Erschienen in: 18th IEEE International Conference On Machine Learning And Applications (ICMLA). - Piscataway : IEEE, 2019. - S. 1800-1805. - ISBN 978-1-72814-551-8
https://dx.doi.org/10.1109/ICMLA.2019.00291

Widened Learning of Index Tracking Portfolios

Iuliia Gavriushina, Oliver Sampson, Michael R. Berthold, Winfried Pohlmeier, Christian Borgelt
University of Konstanz, Germany
{iuliia.gavriushina; oliver.sampson; michael.berthold; winfried.pohlmeier; christian.borgelt}@uni-konstanz.de

***Abstract*—Index investing has an advantage over active investment strategies, because less frequent trades results in lower expenses, yielding higher long-term returns. *Index tracking* is a popular investment strategy that attempts to find a portfolio replicating the performance of a collection of investment vehicles. This paper considers index tracking from the perspective of solution space exploration. Three search space heuristics in combination with three portfolio tracking error methods are compared in order to select a tracking portfolio with returns that mimic a benchmark index. Experimental results conducted on real-world datasets show that *Widening*, a metaheuristic using diverse parallel search paths, finds superior solutions than those found by the reference heuristics. Presented here are the first results using *Widening* on time-series data.**

***Index Terms*—Widening, FinTech, algorithmic trading, machine learning, time-series, heuristic search, parallelism, diversity, index tracking, tracking portfolio**

I. Introduction

Wise investing is the key to building wealth. In the long-term, inflation can significantly degrade the value of cash savings, whereas investing in securities has the potential for higher returns. *Index investing* (*passive investing*) is a well-known investment approach that can outperform *active portfolio strategies*. Automating investment decisions (*algorithmic trading*) for decreasing investment costs is a popular research area in FinTech. Several studies show that many investors fail to outperform the market in the long-term, where a significant part of the loss is often related to the high fees associated with active trading [1]. According to the *efficient-market hypothesis*, a stock's price fully reflect all information available to the market [2]. Any new information about the market is rapidly incorporated into the stock's price. Therefore, to beat the market, an investor must obtain superior and/or faster information. Stock index tracking is a very popular strategy for those investors, who instead of trying to beat the market, aim to match the market's returns. The main advantage of index investing is lower expenses due to less frequent trading when compared to active investment strategies.

An *index* is a basket of securities which represents a whole market or submarket. Because it is not possible to buy an index itself directly, to "invest in an index," one needs to approximate its performance. There are two approaches to track an index: *physical* and *synthetic* replication. The first is realized by direct investment in the assets and aims to mimic the performance of the target index by holding all (*full replication*) or a representative sample (*sample replication*) of the underlying securities making up the index. Instead of physically holding the equities in the constituents of the benchmark, synthetic replication relies on derivatives which are linked to those equities.

When using an index replication approach, it is crucial that the reward is high enough to mitigate the investing risks. Sample replication is characterized by being both sufficiently transparent and sufficiently flexible. In contrast to syntetic replication, sample replication does not involve counterparty risk. Furthermore, it reduces expenses, because in contrast to full replication, the tracking portfolio does not require holding every asset in the index. Sample replication can be used if the index does not have an investable structure, the assets in the index are illiquid, or the number of assets is high.

Given an investment target $\boldsymbol{I}$, this work explores ways of selecting a *tracking portfolio*, $\boldsymbol{P}$, of assets with returns that mimic those of $\boldsymbol{I}$. We consider index tracking from the perspective of a common solution space exploration problem in machine learning, where the aim is to find a model (a tracking portfolio) which accurately represents a set of observed data (the valuation of an index over time) in order to predict its future performance. The goal is to find a minimum number of index constituents that can replicate a stock market index as closely as possible.

The number of possible solutions is prohibitively large for an index of even modest size. Various search heuristics can be used to limit the size of the search space, and portfolio tracking error methods can be used to score the models. Widening is an inherently parallel metaheuristic for use with greedy heuristics that uses diversity between solution paths followed by parallel workers. This results in regions of a solution space being explored that otherwise may not have been explored with the greedy heuristic alone, potentially finding superior solutions [3].

In this paper we demonstrate for the first time the use of Widening on time-series data. We evaluate three search space heuristics (Hill-Climbing Index Tracking, Top-k Index Tracking, and Top-k Widened Index Tracking) and show that using parallel resources and diversity yields better tracking portfolios.

II. Related work

To find a tracking portfolio using sample replication, two problems must be solved simultaneously: selecting a subset of assets from the index and calculating their optimal corresponding investment weights. The selection of assets for the tracking portfolio is NP-hard [4]. When an exhaustive search is computationally intractable, *heuristics* are used to limit the *search space*, by sacrificing optimality for operationality. The

most straightforward local search algorithm, which has been used for the index-tracking problem, is HILL-CLIMBING [5]. It is a greedy evaluation heuristic which selects the best index member in each round until a certain criterion is met. Often a greedy search does not produce an optimal solution, but nevertheless, a greedy heuristic may yield locally optimal solutions that approximate a global optimal solution in a reasonable time [6].

