Interpretable Machine Learning for Diversified Portfolio Construction

Markus Jaeger **Stephan Krügel** **Dimitri Marinelli**
Jochen Papenbrock **Peter Schwendner**

Markus Jaeger
Munich Re Markets, Munich.
majaeger@munichre.com
Königinstraße 107, 80802 Munich, Germany.

Stephan Krügel
Munich Re Markets, Munich.
skruegel@munichre.com
Königinstraße 107, 80802 Munich, Germany.

Dimitri Marinelli
FinNet project - MSCA-H2020-EU and Munich Re Markets, Munich.
dmarinelli@munichre.com
Königinstraße 107, 80802 Munich, Germany.

Jochen Papenbrock
Firamis GmbH, Oberursel
jp@firamis.de
Robert-Kempner-Ring 27, 61440 Oberursel, Germany.

Peter Schwendner
Institute of Wealth & Asset Management, Zurich University of Applied Sciences, Winterthur.
scwp@zhaw.ch
Technoparkstrasse 2, 8400 Winterthur, Switzerland.

Abstract

In this paper, the authors construct a pipeline to benchmark Hierarchical Risk Parity (HRP) relative to Equal Risk Contribution (ERC) as examples of diversification strategies allocating to liquid multi-asset futures markets with dynamic leverage ("volatility target"). The authors use interpretable machine learning concepts ("explainable AI") to compare the robustness of the strategies and to back out implicit rules for decision making. The empirical dataset consists of 17 equity index, government bond and commodity futures markets across 20 years. The two strategies are backtested for the empirical dataset and for about 100'000 bootstrapped datasets. XGBoost is used to regress the Calmar ratio spread between the two strategies against features of the bootstrapped datasets. Compared to ERC, HRP shows higher Calmar ratios and better matches the volatility target. Using Shapley values, the Calmar ratio spread can be attributed especially to univariate drawdown measures of the asset classes.

Keywords: asset allocation, portfolio construction, explainable artificial intelligence, Hierarchical Risk Parity
Classification: C15, G11, G17, G0, G1, G2, G15, G24, E44

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, as the resulting portfolios successfully weathered the 2008 equity and credit drawdowns due to their high sovereign bond allocation. Five years later, funds based on the risk parity principle experienced a sharp drawdown (see, e.g. Corkery et al., *Wall Street Journal* of June 27, 2013). An explanation for this event was sudden correlated drawdowns across asset classes ("correlation breakdown") as a reaction to the "tapering" attempt of the Fed to leaver their Quantitative Easing (QE) policies. Funds that base their strategies on risk budgeting, to achieve a higher return per unit of risk, leverage their portfolios to achieve a higher portfolio return (Asness, Frazzini, and Pedersen (2012)) as they also allocate to low-risk asset classes. 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, after an adverse market movement, a dynamic leverage strategy with a "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 use 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)). Pozzi, Di Matteo, and Aste (2013), Papenbrock and Schwendner (2015), Baitinger and Papenbrock (2017) and Huettner, Mai, and Mineo (2018) discuss applications of correlation networks in portfolio management and market risk analysis. One of the first practical applications of correlation clustering in portfolio construction (Papenbrock (2011)) uses a dendrogram structure to allocate the capital to the positions in the portfolio.

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 are also discussed in Lopez de Prado (2016b) and Focardi and Fabozzi (2016). The idea is to utilize representation learning like clustering to filter relevant information in noisy data. The HRP approach uses hierarchical clustering in order to rearrange the correlation matrix into a hierarchical structure. In a second step, this information is used to allocate portfolio weights proportional to the inverse variance at each split of a recursive 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 optimize 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 and poor out-of-sample performances of allocation schemes that depend on an optimization algorithm. Especially Mean-Variance and Minimum Variance portfolios but also optimized Risk Parity strategies tend to perform poorly out-of-sample. The Markowitz' curse (Michaud and Michaud (2008)) is that optimization 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 back-office process automation and 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 contributes to fulfil regulatory compliance requirements and to foster trust to accelerate adoption. Bussmann et al. (2020) reports on the recent development and presents a practical credit risk application of XAI.

In this paper, we introduce a novel way of using XAI in asset allocation. We use the XAI explanations to investigate in hindsight how sophisticated strategies like HRP perform relative to a classical approach 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 summarize the properties of these datasets and train a machine learning model to the performance spread between HRP and ERC for datasets that reflect certain features. The selected model can adapt to non-linear relations among the features 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.

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 the covariance matrix 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 volatility target.

The dataset

In this work, we use a multi-asset investment universe of commodity, equity index and fixed income (i.e. sovereign bond) futures (Exhibit 1). Our time series of rolled over futures contracts spans the period 2000-05-03 to 2020-06-30 with daily frequency, more than twenty years that cover the dot-com crisis, the global financial crisis, the European debt crisis, the subsequent bull markets and the drawdown of the COVID pandemic. 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.

Exhibit 1: Investment universe

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 BOVESPA
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
RXA Comdty	Fixed Income	EUR	Eurex 10Y Euro-Bund
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. We restrict our analysis to the futures portfolio and do not take into account the funding or currency fluctuations of the futures variation margins. The futures portfolio is rebalanced every month and leveraged to realize 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 covariance matrix Σ, respectively Equal Risk Contribution and Hierarchical Risk Parity.

