Understanding Machine Learning for Diversified Portfolio Construction by eXplainable AI

Markus Jaeger[1*] Stephan Krügel[1†] Dimitri Marinelli[2,3‡] Jochen Papenbrock[2§]
Peter Schwendner[4¶]

[1] Munich Reinsurance Company, Financial Solutions, Königinstraße 107, 80802 Munich, Germany
[2] Firamis GmbH, Robert-Kempner-Ring 27, 61440 Oberursel, Germany
[3] FinNet project - MSCA-H2020-EU
[4] Zurich University of Applied Sciences, Center for Asset Management, Technoparkstrasse 2, CH-8400 Winterthur

Abstract

In this paper, we construct a pipeline to investigate heuristic diversification strategies in asset allocation. We use machine learning concepts ("explainable AI") to compare the robustness of different strategies and back out implicit rules for decision making.

In a first step, we augment the asset universe (the empirical dataset) with a range of scenarios generated with a block bootstrap from the empirical dataset.
Second, we backtest the candidate strategies over a long period of time, checking their performance variability. Third, we use XGBoost as a regression model to connect the difference between the measured performances between two strategies to a pool of statistical features of the portfolio universe tailored to the investigated strategy. Finally, we employ the concept of Shapley values to extract the relationships that the model could identify between the portfolio characteristics and the statistical properties of the asset universe. On the basis of this information, we discuss the similarity between the bootstrapped datasets characterized by their Shapley values using a network technique.

We test this pipeline for studying risk-parity strategies with a volatility target, and in particular, comparing the machine learning-driven Hierarchical Risk Parity (HRP) to the classical Equal Risk Contribution (ERC) strategy. In the augmented dataset built from a multi-asset investment universe of commodities, equities and fixed income futures, we find that HRP better matches the volatility target, and shows better risk-adjusted performances. Finally, we train XGBoost to learn the difference between the realized Calmar ratios of HRP and ERC and extract explanations. The explanations provide fruitful ex-post indications of the connection between the statistical properties of the universe and the strategy performance in the training set. For example, the model confirms that features addressing the hierarchical properties of the universe are connected to the relative performance of HRP respect to ERC.

Contents

*majaeger@munichre.com
†skruegel@munichre.com
‡dm@firamis.de
§jp@firamis.de
¶scwp@zhaw.ch

1 Introduction

After the financial crisis of 2008, the interest, both among practitioners and academics, grew for allocations that equally budget the risk for the assets in a portfolio, the so-called Risk Parity allocation strategies. Later in the last decade, funds based on the risk parity principle experienced a hit (see, e.g. WSJ[1]). An explanation for this event was sudden correlated drawdowns across asset classes ("correlation breakdown"). Funds that base their strategies on risk budgeting, to achieve a higher return per unit of risk, leverage their portfolios (Asness, Frazzini, and Pedersen (2012)). This practice leads to better returns, in some cases, higher Sharpe ratios (Moreira and Muir (2017)) and in general, a lower likelihood of extreme returns (Harvey et al. (2018)). However, a disadvantage of leveraged strategies is the requirement for dynamic risk management. After an adverse market movement, a dynamic leverage strategy like a "risk control with volatility target" leads to a reduction of the portfolio positions, thus realizing the loss and reducing the probability of subsequent recovery.

Some recent advances in portfolio construction are based on new diversification approaches, complex information filtering and graph theory. The use of correlations, hierarchies, networks and clustering in financial markets has become a mature research field since its inception 20 years ago (see an overview in Marti et al. (2017)). Some practical applications of correlation networks in portfolio management and market risk analysis can be found in Pozzi, Di Matteo, and Aste (2013), Papenbrock and Schwendner (2015), Baitinger and Papenbrock (2017) and Huettner, Mai, and Mineo (2018). One of the first practical applications of correlation clustering in portfolio construction uses a dendrogram structure to allocate the capital to the positions in the portfolio: Papenbrock (2011).

Another recent approach using dendrogram structures is called Hierarchical Risk Parity (HRP) (Lopez de Prado (2016a)). It uses graph theory and machine learning whose benefits have also been discussed in Lopez de Prado (2016b) and Focardi and J Fabozzi (2016). The idea is to utilise representation learning like clustering to filter relevant information in noisy data. The HRP approach uses hierarchical cluster structure in order first to rearrange the correlation matrix into a hierarchical structure. In a second next step, this information is used to allocate portfolio weights proportional to the inverse variance at each split of a bisection. As we will see later, such portfolio allocations can result in more robust investment performance, less prone to noise. The reason is that estimating and inverting covariance matrices to optimise a risk measure often leads to errors of such magnitude that they entirely offset the benefits of diversification. Small estimation errors from short samples and of course structural breaks in the market dynamics lead to grossly incorrect inversions. This is a problem for Mean-Variance and Minimum Variance portfolios which are optimal in-sample but tend to perform poorly out-of-sample, but also for optimised Risk Parity strategies. The Markowitz's curse (Michaud and Michaud (2008)) is that optimisation is likely to fail precisely when there is a greater need for finding a diversified portfolio.

Machine learning is becoming increasingly important in the financial industry, see for example Lopez de Prado (2018) and López de Prado (2019). But in many decision-making applications, regulatory and transparency concerns slowed down the industry from embracing these new technologies, despite their massive success in backoffice process automation and in other domains like computer vision (see, e.g. LeCun, Bengio, and Hinton (2015)). One approach to overcome this limitation is eXplainable AI (XAI) (Murdoch et al. (2019) and Du, Liu, and Hu (2020)). XAI not only delivers the desired quantitative result, but also reports the reasoning to make its functioning clearer to understand by humans. The ability to explain model decisions to stakeholders

[1] https://www.wsj.com/articles/SB10001424127887323689204578572050047323638

contributes to fulfil regulatory compliance requirements and to foster trust to accelerate adoption. More on the recent development and a practical credit risk application of XAI can be found in Bussmann et al. (2019).

In this paper, we introduce a novel way of using XAI in asset allocation. We use the XAI explanation to investigate in hindsight how sophisticated strategies like HRP perform relative to a classical strategy like ERC (Equal Risk Contribution). To do so, we augment the empirical asset time series dataset with a large number of bootstrapped datasets to explore a wide range of plausible scenarios. We summarise the properties of these datasets and train a machine learning model to learn which strategy, either ERC or HRP, would have performed better with datasets that reflect certain features. The selected model can adapt to non-linear relations among the returns of these artificial datasets. Finally, we use XAI explanations to show the relationships discovered by the model to make the strategies more transparent for a financial market practitioner.

Original contribution

The original contribution of this paper is, first, a method to link the statistical properties of an investment universe to the outperformance of risk-based investment strategies. Risk-based strategies only use a point estimate of VCV derived using historical returns of the considered portfolio components. The link is established by explanation technologies of trained supervised learning models. Explanations are delivered in terms of importance metrics of features (the statistical properties) that are directly linked to outperformance probabilities. Second, the underlying data set used to train the supervised learning model is augmented by a collection of bootstrapped scenarios of investment universes. This augmentation lowers the dependence on a specific empirical dataset representing a certain time frame for the defined investment universe. Third, we apply the system for a horse race between the HRP and ERC allocation strategies, both subject to a typical dynamic leverage rebalancing mechanism in the form of a typical risk control concept using a volatility target.

2 Preliminaries

The dataset

In this work, we focus on a multi-asset investment universe of commodities, equities and fixed income futures. Our time series spans the period 2000-05-03 to 2019-10-04 with daily frequency, more than twenty years that cover the dot-com crisis, the global financial crisis, the European debt crisis and the subsequent bull markets. We use listed futures as instruments as this is the most cost-efficient way of obtaining a global cross-asset allocation. Furthermore, futures as unfunded derivatives enable dynamic leverage approaches like a volatility target concept in contrast to fully funded instruments like ETFs.

