Hierarchical Clustering based Asset Allocation

Thomas Raffinot *

September 2016

Abstract

A hierarchical clustering based asset allocation method, which uses graph theory and machine learning techniques, is proposed. Classical and more modern hierarchical clustering methods are tested, such as Simple Linkage or Directed Bubble Hierarchical Tree for example. Once the assets are hierarchically clustered, a simple and efficient capital allocation within and across clusters of assets at multiple hierarchical levels is computed. The out-of-sample performances of hierarchical clustering based portfolios and more traditional risk-based portfolios are evaluated across three disparate datasets. To avoid data snooping, the comparison of profit measures is assessed using the bootstrap based model confidence set procedure. The empirical results indicate that hierarchical clustering based portfolios are robust, truly diversified and achieve statistically better risk-adjusted performances than commonly used portfolio optimization techniques.

JEL classifications: G00, G10, G11
Keywords: Hierarchical Clustering; Asset Allocation; Model Confidence Set; Portfolio Construction; Graph Theory; Financial Networks; Machine Learning; Risk Parity

* *Millesime-IS/ PSL Research University, Université Paris Dauphine, LEDa-SDFi (email: traffinot@gmail.com)*

Introduction

Nobel Prize winner Harry Markowitz described diversification, with its ability to enhance portfolio returns while reducing risk, as the only the "only free lunch" in investing (Markowitz (1952)). Yet diversifying a portfolio in real life is easier said than done.

Investors are aware of the benefits of diversification but form portfolios without giving proper consideration to the correlations (Goetzmann and Kumar (2008)). Moreover, modern and complex portfolio optimisation methods are optimal in-sample, but often provide rather poor out-of-sample forecast performance. For instance, DeMiguel et al. (2009) demonstrate that the equal-weighted allocation, which gives the same importance to each assets, beats the entire set of commonly used portfolio optimization techniques. In fact, optimized portfolios depend on expected returns and risks, but even small estimation errors can result in large deviations from optimal allocations in an optimizer's result (Michaud (1989)).

To overcome this issue, academics and practitioners have developed risk-based portfolio optimization techniques (minimum variance, equal risk contribution, risk budgeting,...), which do not rely on return forecasts (Roncalli (2013)). However, the inversion of a positive-definite covariance matrix remains needed, which lead to errors of such magnitude that they entirely offset the benefits of diversification (López de Prado (2016b)).

Exploring a new way of capital allocation, López de Prado (2016a) introduces a portfolio diversification technique called "Hierarchical Risk Parity" (HRP), which uses graph theory and machine learning techniques. One of the main advantages of HRP is to manage to compute a portfolio on an ill-degenerated or even a singular covariance matrix.

The starting point of his analysis is that a correlation matrix is too complex to be properly analysed and understood. If you have N assets of interest, there are $\frac{1}{2}N(N-1)$ pairwise correlations among them and that number grows quickly. For example, there are as many as 4950 correlation coefficients between stocks of the FTSE 100 and 124750 between stocks of the S&P 500. More importantly, López de Prado (2016a) points out that correlation matrices lack the notion of hierarchy. Actually, Nobel prize laureate Herbert Simon has demonstrated that complex systems can be arranged in a natural hierarchy, comprising nested sub-structures (Simon (1962)). But, a correlation matrix makes no difference between assets. Yet, some assets seem closer substitutes of one another, while others seem complementary to one another. This lack of hierarchical structure allows weights to vary freely in unintended ways (López de Prado (2016a)).

To simplify the analysis of the relationships between this large group of relative prices, López de Prado (2016a) applies a correlation-network method known as "Minimum Spanning Tree (MST)"[1]. Its main principle is easy to understand: the heart of correlation analysis is choosing which correlations really matter; in other words, choosing which links in the network are important, and removing the rest, keeping $N-1$ links[2].

The MST is strictly related to a hierarchical clustering algorithm, named the "Single Linkage" (Tumminello et al. (2010)). Graph theory is indeed linked to unsupervised machine learning, hierarchical clustering especially. For instance, another hierarchical clustering method, the "Average Linkage", has been shown to be associated to a slightly different version of spanning tree called Average Linkage Minimum Tree (Tumminello et al. (2007)). Hierarchical clustering refers to the formation of a recursive clustering, suggested by the data, not defined a priori. The objective is to build a binary tree of the data that successively merges similar groups of points. Hierarchical clustering is thus another way to filter correlations. Another variants of hierarchical clustering algorithm (Complete Linkage, Ward's method) are not associated to a spanning tree representation. Yet, they may provide interesting results.

At last, there are many generalizations of the MST. The Planar Maximally Filtered Graph (Tumminello et al. (2005)) is a recent and prominent one. It is associated with a hierarchical clustering method, the Directed Bubble Hierarchical Tree (Musmeci et al. (2015)).

Building upon López de Prado (2016a) and Simon (1962), this paper exploits the notion of hierarchy. Different hierarchical clustering methods are presented and tested, namely Simple Linkage, Complete Linkage, Average Linkage, Ward's Method and Directed Bubble Hierarchical Tree (DBHT). Once the assets are hierarchically clustered, a simple and efficient capital allocation within and across clusters of assets at multiple hierarchical levels is computed.