Naive Risk Parity Naive Risk Parity (RP), is here called naive because it ignores the correlation among the assets. In an RP portfolio, an asset weight is indirectly proportional to its historical volatility as explained in Roncalli (2013). More formally, the weight w_i for the i-th-asset with i spanning the portfolio universe $i = 1,\ldots,N$ is

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

where $\sigma_i = \sqrt{\Sigma_{ii}}$ denotes 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 covariance matrix to budget equally the risk among the assets. In an ERC portfolio with asset weightings w_i, 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 defined by the solution of the optimization problem

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

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 $w_i = \frac{1/\sigma_i^2}{\sum_j \left(1/\sigma_j^2\right)}$ 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 covariance 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. In this work, we restrict our analysis to the standard HRP approach, however, 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 for a given task. 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 into a matrix that describes the distance or dissimilarity of the assets. In the literature, the Gower distance is often used. Resulting distance matrices can then be further processed 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-diagonalization) 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.

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 quantities of each instrument. The estimation of realized volatility used for the updated leverage number 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)) to increase the probability the strategy will not show a higher out-of-sample volatility than the ex ante volatility target. 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 $\widetilde{w_i} = W^{\text{target}} w_i$.

We considered a 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 covariance matrix is calculated as the sample covariance matrix of the last 252 observations. Exhibit 2 describes the performance statistics we are using in this paper.

Exhibit 2: 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 shown in Exhibit 3.

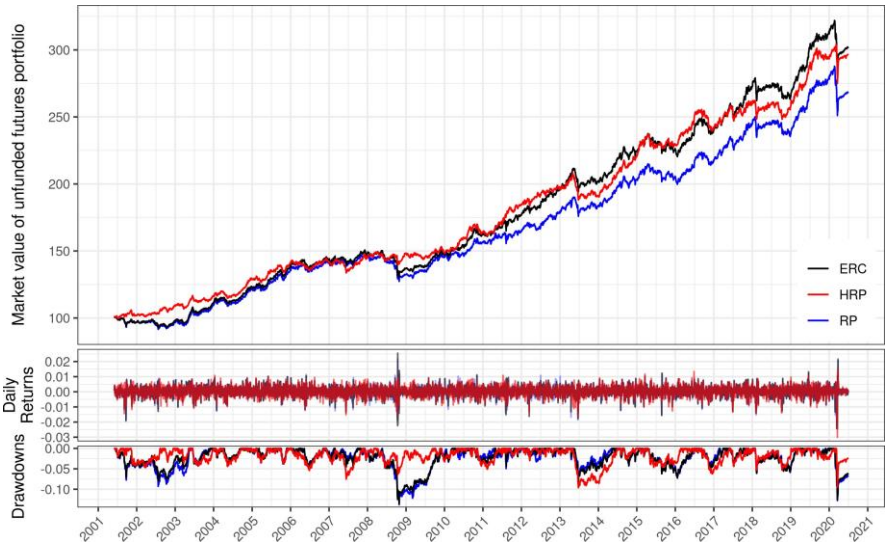
Exhibit 3: Performance results

	RP	ERC	HRP
SD	0.0522	0.0523	0.0505
RET	0.0530	0.0596	0.0585
MDD	0.1377	0.1356	0.0962
Calmar	0.3851	0.4394	0.6083
CVaR	-0.0102	-0.0107	-0.0094
SR	1.0153	1.1383	1.1578

We notice that HRP better complies with the volatility target than ERC and RP. ERC with a higher volatility 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

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. Exhibit 4 shows the performance of unfunded RP, HRP and ERC strategies applied to the empirical dataset of a multi-asset futures portfolio with a dynamic 5% volatility target, and also their daily returns and drawdowns.

Exhibit 4: Strategy performance



Robustness of the strategies

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 corresponds to 60 business days. This block length is motivated by a typical monthly or quarterly rebalancing frequency of dynamic rule-based strategies and by the empirical market dynamics that happen on this time scale. Papenbrock and Schwendner (2015) found multi-asset correlation patterns to change at a typical frequency of a few months.
- 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.

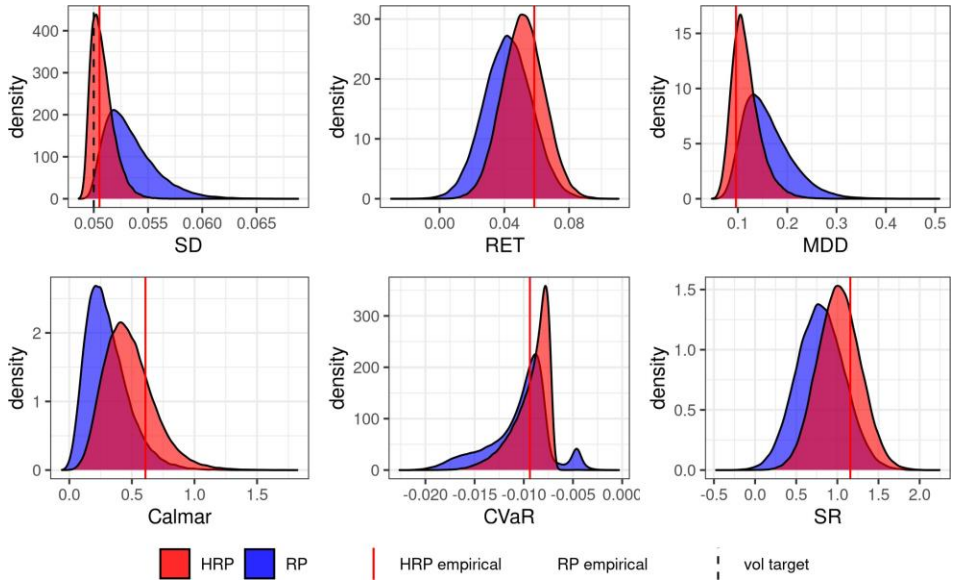
We generate 99'974 bootstrapped return time-series for each of the 17 multi-asset futures markets with a block length of 60 days. The reason for the specific number of resamplings is the organization of computing workload across CPUs.