Ticker	Asset class	Currency	Name
CLA Comdty	Commodities	USD	NYMEX WTI Light Sweet Crude Oil
GCA Comdty	Commodities	USD	COMEX Gold
SIA Comdty	Commodities	USD	COMEX Silver
BZA Index	Equities	BRL	BM&F IBOVERSPA
ESA Index	Equities	USD	CME E-mini S&P 500
HIA Index	Equities	HKD	HKFE Hang Seng
NKA Index	Equities	JPY	OSE Nikkei 225
NQA Index	Equities	USD	CME E-mini NASDAQ-100
SMA Index	Equities	CHF	Eurex SMI
VGA Index	Equities	EUR	Eurex EURO STOXX 50
XPA Index	Equities	AUD	ASX SPI 200
Z A Index	Equities	GBP	ICE FTSE 100
CNA Comdty	Fixed Income	CAD	10Y Canadian GB
G A Comdty	Fixed Income	GBP	ICE Long Gilt
JBA Comdty	Fixed Income	JPY	OSE 10Y JGB
RXA Comdty	Fixed Income	EUR	Eurex 10Y Euro-Bund

Ticker	Asset class	Currency	Name
TYA Comdty	Fixed Income	USD	CBOT 10Y US T-Note
XMA Comdty	Fixed Income	AUD	ASX 10Y Australian T-Bonds

The Strategies

We implemented several industry-standard strategies that focus on diversifying the risk among the assets. The portfolio is rebalanced every month and leveraged to realise the target volatility. We begin our discussion with two strategies that do not make use of the correlation among the assets (Inverse Variance and Naive Risk Parity), and two that use the full and filtered information of the VCV matrix Σ, respectively Equal Risk Contribution and Hierarchical Risk Parity. In the literature, and among practitioners, there are many other strategies and variations of these strategies, but their analysis goes beyond the scope of this paper.

Naive Risk Parity Naive Risk Parity (RP), here called naive because it ignores the correlation among the assets, distributes the weights proportional to the inverse of the volatility of each asset in the universe. In fact, if we consider volatility as the risk measure, RP portfolios are composed of assets, each contributing equally to the full portfolio once assumed that they are equally correlated with each-other Roncalli (2013). More formally, the weight w_i for the i-th-asset with i spanning the portfolio universe $i = 1, ..., N$ is

$$w_i = \frac{\sigma_i^{-1}}{\sum_{j=1} \sigma_j^{-1}}$$

where $\sigma_i = \sqrt{\Sigma_{ii}}$ is the volatility of asset i.

Equal Risk Contribution ERC portfolios (Maillard, Roncalli, and Teiletche (2010), Qian (2005), Neukirch (2008)) use the full information in the VCV matrix to budget equally the risk among the assets. For ERC portfolios w, the percentage volatility risk contribution of the i-th asset in the portfolio is given by:

$$\mathcal{RC}_i = \frac{w_i \ [\Sigma\, w]_i}{\sqrt{(w'\Sigma w)}}$$

The ERC portfolio is achieved by solving the following optimization problem:

$$\underset{w}{\text{argmin}} \left[\sum_{i=1}^{N} (\frac{\mathcal{RC}_i}{\sqrt{(w'\Sigma w)}} - \frac{1}{N})^2 \right].$$

Inverse Variance Inverse variance corresponds to minimum variance when correlation among assets is negligible. The portfolio weight of each asset is proportional to the inverse of its variance, namely

$$w_i = \frac{1/\sigma_i^2}{\sum_j \left(1/\sigma_j^2\right)} \quad \text{with } \sigma_i^2 = \Sigma_{ii}$$

HRP The standard HRP approach (Lopez de Prado (2016a)) uses a tree clustering algorithm to perform a quasi-diagonalization of the the VCV matrix. After the quasi-diagonalization is carried out a recursive bi-sectioning method is used to define the weights of each asset within the portfolio. The details of this process can be found in the appendix. Variations of this approach use some well-known additional building blocks for processing the time series data. These blocks are executed in a sequential way. Each block can be replaced by appropriate methods which might be more suitable. This, in turn, leads to a large variety of HRP-like approaches.

An example of the first step in information filtering is the choice of the correlation function, which is the basis for the hierarchical clustering step in HRP. Many papers in literature use the Pearson correlation coefficient matrix but obviously, there can be more robust and non-linear alternatives.

The next step is the choice of distance function, which transforms the correlation information to a matrix that describes the distance or dissimilarity of the assets. In the literature, there is often the Gower distance which results in euclidean distance matrices. Also, there are papers on feature-based distances where the features capture several statistical properties and stylised facts. Resulting distance matrices can further be made more robust by using the distance of distance approach by de Prado.

The third step is the choice of the hierarchical clustering procedure (HRP uses the single linkage clustering). More generally speaking, the hierarchical clustering is used to reorder the correlation matrix (quasi-diagonalisation) to process it with a bisectioning method later, and this rearrangement could be done in numerous alternative ways. Alternatives to single linkage clustering are absolute linkage and complete linkage. There could also be an adaptive procedure that chooses among the best linkage methods in each step in time. Some approaches also use a mixture of hierarchical clustering up to a certain tree cutting level and then proceed with a discrete/flat clustering.

Finally, the last step of HRP uses the reordered VCV matrix to come up with a portfolio following inverse variance.

The backtests

The strategies are rebalanced every month. At every rebalancing date, the portfolio leverage is set to reach the volatility target of $\sigma_{\text{target}} = 5\%$ annualized in a hindsight. The portfolio leverage determines the total market value of the portfolio and thus the position sizes of each instrument. The estimation of realized volatility used for the leverage is the maximum of the volatilities of the portfolio measured over 20 and 60 trading days, respectively $\sigma_{t=20}$ and $\sigma_{t=60}$. This is a popular approach in the industry (see for example ("Guide to the Strategy Indices of Deutsche Börse AG" 2018)). The target weight is calculated as

$$W^{\text{target}} = \frac{\sigma_{\text{target}}}{\max(\sigma_{t=20}, \sigma_{t=60})}.$$

And the unnormalized portfolio weights are $\tilde{w}_i = W^{\text{target}} w_i$.

We considered half-turn transaction cost of 2 bp (flat) in the performance evaluation.

At every rebalancing date, the parameters for the strategies are estimated on the last 252 trading days. The estimation of the VCV is a crucial ingredient of this work and is done employing the exponentially weighted estimation with decay parameter $\lambda = 0.97$ (Jacques Longerstae and Martin Spencer (1996)).

Performance statistics

Statistics	Short	Description
Volatility	SD	Annualized volatility
Returns	RET	Annualized returns
Maximum Drawdown	MDD	Drawdowns percentage
Conditional Value-at-Risk	CVaR	Conditional Value-at-Risk with confidence interval p=0.95
Sharpe ratio	SR	The ratio between returns and volatility (annualized)
Calmar Ratio	Calmar	The ratio between annualized returns and max drawdown

Results for the empirical dataset For the futures universe, the strategies performed as follows

	RP	ERC	HRP
SD	0.0518	0.0556	0.0509
RET	0.0488	0.0576	0.0529

	RP	ERC	HRP
MDD	0.1645	0.1993	0.0995
Calmar	0.2968	0.2890	0.5320
CVaR	-0.0097	-0.0112	-0.0099
SR	0.9432	1.0352	1.0389

We notice that HRP and RP better reach the volatility target than ERC. ERC with an higher volatililty reaches also higher returns, but a lower Sharpe ratio. HRP dominates in terms of Calmar ratio, which takes the maximum drawdown into account. The drawdowns (see Fig. 1) often determine if a buy-side investor can keep an investment or will have to unwind and thus will miss subsequent recoveries. For this reason, the Calmar ratio is of specific interest both for the buy-side investor and for the manager.

