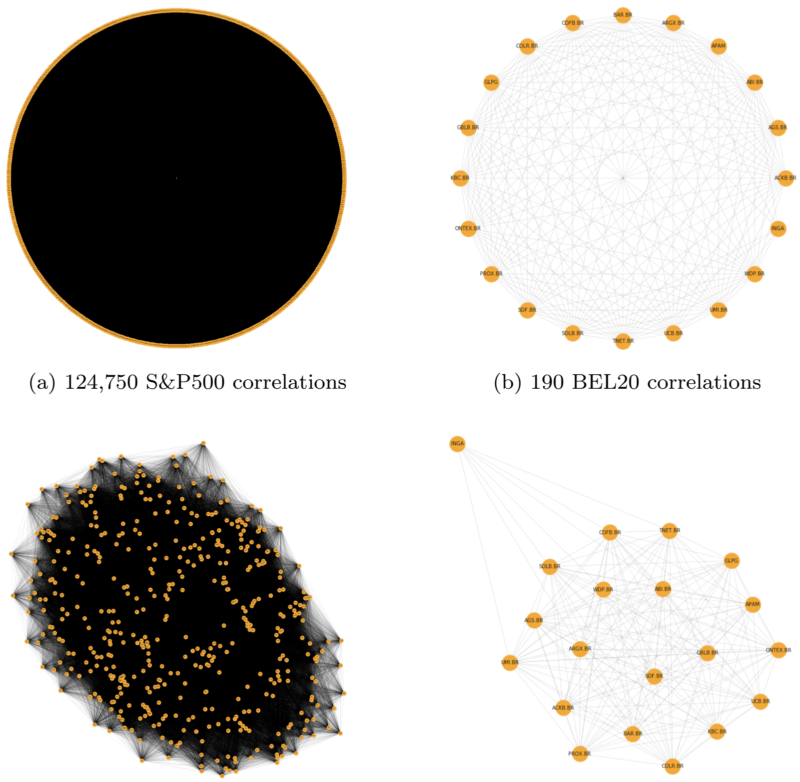
[19] We are not going to engage in lengthy discussions and formulae on eigenvalues here, but Appendix 4 reduces previous comments to a brief discussion on some stylized facts of financial correlation matrices.

[20] For instance using least-squares.

[21] Squared values of coefficients. This comes down to shifting each eigenvalue with a given offset, see Appendix 4 and 6.

[22] Absolute values of coefficients.

[23] The reason is basically that from the first-order condition derived from an objective function with squared coefficients, we can never set that coefficient to exactly zero.



(a) 124,750 S&P500 correlations

(b) 190 BEL20 correlations

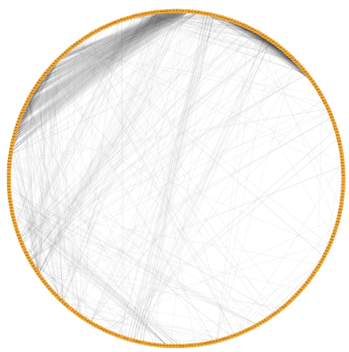
(c) S&P500 full cluster

(d) BEL20 full cluster

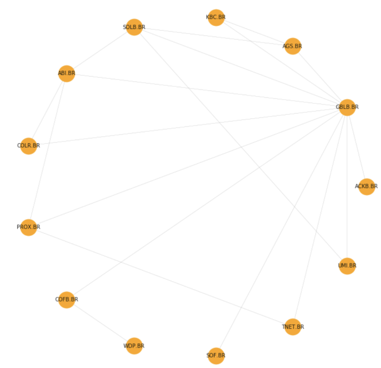
Figure 3: $\mathbf{\Sigma}$ according to Markowitz

largest stocks in Belgium by market cap) respectively. Remember, the fully specified $\mathbf{\Sigma}$ requires N(N-1)/2 estimations of correlation for N assets, that is 124,750 and 190 correlations respectively. Figure 3 (a) and (b) show the full graph we introduced in the introduction of this chapter, where every edge corresponds to one number (the correlation) which summarizes the relationship between the two stocks. Because of their vast number, it was impossible to annotate them. Moreover, for the S&P500 the lines are so dense that we just see one black surface. Figure 4 shows the sparse counterparts of these graphs. Only the correlations that really matter were kept, corresponding to using an L1 penalty with, in this case, a rather large penalty tuning parameter. We now see that these relationships become more manageable to observe, and more importantly, they can be estimated with more confidence.

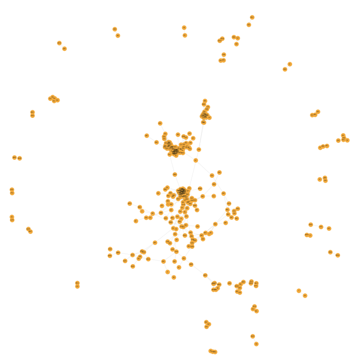
What is more interesting is that these figures aptly illustrate the link with clustering (we will delve into clustering in h. and Appendix 7). Reducing the full covariance matrix to only those relationships that are relatively important is similar to defining discrete sets of groups of assets that are more correlated. Figure 3 (c) and (d) show the full clusters of S&P500 and BEL20, i.e. the full graph where the highest correlated assets are put together, rather than on a radial axis. Figure 4 then shows the sparse cluster counterparts. We indeed see that only dense clusters or land masses of assets are kept, while the islands in between them are flooded. For a small number of assets such as the BEL20, this looks like a web structure, where assets close to each other and close to the center are most correlated[24].



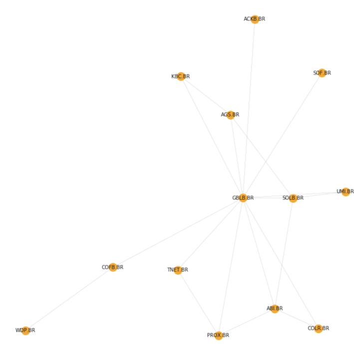
(a) Sparse S&P 500 graph



(b) Sparse BEL20 graph



(c) Example S&P500 cluster



(d) Example BEL20 cluster

Figure 4: Sparse Σ

b. Equally-weighted portfolios (EWP)

"*All of a sudden everything was highly correlated.*" (FT.com, April 2009)

Equally-weighted portfolios or maximum deconcentration portfolios are extremely simple in their set-up. With N assets the weights simply are:

[24]Important remark: assets that are not significantly correlated with any other asset are clusters on their own, and dropped out of this visualization.

$$\mathbf{w} = \begin{pmatrix} 1/N & ... & 1/N \end{pmatrix} \tag{16}$$

Their motivation is similar to the 'Talmudic rule': when you don't know how to distribute your wealth, just make sure you're invested in as many types of resources as sensibly possible (e.g. land, a house, a business, ...) and then allocate equally among these resources. This is extremely simple and sometimes very effective. Literature [24][36][37] suggests that EWP often outperforms MPT out-of-sample because of the overfitting inherent to covariance misspecification. It is only when we have a relatively low number of assets compared to the number of observations, and when we estimate covariance using the correct estimator [17], that MPT significantly outperforms EWP.

The idea of diversification in this setting is that we support on naive diversification (the first term in (15)), and that we invest in many assets and distribute equally because remaining reduction in risk might be spurious — think about the FT.com quote. Maximum deconcentration deliberately neglects a risk measure, i.e. concentration in the sense of risk contribution. EWP establishes diversified portfolios as we will never have concentrations that stem from spuriously uncorrelated assets like we could have in more sophisticated methods. In terms of the graph, EWP tells us to: (1) make it reasonably large and consider many nodes, (2) ignore information in the edges, and just (3) give every node the same piece of the cake.

The main shortcomings of this perspective is that it is naive, ignoring asymmetric returns and risks — however we measure it — between assets, etc. Therefore EWP will never give us an optimal solution, but it serves as a most transparent benchmark.

c. Minimum variance portfolios (MVP)

With minimum variance portfolios the idea is to minimize portfolio variance given a variance-covariance matrix Σ. Since returns are extremely hard to forecast, and the mean historical returns all but suffice to measure expected return, we minimize risk (the denominator of SR) while ignoring returns (the numerator of SR):

$$\min_{w}(\mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w}) \tag{17}$$

Minimum-variance optimization often outperforms mean-variance optimization out-of-sample. Ignoring $\mathbf{r}$ means we do not need its (spurious) estimation. From the perspective of diversification, this corresponds to the intuition of the second term of equation (15). We do not look at returns, so all we do is minimize $c\bar{o}v$. We look for the lowest (ideally negative) covariance terms and combine them to minimize σ_p^2. In terms of our graph, we try to detect the edges with the smallest correlation in the full graph (in case of no shrinkage/clustering) or a more sparse graph (in case of shrinkage or clustering). This is what we did for MPT as well, but with MPT there was still a tradeoff between finding the edges with the smallest value and the return their nodes provided. Now we just plot the graph and look for the smallest edges, without any constraints.

The major drawback of MVP, however, is that not looking at $\mathbf{r}$ is always suboptimal from a theoretical stance. From a practical one, there could be reasons why we can outpredict $\mathbf{r}$. With outpredict we mean that we do not necessarily have to be always correct to outperform MVP, as long as we better predict returns than average, we could expect some outperformance. A second drawback is related to the remarks about the uncertainty in $\mathbf{\Sigma}$, which are completely analogous to MPT.

d. Inverse variance portfolios (IVP)

Inverse variance portfolios are similar to MVP and also avoid the hassle of return forecasting, but they try to provide an answer to covariance misspecification by reducing the amount of inputs to the optimizer. We simply assume the covariance matrix is a diagonal matrix, i.e. all covariances are zero, and give weights that are inversely proportional to the variance of each asset: $w_i \propto 1/\sigma_i$.

$$\mathbf{w} = \begin{pmatrix} \frac{\frac{1}{\sigma_1}}{\sum_{j=1}^{N} \frac{1}{\sigma_j}} & \cdots & \frac{\frac{1}{\sigma_N}}{\sum_{j=1}^{N} \frac{1}{\sigma_j}} \end{pmatrix} \tag{18}$$

Inverse variance portfolios are also known as risk parity portfolios. Risk parity [29], pioneered and popularized by Bridgewater Associates, states that if we consider a broad investment universe (with a wide range of classes such as stocks, bonds, credit-related securities, commodities, real estate, inflation-protected bonds,...) the correlation over these classes is typically low, such that we can leapfrog the full covariance estimation and only look at the diagonal (variance). Inverse variance, if properly implemented (this hinges on the universe) provides a relatively simple-to-grasp, low-cost and low-risk portfolio.

Risk parity is thus very simple but crude. It is only a reasonably efficient allocation if we indeed have a diagonal covariance matrix, and can ignore covariances. Therefore, on first sight, risk parity neglects diversification benefits. This is only on first sight because of the way we set this up. In term of our graph, risk parity assumes an asset manager considers the full graph and is skilled enough to only keep those assets that are not linked to each other and reduce it to a sparse graph with no links. This is a pretty heroic assumption. This is an ad hoc, non-systematic approach and completely hinges on the skill of the manager. HRP takes the attractions of IVP and tries to remediate this limitation using machine learning (cf. h.), where we rearrange our non-diagonal covariance matrix in a way that it is quasi-diagonal.

In summary, is there no link between diversification and risk parity? Quite the reverse, if we are able to learn the full graph and know which assets are sufficiently dissimilar, inverse variance is the appropriate way to spread risk over these uncorrelated groups. How we determine these subsets, (1) manually/qualitatively using expert judgement (like risk parity), or (2) by using quantitative tools like machine learning, that is what makes it interesting.

e. Maximum decorrelation portfolios (MCP)

Maximum decorrelation portfolios (MCP) were introduced by Christoffersen et al. in 2011 [31], following a series of developments in dynamic correlation modeling. Similar to autoregressive variance (such as the generalized autoregressive conditional heteroskedasticity or GARCH models), researchers developed time-varying correlation models such as Robert Engle's DCC-GARCH (dynamic conditional correlation GARCH) or BEKK (Baba, Engle, Kraft and Kroner) models. Christoffersen (2011) proposes that with a sufficiently sophisticated treatment of correlation, portfolios can be constructed as follows:

$$\min_{w}(\mathbf{w}^T\mathbf{C}\mathbf{w}) \tag{19}$$

where $\mathbf{C}$ is the correlation matrix. Hence MCP is similar to MVP where we replace the covariance matrix by the correlation matrix. Similar to IVP, a maximum decorrelation portfolio attempts to reduce the number of inputs to the optimizer, but it uses the opposite assumption. Instead of focusing on volatility and ignoring correlation, the strategy assumes that individual asset volatilities are identical, such that the covariance matrix can be reduced to a correlation matrix.

In terms of our graphs, we can say similar things as with MVP. MCP portfolios look for edges with the lowest correlation. The difference with MVP is that MVP looks for minimum covariance, or pairwise correlation multiplied with the two individual variances. MVP therefore sometimes prefers more risky assets if they are negatively correlated, while obviously preferring low-variance assets if they are more (positively) correlated. MCP is ignorant for variance and assumes they are identical. Again from a theoretical stance, this is less optimal, but as we need fewer parameters it can give us better and more robust results. A quick comparison between MVP with dynamic Σ matrix versus MCP with dynamic $\mathbf{C}$ matrix makes an important case for MCP [31].

f. Equal risk contribution (ERC)

Equal risk contribution portfolios choose the weights in such a manner that all positions contribute equal amounts of risk $w_i[\mathbf{\Sigma}\mathbf{w}]_i$ to the portfolio risk $\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}$. Intuitively, this may sound similar to IVP, as we expect the weights to be inversely proportional to their risk. However, ERC takes into account diversification benefits implied by $\mathbf{\Sigma}$. High variance assets with low correlation can have a relatively low risk contribution. This concept of risk contribution is the very essence of what we intuitively understood by diversification in the MPT framework. Where ERC differs, however, is how it uses these contributions. ERC neglects expected returns and just imposes that every asset contributes to the portfolio risk equally. Therefore, the percentage risk contribution ($\%RC$) is compared with $1/N$ in the objective function, and the squared deviation is minimized.

$$\%RC_i = \frac{w_i[\mathbf{\Sigma}\mathbf{w}]_i}{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}} \tag{20}$$

$$\min_{w}(\sum_{i}^{N}(\%RC_i - \frac{1}{N})^2) \tag{21}$$

The resulting portfolio is similar to a minimum-variance portfolio subject to a diversification constraint on the weights of its components. The meaning of diversification in this setting is that diversification is the opposite of concentration. We penalize concentrations in terms of $\%RC$ in our objective function. An alternative could be to penalize high weights $\mathbf{w}$ in our objective function (e.g. a L1/L2 shrinkage on $\mathbf{w}$), but these weights do not tell us anything about the graph or our risk sources, and the portfolio can be well-diversified even if some are relatively large. They are the product of our optimization, and should not be confounded with the input. In short, the attraction of ERC is that risk contributions are at the heart of diversification. The drawback is that the $1/N$ is arbitrary, and there is no theoretical basis why equal would be the best, we just want to avoid too high $\%RC$.

g. Most diversified portfolios (MDP)

Most diversified portfolios were introduced by Choueifaty and Coignard in 2008 in their paper 'Toward Maximum Diversification' [20]. They provide us with a definition of diversification benefits: it is the reduction in risk between holding assets together in a diversified portfolio, versus holding them separately in fully-concentrated or undiversified ones. Therefore, for any portfolio w, they define the diversification ratio $D(w)$ as the following ratio:

$$D_{var}(w) = \frac{\mathbf{w}^T \sigma}{\sqrt{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}} \tag{22}$$

The numerator is simply the weighted average volatility of the instruments in the portfolio w, where σ is the square root of the diagonal of the covariance matrix $\mathbf{\Sigma}$. The denominator is the portfolio risk or standard deviation. Intuitively, the numerator ignores the benefits of low or negative covariance terms for portfolio variance (cf. a) and sets them all equal to one. This would be the risk of a hypothetical undiversified portfolio: if all correlations are one, the sum of the standard deviations is the standard deviation of the sum. The higher the risk in this hypothetical scenario compared to the actual portfolio that constitutes of these same assets, the higher the diversification benefits. The MDP portfolio simply maximizes $D(w)$ by altering w:

$$\max_w D_{var}(w) \tag{23}$$

The intuition behind maximizing this ratio is that the denominator makes sure risk is minimized, while the numerator maximizes individual volatilities (which we can reasonably assume is proportional with expected return). Indeed, maximizing diversification benefits effectively means minimizing portfolio risk without being ignorant for returns.

The attractions of MDP are obvious: (1) it provides us with a definition of diversification, it (2) minimizes portfolio risk, while (3) not being ignorant for returns. However, its results are again dependent on the quality of our covariance estimates. Evidence [37] suggests that it is most sensitive to covariance misspecification.

MDP's perspective on diversification is quite obvious. Diversification is the reduction in risk of holding assets together. Notice it is a 'positive' definition that focuses on benefits rather than the disadvantages of concentration (a 'negative' definition assumed in ERC). In terms of our graph, MDP has a rather different interpretation. The denominator tries to minimize the risk of our portfolio by looking for edges with small values in the graph (similar to what we discussed before), while the numerator reduces a scattered graph of islands to one pooled mainland. It assumes all edges have value one, or intuitively there is only one risk source and all assets are affected by it, the exact extent determined by their different variances. There is no way to diversify in such an unfortunate setting. So we just pick the ones with the maximal variance, because it is reasonable to assume they give us a higher expected return. The combined portfolio is the optimal tradeoff between these two effects, and maximizes the difference between the risk in the two perspectives.

h. Hierarchical risk parity (HRP)

Hierarchical risk parity was introduced by Marcos López de Prado in his 2016 paper 'Building diversified portfolios that outperform out-of-sample' [33]. The title gives much away of the gist of the paper as HRP tries to address three major concerns of quadratic optimizers in general and Markowitz' CLA in particular: in-sample it provides a theoretical optimal portfolio, but out-of-sample they (1) lack stability[25], (2) show concentrations[26] and underperformance[27]. HRP applies graph theory and machine learning techniques to build portfolios based on clustering the covariance matrix. In contrast to CLA, HRP is not prone to the conditioning of this matrix, as it does not require its invertibility. After delving into the CLA algorithm, its shortcomings and explaining Markowitz Curse, López de Prado introduced a new concept for diversification: *hierarchy*. The paper uncovers the root cause for instability, namely the fact that correlations are based on geometric or continuous relationships, rather than hierarchical ones. The essential idea behind hierarchical clustering is that not all assets should compete for allocation in a continuous manner. Some investments seem closer substitutes to another, and other investments seem complementary to one another. For instance, stocks could be grouped in terms of liquidity, size, industry or region, where stocks within a given group compete for allocations. To a correlation matrix, all investments are potential substitutes to each other. They lack the notion of hierarchy, and the very root of instability is that weights are thus allowed to vary freely in unintended (and often concentrated) ways. HRP's algorithm constitutes of three steps: (1) hierarchical clustering of the correlation matrix to obtain groups of similar assets at different levels, (2) quasi-diagonalization, which reorders the covariance matrix based on the groups of (1), and (3) recursive bisection, which recursively breaks up the covariance matrix into subsets and allocates weights over these clusters using inverse cluster variance.

1. Hierarchical clustering, and more specifically agglomerative or *bottom-up* hierarchical clustering (e.g. Fig. 5 on BEL20), is a form of clustering where we initially start with every unit being a cluster of itself, and then recursively add new units or clusters to our initial clusters until every unit falls under a single cluster. HRP applies linkage techniques on the correlation matrix. Linkage means we first transform our correlation matrix into a distance matrix, where we define a pairwise metric of dissimilarity $d_{i,j}$ proportional with one minus the correlation. Intuitively, distance or dissimilarity is one minus correlation or similarity. Next, the Euclidean distance $\tilde{d}_{i,j}$ between every two assets in this metric space is taken. The next step is defining the linkage criterion: which assets will we recursively add to a cluster? If we add two clusters from the previous step together, will we take the most similar or closest ones (minimum distance), the on average most similar ones (average distance), or the ones with the least dissimilar ones (complete linkage)? López de Prado proposes to use single linkage or minimum distance:

$$\tilde{d}(u,v) = min(\tilde{d}(u_i, v_j)) \tag{24}$$

[33] does not provide motivations as to why to use single linkage, but instead invites the reader to see Rokach and Maimon [22] for a more detailed description

[25] cf. Curse of Dimensionality
[26] cf. Curse of Diversification
[27] As returns are hard, if not impossible, to predict

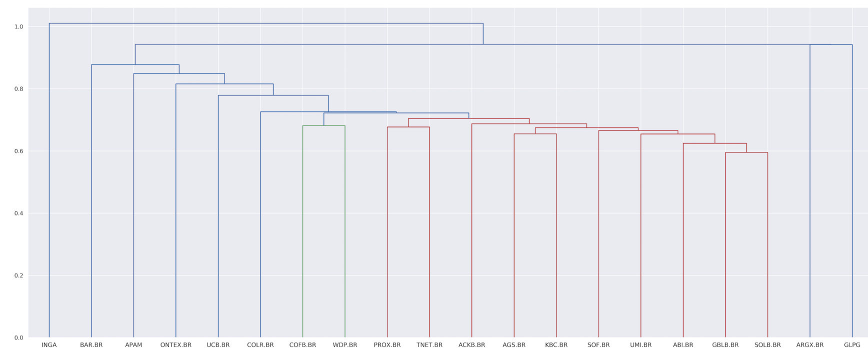


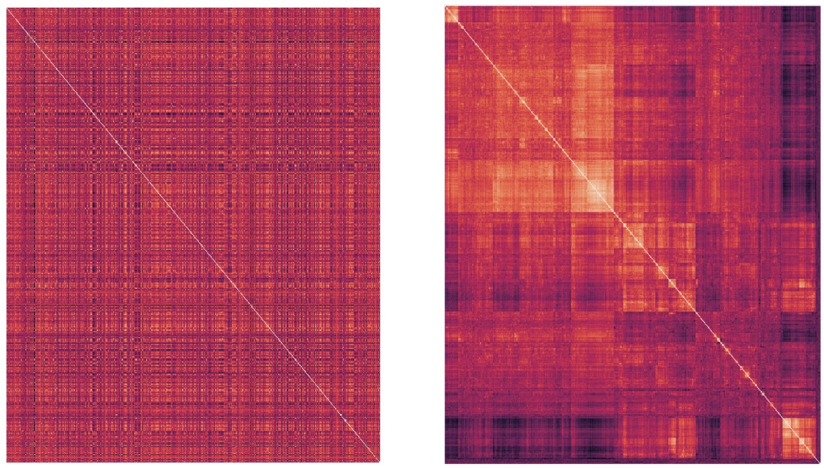
Figure 5: Hierarchial clustering on BEL20, the vertical axis is the distance metric $\tilde{d}_{i,j}$ at cutpoint

of the algorithms we described above, and to experiment with them in the Scipy library. This is the approach we will take in section G. and the code.