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. Vertical lines point to the values for the empirical datasets. HRP (in red) 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, in blue). This leads to higher Sharpe and Calmar ratios for HRP compared to RP. The CVar distribution is wider for RP than for HRP. Exhibit 5 shows density plots of annualized performance statistics of HRP (red) and RP (blue) strategies applied to the block bootstrapped portfolios.

Exhibit 5: HRP and RP bootstrap results

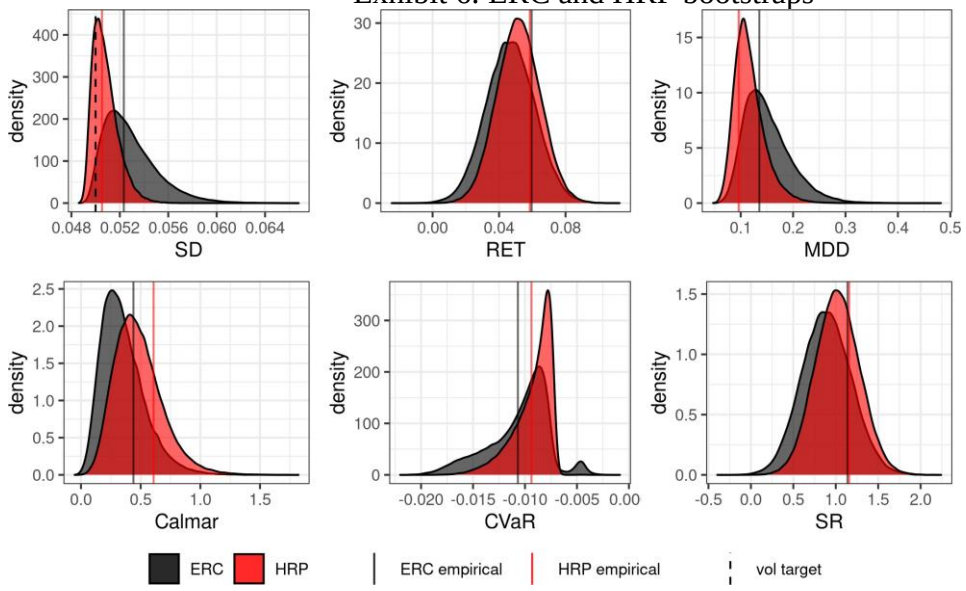


To have a benchmark against a method that also accounts for the full covariance matrix as HRP does, we consider Equal Risk Contribution (ERC). Exhibit 6 shows the performance and risk density plots for HRP (in red) versus ERC (in green).

The return densities of HRP versus ERC are closer than HRP versus RP in the panels above. HRP is still more attractive 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 prominent tail on the left-hand side of the distribution for ERC versus HRP. A reason might be the amplification of covariance estimation errors in the matrix inversion step of the ERC optimization algorithm.

Calmar Ratio To assess the explanatory power of the XAI, we focus on the Calmar ratio. The Calmar ratio is a non-linear and even path-dependent performance measure that reflects the interests of an investor who looks for returns but is also concerned by drawdowns, i.e. cumulative returns below the recent "high" of the cumulative performance.

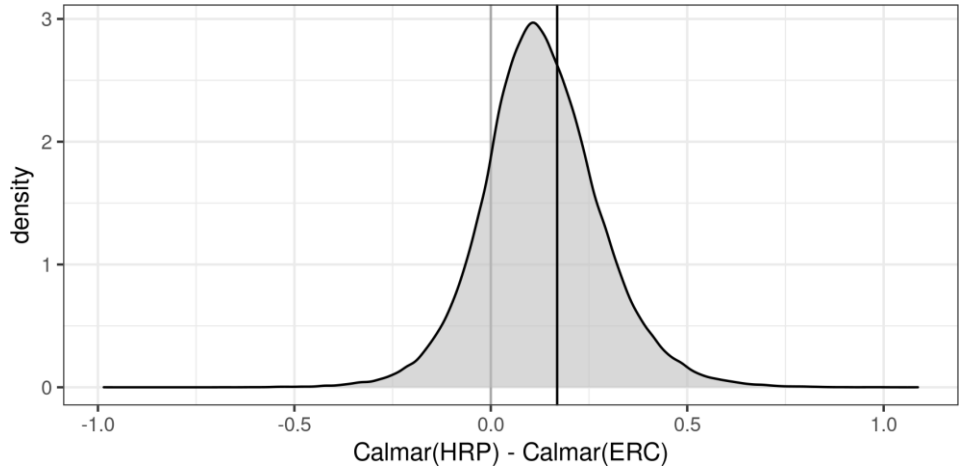
Exhibit 6: ERC and HRP bootstraps



Typically, an institutional investor subject to a stop-loss risk management rule has to unwind 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.

In the rest of the paper, we focus on a horse race between the Calmar ratios of the ERC and HRP strategies as those make use of the correlation between the assets. Exhibit 7 shows the density plot of the spread between the Calmar ratios of HRP and ERC across the bootstrapped datasets. The vertical line at a Calmar ratio spread of 0.169 marks the advantage of HRP versus ERC on the empirical dataset. The mean of the distribution is larger than zero, but clearly lower than the result from the empirical distribution.