The out-of-sample performances of hierarchical clustering based portfolios and risk-based portfolios are evaluated across three empirical datasets, which differ in terms of number of assets and composition of the universe ("S&P

[1] Since the seminal work of Mantegna (1999), correlation-networks have been extensively used in Econophysics as tools to filter, visualise and analyse financial market data (see Baitinger and Papenbrocky (2016) for a review)

[2] One concrete example would be a telecommunications company which is trying to lay out cables in new neighborhood. In any case, the easiest possibility to install new cables is to bury them along roads. Some of those paths might be more expensive, because they are longer, or require the cable to be buried deeper. These paths would be represented by edges with larger weights. A spanning tree for that graph would be a subset of those paths that has no cycles but still connects to every house. There might be several spanning trees possible. A minimum spanning tree would be one with the lowest total cost would then represent the least expensive path for laying the cable.

sectors", multi-assets and individual stocks). To avoid data snooping, which occurs when a given set of data is used more than once for purposes of inference or model selection, the comparison of profit measures is assessed using the bootstrap based model confidence set procedure proposed by Hansen et al. (2011). It prevents strategies that perform by luck to be considered as effective.

The findings of the paper can be summarized as follows: hierarchical clustering based portfolios are robust, truly diversified and achieve statistically better risk-adjusted performances than commonly used portfolio optimization techniques. Among clustering methods, there is no clear winner. DBHT based portfolios produce slightly superior risk-adjusted returns but Average Linkage based portfolios are clearly more robust.

The rest of the paper proceeds as follows. Section 1 describes the risk budgeting approach. Section 2 introduces hierarchical clustering methods and their application to asset allocation. Section 3 presents the empirical set up: the datasets and the comparison criteria. Section 4 analyses the empirical results.

1 Risk Budgeting Approach

This section briefly describes risk budgeting portfolios. Refer to Roncalli (2013) for a detailed exposition of this approach. In a risk budgeting approach, the investor only chooses the risk repartition between assets of the portfolio, without any consideration of returns, thereby partially dealing with the issues of traditional portfolio optimization methods.

1.1 Notations and definitions

Consider a portfolio invested in N assets with portfolio weights vector $w = (w_1, w_2, \cdots, w_N)'$. Returns are assumed to be arithmetic: $r_{t,i} = (p_{t,i} - p_{t-1,i})/p_{t-1,i} = p_{t,i}/p_{t-1,i} - 1$. The portfolio return at time t is thus:

$$r_{P,t} = \sum\nolimits_{i=1}^{N} w_i r_{t,i}$$

Let σ_i^2 be the variance of asset i, σ_{ij} be the covariance between assets i and j and Σ be the covariance matrix.

The volatility is defined as the risk of the portfolio:

$$\mathcal{R}_w = \sigma_w = \sqrt{w'\Sigma w}$$

and μ is the expected return:

$$\mu = E(r_P) = \sum\nolimits_{i=1}^{N} w_i E(r_i)$$

.

1.2 Risk budgeting portfolios

In a risk budgeting portfolio, the risk contribution from each components is equal to the budget of risk defined by the portfolio manager.

Since the risk measure is coherent and convex, the Euler decomposition is verified:

$$\mathcal{R}_w = \sum_{i=1}^{N} w_i \frac{\partial \mathscr{R}_w}{\partial w_i}$$

With the volatility as the risk measure, the risk contribution of the i^{th} asset becomes:

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

A long-only, full invested risk budgeting portfolio is defined as follows (Roncalli (2013)):

$$\begin{cases} \mathcal{RC}_{w_i} = b_i \mathcal{RC}_w \\ b_i > 0 \\ \sum_{i=1}^{N} b_i = 1 \\ w_i \geq 0 \\ \sum_{i=1}^{N} w_i = 1 \end{cases}$$

Once a set of risk budgets is defined, the weights of the portfolio are computed so that the risk contributions match the risk budgets.

In this paper, four risk budgeting portfolios are considered[3]:

- The minimum variance (MV) portfolio is a risk budgeting portfolio where the risk budget is equal to the weight of the asset:

$$b_i = w_i$$

- The most diversified portfolio (MDP) (Choueifaty et al. (2013)) is a risk budgeting portfolio where the risk budgets are linked to the product of the weight of the asset and its volatility:

[3] Five if the equal weighted portfolio is considered as a risk budgeting portfolio.

$$b_i = \frac{w_i \sigma_i}{\sum_{i=1}^{N} w_i \sigma_i}$$

- The equal risk contribution portfolio (ERC) (Maillard et al. (2010)) is a risk budgeting portfolio where the risk contribution from each asset is made equal:

$$b_i = \frac{1}{N}$$

- The inverse-variance (IVRB) risk budgeting portfolio defines risk budgets as follows:

$$b_i = \frac{\sigma_i^{-2}}{\sum_{i=1}^{N} \sigma_i^{-2}}$$

The cyclical coordinate descent (CCD) algorithm for solving high dimensional risk parity problems (Griveau-Billion et al. (2013)) is employed to estimate the risk-based models.

2 Hierarchical clustering and asset allocation

2.1 Notion of hierarchy

Nobel Prize winner Herbert Simon has demonstrated that complex systems, such as financial markets, have a structure and are usually organized in a hierarchical manner, with separate and separable sub-structures (Simon (1962)). The hierarchical structure of interactions among elements strongly affects the dynamics of complex systems. The need of a quantitative description of hierarchies to model complex systems is thus straightforward (Anderson (1972)).