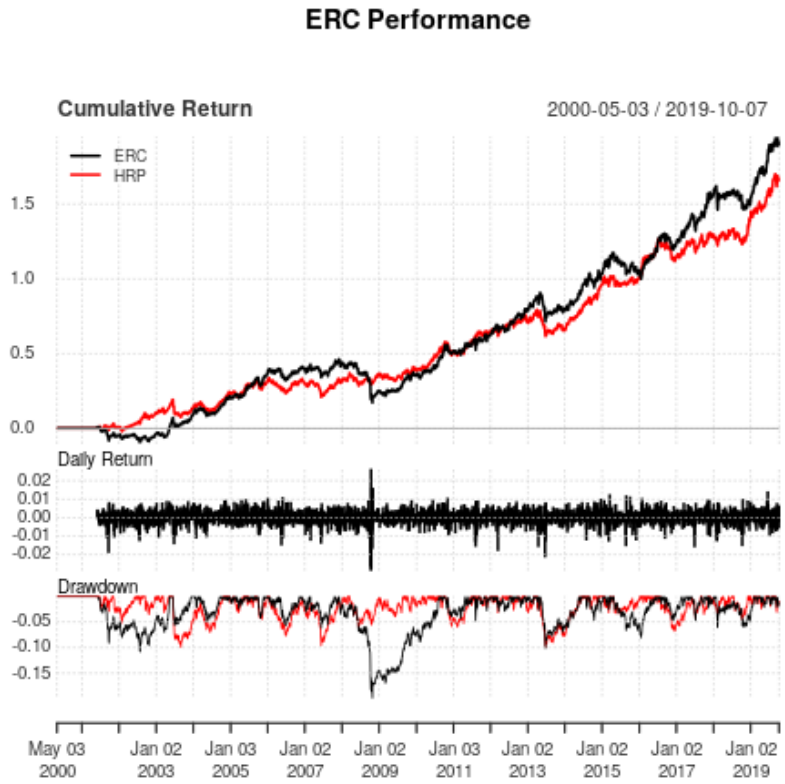


Figure 1: Cumulative log returns of HRP and ERC strategies applied to the empirical dataset of a multi-asset futures portfolio with a dynamic 5% volatility target.

3 Robustness of the strategies

Data augmentation

Bootstrapped dataset To account for the non-stationarity of futures return time series, we generate an additional dataset of time-series by block bootstrapping (Hall (1985), Carlstein and others (1986), Fengler and Schwendner (2004) and Lohre, Rother, and Schaefer (2020)):

- Blocks with a fixed length, but a random starting point in time are defined from the futures return

time-series. One block corresponding to 60 business days. This block length is motivated by a typical rebalancing frequency of dynamic rule-based strategy and by the empirical market dynamics that happen on this time scale (Papenbrock and Schwendner (2015)).
- A new return time-series is constructed by sampling the blocks with replacement to reconstruct a time-series with the same length of the original time-series.

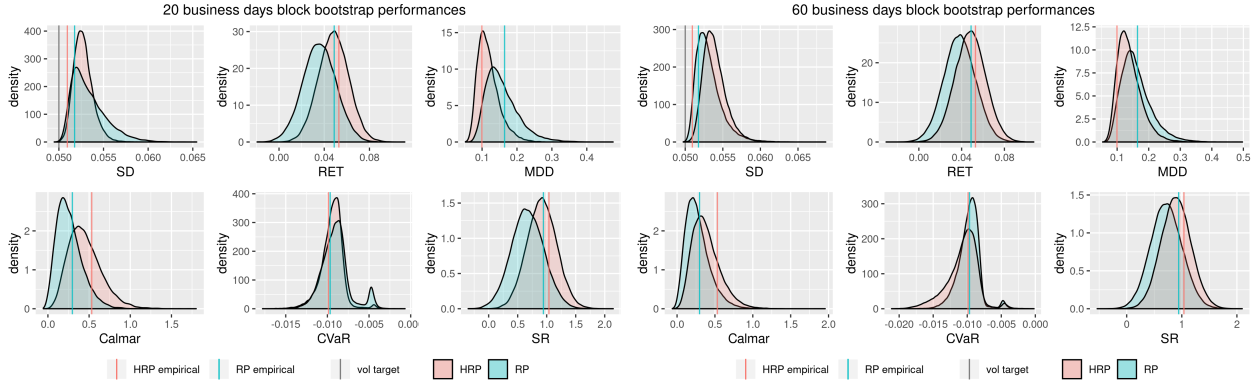
We generate 100.000 bootstrapped return time-series for each of the 18 multi-asset futures markets with a block length of 60 days. To enable a discussion about the impact of the block length, we also generated 25.000 additional bootstrap samples with a block length of 20 days.

Computational effort

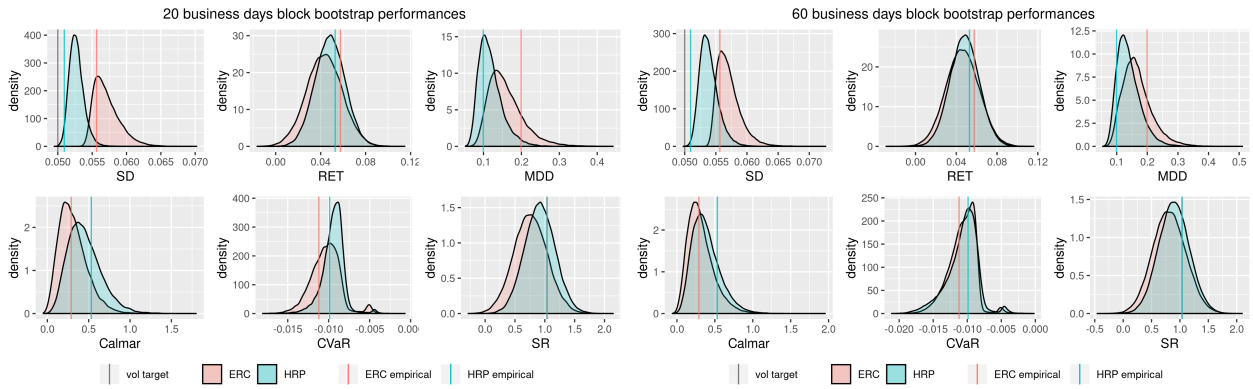
For the simulations and the backtests we employ 15 high-performance computers with 96 CPUs each in a highly parallelized environment.

Results

We display our results in the form of panels with densities for various performance and risk measures across the bootstrapped datasets to compare first HRP versus RP, and second HRP versus ERC. On the left hand panel, we show the densities across the bootstrapped datasets with a 20 day block length, and on the right hand panel with a block length of 60 days. Vertical lines point to the values for the empirical datasets. For the 20 day block length (panel on the left), HRP delivers lower standard deviations (SD) of returns and a better compliance with the 5% volatility target, higher returns (RET) and less pronounced maximum drawdowns (MDD) than Naive Risk Parity (RP). This leads to higher Sharpe and Calmar ratios for HRP compared to RP. On the other hand side, the distributions for CVaR look similar for HRP versus RP. The shape of the densitites for a specific performance or risk measure is very similar for the two different block lengths. However, there are subtle changes: the advantage of HRP versus RP in terms of performance and risk seems to decrease at the increased block length. The standard deviation of HRP exceeds that of RP, and the CVaR even develops a long left tail for HRP versus RP at the larger block length. In terms of risk adjusted return (Sharpe and Calmar ratios), HRP still outperforms RP also at the larger block length.



To have a competition with a method that also accounts for the full covariance matrix like HRP, we consider Equal Risk Contribution (ERC). The two panels below show the performance and risk density plots for HRP versus ERC. Please note the inverted colours: the colour of HRP does not match the colour in the panels above.



