Interpretable Machine Learning for Diversified Portfolio Construction

Markus Jaeger, Stephan Krügel, Dimitri Marinelli, Jochen Papenbrock, and Peter Schwendner

Markus Jaeger
is a senior investment expert at Munich Re Markets in Munich, Germany.
majaeger@munichre.com

Stephan Krügel
is a senior investment expert at Munich Re Markets in Munich, Germany.
skruegel@munichre.com

Dimitri Marinelli
is a financial data scientist and Marie Skłodowska-Curie Fellow at Munich Re Markets in Munich, Germany.
dmarinelli@munichre.com

Jochen Papenbrock
is the CEO of Firamis GmbH in Oberursel, Germany.
jp@firamis.de

Peter Schwendner
is the director of the Institute of Wealth & Asset Management at Zurich University of Applied Sciences in Winterthur, Switzerland.
scwp@zhaw.ch

KEY FINDINGS

- The authors introduce a procedure to benchmark rule-based investment strategies and to explain the differences in path-dependent risk-adjusted performance measures using interpretable machine learning.
- They apply the procedure to the Calmar ratio spread between hierarchical risk parity (HRP) and equal risk contribution (ERC) allocations of a multi-asset futures portfolio and find HRP to have superior risk-adjusted performance.
- The authors regress the Calmar ratio spread against statistical features of bootstrapped futures return datasets using XGBoost and apply the SHAP framework by Lundberg and Lee (2017) to discuss the local and global feature importance.

ABSTRACT

In this article, 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 back tested 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.

TOPICS

Quantitative methods, statistical methods, big data/machine learning, portfolio construction, performance measurement*

*All articles are now categorized by topics and subtopics. **View at PM-Research.com**.

After the financial crisis of 2008, interest among both practitioners and academics grew for allocations that equally budget the risk for the assets in a portfolio, the so-called *risk parity* (RP) allocation strategies, because the resulting portfolios successfully weathered the 2008 equity and credit drawdowns owing to their high

sovereign bond allocation (Denis et al. 2011). Five years later, funds based on the RP principle experienced a sharp drawdown (see, e.g., Corkery, Cui, and Grind 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 lever its quantitative easing 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) because they also allocate to low-risk asset classes. This practice leads to better returns, higher Sharpe ratios in some cases (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 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 by Marti et al. 2017). Pozzi, Di Matteo, and Aste (2013), Papenbrock and Schwendner (2015), Baitinger and Papenbrock (2017), and Huettner, Mai, and Mineo (2018) discussed 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) used a dendrogram structure to allocate 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 (ML), whose benefits are also discussed by Lopez de Prado (2016b) and Focardi and Fabozzi (2016). The idea is to use representation learning such as clustering to filter relevant information in noisy data. The HRP approach uses hierarchical clustering 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 that is less prone to noise—estimating and inverting covariance matrixes to optimize a risk measure often leads to errors of such magnitude as to 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 performance of allocation schemes that depend on an optimization algorithm. Mean–variance and minimum variance portfolios in particular, but also optimized RP 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 a diversified portfolio.

ML is becoming increasingly important in the financial industry; see, for example, Lopez de Prado (2018, 2019). In many decision-making applications, however, regulatory and transparency concerns slowed the industry in embracing these new technologies, despite their massive success in back-office process automation and other domains such as computer vision (see, e.g., LeCun, Bengio, and Hinton 2015).

One approach to overcome this limitation is explainable AI (XAI) (Murdoch et al. 2019; Du, Liu, and Hu 2020). XAI not only delivers the desired quantitative result but also reports its reasoning to make its functioning easier to understand by humans. The ability to explain model decisions to stakeholders contributes to fulfilling regulatory compliance requirements and fostering trust to accelerate adoption. The underlying concept of Shapley values (Shapley 1953) has also previously been applied to finance, as documented by Mussard and Terraza (2008) and Simonian (2012, 2014, 2019). Bussmann et al. (2020) reported on the recent development and presented a practical credit risk application of XAI.

In this article, we introduce a novel way of using XAI in asset allocation. We use the XAI explanations to investigate in hindsight how sophisticated strategies such as HRP perform relative to a classical approach such as equal risk contribution (ERC). 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 an ML model on the performance spread between HRP and ERC for datasets that reflect certain features. The selected model can adapt to nonlinear relations among the features of these artificial datasets. Finally, we use XAI explanations to describe the relationships discovered by the model and make the strategies more transparent for a financial market practitioner.

The original contribution of this article is, first, a method to link the statistical properties of an investment universe to the outperformance of risk-based investment strategies. Risk-based strategies use only 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 the importance metrics of features (the statistical properties) that are directly linked to outperformance probabilities. Second, the underlying dataset 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 process to conduct 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 May 3, 2000 to June 30, 2020, with daily frequency—more than 20 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-19 pandemic. We use listed futures as instruments because this is the most cost-efficient way of obtaining a global cross-asset allocation. Furthermore, futures as unfunded derivatives enable dynamic leverage approaches such as a volatility target concept, in contrast to fully funded instruments such as exchange-traded funds.

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 naïve RP) and two that use the full and filtered information of the covariance matrix Σ (ERC and HRP, respectively).

Naïve RP

Naive RP is here called *naive* because it ignores the correlation among the assets. In an RP portfolio, an asset's weight is indirectly proportional to its historical volatility,

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

as explained by Roncalli (2013). More formally, the weight w_i for the ith 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.

ERC

ERC portfolios (Qian 2005; Neukirch 2008; Maillard, Roncalli, and Teiletche 2010) use the full information in the covariance matrix to budget the risk among the assets equally. In an ERC portfolio with asset weights w_i, the percentage volatility risk contribution of the ith 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 $w_i = \frac{1/\sigma_i^2}{\sum_j (1/\sigma_j^2)}$ 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 correlation matrix. After the quasi-diagonalization is carried out, a recursive bisectioning method is applied to the covariance matrix 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 sequentially. Each block can be replaced by appropriate methods that 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 the literature use the Pearson correlation coefficient matrix, but obviously there can be more robust and nonlinear 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 matrixes can then be further processed by using the distance of distance approach by Lopez de Prado.

The third step is the choice of the hierarchical clustering procedure (HRP uses single-linkage clustering). More generally speaking, hierarchical clustering is used to reorder the correlation matrix (quasi-diagonalization) for later processing with a bisectioning method, and this rearrangement can be done in numerous alternative ways. Alternatives to single-linkage clustering include 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 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, e.g., Deutsche Börse AG 2018) to increase the probability that the strategy will not show 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 bps (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 use in this article.

EXHIBIT 2

Performance Statistics

Statistics	Short	Description
Volatility	SD	Annualized volatility
Returns	RET	Annualized portfolio excess returns
Maximum Drawdown	MDD	Most severe drawdown as percentage of previous highest market value
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 Returns and max drawdown

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

Results for the Empirical Dataset

For the futures universe, the strategies performed as shown in Exhibit 3.

We note that HRP complies with the volatility target better than ERC and RP. ERC with a higher volatility reaches also higher returns but has a lower Sharpe ratio. HRP dominates in terms of Calmar ratio, which takes the maximum drawdown into account. The drawdowns often determine whether a buy-side investor can keep an investment or will have to unwind and thus 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, as well as their daily returns and drawdowns.