To track an index, the search heuristics are used in combination with different tracking error methods that find investment weights $\boldsymbol{\omega}$, $|\boldsymbol{\omega}| = |\boldsymbol{P}|$, for the assets in portfolio $\boldsymbol{P}$, where the calculation of the weights is a multivariate optimization problem of size $|\boldsymbol{P}|$. There are a lot of *tracking measures* that are used in the objective function of portfolio optimization methods for the index replication. The most popular are tracking error variance (TEV), mean squared error (MSE), and mean absolute error (MAE). MAE is more robust to outliers than MSE but significantly more computationally complex. Additionally, MSE is characterized by faster convergence, is continuously differentiable and, therefore, allows for gradient-based methods [7].

In the literature various methods are found to restrict the overall search space, such as limiting the size of the tracking portfolio (cardinality constraint) and specifying the asset classes (class constraint), but there is no specific algorithm best-suited for finding a tracking portfolio [1], [5].

III. SAMPLE REPLICATION HEURISTICS

Given an index $\boldsymbol{I}$ with constituent assets I_i, $i \in \{1, \ldots, |\boldsymbol{I}|\}$, the set of all possible portfolios irrespective of the corresponding weights is $\Theta = \mathbb{P}(\boldsymbol{I}) \setminus \emptyset$, where $\mathbb{P}(\boldsymbol{I})$ is the powerset of $\boldsymbol{I}$. According to the sample replication, a tracking portfolio $\boldsymbol{P} \subseteq \boldsymbol{I}$ with constituent assets P_j, where $j \in \{1, \ldots, |\boldsymbol{P}|\}$.

A. HILL-CLIMBING INDEX TRACKING

HILL-CLIMBING is a type of breadth-first search, which can be presented as an iterative application of a *refinement operator,* $r(\cdot)$, and a *selection operator,* $s(\cdot)$.

Definition 1. *A **refinement operator**, $r(\cdot)$, generates a set of tracking portfolios (called refinements) by adding one asset from $\boldsymbol{I} \setminus \boldsymbol{P}$ to $\boldsymbol{P}$ [3].*

$$\mathcal{P} = r(\boldsymbol{P}) \tag{1}$$

where $\mathcal{P} \subseteq \Theta$ is a set of unique portfolios from Θ.

Definition 2. *A **selection operator**, $s(\cdot)$, chooses the best tracking portfolio $\boldsymbol{P}'$ from a set of portfolios $\mathcal{P}$ [3]:*

$$\boldsymbol{P}' = s(\mathcal{P}) = s(r(\boldsymbol{P})) \tag{2}$$

At each iteration, a refinement operator returns a new set of tracking portfolios. Using a tracking quality measure, we score all of the new refinements, and select the best portfolio $\boldsymbol{P}' \in \mathcal{P}$ with the selection operator $s(\cdot)$. The scoring function used in this work is the mean squared error (MSE). The number of potential refinements decreases with each iteration, because the number of remaining assets decreases. The stop criterion for HILL-CLIMBING INDEX TRACKING is when the addition of a new asset no longer improves the MSE or all of the assets in $\boldsymbol{I}$ have been added to $\boldsymbol{P}$.

B. TOP-k INDEX TRACKING

The drawback to the HILL-CLIMBING algorithm is the likelihood of becoming stuck at a local optimum of poor quality. TOP-k search is a common heuristic for use with prohibitively large search spaces that impoves on the results of HILL-CLIMBING [9]. It evaluates all models refined at a particular iteration, and keeps k, the *search width*, of the best performing models for further refinement. The rest of the models are discarded and no longer explored. TOP-k search iteratively explores k solution paths in parallel. The larger the search width k, the more likely it is to find a better solution.

Similar to the HILL-CLIMBING algorithm, TOP-k INDEX TRACKING can be presented using the refine-and-select process. The difference is that after a model is refined, the selection operator, $s(\cdot)$, is modified to choose the best k portfolios at each step until a stopping criterion met:

$$\mathcal{P}' = s_{Top-k}(\mathcal{P}) = s_{Top-k}(\{r(\boldsymbol{P}_1), \ldots, r(\boldsymbol{P}_{|\mathcal{P}|})\}) \tag{3}$$

where $\mathcal{P}' = \{\boldsymbol{P}_1', \ldots, \boldsymbol{P}_k'\}$. TOP-$k$ INDEX TRACKING stops when either none of the k models improves or all of the assets from $\boldsymbol{I}$ are included in one of the k portfolios.