On the left hand side, we see for a block length of 20 days the two strategies (HRP versus ERC) having closer performances than HRP versus RP in the panels above. HRP is still more promising in terms of risk-adjusted performance (Sharpe and Calmar ratios) due to the lower standard deviation of returns and due to the less pronounced maximum drawdown (MDD). Also the CVaR shows a more pronounced tail on the left hand side of the distribution for ERC versus HRP. One of the problems that may lead to ERC strategies having worse performances is the reliance of the allocation on the estimation of the covariance matrix whose errors get amplified by the optimization. The results for the block bootstrap with the 60 day block length on the right hand side look less differentiated between HRP and ERC than with the lower 20 day block length on the left hand side. This mirrors our experience from the above comparison between HRP and RP.

Calmar Ratio To assess the explanatory power of the XAI, we focus on the Calmar ratio, a performance measure that can express the interests of an investor who looks for returns but is also concerned by drawdowns, i.e. cumulative returns below the recent performance top. Typically, an investor unwinds an existing exposure to a live strategy at a significant drawdown. This makes the Calmar ratio especially relevant for practitioners. Moreover, due to the path-dependency of the drawdowns, the Calmar ratio is very far from being easy to be extracted from the universe properties, making it a suitable challenge for the supervised learning model. We focus on a horse race between the Calmar ratios of the two strategies that make use of the correlation between the assets: ERC and HRP. Fig. 2 shows the density plot of the spread between the Calmar ratios of HRP and ERC across the two bootstrapped datasets. For the 20-day bootstrap, the advantage of HRP versus ERC is larger than for the 60-day bootstrap. This is consistent with the discussion above.

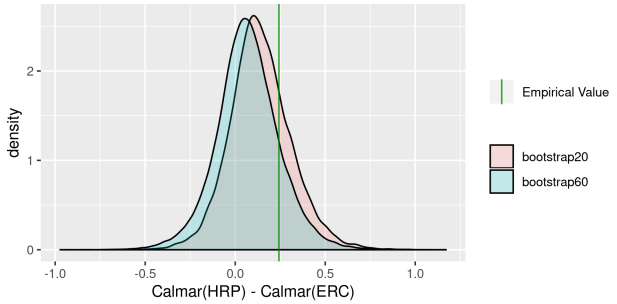


Figure 2: Density plot of the difference between the Calmar ratio of HRP and the Calmar ratio of ERC.

4 XAI

In this section, we train a supervised learning model to fit the difference between the performance of HRP and the classical ERC using the empirical dataset and the 100.000 bootstrapped datasets with a block length of 60 days. Can the model attribute the ex-post Calmar ratio spread between HRP and ERC to the statistical features of the investment universe?

Hierarchical Risk Parity has been widely considered one of the first use of AI technologies in risk-based asset management. Its strength is usually associated with the "hierarchization" of the investment universes. Here we select a set of features that can measure different aspects associated with the hierarchical structure of the time series, and other properties of the clustering method that HRP uses for the quasi-diagonalization of the correlation matrix. Moreover, we combine these features with more traditional ones, able to statistically characterize the investment universe.

The features

To characterize the portfolio universe, we select a set of classical statistical features, that any asset manager would look up before choosing a strategy, plus a set of quantities that can be measured from the portfolio universe and that can indicate properties of the hierarchical structure of the asset universe; This particular set of features is tailored to the HRP strategy, and without the help of ML it would be quite difficult to link them to the performances of the strategy.

We also look at some features that encode non-stationarity properties. Whenever the feature name has the extension `.sd` , that means that we take the standard deviation of the statistical property, measured across time. That helps to identify the heterogeneity of that property across the years.

In total, we use 24 features associated with the portfolio universe. Please see the Appendix for a complete list and a description. Here we focus on the ones that, as we will see later, the model selected as more relevant.

For example, `meanRET` identifies the mean across assets of the mean returns across time. In other words, it provides information regarding the overall trend of the returns of the full portfolio. The `meanRET.sd` instead represents how the overall trend changes across years and is measured by the standard deviation of the `meanRET`s measured year by year. Another feature is `sdRET` that measures the heterogeneity of the returns across the assets. A high value of this quantity means that the overall trend of the returns is characterized by a very heterogeneous behaviour across assets (in general features that have names starting with `sd`X have been measured with the standard deviation of X across assets). We also introduced quantities associated with the overall risk of the portfolio universe. `meanCORR` is the mean of the entries of the correlation matrix (only the lower diagonal terms) and together with `sdCORR` (their standard deviation) they provide information on the independence of the asset from the rest of the universe. For example, a negative value of `meanCORR` suggests that there is a high number of assets that are anti-correlated. A value close to zero can represent either a portfolio with independent assets or one with the same degree of positive and negative correlations. In this case, `sdCORR` would discriminate between the two possibilities. Finally, HRP bases its strategy on a clustering algorithm applied to the correlation among assets. The practitioner can wonder if the portfolio universe is or is not composed of subgroups of assets. To quantify these kinds of questions, we introduce, e.g. `CopheneticCorrelationCoefficientsingle` that measures how much a distance among clusters in the correlation are correlated with the initial correlation distance among the assets. In this case, the distance is the Euclidean distance used by the HRP algorithm. A high value of `CopheneticCorrelationCoefficientsingle` would suggest that the cluster structure well approximates the original correlation structure.

The ML learning model

For the supervised learning algorithm, we selected XGBoost (Chen and Guestrin (2016)), a gradient tree boosting library that is fast and accurate. This algorithm can construct non-linear relations among the features. Moreover, for large datasets, it can scale across GPUs to speed-up the learning process. Another benefit of using XGBoost is that it produces fast explanations, as we will see later.

To assess the stability of the explanations, the set of 100.000 bootstrapped datasets, each across 18 multi-asset futures, is split into 50% training and 50% test set. We trained the model as a regression, to learn the difference between the Calmar ratio obtained with HRP minus the Calmar ratio obtained by ERC as shown in Fig. 2. We train the model only on half of the bootstrapped datasets. A better accuracy both in the training and in the test set can be reached if we increase the number of samples. But we do not focus here on predictive accuracy. We want to show how the explanation can be used as a discovery tool. Please note that the training and test set span across the full time window of the empirical set, so they do not constitute an "out-of-sample" test in the sense of a strategy backtest.

The training leads to a root mean square error (RMSE) for the Calmar ratio spread of 0.110 in training and 0.135 in test sets. The R2 are 0.625 in the training, but only 0.410 in the test sets. The weaker R2 in the test set means the results being more relevant within the training set. Fig. 3 and 4 show frequency plots of the predicted Calmar ratio spreads against the true values in the train and test sets. Compared to the training set, the test set shows a less pronounced "cigar-shape" with more outliers and a stronger bias from the perfect diagonal.

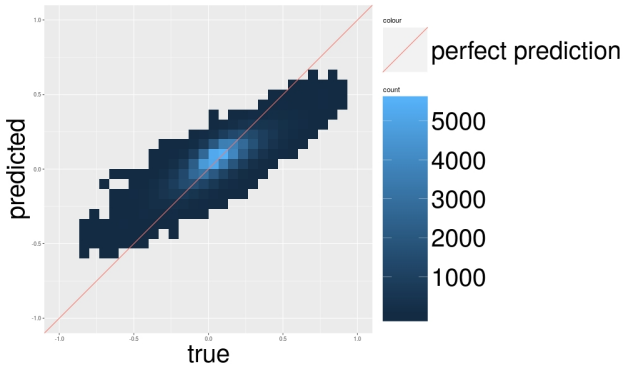


Figure 3: Frequency plot of the true and predicted ML outcomes of the training set

Computational effort

For the model learning we used an 8 CPUs machine with an NVIDIA Tesla V100 GPU.

The explanation method

The main objective of the explanation step is to explore the relations that the algorithm discovers between the statistical properties of the portfolio universe and the strategies performances within the in-sample training set. This can be achieved by looking at a set of measures that have been included into the umbrella terms of "eXplainable AI" (XAI) or "interpretable machine learning". We will focus on a particular one that revealed to be quite promising because of its generality, and comes with the name of Shapley values of feature contribution (see Lundberg and Lee (2017) and references therein).