Exhibit 7: Calmar ratio spread



Interpretable Machine Learning

In this section, we train a supervised learning model to fit the spread between the Calmar ratios of HRP and the classical ERC using statistical features of 90% of the 99'974 bootstrapped datasets. We test how well the model attributes the ex-post Calmar ratio spread to the statistical features of the other 10% of the bootstrapped datasets.

Hierarchical Risk Parity has been widely considered one of the first applications of machine learning 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 plus a set of quantities that can indicate properties of the hierarchical structure of the asset universe. This particular set of features is tailored to both strategies, and without the help of ML it would be quite difficult to link them to the performances of the strategies.

We also look at some features that encode non-stationarity properties. Whenever the feature name has the prefix sd., we measure the standard deviation of the statistical property across time. That helps to identify the heterogeneity of that property across the years. We also included measures restricted to each asset class in the portfolio.

In total, we use 96 features associated with the portfolio universe. Please see the Appendix for a detailed description.

For example, *mean_X_mean* 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 *sd.mean_X_mean* instead represents how the overall trend changes across years and is measured by the standard deviation of the *mean_X_means* measured year by year. Another feature is *mean_X_sd* 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 ending with *X_sd* have been measured with the standard deviation of X across assets). We also introduced quantities associated with the overall risk of the portfolio universe. corr_mean is the mean of the entries of the correlation matrix (only the lower diagonal terms) and together with *corr_sd* (their standard deviation) they provide information on the independence of the asset from the rest of the universe. For example, a negative value of *corr_mean* 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, *corr_sd* 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.

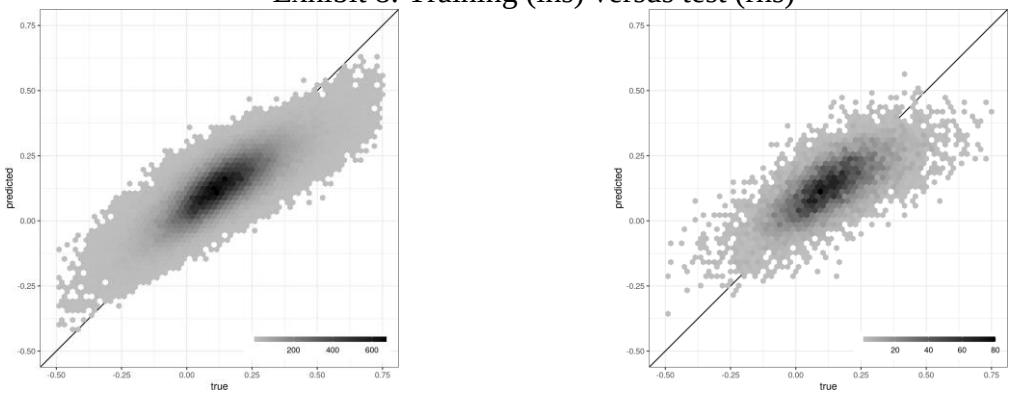
1.2 The ML learning model

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 99'974 bootstrapped datasets, each across 17 multi-asset futures, is split into 90% training and 10% 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 Exhibit 7. 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 0.0902 in training and 0.1059 in test sets. The R2 are 0.6520 in the training, and 0.5105 in the test sets. The weaker R2 in the test set means the results being more relevant within the training set. Exhibit 8 shows frequency plots of the predicted Calmar ratio spreads against the true values in the training (left) and test (right) 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.

Exhibit 8: Training (lhs) versus test (rhs)



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 relevance (see, e.g. Joseph (2019)), 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. The Shapley value is a quantity introduced in co-operative game theory to provide the fair payout to a player (the features) with 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 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 $z \in \{0,1\}^M$, with M the number of input features, and the quantities $\phi_i \in$ |R:

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

Lundberg and Lee (2017) proposed a set of desirable properties that the feature attribution method should have: *local accuracy* that connects the explanation model to the prediction we want to explain by stating that the sum of the feature attributions is equal to the prediction output, the *missingness* property assuring that missing features have no contributions and *consistency* that assures that if in a second model the contribution of a feature is higher, so will be its feature attribution. They proved that only one feature attribution has these desirable properties and is the classic Shapley value from game theory

$$\phi_i(f, x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

where N is the set of all input features. In this context, the quantity $f_x(S \cup \{i\}) - f_x(S)$ is the contribution of a feature i, where f_x is the model prediction observing the features in S with or without including i.

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)).

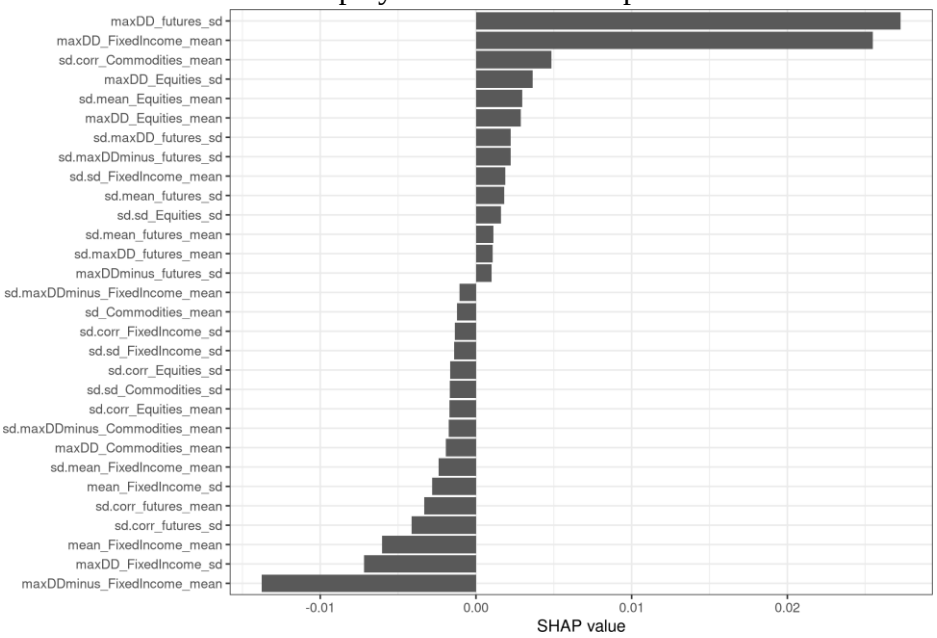
The Shapley values in equation (2) are computationally too expansive for any reasonable ML experiment, and therefore Lundberg and Lee (2017) proposed a set of efficient methods able to get reliable approximations of the explanatory model (SHAP). In particular, in this work we relied on the model tailored for tree ensembles (TreeSHAP) introduced by Lundberg et al. (2018). SHAP can be computed in an accelerated way by the approach of Mitchell, Frank, and Holmes (2020) for tree-based machine learning models like XGBoost which in turn can be accelerated by multiple CPUs and GPUs.