2. Quasi-diagonalisation means we rearrange our covariance matrix according to our clusters. If we reorganize $\mathbf{\Sigma}$ in such a way that highly correlated assets or similar assets occur together and dissimilar assets are placed far apart, the largest values of covariance will lie close to the diagonal. This has a useful property, since we know from d. that inverse variance allocation can be used for diagonal matrices. Clusters we obtain from subsets of this matrix can be used to get more robust inverse variance allocations, much in line with what we discussed in d.. Figure 6 illustrates the difference between an original, unclustered correlation matrix on the left, and the quasi-diagonalized one on the right. The interpretation of these plots is that the brighter the color, the higher the correlation. We can see that the most similar assets are placed close to each other, creating a bright box in the top left corner. The lower and right edges are much darker. Moreover, we can distinguish other boxes along the diagonal. These are our main clusters.

3. Recursive bisection is used in the next step to recursively assign weights over the tree. We use inverse-vol allocation on two subsets or clusters of our covariance matrix, and the weights are sequentially determined from root to leaf nodes as each cluster is bisected until we have a single asset in each cluster. To provide some intuition, Fig. 7 illustrates how weights could be determined from left to right[28]. First we compare the variance of ING bank (the most dissimilar asset) with the other cluster, and assign a weight according to inverse vol. Next, we compare the biotechs Argenx and Galapagos with the remaining assets in that cluster. We again apply risk parity, and so we continue. We understand that the assets that are most similar to all other assets in the graph will lie deeper in the tree. The deeper an asset lies in the tree, the more its eventual weight is a product of these recursive weights over subclusters. This makes sure assets that

[28] Remark this is not exactly how traditional HRP works, but this is closer to the cluster-based 'waterfall' approach in [27]. We will extend HRP and make it more similar to this logic in Appendix 7.



(a) Original correlation matrix

(b) Quasi-diagonalized matrix

Figure 6: Quasi-diagonalization

do not provide diversification benefits get on average lower weights. This is the most intuitive way to explain recursive weight allocation, but it is not strictly the recursive bisection proposed by HRP. As Thomas Raffinot remarks in his paper [36], the original HRP paper (1) treats the dendrogram as equivalent to its *minimum spanning tree*[29], and (2) uses only the order of the assets, not the number of assets in each cluster as implied by the dendrogram. Therefore, in traditional HRP the single linkage algorithm provides us with an order, and these split decisions can be visualized by a dendrogram, but the recursive bisection in HRP does not use this dendrogram. This is visualized in Fig. 8. The solution to this problem is a simple tweak proposed in the same paper [36]. Another blatant limitation of single linkage that may have become obvious from the Fig. 7, is that it suffers from chaining. Reading from right to left, we only need a single observation that is close to an existing cluster to add the whole cluster to it. As a result some clusters are really large, while others are extremely small (e.g. ING). In case ING would be a relatively low-risk asset compared to the diversified pool of other assets, one would end up with a concentrated exposure in ING. The solution for this problem is using a different linkage criterion, as proposed before, or using robust single linkage[30] [30].

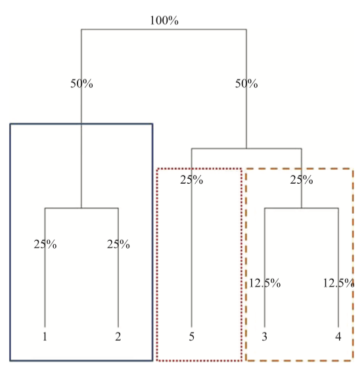
To sum up Hierarchical Risk Parity (HRP) and make the link with our graph, we can say that graph theory inspired HRP is intuitively the closest technique to our

[29] A clustering tree strictly related to single linkage.

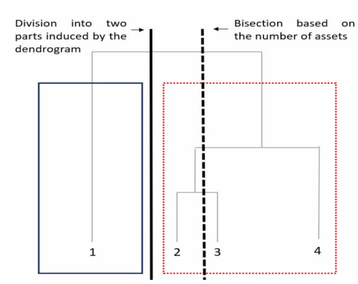
[30] Robust single linkage uses the minimum distance or single linkage criterion, but tweaks the distance measure. The distance is the larger of the Euclidean distance between two points and the maximum distance to k neighbours of each of the two points. This makes sure that we correct for the *density* of these points, and no isolated island of a point can serve as a bridge between two mainlands or clusters. Alternative clustering options are also extensively discussed in Appendix 7.

```
   /-INGA
  |
  |      /-BAR.BR
  |     |
  |   /-|   /-APAM
  |  |  |  |
  |  |   \-|   /-ONTEX.BR
  |  |     |  |
  |  |      \-|   /-UCB.BR
  |  |        |  |
--|  |         \-|   /-COLR.BR
  |  |           |  |
  |  |            \-|      /-COFB.BR
  |  |              |   /-|
  |  |              |  |   \-WDP.BR
  |  |              |  |
  |  |               \-|      /-PROX.BR
  |  |                 |   /-|
  |  |                 |  |   \-TNET.BR
  |  |                  \-|
   \-|                    |   /-ACKB.BR
     |                    |  |
     |                     \-|      /-AGS.BR
     |                       |   /-|
     |                       |  |   \-KBC.BR
     |                        \-|
     |                          |   /-SOF.BR
     |                           \-|
     |                             |   /-UMI.BR
     |                              \-|
     |                                |   /-ABI.BR
     |                                 \-|
     |                                   |   /-GBLB.BR
     |                                    \-|
     |                                       \-SOLB.BR
     |
     |   /-ARGX.BR
      \-|
         \-GLPG
```

Figure 7: Recursive weight allocation BEL20



(a) Diversification benefits of clustering (Raffinot, 2017 [34])



(b) Limitation of recursive bisection (Raffinot, 2018 [36])

Figure 8: Diversification benefits and limitations of HRP, according to Thomas Raffinot

Inspiration	What	Solution/Implementation
ERC	Diversification = opposite of concentration	Penalize concentration in objective: (1) Penalize high %RC, (2) Penalize %RC that deviate from 1/N, (3) Shrinkage on weights in objective (L1 / L2 type penalty)
MDP	Diversification = the benefits, or reduction in risk, of holding assets together in a portfolio	Exploit the measured benefits: (1) Maximize difference sum weighted risk with portfolio risk (maximize diversification ratio), (2) Penalize adding low-benefit assets (weighted sum correlations) in objective, (3) Naively boost your model to take more individual risk, while minimizing portfolio risk
HRP	Diversification = spread risk over many *sources*, these determine discrete sets of assets (clusters)	*Optimally determine* these clusters, assume cross-cluster correlation is spurious and use inverse *risk* to allocate between clusters: HRP and derived methods of recursive weight allocation over clusters using different linkage criteria, clustering algorithms and risk measures (e.g. HERC by Raffinot [36])

Table 1: Diversification in standard construction frameworks

starting premise. HRP starts from hierarchical relationships between assets, rather than geometric ones. This corresponds to the most important similarities in the graph. In a previous subsection a., we already made the link between clusters and the sparse graph. HRP provides us with a *systematic* way to reduce the full graph to a sparse graph and spread risk over these clusters. It is analogous to the philosophy behind traditional risk parity, where positions were inversely proportional with risk and where we ignored (spurious) correlation across clusters. This was achieved through managerial skill in IVP versus clustering and quasi-diagonalisation in HRP.

i. Overview

In summary, what insights w.r.t. diversification do we take away from the most tractable construction frameworks in literature, i.e. the variance and hierarchical clustering-based techniques? This is summarized very briefly in Table 1. We will project these views on an iVaR framework in the next section.

B. iVaR and diversification

"*Make everything as simple as possible, but not simpler.*" (Albert Einstein)

What is the link between iVaR and diversification? How can we make a smooth ride diversified at the same time? From the introduction we know that a diversified allocation means a more robust allocation, as it is less prone to misspecification errors of 'spuriously precise' risk and especially correlation measures. Recall that robustness by design, rather than correction, is vital for rational investment decision making, especially in times of automated investment solutions where algorithms end up in live environments. Relying solely on a single objective such as 'minimize iVaR' can therefore only yield a mathematical optimum, but in practice it might lead to biased decisions. Imagine assets with close-to-perfect monotonic growth. These unfortunately do not exist, but let's say we do have an extreme outlier in our investible horizon in terms of 'smoothness'. Unrestricted iVaR portfolios will tend to show a concentrated position in this asset because nothing is preventing it to do so. The easy way to cope with this issue is hard-coded constraints, which always lead to suboptimal allocations. Even in less extreme cases an unrestricted iVaR algorithm might prefer some geographies or industries which will make sense in-sample, but lead to excessive concentration out-of-sample. For instance, large-cap stocks, industries with matured cash cows, geographies like US, safe commodities such as gold, and so forth might be preferred by an iVaR optimizer, while a robust allocation should be spread over many of these characteristics. So an unrestricted iVaR portfolio might fail in practice, but what about in theory ?

In theory, we want to achieve three things: a portfolio that (1) minimizes stress (or minimizes iVaR), (2) exploits diversification benefits (risk reduction by combining assets) and (3) avoid concentrations (in terms of individual weights or risk contributions). This and Table 1 give us hints about how we can change the minimum iVaR objective to account for (2) and (3).

In what follows we will mirror the different perspectives we took in A. on an iVaR framework. First, we define a measure of association in an iVaR framework: coiVaR. What is the risk of a portfolio of two equally weighted assets versus the sum of the two individual risks? This gives us a ratio between 0 and 1 which tells us how useful it is to add assets to a portfolio based on a pairwise assessment with the assets already in the portfolio. 0 means adding the asset to the portfolio will not increase portfolio iVaR. Given a non-zero expected return of this asset, this would clearly increase the efficiency of the portfolio. 1 means adding the asset to the portfolio increases iVaR as if we would just buy more from the assets we already have. In this case, for many obvious reasons (transaction costs, practical reasons, etc.) adding this asset will not improve our portfolio's performance, nor its robustness.

On the other hand, this *pairwise* assessment might make concentrations worse, for instance if our horizon is ill-specified. Imagine we start from a very limited horizon — limited not in number, but imagine they are all closely linked to a similar source in the graph. In such a case, average coiVaR would be high, and minimum iVaR would lead to concentrations in the few assets with lowest iVaR. Now we could penalize for high coiVaR assets, but we would consequently add few or no assets to the minimum iVaR portfolio, not solving the issue and possibly making concentrations worse. It is clear that the idea of coiVaR is interesting, but we need something more.

That is why we tweak the coiVaR measure a bit to redefine a *diversification ratio*. The diversification ratio $D_{iVaR}(w)$ is not a function of two assets, but of the weights of a portfolio w. Rather than a pairwise assessment, we check the risk of the portfolio in iVaR terms, compared to the risk of holding the instruments individually. Hence, all we do in this subchapter is a simple mapping of two essential concepts of diversification on iVaR: correlation and $D_{var}(w)$ become coiVaR and $D_{iVaR}(w)$. Indeed, we try to make things as simple and intuitive as possible, while trying not to be oversimplistic. Next, we use these concepts in our objective to optimize for both iVaR (MIP, IIP) and diversification (MDIP, HiVaR, PiVaR and MVMIP).

a. coiVaR and diversification ratio $\mathbf{D}_{iVaR}(w)$

The first measure we introduce to construct diversified iVaR portfolios is coiVaR. coiVaR is a pairwise association measure between assets which compares the risk of an equally weighted portfolio of asset A and B with the risk of holding A and B individually:

$$coiVaR(A,B) = \frac{iVaR(A+B)}{iVaR(A) + iVaR(B)} \tag{25}$$

This gives us a pairwise measure of potential diversification benefits, much like an inverse diversification ratio for a universe with two assets. As said before, if iVaR is subadditive and non-negative, coiVaR ranges between 0 and 1,

$$0 \leq coiVaR(A,B) \leq 1 \tag{26}$$

where adding low coiVaR assets to an existing portfolio will not increase overall iVaR much, and vice versa. The importance of the subadditivity (or a fortiori convexity) of our measure here can not be overemphasized. That is why we provide an argument for iVaR's subadditivity in Appendix 3. Clearly if iVaR would not be subadditive this ratio could be any number exceeding 1, and would be rendered useless for further analysis or optimization.

Low coiVaR assets essentially means that if we combine two time series they are close to monotonic growth, while the individual ones are not. This might sound counterintuitive. In practice, it means the drawdowns of A and B are lowly correlated in terms of size, in terms of timing of their occurrence and in terms of the time to earn them back. An extreme example might help: when one asset A shows relatively small drawdowns that takes ages to recover, temporarily adding a volatile asset B that can bring the portfolio above water more quickly, and is not correlated with the drawdowns in A, will be a great choice to lower portfolio iVaR. Mathematically, coiVaR is a very simple approach to capturing both the correlation process between the returns series and their volatility processes[31].

Analogous to correlation we can now estimate a coiVaR matrix by assessing all two-by-two coiVaRs. This matrix can be used for many portfolio analysis ends, such as

[31] For diversification benefits, the returns might be lowly (ideally negatively) correlated, but the volatility might be negatively correlated as well, such that drawdowns are correlated with more volatility or — similar to option pricing intuition — higher odds of recovering fast. In principle one could try to model the correlation structure of those returns and dispersion for every two assets. Essentially coiVaR captures a similar quantity. coiVaR is simpler and more crude than more elaborate models, but not necessarily an approximation. In the eloquent words of Emanuel Derman: '*Handwriting and typewriting are two ways to express ideas. But one is not an approximation of the other. Using a GPS system or stars to orientate are two ways to get to a destination. One might be more sophisticated than the other, but the one is not an approximation of the other*', as long as you in all cases achieve what you want to achieve: convey information, get to your destination or 'decorrelate' returns as well as time-under-water.

weighted average portfolio coiVaR as a measure of concentration or as a similarity matrix for clustering:

$$\mathbf{C} = \begin{pmatrix} 1 & coiVaR_{1,2} & ... & coiVaR_{1,N} \\ coiVaR_{2,1} & 1 & & \\ ... & & ... & \\ coiVaR_{N,1} & ... & & 1 \end{pmatrix} \tag{27}$$

It is clear that as an analysis tool coiVaR can be an interesting addition to our toolbox. However, in portfolio optimization it might prove insufficient. Firstly, we have the issue of inevitable concentrations when we apply a coiVaR penalized algorithm on ill-specified horizons. Therefore, coiVaR does not suffice to get all three properties of a diversified iVaR portfolio. Secondly, a matrix with N(N-1)/2 unique coiVaRs raises new versions of old problems, such as estimation error, stability and issues of dimensionality. Therefore, a second very simple measure of diversification is coined the iVaR diversification ratio:

$$D_{iVaR}(w) = \frac{\sum_i^N w_i iVaR_i}{iVaR_p(w)} \tag{28}$$

D_{iVaR} can be seen as an iVaR analogue of D_{var}, where we simply replaced volatility by iVaR. It compares weighted average iVaR of a portfolio with portfolio iVaR.

b. Undiversified iVaR: MIP and IIP

The initial optimization problem we start from is the minimum iVaR portfolio (MIP):

$$w = \min_w (iVaR_p(w)) \tag{29}$$

The resulting portfolio w promises us the smoothest ride, i.e. by combining assets in such a way that their combined time series is closest to monotonic growth. However, we do not impose any hard-coded constraints on maximum weights, industry exposures and so forth, which makes it prone to concentrations. This is nothing but a potential overfitting problem, where few smooth paths in-sample are used to cause relatively high weights out-of-sample. So again the question is how we can prevent the MIP from overfitting by tweaking the objective, rather than by imposing constraints. We firstly propose to always keep track of diversification measures when constructing portfolios: the so-called Herfindahl index (sum of squared weights, see below), D_{var} and D_{iVaR}. If after running the optimization excessive concentration exists, we would have to rerun the optimization with additional constraints, concentration penalties or by adding D into the objective. However, in times of roboadvisory and automated investment solutions, we cannot or do not want to manually assess diversification and have to rely on the model being diversified by construction. Possible answers to this question are given in c..

Before delving into these models, we briefly introduce another mirrored concept of what we have covered before, an inverse iVaR portfolio (IIP):

$$w = \left(\frac{1/iVaR_1}{\sum_i^N (1/iVaR_i)}, ..., \frac{1/iVaR_N}{\sum_i^N (1/iVaR_i)}\right) \tag{30}$$

Inverse iVaR is extremely naive, because it is in the combination of assets that the portfolio forms smoother paths. It ignores this fact completely and just picks the assets with individual low-iVaR paths. Nevertheless, IIP can be calculated really quickly without any solver and is closer to MIP than e.g. EWP or random starting weights. IIP initiated optimization allows us to speed up further solving, which is especially important in recursive algorithms like the hierarchical allocation. Moreover, IIP can serve as another very transparent benchmark.

c. Diversified iVaR portfolios: MDIP, HiVaR, PiVaR and MVMIP

In this subchapter, we will briefly describe the objective functions of four diversified iVaR portfolios: most-diversified iVaR portfolio (MDIP), Hierarchical iVaR portfolio (HiVaR), Penalized iVaR portfolio (PiVaR) and maximum-variance, minimum-iVaR portfolio (MVMIP).

The most-diversified iVaR portfolio (MDIP) maximizes the iVaR diversification ratio:

$$\max_{w}(D_{iVaR}(w)) \tag{31}$$

The idea is that when smooth portfolios can be created from rough individual paths, we are truly diversifying risk in the three dimensions that iVaR measures. Identically to the MDP intuition, the denominator pushes the solution towards the MIP, while the numerator pushes the weights away from concentrated low-iVaR positions. This is the most intuitive answer to our three criteria: (1) the denominator assures a smooth ride, (2) the maximized difference (or risk reduction) between numerator and denominator makes sure we exploit diversification benefits, while (3) the numerator also pushes the weights away from individual concentrations.

Hierarchical iVaR portfolios (HiVaR) spread out portfolio weights over clusters inversely proportional to cluster iVaR. The steps are the following:

- Step 1: Hierarchical clustering on correlation matrix (HiVaR-v and HCAA-iv) or coiVaR matrix (HiVaR-i and HCAA-ii)
- Step 2: Selection of the optimal number of clusters using the Gap index (cf. Appendix 5) in case of HCAA.
- Step 3: Top-down recursive division into subuniverses, determining MIP for each, and following an inverse portfolio iVaR allocation of each cluster: $w_1 = iVaR_1/(iVaR_1 + iVaR_2), w_2 = 1 - w_1$, where $iVaR_i$ is the portfolio iVaR of cluster i using MIP weights for that subuniverse. We follow the shape of the dendrogram in case of HCAA-v and HCAA-i, and simple recursive bisection in case of HiVaR-i and HiVaR-v.

We borrow from the intuition we developed in Table 1 and determine clusters of assets that reflect similar sources of risk, using both correlation (-v) and coiVaR (-i) matrices. Next, we split these clusters recursively, each time finding the MIP for these subuniverses. Then we give a total weight to that cluster inversely proportional to the portfolio iVaR of the cluster using these MIP weights. The use of the minimum iVaR weights makes sure that the portfolio iVaR is reduced by combining assets, forcing smooth trajectories, while the recursion itself makes sure that the most similar assets that lie deeper in the tree will have on average lower weights, forcing diversification.

Penalized iVaR portfolios (PiVaR) measure the concentrations of assets in terms of iVaR and penalize these explicitly in the objective function. Generally, we could define PiVaR as,

$$\min_{w}(iVaR(w) + \lambda C(w)) \tag{32}$$

while a logical choice for C would be the weighted average coiVaR:

$$C(w) = \sum_{i}^{N}\sum_{j}^{N} w_i w_j coiVaR(i,j) \tag{33}$$

This portfolio implements the third idea behind diversified iVaR portfolios and uses the intuition behind penalizing concentrations from Table 1. It is natural in a min-iVaR framework that we do not use correlation but coiVaR for this penalty. Moreover, PiVaR is superior to a simple L1- or L2-like shrinkage of the portfolio weigths, since coiVaR tells us something about shared risk sources, while weights do not. However, its major disadvantage (which it shares with traditional penalized methods) is that we need to determine an amount of penalization as implied by a hyperparameter λ. This parameter can either be determined through cross-validation, i.e. by determining which λ has best out-of-sample diversification and performance haracteristics, or it can be set by requiring some minimum level of diversification (e.g. a min D_{iVaR}) in the solution.

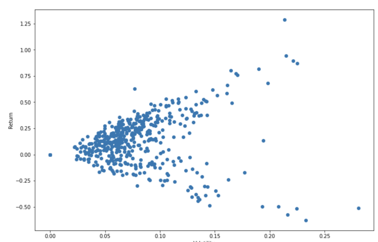
Finally, we introduce the maximum variance - minimum iVaR portfolio (MVMIP), which minimizes iVaR while maximizing for variance. MVMIP can be seen as a hybrid between MIP and MDP.

$$\max_{w}\left(\frac{\sum_{i}^{p} w_i\sigma_i}{iVaR_p(w)}\right) \tag{34}$$

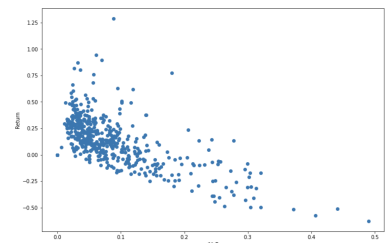
Recall that we argued that MDP requires the numerator to be proportional with expected return in order to be effective. There we argued that this was the extent to which symmetric risk measures, in contrast to our remarks in A., are actually very useful. In that sense our discussion of MDIP still misses an important point: what is the predictive power of an asset's individual iVaR for its expected return and what is the relationship between portfolio iVaR and portfolio return?

Figure 9 illustrates the average return, measured volatilities and measured iVaR for all the individual S&P500 stocks between 2015 and 2020 in panel (a) and (b). Panel (c) and (d) show simulated portfolio returns versus vol and iVaR characteristics for 10,000 random portfolios constructed from these stocks. It is clear from (a) that volatility is a symmetric risk measure, such that high individual risk corresponds with both higher and lower returns in panel (a). From panel (c) we can say that if we are on the efficient frontier (top half, and top left edge of the graph) we will indeed find a higher expected return in exchange for additional volatility. MPT will find the portfolio that maximizes the ratio of the two. Note that the color indicates the SR.

