Mateusz Andrzejewski[a], *Piotr Wójcik*[b]
University of Warsaw, Faculty of Economic Sciences
00-241 Warszawa, ul. Długa 44/50
[a] mat.jan.andrzejewski@gmail.com
[b] pwojcik@wne.uw.edu.pl, ORCID: 0000-0003-1853-8784

Sparse Tracking of Stock Indices with Principal Component Analysis and LASSO

1. Introduction

Exchange Traded Funds, also known as ETFs, form a class of open-end investment funds distinguished by the fact that their shares are publicly listed and traded on stock exchanges in the same way as shares in publicly traded companies (Investment Company Institute 2022). This feature differentiates ETFs from mutual funds, whose shares are only traded at one moment during the day, usually by the time markets close and Net Asset Value of holdings is calculated. ETFs were initially developed in the United States to track performance of specific stock indices, such as S&P 500 (tracked by the very first ETF in the world, approved by the Securities and Exchanges Commission in 1993) or NASDAQ; however, the industry has since expanded into other asset classes, including bonds, commodities, and currencies (Investment Company Institute 2022). Additionally, investors may now access active ETFs – customised indices which aim to track specific investment themes (such as robotics, fintech, green energy) or segments of the market such as growth, value, high or low volatility stocks, etc. (Davis 2023).

ETFs offer a range of advantages over mutual funds, including lower costs (Investment Company Institute 2022), diversification benefits and favourable tax treatment (Culloton 2006) as well as increased transparency (Investment Company Institute 2022). As a result, passive investment is reported to yield higher returns for investors (Rompotis 2009), and it has enjoyed dynamic growth in both the number of ETFs and their Assets Under Management: by the end of 2021, $10.1 trillion had been invested in ETFs worldwide, with the majority ($7.2 trillion) in the US market, constituting 21% of all institutionally invested funds on the largest stock market in the world (Investment Company Institute 2022). In Europe, however, while growth has also been dynamic, achieving the Compound Annual Growth Rate of 16% between 2010 and 2020 (O'Dwyer 2020), the share of exchange-traded products remains much lower, at ca. 7% in 2022 (Gordon 2022). Polish market remains relatively underdeveloped as well, with only 11 ETFs available on the Warsaw Stock Exchange (GPW 2022), mostly due to cumbersome administrative procedures required both at the point of fund origination and on an ongoing basis while the fund is operational (Żuławiński 2022).

Since the majority of ETFs attempt to replicate the behaviour of a specific index, they are constructed via either physical or synthetic replication of its returns (Corsi et al. 2020). Synthetic ETFs replicate the index via swap agreements, while physical replication can be achieved either by holding all the individual components in the portfolio, which is most feasible in the case of funds tracking indices with large capitalisation, by liquid stocks, or by creating a representative sample. The latter approach, in the case of equities, is used when the underlying index consists of a large number of stocks (such as S&P Global BMI, including over 14,000 stocks [S&P Dow Jones Indices 2023a]), or when its components are relatively small and infrequently traded (such as S&P Global SmallCap, focused the lowest 15% of market cap in each country [S&P Dow Jones Indices 2023b]). In addition, this approach is often selected for bond ETFs, where underlying indices may contain thousands of bonds, many of which are held by investors until maturity, which consequently limits their liquidity (Corsi et al. 2020).

In order to achieve a representative sample, two main solutions are employed: stratified sampling and optimisation (Corsi et al. 2020). Stratified sampling, which allows for a degree of managerial discretion with allocation decisions, is based on dividing the index into a number of groups (strata) based on a selected characteristic, which for equities is usually the market capitalisation, while for bonds – the amount outstanding. Then, only a part of assets in each stratum is purchased (Guo and Leung 2015). As this approach does not take into account the correlation structure within each group, it is nowadays widely employed mostly for bond ETFs, while sampling-based equity ETFs rely on purely model-driven optimisation methods (Snowden 2021). Optimisation approaches, aimed at approximating the risk structure of an index, range from deterministic algorithms through stochastic algorithms to machine learning models.

Methods based on machine learning, which are the main focus of our study, usually attempt to achieve a tracking error on par with more classical optimisation approaches while reducing the computational effort required to perform model rebalancing (Satpathy and Shah 2021). In addition, machine learning models may discover more complex interdependencies in data, as well as allow fund managers to incorporate additional information into their analysis (Zheng et al.