Results

In our analysis, the Shapley values provide insightful 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 additive quantities, therefore, for example, for a particular asset universe *i*, a Shapley value of $\phi_{\text{meanRET}}^{(i)} = -0.02$ means that the model attributes a contribution of - 0.02 to the average outcome of the ML model in favor 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, Exhibit 9 shows the ordered Shapley values for the empirical data set:

Exhibit 9: Shapley values for the empirical data set



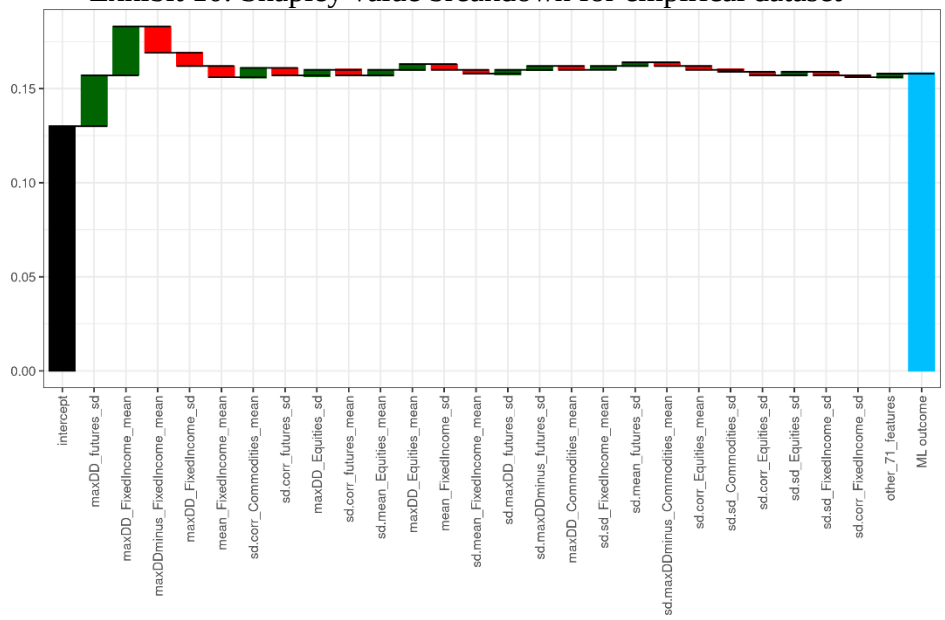
We can see that for the model estimation result for the empirical dataset of Calmar(HRP)

- Calmar(ERC)= 0.158 (at a true value of this spread of 0.169), the most important feature is maxDD_futures_sd, the heterogeneity of the drawdowns across the futures, is contributing in favour of HRP. *maxDDminus_FixedIncome_mean*, the mean drawups of fixed income futures, that, on the other hand, contributed negatively (i.e. in favour of ERC strategy). This interplay can best be observed with a break-down plot (Exhibit 10) that adds intercept (bias) and local Shapley values results in predicted value.

As ERC relies more on the negative correlation between fixed income instruments and the other two asset classes than HRP, the fixed income-related features get a high factor loading in the breakdown.

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. Exhibit 11 shows the features with the highest means of the absolute Shapley value across the training dataset. The nine most important features are *maxDD_futures_sd, mean_FixedIncome_mean, maxDD_FixedIncome_mean, maxDDminus_FixedIncome_mean, maxDD_FixedIncome_sd, maxDD_Equities_mean, sd.corr_Equities_sd, mean_FixedIncome_sd* and *sd.mean_FixedIncome_mean.*