TOP-k search is not optimal; there is no guarantee that it will find the best solution. The choice of k best solutions at each iteration based on the quality measure does not ensure exploration of *different* regions of the search space. In contrast, it is likely that we are exploring only closely related variations of the locally best model [9].

C. Sets of Diverse Portfolios

The general problem of selecting diverse subset of elements from a larger set, where some distance between each pair of points is maximized known as the *p-dispersion problem* [10]. Diversity is an important issue in bio- and chem-informatics and has been studied regarding protein and molecular similarity in [11]. In data mining the effect of diversity on the parallel exploration of the solution space was studied in [12]–[16].

There are several diversity measures commonly used in the *p-dispersion problem*. The most straightforward is *p-dispersion-sum*. Sampson and Berthold in [12], [13] suggest using *p-dispersion-min-sum* because it shows a tendency towards a more representative subset. In this paper we explore both diversity measures for the index tracking.

Definition 3. *Given a set of portfolios, $\mathcal{P}$, the **p-dispersion-sum** problem is defined as the selection of the set $\hat{\mathcal{P}}$ with $|\hat{\mathcal{P}}| \leq |\mathcal{P}|$ items by maximizing the sum of all pairwise distances $d(\boldsymbol{P}_i, \boldsymbol{P}_j)$ with $\boldsymbol{P}_i, \boldsymbol{P}_j \in \mathcal{P}$ [11]:*

$$\hat{\mathcal{P}} = \underset{|\hat{\mathcal{P}}|=k}{\operatorname{argmax}} \sum_{i=1}^{|\mathcal{P}|} \sum_{j=1}^{i-1} d(\boldsymbol{P}_i, \boldsymbol{P}_j) \tag{4}$$

The *p-dispersion-sum* measure provides a subset whose elements located maximally far away from each other and

are concentrated at the edges of the data space. In contrast, *p-dispersion-min-sum* provides a subset with the largest sum of minimum distances between each pair of elements in the set, resulting in an even distribution of selected points over the whole space, e.g. in a better coverage of the space.

Definition 4. *Given a set of portfolios, $\mathcal{P}$, the **p-dispersion-min-sum** problem is defined as selection of the set $\hat{\mathcal{P}}$ with $|\hat{\mathcal{P}}| \leq |\mathcal{P}|$ items by maximizing the sum of minimal distances $d(\boldsymbol{P}_i, \boldsymbol{P}_j)$ with $\boldsymbol{P}_i, \boldsymbol{P}_j \in \mathcal{P}$ [11]:*

$$\hat{\mathcal{P}} = \underset{|\hat{\mathcal{P}}|=k}{\operatorname{argmax}} \sum_{j=1}^{|\boldsymbol{\mathcal{P}}|} \min_{1 \leq j \leq |\mathcal{P}|, i \neq j} d(\boldsymbol{P}_i, \boldsymbol{P}_j) \tag{5}$$

Finding a distance metric $d(\boldsymbol{P}_i, \boldsymbol{P}_j)$ which is able to describe portfolios' dissimilarity is not commonly found in the literature, where the main focus is diversification within a portfolio. Goetzmann and Kumar in [17] use the sum of squared portfolio weights (SSPW) to measure the diversification of retail investors:

$$SSPW(\boldsymbol{P}) = \sum_{j=1}^{|\boldsymbol{P}|} (\omega_j - \Omega)^2 = \sum_{j=1}^{|\boldsymbol{P}|} (\omega_j - \frac{1}{|\boldsymbol{I}|})^2 \approx \sum_{j=1}^{|\boldsymbol{P}|} \omega_j^2 \tag{6}$$

where ω_j is the weight of security P_j with $j \in \{1, \ldots, |\boldsymbol{P}|\}$, $\omega_j \neq 0$, and $\Omega = \frac{1}{|\boldsymbol{I}|}$ is the normalized weight assigned to each security in the index.

To track an index, Focardi and Fabozzi in [18] suggest using Euclidean distances between stock price series as a basis for hierarchical clustering to obtain a diversified portfolio. With the same aim, several studies consider the Pearson distance between stock prices [18], [19].

In preliminary experiments, we evaluated different portfolio characteristics such as SSPW, portfolio variation, Sharpe ratio, highest/lowest portfolio returns with several distance measures such as Pearson and Euclidean distances. The experiments using SSPW with the Euclidean distance (see (7)) lead the best results and are presented in this paper.

$$d_{\ell_2}(SSPW(\boldsymbol{P}_i), SSPW(\boldsymbol{P}_j)) = \sqrt{(SSPW(\boldsymbol{P}_i) - SSPW(\boldsymbol{P}_j))^2} \tag{7}$$

D. Top-k Widened Index Tracking

The iterative refine-and-select process of the Top-k Widened Index Tracking modifies Top-k Index Tracking by including a diversity measure, δ, when refining models:

$$\mathcal{P}' = s_{Top-k}(r_\delta(\mathcal{P})) \tag{8}$$

The diversity measure δ describes differences between the resulting refined models and the selection operator $s_{Top-k}(\cdot)$ chooses the k best portfolios at each step until a stopping criterion met. This ensures exploration of different regions of the solution space. In this paper we compare both *p-dispersion-sum* (see Definition 3) and *p-dispersion-min-sum* (see Definition 4) for use as the diversity measure in $r_\delta(\cdot)$.