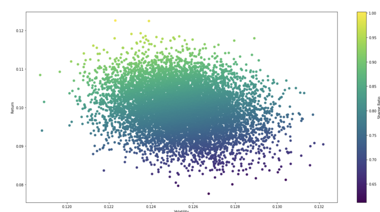
This is not the case for iVaR. The cloud with return-iVaR tradeoffs between individual assets in panel (b) is clearly not symmetrical. We can see that the highest return assets generally have low iVaR, while high iVaR assets have the lowest returns. The 10,000 simulated iVaR portfolios in panel (d) tell us that the return-iVaR tradeoff is maximized when iVaR is minimal. The elliptic opportunity set is rotated, such that more of the higher returns correspond to lower iVaR. That is why we need to be careful with MDIP. It will maximize diversification benefits in iVaR terms, and push MIP away from concentrated positions, but the argument that the numerator maximizes risk because this is proportional with expected return is no longer valid. For variance, however, this is the case. That is why we introduce MVMIP, which minimizes iVaR in the denominator and pushes MIP away from concentrations in the numerator using volatility.



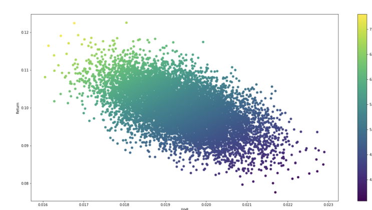
(a) Vol asset cloud



(b) iVaR asset cloud



(c) Vol portfolio cloud



(d) iVaR portfolio cloud

Figure 9: Asymmetry of drawdowns and return shown for S&P 500

In conclusion, there is no reward in managing portfolios without taking risk, but there is no such thing as a single risk. There are good risks (symmetrical or two-dimensional risks), and there are bad risks (downside risks). If the rational investor is on the efficient frontier, (s)he is rewarded for taking on additional ex-ante volatility, because (s)he has correctly identified that the investment can go in both ways at that point in time. However, taking on additional iVaR is not rewarded since (s)he is simply picking stocks whose expected[32] *losses* are higher[33]. That's why we also test the MVMIP portfolio. In essence because we minimize iVaR while at the same time (1) using the useful property that the average vol in the numerator is proportional with expected return in the efficient solution region enforced by the denominator, and (2) we need not specify covariances. From an MDP perspective it inherits its attractions while (1) using a more realistic risk measure and (2) avoiding covariance estimation. However, we should be wary of the fact that the unit of our numerator and denominator are not the same anymore. Volatility is expressed as a percentage return,

[32]Ex-ante provided sufficient persistence in the loss measure, which is likely the case with a by construction dynamic measure such as iVaR.

[33]It is still unclear where this fits in into financial theory and utility theory, but clearly volatility indirectly *provides* utility — although $U = E(r) - \lambda\sigma$ — because σ can be written as a function of $E(r)$ in the efficient part of the feasible region. iVaR, or drawdowns in general, do not seem to provide this 'indirect utility'. Anecdotally, 'volatility is a PM's best friend and enemy', for iVaR we could drop the friend.

while portfolio iVaR is a percentage drawdown. This underscores the importance of proper care[34] in order for the solver not to 'game' this conceptual mismatch.

C. Approaches and algorithmic considerations: practicalities and optimalities

So far we have only delved into the objective functions and their intuitions. In terms of implementation there are many approaches and algorithmic considerations (and complications).

a. Generic nonlinear programming approach

Here we directly utilize a nonlinear programming solver such as Scipy optimizer to find the maximum Sharpe ratio, maximum D_{var}, D_{iVaR} and so forth. We can for instance use (reduced) gradient algorithms and (quasi) Newton algorithms such as the Newton-CG (conjugate gradient) solver in the Scientific Python (Scipy) package. Moreover, we can use sequential methods for calculating gradients and finding optimal weights, such as the SLSQP algorithm. The advantage of this generic and direct nonlinear programming approach is that it has essentially no limitations in objective function, which allows for more easy implementation and experimentation with objectives[35]. The disadvantage is that, because of the lack of mathematical rigour in defining any nonlinear objective function, in some cases[36] the problem might become non-convex such that there is no optimality guarantee. By applying some more mathematical rigour (see below) we can obtain such an optimality guarantee and avoid numerical difficulties inherent to the more naive direct methods.

b. Formulation as a convex quadratic problem

The prototypical example of a convex quadratic programme (QP) is the mean-variance optimization we already discussed before:

$$\begin{aligned} \text{Maximise } \mathbf{r}^T\mathbf{w} - \frac{\lambda}{2}\sqrt{\mathbf{w}^T\Sigma\mathbf{w}} \\ \text{s.t. } \mathbf{A}\mathbf{w} \geq \mathbf{b} \\ \mathbf{e}^T\mathbf{w} = 1 \end{aligned} \tag{35}$$

where the objective function is the same one as Eq. (11), the first constraint matrix $\mathbf{A}$ and bounds $\mathbf{b}$ can be used for additional linear constraints on $\mathbf{w}$, and the second makes sure that the weights sum up to one as $\mathbf{e}$ is a vector of ones with as size the length of the weight vector. The other classical variance-based portfolios can easily be obtained by tweaking this objective to, for instance, minimum-variance, maximum decorrelation, etc. Moreover, Appendix 8 illustrates that (35) corresponds to a maximum Sharpe ratio portfolio for a certain optimal level of risk aversion $\hat{\lambda}$. However, solving the programme (35) still requires inputting λ, e.g. in a sequential procedure (which has the limitations of a.).

[34] Like proper standardization of inputs when e.g. applying MVMIP to different horizons.
[35] i.e. small tweaks instead of defining a whole new MILP problem (see below).
[36] For instance when expected returns are all negative in the SR approach and the objective function can never assume positive values.

Alternatively, we can define the maximum Sharpe ratio portfolio literally as:

$$\begin{aligned} \text{Maximise } & \frac{\mathbf{r}^T\mathbf{w}}{\sqrt{\mathbf{w}^T\Sigma\mathbf{w}^T}} \\ \text{s.t. } & \mathbf{A}\mathbf{w} \geq \mathbf{b} \\ & \mathbf{e}^T\mathbf{w} = 1 \end{aligned} \tag{36}$$

which is not a standard QP any more. We can now essentially do two things: (1) factorize Σ as in [40] or (2) use auxiliary variables as in [25]. The second approach reduces (36) to the following problem:

$$\begin{aligned} \text{Minimize } & \mathbf{g}^T\Sigma\mathbf{g} \\ \text{s.t. } & \mathbf{r}^T\mathbf{g} = 1 \\ & \mathbf{A}\mathbf{g} - \zeta\mathbf{b} \geq 0 \\ & \mathbf{e}^T\mathbf{g} - \zeta = 0 \end{aligned} \tag{37}$$

which is essentially a minimum variance problem on scaled weights with additional constraints using the auxiliary variable $\zeta = \mathbf{g}/\mathbf{w}$ (see proof based on [25] in Appendix 9). This formulation allows us to also determine the optimal MDP ratio and weights in a similar manner where $\mathbf{r}$, expected return, is replaced by the volatilities σ.

Let us now introduce slightly more formal formulations of the iVaR problems. The minimum iVaR portfolio (MIP) minimizes the residuals with regard to monotonic growth, whose expected value we have called iVaR:

$$\begin{aligned} \text{Minimize } & \mathbf{r}^T\mathbf{w} \\ \text{s.t. } & \mathbf{m}_+ \geq \mathbf{m} \\ & \mathbf{m} - \mathbf{t}\mathbf{w} = \mathbf{r} \\ & \mathbf{e}^T\mathbf{w} = 1 \\ & \mathbf{l}_b \leq \mathbf{w} \leq \mathbf{u}_b \end{aligned} \tag{38}$$

where $\mathbf{t}$ are pre-estimated simulated trajectories, $\mathbf{m}$ is the problem variable that measures monotonic growth, $\mathbf{r}$ is the problem variable that measures the residuals w.r.t. monotonic growth. $\mathbf{m}_+$ is a simple transformation of $\mathbf{m}$ where we shift the index by one. Therefore we enforce in the first inequality that $\mathbf{m}$ at any date of the trajectory is equal to the running max or high-water mark. Notice that the average or expected $\frac{1}{T}\mathbf{r}^T\mathbf{w}$ for the optimal w^* and trajectory date size T corresponds to the iVaR of the portfolio.

Similarly, the most-diversfied iVaR portfolio can be formulated as:

$$\begin{aligned} \text{Maximize } & \frac{\mathbf{i}^T\mathbf{w}}{\mathbf{r}^T\mathbf{w}} \\ \text{s.t. } & \mathbf{m}_+ \geq \mathbf{m} \\ & \mathbf{m} - \mathbf{t}\mathbf{w} = \mathbf{r} \\ & \mathbf{e}^T\mathbf{w} = 1 \\ & \mathbf{l}_b \leq \mathbf{w} \leq \mathbf{u}_b \end{aligned} \tag{39}$$

where $\mathbf{i}$ are the known, pre-estimated individual iVaRs. This problem is clearly not an LP, but a linear-fractional programming problem (LFP). The constraints for this problem have to restrict the feasible region to those $\mathbf{w}$ where $\mathbf{r}^T\mathbf{w}$ is strictly positive. This is clearly the case because our denominator corresponds to the portfolio iVaR which is strictly positive per our definition of $\mathbf{m}$ and $\mathbf{r}$ in the constraints. To solve (39) we require a variable transformation to translate the problem into a standard LP program. We first use a variable transformation called the Charnes-Cooper transform [4] that holds if the feasible region is non-empty and bounded:

$$\begin{aligned} \mathbf{y} &= \frac{1}{\mathbf{r}^T\mathbf{w}}.\mathbf{w}, \\ \zeta &= \frac{1}{\mathbf{r}^T\mathbf{w}} \end{aligned} \tag{40}$$

such that $\mathbf{w} = \frac{\mathbf{y}}{\zeta}$, i.e. the scaled weights. Now the problem can be reformulated as:

$$\begin{aligned} \text{Maximize } & \mathbf{i}^T\mathbf{y} \\ \text{s.t. } & \mathbf{m}_+ \geq \mathbf{m} \\ \zeta\mathbf{m} - \mathbf{t}\mathbf{y} &= \mathbf{r}\zeta \\ \mathbf{e}^T\mathbf{y} &= \zeta \\ \mathbf{r}^T\mathbf{y} &= 1 \\ \zeta\mathbf{l}_b \leq \mathbf{y} &\leq \mathbf{u}_b\zeta \\ \zeta &\geq 0 \end{aligned} \tag{41}$$

Analogously, the maximum-variance minimum-iVaR (MVMIP) portfolios can be computed using (41), setting $\mathbf{i}$ equal to the vector of individual volatilities.

Finally, the penalized iVaR portfolio (PiVaR) can be formulated as:

$$\begin{aligned} \text{Minimize } & \mathbf{r}^T\mathbf{w} + \lambda\mathbf{w}^T\mathbf{C}\mathbf{w} \\ \text{s.t. } & \mathbf{m}_+ \geq \mathbf{m} \\ \mathbf{m} - \mathbf{t}\mathbf{w} &= \mathbf{r} \\ \mathbf{e}^T\mathbf{w} &= 1 \\ \mathbf{l}_b \leq \mathbf{w} &\leq \mathbf{u}_b \end{aligned} \tag{42}$$

where $\mathbf{C}$ is our pre-estimated coiVaR matrix and the other variables ($\mathbf{r}$, $\mathbf{w}$, $\mathbf{m}$ and $\mathbf{t}$) are identical to the previous problems.

Now that we have all the necessary mathematical formulations[37] we can use non-generic solvers such as the mixed-integer linear and quadratic (MILP/MIQP) solver MOSEK, a package which will allow us to check the optimality of the solution.

[37] Note that the hierarchical portfolios are just a recursive use of (38) over clusters

D. Overview

Name (Abbreviation)	Risk	Similarity	Objective
Modern Portfolio Theory (MPT)	σ	ρ	$\max_w (SR(w))$
Minimum Variance Portfolio (MVP)	σ	ρ	$\min_w (w^T \Sigma w)$
Equally-weighted Portfolio (EWP)	None	None	$w_i = 1/N$
Inverse-variance Portfolio (IVP)	σ	None	$w_i = \frac{\frac{1}{\sigma_i}}{\sum_i \frac{1}{\sigma_i}}$
Equal Risk Contributions (ERC)	σ	ρ	$\min_w (\sum_i (\%RC_i - 1/N)^2)$
Most decorrelated Portfolio (MCP)	σ	ρ	$\min_w (w^T C w)$
Most diversified Portfolio (MDP)	σ	ρ	$max_w (D_{var}(w))$

Table 2: Overview variance models

Name (Abbreviation)	Risk	Similarity	Objective
Hierarchical Risk Parity (HRP)	σ	ρ	inverse σ - inverse cluster σ
Hierarchical iVaR - variance (HiVaR-v)	$iVaR$	ρ	MIP - inverse portfolio iVaR
Hierarchical iVaR - iVaR (HiVaR-i)	$iVaR$	$coiVaR$	MIP - inverse portfolio iVaR
Hierarchical Clustering-based Allocation - variance (HCAA-v)	σ	ρ	inverse σ - inverse cluster σ
Hierarchical Clustering-based Allocation - iVaR (HCAA-i)	$iVaR$	$coiVaR$	MIP - inverse portfolio iVaR

Table 3: Overview hierarchical models [38]

Name (Abbreviation)	Risk	Similarity	Objective
Minimum iVaR Portfolio (MIP)	$iVaR$	None	$\min_w (iVaR(w))$
Inverse iVaR Portfolio (IIP)	$iVaR$	None	$w_i = 1/iVaR_i$
Most-diversified iVaR Portfolio (MDIP)	$iVaR$	None	$\max_w (D_{iVaR}(w))$
Penalized iVaR Portfolio (PiVaR)	$iVaR$	$coiVaR$	$\min_w (iVaR(w) + C(w))$
Max-var min-iVaR Portfolio (MVMIP)	$iVaR$	None	$max_w (w^T \Sigma w / iVaR_p)$

Table 4: Overview iVaR models

III. Model construction and comparison

"*My imagination has always been much broader than my skills.*" (Freeman Dyson)

Now that we have covered the different models in our model set, we will construct and compare them on real-world data. Firstly, we run some shorter experiments for the full model set on a smaller universe, the 30 stocks of the DOW Jones Industrial Average. Then, we select a large universe of ETFs and assess the diversified iVaR portfolios in more detail. Each time we run the model set over multiple periods and use the Hansen bootstrapped model confidence set (MCS, [26] and Appendix 10) to test which models significantly out- or underperform out-of-sample (OOS) in terms of (1) performance, (2) risk and, importantly, (3) diversification measures.

A. Backtesting on the DJIA using nonlinear solvers

a. Data

Our first experiments are done on the DOW Jones Industrial Average (DJIA) using its constituents on the 1st of Jan 2020[39]. Data is obtained from Thomson Reuters Eikon for the previous 5 years or approximately 1250 daily observations per asset. In the following backtests we use a recursive window approach to update portfolios on a monthly basis from 2017 on, with no forward-looking biases[40]. Each month of observations is added to the sample after rebalancing. Note that rebalancing and reoptimization happen with the same monthly frequency in these experiments[41], and transaction costs are neglected for now[42].

b. Results

– **A first glance at the portfolios**

To get a sense of how these portfolios look and differ, we firstly run the model set over the full sample and obtain the following weights (Table 7 and 8 in Exhibits VI.). It is clear that traditional MPT portfolios are prone to high concentrations (Table 7, in **bold**), which might be indicative of the overfitting and misspecification issues we discussed before. Similar remarks can be made for MVP. The weights are more spread out for HRP, MDP and the other more naive portfolios. Since the latter drastically underperform (see below) this might come down to

[39] As such the backtests are not taken *point-in-time*, such that comparison with DIA is complicated by survivorship biases. However, comparison across models remains valid.

[40] Meaning we use a 2-year offset for our recursive window to avoid overfitting the short window in the beginning and end up with instable results for all portfolios.

[41] Obviously this is an experimental set-up and the rebalancing frequency is an important strategic parameter for investors. Fixed monthly rebalancing is not optimal and would need to be accounted for in practical applications. We should consider time versus threshold rebalancing [32].

[42] Therefore giving a positive bias to instable methods with high turnover for each rebalancing.

overdiversification. We notice some high individual weights for HCAA as well, which can be explained by very small cluster sizes in a limited horizon[43].

The iVaR portfolios are generally more diversified, although the minimum iVaR portfolio seems to prefer some individual low-iVaR paths, such as PG[44]. Again we notice that variance-clusters[45] result in higher weight concentration (in JNJ[46]). PiVaR seems to be the most concentrated iVaR portfolio, which is related to the aforementioned fact that the additional penalty for similar positions might be counterproductive from a diversification perspective.

Table 9 shows the D_{var} and D_{iVaR} of the respective portfolios. It is clear that from both perspectives MPT is the most concentrated portfolio. As we expected, MDP and MDIP offer the highest D_{var} and D_{iVaR} respectively. However, one table of average D ratios does not tell us much. Let us therefore see how this generalizes over multiple periods in a dynamic backtest.

– **Backtests**

Table 5: Overview portfolio monthly OOS performance

	Return (%)	Volatility (%)	Sharpe	D_{var}	D_{iVaR}
MPT	*0.44051*^{M}	**0.884687**	0.497930	**1.240483**	2.200563
MVP	**0.20133**	*0.630451*^{M}	**0.319352**	1.369099	2.381367
MIP	0.35042	0.718709	0.487567	1.370251	4.236785
IVP	0.23545	0.694178	0.339184	1.391277	3.365166
IIP	0.29515	0.731306	0.403592	1.364977	3.417444
HRP	0.25491	0.683379^M	0.373014	1.402867	3.380617
HiVaR-i	0.25040	0.757109	0.330732	1.372132	3.274569
HiVaR-v	0.27627	0.738558	0.374071	1.373516	3.360227
HCAA-v	0.30818	0.801462	0.384525	1.276088	**1.630977**
HCAA-iv	0.29604	0.752568	0.393384	1.331814	1.885577
HCAA-ii	0.36341	0.739866	0.491188	1.340989	2.987645
MDP	0.34270	0.686219^M	0.499401	*1.448595*^{M}	3.863265
MDIP	0.42846^M	0.777709	0.550936^M	1.390421	*5.555551*^{M}
MCP	0.35447	0.713991	0.496474	1.443159	3.827174
EWP	0.26518	0.743198	0.356814	1.381726	3.487878
ERC	0.25797	0.706866	0.364952	1.397043	3.571168
PiVaR	0.35734	0.722096	0.494873	1.372388	4.292657
MVMIP	0.43428^M	0.780791	*0.556201*^{M}	1.364052	4.424851

1 Items indicated with ^{M} are in the $M_{90\%}$

2 Items in **bold** are worst performers, while items in *italics* are best-in-class

Table 5 shows the simple average of monthly OOS performance metrics of our strategies over the simulated 36 months. From a pure return perspective MPT is the superior model, while the $M_{90\%}$ includes MDIP and MVMIP as well. This

[43] In this case, the concentration is due to another type of overfitting (cf. Appendix 7) linked to individual low vol, low correlation assets (here JNJ) resulting in small clusters. This also makes correlation-based HCAA instable over multiple horizons (cf. infra). The linkage criteria also has an important impact on this. In these tests we used *average linkage* which proved more robust than e.g. *single* or *complete linkage* over multiple periods (cf. Appendix 7). Nevertheless, correlation-based clustering in the HCAA context was very instable due to our limited horizon and the instability of the similarity matrix itself.

[44] Proctor Gamble

[45] i.e. now using iVaR allocation for intracluster weights, while using correlation as similarity.

[46] Johnson & Johnson

can also be seen in Figures 12-14 in Exhibits VI.. Clearly, MPT is the only variance portfolio that is competitive in terms of returns. However, looking at diversification measures such as the sum of squared weights (Herfindahl index, [9]), D_{var} and D_{iVaR} we again find that MPT is by far the least diversified portfolio. Outperformance compared to other variance portfolios is most likely due to taking concentration risks, and would disappear when accounted for excessive turnover, i.e. transaction costs. The MCS for volatility suggests MVP as the superior model, while MDP and HRP are also in the $M_{90\%}$. This means a minimum risk investor who picks volatility as his measure can pick any of these models and will obtain statistically insignificantly different volatility. However, since variance can be written as a function of expected return in the efficient part of the feasible region, the global minimum variance problem corresponds to a minimization of returns. Interestingly MPT is the most risky strategy w.r.t. volatility. Additionally, the difference between iVaR portfolio volatility is small. In terms of the Sharpe ratio, MVMIP is the preferred model, while MDIP is also in the $M_{90\%}$. MVP minimizes the Sharpe ratio, while the max SR objective in MPT leads to an average SR out-of-sample, indicating the max SR in-sample does not generalize well OOS[47].

In terms of diversification measures, MPT is the most concentrated portfolio w.r.t. variance, while HCAA correlation-cluster overfitting leads to the highest iVaR concentration. The average diversification ratios of our different allocation strategies over the full backtest is summarized in Figure 11. As we ideally want high D ratios with low individual weights, we generally desire sharpe V shapes in this plot, while red flags should be raised when it flattens out for a certain portfolio. In one glance we can tell MDIP, MDP and MVMIP portfolios are generally well diversified, while variance-based HCAA and MPT portfolios suffer from concentration. Finally, it is no surprise that the MCS shows MDP and MDIP are the most-diversified portfolios for the D_{var} and D_{iVaR} measures respectively, while no other models are in their respective $M_{90\%}$.

– **Robustness and date breaks**
  We included similar performance tables to Table 5 in Exhibits VI. over individual years. Obviously, these small subsamples are of limited use given the small initial sample. Nevertheless, it splits up model performance for good versus bad market environments. Generally the relative performance of the models stays the same as we discussed over good market years (2017, 2019) and bad ones (2018).

c. Conclusion

In short, MVMIP and MDIP outperform on an overall basis, being in most model confidence sets (i.e. outperformance in terms of both return, risk and diversification). MPT is the only variance method competitive on returns, but this was linked to concentration levels beyond any iVaR model. PiVaR seems to be the least useful viewpoint, as a weighted average coiVaR penalty increases weight concentration significantly for a marginal gain in D_{iVaR}. Cluster algorithms using correlation lacked stability which interestingly makes coiVaR, as a dynamic measure more persistent by construction, a better choice for hierarchical iVaR portfolios. However, these conclusions are only preliminary given: (1) the short time sample, (2) the small universe and

[47] This is clearly because, although offering superior returns, the max SR strategy has highest OOS volatility.