López de Prado (2016a) points out that correlation matrices lack the notion of hierarchy, which allows weights to vary freely in unintended ways. Moreover, he provides a concrete example to highlight the interest of the notion of hierarchy for asset allocation: "stocks could be grouped in terms of liquidity, size, industry and region, where stocks within a given group compete for allocations. In deciding the allocation to a large publicly-traded U.S. financial stock like J.P. Morgan, we will consider adding or reducing the allocation to another large publicly-traded U.S. bank like Goldman Sachs, rather than a small community bank in Switzerland, or a real estate holding in the Caribbean". To sum up, a correlation matrix makes no difference between assets. Yet, some assets seem closer substitutes of one another, while others seem complementary to one another.

2.2 Hierarchical clustering

The purpose of cluster analysis is to place entities into groups, or clusters, suggested by the data, not defined a priori, such that entities in a given cluster tend to be similar to each other and entities in different clusters tend to be dissimilar.

Hierarchical clustering refers to the formation of a recursive clustering. The objective is to build a binary tree of the data that successively merges similar groups of points. The tree based representation of the observations is called a dendrogram. Visualizing this tree provides a useful summary of the data.

Hierarchical clustering requires a suitable distance measure. The following distance is used (Mantegna (1999)):

$$D_{i,j} = \sqrt{2(1 - \rho_{i,j})}$$

where $D_{i,j}$ is the correlation-distance index between the i^{th} and j^{th} asset and $\rho_{i,j}$ is the respective Pearson's correlation coefficients.

Four agglomerative clustering variants are tested in this study, namely: Single Linkage (SL), Average Linkage (AL), Complete Linkage (CL), Ward's Method (WM).

An agglomerative clustering starts with every observation representing a singleton cluster and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. At each of the $N - 1$ steps the closest two (least dissimilar) clusters are merged into a single cluster, producing one less cluster at the next higher level. Therefore, a measure of dissimilarity between two clusters must be defined and different definitions of the distance between clusters can produce radically different dendrograms. The clustering variants are described below:

- Single Linkage: the distance between two clusters is the minimum of the distance between any two points in the clusters. For clusters C_i, C_j:

  $$d_{C_i,C_j} = min_{x,y}\{D(x,y) \mid x \in C_i, y \in C_j\}$$

  This method is relatively simple and can handle non-elliptical shapes. Nevertheless, it is sensitive to outliers and can result in a problem called chaining whereby clusters end up being long and straggly. The SL algorithm is strictly related to the one that provides a Minimum Spanning Tree (MST). However the MST retains some information that the SL dendrogram throws away.

- Complete Linkage: the distance between two clusters is the maximum of the distance between any two points in the clusters. For clusters C_i, C_j:

$$d_{C_i,C_j} = max_{x,y}\{D(x,y) \mid x \in C_i, y \in C_j\}$$

  This method tends to produce compact clusters of similar size but, is quite sensitive to outliers.

- Average linkage: the distance between two clusters is the average of the distance between any two points in the clusters. For clusters C_i, C_j:

$$d_{C_i,C_j} = mean_{x,y}\{D(x,y) \mid x \in C_i, y \in C_j\}$$

  This is considered to be a fairly robust method.

- Ward's Method (Ward (1963)): the distance between two clusters is the increase of the squared error that results when two clusters are merged. For clusters C_i, C_j with sizes m_i, m_j, respectively,

$$d_{C_i,C_j} = \frac{m_i m_j}{m_i + m_j} ||c_i - c_j||^2.$$

  where c_i, c_j are the centroids for the clusters.

This method is biased towards globular clusters, but less susceptible to noise and outliers. It is one of the most popular methods.

To determine the number of clusters, the Gap index (Tibshirani et al. (2001)) is employed. It compares the logarithm of the empirical within-cluster dissimilarity and the corresponding one for uniformly distributed data, which is a distribution with no obvious clustering.

The last approach differs completely from the agglomerative one: the idea of Directed Bubble Hierarchical Tree (DBHT) is to use the hierarchy hidden in the topology of a Planar Maximally Filtered Graph (PMFG) (Tumminello et al. (2005)).

The PMFG network keeps the hierarchical structure of the MST network but contains a greater amount of information by connecting N nodes (assets) with $3(N-2)$ edges. The basic elements of a PMFG are three-cliques (subgraphs made of three nodes all reciprocally connected). For a detailed introduction of MST and PMFG, see Tumminello et al. (2005) and Aste et al. (2010).

Musmeci et al. (2015) explains that the DBHT exploits this topological structure, and in particular the distinction between separating and non-separating three-cliques, to identify a clustering partition of all nodes in the PMFG. First of all the clusters are identified by means of topological considerations on the planar graph, then the hierarchy is constructed both inter-clusters and intra-clusters. The difference involves therefore both the kind of information exploited and the methodological approach. It is important to note that the "optimal" number of clusters is determined during the process (see Song et al. (2012) for more on this subject).

2.3 Asset allocation weights

Once the clusters have been determined, the capital should be efficiently allocated both within and across groups. Indeed, a compromise between diversification across all investments and diversification across clusters of investments at multiple hierarchical levels has to be found.

Since asset allocation within and across clusters can be based on the same or different methodologies, there are countless options.

The chosen weighting scheme attempts to stay very simple and focuses not only on the clusterings, but on the entire hierarchies associated to those clusterings. The principle is to find a diversified weighting by distributing capital equally to each cluster hierarchy, so that many correlated assets receive the same total allocation as a single uncorrelated one. Then, within a cluster, an equal-weighted allocation is computed.