ROBUSTNESS OF THE STRATEGIES—BOOTSTRAPPED DATASET

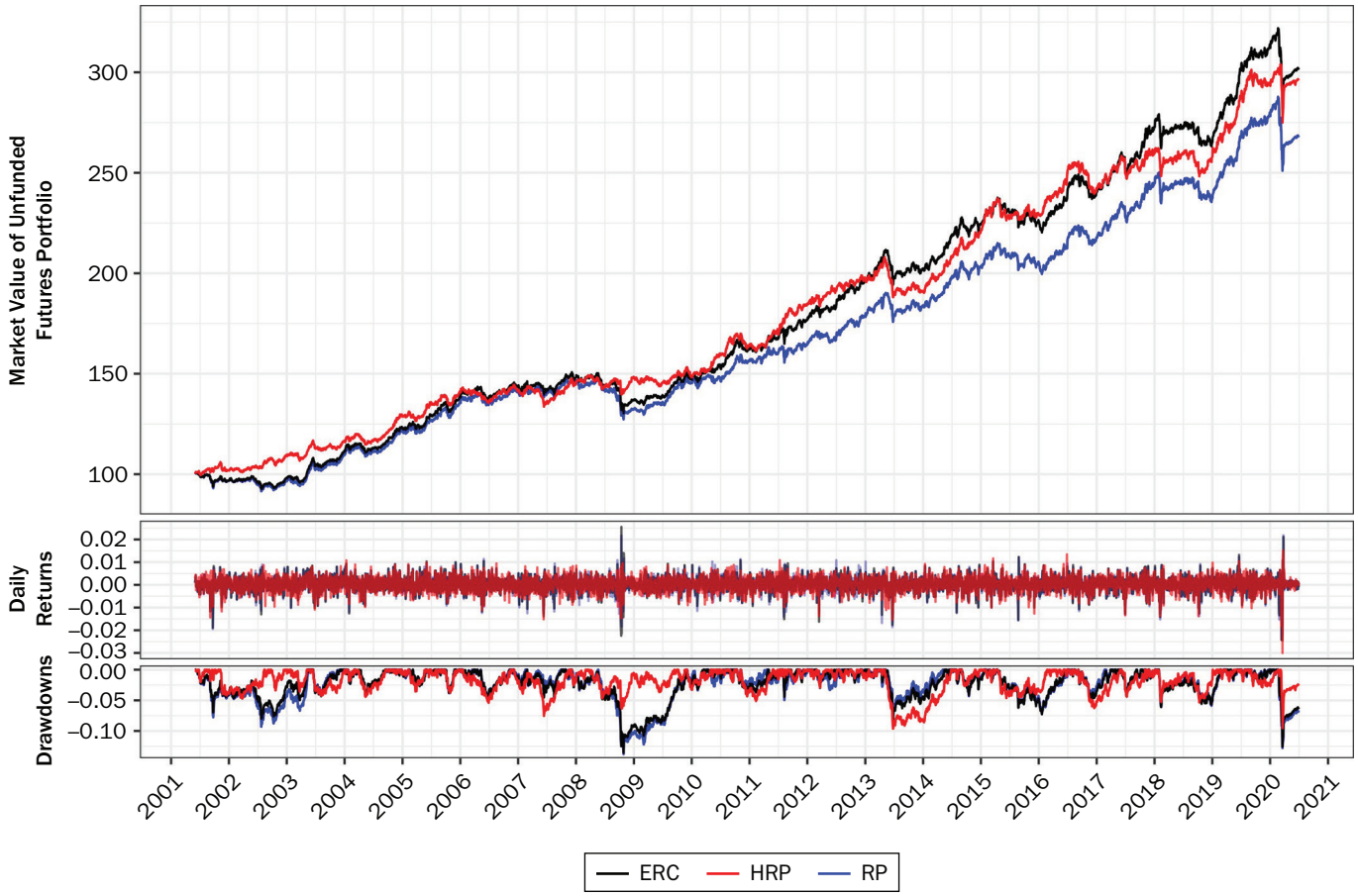
To account for the nonstationarity of futures return time series, we generate an additional dataset of time series by block bootstrapping (Hall 1985; Carlstein 1986; Fengler and Schwendner 2004; Lohre, Rother, and Schaefer 2020):

- 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 choice of block length is motivated by the 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 as 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

EXHIBIT 4

Strategy Performance



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

Backtest Results for the Bootstrapped Datasets

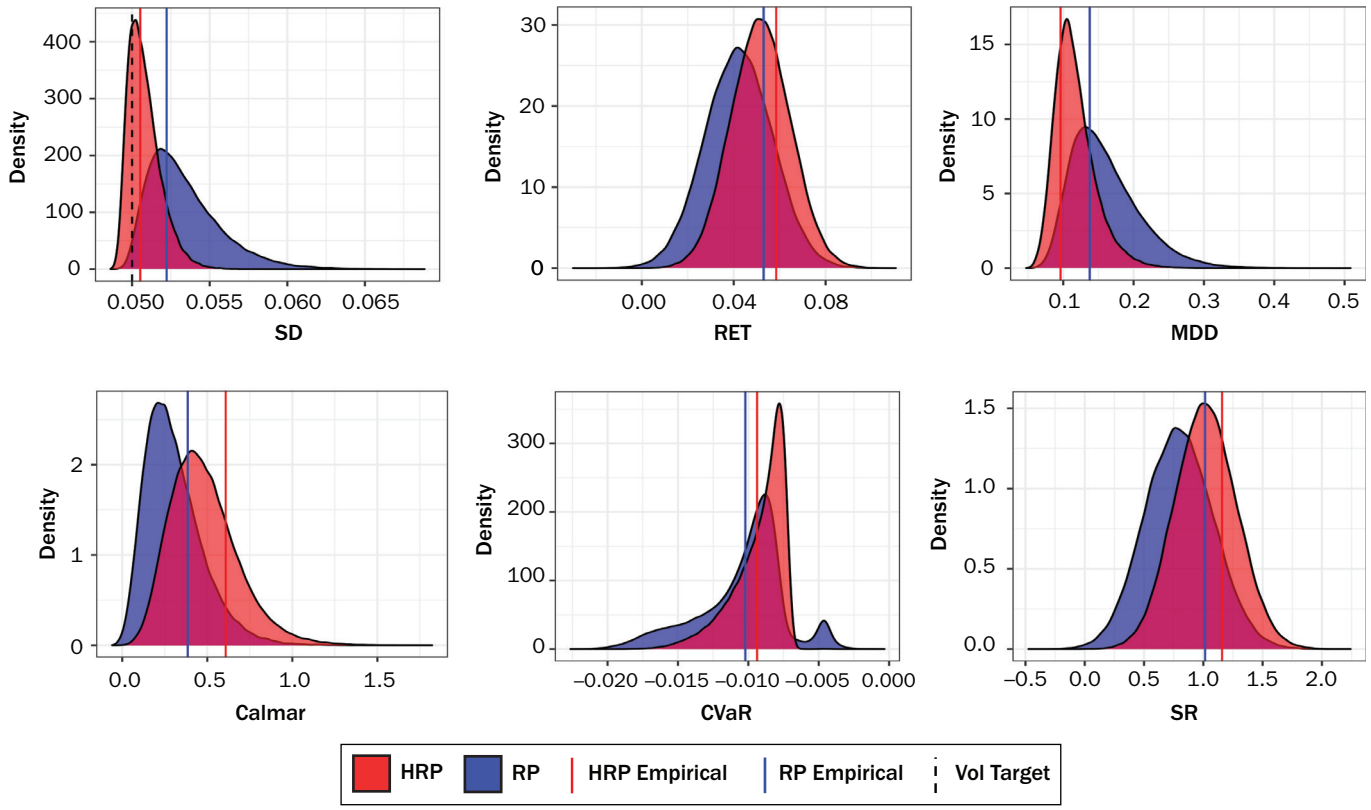
We display our results in the form of panels, with densities for various performance and risk measures across the bootstrapped datasets to compare (Exhibit 5) HRP versus RP and (Exhibit 6) HRP versus ERC. Vertical lines point to the values for the empirical datasets. HRP (in red) delivers lower standard deviations (SDs) of returns and better compliance with the 5% volatility target, higher returns (RET), and less-pronounced maximum drawdowns (MDDs) than naïve RP (in blue). This leads to higher Sharpe and Calmar ratios for HRP compared to RP. The conditional value-at-risk (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.