(3) no optimality guarantee with the generic solvers[48].

B. MIP, MDIP, MVMIP or HiVaR? Which perspective prevails using the MILP solver?

a. Data

We constructed a diversified data set of 90 exchange-traded funds (ETFs), including 54 stock, 31 bond and 5 commodity ETFs from 1999 up to 2016. Importantly for this dissertation, the lower correlations between these asset classes should allow the competing models in the set to exploit more diversification benefits than for equity index constituents only. In the following backtest we simulate 10 years of rebalancing from 2006 to 2016. We included the 2008 market collapse to see which algorithm smartly diversifies and outperforms during market turmoil, without overdiversifying or resorting to a flight to safety in concentrated low-risk assets, missing the recovery afterwards. We again update portfolios on a monthly basis via a recursive window, i.e. we sample from our first observation up to the date where we reoptimize. We now include a variable transaction cost in the formulation of the MILP problem.

b. Results

Figure 10 highlights the power of diversification in an iVaR framework, i.e. from the *combination* of different assets from different asset classes one can create a genuinely smooth ride. The different iVaR perspectives are indicated in red, while a benchmark fund is given in blue[49]. In this large universe MIP is able to construct portfolios that very well approximate monotonic growth. It is very able to do so, but it does so through (1) relative concentration and (2) missing out on returns. The other portfolios are essentially a tradeoff of smoothness with additional return (either through iVaR or volatility boosts, or larger equity exposures as a consequence of clustering).

During the initial economically benign years of 2006-2007, the iVaR portfolios seem to miss out on the best performing securities (i.e. equities), and overdiversifying across bonds and commodities. During the market turmoil of 2008-2009 MIP, MVMIP and MDIP profit from large bond and commodity exposures. During the rapid recovery of the markets after the GFC HiVaR seems to outperform because of its large equity cluster[50]. In summary, our different models naturally all have periods where they outperform, although MDIP is the most consistent outperformer.

This can be seen from Table 6. MDIP offers competitive overall return with HiVaR, which clearly only outperformed in the second half of the simulation. Moreover, HiVaR is twice as volatile as MDIP, and compared to HiVaR the tail characteristics of MDIP returns are much closer to those of safe haven MIP (see Fig. 22).

[48] Given our large model set, a computationally demanding solver, and the vast number of iterations over clusters, more extended backtests over longer time windows and on a larger universe were computationally infeasible on a desktop computer. The following paragraphs will zoom in on the iVaR perspectives and use a longer and broader universe but on a reduced model set.

[49] More comparable benchmarks are plotted in Fig. 19.

[50] Instead of MIP on the full universe including bonds and commodities, we recursively use MIP over clusters, where equities are likely to form relatively big chunks of assets in this universe.

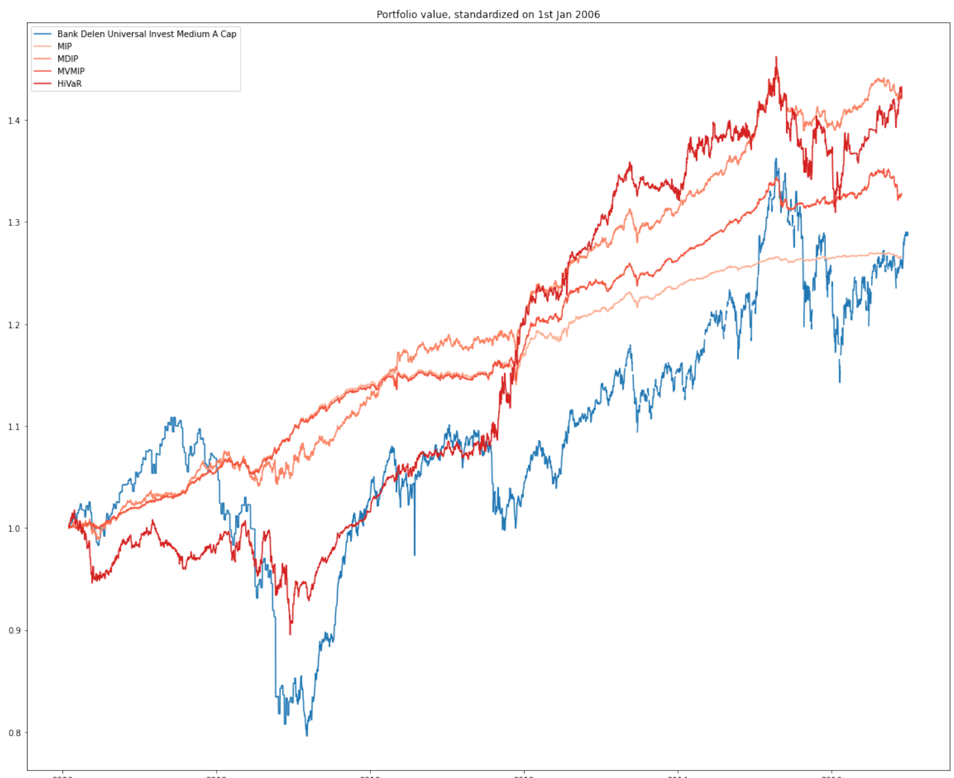


Figure 10: Long-term portfolios values for ETF universe

Table 6: Monthly OOS performance ETF universe

	Return (%)	Volatility (%)	Sharpe	D_{var}	D_{iVaR}
MIP	**0.204503**	*0.308152*^{M}	0.728379	**1.00525**	4.36087
MDIP	0.323616^M	0.686290	0.597445	1.03708	*5.67482*^{M}
MVMIP	0.252064	0.410271	*0.803194*^{M}	1.01888	5.43738^M
HiVaR	*0.332677*^{M}	**1.342111**	**0.388532**	*1.05305*^{M*}	**3.11179**

Items indicated with ^{M} are included in the $M_{90\%}$

In terms of diversification, all portfolios seem to be concentrated from a variance perspective. Given their low out-of-sample portfolio volatility, this means they are investing in low OOS volatility assets, therefore not exploiting any diversification benefits from a traditional portfolio theory perspective. At the same time the portfolio risk in iVaR terms is only a small fraction of the individual trajectory iVaRs. This implies that, on average, trajectories are combined to minimize iVaR rather than low-iVaR concentration. However, these are again only average D ratios. From the V-plot for this universe (see Fig. 20), we notice MIP has high weight concentration in low iVaR assets at least during some periods[51]. Taking a closer look at the portfolio constituency of MIP we notice high individual exposures to commodities and bonds during the GFC[52]. This proves to be optimal during the crisis, but also explains why MIP underperforms during the bull market thereafter. It is clear that from a diversification perspective MDIP does the best job with an especially low Herfindahl index and high D_{iVaR} ratios, while providing the most consistent and competitive returns afterwards.

Robustness checks in the form of date breaks are included for 3 year subperiods in Exhibits VI.. We have already briefly discussed the relative performance of our models over time, and the statistics in Table 12 confirm our view with MDIP being the most consistent outperformer.

c. Conclusion

Over an extended period of time and a large multi-asset class universe the most-diversified iVaR portfolio outperforms the other perspectives both in terms of consistency and size of returns, risk and especially diversification characteristics. Initially, it underperforms MIP and MVMIP because it pushes for higher iVaR assets in a down movement, but the underperformance is largely compensated for in the bull markets afterwards. MIP does a great job in realizing consistent returns and monotonic growth. This results in capital preservation in bear markets, which is then capitalized upon in the years after. However, it does this through relative concentration and ignorance for excess returns during these bull markets. The other portfolios leverage conceptually little tweaks to MIP to profit from these bullish periods. Clustering-based HiVaR applies MIP to subuniverses or clusters, in this case largely corresponding to the asset classes. As a result, during the market crash it goes sideways, as a middle way between commodity- and bond-driven increase of MIP and the equity-induced implosion of the benchmark. Afterwards, it provides the most expressed excess returns of all strategies. However, it does so with high volatility and low iVaR diversification,

[51] Exhibits VI. also includes a breakdown in multiple subperiods (including the downturn versus recovery) which clearly accentuates this phenomenon.

[52] i.e. a flight to safety as discussed in Appendix 11.

effectively giving away most of the attractive properties of the iVaR framework by introducing subuniverses. MVMIP is in the beginning of the simulation almost perfectly aligned with MIP, where alteration of MIP weights stays out during decline because overall high volatility does not push individual weights away from the MIP allocation. Afterwards, the relatively high-vol assets clearly do not necessarily correspond with high-iVaR assets, i.e. during the up movement, and large deviations between MIP and MVMIP exist. MVMIP is in all its characteristics (return, risk and diversification) literally a hybrid between MIP and MDIP. Apart from a maximum Sharpe ratio, it is the 'in-between' model in many respects.

IV. Conclusion, limitations and recommended further research

"*There is enough math in finance already, what is missing is imagination.*" (Emanuel Derman)

Diversification is a delicate topic, more than it appears on first sight. On the one hand it is conceptually simple and transparent, on the other hand it is easier said than done to guarantee optimally diversified portfolios out-of-sample. Markowitz argued that diversification is the only 'free lunch' in finance, where one can improve portfolio returns while reducing risk. In his framework, however, we have shown that (1) the nature of the risk measure and (2) issues of covariance misspecification, linked to the idiosyncracies of financial data[53], makes us often end up with undiversified and irrationally behaving portfolios, to the point of entirely offsetting the benefits of optimization [24].

The research question of this paper was to determine which perspective on diversification, assumed in over sixty years of portfolio theory and variance framework optimization, would work best in an iVaR framework. One classical option is to define and penalize for concentration. We argued focusing on weight concentration does not tell us anything about risk concentration, as a well spread portfolio in terms of weights might be undiversified and vice versa, such that traditional shrinkage is suboptimal. Penalizing for weighted average covariance is simply an MV-optimization, while penalizing for average coiVaR is not effective as it increased weight concentration for a marginal improvement in D_{iVaR}. A second perspective is to define and optimize for diversification benefits. Benefits are defined as the reduction in risk by combining assets. A straightforward measure is comparing the portfolio risk with the weighted instrument risk in a ratio, both from a variance and iVaR perspective. This viewpoint, we believe, proves most valuable in an iVaR framework. The third and more sophisticated approach is to learn a structure in the input data that reveals clusters or subuniverses that correspond to the most relevant risk sources in the graph. Although conceptually appealing, recursive use of an optimizer might give away its most attractive properties and requires additional qualitative validation of the clusters, something that is not desirable or even possible in e.g. an automated investment solution.

After all the previous comments and tests, we believe MDIP is a valuable enhancement compared to MIP. However, we clearly emphasized every model has its market environment in which it prevails. During rainy days or extended bear markets MIP outperforms, while on sunny days or bull markets HiVaR might come out on top.

[53] Related to both noise and signal/structure ([35] and Appendix 4).

If the art of diversification is — like in our more philosophical discussion — about constructing *all-weather* portfolios the MDIP and MVMIP, with a (purely empirical) preference for MDIP, are the most reliable portfolio models.

A more nuanced and probably better answer to the research question is that now that we have constructed a model set with clear insights in when these models work best, the optimal way to diversify is not to select a single best model but to combine these models dynamically over time, proportionally to their aptness at that point in time. Clearly, this increases model complexity assuming we have some predictive ability of the market environment. The general aim for anyone using these insights for portfolio construction should be to construct the best data set for estimation and calibration of the portfolio models, the clustering models, and so forth. Diversified iVaR portfolios should not be seen as siloed techniques to optimize portfolios but rather as a framework in which we can embed other financial predictive analytics techniques to, to continue the analogy, 'predict the weather'.

As a general conclusion, therefore, the power of having an alternative and potentially better portfolio algorithm is not merely in the algorithm, it is in the data. In the end most algorithms will be commoditised, but superior data will always exist. Data with the characteristics of the graph can help us with constructing better clusters, data discriminative for the market environment can assist us in attaching posterior probabilities to trajectories for iVaR calculation, as well as be used to select or combine the portfolio model(s) that would prevail in that environment, and so and so forth. It is clear, in line with the informational argument underpinning all rational portfolios from the introduction, that iVaR optimization is at most a necessary condition for more rational portfolios, certainly not a sufficient one. If iVaR portfolios do not prove to give more robust portfolios in practical applications, at least it is a step in the right direction. Especially with Figure 10 in mind it became apparent that if iVaR can not make you any money, it will at least save you some.

In terms of limitations, there are a great many more complications and parameters to take into account once we take this diversified iVaR framework from an experimental setting to a real-world application like a roboadvisor. These aspects were mentioned in the text, such as making the algorithm consistent with policies (constraints linked to mandates or regulations) while avoiding losing the diversification benefits when applying these hard-coded constraints. Additionally, many other limitations and complications were shortly mentioned in the text: the rebalancing strategy, performance fees, and other practical humdrum that complicates portfolio models in real-life. Moreover, before deploying the algorithms we would need validation of the results on many different universes with more assets, such as a multi-asset class universes with individual securities instead of (exchange-traded) funds. Recommended further research should start with tackling the above limitations and tilting the experimental use of the collection of portfolio algorithms, clustering algorithms, coiVaR and covariance estimators, et cetera to a more coherent framework. A start for this is given in the appendices. We could for a start dive deep into dimension reduction and clustering algorithms, and delve deeper than was possible for an MSc thesis. More research should be done on the mathematical formulations of the problems and the iVaR measures. How do we optimally calculate a coiVaR matrix? What are the more mathematically rigorous properties of iVaR, coiVaR and their relationship? If one thing is conclusive in this paper, then it is that the paper raises more questions than it solves. To quote Karl Popper: "*With every hypothetical solution of a scientific problem both the number* *of unsolved problems and the degree of their difficulty increase. They increase much faster than do the solutions. And it would be correct to say that whilst our conjectural knowledge is finite, our ignorance is infinite.*"

V. List of Acronyms

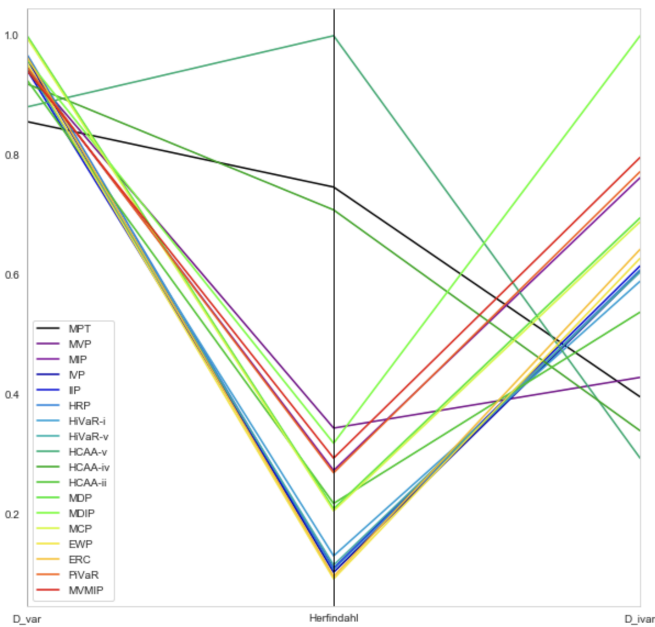
ADD	Average drawdown
BTP	Below-target probability
CAPM	Capital asset pricing model
CDD	Conditional drawdown
CLA	Critical line algorithm
coiVaR	iVaR codependence measure
CRM	Coherent risk measure
CTA	Commodity Trading Advisors
CUNY	City University of New York
CVaR	Conditional value-at-risk (= ES)
$\mathbf{D}_{iVaR}$	iVaR diversification ratio
$\mathbf{D}_{var}$	Variance diversification ratio
DJIA	Dow Jones Industrial Average
EMH	Efficient market hypothesis
ERC	Equal risk contribution
ES	Expected shortfall (= CVaR)
ETF	Exchange-traded fund
EWP	Equally weighted portfolio
FRTB	Fundamental Review of the Trading Book
GFC	Global Financial Crisis
GICS	Global Industry Classification Standard
HCAA	Hierarchical clustering-based asset allocation
HERC	Hierarchical Equal Risk Contribution
HiVaR	Hierarchical iVaR portfolio
HRP	Hierarchical risk parity
IID	Independent and identically distributed
IIP	Inverse iVaR portfolio
iVaR	InvestSuite value-at-risk
IVP	Inverse variance portfolio or risk parity
LFP	Linear-fractional program
LnP	Loss-and-profit distribution
LPM	Lower partial moment
MCP	Most decorrelated portfolio
MCS	Model confidence set
MDD	Maximum drawdown
MDIP	Most diversified iVaR portfolio
MDP	Most diversified portfolio
MIP	Minimum iVaR portfolio
MP	Marcenko-Pastur distribution
MPT	Modern portfolio theory
MV	Mean-variance criterion/optimization
MVMIP	Maximum variance, minimum iVaR portfolio
MVP	Minimum variance portfolio
OOS	Out-of-sample
PiVaR	Penalized iVaR portfolio
PnL	Profit-and-loss distribution
SR	Sharpe ratio
US	United States
VaR	Value-at-risk

VI. Exhibits

Table 7: DJIA weights variance portfolios

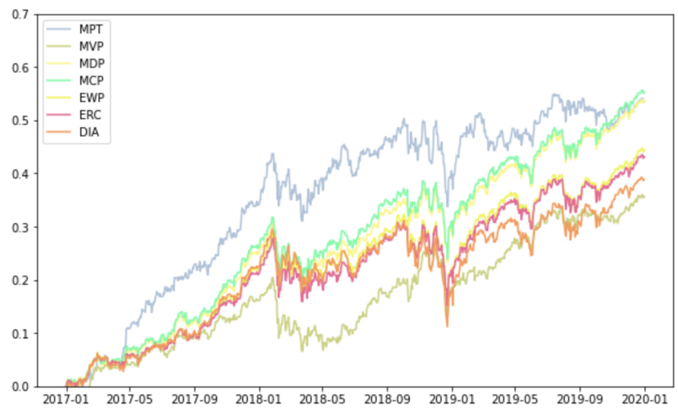
	MPT	MVP	IVP	HRP	HCAA-v	MDP	MCP	EWP	ERC
MMM	0.000	0.020	0.040	0.033	0.001	0.000	0.000	0.036	0.036
AXP	0.000	0.000	0.023	0.019	0.001	0.000	0.000	0.036	0.027
AAPL	0.082	0.000	0.027	0.025	0.037	0.040	0.046	0.036	0.032
BA	0.000	0.000	0.014	0.013	0.001	0.078	0.125	0.036	0.023
CAT	0.000	0.000	0.024	0.020	0.001	0.063	0.077	0.036	0.030
CVX	0.000	0.000	0.023	0.018	0.001	0.035	0.043	0.036	0.028
CSCO	0.000	0.000	0.031	0.045	0.043	0.000	0.000	0.036	0.032
KO	0.000	0.199	0.063	0.062	0.053	0.000	0.000	0.036	0.048
XOM	0.000	0.000	0.033	0.028	0.001	0.000	0.000	0.036	0.033
GS	0.000	0.000	0.025	0.031	0.002	0.000	0.000	0.036	0.027
HD	0.001	0.000	0.035	0.029	0.003	0.000	0.000	0.036	0.033
IBM	0.000	0.000	0.036	0.030	0.001	0.000	0.000	0.036	0.034
INTC	0.000	0.000	0.022	0.020	0.031	0.028	0.035	0.036	0.028
JNJ	0.000	0.121	0.058	0.090	**0.205**	0.000	0.000	0.036	0.047
JPM	0.000	0.000	0.027	0.021	0.002	0.000	0.000	0.036	0.027
MCD	0.074	0.092	0.042	0.034	0.004	0.064	0.059	0.036	0.040
MRK	0.000	0.024	0.045	0.043	0.159	0.086	0.077	0.036	0.044
MSFT	**0.500**	0.000	0.029	0.025	0.040	0.000	0.000	0.036	0.030
NKE	0.000	0.000	0.030	0.044	0.003	0.065	0.071	0.036	0.034
PFE	0.000	0.097	0.049	0.047	0.173	0.060	0.051	0.036	0.045
PG	0.000	0.038	0.055	0.054	0.046	0.031	0.025	0.036	0.048
RTX	0.000	0.000	0.029	0.027	0.002	0.000	0.000	0.036	0.030
TRV	0.000	0.000	0.037	0.030	0.004	0.000	0.000	0.036	0.035
UNH	**0.266**	0.000	0.028	0.029	0.007	0.033	0.037	0.036	0.033
VZ	0.000	**0.233**	0.060	0.068	0.051	0.167	0.128	0.036	0.056
V	0.077	0.000	0.034	0.029	0.046	0.000	0.000	0.036	0.031
WMT	0.000	0.148	0.045	0.051	0.081	**0.203**	0.179	0.036	0.054
DIS	0.000	0.030	0.034	0.034	0.002	0.047	0.048	0.036	0.035

Items in **bold** denote high concentration (>20%)

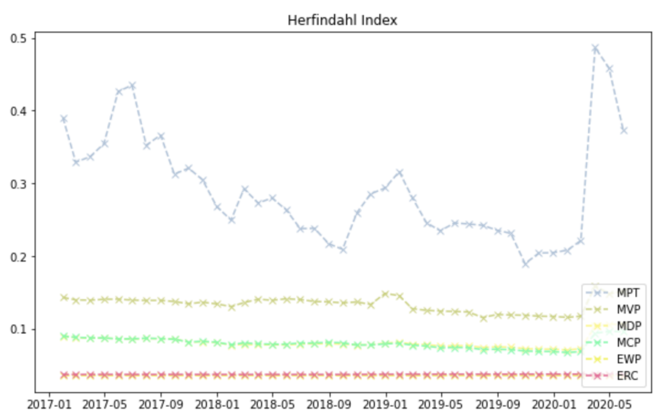


(a)

Figure 11: Diversification summary



(a) Cumulative returns



(b) Herfindahl concentration

Figure 12: Backtesting variance strategies on DJIA (1/2)