IV. Portfolio Tracking Error Methods

Evaluating portfolio models at each step in the refine-and-select process requires an appropriate scoring method, where the optimal weights for a portfolio are calculated to minimize the tracking error relative to the target index. Denote the return of an asset $I_i \in \boldsymbol{I}$ with $i \in \{1, \ldots, |\boldsymbol{I}|\}$ during a trading period $t \in \{1, \ldots, T\}$ by $\eta_t(I_i)$. Then the return of an asset I_i during a studied time interval is $\boldsymbol{\eta}(I_i) = [\eta_1(I_i), \ldots, \eta_T(I_i)]^\top$. Let $P_j \in \boldsymbol{P}$ with $j \in \{1, \ldots, |\boldsymbol{P}|\}$ be a component asset of $\boldsymbol{P}$ with $\boldsymbol{\eta}(P_j) = [\eta_1(P_j), \ldots, \eta_T(P_j)]^\top$. A portfolio's return is $\boldsymbol{\eta}(\boldsymbol{P}) = \sum_{j=1}^{|\boldsymbol{P}|} \omega_j \boldsymbol{\eta}(P_j)$.

Portfolio $\boldsymbol{P}$ reproduces the performance of the benchmark index $\boldsymbol{I}$ if the portfolio return $\boldsymbol{\eta}(\boldsymbol{P})$ follows the return of the index $\boldsymbol{\eta}(\boldsymbol{I})$ at every unit time period closely. Let tracking error $\boldsymbol{TE}(\boldsymbol{I}, \boldsymbol{P}) = \boldsymbol{\eta}(\boldsymbol{I}) - \boldsymbol{\eta}(\boldsymbol{P}) = \boldsymbol{\eta}(\boldsymbol{I}) - \sum_{j=1}^{|\boldsymbol{P}|} \omega_j \boldsymbol{\eta}(P_j)$. To find the optimal investing weights, $\boldsymbol{\omega}^*$, for $\boldsymbol{P}$, the general optimization method for the index tracking problem is formulated as the minimization of a tracking quality measure based on the tracking error $\boldsymbol{TE}$ [20]:

$$\begin{aligned} \boldsymbol{\omega}^* &= \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(f(\boldsymbol{TE})) \\ &\sum_{j=1}^{|\boldsymbol{P}|} \omega_j = 1 \end{aligned} \tag{9}$$

where $\sum_{j=1}^{|\boldsymbol{P}|} \omega_j = 1$ is a typical constraint which implies that all of the available budget must be invested, and $f(\boldsymbol{TE})$ is a tracking quality measure.

A. Tracking-Error Variance Minimization

One of the most popular tracking quality measures is the tracking error variance (TEV) [5], [21]:

$$TEV = Var(\boldsymbol{TE}) = \sigma_{\boldsymbol{P}}^2 + \sigma_{\boldsymbol{I}}^2 - 2\sigma_{\boldsymbol{I}}^2 \beta_{\boldsymbol{P}}, \tag{10}$$

where $\sigma_{\boldsymbol{P}}^2$ is the portfolio variance, $\sigma_{\boldsymbol{I}}^2$ is the index variance, and $\beta_{\boldsymbol{P}} = \sum_{j=1}^{p} \beta_j \omega_j = \boldsymbol{\beta}^\top \boldsymbol{\omega}$ is the portfolio beta, which measures portfolio volatility relative to the index.

With TEV the goal is to find optimal investment weights, $\boldsymbol{\omega}^*$, of the portfolio, $\boldsymbol{P}$ that minimizes the variance of the difference between $\boldsymbol{\eta}(\boldsymbol{P})$ and $\boldsymbol{\eta}(\boldsymbol{I})$. Because the index variance $\sigma_{\boldsymbol{I}}^2$ is independent of portfolio positions:

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(TEV) = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega} - 2\sigma_{\boldsymbol{I}}^2 \boldsymbol{\beta}^\top \boldsymbol{\omega}) \tag{11}$$

where $\sum_{j=1}^{|\boldsymbol{P}|} \omega_j = 1$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the assets' returns.