To have a benchmark against a method that also accounts for the full covariance matrix, as HRP does, we consider 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 those of HRP versus RP in Exhibit 5. HRP is still more attractive in terms of risk-adjusted performance

EXHIBIT 5

HRP and RP Bootstrap Results



(Sharpe and Calmar ratios) due to the lower SD of returns and the less pronounced MDD. Furthermore, the CVaR shows a more prominent tail on the left-hand side of the distribution for ERC versus HRP. A reason for this might be the amplification of covariance estimation errors in the ERC optimization step.

Calmar Ratio

To assess the explanatory power of the XAI, we focus on the Calmar ratio. The Calmar ratio is a nonlinear 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).

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 extract from the universe properties, making it a suitable challenge for the supervised learning model.

In the rest of the article, we focus on a horse race between the Calmar ratios of the ERC and HRP strategies because they 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.

EXHIBIT 6

ERC and HRP Bootstraps

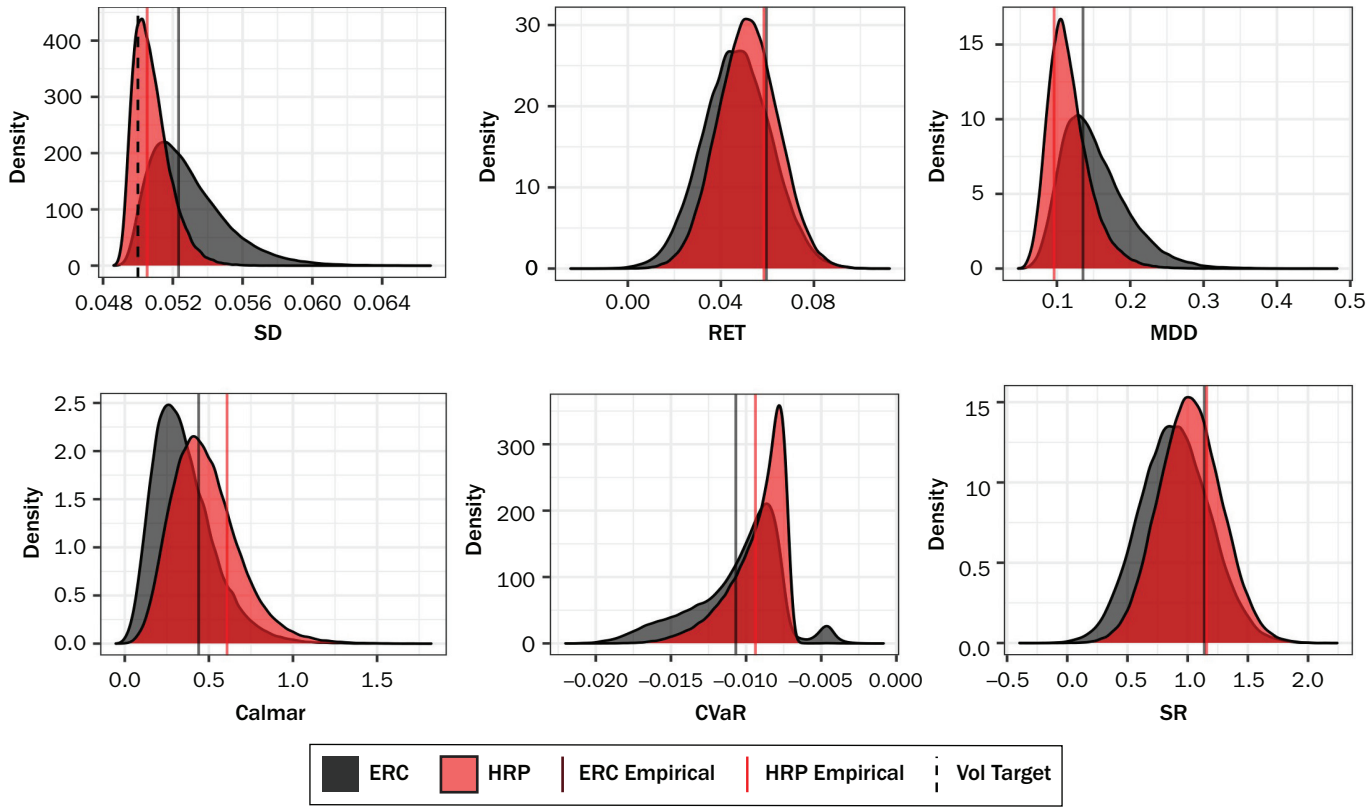
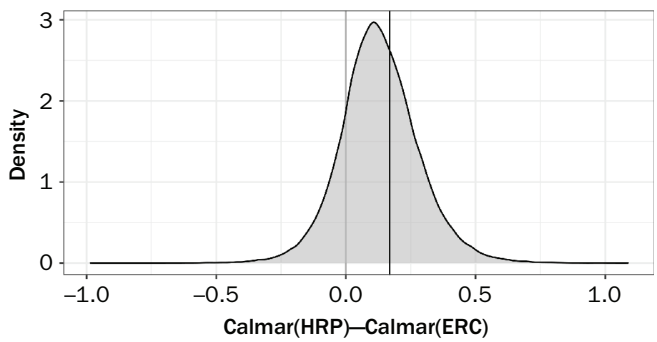


EXHIBIT 7

Calmar Ratio Spread



The mean of the distribution is larger than zero but clearly lower than the result from the empirical distribution.

INTERPRETABLE ML

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.

HRP has been widely considered one of the first applications of ML 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 that can 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 performance of the strategies.