For example, Figure 1 exhibits a small dendrogram with five assets and three clusters. The first cluster is made up of assets 1 and 2, asset 5 constitutes the second cluster and the third cluster consists of assets 3 and 4. Based on the hierarchical clustering weighting, weights for cluster number one is 0.5 ($\frac{1}{2} = 0.5$) and weights for clusters 2 and 3 are 0.25 ($\frac{0.5}{2} = 0.25$). Since there are two assets in the cluster number one, final weights for assets 1 and 2 are $\frac{0.5}{2} = 0.25$. Asset 5 would have a weight of $\frac{0.25}{1} = 0.25$. At last, assets 3 and 4 would get a weight of $\frac{0.25}{2} = 0.125$.

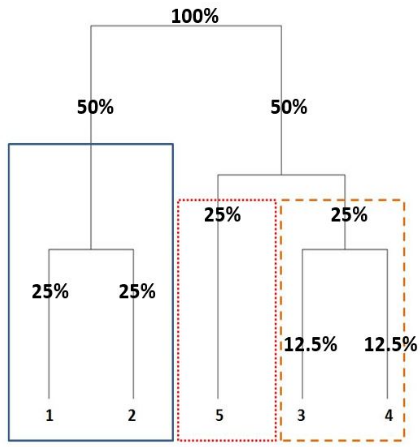

Figure 1: Asset allocation weights: a small example

This weighting scheme should guarantee the diversification and the robustness of the portfolio. For instance, since at least two clusters are considered, the weights are constrained: $\forall i : 0 \leq w_i \leq 0.5$. Moreover, if clusters are lasting, then weights should be very stable. At last, neither expected returns nor risk measures are required, thereby making the method more robust.

The will to exploit the nested clusters or in other words the notion of hierarchy explains why clustering methods such as K-means or K-medoids have not been tested. Indeed, these algorithms provide a single set of clusters, with no particular organisation or structure within them[4].

2.4 Future research

The aim of this paper is to introduce hierarchical clustering based asset allocation. Nevertheless, there may be several way of improving the out-of-sample performance. Improvements are left for future research.

- Typical machine learning issues have to be investigated, such as the choice of the distance measure and the criteria used to select the number of clusters.
- Another weighting strategy is obviously possible. For instance, López de Prado (2016a) employs an inverse-variance weighting allocation.

[4]The results of applying K-means or K-medoids clustering algorithms depend on the choice for the number of clusters to be searched and a starting configuration assignment. In contrast, hierarchical clustering methods do not require such specifications.

- Even if the proposed method does not require a matrix to be inverted, the correlation matrix has still to be estimated. Historical asset returns are commonly used. But, in many cases, the length of the asset returns' time series used for estimation is not long enough compared to the number of assets considered (Jobson and Korkie (1980)). As a result, the estimated correlation matrix is unstable. One approach to improving estimation is to use "shrinkage". The general idea is that a compromise between a logical/theoretical estimator and a sample estimator will yield better results than either method (see Ledoit and Wolf (2004), Ledoit and Wolf (2014) and Gerber et al. (2015)).

- More importantly, asset classes behave differently during different phases of the economic cycles (Raffinot (2014)): asset classes are more correlated during bad times than during good times. An important implication of higher correlations is that otherwise diversified portfolio lose some of diversification benefit during bad times, when most needed. Taking into account the current phase of the economic cycles, especially the growth cycle, would constitute a real breakthrough.

3 Investment strategies comparison

Portfolios are updated on a daily basis via a 252 days rolling window approach, with no forward-looking biases. This approach differs from the traditional one, where portfolio are rebalanced on a more realistic monthly basis[5]. Nevertheless, the main objective of this paper is not to create a real investment strategy, but to compare asset allocation methods. The daily rebalancing framework should help highlighting the strengths and weaknesses of the different approaches, especially the robustness.

3.1 Datasets

The out-of-sample performances of models are evaluated across three very disparate datasets. The three considered datasets differ in term of assets' composition and number of assets[6]:

- The "S&P sectors" dataset consists of daily returns on 10 value weighted industry portfolios formed by using the Global Industry Classification

[5]For investors, the choice of the rebalancing strategy is crucial. The periodic rebalancing is not optimal and others options should be investigated (Sun et al. (2006)).

[6]Data are available from the author upon request.

Standard (GICS) developed by Standard & Poor's. The 10 industries considered are Energy, Material, Industrials, Consumer-Discretionary, Consumer-Staples, Healthcare, Financials, Information-Technology, Telecommunications, and Utilities. The data span from January 1995 to August 2016.

- The multi-assets dataset is constituted of asset classes exhibiting different risk-return characteristics (in local currencies): S&P 500 (US large cap), Russell 2000 (US small cap), Euro Stoxx 50 (EA large cap), Euro Stoxx Small Cap (EA small cap), FTSE 100 (UK large cap), FTSE Small Cap (UK small cap), France 2-Year bonds, France 5-Year bonds, France 10-Year bonds, France 30-Year bonds, US 2-Year bonds, US 5-Year bonds, US 10-Year bonds, US 30-Year bonds, MSCI Emerging Markets (dollars), Gold (dollars).

  France has been chosen over Germany for data availability reasons. A difficult decision has been made for fixed-income indices: coupons are not reinvested. The reasoning is the following: rates are low and are expected to stay low for a long time[7]. It implies that performances in the future will not come from coupons. As the aim is to build portfolios that will perform and not that have performed, this solution has been preferred. As a consequence, no dividends are reinvested. The data span from February 1989 to August 2016.