When trying to obtain a precise solution with specific characteristics, one can impose a variety of constraints on the portfolio optimization formulations. The Mean Constrained TEV method (TEMV) allows to specify the target expected returns of the portfolio [21]:

$$\begin{aligned} \boldsymbol{\omega}^* &= \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(TEV) = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(\boldsymbol{\omega}^\top \boldsymbol{\Sigma} \boldsymbol{\omega} \ - \ 2\sigma_{\boldsymbol{I}}^2 \boldsymbol{\beta}^\top \boldsymbol{\omega}) \\ &\boldsymbol{\mu}^\top \boldsymbol{\omega} = r_0 \end{aligned} \tag{12}$$

TABLE I: Out-of-sample statistics for the tested datasets. The search width for TOP-k INDEX TRACKING and TOP-k WIDENED INDEX TRACKING is k = 5. The smallest average MSE between the index and portfolio returns for each heuristic appears in bold. Underlined text shows the minimum average number of assets included in the best tracking portfolio.

Index	**Method**	HILL-CLIMBING		TOP-k		TOP-k WIDENING			
						p-dispersion-sum		*p-dispersion-min-sum*	
	Model	$\|\boldsymbol{P}\|$	MSE $\times 10^4$	$\|\boldsymbol{P}\|$	MSE $\times 10^4$	$\|\boldsymbol{P}\|$	MSE $\times 10^4$	$\|\boldsymbol{P}\|$	MSE $\times 10^4$
STOXX50 ($\|\boldsymbol{I}\| = 50$)	TEV	25.27	0.0168	28.36	0.0193	44.73	0.0045	44.91	0.0055
	MSE	25.36	**0.0161**	29.55	**0.0135**	44.64	**0.0045**	<u>47.45</u>	**0.0032**
	TEMV	22.55	0.0662	27.09	0.0198	42.00	0.0092	35.36	0.0570
NASDAQ ($\|\boldsymbol{I}\| = 101$)	TEV	23.91	**0.0483**	30.36	**0.0282**	<u>65.18</u>	**0.0070**	55.64	**0.0112**
	MSE	26.18	0.0487	29.36	0.0292	51.27	0.0108	50.55	0.0116
	TEMV	28.91	0.0387	30.73	0.0390	49.27	0.0230	52.64	0.0206
S&P500 ($\|\boldsymbol{I}\| = 495$)	TEV	33.27	0.0813	35.64	0.0673	61.00	0.0132	57.20	**0.0154**
	MSE	31.82	**0.0763**	37.17	**0.0616**	<u>59.60</u>	**0.0124**	53.80	0.0174
	TEMV	28.91	0.0958	37.55	0.0689	56.00	0.0175	51.60	0.0207

where $\sum_{j=1}^{|\boldsymbol{P}|} \omega_j = 1$, $\boldsymbol{\mu}^\top$ is a vector of the expected returns of the assets with $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_{|\boldsymbol{P}|}]^\top$, and r_0 is a desirable expected return specified by the investor, which in this work is equal to the index mean return.

B. Mean Squared Error Minimization

Although TEV is commonly used as a tracking measure, it has a disadvantage: if the difference between portfolio and index returns is constant over time, $Var(\boldsymbol{TE}) = 0$, there could be still a deviation from the index. Therefore, another popular tracking-error-based measure, mean squared error (MSE), is often used instead:

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\eta_t(\boldsymbol{I}) - \eta_t(\boldsymbol{P}))^2 = \frac{1}{T}\sum_{t=1}^{T}(TE_t)^2 \quad (13)$$

In this case the portfolio tracking error method has the following representation [4]:

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(MSE) = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}(\boldsymbol{\omega}^\top \mathbf{H}\boldsymbol{\omega} - 2\boldsymbol{q}^\top\boldsymbol{\omega}) \quad (14)$$

where $\sum_{j=1}^{|\boldsymbol{P}|}\omega_j = 1$, $\mathbf{H}$ is the matrix with $H_{ji} = \frac{1}{T}\sum_{t=1}^{T}\eta_t(P_j)\eta_t(P_i)$, and $\boldsymbol{q}$ with $q_j = \frac{1}{T}\sum_{t=1}^{T}\eta_t(P_j)\eta_t(\boldsymbol{I})$ for $j, i \in \{1, \ldots, |\boldsymbol{P}|\}$.

V. EXPERIMENTAL RESULTS

In this work we evaluate three search heuristics (HILL-CLIMBING INDEX TRACKING, TOP-k INDEX TRACKING, and TOP-k WIDENED INDEX TRACKING) in combination with three tracking-error-based methods (minimization of TEV, TEMV, and MSE).

Index	**Area**	$\boldsymbol{n}$	$\boldsymbol{n^*}$
STOXX50	Eurozone	50	50
NASDAQ	USA	103	101
S&P500	USA	505	495

TABLE II: Summary of the replicated indices. The number of index constituents n is taken based on 31.01.2018. n^* denotes the number of index constituents after data preprocessing.