Table 8: DJIA weights iVaR portfolios

	MIP	IIP	HiVaR-i	HiVaR-v	HCAA-iv	HCAA-ii	MDIP	PiVaR	MVMIP
MMM	0.000	0.020	0.051	0.072	0.000	0.053	0.057	0.000	0.000
AXP	0.000	0.029	0.034	0.048	0.001	0.008	0.000	0.000	0.000
AAPL	0.054	0.026	0.069	0.036	0.052	0.000	0.034	0.000	0.052
BA	0.000	0.013	0.049	0.028	0.001	0.000	0.000	0.000	0.000
CAT	0.000	0.019	0.037	0.047	0.000	0.032	0.000	0.000	0.000
CVX	0.000	0.033	0.056	0.049	0.001	0.030	0.000	0.000	0.038
CSCO	0.000	0.034	0.026	0.044	0.059	0.000	0.000	0.000	0.000
KO	0.000	0.059	0.033	0.051	0.041	0.001	0.000	0.000	0.000
XOM	0.000	0.025	0.022	0.036	0.001	0.044	0.015	0.000	0.000
GS	0.002	0.021	0.017	0.028	0.001	0.033	0.030	0.000	0.012
HD	0.000	0.042	0.015	0.050	0.003	0.065	0.000	0.000	0.000
IBM	0.000	0.023	0.026	0.050	0.000	0.048	0.072	0.188	0.000
INTC	0.045	0.032	0.032	0.014	0.042	0.007	0.070	0.000	0.096
JNJ	0.000	0.053	0.026	0.033	**0.207**	0.001	0.000	0.000	0.000
JPM	0.000	0.036	0.062	0.055	0.001	0.000	0.000	0.000	0.000
MCD	0.052	0.045	0.022	0.040	0.004	0.078	0.062	0.119	0.056
MRK	0.110	0.047	0.040	0.050	0.160	0.015	0.080	0.000	0.101
MSFT	0.049	0.056	0.033	0.039	0.055	0.077	0.037	0.106	0.041
NKE	0.097	0.031	0.034	0.036	0.003	0.176	0.154	**0.232**	0.116
PFE	0.078	0.035	0.043	0.040	0.174	0.004	0.097	0.000	0.093
PG	**0.207**	0.053	0.056	0.028	0.036	0.018	0.092	0.000	0.167
RTX	0.000	0.033	0.063	0.023	0.001	0.000	0.000	0.000	0.000
TRV	0.000	0.041	0.045	0.027	0.003	0.000	0.000	0.000	0.000
UNH	0.042	0.034	0.028	0.025	0.007	0.002	0.025	0.113	0.031
VZ	0.187	0.046	0.024	0.017	0.039	0.005	0.026	0.018	0.111
V	0.020	0.056	0.022	0.015	0.064	0.089	0.000	0.000	0.017
WMT	0.047	0.027	0.010	0.011	0.041	0.015	0.121	**0.224**	0.045
DIS	0.010	0.028	0.025	0.011	0.002	0.199	0.027	0.000	0.023

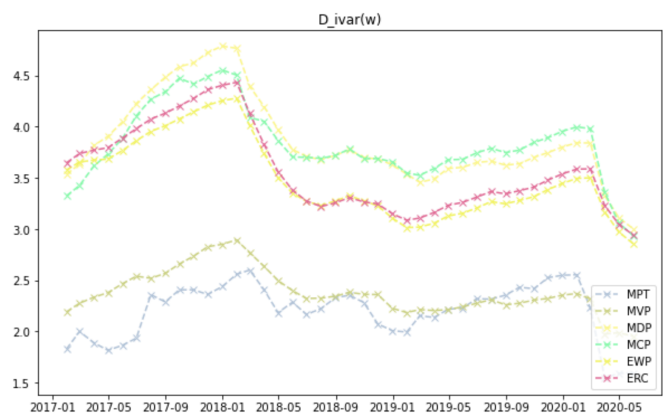
Items in **bold** denote high concentration (>20%)

Table 9: Average OOS diversification measures

	D-Variance	D-iVaR
MPT	**1.17407**	**1.68108**
MVP	1.37089	1.96133
MIP	1.40667	3.34261
IVP	1.4005	2.92405
IIP	1.37986	2.84318
HRP	1.40623	2.94113
HiVaR-i	1.36935	2.64163
HiVaR-v	1.36139	2.76254
HCAA-v	1.34698	2.37866
HCAA-iv	1.33296	2.46934
HCAA-ii	1.31312	2.51577
MDP	*1.48367*	2.9783
MDIP	1.41751	*3.71356*
MCP	1.47321	2.9003
EWP	1.38173	2.84531
ERC	1.40355	2.93217
PiVaR	1.35764	3.25967
MVMIP	1.41748	3.44135

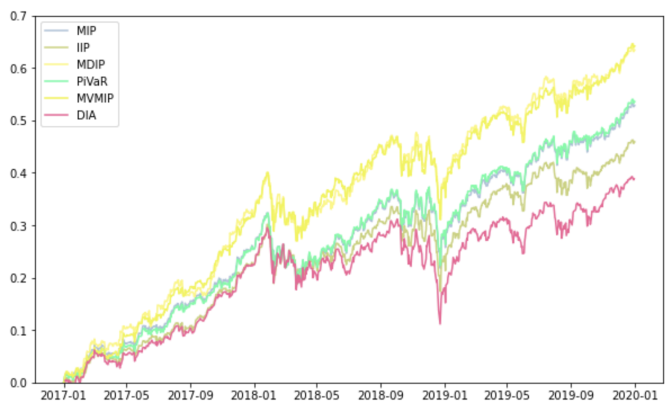
Items in **bold** denote worst performers
Items in *italics* denote best-in-class

(a) OOS D_{var}

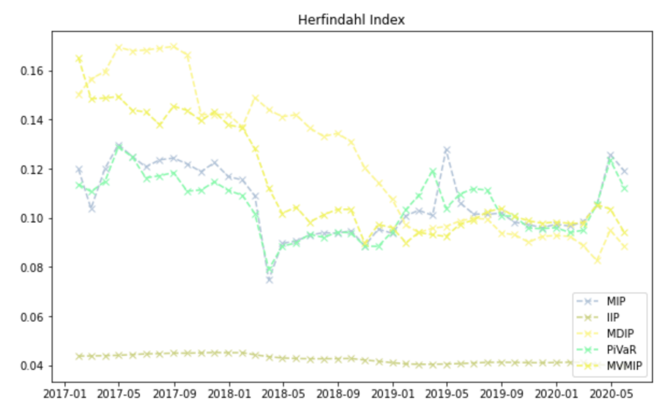


(b) OOS D_{iVaR}

Figure 13: Backtesting variance strategies on DJIA (2/2)



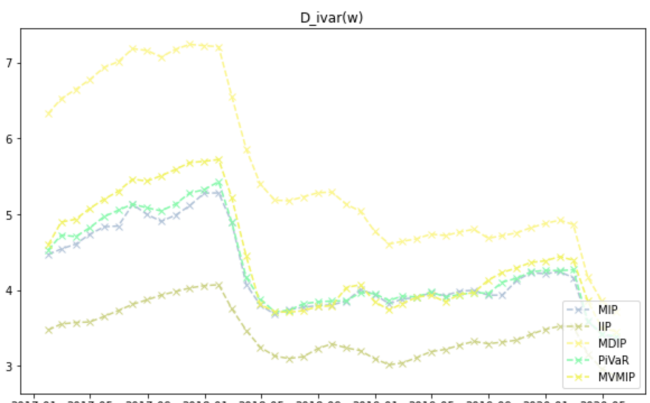
(a) Cumulative returns



(b) Herfindahl concentration

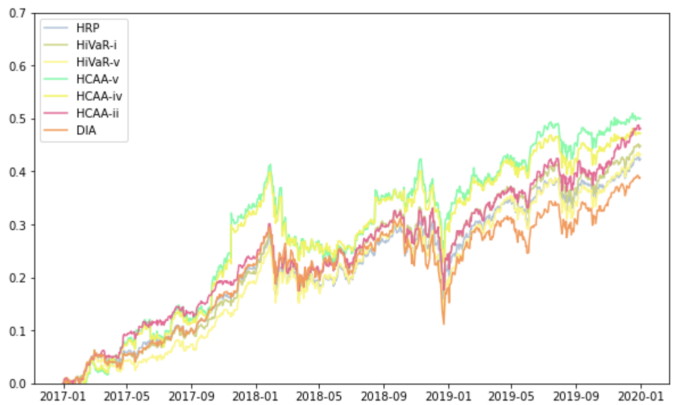
Figure 14: Backtesting iVaR strategies on DJIA (1/2)

(a) OOS D_{var}

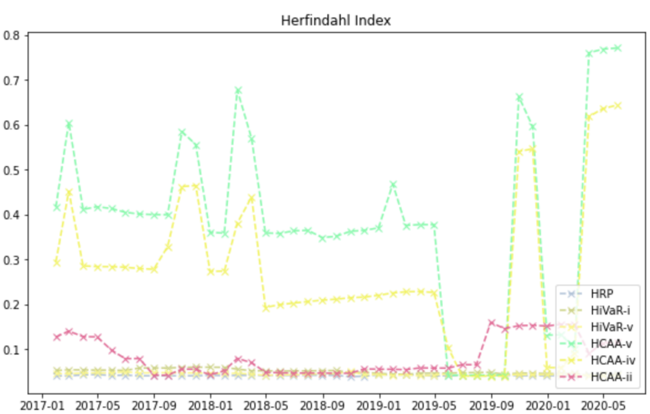


(b) OOS D_{iVaR}

Figure 15: Backtesting iVaR strategies on DJIA (2/2)

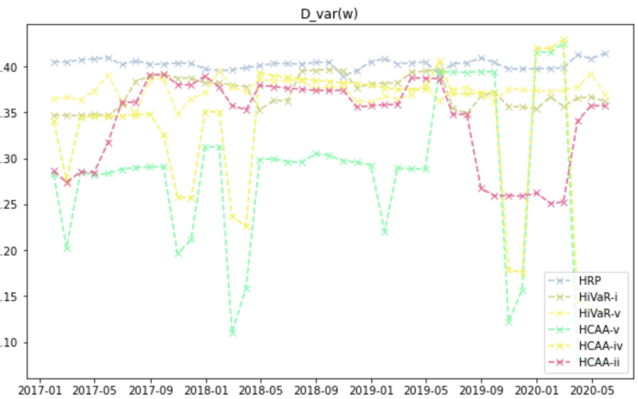


(a) Cumulative returns

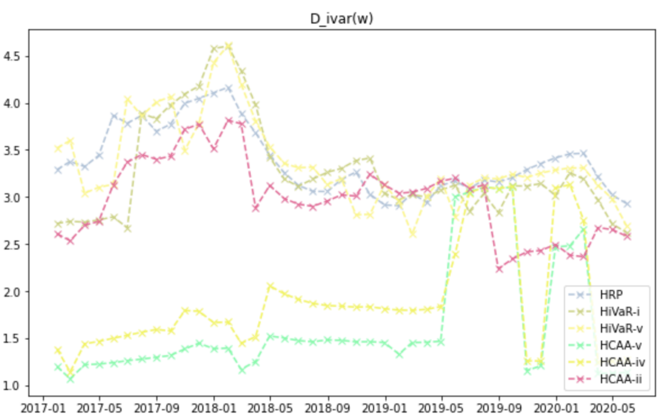


(b) Herfindahl concentration

Figure 16: Backtesting cluster strategies on DJIA (1/2)

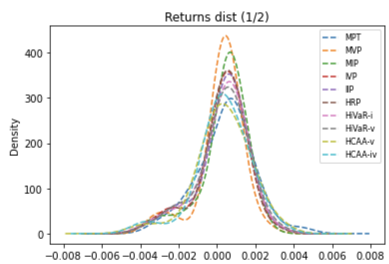


(a) D_{var}

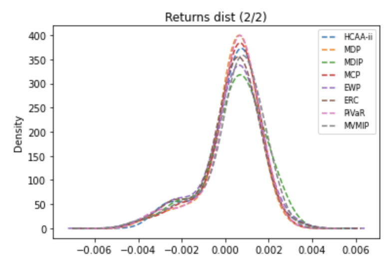


(b) D_{iVaR}

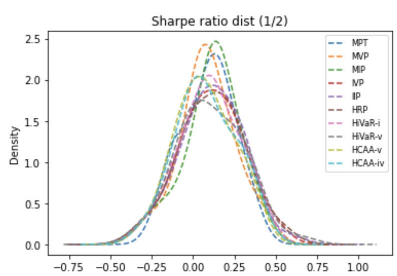
Figure 17: Backtesting cluster strategies on DJIA (2/2)



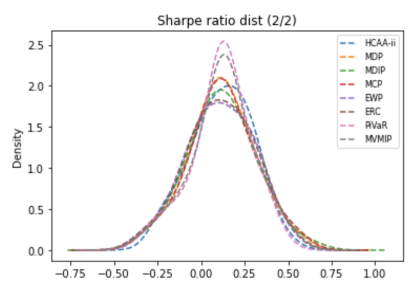
(a)



(b)



(c)



(d)

Figure 18: Simulated return and Sharpe distributions of all portfolios on DJIA

Table 10: Monthly OOS Performance DJIA, breakdown per year

		Return (%)	Volatility (%)	Sharpe	D_{var}	D_{iVaR}
2017	MPT	0.918	0.473	1.940803	1.229	2.133
	MVP	0.458	0.301	1.521595	1.365	2.529
	MIP	0.683	0.356	1.918539	1.367	4.870
	IVP	0.522	0.292	1.787671	1.393	3.718
	IIP	0.587	0.309	1.899676	1.362	3.768
	HRP	0.575	0.293	1.962457	1.405	3.714
	HiVaR-i	0.531	0.337	1.575668	1.368	3.413
	HiVaR-v	0.489	0.331	1.477341	1.369	3.674
	HCAA-v	0.942	0.597	1.577889	1.268	1.277
	HCAA-iv	0.890	0.520	1.711538	1.324	1.537
	HCAA-ii	0.618	0.322	1.919255	1.342	3.198
	MDP	0.674	0.345	1.953623	1.439	4.229
	MDIP	0.882	0.435	2.027586	1.374	6.942
	MCP	0.702	0.357	1.966387	1.431	4.054
	EWP	0.565	0.318	1.776730	1.382	3.907
	ERC	0.553	0.302	1.831126	1.397	4.024
	PiVaR	0.677	0.353	1.917847	1.371	4.983
	MVMIP	0.828	0.407	2.034398	1.351	5.282
2018	MPT	0.111	0.972	0.114198	1.254	2.289
	MVP	-0.042	0.663	-0.063348	1.362	2.460
	MIP	0.021	0.795	0.026415	1.342	4.052
	IVP	-0.071	0.735	-0.096599	1.389	3.268
	IIP	-0.013	0.792	-0.016414	1.360	3.326
	HRP	-0.053	0.733	-0.072306	1.400	3.340
	HiVaR-i	0.034	0.783	0.043423	1.380	3.521
	HiVaR-v	-0.043	0.767	-0.056063	1.378	3.435
	HCAA-v	-0.152	0.882	-0.172336	1.272	1.426
	HCAA-iv	-0.153	0.811	-0.188656	1.357	1.801
	HCAA-ii	-0.047	0.736	-0.063859	1.369	3.147
	MDP	0.052	0.738	0.070461	1.444	3.917
	MDIP	0.105	0.826	0.127119	1.386	5.512
	MCP	0.046	0.761	0.060447	1.440	3.846
	EWP	-0.054	0.782	-0.069054	1.382	3.467
	ERC	-0.050	0.750	-0.066667	1.396	3.505
	PiVaR	0.042	0.796	0.052764	1.345	4.091
	MVMIP	0.119	0.866	0.137413	1.340	4.155
2019	MPT	0.332	0.653	0.508423	1.260	2.304
	MVP	0.475	0.511	0.929550	1.377	2.264
	MIP	0.648	0.557	1.163375	1.388	3.991
	IVP	0.559	0.531	1.052731	1.390	3.172
	IIP	0.590	0.554	1.064982	1.369	3.248
	HRP	0.544	0.516	1.054264	1.402	3.156
	HiVaR-i	0.552	0.574	0.961672	1.372	3.025
	HiVaR-v	0.649	0.586	1.107509	1.371	3.088
	HCAA-v	0.434	0.566	0.766784	1.313	2.158
	HCAA-iv	0.433	0.544	0.795956	1.347	2.299
	HCAA-ii	0.629	0.589	1.067912	1.324	2.808
	MDP	0.618	0.505	1.223762	1.452	3.626
	MDIP	0.639	0.570	1.121053	1.403	4.733
	MCP	0.649	0.525	1.236190	1.449	3.732
	EWP	0.604	0.572	1.055944	1.382	3.211
	ERC	0.579	0.538	1.076208	1.397	3.307
	PiVaR	0.652	0.561	1.162210	1.387	4.015
	MVMIP	0.675	0.592	1.140203	1.384	4.045

Table 11: Monthly average risk measures on the ETF universe

	10-d VaR (%)	10-d CVaR (%)	MDD	AvDD	Calmar	Burke
MIP	-0.585566	-0.811834	0.027725	0.002418	8.109703	1.553425
MDIP	-1.180096	-1.625289	0.084793	0.012427	7.402861	1.264630
MVMIP	-0.753354	-1.164694	0.044631	0.006234	11.577884	1.876739
HiVaR	-2.765513	-3.601728	0.225186	0.037574	4.646394	0.959645

Table 12: iVaR model OOS performance over ETF universe: subsamples

		Returns (%)	Volatility (%)	Sharpe	D_{var}	D_{iVaR}
2006	MIP	0.176406	0.246141	0.800961	1.013734	7.087910
	MDIP	0.221268	0.638560	0.490237	1.031074	4.652842
	MVMIP	0.174673	0.276938	0.744123	1.022679	8.951333
	HiVaR	-0.157494	1.222364	0.074257	1.120880	2.640861
2007-2009	MIP	0.333402	0.344077	1.017839	1.006914	4.608192
	MDIP	0.282562	0.721174	0.508734	1.030106	4.794792
	MVMIP	0.326265	0.344235	1.021644	1.012947	5.089815
	HiVaR	0.092080	1.068940	0.414732	1.032440	3.760455
2010-2012	MIP	0.202989	0.443208	0.475500	1.003429	4.690641
	MDIP	0.424070	0.660795	0.698146	1.042997	5.878120
	MVMIP	0.287506	0.472158	0.604259	1.019440	6.183981
	HiVaR	0.786425	1.182576	0.715720	1.044778	3.699683
2013-2015	MIP	0.144933	0.209665	0.848133	1.003616	3.836165
	MDIP	0.325959	0.706086	0.661056	1.041923	6.517812
	MVMIP	0.218098	0.398332	0.959033	1.025575	4.841524
	HiVaR	0.199003	1.702447	0.129946	1.060772	2.413113
2015-2017	MIP	0.010662	0.132904	0.144158	1.001814	1.214443
	MDIP	0.223900	0.638502	0.457232	1.031228	6.245543
	MVMIP	0.081785	0.596254	0.288379	1.010434	2.216493
	HiVaR	0.562739	1.698658	0.392529	1.048287	2.031052

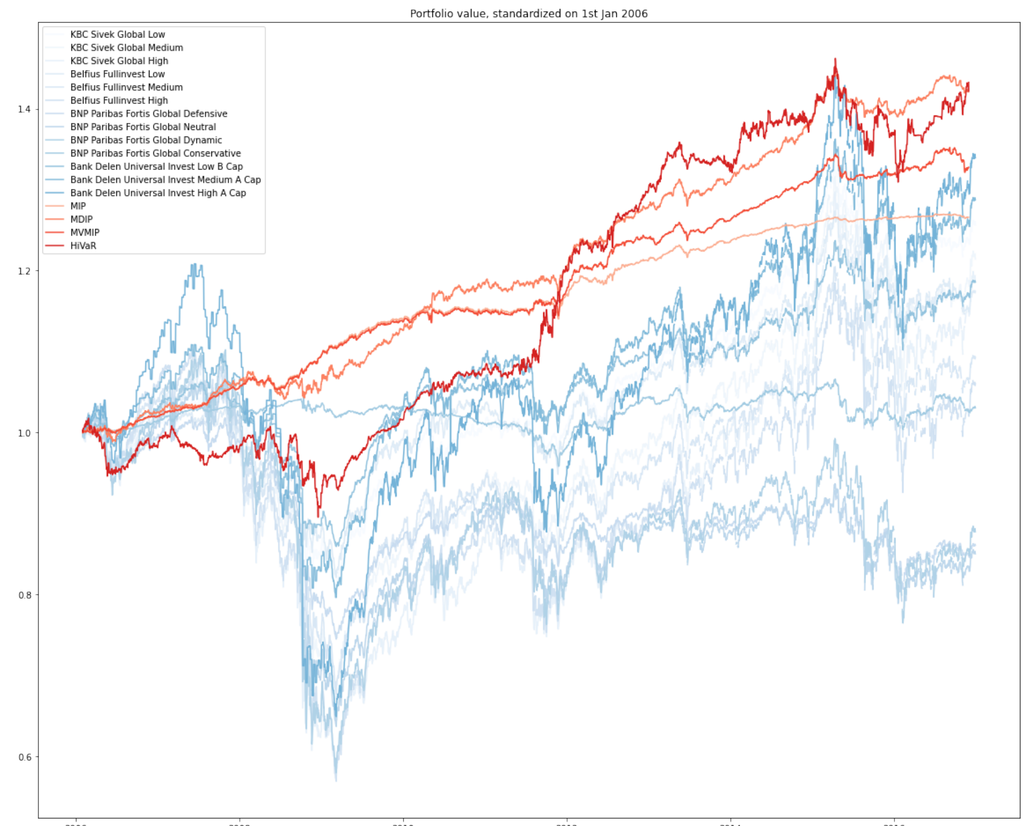


Figure 19: iVaR models (red) versus large pool of benchmark funds (blue)

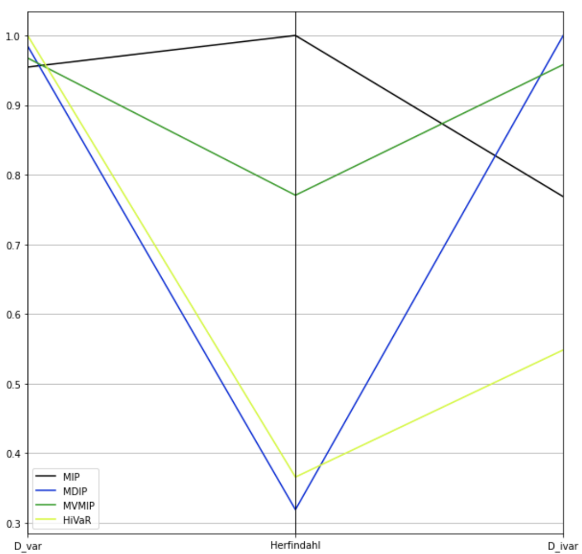
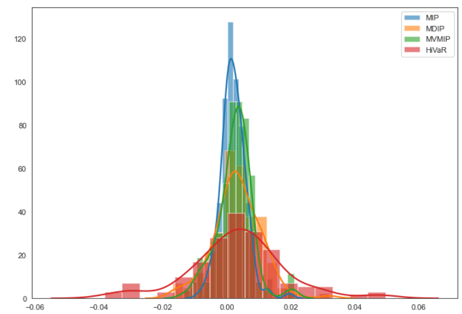
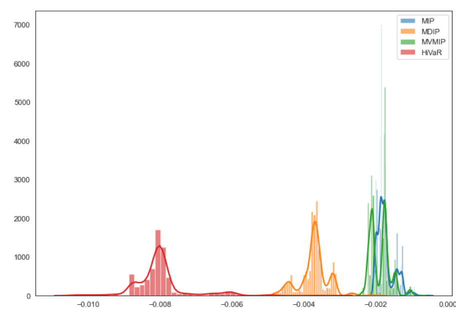


Figure 20

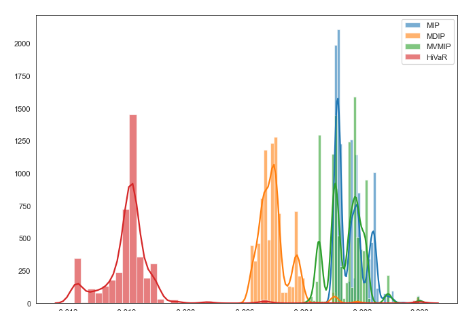
Figure 21: V-plot for ETF universe: MIP has highest relative concentration, MDIP offers optimal diversification.