We also look at some features that encode nonstationarity properties. Whenever the feature name has the prefix *sd.*, we measure the SD of the statistical property across time. This helps to identify the heterogeneity of that property across the years. We also include 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 SD of the *mean_X_mean* measured year by year. Another feature is *mean_X_sd*, which measures the heterogeneity of the returns across the assets; a high value means the overall trend of the returns is characterized by very heterogeneous behavior across assets (in general, features that have names ending with *X_sd* have been measured with the SD of X across assets).

We also introduced quantities associated with the overall risk of the portfolio universe. For example, *corr_mean* is the mean of the entries of the correlation matrix (only the lower diagonal terms), and together with *corr_sd* (their SD) 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 a high number of assets 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 may wonder whether the portfolio universe is or is not composed of subgroups of assets. To quantify these kinds of questions, we introduce, for example, *CopheneticCorrelationCoefficientsingle*, which measures how much the distance among clusters in the correlation is 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.

Training the ML Model

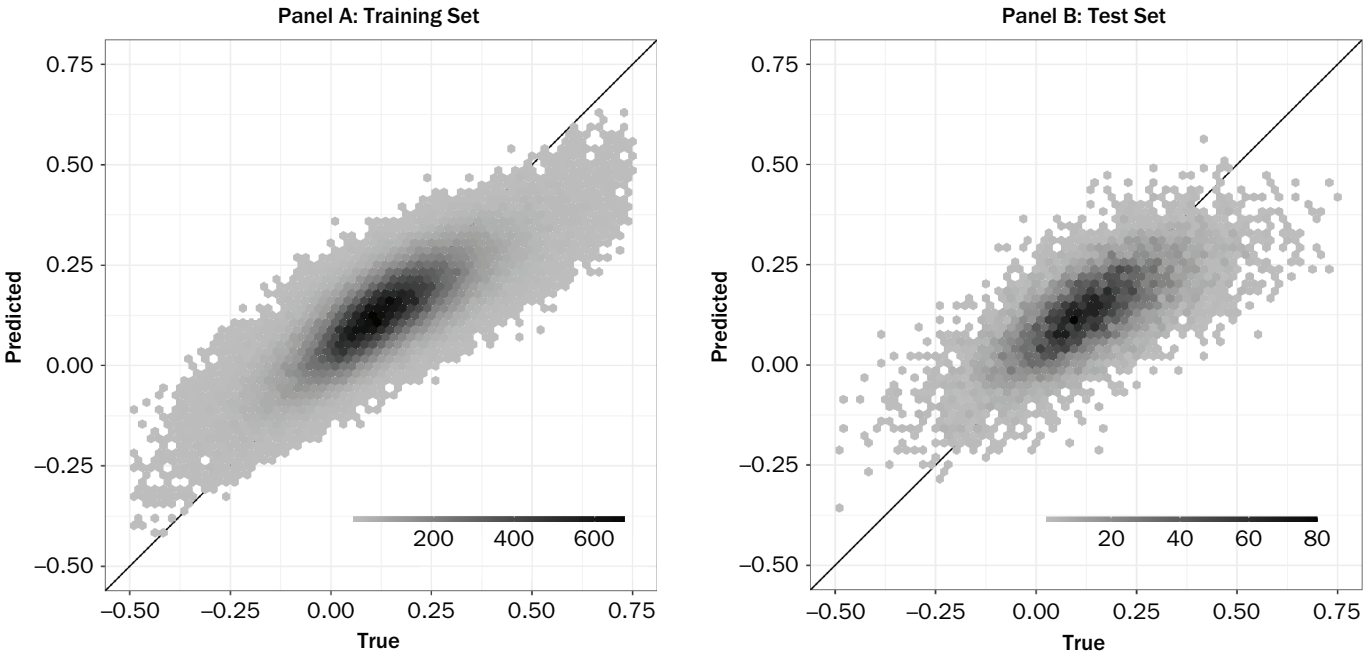
For the supervised learning algorithm, we selected XGBoost (Chen and Guestrin 2016), a gradient tree boosting library that is fast and accurate. This algorithm can construct nonlinear relations among the features. Moreover, for large datasets, it can scale across GPUs to speed up the learning process. 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 a 90% training and a 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. Better accuracy both in the training and in the test set can be reached if we increase the number of samples. However, we do not focus here on predictive accuracy. We instead want to show how the explanation can be used as a discovery tool. Please note that the training and test sets span 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 the training set and 0.1059 in the test set. The R^2 values are 0.6520 in the training and 0.5105 in the test set. The weaker R^2 in the test set means the

EXHIBIT 8

Frequency Plots of Predicted Calmar Ratio Spreads



results are more relevant within the training set. Exhibit 8 shows frequency plots of the predicted Calmar ratio spreads against the true values in the training (Panel A) and test (Panel B) 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.

For the model learning, we used an eight-CPU 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' performance within the in-sample training set. This can be achieved by looking at a set of measures that have been included in the umbrella terms of XAI or interpretable ML. We will focus on a particular one that was revealed to be quite promising because of its generality and relevance (see, e.g., Joseph 2019), called the Shapley values of feature contribution (see Lundberg and Lee 2017 and references therein).

Stated simply, Shapley values tell us how much each feature (the statistical properties of the asset universe described earlier) has contributed to a specific outcome of the ML model. Because of the complexity (nonlinearity) of the model, this is a nontrivial task. The Shapley value is a quantity introduced in cooperative game theory to provide a fair payout to a player (the features) with respect to its contribution toward 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 us to compare 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 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*, which 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, ensuring that missing features make no contributions; and *consistency*, which ensures that if the contribution of a feature is higher in a second model, its feature attribution will be as well. They proved that only one feature attribution has these desirable properties, the classical 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 of working with statistical tests and econometrics analysis (Joseph 2019).

The classical Shapley values are computationally too expensive for any reasonable ML experiment; therefore, Lundberg and Lee (2017) proposed a set of efficient methods that can reliably approximate the explanatory model (SHAP). In particular, in this work we relied on the model tailored for tree ensembles (TreeSHAP) introduced by Lundberg, Erion, and Lee (2018). SHAP can be computed in an accelerated way using the approach of Mitchell, Frank, and Holmes (2020) for tree-based ML models such as XGBoost that, 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 better HRP performance but a negative ERC value. Shapley values are additive quantities; therefore, for example, for a particular asset universe, a Shapley value of $\phi_{\text{meanRET}} = -0.02$ means 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 datasets to obtain a global variable importance that is consistent with the local Shapley values.