The presented algorithms are implemented in KNIME v3.5 [22] and tested on real-world market indices of different sizes (see Table II). Daily returns are obtained for 01.01.15–31.01.18 from *Thomson Reuters Datastream*.[1] The stocks with missing historical data from the beginning of the testing periods are excluded. The one-period daily returns (or simple returns) for the index $\boldsymbol{I}$ and each component of the index I_i are calculated:

$$\eta_t^*(\boldsymbol{I}) = \frac{\eta_{t+1}(\boldsymbol{I}) - \eta_t(\boldsymbol{I})}{\eta_t(\boldsymbol{I})},\ \eta_t^*(I_i) = \frac{\eta_{t+1}(I_i) - \eta_t(I_i)}{r_t(I_i)} \quad (15)$$

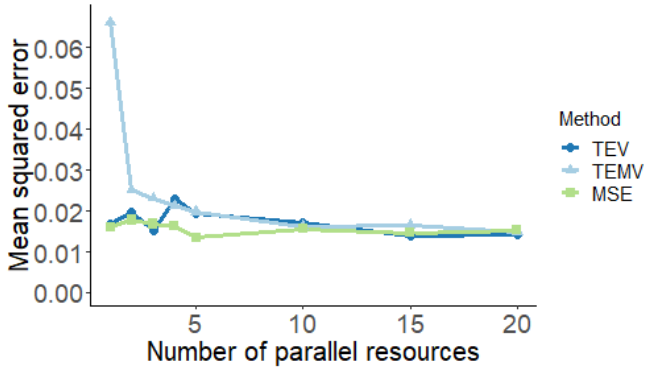
Evaluation of the datasets is configured to mimic real-world quarterly reviews by creating a test partition of the data ($H = 60$) of approximately three months' trading days. Correspondingly, the training partition is set to be 440 to create a ($T = 500$) total train/test. Additionally a rolling window of step size ($h = 20$) is used to verify reproducability.

A. Heuristics' Comparison

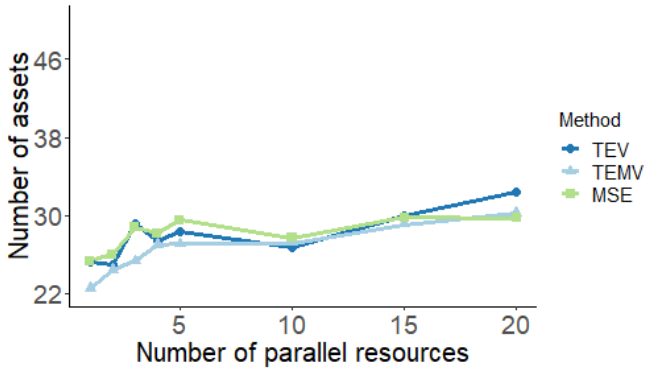
The experimental results show that both TOP-k and TOP-k WIDENING provides better index replication than that of the reference heuristic HILL-CLIMBING (see Table I). Adding diversity to the set of parallel search paths allows broader exploration of the solution space and finds a better solution for each tracking-error-based method. TOP-k WIDENED INDEX TRACKING with width $k = 5$ provides a tracking portfolio with the smallest MSE. However, the number of assets in the portfolio increases. In contrast to the results of [12], we found no discernible difference between the *p-dispersion-sum* and *p-dispersion-min-sum*; this may be due to the multiobjective optimization nature of the problem. The use of TEV and MSE methods shows very similar results for each heuristic, whereas minimization of TEMV demonstrates the worst performance among suggested tracking error methods.

When replicating the STOXX50 index, TOP-k INDEX TRACKING with width $k = 5$ used with TEV minimization method was not able to find a better tracking portfolio than that obtained by HILL-CLIMBING. Additionally, for this index,

[1] https://eikon.thomsonreuters.com

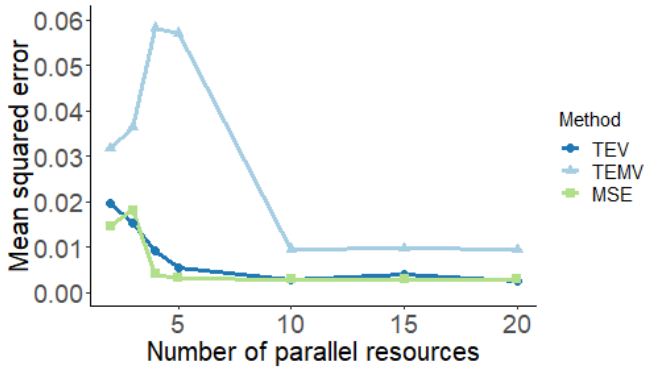


(a) Increasing the search width k decreases MSE $\times 10^4$.