- 357 individual stocks with a sufficiently long historical data from the current S&P 500 compose the last dataset. The objective is to get "real" correlations between stocks. Obviously, this dataset does not incorporate information on delistings. Since there is a strong survivor bias, comparisons with the S&P 500 are meaningless. Nevertheless, comparisons between different models are meaningful. The data span from January 1996 to August 2016.

Although more data history would have been desirable, the different periods cover a number of different market regimes and shocks to the financial markets and the world economy, including the "dot-com" bubble, the Great Recession and the 1994 and 1998 bond market crashes as regards the multi-asset dataset.

[7] Off topic: low interest rates are not the symbol of easy monetary policy, but rather an outcome of excessively tight monetary policy (see Friedman (1992) and Friedman (1997)).

3.2 Comparison measures

Given the time series of daily out-of-sample returns generated by each strategy in each dataset, several comparison criteria are computed:

- The Adjusted Sharpe Ratio ($\mathcal{ASR}$) (Pezier and White (2008))[8] explicitly adjusts for skewness and kurtosis by incorporating a penalty factor for negative skewness and excess kurtosis:

$$\mathcal{ASR} = \mathcal{SR}[1 + (\frac{\mu_3}{6})\mathcal{SR} + \frac{(\mu_4 - 3)}{24})\mathcal{SR}^2]$$

  where μ_3 and μ_4 are the skewness and kurtosis of the returns distribution and $\mathcal{SR}$ denotes the traditional Sharpe Ratio ($\mathcal{SR} = \frac{\mu - r_f}{\sigma}$, where r_f is the risk-free rate[9]).

- The certainty-equivalent return ($\mathcal{CEQ}$) is the risk-free rate of return that the investor is willing to accept instead of undertaking the risky portfolio strategy. DeMiguel et al. (2009) define the $\mathcal{CEQ}$ as:

$$\mathcal{CEQ} = (\mu - r_f) - \frac{\gamma}{2}\sigma^2$$

  where γ is the risk aversion. Results are reported for the case of $\gamma = 1$ but other values of the coefficient of risk aversion are also considered as a robustness check. More precisely, the employed definition of $\mathcal{CEQ}$ captures the level of expected utility of a mean-variance investor, which is approximately equal to the certainty-equivalent return for an investor with quadratic utility (DeMiguel et al. (2009)). It is the most important number to consider to build profitable portfolios (Levy (2016)).

- The Max drawdown ($\mathcal{MDD}$) is an indicator of permanent loss of capital. It measures the largest single drop from peak to bottom in the value of a portfolio. In brief, the $\mathcal{MDD}$ offers investors a worst case scenario.

- The average turnover per rebalancing ($\mathcal{TO}$):

$$\mathcal{TO} = \frac{1}{F}\sum_{t=2}^{F}|w_{i,t} - w_{i,t-1}|$$

  where F is the number of out-of-sample forecasts.

[8] Similar to the Adjusted Sharpe Ratio, the Modified Sharpe Ratio uses Modified VaR adjusted for skewness and kurtosis as a risk measure.

[9] A risk-free interest rate of zero is assumed when calculating the $\mathcal{ASR}$ and $\mathcal{CEQ}$.

- The Sum of Squared Portfolio Weights ($\mathcal{SSPW}$) used in Goetzmann and Kumar (2008) exhibits the underlying level of diversification in a portfolio and is defined as follows:

$$\mathcal{SSPW} = \frac{1}{F} \sum_{t=2}^{F} \sum_{i=1}^{N} w_{i,t}^2$$

  $\mathcal{SSPW}$ ranges from 0 to 1, where 1 represents the most concentrated portfolio.

No transactions costs or economic costs generated by the turnover are reported. Indeed, the study of transactions costs is difficult because investors face different fees and the same strategy can be implemented via Futures or ETF or CFD or cash. Moreover, taxes and the chosen rebalancing strategy influence costs. Nevertheless, high average turnover per rebalancing leads to expensive strategy.

3.2.1 Data snooping

To avoid data snooping (White (2000)), the model confidence set (MCS) procedure proposed by Hansen et al. (2011) is computed. The MCS procedure is a model selection algorithm, which filters a set of models from a given entirety of models. The resulting set contains the best models with a probability that is no less than $1 - \alpha$ with α being the size of the test (see Hansen et al. (2011)).

An advantage of the test is that it not necessarily selects a single model, instead it acknowledges possible limitations in the data since the number of models in the set containing the best model will depend on how informative the data are.

More formally, define a set M_0 that contains the set of models under evaluation indexed by: $i = 0, ..., m_0$. Let $d_{i,j,t}$ denote the loss differential between two models by

$$d_{i,j,t} = L_{i,t} - L_{j,t}, \forall i, j \in M_0$$

L is the loss calculated from some loss function for each evaluation point $t = 1, ..., F$. The set of superior models is defined as:

$$M^* = \{i \in M_0 : E[d_{i,j,t}] \leq 0 \, \forall j \in M_0\}$$

The MCS uses a sequential testing procedure to determine M^*. The null hypothesis being tested is:

$$\begin{cases} H_{0,M} : E[d_{i,j,t}] = 0 \; \forall i, j \in M \text{ where } M \text{ is a subset of } M_0 \\ H_{A,M} : E[d_{i,j,t}] \neq 0 \text{ for some } i, j \in M \end{cases}$$

When the equivalence test rejects the null hypothesis, at least one model in the set M is considered inferior and the model that contributes the most to the rejection of the null is eliminated from the set M. This procedure is repeated until the null is accepted and the remaining models in M now equal $\widehat{M}^*_{1-\alpha}$.