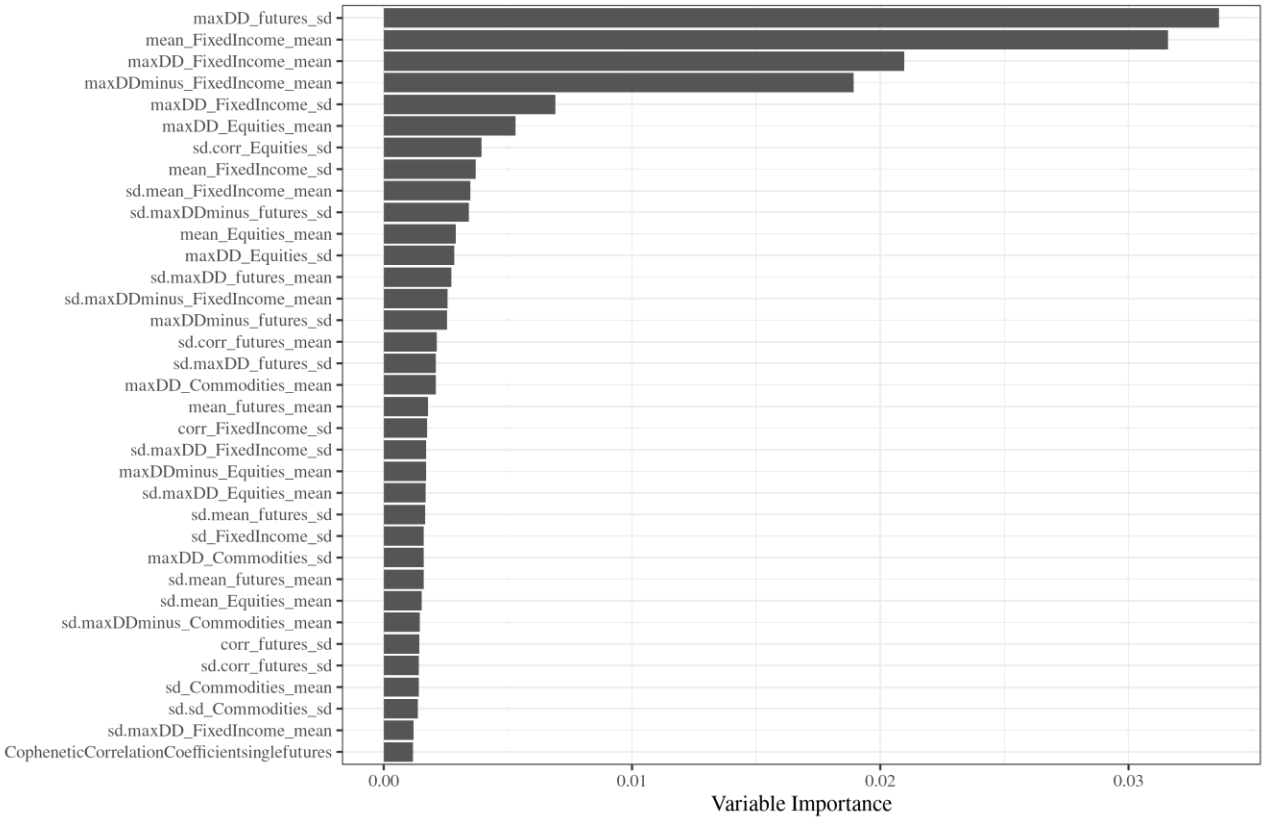
Exhibit 10: Shapley value breakdown for empirical dataset



Now if we compare these main feature values with their respective Shapley value, for each point of the training set, Exhibit 12 shows interesting patterns:

The features are sorted according to descending feature importance (Exhibit 11). The most important feature is *maxDD_futures_sd*. Higher values of the standard deviation of the max drawdowns across assets lead to a higher Shapley value of this feature, i.e. to a higher Calmar ratio spread of HRP versus ERC. The sensitivity of the Shapley value for *maxDD_FixedIncome_mean* is antisymmetric to *maxDDminus_FixedIncome_mean*: at higher fixed income drawdowns, the advantage of HRP decreases, but at higher "drawup", it increases. The *maxDD_FixedIncome_mean* sensitivity is consistent to *maxDDminus_FixedIncome_mean*: it confirms the advantage of HRP versus ERC at higher fixed income returns. The first feature related to the cross-asset correlation is sd.corr_Equities_sd, in position 7. The feature measures the annual variability of the standard deviation of the cross-asset correlation parameters of the Equity futures.

Exhibit 11: Global feature importance



For the nine most important statistical properties identified by the ML model, Exhibit 12 shows on the y-axis the Contribution to Calmar(HRP) - Calmar(ERC) as a function of the statistical property. Or in other words, the Shapley values as a function of their feature value. Each point corresponds to one bootstrap sample within the training dataset. The green points reflect the empirical dataset.

The plot in the 7th panel of Exhibit 12 does not provide an explicit interpretation. However, once one considers its interplay with another feature, *maxDD_futures_sd*, whose value is encoded by the colour in Exhibit 13, the explanations reveal that the model considers the contribution of *sd.corr_Equities_sd* differently for different values of the concurrent statistical property maxDD_futures_sd: If the differences in correlation values are stable in time, and there are great differences between asset drawdowns, the contribution of the feature is in favor of HRP. The opposite happens when the differences in correlation parameters are more variable among the years. We do not have to forget that *maxDD_futures_sd* can have a much higher contribution to the model outcome (see the first panel in Exhibit 12) and that we are considering contributions much smaller as the RMSE of the data.

The features measuring the return correlations and the features *ClusterCoefficientsingle* and *CopheneticCorrelationCoefficientsingle* measuring the clustering structure do not show up in a prominent position of the ranked feature importance (Exhibit 11).