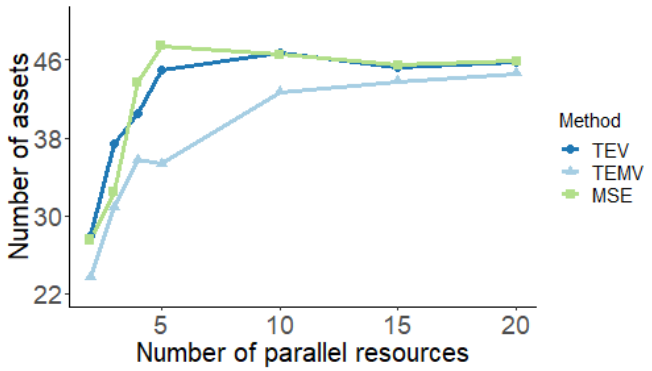


(b) The number of assets in the tracking portfolio increases while increasing the search width k.

Fig. 1: Replication of the STOXX50 with TOP-k INDEX TRACKING.



(a) Increasing the search width k decreases MSE $\times 10^4$.



(b) The number of assets in the tracking portfolio grows while increasing the search width k.

Fig. 2: Replication of the STOXX50 with TOP-k WIDENING using *p-dispersion-min-sum* measure.

TOP-k WIDENED INDEX TRACKING used with TEMV minimization method and *p-dispersion-min-sum* measure provides worse results than that obtained by TOP-k INDEX TRACKING. Therefore, with the example of STOXX50, we demonstrate that increasing the number of parallel resources provides better index replication.

B. The Effect of Width

Fig. 1 and Fig. 2 show that a larger search width k in TOP-k INDEX TRACKING and TOP-k WIDENED INDEX TRACKING allows finding a portfolio with the smaller tracking error. However, the number of assets in the tracking portfolio has general tendency to increase with the wider search. This may be a result of a lack of any restrictions on tracking error convergence, i.e., we aimed to get the smallest possible error.

For the TOP-k WIDENED INDEX TRACKING with the width $k = 15$ the number of assets for the TEV and MSE methods decreases, whereas the difference between an index and portfolio returns increases insignificantly. This shows that broader search allows TOP-k WIDENED INDEX TRACKING not only finding a tracking portfolio with the smaller tracking error than that obtained by the other two methods, but also finding a smaller number of assets for the portfolio.

VI. CONCLUSION AND FUTURE WORK

This paper explores methods of heuristically choosing an index tracking portfolio. We demonstrate that WIDENING, which was used in this work with time series for index tracking for the first time, when applied to a greedy algorithm is able to find superior solutions than that algorithm would have found alone.

The experimental results show that it is possible to find a smaller set of assets which is able to mimic the index performance within some error range. Presented here are indices of sizes 50 to 500. Other source indices of even larger sizes (e.g., Russell 3000 or Wilshire 5000) could be used, but would require significantly larger computing resources.

We show that the use of parallel resources for the index replication allows finding better tracking portfolios than the most straightforward HILL-CLIMBING algorithm. Adding diversity to the parallel search paths explores otherwise unexplored regions of the solution space and improves the results

obtained by a greedy search heuristic. Future work includes finding a better diversity measure which is able to provide wider exploration of the portfolio solution space. One could use different portfolio optimization models and different portfolio measurements which may yield better results. The same portfolio could be evaluated with different sets of investment weights resulting in a diverging distances and correspondingly affecting the diversity measures. Additionally, it would be interesting to find out how non-traditional metrics, such as the number of women on corporate boards used as a distance metric or as a component of a composite measurement [23], can represent out-of-channel asset performance.

Increasing the width of a search explores more of the solution space and finds a better solution. However, we can expect that at some point wider exploration of the solutions space will show diminishing returns. As shown by these experiments, increasing the width of a search often increases the number of assets in the portfolio while only marginally reducing the tracking error. Future work should specifically explore the Pareto front for this multi-objective optimization problem and explore the number of parallel paths after which no further improvement in performance is seen.

In this work we constructed a tracking portfolio from the set of index constituents. However, different combination are possible. One can use a subset of the index components or choose completely arbitrary set of assets. Future work could explore whether it is still possible to replicate an index. Furthermore, future work can explore whether it is possible to track different types of indices, such as those of commodities.

This paper focuses on the tracking portfolio construction i.e., on single-stage index tracking, and demonstrates positive results with the metaheuristic, WIDENING. The other important part of the index replication problem is *portfolio rebalancing* that allows incorporating transaction costs associated with trading (embedded after the creation phase) and requires development of a rebalancing strategy. The weights of the securities in indices change over time. Because indices are not static, index providers make a regular check of the index components, which takes place at fixed time intervals determined in the index rules. If any changes are made to the index, for example, when it is rebalanced or reconstituted, a tracking portfolio has to adjust its holdings accordingly. The influence of transaction costs on portfolio performance in practice can be significant. If the transaction costs are not considered during rebalancing, the performance of the portfolio could be poor [24]. One can analyze the difference between the index and portfolio returns over time and make portfolio rebalancing when it is reasonable. Future work should include developing a rebalancing strategy and embedding it in TOP-k and TOP-k WIDENING for the multiperiod index tracking.