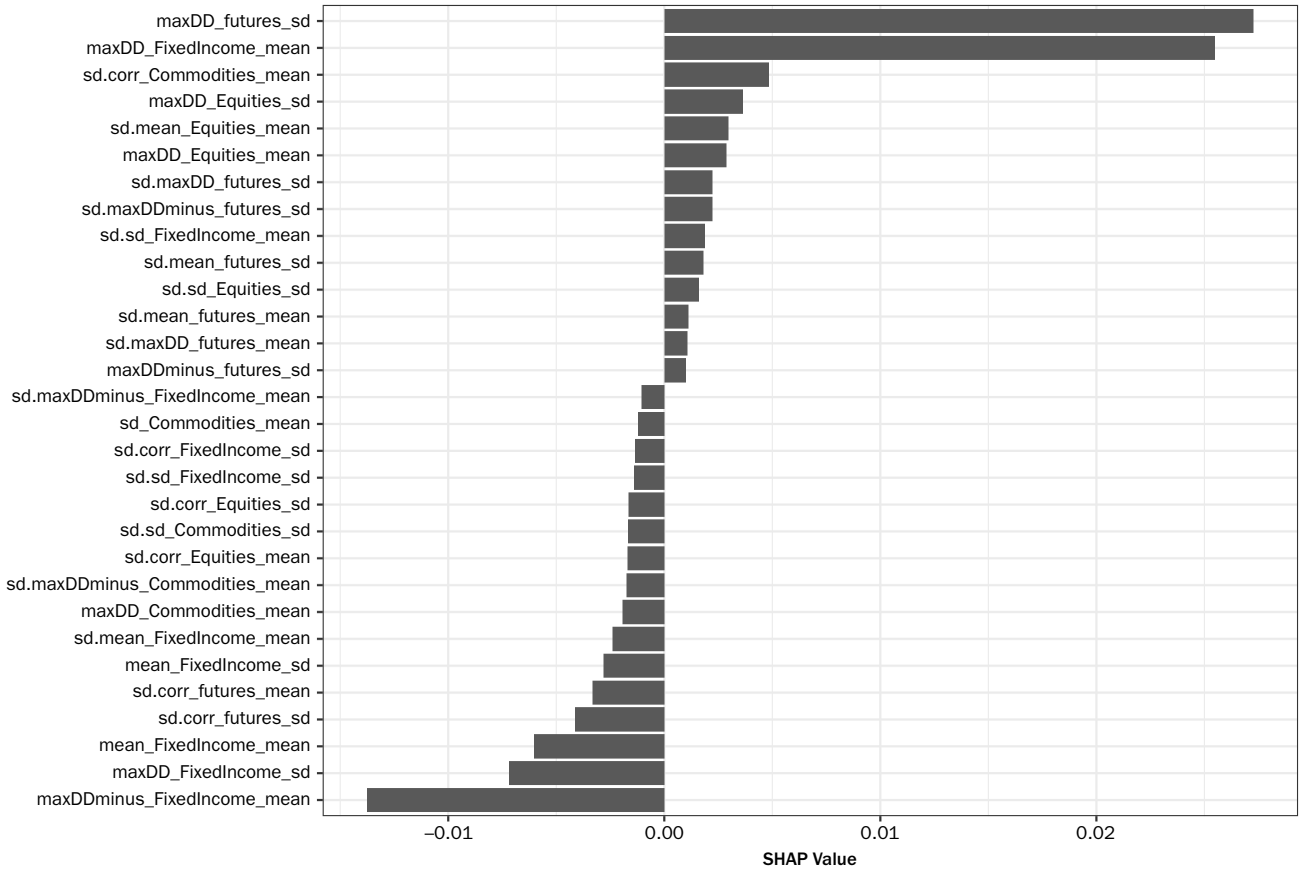
As an example, Exhibit 9 shows the ordered Shapley values for the empirical dataset.

We can see that for the model estimation result of Calmar(HRP) − Calmar(ERC) = 0.158 for the empirical dataset (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, which contributes in favor of HRP. On the other hand, maxDDminus_FixedIncome_mean, the mean drawups of fixed-income futures, contributed negatively (i.e., in favor of ERC strategy). This interplay can best be observed with a breakdown plot (Exhibit 10) that adds intercept ϕ_0 and local Shapley values to yield predicted value. The intercept reflects the mean predicted Calmar ratio spread across all samples in the training set.

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

EXHIBIT 9

Shapley Values for the Empirical Dataset



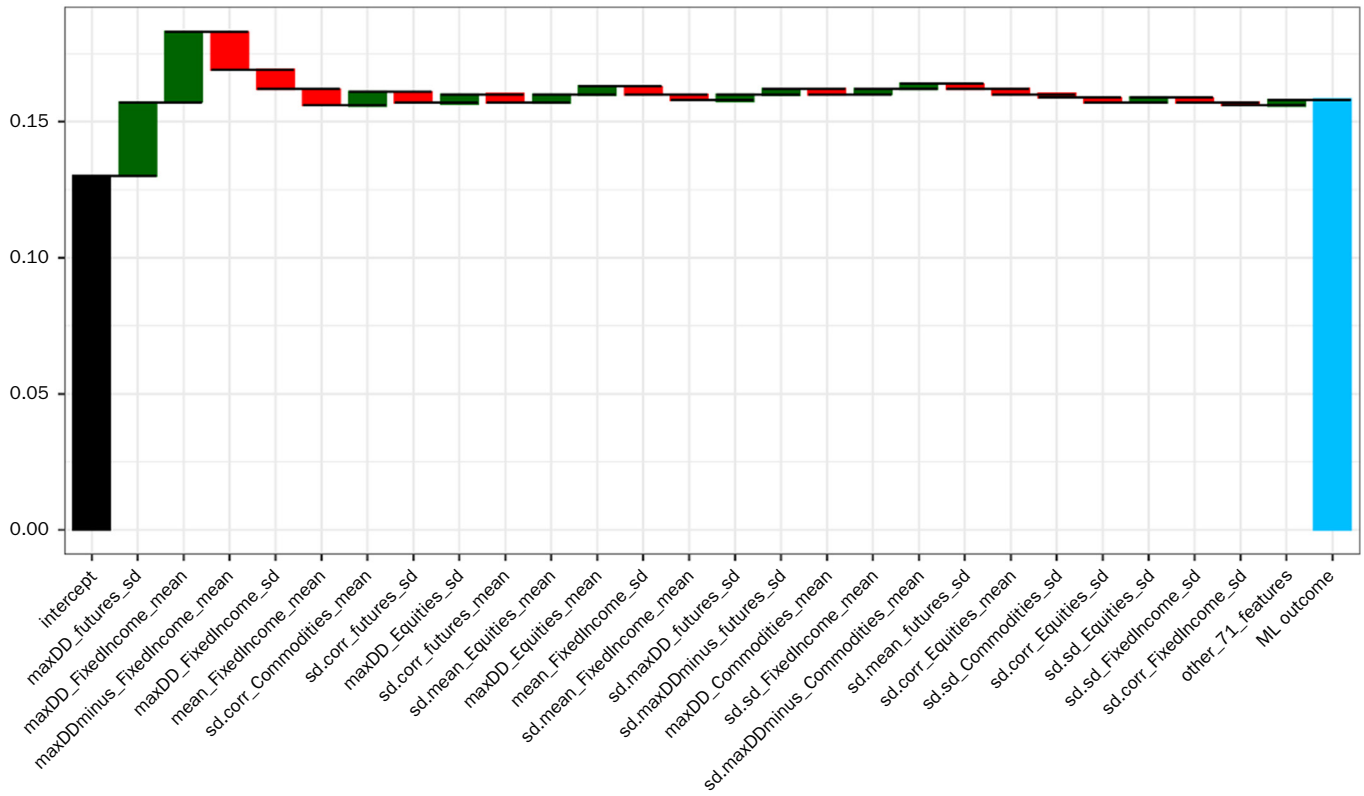
To better understand the relations constructed by the model, it is fruitful to compare the Shapley values with the feature value that actually generated them. Exhibit 11 shows the features with the highest mean absolute Shapley value across the training dataset. The nine most important features are *maxDD_futures_sd, 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.*

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 SD of the MDDs 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 with *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 SD of the cross-asset correlation parameters of the equity futures.