In simple words, what Shapley values tell us is how much each feature (the statistical properties of the asset universe described above) has contributed to a specific outcome of the ML model. Because of the complexity (non-linearity) of the model, this is a non-trivial task. Shapley values is a quantity introduced in co-operative game theory to provide the fair payout to a player (the features) respect to its contribution to the common goal (ML prediction). The SHAP framework (Lundberg and Lee (2017)) provides a tool to evaluate this quantity even in a model agnostic way. It allows comparing these quantitative explanations among different models.

More formally, the explanation model $g(x')$ for the prediction $f(x)$ is constructed by an additive feature attribution method, which decomposes the prediction into a linear function of the binary variables $z' \in \{0,1\}^M$

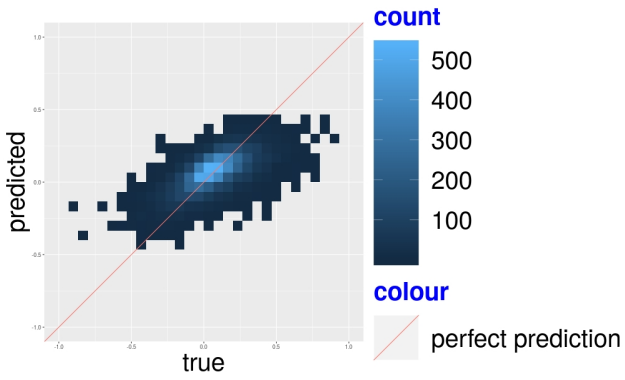


Figure 4: Frequency plot of the true and predicted ML outcomes of the test set

and the quantities $\phi_i \in \mathbb{R}$:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i. \tag{1}$$

In other terms, $g'(z') \approx f(h_x(z'))$ is a local approximation of the predictions where the local function $h_x(x') = x$ maps the simplified variables x' into x, $z' \approx x$ and M is the number of the selected input variables.

Indeed, Lundberg and Lee (2017) prove that the only additive feature attribution method that satisfies the properties of `local accuracy`, `missingness` and `consistency` is obtained attributing to each feature x'_i an effect ϕ_i called Shapley value, defined as

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{2}$$

where f is the trained model, x the vector of inputs (features), x' the vector of the M selected input features. The quantity $f_x(z') - f_x(z' \setminus i)$ is the contribution of a variable i and expresses, for each single prediction, the deviation of Shapley values from their mean.

In other words, a Shapley value represents a unique quantity able to construct an explanatory model that locally linearly approximate the original model, for a specific input x,(`local accuracy`). With the property that, whenever a feature is locally zero, the Shapley value is zero (`missingness`) and if in a second model the contribution of a feature is higher, so will be its Shapley value (`consistency`).

Another essential property of this explanatory model is that it embeds the feature space into a linear space, opening the possibility to work with statistical tests and econometrics analysis (Joseph (2019)).

Results

In our analysis, the Shapley values provide straightforward explanations. Let's look at an example. Let's recall first that our model learns the difference between the Calmar(HRP) and Calmar(ERC). Therefore, a positive outcome is associated with a better performance of HRP while a negative value to ERC. Shapley values are linear and additive quantities, therefore, for example, for a particular asset universe i, a Shapley value $\phi^{(i)}_{\text{meanRET}} = -0.02$ means that the model attributes a contribution of -0.02 to the average outcome of

the ML model in favour of ERC. Due to these properties, the absolute Shapley values can be added across all data sets to get a global variable importance that is consistent with the local Shapley values.

As an example, the following graph shows the ordered Shapley values for the empirical data set:

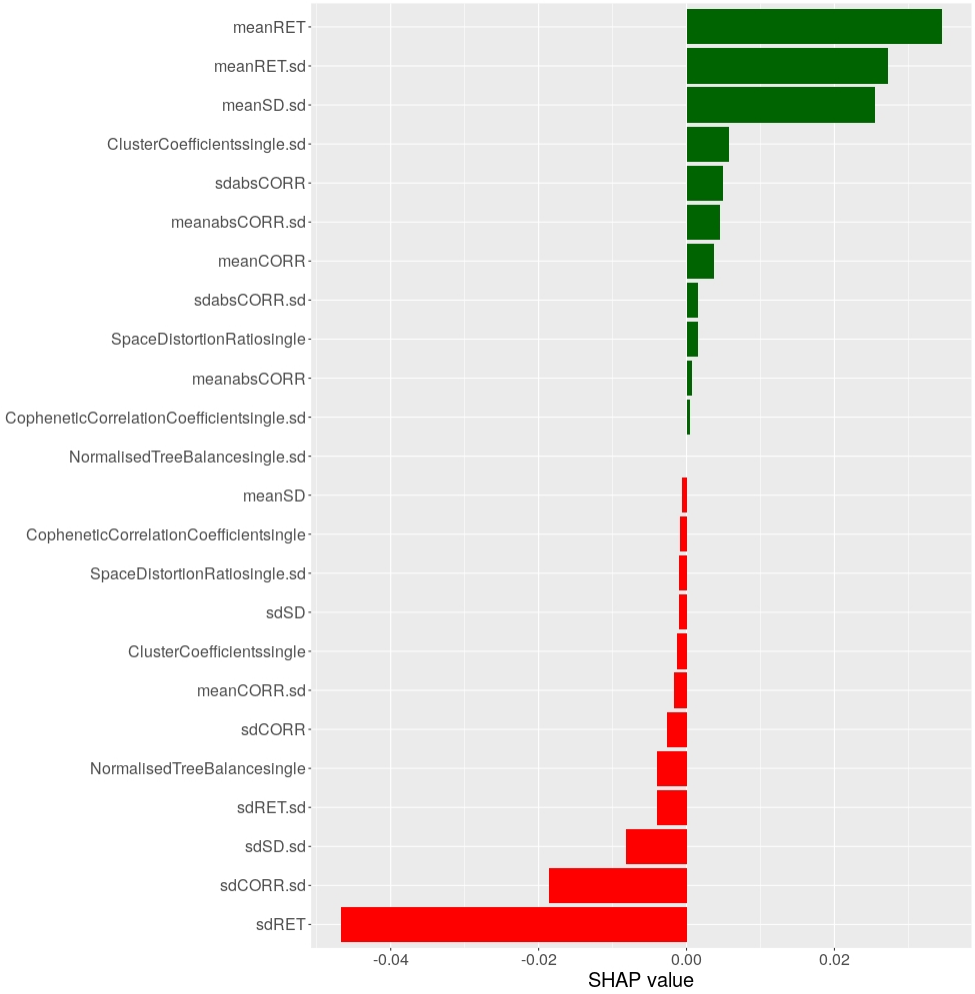


Figure 5: Shapley values for the empirical data set

We can see that for the model estimation result for the empirical dataset of Calmar(HRP) - Calmar(ERC)= 0.090 (at a true value of this spread of 0.243), the most important feature is `sdRET`, the heterogeneity of the returns across the assets, that contributed negatively (namely in favour of ERC strategy). `meanRET.sd`, the heterogeneity of the returns across the years, is, on the other hand, contributing in favour of HRP. This interplay can best be observed with a break-down plot:

To better understand the relations constructed by the model, it is fruitful to compare the Shapley values with the feature value that actually generated them. If we look at the global explanation, in our analysis, the model identified as the eight most important features: `meanRET`, `sdRET`, `meanRET.sd`, `meanSD.sd`, `meanSD`, `meanCORR`, `ClusterCoefficientssingle` and `sdSD` as one can see from the Fig. 7.