This first demonstration of WIDENING on time-series data can also be extended to other time-series optimization problems, including other portfolio management optimization objectives, e.g., instead of matching returns, maximizing returns, minimizing risk, or multi-objective optimization problems matching individual investors' profiles could be addressed.

REFERENCES

[1] D. Maringer and O. Oyewumi, "Index tracking with constrained portfolios," *Intelligent Systems in Accounting, Finance and Management*, vol. 15, no. 1-2, pp. 57–71, 2007.

[2] B. G. Malkiel and E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[3] Z. Akbar, V. N. Ivanova, and M. R. Berthold, "Parallel data mining revisited. better, not faster," in *International Symposium on Intelligent Data Analysis*. Springer, 2012, pp. 23–34.

[4] R. Ruiz-Torrubiano and A. Suárez, "A hybrid optimization approach to index tracking," *Annals of Operations Research*, vol. 166, no. 1, pp. 57–71, 2009.

[5] U. Derigs and N.-H. Nickel, "On a local-search heuristic for a class of tracking error minimization problems in portfolio management," *Annals of Operations Research*, vol. 131, no. 1-4, pp. 45–77, 2004.

[6] S. Edelkamp and S. Schroedl, *Heuristic search: theory and applications*. Elsevier, 2011.

[7] A. Preminger and R. Franck, "Forecasting exchange rates: A robust regression approach," *International Journal of Forecasting*, vol. 23, no. 1, pp. 71–84, 2007.

[8] A. Goel, A. Sharma, and A. Mehra, "Index tracking and enhanced indexing using mixed conditional value-at-risk," *Journal of Computational and Applied Mathematics*, vol. 335, pp. 361–380, 2018.

[9] V. N. Ivanova and M. R. Berthold, "Diversity-driven widening," in *International Symposium on Intelligent Data Analysis*. Springer, 2013, pp. 223–236.

[10] E. Erkut, "The discrete p-dispersion problem," *European Journal of Operational Research*, vol. 46, no. 1, pp. 48–60, 1990.

[11] T. Meinl, C. Ostermann, and M. R. Berthold, "Maximum-score diversity selection for early drug discovery," *Journal of chemical information and modeling*, vol. 51, no. 2, pp. 237–247, 2011.

[12] O. Sampson and M. R. Berthold, "Widened KRIMP: better performance through diverse parallelism," in *International Symposium on Intelligent Data Analysis*. Springer, 2014, pp. 276–285.

[13] O. R. Sampson and M. R. Berthold, "Widened learning of bayesian network classifiers," in *International Symposium on Intelligent Data Analysis*. Springer, 2016, pp. 215–225.

[14] O. R. Sampson, C. Borgelt, and M. R. Berthold, "Communication-free widened learning of Bayesian network classifiers using hashed Fiedler vectors," in *Advances in Intelligent Data Analysis XVII*. Springer, October 2018.

[15] A. Fillbrunn and M. R. Berthold, "Diversity-driven widening of hierarchical agglomerative clustering," in *International Symposium on Intelligent Data Analysis*. Springer, 2015, pp. 84–94.

[16] A. Fillbrunn, L. Wörteler, M. Grossniklaus, and M. R. Berthold, "Bucket selection: A model-independent diverse selection strategy for widening," in *International Symposium on Intelligent Data Analysis*. Springer, 2017, pp. 87–98.

[17] W. N. Goetzmann and A. Kumar, "Equity portfolio diversification," *Review of Finance*, vol. 12, no. 3, pp. 433–463, 2008.

[18] S. M. Focardi and F. J. Fabozzi, "A methodology for index tracking based on time-series clustering," *Quantitative Finance*, vol. 4, no. 4, pp. 417–425, 2004.

[19] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, and A. Kanto, "Dynamics of market correlations: Taxonomy and portfolio analysis," *Physical Review E*, vol. 68, no. 5, p. 056110, 2003.

[20] T. F. Coleman and Y. Li, "Minimizing tracking error while restricting the number of assets," *The Journal of Risk*, vol. 8, no. 4, pp. 33–55, 2006.

[21] N. C. Edirisinghe, "Index-tracking optimal portfolio selection," *Quantitative Finance Letters*, vol. 1, no. 1, pp. 16–20, 2013.

[22] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[23] M. Noland, T. Moran, and B. R. Kotschwar, "Is gender diversity profitable? evidence from a global survey," *Peterson Institute for International Economics Working Paper*, no. 16-3, 2016.

[24] F. J. Fabozzi and D. A. Pachamanova, *Portfolio construction and analytics*. John Wiley & Sons, 2016.