Exhibit 12: Shapley values as a function of the feature values

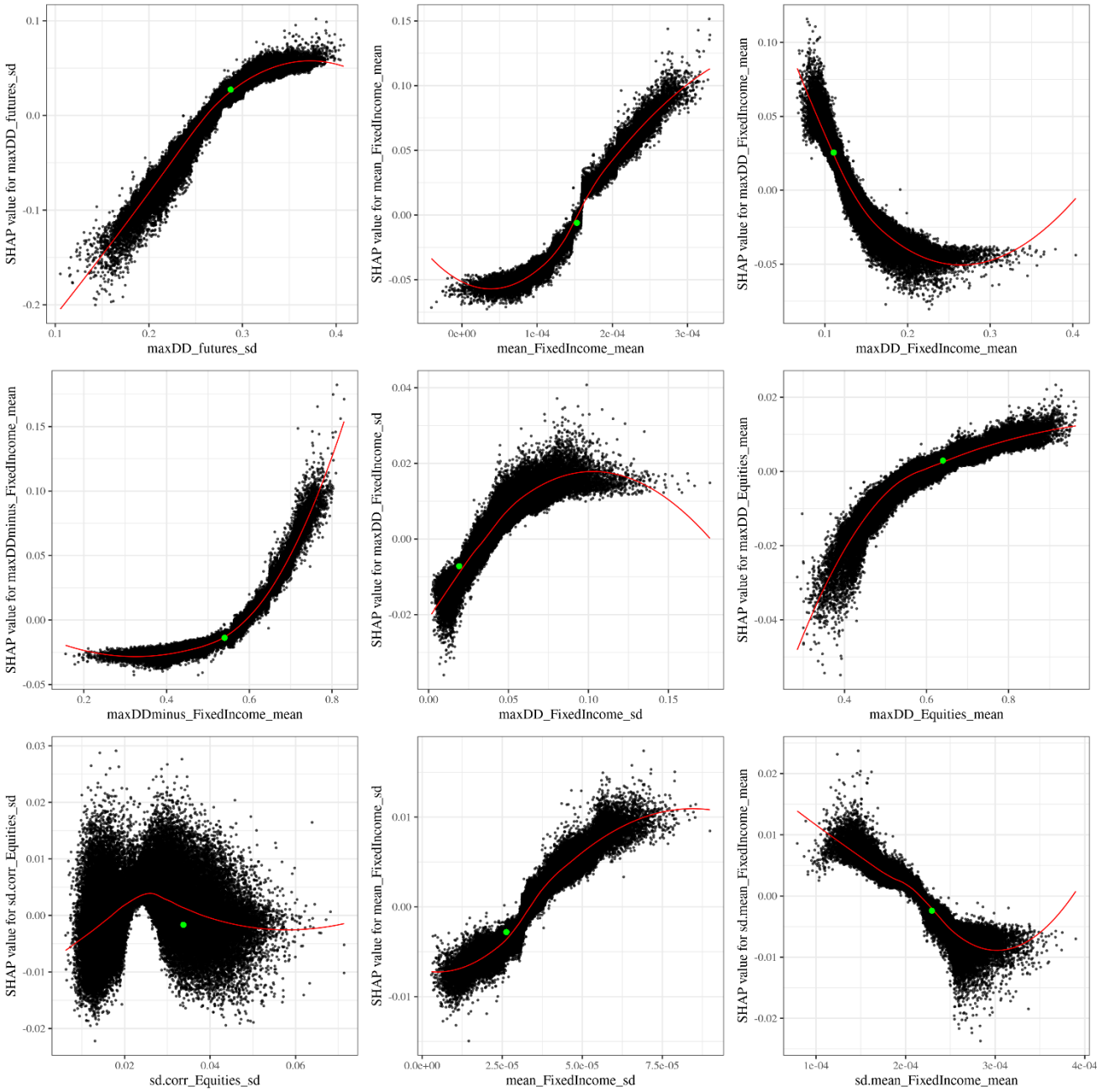
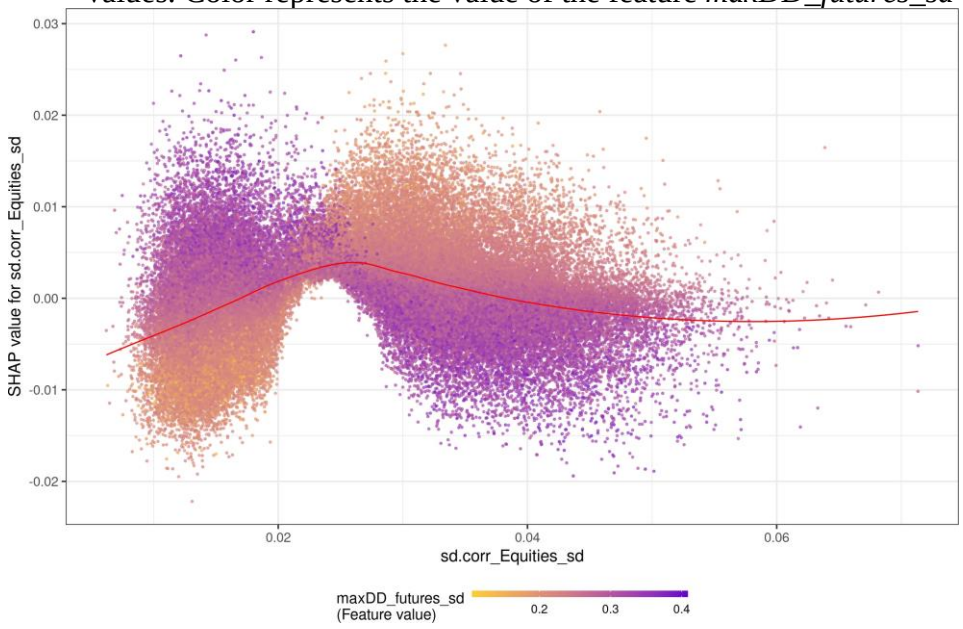


Exhibit 13: Contribution to Calmar(HRP) - Calmar(ERC) as function of the *sd.corr_Equities_sd* values. Color represents the value of the feature *maxDD_futures_sd*



Conclusions and outlook

In this work, we presented a consistent pipeline able to challenge, inspect and study the behavior 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 17 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 and study their relations with the properties of their asset universe that otherwise would be hidden under non-linear relationships or complex statistical dependencies. Our approach can explore the implicit rules HRP and similar ML models internally construct on a specific training dataset.