Now if we compare these main feature values with their respective Shapley value, for each point of the train set, we can find interesting patterns:

The features are sorted according to descending feature importance (Fig. 7). The most important feature is `meanRet`. Higher values of the mean asset returns lead to a lower Shapley value of `meanRet`, i.e. to a lower Calmar ratio spread of HRP versus ERC. But higher values of `sdRet`, the standard deviation of returns

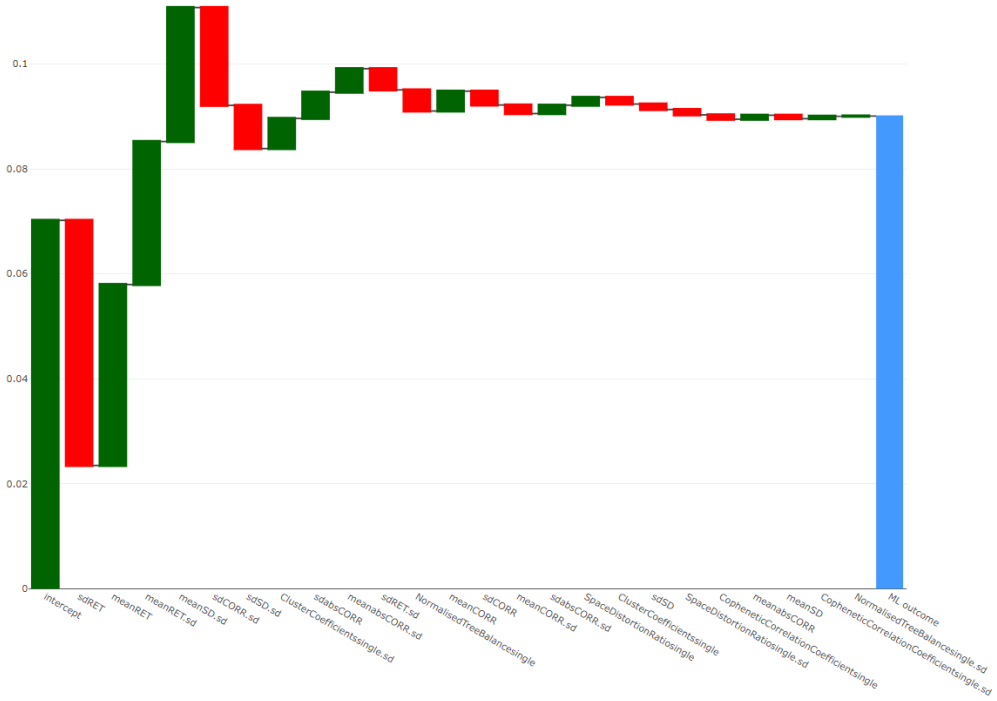


Figure 6: Shapley value breakdown: adding intercept (bias) and local Shapley values results in predicted value.

across assets, leads to a higher Shapley value for `sdRet`, i.e. to a higher Calmar ratio spread of HRP versus ERC. The impact of the variation in time of mean returns `meanRet.sd` and standard deviation `meanSD.sd` across assets on the corresponding Shapley values is consistently negative.

Lower values of `ClusterCoefficientsingle` - which measures the clustering structure - tend to penalize the Calmar ratio of HRP. Correlation matrices having a higher number of negative correlations (low `meanCORR`) get ERC to dominate. In fact, ERC risk parity products specifically rely on a negative bond-equity correlation as a statistical hedge to allow high portfolio leverage in the volatility target framework. They also seem to be sensitive to the volatility across assets (`sdSD`). On the other hand side, according to the decreasing form of the `meanSD` Shapley graph, the ERC risk parity strategies also seem to suffer less from asset volatility than the HRP strategies.

5 Conclusions and outlook

In this work, we presented a consistent pipeline able to challenge, inspect and study the behaviour of investment strategies with a complex target. As an example, we discussed the Calmar ratio spread of the Hierarchical Risk Parity (HRP) allocation method versus the Equal Risk Contribution (ERC) allocation method. Both allocation methods were applied to a multi-asset futures universe of 18 markets and a dynamical rebalancing scheme based on a 5% volatility target. The claim of HRP is to better address the hierarchical correlation structure of real markets than ERC that relies on an inversion of the covariance matrix. ERC has been scrutinized for its reliance towards a negative correlation assumption between equity and bond markets. However, adverse scenarios where this assumption breaks down did not happen often in the empirical data, so they are not easy to study.

First, in our pipeline, we make use of non-parametric bootstrapping to construct different cross-sectional market scenarios that mimic plausible and possibly problematic correlation structures. Second, we apply explainable AI (XAI) methods to discover weaknesses and implicit rules of the complex investment strategies within the bootstrapped training set. This discovery tool opens the possibility to challenge heuristic strategies