(a) Simulated return of portfolios



(b) Simulated VaR of portfolios



(c) Simulated CVaR of portfolios

Figure 22: Return and tail distributions

VII. Appendix

A. Appendix 1: LPM and VaR

We avoided lengthy formulae in the introduction on risk measures to give the reader an idea about the history of standard risk measures without getting lost in the details. We defined LPM, VaR and CVaR more intuitively as:

$$(\alpha, R) - LPM = \frac{1}{T}\sum_{t=1}^{T} \max\left(R - r_{p,t}, 0\right)^{\alpha} \tag{43}$$

$$p = 1 - cl = Pr(r < VaR_{1-cl}(F)) = F(VaR_{1-cl}(F)) \tag{44}$$

$$CVaR_{1-cl}(F) = E(r|r < VaR_{1-cl}(F)) \tag{45}$$

with returns r, return distribution or PnL $F(r)$ and confidence level cl. More formally, we should define LPM as an integral over F as well:

$$LPM_{\alpha}(R, F) = \int_{-\infty}^{R} (R - r)^{\alpha} dF(r) \tag{46}$$

and CVaR as:

$$CVaR_{1-cl}(F) = \frac{1}{1-cl}\int_{0}^{1-cl} VaR_{\gamma}(F)d\gamma \tag{47}$$

Now we find that if $\alpha = 0$,

$$LPM_0(R, F) = \int_{-\infty}^{R} (R - r)^0 dF(r) = \int_{-\infty}^{R} dF(r) = F(R) \tag{48}$$

and if we set R = VaR then we can see:

$$F(R) = 1 - cl = VaR_{1-cl}^{-1}(F) \tag{49}$$

Moreover, if $\alpha = 1$, and we again set R = VaR,

$$LPM_1(R, F) = \int_{-\infty}^{R} (R - r) f(r)dr \tag{50}$$

$$CVaR_{1-cl} = \frac{1}{1-cl}\int_{-\infty}^{VaR_{1-cl}(F)} r f(r)dr \tag{51}$$

we more clearly see the resemblance between CVaR and LPM of order 1. LPM will measure the average deviation from the VaR return, while CVaR will measure the average return below VaR. They are clearly not identical — the denominator $1 - cl$ and deviation versus actual losses — but they measure the same general idea. Now the generality of LPM measures is that α can be any rational number, and this will reflect the investor's risk aversion (see Fishburn 1977, [7]). The link with drawdown measures is that they can be seen as dynamic generalizations of this idea, where VaR and CVaR are mapped on MDD and CDD measures, as to replace VaR functionals on a static F, by a dynamic AD functional (see Appendix 3 below).

B. Appendix 2: Coherent risk measures

As we discussed in the main text, coherence is a key concept related to risk measures that gained wide academic acceptance as well as adoption by practioners. For instance, the change from VaR to ES in new regulation was mainly motivated by VaR's lack of subadditivity. Artzner and Delbaen [10] define 5 desirable properties that a quantitative risk measure $\rho(V)$ should have — convexity was later added as a 6th and replaces subadditivity and positive homogeneity:

- Normalisation: if $V = 0$ then $\rho(0) = 0$, or the risk of an empty portfolio is zero.
- Monotonicity: if $V1 \leq V2$ in almost all cases then $\rho(V1) \geq \rho(V2)$, or when portfolio values V of one portfolio dominate another portfolio almost surely then the risk should be lower.
- Positive homogeneity: $\rho(\lambda V) = \lambda\rho(V), \forall\lambda \geq 0$, or the risk of a portfolio should be proportional with its size.
- Translational invariance: $\rho(V + const) = \rho(V)$, or adding certain values (such as cash) to V should decrease risk with the same amount.
- Subadditivity: $\rho(V1 + V2) \leq \rho(V1) + \rho(V2)$, or the risk of portfolios held together can never be higher than the risk of individual portfolios.
- Convexity: $\forall\lambda \in [0, 1], \rho(\lambda V1 + (1 - \lambda)V2) \leq \lambda\rho(V1) + (1 - \lambda)\rho(V2)$, or the risk of a linear combination of V1 and V2 should never be more than the linear combination of risks. It can be seen from the nature of linear combinations that the notion of convexity effectively replaces the notion of both subadditivity (sum) and positive homogeneity (scalar product).

We can conclude from their interpretation that normalisation, monotonicity and translational invariance are very natural requirements for risk measures, and most of the classical risk measures (variance, VaR, etc.) have these properties. Convexity, on the other hand, is not obvious and this has crucial implications for diversification. Convexity tells us that combining assets into a portfolio will never be more risky than holding them seperately. This was exactly the problem we pinpointed about VaR, as VaR lacks subadditivity. Variance, CVaR and drawdowns are convex measures of risk (for which we will provide arguments below) and can therefore be used in our considerations (e.g. for the calculation of diversification ratios), while incoherent risk measures cannot.

C. Appendix 3: Drawdown measures are subadditive

Here we will introduce conditional drawdowns (CDD) as a dynamic generalization of CVaR and use a CVaR-functional's properties to argue that drawdown measures are subadditive. The aim of this appendix is not to provide complete, exhaustive proofs on the properties of drawdown measures, we just want to provide the intuition behind them as to provide an argument for drawdown's subadditivity (or a fortiori convexity). This in order to guarantee that optimizing for diversification, calculating diversification ratios and so forth would make sense. Indeed, if the sum of the risks could be smaller than the risk of the sum, all the concepts above would be completely useless.

Recall the AD-functional:

$$AD(V) = (\xi_1, ..., \xi_T), \xi_t = \max_{s<t}(V_s) - V_t \tag{52}$$

Let's now define the $\alpha - CDD$ as the average of the $\alpha\%$ worst drawdowns. To that extent, we need a threshold $\zeta(\alpha)$ such that $\alpha\%$ of elements in AD(V) exceed $\zeta(\alpha)$. This can be seen as a VaR-functional on AD(V) instead of $F(r)$. We now simply need the mean of all ξ in a set where ξ is larger than $\zeta(\alpha)$. This is the CVaR functional. However, a series of drawdowns has a discrete number of elements T, such that it is very unlikely that we can find the exact $\alpha\%$ largest drawdowns. Let's use $\zeta(\alpha)$ and define a CVaR-functional from [14]:

$$CVaR(\xi) = \frac{1}{\alpha N} \sum_{\xi_t \in \Theta} \xi_t + \pi(\alpha, \xi) \tag{53}$$

where $\Theta = \{\xi_t | \xi_t > \zeta(\alpha)\}$, and $\pi(\alpha, \xi)$ an additional term that depends on how much we have to extrapolate the $\alpha\%$ largest drawdown from the discrete series (for more explanation see [18] and [14]). Now we can just use this logic and say the $\alpha - CDD$ is the CVaR-functional applied to the drawdown functional:

$$\alpha - CDD = CVaR(AD(V)) \tag{54}$$

In other words, $\alpha - CDD$ is the $CVaR$ of a loss function $AD(V) = \xi$ which is dynamically obtained, while CVaR normally uses a static PnL function F(r). CDD is thus an example of a functional generalizing properties of deviation measures to a dynamic case. Moreover, expression (54) makes our lives considerably easier in the sense that we just need the properties of AD and CVaR to discuss those of CDD (and therefore MDD and ADD).

For AD(V) we can tell from definition (52) that AD is nonnegative, insensitive to constant shifts, positive homogeneous and convex. The first three properties follow directly from (52). Convexity means $\forall \lambda \in [0, 1], AD(\lambda V1 + (1 - \lambda)V2) \leq \lambda AD(V1) + (1 - \lambda)AD(V2)$. Again looking at (52) we just have to show that $\max_{t<T}(\lambda V1 + (1 - \lambda)V2) \leq \lambda \max_{t<T}(V1) + (1 - \lambda) \max_{t<T}(V2)$, which is clearly true based on the properties of the max functional.

We argued $CVaR(\xi)$ is coherent, having the following properties: normalisation, monotonicity, positive homogeneity, translational invariance, subadditivity and convexity. For complete proofs see [14]. The point is not to prove this in depth here, the point is to show that its convexity, together with the convexity of AD and equation (54), makes the proof that CDD is convex rather trivial: $CDD(\lambda V1 + (1 - \lambda)V2) = CVaR(AD(\lambda V1 + (1 - \lambda)V2) \leq CVaR(\lambda AD(V1) + (1 - \lambda)AD(V2)) \leq \lambda CVaR(AD(V1)) + (1 - \lambda)CVaR(AD(V2)) = \lambda CDD(V1) + (1 - \lambda)CDD(V2)$.

Now we can easily see that (1) MDD is just the special case of CDD with $\alpha = 0$ — or we let the average of the 0% worst drawdowns correspond to the single worst drawdown — and (2) ADD is the special case where $\alpha = 1$ — or the average of *all* worst drawdowns corresponds with the simple average. We now know drawdown measures are convex, therefore subadditive, such that we can calculate meaningful diversification ratios, optimize for them and so forth.

D. Appendix 4: Some stylized facts about financial correlation matrices

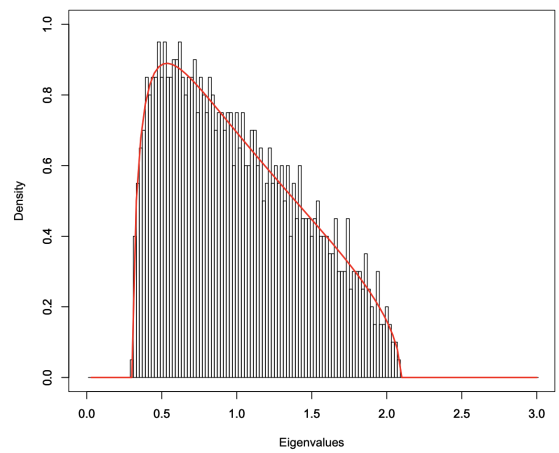
In section II we discussed Markowitz' Curse and its implications for diversification. We mentioned matrix conditioning and eigenvalues but avoided lengthy discussions or formulae. Even without formulae, the key characteristics of financial correlation matrices (compared to random matrix theory) can be brought back to some stylized facts [39]:

- The distribution of financial correlations is significantly shifted to the positive, i.e. most assets are positively correlated.
- Perron-Frobius property: the first eigenvector has positive entries, i.e. all assets typically have positive exposure to the market.
- Eigenvalues follow the Marcenko-Pastur distribution, with the exception of the first very large eigenvalue (the market) and a couple of other large eigenvalues (the industries).

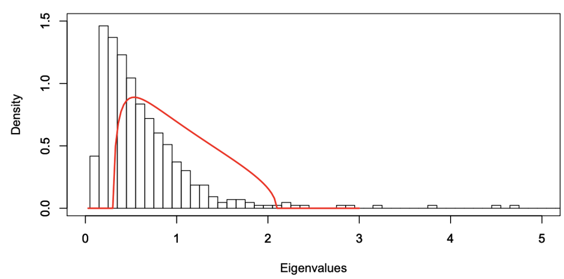
The first two characteristics imply most assets have positive pairwise correlation and positive *betas* with the market. From a diversification perspective, this makes 'smart' diversification (15) especially important and difficult. The third characteristic is closely related to the conditioning of the matrix. According to random matrix theory, the Marcenko-Pastur limit theorem describes the asymptotic distribution of eigenvalues for the correlation matrix of random independent and identically distributed (IID) data. The theorem and its eponymous density is illustrated in Figure 23 (a). For multivariate financial time series, we typically find that the eigenvalues of their correlation matrix deviate from the MP density as the first few eigenvalues (market and industries) are much higher. This is clearly visible in 23 (b). Before, we defined the condition number as the ratio of maximum and minimum eigenvalues. In our setting, we said the condition number can be seen as how much the output values of our quadratic program will change for a small change in the input covariance matrix. The higher this number, the more instable variance-based methods will be. This third fact thus explains why this is particularly problematic on financial data.

On the other hand, plot 23 also illustrates the power of shrinkage as a correlation filtering method. The distribution of the eigenvalues exhibits a property called approximate sparsity[54] or a rapid decay in value. This essentially means that there exists a small group of eigenvectors with large eigenvalues that capture most of the variance in the data. As a filter, we can leverage this property and shrink covariances and eigenvalues using the mentioned techniques as to improve the conditioning of the problem. This is illustrated in Figure 23 (c). Lastly, MP is the distribution for random data, meaning it also provides us with a upper bound('λ_+') for random eigenvalues [5]. This implies that values higher than this bound can be considered non-random or signal, and the ones below the bound as noise [35].

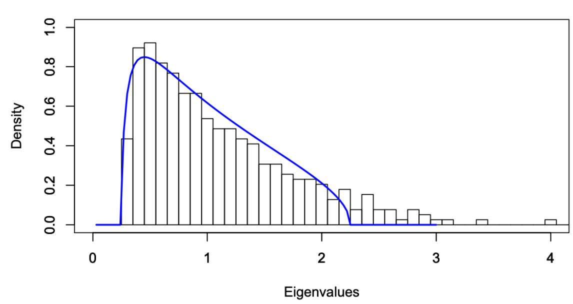
[54] The sorted absolute values of the eigenvalues decay fast enough in values, i.e. the j^{th} largest absolute value $|\lambda|_j \leq \frac{A}{j^a}, a \geq 1/2, \forall j$



(a) Distribution of eigenvalues for correlation matrix of random IID data with MP density.



(b) Distribution of eigenvalues for sample financial correlation matrix with MP density.



(c) Correlation filtering

Figure 23: Marcenko-Pastur density of random versus financial correlation matrix [41]

E. Appendix 5: Tibshirani's Gap index

For the construction of HCAA clusters, we first determine the optimal number of clusters using the Gap index [12][55]. We briefly mentioned that this index maximizes the gap between the learned structure and the structure obtained on uniformly distributed data. This was the intuition, because in practice we look for a minimum amount of clusters such that the reduction of Gap index compared to adding one more cluster is greater than a standard deviation of all Gap indexes. This is called the Gap inertia of adding new clusters. So in terms of implementation we need random uniform distribution generation, linkage on our data and random uniform, and select the least number of clusters that satisfies this inertia idea.

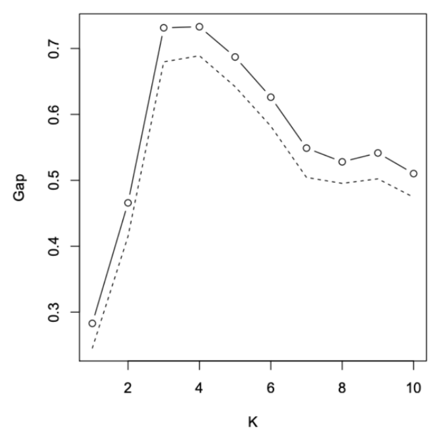


Figure 24: Gap index [36]

The exact calculation of the index happens as follows [12]:

- Cluster the observed data, varying the number of clusters from $k = 1, \ldots, k_{max}$, and compute the corresponding total within intra-cluster variation W_k.
- Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters $k = 1, \ldots, kmax$, and compute the corresponding total within intra-cluster variation W_{kb}.
- Compute the estimated Gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k) = \frac{1}{B}\sum_{b=1}^{B} log(W_{kb}) - log(W_k)$. Compute also the standard deviation of the statistics.
- Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k) \geq Gap(k+1) - s_{k+1}$.

[55] For HRP and simple HiVaR we did not need this as we bisected down to the leaf nodes.

F. Appendix 6: Alternatives to the sample covariance matrix

"*Anything merely based on correlation is charlatanism.*" (Nassim N. Taleb)

In previous sections, we already delved into the importance of the covariance matrix, and the hazards of covariance misspecification. We avoided going too deep into the technicals, not to lose track of the message there (the link between portfolio construction frameworks and diversification). Now that we hopefully succesfully covered that, and before we can specify what we did in the code, let's delve into the options we have as alternatives to the sample covariance matrix.

We know covariance and correlation modeling is a particularly tricky business. Financial mathematician Paul Wilmott calls correlation modeling 'accidents waiting to happen', while Nassim Nicholas Taleb refers to it as 'charlatanism'. Gunter Meissner, author of the book 'Correlation Risk Modeling and Management', ironically calls it 'work of the devil', after the role of copula models in the 2008 CDO crisis. It is clear why they think so: we reduced all the information in the edges of the full graph to a single number, causing all kinds of issues with our allocation (cf. supra). However, we will need a correlation matrix for the comparison of iVaR portfolios with classical frameworks and as input for hierarchical clustering if we use correlation as similarity. So what is the best we can do?

Ledoit and Wolf wrote a famous paper *Honey, I shrunk the sample covariance matrix*, which introduced shrinkage for the estimation of covariances for asset allocation (in an MPT framework). It followed their earlier work, where they defined a general shrinkage approach as estimating a convex transformation of the sample covariance matrix $\hat{\Sigma}$:

$$\Sigma_{shrunk} = (1-\alpha)\hat{\Sigma} + \alpha \frac{Tr\hat{\Sigma}}{p}\mathbb{1} \tag{55}$$

Remember from before the issue of conditioning[56] was related to the ratio of the largest and smallest eigenvalue of the empirical or sample covariance matrix. When this ratio is high, we called $\hat{\Sigma}$ ill-conditioned and inverting $\hat{\Sigma}$ into the matrix inverse — also called the precision matrix K — causes high errors such that our eventual solution is extremely instable. We can reduce this ratio by simply shifting every eigenvalue according to a given offset[57]. This is equivalent to finding the L2-penalized Maximum Likelihood estimator of the covariance matrix [42]. This just boils down to the expression above, where we choose the amount of shrinkage by setting α. This depends on a *bias-variance trade-off*. A high penalty will reduce variance in the estimates drastically by reducing the average covariance, but will introduce a lot of bias as well. A small penalty has the opposite effect. The overall impact of these two effects should be optimal. Ledoit-Wolf (2003) defines optimal as the minimum Mean-Squared Error (MSE) between the estimated and real covariance matrix. Moreover, in their aforementioned paper on asset allocation they define the shrinkage estimator generally as:

$$\Sigma_{shrunk} = (1-\alpha)\hat{\Sigma} + \alpha F \tag{56}$$

where F is any 'highly structured' covariance estimator, and the optimal α is determined through minimizing MSE. Ledoit and Wolf propose the constant correlation

[56] Also see Appendix 4

[57] This is the intuition behind the MP plot in 23 (c)

model for F. This is the covariance matrix where all correlations are identical.

As an alternative, we could use L1 shrinkage. Remember from a previous discussion that this allows us to set covariances to exactly zero (or cut edges in the graph), while L2 penalties cannot. That is why L1 shrinkage on Σ is called *covariance selection*. As we know from the Curse of Dimensionality, especially in the situation in which N(N-1)/2 is relatively large compared to the TxN observations, sparse L1 covariance estimators tend to work better than shrunk L2 covariance estimators. In the opposite situation, or for very correlated data, they can be numerically instable [42][21]. The L1 shrinkage for covariance is implemented in Scikit Learn as a Graphical LASSO (CV) [21]:

$$\hat{K} = \min_{k} (Tr\hat{\Sigma}K - logdetK + \alpha||K||_1) \tag{57}$$

where K is the precision matrix to be estimated, $\hat{\Sigma}$ is the sample covariance matrix, $\alpha||K||_1$ is the sum of the absolute values of off-diagonal coefficients of K [42]. Again the alpha parameter tunes the sparsity of the precision matrix. In Scikit Learn, we use GraphicalLassoCV to automatically set the hyperparameter using 10-fold cross-validation [42].

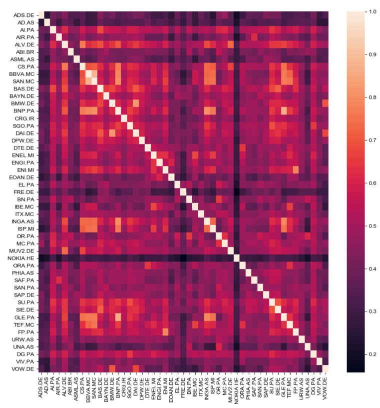
In conclusion, our code uses and compares three approaches to covariance estimation: empirical or sample covariance, L2-like Ledoit-Wolf shrinkage and L1-like Graphical LASSO shrinkage. These alternatives are visualised for the Eurostoxx 50 in Figure 25.