For the multi-asset futures universe, we saw that HRP is more robust than Naive Risk parity and ERC. On average HRP has better compliance with the volatility target and an improved worst drawdown. However, XAI points to the univariate but path-dependent drawdown measures as drivers for the success of HRP over ERC strategies.

Practitioners have proposed many variations of HRP. The framework we introduced in this work would be a suitable 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), Marti (2019) or using the "matrix evolutions" scheme of Papenbrock et al. (2020).

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 full time horizon into account. A "model selection" scheme would be the scope of a future study.

Shortly we plan to extend this work using the Shapley value similarity network concept introduced in Bussmann et al. (2020) and by synthetically generated scenarios to stress-test the allocation strategies.

Acknowledgement

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement N.750961. The implementation was sponsored by Munich Re Markets. We appreciate the infrastructure by Open Telekom Cloud and the NVIDIA GPU resources provided for this research. The work was also supported by the European Union's Horizon 2020 research and innovation program FIN-TECH: A Financial supervision and Technology compliance training programme" under the grant agreement No 825215 (Topic: ICT-35-2018, Type of action: CSA).

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 $\rho_{i,j} = \Sigma_{i,j}/(\sigma_i\sigma_j)$, construct the distance matrix D using the Gower metric $d_{i,j} = \sqrt{\frac{1}{2}(1-\rho_{i,j})}$.
   - As a second step, construct an Euclidean distance between assets as $\widetilde{d_{i,J}} = \sqrt{\sum_{n=1}^{N}(d_{n,i} - d_{n,j})^2}$.
   - Reorganize the matrix to minimize the distance between columns and construct a linkage matrix (quasi-diagonalization).
2. Recursive bisection. The matrix is now ordered by the previous step. It gets split in half. A "split factor" $\alpha = 1 - \frac{\sigma^2(w^{(1)})}{\sigma^2(w^{(1)})+\sigma^2(w^{(2)})}$ is associated with one of the two blocks, and 1-α for the other. The split factor reflects the minimum variance paradigm for the blocks (namely it neglects the off diagonal blocks). For evaluating the variance of each block, the scheme uses $\sigma^2(w^{(j)}) = w^{*(j)T}\Sigma^{(j)}w^{(j)}$ and $x^{(j)} = \frac{1/\mathrm{diag}[\Sigma^{(j)}]}{\mathrm{tr}\left(\mathrm{diag}[\Sigma^{(j)}]^{-1}\right)}$, so the internal weights of each block are assigned (temporarily) 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

We have introduced 96 features describing the statistical properties of the multivariate time-series. The features can be reconstructed from their name as follows:

$$\underbrace{[(sd.)]}_{\text{over time}} \quad \underbrace{maxDD}_{\text{statistical measure}} \quad \underbrace{(\_futures)}_{\text{asset class}} \quad \underbrace{\_mean}_{\text{aggregated}}$$

statistical measure
mean mean
sd standard deviation
corr correlation coefficients
maxxDD maximum drawdown
maxDDminus maxxDD of minus log-returns

asset class
futures
FixedIncome
Commodities
Equities

aggregated
_mean
_sd

The statistical measure is applied to each asset in the class defined in "asset class" (where futures stands for all the assets, regardless of their asset class). The correlation coefficients are instead the upper triangular part of the correlation matrix. *maxDDminus* refers to the opposite of a "drawdown", i.e. a "drawup" from the previous all-time-low or the "trough-to-peak-performance".

The quantities are then aggregated across assets to a scalar value by taking their mean or standard deviation (aggregated). This construction leads to $5 \times 4 \times 2 = 40$ different quantities. Moreover, we consider two additional statistical measures associated with "clusterability" of the correlation matrix:

statistical measure	
ClusterCoefficientssingle	specifies the agglomerative coefficient as defined in Kaufman and Rousseeuw (2009) measuring the clustering structure of the dataset
CopheneticCorrelationCoefficientsingle	correlation between the distance matrix and the ultrametric distance matrix

that are calculated for all the assets and for all the assets restricted per asset class. This results in $40 + 8 = 48$ combinations so far.

If the feature name starts with *sd.*, the scalar values are evaluated for each year, and the measured feature reports their standard deviation across the years. These quantities, therefore, identify the variability of the statistical measures over time. In addition to the features evaluated for the entire time series, it results in a total of $48 + 48 = 96$ features.

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. 2020. "Explainable AI in Credit Risk Management." *Comput Econ.* https://doi.org/10.1007/s10614-020-10042-0.

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.

Corkery, Michael, Cui, Carolyn, and Kirsten Grind. "Fashionable' Risk Parity' Funds Hit Hard." Wall Street Journal, 27.6.2013. https://www.wsj.com/articles/SB10001424127887323689204578572050047323638

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 Multi-asset 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 Introduc- tion 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 M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." *arXiv preprint* arXiv:1802.03888 (2018).

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.

Mitchell, Rory, Eibe Frank, and Geoffrey Holmes. 2020. *GPUTreeShap: Fast Parallel Tree Interpretability*. http://arxiv.org/abs/2010.13972

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.

Papenbrock, Jochen, Peter Schwendner, Markus Jaeger, and Stephan Krügel. 2020. "Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios." *Journal of Financial Data Science*, under review.

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.