According to Hansen et al. (2011), the following two statistics can be used for the sequential testing of the null hypothesis:

$$t_{i,j} = \frac{\overline{d}_{i,j}}{\sqrt{\widehat{var}(\overline{d}_{i,j})}} \text{ and } t_i = \frac{\overline{d}_i}{\sqrt{\widehat{var}(\overline{d}_i)}}$$

where m is the number of models in M, $\overline{d}_i = (m-1)^{-1} \sum_{j \in M} \overline{d}_{i,j}$, is the simple loss of the i^{th} model relative to the averages losses across models in the set M, and $\overline{d}_{i,j} = (m)^{-1} \sum_{t=1}^{m} d_{i,j,t}$ measures the relative sample loss between the i^{th} and i^{th} models. Since the distribution of the test statistic depends on unknown parameters a bootstrap procedure is used to estimate the distribution.

In this paper, the MCS is applied with profit maximization loss function ($\mathcal{ASR}$ and $\mathcal{CEQ}$). It should be noted that the MCS aims at finding the best model and all models which are indistinguishable from the best, not those better than a benchmark. To determined if models are better than a benchmark, the stepwise test of multiple reality check by Romano and Wolf (2005) and the stepwise multiple superior predictive ability test by Hsu et al. (2013) should be considered. However, a small trick is possible: if the benchmark is not selected in the best models set, investors can conclude that their strategies "beat" the benchmark.

4 Empirical results

4.1 S&P sectors

Table 1 highlights the interest of hierarchical clustering based portfolios, especially the DBHT based model. It is the only model selected in the best models set $\widehat{M}^*_{70\%}$ for both $\mathcal{ASR}$ and $\mathcal{CEQ}$. This portolio is a diversified ($\mathcal{SSPW}$=0.122), but the average turnover per rebalancing is elevated in comparison with other models ($\mathcal{TO}$= 3.45%).

The MV is included in $\widehat{M}^*_{\mathcal{ASR}-20\%}$ but its diversification ratio $\mathcal{SSPW}$ is by far the highest of all models: the portfolio is concentrated instead of being diversified.

CL belongs to $\widehat{M}^*_{\mathcal{CEQ}-20\%}$. The portfolio is diversified ($\mathcal{SSPW}$=0.114) and the turnover is quite low ($\mathcal{TO}$=0.817%).

Table 1: **Investment strategies comparison: S&P 500 sectors (January 1996-August 2016)**

	$\mathcal{ASR}$	$\mathcal{CEQ}$	$\mathcal{MDD}$	$\mathcal{TO}$	$\mathcal{SSPW}$
EW	0.422	6.81	54.2	-	0.100
MV	0.448*	5.75	37.8	1.78	0.480
MDP	0.344	5.28	57.6	1.83	0.217
ERC	0.442	6.40	51.6	0.294	0.107
IVRB	0.428	6.51	49.0	0.474	0.121
SL	0.418	6.53	53.4	0.669	0.115
CL	0.430	6.87*	51.0	0.817	0.114
AL	0.421	6.62	52.6	0.762	0.114
WM	0.415	6.49	52.9	0.883	0.149
DBHT	0.449**	7.43**	59.0	3.45	0.122

Note: This table reports comparison criteria used to evaluate the quality of the models: the Adjusted Sharpe Ratio ($\mathcal{ASR}$), the certainty-equivalent return ($\mathcal{CEQ}$) in percent, the Max drawdown ($\mathcal{MDD}$) in percent, the average turnover per rebalancing (($\mathcal{TO}$) in percent, the Sum of Squared Portfolio Weights ($\mathcal{SSPW}$). * and ** indicate the model is in the set of best models $\widehat{M}^*_{20\%}$ and $\widehat{M}^*_{70\%}$, respectively. EW refers to the equal weight allocation, MV refers to the minimum variance allocation, MDP refers to the most diversified portfolio allocation, ERC refers to the equal risk contribution allocation ,IVRB refers to the inverse-volatility risk budget allocation, SL refers to the simple linkage based allocation, CL refers to the complete linkage based allocation, AL refers to the average linkage based allocation, WM refers to the Ward's method based allocation, DBHT refers to the Directed Bubble Hierarchical Tree based allocation.

4.2 Multi-assets dataset

Table 2 paints a contrasting picture: risk-based portfolios achieve impressive $\mathcal{ASR}$ along with ridiculous low $\mathcal{CEQ}$. For instance, IVRB constitutes the best models set $\widehat{M}^*_{\mathcal{ASR}-70\%}$. To highlight the interest of the MCS, it is important to note that IVRB does not obtain the higher $\mathcal{ASR}$, yet it is the best model. Moreover, MDP and ERC are selected in $\widehat{M}^*_{\mathcal{ASR}-20\%}$. That said, risk-based portfolios attain very low $\mathcal{CEQ}$, especially IVRB ($\mathcal{CEQ}$=0.951). Above all, they do not produce diversified portfolios. It implies that portfolios are almost only invested in bonds, thereby being very exposed to shocks from this asset class. This is not what diversified portfolios aim at.

Hierarchical clustering based portfolios do not face the same problems. SL, SL and DBHT compose $\widehat{M}^*_{\mathcal{CEQ}-70\%}$, while delivering reasonably good $\mathcal{ASR}$. All portfolios are diversified and the average turnover per rebalancing is low for SL and SL. Again, DBHT's average turnover per rebalancing is elevated in comparison with other models.