G. Appendix 7: Alternatives to single linkage hierarchical clustering

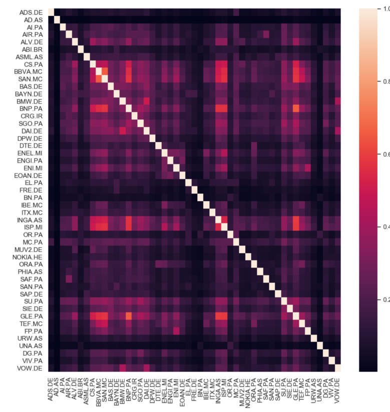
"*Assets are characterized in a portfolio by the company they lack.*"

Linguist John R. Firth popularized the idea that words in a document are characterized by the company they keep. Words only really have a meaning if they have a context. This idea is what underlies recent developments in so-called word embeddings, where we try to represent words as vectors where words that are often found close to each other (for semantic or syntactic reasons) are close to each other in a vector space. Word embeddings are essentially a clustering and dimension reduction problem. This subchapter zooms in on similar techniques for assets, which we can use for the implementation of our clusterd-based iVaR portfolios. In contrast to words, assets in a (diversified) portfolio are characterized by the company they *lack*. We already iterated this point many times, but the value of adding an asset to a portfolio is inversely proportional with how often we find this asset in the company of the assets already in our portfolio. We could find them in the same supply chain, the same industry, the same country, the same liquidity bucket and so forth. However, identifying which assets are similar is much more difficult than words. The input characteristics to word vectors are simply word counts. The input characteristics to firms is all the information in the edges of our graph.

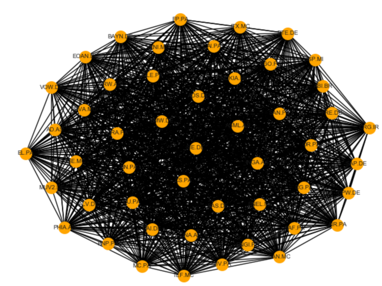
Recall the goal of clustering analysis: we observe data from N firms in the graph, and for each firm $i = 1, \dots, N$, we observe a tensor ('vector of vectors') of m characteristics $x_i = (x_{i,1}, \dots, x_{i,m})'$, where every $x_{i,j}$ is a time series of T observations of characteristic j over time for firm i. The goal of clustering analysis is to group the firms into a relatively small number of clusters such that:



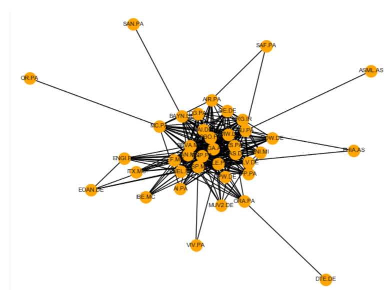
(a) Emp corr matrix



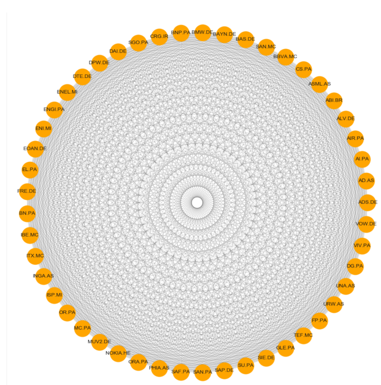
(b) L1 corr matrix



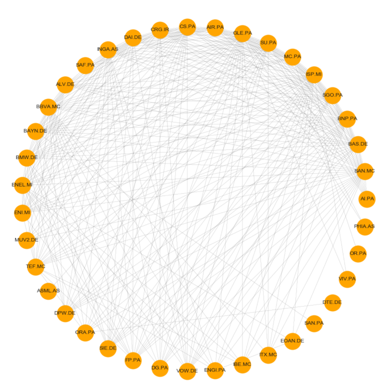
(c) Emp corr cluster



(d) L1 corr cluster



(e) Emp radial cluster



(f) L1 radial cluster

Figure 25: Alternatives to the sample covariance matrix

– every firm i = 1,..., N is a member of one and only one cluster

– two firms i and j in the same cluster have similar observed characteristics x_i and x_j.

What these characteristics are, how similarity is measured and assessed over the course of the clustering task, needs further specification. Recall that so far in HRP, we assumed (1) our characteristics tensor comprises the returns of our firms over time, (2) similarity is correlation ρ, and (3) dissimilarity a transformation of correlation (i.e. proportional to $1-\rho$). Let's now first generalize — go broad - and then specify — go deep.

There are many approaches to clustering: connectivity-based clustering (or hierarchical clustering), centroid-based clustering (or k-means clustering) and distribution-based clustering (e.g. Gaussian mixture models). Table 13 provides an overview of these types of clustering. The two most commonly used approaches to clustering are:

– K-means clustering: we minimize a least-squares objective function of the distance between firms and a mean vector (centroids) of characteristics. The centroid or mean vector can be seen as a prototypical asset in that cluster.

– Hierarchical Clustering: we use a hierarchical procedure of sequentially forming clusters (cf. h.).

a. Extending HRP

In h., we took the agglomerative hierarchical clustering approach. This was a bottom-up approach, where we started with all firms as their own group and then merged them sequentially. The measure of dissimilarity we used was:

$$\tilde{d}_{i,j} = ||d_i - d_j|| \tag{58}$$

or the Euclidean distance between the transformed[58] correlation of asset i and j. Given the dissimilarity measure, we needed a linkage criterion. The linkage is the dissimilarity we minimize between groups each time we merge two groups[59]. Recall HRP uses single linkage, but we realized the Nearest Point Algorithm suffers from chaining. Alternatively, complete linkage or the Farthest Point Algorithm could be considered. On its turn, complete linkage suffers from crowding. Because its score is based on the worst-case dissimilarity between pairs, a point can be closer to points in other clusters than to points in its own cluster. Therefore, clusters can be too close to each other. The third 'canonical' choice would be average clustering. It gives us very robust results [34], but average distance is more difficult to interpret. In the code, we tested these methods — as well as weighted, median and Ward's linkage — as small extensions to the HRP model. In summary, the alternatives are:

– Single linkage or the Nearest Point Algorithm:

$$\tilde{d}(u, v) = min(\tilde{d}(u_i, v_j)) \tag{59}$$

[58] $d = \sqrt{2(1-\rho)}$

[59] Just to be sure: we always minimize this criterion describing the two groups, but the definition of this distance can either be the minimum distance between two points from the groups, maximum distance, average distance, and so and so forth.

– Complete linkage or the Farthest Point Algorithm:

$$\tilde{d}(u,v) = max(\tilde{d}(u_i, v_j)) \tag{60}$$

– Average linkage:

$$\tilde{d}(u,v) = \sum_{i,j} \frac{\tilde{d}(u_i, v_j)}{|u| * |v|} \tag{61}$$

– Weighted linkage:

$$\tilde{d}(u,v) = (\tilde{d}(s,v) + \tilde{d}(t,v))/2 \tag{62}$$

where cluster u was formed with cluster s and t, and v is a remaining cluster.

– Centroid linkage:

$$\tilde{d}(u,v) = ||c_u - c_v||_2 \tag{63}$$

where c_u and c_v are the centroids of clusters u and v, respectively. Centroids are the average observations or mean vectors. When two clusters u and v are combined into a new cluster w, the new centroid is computed over all the original objects in clusters and u and v. The distance then becomes the Euclidean distance between the centroid of w and the centroid of a remaining cluster [42].

– Median linkage assigns $\tilde{d}(u,v)$ like the centroid method. When two clusters u and v are combined into a new cluster w, the average of centroids u and v give the new centroid w.

– Ward's method minimizes variance as the new entry is computed as follows:

$$\tilde{d}(u,v) = \sqrt{\frac{|u|+|s|}{T} d(v,s)^2 + \frac{|u|+|s|}{T} d(v,t)^2 - \frac{|v|}{T} d(s,t)^2} \tag{64}$$

where u is the newly joined cluster consisting of clusters s and t, v is an unused cluster in the forest, $T = |v| + |s| + |t|$, and $|*|$ is the cardinality of its argument. This is also known as the incremental algorithm [42].

More important than their plain definition is their direct link with diversification or weight concentration. Figure 26 from Papenbrock's thesis 'Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization' [27] summarizes weight concentration as a spectrum from naive maximum deconcentration portfolios to cluster-based 'waterfall' portfolios using single linkage.

Apart from the issue of chaining inherent to single linkage, we identified two other issues with the original HRP algorithm in section h.: in the recursive bisection step the original HRP algorithm (1) only uses the order of assets after clustering, not the true shape of the dendrogram, and (2) bisects top-down, all the way until every cluster is an individual asset, which makes it prone to overfitting [34]. In brief, we do not use the true shape of the dendrogram, nor an optimal number of clusters implied by the dendrogram. These two issues were resolved by Raffinot [34][36] with the introduction of a more general HCAA (Hierarchical Clustering-based Asset Allocation) algorithm. This model essentially makes two tweaks to the original HRP: (1) perform a top-down bisection based on the number of assets in the clusters, not just splitting the covariance

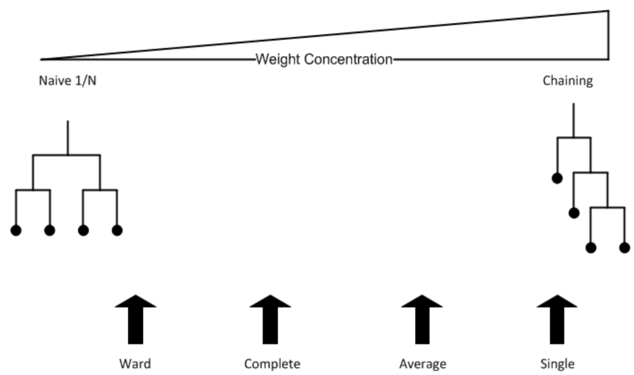


Figure 26: Linkage criterion and concentration [27]

matrix in half each time — which is the natural interpretation of the 'waterfall' we discussed in h. — and (2) calculate an optimal number of clusters using the Gap index. Essentially, HRP uses the hierarchial tree's full complexity. Or as Raffinot argues, bisection all the way to the leaf nodes is overfitting, in the sense that we do not only use the *structure* implied by the data, but also attach importance to individual units.

The determination of an optimal number of clusters C is a common issue in clustering analysis. Common approaches are the Elbow method (tradeoff C with total within-group sum-of-squares), silhouette methods (maximizing silhouette values) and Gap index. The Gap index is generally applicable and a natural choice here. The intuition behind Tibshirani's Gap index [12] is essentially that we should maximize the gap between the learned structure and a structure obtained from uniformly distributed data. This idea is explained in more detail in Appendix 5.

b. Looking beyond

We now covered and remedied the most important issues about original HRP. But we did not address the most pertinent question: what exactly are we clustering? Why do we use correlations and coiVaRs? Why transform historical returns to linear comovements or codrawdowns in general? Why not try to detect patterns in the original high-dimensional return or any 'asset characteristics space'? Essentially, we are coping with a dimensionality problem. Hierarchical clustering does not fare well on very high-dimensional spaces (see Table 13). That is one of the main reasons why we apply clustering on the correlation matrix. But why reduce the similarity of return vectors to a single number that measures linear comovement?

Essentially, we want to find a structure in the high-dimensional graph data, that gives hints about the sources in the graph. Before we can deploy that structure in a hierarchical asset allocation, we need to embed that high-dimensional graph in a low-dimensional space. The 'easy' way to do this is summarize the TxN-dimensional return space into a NxN covariance matrix with N + N(N-1)/2 unique dimensions. There are good reasons to do this (as we explained), as well as major drawbacks (as

Type	Examples	Comments
Connectivity-based	Hierarchical clustering: single linkage/MST, complete, average, WPGMA, Ward's, DBHT	+ Practical and interpretable - Distance measure (e.g. Euclidean) preferably in low-dimensional space
Centroid-based	K-means, K-SVD	+ Simple, practical and interpretable - No notion of hierarchy, a single set of clusters with no organization or structure within and the choice of the number of clusters always happens ex-ante
Distribution-based	Gaussian Mixtures, Multimodel Mixtures	+ Statistical properties, interpretability parameters - Parametrical approaches: choice of distribution and parameters and often highly complex in nature
Density-based	DBSCAN, HDBSCAN, OPTICS	+ Does not require a-priori specification of number of clusters (like hierarchical clustering), works well on noisy data - Does not work well on high-dimensional data
Embeddings	(kernel-)PCA, Isomap, (M)LLE, LTSA, LDS t-SNE, UMAP, Neural Networks	+ Powerful techniques, especially suited for high-dimensional data. Possible to leapfrog correlation as essentially a means of dimension reduction - Information loss: be careful on how to select the information/distances to preserve (example in text).

Table 13: Overview clustering techniques

we also extensively explained), but there are also alternatives. In essence, we want to embed the TxN space on a lower-dimensional e.g. SxN space ($S << T$), that preserves some of the 'essence', i.e. valuable information from the original space. We inevitably lose information and we have to make an important decision as to what information we want to keep.

A first natural choice is a simple Principle Component Analysis (PCA). PCA extracts S linear combinations from the original space that maximize the variance of these combinations[60], while being mutually orthogonal or independent. Therefore, PCA extracts common factors that explain shared variance of stocks (such as common *market factors* influencing many stocks at the same time). Clustering on PCA output essentially uncovers which assets are sensitive to similar market factors. Therefore, identifying assets' exposure to the first couple of principle components is an often used technique to assess diversification. However, PCA is a very simple linear method. Actually, these common factors can be seen as a simple linear econometric panel model with interactive fixed effects[61]. There are 2 major limitation in how PCA sees the 'essence' (i.e. variance) of returns: (1) the linearity assumption provides insufficient structure to capture the non-linear interactions between the data — although we can use kernel-PCA, that applies linear PCA after non-linear 'kernel trick' — and (2) maximizing variance essentially means maximizing the distance between points in the low-dimensional subspace for points that were far away in the original space. The latter limitation is of utmost importance here. Within clusters neighbours should have similar return characteristics, we do not merely want to to a good job at explaining differences in returns.

Since dimension reduction essentially means embedding high-dimensional spaces in lower-dimensional ones, we have many options that we can borrow from manifold and even deep learning to tackle this problem: Isomap, LLE, MLLE, LTSA, LDS, t-SNE, UMAP and neural networks. As we argued, the choice vastly depends on the topological structures we want to preserve, and the ones we are willing to give up. In that regard, an interesting approach which we will briefly illustrate here is called t-distributed stochastic neighbour embedding (t-SNE).

t-SNE uses the probability that two assets are close to each other in terms of returns in the original space and the probability of finding those assets close to each other in the lower-dimensional space. Put very briefly, t-SNE minimizes the divergence between these two probabilities P and Q using the most frequently used divergence measure for distributions in information theory, the Kullback-Leibler divergence:

$$D_{KL}(P|Q) = \sum_x P(x) log(\frac{P(x)}{Q(x)}) \tag{65}$$

In more intuitive terms, t-SNE preserves neighbours in the low-dimensional space, much in contrast to PCA. Stocks with returns that were persistently close to each other on particular days will be put together. However, further distances in the new space might not be indicative of far distances in the original space. In terms of our clusters, this means 'leaf' cluster members or neighbours are likely to be preserved while cross-cluster distances may be more difficult to interpret. Some of the general topological features can be preserved by initializing t-SNE with PCA values.

[60] This is closely related to the eigenvalues and -vectors discussed in Appendix 4.
[61] Such that the conceptual edge over linear correlation is low to non-existent.

Figure 27 shows a two-dimensional embedding of all S&P500 returns over 5 years (Jan 2015-2020) using t-SNE with PCA initialization. Figure 28 shows the same plot with the CIGS sectors as colors. We can tell many neighbours belong to the same sector, while this is clearly not the only factor to explain return comovement, or a more adequate term would be 'co-location' in this use case.

Next, Figure 30 shows a hierarchical clustering on the embedded space. This allows us to determine an alternative order in our assets that regroups assets as learned by our embedding, rather than by correlation. This order can then be used in subsequent steps such as recursive bisection. However, clustering on embedded output has its limitations as assets in distant clusters lose their meaning after neighbour embedding. Although the PCA initiation seems like a quick fix, it is generally unreliable to obtain a hierarchy from a transformed space. However, the embedding should be seen as an alternative for correlation modeling which in our portfolio construction framework merely serves as a means for dimension reduction for similarity measures. Embeddings thus have obvious limitations, but as a means of reducing some high-dimensional characteristics space (such as multivariate timeseries of returns) to low-dimensional similarity measures, they have their utility[62]. Moreover, alternatives such as UMAP — based on Riemannian geometry and algebraic topology — have proven better in preserving overall topology, i.e. intuitively 'hacking' the tradoff of what information to keep versus throw away.

Figure 29 shows how the correlation matrix of the original, unordered universe is reordered using the order predicted by embeddings. We see some structure revealing itself along the diagonal without having calculated correlations to determine this order. Embeddings learn at least some structure in the data that captures but is not limited to linear correlations. Whether that is a more meaningful approach than clustering on correlation or coiVaR is an open question.

In conclusion, embeddings allow us to rearrange and cluster our assets in our horizon without having to calculate correlations. Their caveats are twofold: (1) we have to decide which information to throw away when reducing dimensions, and (2) need to be careful in interpreting distances that were not fully preserved. As a counterargument to this, however, think about the amount of information we throw away by relying on linear correlations. Essentially, correlation modeling for hierarchical asset allocation where we do not use variance as a risk measure is merely a means of dimension reduction. The discussion and testing of more clustering methods from Table 13 goes beyond the scope of this paper, but the clear advantages of these methods open paths to further research in machine learning diversification. The implemented clustering-based algorithms can therefore be seen as a framework rather than a ready-to-use package, and they are easily compatible with more advanced clustering and dimension reduction techniques, as well as more and better characteristics data to train them on.

[62]The pros however are clear: dimension reduction can be applied on many more characteristics than merely (co-)returns or (co-)drawdowns. And the clusters can therefore by design have more meaning and be more stable, therefore resolving the main issue with variance clusters.

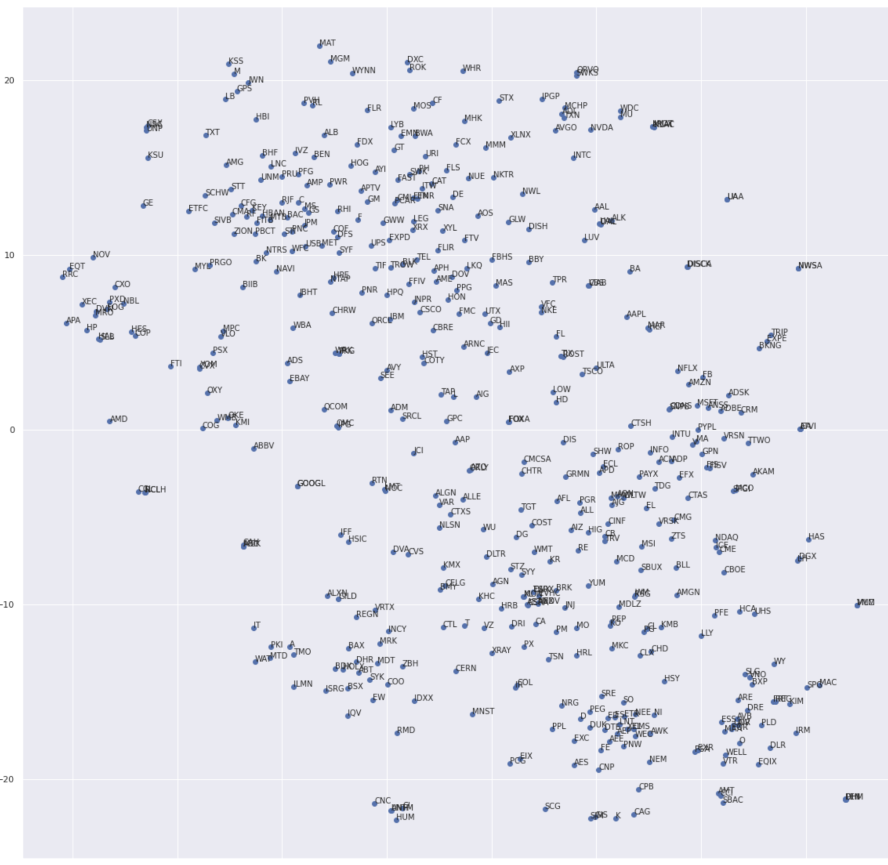
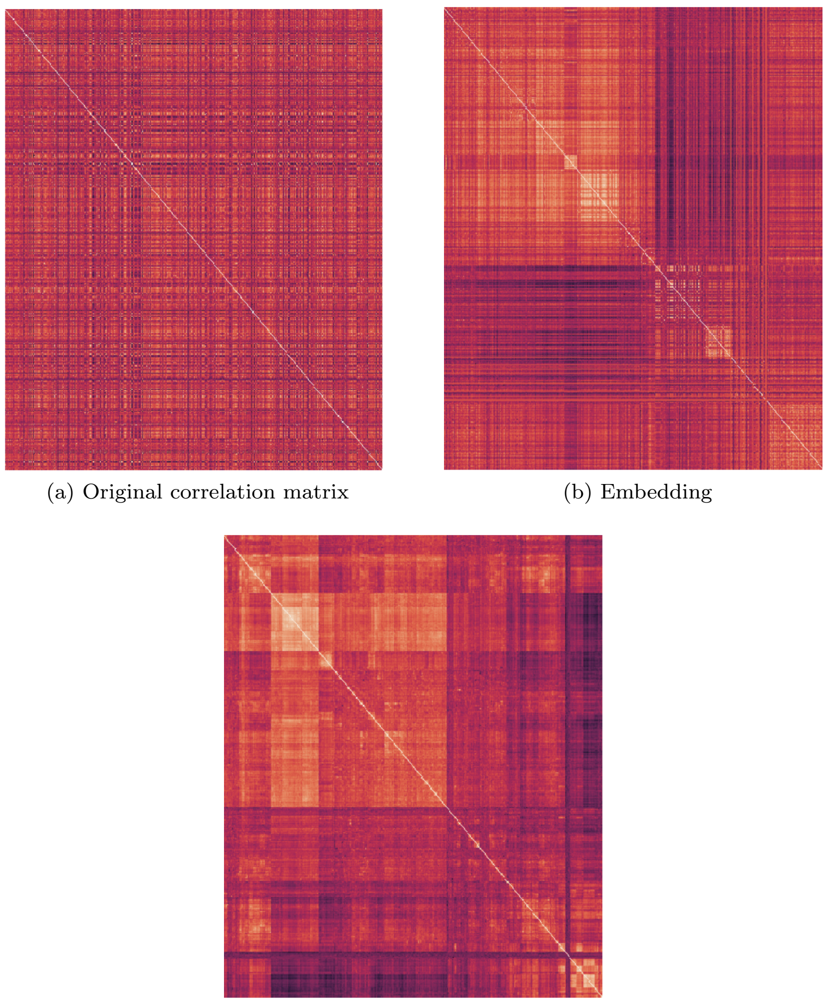