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

EXHIBIT 10

Shapley Value Breakdown for Empirical Dataset



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 dots reflect the empirical dataset.

The plot in the seventh 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 color 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 is true when the differences in correlation parameters are more variable among the years. We do not have to forget that *maxDD_futures_sd* can make a much greater 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.

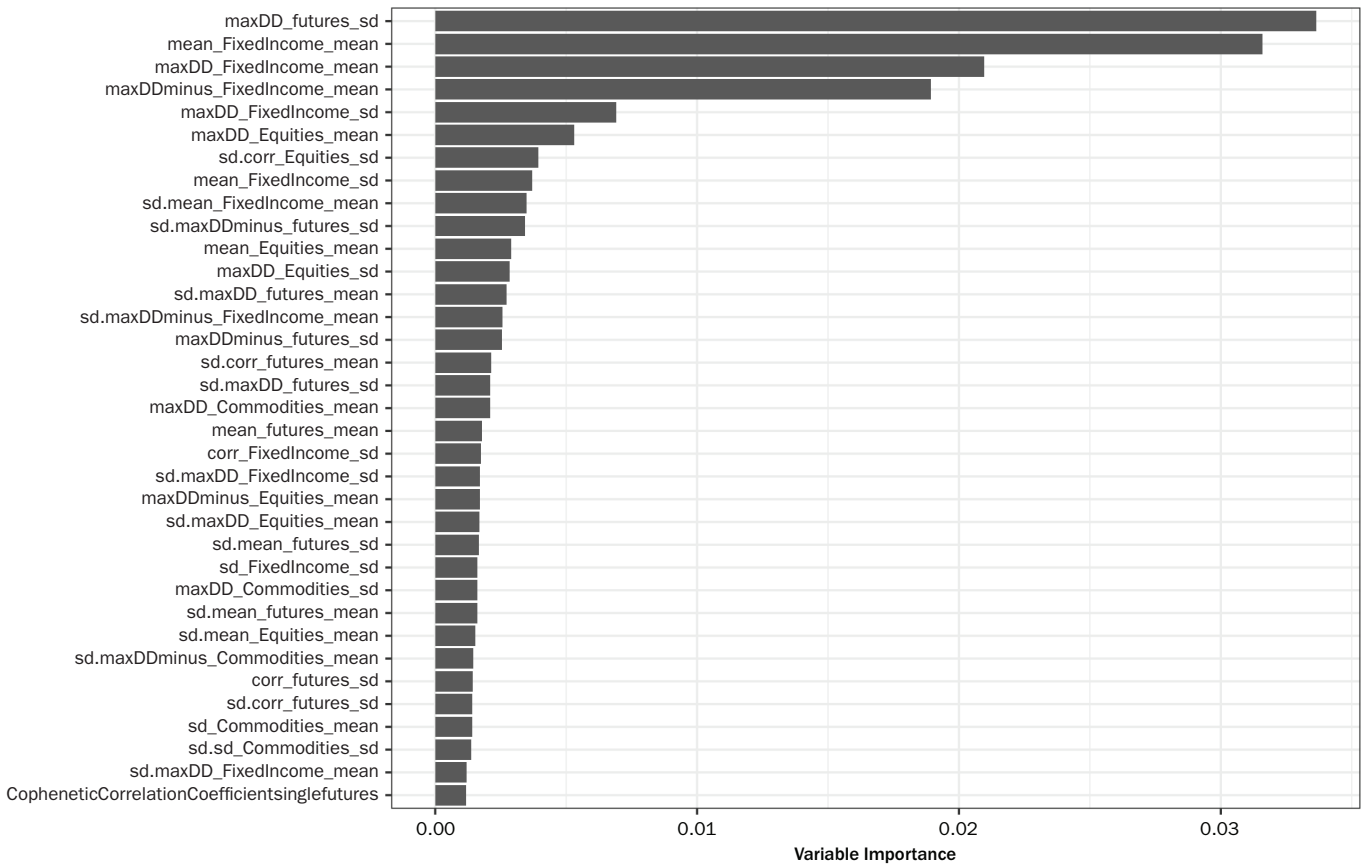
The features measuring the return correlations and the features *ClusterCoefficientsingle* and *CophenticCorrelationCoefficientsingle* measuring the clustering structure do not appear in a prominent position in the ranked feature importance (Exhibit 11).

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 HRP allocation method versus the ERC

EXHIBIT 11

Global Feature Importance



allocation method. Both allocation methods were applied to a multi-asset futures universe of 17 markets and a dynamic rebalancing scheme based on a 5% volatility target. HRP has been claimed to better address the hierarchical correlation structure of real markets than ERC, which relies on an inversion of the covariance matrix. ERC has been scrutinized for its reliance on a negative correlation assumption between equity and bond markets. However, adverse scenarios in which this assumption breaks down did not occur often in the empirical data, so they are not easy to study.