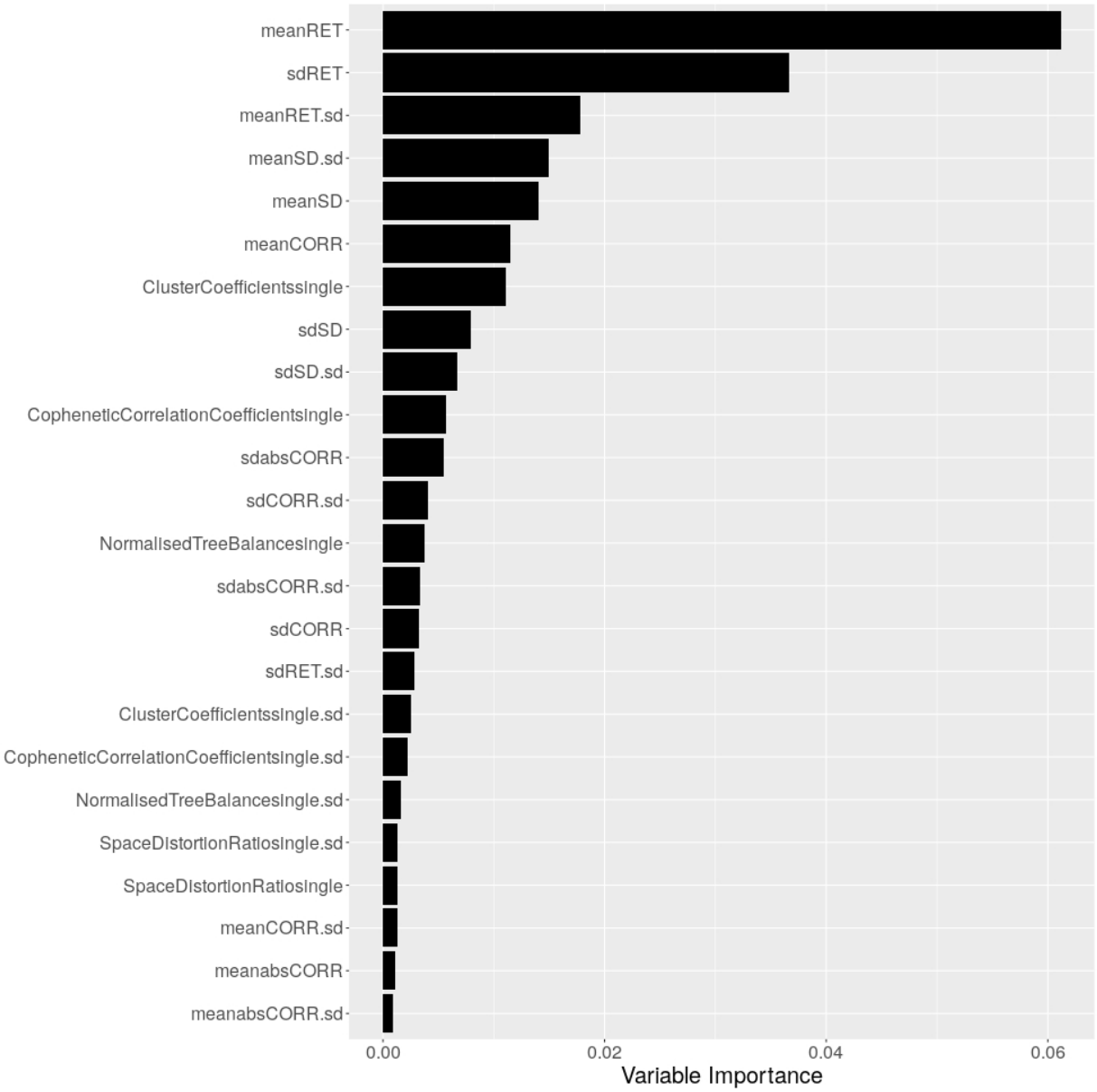


Figure 7: Features importance

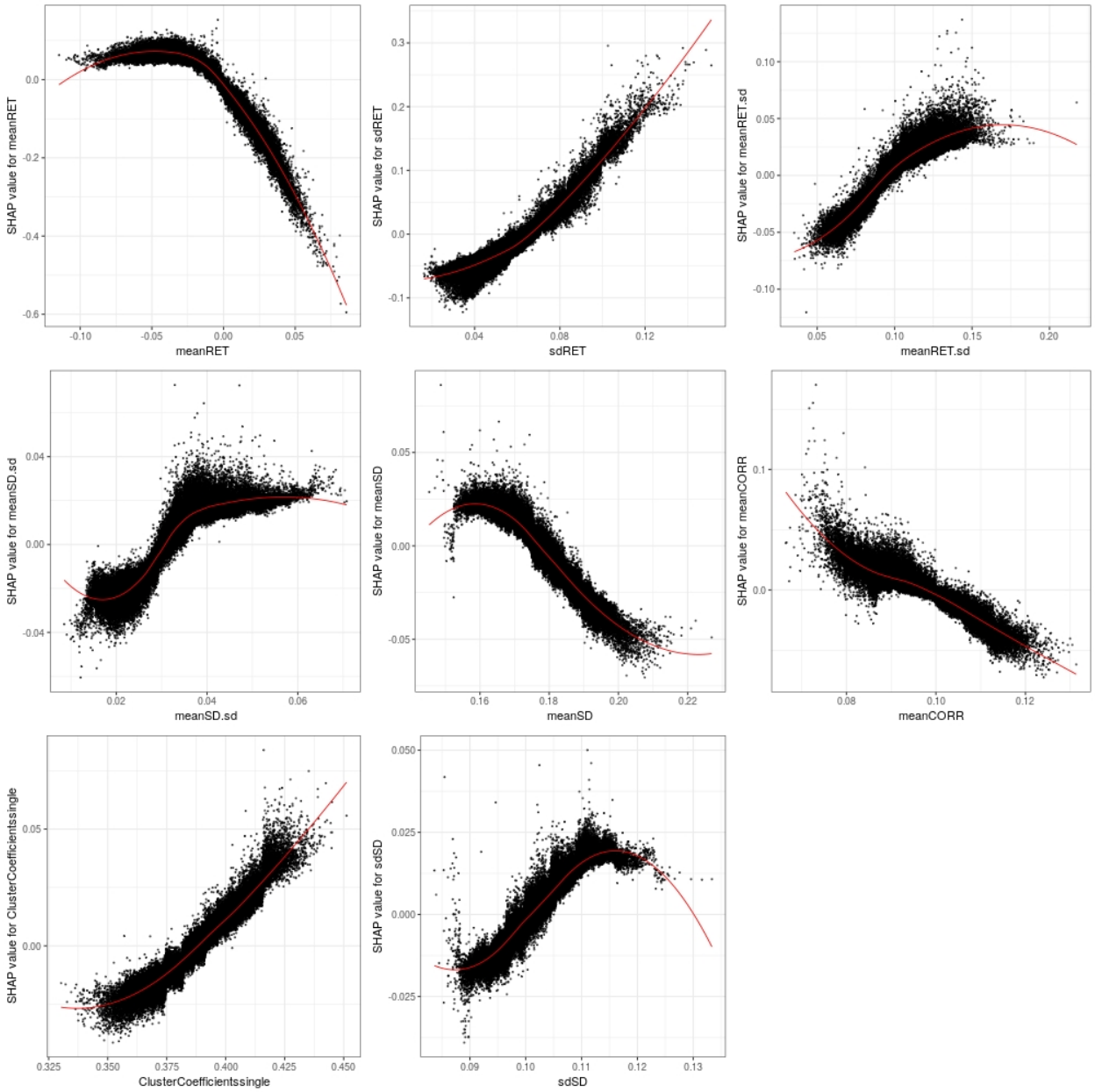


Figure 8: For the eight most important statistical properties identified by the ML model, on the y-axis the Contribution to Calmar(HRP) - Calmar(ERC) as function of the asset universe statistical property. Or in other words, the Shapley values as a function of their feature value.

and study their relations with the properties of their asset universe that otherwise would be hidden under very non-linear relationships or complex statistical dependencies. These rules that a ML model like HRP constructs internally on a specific training dataset can be explored via our setup.

For the multi-asset futures universe, we saw that HRP is more stable than Naive Risk parity and ERC. On average HRP has better compliance with the volatility target and an improved worst drawdown. XAI indeed points to the average correlation as a driver for the success of HRP over ERC strategies, but not as the main driver.

Practitioners have proposed many variations of HRP. The framework we introduced in this work would be a good testbed to challenge them, against the classical HRP strategy from López de Prado. Moreover, the analysis can be enhanced by also comparing other strategies or enriching the training dataset generating more complex simulations using AI like GAN as in (Wiese et al. (2019)) and (Marti (2019)).

In the data science life-cycle (Murdoch et al. (2019)) we can challenge the model itself with more accurate simulations. Our explainable machine is also able to show whether our dataset is a good representation of the empirical dataset, as explained in the previous sections. Of course, we do not claim to be able to predict what strategy should be applied for a certain portfolio universe for the future, as the features used in the supervised learning step are derived from the empirical sample that takes the the full time horizon into account. A "model selection" scheme would be the scope of a future study.

In the near future we plan to extend this paper using the Shapley value similarity network concept introduced in Bussmann et al. (2019) and by synthetically generated scenarios to stress-test the allocation strategies.

---

6 Acknowledgement

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement N.750961. Firamis acknowledges the NVIDIA Inception DACH program for the computational GPU resources.

7 Appendix

HRP algorithm:

In this section we report a more detailed description of the HRP strategy employied in this work. HRP is composed of three different stages: Tree clustering, quasi-diagonalization and recursive bisection.

1. Tree Clustering
   - From correlation matrix $\boldsymbol{\rho}$ with entries $\rho_{i,j} = \Sigma_{i,j}/(\sigma_i \sigma_j)$ one constructs the distance matrix D using the Gower metric $d_{i,j} = \sqrt{\frac{1}{2}(1-\rho_{i,j})}$
   - As a second step, one construct an euclidean distance between assets as $\widetilde{d}_{i,j} = \sqrt{\sum_{n=1}^{N}(d_{n,i} - d_{n,j})^2}$ (a distance of distances)
   - Reorganize the matrix to miminimize the distance between columns and construct a linkage matrix (quasi-diagonalization)
2. Recursive bisection. The matrix now ordered by the previous step. It gets split in half. A "split factor" α is associated with one of the two blocks ($1-\alpha$ for the other), and the split factor reflects the minimum variance paradigm for the blocks (namely it neglects the off diagonal blocks),

$$\alpha = 1 - \frac{\sigma^2\left(w^{(1)}\right)}{\sigma^2\left(w^{(1)}\right) + \sigma^2\left(w^{(2)}\right)}$$

while for evaluating the variance of each block it uses

$$\sigma^2\left(w^{(j)}\right) = w^{*(j)T}\Sigma^{(j)}w^{(j)} \text{ and } x^{(j)} = \frac{1/\text{diag}\left[\Sigma^{(j)}\right]}{\text{tr}\left(\text{diag}\left[\Sigma^{(j)}\right]^{-1}\right)}$$

therefore the internal weights of each block are assigned (temporarely) ignoring the off diagonal terms in the block, while the volatility uses the correlation for its estimation. The weights are just dummy variables, the final weight of each asset is provided by the series of split factors. In López de Prado words: *"takes advantage of the quasi-diagonalization bottom-up because it defines the variance of the partition... using inverse-variance weightings....takes advantage of the quasi-diagonalization top-down, because it splits the weight in inverse proportion to the cluster's variance".*