Figure 27: Embedding output (Neighbour embedding using t-SNE)

Figure 28: Embedded neighbours versus sectors



(a) Original correlation matrix

(b) Embedding

(c) HRP

Figure 29: Embeddings and correlation structures

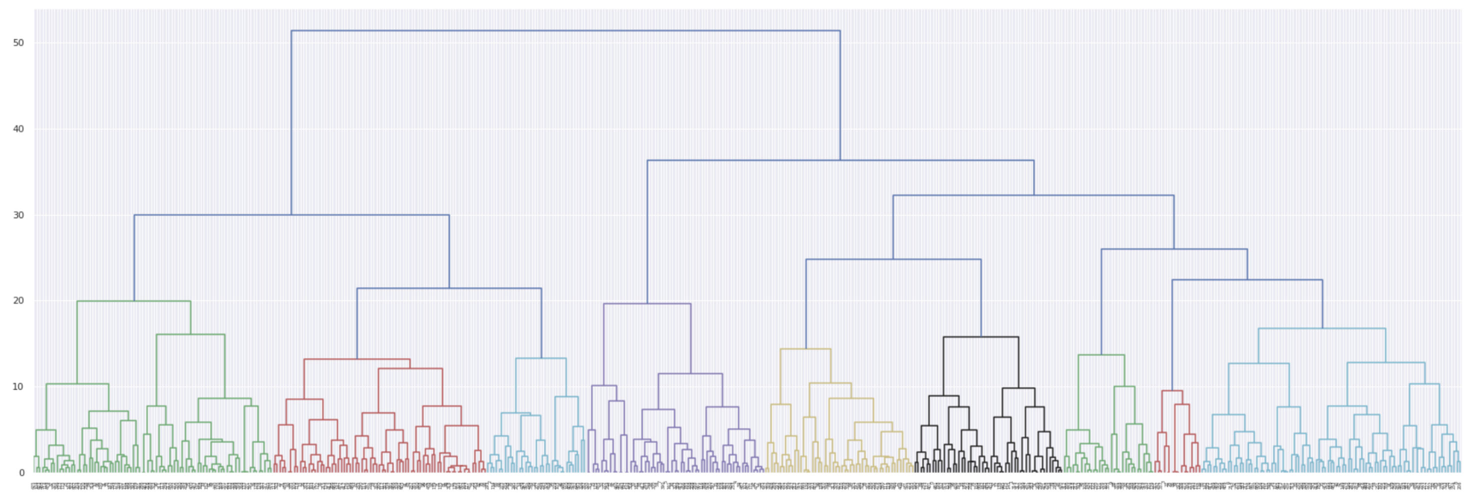


Figure 30: Embeddings and hierarchical clustering

H. Appendix 8: Quadratic utility and the max Sharpe ratio portfolio

In our discussion of modern portfolio theory we argued that mean-variance optimization, the maximization of quadratic utility for each investor's risk aversion λ, corresponds to the maximum Sharpe ratio portfolio for some critical level of risk aversion $\hat{\lambda}$. This level can be found by solving (11) iteratively, finding the parameter $\hat{\lambda}$ that maximizes the objective function. As it is key in the build-up of our argument in Chapter II, we borrow the proof of [25] to show that this perspective is equivalent to a direct maximum Sharpe ratio formulation, solvable using the program in 36. Say we want to obtain a portfolio using (11):

$$\text{Maximise } \mathbf{r}^T\mathbf{w} - \frac{\lambda}{2}\mathbf{w}^T\Sigma\mathbf{w} \tag{66}$$

for some constraints $g_i(\mathbf{w}) \leq 0$, $i = 1, ..., N$. Let us now call the expected return $\mu(\mathbf{w})$ as a function of $\mathbf{w}$ and the portfolio variance $\sigma(\mathbf{w})$ as a function of $\mathbf{w}$ as well. As long as $\mu(\mathbf{w})$, $\sigma(\mathbf{w})$ and $\frac{\mu(\mathbf{w})}{\sigma(\mathbf{w})}$ are convex functions, we can show that the problem:

$$\text{Maximise } \frac{\mu(\mathbf{w})}{\sqrt{\sigma(\mathbf{w})}} \tag{67}$$

has the same solution as the problem:

$$\text{Maximise } \mu(\mathbf{w}) - \hat{\lambda}\sigma(\mathbf{w}) \tag{68}$$

for the same constraints $g_i(\mathbf{w}) \leq 0$, $i = 1, ..., N$, i.e. that $\hat{\lambda}$ exists and the two problems are equivalent.

The gradient of the problem becomes:

$$\frac{1}{\sqrt{\sigma(\mathbf{w})}}(\mu'(\mathbf{w}) - \frac{1}{2}\sigma'(\mathbf{w})\frac{\mu(\mathbf{w})}{\sigma(\mathbf{w})} \tag{69}$$

Say $\mathbf{w}^*$ are the optimal weights for problem (67) then the objective part of the Karush-Kuhn-Tucker conditions,

$$\frac{\mu'(\mathbf{w}^*)}{\sqrt{\sigma(\mathbf{w}^*)}} - \frac{1}{2}\frac{1}{\sqrt{\sigma(\mathbf{w}^*)}}\frac{\mu(\mathbf{w}^*)}{\sigma(\mathbf{w}^*)}\sigma'(\mathbf{w}^*) + \sum_i \pi_i^1 g_i'(\mathbf{w}) = 0 \tag{70}$$

with π_i^1 the optimal dual vector for $\mathbf{w}^*$, is the same as the objective part of the KKT conditions for (68) by setting is dual vector $\pi^2 = \sqrt{\sigma(\mathbf{w})}\pi^1$ and our risk aversion λ equal to $\frac{1}{2}\frac{\mu(\mathbf{w}^*)}{\sigma(\mathbf{w}^*)}$.

In summary, it is clear that within modern portfolio theory the mean-variance criterion and maximizing Sharpe ratios are two seperate things, and jumping from the MV criterion to maximizing Sharpe ratios was a bit sloppy mathematically in the build-up of our argument. However, we wanted to introduce the MPT problem setting in an intuitive way without dropping formulas on QPs, efficient frontiers and so forth. Hopefully this proof shines some light on this matter: maximizing the Sharpe ratio is a particular MV problem for a critical level of risk aversion, which in portfolio theory can be seen as the risk aversion of the market as the maximum Sharpe ratio or MPT portfolio is essentially the optimal market portfolio.

I. Appendix 9: Ratio objectives as quadratic problems

One small technical problem remaining is that a QP/LP generally does not take a ratio as an objective function, but only linear and quadratic transformations of the problem variables — hence the name. Therefore, we again borrow from [25] and show how we can redefine a ratio problem into a standard QP using a simple variable transformation.

Say we have a maximum Sharpe ratio problem:

$$\begin{aligned} \text{Maximise } & \frac{\mathbf{r}^T\mathbf{w}}{\sqrt{\mathbf{w}^T\Sigma\mathbf{w}^T}} \\ \text{s.t. } & \mathbf{A}\mathbf{w} \geq \mathbf{b} \\ & \mathbf{e}^T\mathbf{w} = 1 \end{aligned} \tag{71}$$

The important condition here is that for the optimal weights $\mathbf{w}^*$ the expected return $\mathbf{r}^T\mathbf{w}^*$ is positive, such that we can set $\mathbf{r}^T\mathbf{w}^* = \frac{1}{\zeta} \geq 0$. Call the risk of the optimal portfolio σ^* then the objective value, the Sharpe ratio, is nothing but $\frac{1}{\zeta\sqrt{\sigma^*}}$.

Now consider the following QP:

$$\begin{aligned} \text{Minimize } & \mathbf{g}^T\Sigma\mathbf{g} \\ \text{s.t. } & \mathbf{r}^T\mathbf{g} = 1 \\ & \mathbf{A}\mathbf{g} - \zeta\mathbf{b} \geq 0 \\ & \mathbf{e}^T\mathbf{g} - \zeta = 0 \end{aligned} \tag{72}$$

Given we enforce $\mathbf{r}^T\mathbf{g} = 1$, $\mathbf{g} = \zeta\mathbf{w}$. The objective value is then $\zeta^2\sigma^* = \frac{1}{(\frac{1}{\zeta\sqrt{\sigma^*}})^2}$ which is the squared inverse of the objective of (71). Hence, maximizing the Sharpe ratio is minimizing this objective. If (71) is feasible for an optimal $\mathbf{w}^*$ and objective SR^* then there exist a solution for $(\mathbf{g}, \zeta)$ in (72) with objective $\frac{1}{SR^2}$.

This formulation paves the road for our required implementations of the MPT and MDP portfolios as a QP.

J. Appendix 10: Hansen's Model Confidence Set

For the assessment of our backtest results we used the Hansen 2011 Model Confidence Set (MCS) approach [26]. Our problem is as follows: we have timeseries of OOS performance, risk and diversification of our different portfolios and we want to know which ones significantly out- or underperform. We have called our total pool of models our model set. Now we want to divide this set into two groups based on a chosen measuren, each model belonging either to the set of models that do not statistically differ from the best-in-class in terms of this measure, i.e. belonging to the set of superior models, or belonging to the inferior group.

So we evaluate our set M with i = 0,...,m models (in our case m=18). Let $d_{i,j,t}$ denote the difference in loss L between two models i and j at some point t in our backtest:

$$d_{i,j,t} = L_{i,t} - L_{j,t} \tag{73}$$

where we can define some loss function L as a negative property of portfolios (e.g. risk measures like volatility, drawdowns, etc.) or the negative of a desirable property (e.g. returns, Sharpe ratios, etc.).

The set of superior models is defined as [34]:

$$M^*_{1-\alpha} = \{i \in M : E[d_{i,j}] \leq 0, \forall j \in M\} \tag{74}$$

The MCS uses a sequential test of loss equivalence [26]. When this test rejects the null hypthesis, one or more models in the set M is found inferior, such that the model with the largest average loss differential (see below) is removed from the set. This sequential procedure is repeated until the null is not rejected anymore and the remaining set is called the model confidence set $M^*_{1-\alpha}$, depending on the chosen confidence level $cl = 1 - \alpha$ for each hypothesis test [36].

In [26] Hansen et al. prove the use of the following hypothesis and test statistics:

$$\begin{aligned} H_{0,M_k} &: E[d_{i,j,t}] = 0, \forall i, j \in M_k, M_k \in M \\ H_{A,M_k} &: E[d_{i,j,t}] \neq 0 \text{ for some } i, j \in M_k \end{aligned} \tag{75}$$

$$\begin{aligned} t_{i,j} &= \frac{\bar{d_{i,j}}}{\sqrt{var(\bar{d_{i,j}})}} \\ t_i &= \frac{\bar{d_i}}{\sqrt{var(\bar{d_i})}} \end{aligned} \tag{76}$$

where $\bar{d_i} = \frac{1}{m'-1}\Sigma_{j \in M_k}\bar{d_{i,j}}$ is the L of the ith model compared to the average loss across the m' models in M_k. $\bar{d_{i,j}} = \frac{1}{m'}\Sigma_{t=1}^{m'} d_{i,j,t}$ measures the relative sample loss between model i and j [34].

The distributions of the test statistics obviously depend on unknown parameters, such as the choice of L. Hansen thus proposes a bootstrapped procedure to estimate these distributions [34].

K. Appendix 11: Corona, the ultimate backtest for diversification

As an additional backtest, we compare diversified iVaR portfolios with the market during the March 2020 coronavirus-induced market turmoil. We look at the badly hit European equities, and pick as universe the 50 constituent of the Eurostoxx 50 index.

We notice that MIP outperformed during the initial stage of the decline in the end of February and March. This can be explained by a flight to safety: during extreme market events traditional diversification does not fair well, because ironically *concentrations* in low risk/iVaR assets preserve capital most effectively during crashes. In essence, truly diversified allocations are superior over an extended period, or multiple periods with market uncertainty like we simulated before. During a collapse, concentrated low risk allocations outperform, but these assets then significantly underperform

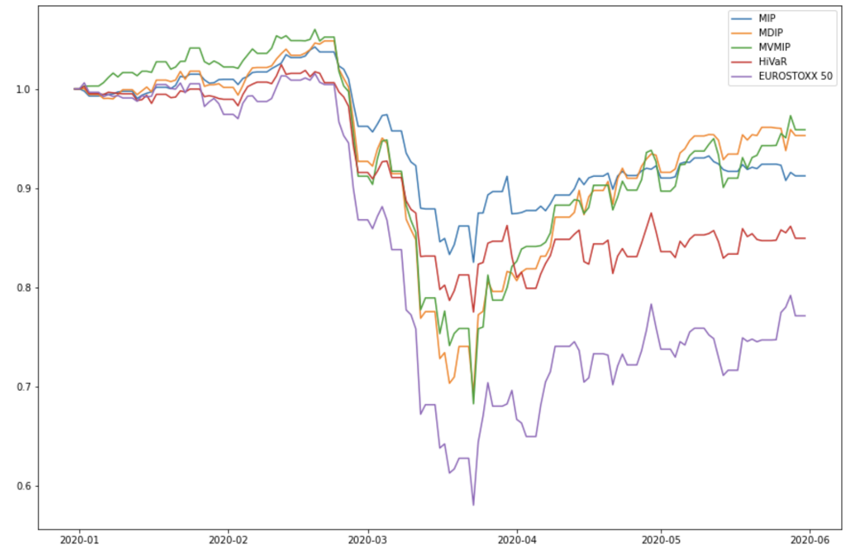
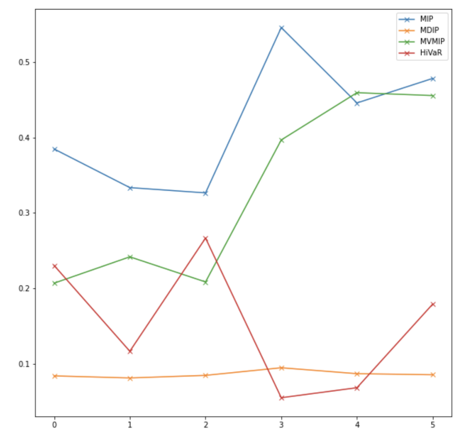


Figure 31: Portfolio values during Covid-19 market turbulence

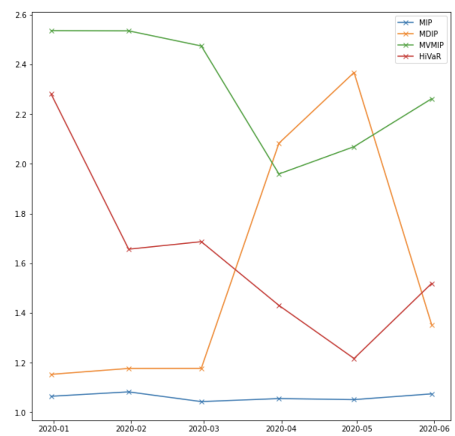
Table 14: Monthly OOS performance iVaR models

	Return (%)	Volatility (%)	Sharpe	D_{var}	D_{iVaR}
MIP	-1.83388	4.82871	-0.030714	1.06216	3.51702
MDIP	-0.96388	8.52898	0.096088	1.55174	4.28219
MVMIP	-0.84091	9.30815	0.213879	2.30610	2.23209
HiVaR	-3.26055	6.96595	-0.489072	1.63201	2.75985

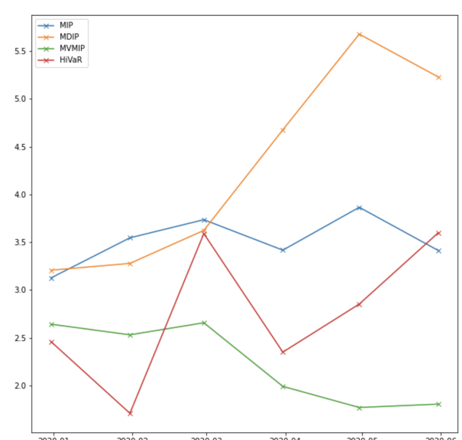
when the markets recover with the same pace as they imploded. In other words, during the contraction the diversified portfolios were overdiversified: (1) MVMIP essentially because volatility breaks down, (2) MDIP because it is simply contraproductive to push for higher drawdown assets during a persistent down movement, and (3) HiVaR is often overdiversified over clusters, even in normal market circumstances. After the dip MDIP and MVMIP profit from their respective iVaR and volatility boost resulting in rapid recovery, while MIP and HiVaR miss out on the rally. From Figure 32 we can say it is remarkable how MDIP is able to achieve similar performance as MVMIP while having way lower weight concentration. It is unable to create diversification benefits during the downturn, but spikes afterwards, which explains the competitive performance with extremely low weight concentration. MVMIP is naturally most diversified from a variance perspective, but is most concentrated from an iVaR perspective, which explains its biggest drawdown.



(a) Herfindahl concentration



(b) D_{var}



(c) D_{iVaR}

Figure 32: Diversification during market turmoil

References

[1] Harry Markowitz. "The utility of wealth". In: *Journal of political Economy* 60.2 (1952), pp. 151–158.

[2] Andrew Donald Roy. "Safety first and the holding of assets". In: *Econometrica: Journal of the econometric society* (1952), pp. 431–449.

[3] Harry Markowitz. *Portfolio selection: Efficient diversification of investments*. Vol. 16. John Wiley New York, 1959.

[4] Abraham Charnes and William W Cooper. "Programming with linear fractional functionals". In: *Naval Research logistics quarterly* 9.3-4 (1962), pp. 181–186.

[5] Vladimir A Marčenko and Leonid Andreevich Pastur. "Distribution of eigenvalues for some sets of random matrices". In: *Mathematics of the USSR-Sbornik* 1.4 (1967), p. 457.

[6] Sanford J Grossman and Joseph E Stiglitz. "Information and competitive price systems". In: *The American Economic Review* 66.2 (1976), pp. 246–253.

[7] Peter C Fishburn. "Mean-risk analysis with risk associated with below-target returns". In: *The American Economic Review* 67.2 (1977), pp. 116–126.

[8] Marc De Clercq. *Economie toegelicht*. Vol. 632. Garant, 1992.

[9] Walt Woerheide and Don Persson. "An index of portfolio diversification". In: *Financial services review* 2.2 (1992), pp. 73–85.

[10] Philippe Artzner et al. "Coherent measures of risk". In: *Mathematical finance* 9.3 (1999), pp. 203–228.

[11] David N Nawrocki. "A brief history of downside risk measures". In: *The Journal of Investing* 8.3 (1999), pp. 9–25.

[12] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[13] Anders Johansen and Didier Sornette. "Large stock market price drawdowns are outliers". In: *Journal of Risk* 4 (2002), pp. 69–110.

[14] R Tyrrell Rockafellar and Stanislav Uryasev. "Conditional value-at-risk for general loss distributions". In: *Journal of banking & finance* 26.7 (2002), pp. 1443–1471.

[15] Galen Burghardt, Ryan Duncan, and Lianyan Liu. "Understanding drawdowns". In: *CARR futures research note www. carrfutures. com* (2003).

[16] Foort Hamelink and Martin Hoesli. "Maximum drawdown and the allocation to real estate". In: *Journal of Property Research* 21.1 (2004), pp. 5–29.

[17] Olivier Ledoit and Michael Wolf. "Honey, I shrunk the sample covariance matrix". In: *The Journal of Portfolio Management* 30.4 (2004), pp. 110–119.

[18] Alexei Chekhlov, Stanislav Uryasev, and Michael Zabarankin. "Drawdown measure in portfolio optimization". In: *International Journal of Theoretical and Applied Finance* 8.01 (2005), pp. 13–58.

[19] John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior (commemorative edition)*. Princeton university press, 2007.

[20] Yves Choueifaty and Yves Coignard. "Toward maximum diversification". In: *The Journal of Portfolio Management* 35.1 (2008), pp. 40–51.

[21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[22] Lior Rokach and Oded Z Maimon. *Data mining with decision trees: theory and applications*. Vol. 69. World scientific, 2008.

[23] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Vol. 28. Princeton University Press, 2009.

[24] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" In: *The review of Financial studies* 22.5 (2009), pp. 1915–1953.

[25] Leonid Kopman and Scott Liu. "Maximizing the sharpe ratio, June 2009". In: *MSCI Barra Research Paper* 2009-22 (2009).

[26] Peter R Hansen, Asger Lunde, and James M Nason. "The model confidence set". In: *Econometrica* 79.2 (2011), pp. 453–497.

[27] Jochen Papenbrock. "Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization". PhD thesis. Karlsruher Institut für Technologie (KIT), 2011.

[28] Zvi Bodie, Alex Kane, and Alan J Marcus. *Essentials of investments*. McGraw-Hill/Irwin Taipei, 2013.

[29] Thierry Roncalli. *Introduction to risk parity and budgeting*. CRC Press, 2013.

[30] Maria-Florina Balcan, Yingyu Liang, and Pramod Gupta. "Robust hierarchical clustering". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3831–3871.

[31] Peter Christoffersen et al. "Correlation dynamics and international diversification benefits". In: *International Journal of Forecasting* 30.3 (2014), pp. 807–824.

[32] Fotis Papailias and Dimitrios D Thomakos. "Covariance averaging for improved estimation and portfolio allocation". In: *Financial Markets and Portfolio Management* 29.1 (2015), pp. 31–59.

[33] Marcos López de Prado. "Building diversified portfolios that outperform out of sample". In: *The Journal of Portfolio Management* 42.4 (2016), pp. 59–69.

[34] Thomas Raffinot. "Hierarchical clustering-based asset allocation". In: *The Journal of Portfolio Management* 44.2 (2017), pp. 89–99.

[35] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[36] Thomas Raffinot. "The hierarchical equal risk contribution portfolio". In: *Available at SSRN 3237540* (2018).

[37] Prayut Jain and Shashi Jain. "Can Machine Learning-Based Portfolios Outperform Traditional Risk-Based Portfolios? The Need to Account for Covariance Misspecification". In: *Risks* 7.3 (2019), p. 74.

[38] Marcos Lopez de Prado. "Beyond Econometrics: A Roadmap Towards Financial Machine Learning". In: *Available at SSRN 3365282* (2019).

[39] Gautier Marti. "CorrGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8459–8463.

[40] Erling D Andersen, Joachim Dahl, and Henrik A Friberg. "Markowitz portfolio optimization using MOSEK. MOSEK Technical report: TR-2009-2". In: ().

[41] Jim Gatheral. "Random matrix theory and covariance estimation". In:

[42] Scipy. *Scipy Hierachy Linkage Docs*. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html.