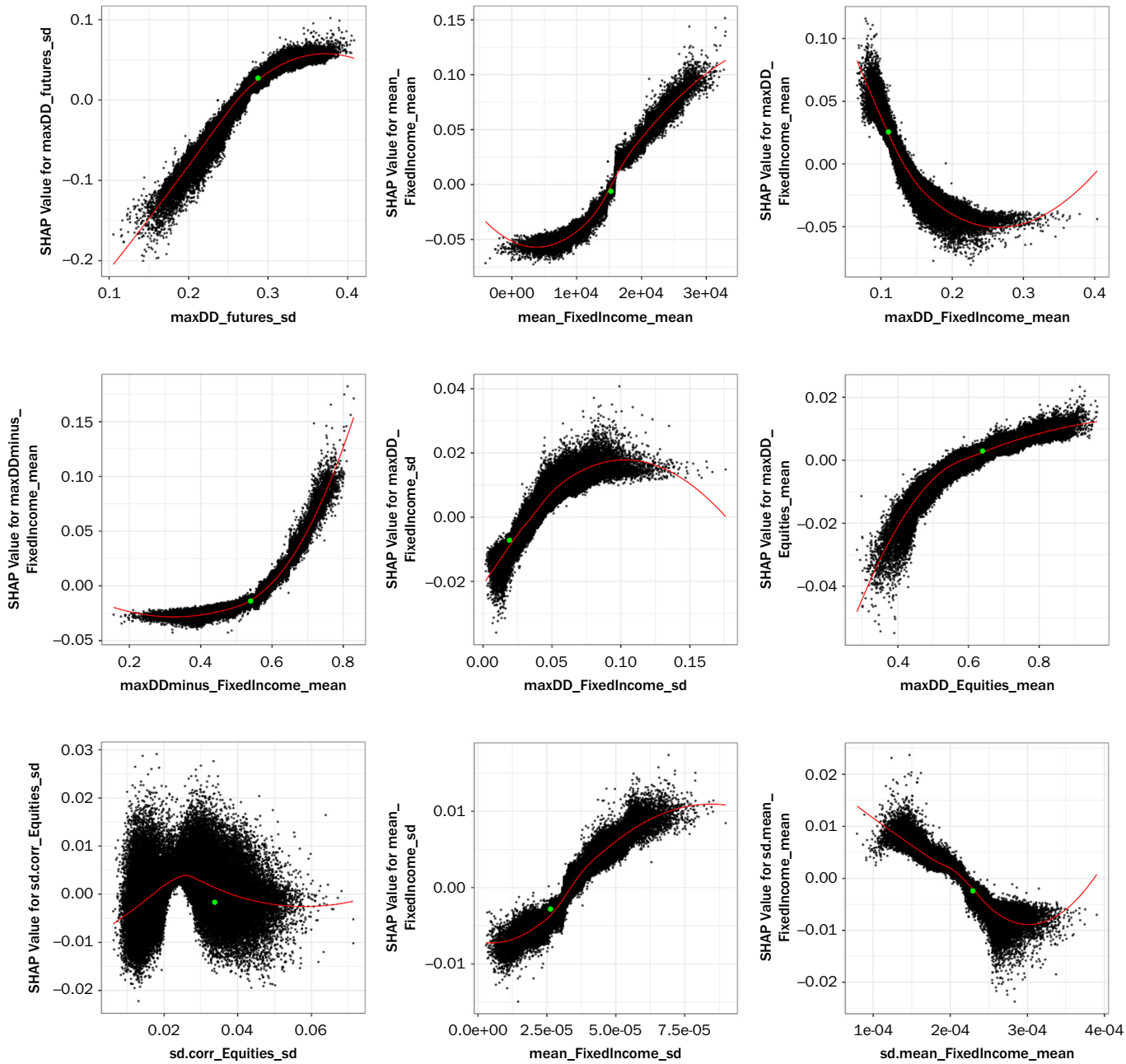
First, in our pipeline, we make use of nonparametric bootstrapping to construct different cross-sectional market scenarios that mimic plausible and possibly problematic correlation structures. Second, we apply XAI methods to discover weaknesses and implicit rules of the complex investment strategies 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 nonlinear 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 RP 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 article would be a suitable testbed to challenge them against the classical HRP

EXHIBIT 12

Shapley Values as a Function of Feature Values

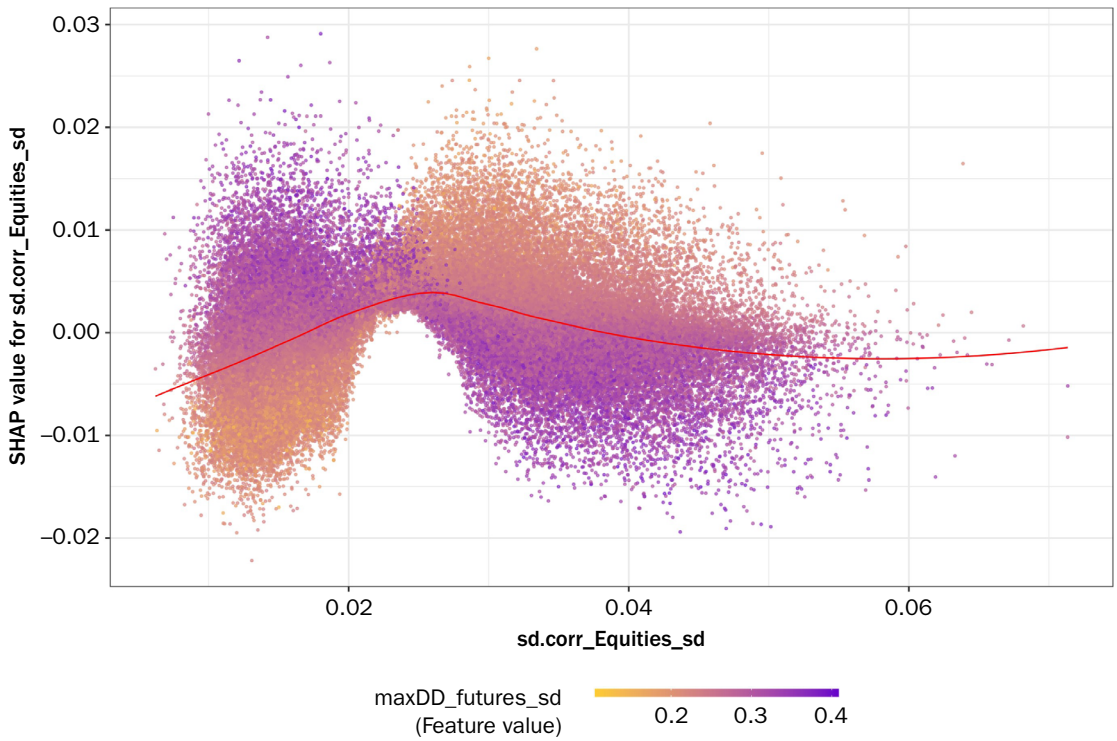


strategy from López de Prado. Moreover, the analysis can be enhanced by comparing other strategies or enriching the training dataset by generating more complex simulations using AI such as generative adversarial networks (see, e.g., Wiese et al. 2019 and Marti 2019) or using the matrix evolutions scheme of Papenbrock et al. (2021).

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 which strategy should be applied for a certain portfolio universe for the future, because the

EXHIBIT 13

Contribution to Calmar(HRP) – Calmar(ERC) as a Function of the *sd.corr_Equities_sd* Values



NOTE: Color represents the value of the feature *maxDD_futures_sd*.

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 within the scope of a future study.

We plan to extend this work using the Shapley value similarity network concept introduced by Bussmann et al. (2020), by testing alternatives to the single-linkage hierarchical clustering in the first step of HRP (Jaeger et al. 2021) and by using synthetically generated scenarios to stress test the allocation strategies.

APPENDIX

HRP ALGORITHM

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

Tree Clustering

From correlation matrix ρ 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 a Euclidean distance between assets as $\widetilde{d_{i,j}} = \sqrt{\sum_{n=1}^{N}(d_{n,i} - d_{n,j})^2}$.

Quasi-Diagonalization

Reorganize the matrix to minimize the distance between columns and construct a linkage matrix.

Recursive Bisection

The matrix is now ordered by the previous step. It is split in half. A split factor of $\alpha = 1 - \frac{\sigma^2(w^{(1)})}{\sigma^2(w^{(1)}) + \sigma^2(w^{(2)})}$ is associated with one of the two blocks, and $1 - \alpha$ is associated with the other. The split factor reflects the minimum variance paradigm for the blocks (i.e., it neglects the off-diagonal blocks). To evaluate the variance of each block, the scheme uses $\sigma^2(w^{(j)}) = w^{*(j)T}\Sigma^{(j)}w^{(j)}$ and $x^{(j)} = \frac{1/\text{diag}[\Sigma^{(j)}]}{\text{tr}(\text{diag}[\Sigma^{(j)}]^{-1})}$, so the internal weights of each block are assigned (temporarily) ignoring the off-diagonal terms in the block, and 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 Lopez de Prado (2016a) words, the technique "takes advantage of the quasi-diagonalization bottom-up because it defines the variance of the partition… using inverse-variance weightings" and it "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:

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

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* (in which futures stands for all 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 SD (aggregated). This construction leads to $5 \times 4 \times 2 = 40$ different quantities.