Features details

Feature Names	Descriptions
`meanCORR` `meanabsCORR` `meanSD` `meanRET`	full data sets, mean across the assets, "average levels"
`sdCORR` `sdabsCORR` `sdSD` `sdRET`	full data sets, sd across the assets, "heterogeneity across assets"
`ClusterCoefficientssingle`	specifies the agglomerative coefficient as defined in Kaufman and Rousseeuw (2009) measuring the clustering structure of the dataset; the amount of clustering structure that has been found
`NormalisedTreeBalancesingle`	Are the dendrogram trees balanced with respect of the number of leaves on the left and the right hand side of a separation? A value below 1 points to a less than perfect balance.
`CopheneticCorrelationCoefficientsingle`	correlation between the distance matrix and the ultrametric distance matrix
`SpaceDistortionRatiosingle`	How "clusterable" are the data sets using single linkage on full data sets?
`meanCORR.sd` `meanabsCORR.sd` `meanSD.sd` `meanRET.sd`	full data sets, sd across time of mean across the assets, "how do average levels change over time?"
`sdCORR.sd` `sdabsCORR.sd` `sdSD.sd` `sdRET.sd`	full data sets, sd across time of sd across the assets, "how does heterogeneity of assets change over time?"
`ClusterCoefficientssingle.sd` `NormalisedTreeBalancesingle.sd` `CopheneticCorrelationCoefficientsingle.sd`	sd across time (year by year) of the quantities `ClusterCoefficientssingle` `NormalisedTreeBalancesingle` `CopheneticCorrelationCoefficientsingle`
`SpaceDistortionRatiosingle.sd`	does the clusterability change over time?

References

Asness, Clifford S., Andrea Frazzini, and Lasse H. Pedersen. 2012. "Leverage Aversion and Risk Parity." *Financial Analysts Journal* 68 (1): 47–59. https://doi.org/10.2469/faj.v68.n1.1.

Baitinger, Eduard, and Jochen Papenbrock. 2017. "Interconnectedness Risk and Active Portfolio Management." *Journal of Investment Strategies* 6 (2): 63–90. https://doi.org/10.21314/JOIS.2017.081.

Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2019. "Explainable Ai in Credit Risk Management." *SSRN*. https://doi.org/http://dx.doi.org/10.2139/ssrn.3506274.

Carlstein, Edward, and others. 1986. "The Use of Subseries Values for Estimating the Variance of a General

Statistic from a Stationary Sequence." *The Annals of Statistics* 14 (3): 1171–9.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94. ACM.

Du, Mengnan, Ninghao Liu, and Xia Hu. 2020. "Techniques for Interpretable Machine Learning." *Communications of the ACM* 63 (1): 68–77. https://doi.org/https://dx.doi.org/10.1145/3359786.

Fengler, Matthias R, and Peter Schwendner. 2004. "Quoting Multiasset Equity Options in the Presence of Errors from Estimating Correlations." *The Journal of Derivatives* 11 (4): 43–54.

Focardi, Sergio, and Frank J Fabozzi. 2016. "Editorial Comments: Mathematics and Economics: Saving a Marriage on the Brink of Divorce?" *The Journal of Portfolio Management* 42 (July): 1–3.

"Guide to the Strategy Indices of Deutsche Börse AG." 2018. Guide Version 2.29. Deutsche Börse AG.

Hall, Peter. 1985. "Resampling a Coverage Pattern." *Stochastic Processes and Their Applications* 20 (2): 231–46.

Harvey, Campbell R, Edward Hoyle, Russell Korgaonkar, Sandy Rattray, Matthew Sargaison, and Otto Van Hemert. 2018. "The Impact of Volatility Targeting." *The Journal of Portfolio Management* 45 (1): 14–33.

Huettner, Amelie, Jan-Frederik Mai, and Stefano Mineo. 2018. "Portfolio Selection Based on Graphs: Does It Align with Markowitz-Optimal Portfolios?" *Depend. Model.* 6: 63–87.

Jacques Longerstae, and Martin Spencer. 1996. "RiskMetrics Technical Document - Fourth Edition." Technical Document. J. P. Morgan.

Joseph, Andreas. 2019. "Shapley Regressions: A Framework for Statistical Inference on Machine Learning Models." Research report 784. Bank of England.

Kaufman, Leonard, and Peter J Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. John Wiley & Sons.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Sci. Rep.* 521 (7553): 436–44.

Lohre, Harald, Carsten Rother, and Kilian Schaefer. 2020. "Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-Asset Multi-Factor Allocations." *SSRN*. https://doi.org/https://dx.doi.org/10.2139/ssrn.3513399.

Lopez de Prado, Marcos. 2016a. "Building Diversified Portfolios That Outperform Out of Sample." *The Journal of Portfolio Management* 42 (4): 59–69. https://doi.org/10.3905/jpm.2016.42.4.059.

———. 2018. *Advances in Financial Machine Learning*. Wiley.

———. 2016b. "Invited Editorial Comment: Mathematics and Economics: A Reality Check." *The Journal of Portfolio Management* 43 (October): 5–8.

López de Prado, Marcos. 2019. "Robots on Wall Street: The Impact of Ai on Capital Markets and Jobs in the Financial Services Industry." Testimony. Testimony before The U.S. House Of Representatives Committee On Financial Services - Task Force On Artificial Intelligence.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. http://arxiv.org/abs/1705.07874.

Maillard, Sebastien, Thierry Roncalli, and Jerome Teiletche. 2010. "The Properties of Equally Weighted Risk Contribution Portfolios." *The Journal of Portfolio Management* 36 (4): 60–70.

Marti, Gautier. 2019. "CorrGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks." *ArXiv E-Prints*, December. http://arxiv.org/abs/1910.09504.

Marti, G., F. Nielsen, M. Binkowski, and P. Donnat. 2017. "A review of two decades of correlations, hierarchies, networks and clustering in financial markets." *ArXiv E-Prints*, March. http://arxiv.org/abs/1703.00485.

Michaud, R. O., and R. O. Michaud. 2008. *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Financial Management Association Survey and Synthesis. Oxford University Press.

Moreira, Alan, and Tyler Muir. 2017. "Volatility-Managed Portfolios." *The Journal of Finance* 72 (4): 1611–44.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. "Definitions, Methods, and Applications in Interpretable Machine Learning." *Proceedings of the National Academy of Sciences* 116 (44): 22071–80. https://doi.org/10.1073/pnas.1900654116.

Neukirch, Thomas. 2008. "Alternative Indexing with the Msci World Index." *Available at SSRN 1106109*.

Papenbrock, Jochen. 2011. "Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization." PhD thesis, Karlsruhe. https://doi.org/10.5445/IR/1000025469.

Papenbrock, Jochen, and Peter Schwendner. 2015. "Handling Risk on/Risk Off Dynamics with Correlation Regimes and Correlation Networks." *Financial Markets and Portfolio Management* 29: 2. 125–47.

Pozzi, F., T. Di Matteo, and Tomaso Aste. 2013. "Spread of Risk Across Financial Markets: Better to Invest in the Peripheries." *Sci. Rep.* 3 (1665).

Qian, Edward. 2005. "Risk Parity Portfolios: Efficient Portfolios Through True Diversification." *Panagora Asset Management*.

Roncalli, Thierry. 2013. *Introduction to Risk Parity and Budgeting*. Edited by Chapman & Hall. CRC Financial Mathematics Series.

Wiese, Magnus, Robert Knobloch, Ralf Korn, and Peter Kretschmer. 2019. "Quant Gans: Deep Generation of Financial Time Series." http://arxiv.org/abs/1907.06673.