Table 2: **Investment strategies comparison: multi-assets (February 1989-August 2016)**

	$\mathcal{ASR}$	$\mathcal{CEQ}$	$\mathcal{MDD}$	$\mathcal{TO}$	$\mathcal{SSPW}$
EW	0.601^{*}	4.02	24.9	-	0.0625
MV	0.611	1.31	7.49	4.22	0.403
MDP	0.717^{*}	1.95	7.7	2.92	0.296
ERC	0.707^{*}	1.91	9.34	0.509	0.164
IVRB	0.581^{**}	0.951	6.85	0.500	0.342
SL	0.597	4.67^{**}	31.4	1.28	0.086
CL	0.586	4.51	29.9	1.15	0.085
AL	0.602	4.71^{**}	29.7	1.19	0.085
WM	0.583	4.47	29.9	1.25	0.084
DBHT	0.525	4.87^{**}	25.4	4.35	0.081

Note: This table reports comparison criteria used to evaluate the quality of the models: the Adjusted Sharpe Ratio ($\mathcal{ASR}$), the certainty-equivalent return ($\mathcal{CEQ}$) in percent, the Max drawdown ($\mathcal{MDD}$) in percent, the average turnover per rebalancing (($\mathcal{TO}$) in percent, the Sum of Squared Portfolio Weights ($\mathcal{SSPW}$). * and ** indicate the model is in the set of best models $\widehat{M}^{*}_{20\%}$ and $\widehat{M}^{*}_{70\%}$, respectively. EW refers to the equal weight allocation, MV refers to the minimum variance allocation, MDP refers to the most diversified portfolio allocation, ERC refers to the equal risk contribution allocation ,IVRB refers to the inverse-volatility risk budget allocation, SL refers to the simple linkage based allocation, CL refers to the complete linkage based allocation, AL refers to the average linkage based allocation, WM refers to the Ward's method based allocation, DBHT refers to the Directed Bubble Hierarchical Tree based allocation.

4.3 Individual stocks

Table 3 points out that hierarchical clustering based portfolios outperform risk-based portfolios. Indeed, DBHT is the only model selected in the best models set $\widehat{M}^{*}_{\mathcal{ASR}-70\%}$ and the best models set $\widehat{M}^{*}_{\mathcal{CEQ}-70\%}$ is only constituted by one model: AL. Both portfolios are diversified.

The main drawback is the surprising elevated average turnover per rebalancing. This point needs to be further investigated, in particular, the impact of the criteria used to select the number of clusters and the consequences of the correlation matrix "shrinkage" on the stability of the clusters .

Table 3: **Investment strategies comparison: Individual stocks (January 1996-August 2016)**

	$\mathcal{ASR}$	$\mathcal{CEQ}$	$\mathcal{MDD}$	$\mathcal{TO}$	$\mathcal{SSPW}$
EW	0.595	13.3	52.2	-	0.0028
MV	0.600	13.0	51.2	0.021	0.0048
MDP	0.658	16.4	45.1	7.21	0.052
ERC	0.570	12.1	49.4	0.79	0.0036
IVRB	0.560	10.8	47.1	0.98	0.0045
SL	0.492	19.2*	43.1	32.4	0.0552
CL	0.473	16.5	47.4	33.7	0.0151
AL	0.520	19.5**	46.1	33.6	0.041
WM	0.572	14.2	51.2	33.4	0.0051
DBHT	0.591**	13.5	47.6	39.4	0.0056

Note: This table reports comparison criteria used to evaluate the quality of the models: the Adjusted Sharpe Ratio ($\mathcal{ASR}$), the certainty-equivalent return ($\mathcal{CEQ}$) in percent, the Max drawdown ($\mathcal{MDD}$) in percent, the average turnover per rebalancing (($\mathcal{TO}$) in percent, the Sum of Squared Portfolio Weights ($\mathcal{SSPW}$). * and ** indicate the model is in the set of best models $\widehat{M}^*_{20\%}$ and $\widehat{M}^*_{70\%}$, respectively. EW refers to the equal weight allocation, MV refers to the minimum variance allocation, MDP refers to the most diversified portfolio allocation, ERC refers to the equal risk contribution allocation ,IVRB refers to the inverse-volatility risk budget allocation, SL refers to the simple linkage based allocation, CL refers to the complete linkage based allocation, AL refers to the average linkage based allocation, WM refers to the Ward's method based allocation, DBHT refers to the Directed Bubble Hierarchical Tree based allocation.

Conclusion

Diversification is often spoken of as the only free lunch in investing. Yet, truly diversifying a portfolio is easier said than done. For instance, modern portfolio optimization techniques often fail to outperform a basic equal-weighted allocation (DeMiguel et al. (2009)).

Building upon the fundamental notion of hierarchy (Simon (1962)), López de Prado (2016a) introduces a new portfolio diversification technique called "Hierarchical Risk Parity", which uses graph theory and machine learning techniques.

Exploiting the same basic idea in a different way, a hierarchical clustering based asset allocation is proposed. Classical and more modern hierarchical clustering methods are tested, namely Simple Linkage, Complete Linkage, Average Linkage, Ward's Method and Directed Bubble Hierarchical Tree. Once the assets are hierarchically clustered, a simple and efficient capital allocation within and across clusters of investments at multiple hierarchical levels is computed. The main principle is to find a diversified weighting by distributing capital equally to each cluster hierarchy, so that many correlated assets receive the same total allocation as a single uncorrelated one.