Moreover, we consider two additional statistical measures associated with the clusterability of the correlation matrix:

ClusterCoefficientssingle	Specifies the agglomerative coefficient, as defined by Kaufman and Rousseeuw (2009), measuring the clustering structure of the dataset
CopheneticCorrelationCoefficientsingle	Correlation between the distance matrix and the ultrametric distance matrix

These are calculated for all assets and for all of the assets as restricted by 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 SD 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.

ACKNOWLEDGMENTS

This research 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. This 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 grant agreement No. 825215 (topic: ICT-35-2018; type of action: CSA). We also would like to thank an anonymous reviewer for helpful comments.

REFERENCES

Asness, C. S., A. Frazzini, and L. H. Pedersen. 2012. "Leverage Aversion and Risk Parity." *Financial Analysts Journal* 68 (1): 47–59.

Baitinger, E., and J. Papenbrock. 2017. "Interconnectedness Risk and Active Portfolio Management." *Journal of Investment Strategies* 6 (2): 63–90.

Bussmann, N., P. Giudici, D. Marinelli, and J. Papenbrock. 2020. "Explainable AI in Credit Risk Management." *Computational Economics* 57: 201–216.

Carlstein, E. 1986. "The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence." *The Annals of Statistics* 14 (3): 1171–1179.

Chen, T., and C. Guestrin. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. New York: ACM, 2016.

Corkery, M., C. Cui, and K. Grind. "Fashionable 'Risk Parity' Funds Hit Hard." *Wall Street Journal*, June 27, 2013, https://www.wsj.com/articles/SB10001424127887323689204578572050047323638.

Denis C., J. Hsu, F. Li, and O. Shakernia. 2011. "Risk Parity Portfolio vs. Other Asset Allocation Heuristic Portfolios." *The Journal of Investing* 20 (1): 108–118.

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

Du, M., N. Liu, and X. Hu. 2020. "Techniques for Interpretable Machine Learning." *Communications of the ACM* 63 (1): 68–77.

Fengler, M. R., and P. Schwendner. 2004. "Quoting Multi-Asset Equity Options in the Presence of Errors from Estimating Correlations." *The Journal of Derivatives* 11 (4): 43–54.

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

Hall, P. 1985. "Resampling a Coverage Pattern." *Stochastic Processes and Their Applications* 20 (2): 231–246.

Harvey, C. R., E. Hoyle, R. Korgaonkar, S. Rattray, M. Sargaison, and O. Van Hemert. 2018. "The Impact of Volatility Targeting." *The Journal of Portfolio Management* 45 (1): 14–33.

Huettner, A., J.-F. Mai, and S. Mineo. 2018. "Portfolio Selection Based on Graphs: Does It Align with Markowitz-Optimal Portfolios?" *Dependence Modeling* 6: 63–87.

Jaeger, M., S. Krügel, J. Papenbrock and P. Schwendner. 2021. "'Adaptive Seriational Risk Parity' and other Extensions for Heuristic Portfolio Construction using Machine Learning and Graph Theory." *The Journal of Financial Data Science* (forthcoming).

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

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

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep Learning." *Scientific Report* 521 (7553): 436–444.

Lohre, H., C. Rother, and K. Schaefer. 2020. "Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations." In Machine Learning for Asset Management, E. Jurczenko (Ed.). https://doi.org/10.1002/9781119751182.ch9.

Lopez de Prado, M. 2016a. "Building Diversified Portfolios That Outperform Out of Sample." *The Journal of Portfolio Management* 42 (4): 59–69.

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

——. *Advances in Financial Machine Learning*. Hoboken: Wiley, 2018.

——. "Robots on Wall Street: The Impact of AI on Capital Markets and Jobs in the Financial Services Industry." Testimony before the US House Of Representatives Committee On Financial Services—Task Force On Artificial Intelligence, 2019.

Lundberg, S. M., G. G. Erion, and S.-I. Lee. "Consistent Individualized Feature Attribution for Tree Ensembles." *arXiv* 1802.03888, 2018.

Lundberg, S., and S.-I. Lee. "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, pp. 4765–4774. Red Hook, NY: Curran Associates, Inc., 2017.

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

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

Marti, G., F. Nielsen, M. Binkowski, and P. Donnat. "A Review of Two Decades of Correlations, Hierarchies, Networks and Clustering in Financial Markets." *arXiv e-Prints*, March 2017, https://arxiv.org/abs/1703.00485.

Michaud, R. O., and R. O. Michaud. *Efficient Asset Management: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. New York: Oxford University Press, 2008.

Mitchell, R., E. Frank, and G. Holmes. "GPUTreeShap: Fast Parallel Tree Interpretability." 2020, http://arxiv.org/abs/2010.13972.

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

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. 2019. "Definitions, Methods, and Applications in Interpretable Machine Learning." *Proceedings of the National Academy of Sciences* 116 (44): 22071–22080.

Mussard, S., and V. Terraza. 2008. "The Shapley Decomposition for Portfolio Risk." *Applied Economics Letters* 15 (9): 713–715.

Neukirch, T. "Alternative Indexing with the MSCI World Index." *SSRN* 1106109, 2008.

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

Papenbrock, J., and P. Schwendner. 2015. "Handling Risk on/Risk Off Dynamics with Correlation Regimes and Correlation Networks." *Financial Markets and Portfolio Management* 29: 125–147.

Papenbrock, J., P. Schwendner, M. Jaeger, and S. Krügel. 2021. "Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios." *The Journal of Financial Data Science* 3 (2): 51–69.

Pozzi, F., T. Di Matteo, and T. Aste. 2013. "Spread of Risk Across Financial Markets: Better to Invest in the Peripheries." *Scientific Reports* 3: 1665.

Qian, E. "Risk Parity Portfolios: Efficient Portfolios through True Diversification." Panagora Asset Management, 2005.

Roncalli, T. 2013. *Introduction to Risk Parity and Budgeting*. Boca Raton, FL: Chapman & Hall, 2013.

Shapley, L. S. "A Value for n-Person Games." In *Contributions to the Theory of Games II*, edited by H. Kuhn and A. W. Tucker, pp. 307–317. Princeton, NJ: Princeton University Press, 1953.

Simonian, J. 2012. "A Formal Methodology for Aggregating Multiple Market Views." *Applied Financial Economics* 22 (14): 1175–1179.

——. 2014. "Copula-Opinion Pooling with Complex Opinions." *Quantitative Finance* 14 (6): 941–946.

——. 2019. "Portfolio Selection: A Game-Theoretic Approach." *The Journal of Portfolio Management* 45 (6): 108–116.

Wiese, M., R. Knobloch, R. Korn, and P. Kretschmer. 2019. "Quant GANs: Deep Generation of Financial Time Series." 2019, http://arxiv.org/abs/1907.06673.