The out-of-sample performances of hierarchical clustering based portfolios and more traditional risk-based portfolios are evaluated across three empirical datasets, which differ in terms of number of assets and composition of the universe ("S&P sectors", multi-assets and individual stocks). To prevent strategies that perform by luck to be considered as effective, the comparison of profit measures is assessed using the bootstrap based model confidence set procedure (Hansen et al. (2011)).

The empirical results point out that hierarchical clustering based portfolios are truly diversified and achieve statistically better risk-adjusted performances, as measured by the the Adjusted Sharpe Ratio (Pezier and White (2008)) and by the Certainty-Equivalent Return on all datasets. The only exception concerns the multi-assets dataset where risk-based portfolios produce impressive $\mathcal{ASR}$ along with ridiculous low $\mathcal{CEQ}$. Among clustering methods, there is no clear winner. DBHT based portfolios attain slightly superior risk-adjusted returns but AL based portfolios are clearly more robust.

Last but not least, this article opens the door for further research. Testing other clustering methods and investigating typical machine learning issues, such as the choice of the distance measure and the criteria used to select the number of clusters, come naturally to mind. Above all, improving the estimation of the correlation matrix seems to be the most important priority. Potential improvements may come from the use of "shrinkage" or/and the detection of current phase of the economic cycles.

References

Anderson, P. W. (1972). More is different. *Science*, 177(4047):393–396.

Aste, T., Shaw, W., and Matteo, T. D. (2010). Correlation structure and dynamics in volatile markets. *New Journal of Physics*, 12(8):085009.

Baitinger, E. and Papenbrocky, J. (2016). Interconnectedness risk and active portfolio management. Technical report.

Choueifaty, Y., Froidure, T., and Reynier, J. (Spring 2013). Properties of the most diversified portfolio. *Journal of Investment Strategies*.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953.

Friedman, M. (1992). Do Old Fallacies Ever Die? *Journal of Economic Literature*, 30(4):2129–32.

Friedman, M. (1997). Rx for japan: Back to the future. *Wall Street Journal*.

Gerber, S., Markowitz, H., and Pujara, P. (2015). Enhancing Multi-Asset Portfolio Construction Under Modern Portfolio Theory with a Robust Co-Movement Measure. *Social Science Research Network Working Paper Series*.

Goetzmann, W. N. and Kumar, A. (2008). Equity portfolio diversification. *Review of Finance*, 12(3):433–463.

Griveau-Billion, T., Richard, J.-C., and Roncalli, T. (2013). A fast algorithm for computing high-dimensional risk parity portfolios. Technical report.

Hansen, P., Lunde, A., and Nason, J. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Hsu, Y.-C., Kuan, C.-M., and Yen, M.-F. (2013). A generalized stepwise procedure with improved power for multiple inequalities testing. IEAS Working Paper : academic research 13-A001, Institute of Economics, Academia Sinica, Taipei, Taiwan.

Jobson, J. D. and Korkie, R. (1980). Estimation for Markowitz Efficient Portfolios. *Journal of the American Statistical Association*, 75.

Ledoit, O. and Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, 30(4):110–119.

Ledoit, O. and Wolf, M. (2014). Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks. *Available at SSRN 2383361*.

Levy, M. (2016). Measuring portfolio performance: Sharpe, alpha, or the geometric mean? Technical report.

López de Prado, M. (2016a). Building diversified portfolios that outperform out of sample. *The Journal of Portfolio Management*, 42(4):59–69.

López de Prado, M. (2016b). Mathematics and economics: A reality check. *Journalof Portfolio Management*, 43(1).

Maillard, S., Roncalli, T., and Teiletche, J. (2010). The properties of equally-weighted risk contribution portfolios. *The Journal of Portfolio Management*, 36(4):60–70.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B-Condensed Matter and Complex Systems*, 11(1):193–197.

Markowitz, H. (1952). Portfolio Selection. *Journal of Finance*, 7(1):77–91.

Michaud, R. (1989). The Markowitz Optimization Enigma: Is Optimized Optimal. *Financial Analysts Journal*.

Musmeci, N., Aste, T., and Di Matteo, T. (2015). Relation between financial market structure and the real economy: comparison between clustering methods. *PloS one*, 10(3):e0116201.

Pezier, J. and White, A. (2008). The relative merits of alternative investments in passive portfolios. *The Journal of Alternative Investments*.

Raffinot, T. (2014). A simple and efficient macroeconomic framework for asset allocation. Technical report.

Romano, J. P. and Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Roncalli, T. (2013). *Introduction to Risk Parity and Budgeting*. Chapman & Hall.

Simon, H. A. (1962). The architecture of complexity. In *Proceedings of the American Philosophical Society*, pages 467–482.

Song, W.-M., Di Matteo, T., and Aste, T. (2012). Hierarchical information clustering by means of topologically embedded graphs. *PLoS One*, 7(3):e31929.

Sun, W., Fan, A., Chen, L.-W., Schouwenaars, T., and Albota, M. A. (2006). Optimal rebalancing for institutional portfolios. *The Journal of Portfolio Management*, 32(2):33–43.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10421.

Tumminello, M., Coronnello, C., Lillo, F., Micciche, S., and Mantegna, R. N. (2007). Spanning trees and bootstrap reliability estimation in correlation-based networks. *International Journal of Bifurcation and Chaos*, 17(07):2319–2329.

Tumminello, M., Lillo, F., and Mantegna, R. N. (2010). Correlation, hierarchies, and networks in financial markets. *Journal of Economic Behavior & Organization*, 75(1):40–58.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

White, H. (2000). A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126.