2020). The most popular family of models employed for sparse index construction is LASSO, a regression method providing variable selection and regularisation. Many extensions to the basic model have been introduced to increase robustness and efficiency of computation, as well as to reduce the tracking error (Wu et al. 2014, Xu et al. 2015, Che et al. 2022). Other solutions, such as genetic algorithms (Beasley et al. 2003, Giuzio 2017) and neural networks (Chen et al. 2020), were also employed.

The machine learning research carried out in the field of index tracking, however, has to this point been focused on supervised models, since stock prices form labelled datasets. While intuitive, this avenue has omitted a useful tool for dimensionality reduction – Principal Component Analysis (PCA). PCA, an unsupervised learning algorithm developed in order to distil information from a multidimensional dataset, does not give predictions on its own, and therefore has been used in finance mostly as a tool for financial statement or industry analysis (Janićijević et al. 2022, Mbona and Kong 2019), with only a small number of studies applying it to portfolio construction (Chen 2014).

In our work, we propose a novel approach to sparse index replication, building on the method described by Jolliffe (1986), and successfully tested on empirical data by Rea et al. (2015). We employ PCA to select a subset of the index explaining the majority of its variance and test our method on small-, mid- and large-capitalisation stock indices from German and Polish markets. The key difference we introduce to the Jolliffe method is our use of the variance to set the hyperparameters explained by Principal Component (PC); however, we also test various methods of selecting different parametrisation windows. Finally, we compare our results with the LASSO-based approach and find that while PCA is quite successful at creating diversified, well-performing sparse portfolios, the tracking error of LASSO-based portfolios is much lower. Tracking the performance of both methods is affected by market development and index capitalisation. The results indicate better outcomes for developed markets and a positive impact of capitalisation on sparse index performance.

The five key research hypotheses we tested were:

1. A PCA-based model can achieve tracking performance similar to the one achieved using portfolio optimisation and regularisation methods. Principal Component Analysis is a well-known method for dimensionality reduction and feature extraction. It is a linear technique that seeks to transform the original data into a new coordinate system. The prices of indices are the weighted sum of underlying assets' prices. Thus, linear techniques, like PCA, are a natural choice to construct sparse portfolios.
2. Sparse portfolio selection achieves better performance for developed markets than for developing markets. Developed markets, due to their maturity, are very often characterised by smaller and more stable price volatilities

compared to less mature markets, which may heavily impact sparse portfolio results.

3. The higher the index capitalisation, the better results sparse portfolio performance can achieve. Large-cap indices consist of bigger, more mature companies, which mostly do not offer high returns, but rather stability and smaller fluctuations of the stock price.
4. The shorter the rebalancing time period and the longer the calibration time window, the better sparse portfolio performance is. A short rebalancing period allows the sparse portfolio to quickly capture changes in the market regime, but this can result in unstable statistical estimators negatively affecting portfolio performance.
5. PCA-based models computed using the covariance matrix outperform PCA-based models computed using the correlation matrix. Covariance matrix provides additional market information about price volatilities. Providing more information allows PCA to exclude not only the most correlated instruments, but also stocks with the smallest contribution to the index's variance.

In addition, an important contribution of our work lies in the analysis of Polish indices, which have not yet been the subject of such research, and in the comparison we make across two dimensions: between a developed and a developing market as well as between large-, mid-, and small-capitalisation indices. Previous research in the field, mostly due to data availability, has been generally focused on the few main indices in the largest stock markets in the world, such as the S&P family (Chen et al. 2020, Benidis et al. 2018, Karlow 2012, Rudd 1980, Chen and Kwon 2012), NASDAQ (Satpathy and Shah 2021), or FTSE100 (Shapcott 1992, Yuen et al. 2021), with a smaller number of studies covering broader indices (Satpathy and Shah 2021, Xu et al. 2015, Beasley et al. 2003) and developing markets (Rea et al. 2015), which in practice are the main target for employing sparse index construction.

The remaining part of the paper is organised as follows: in Section 2 we discuss the relevant topic literature, reviewing both historical and state-of-the-art approaches. Section 3 provides a description of the dataset used in our study and a discussion on stylised facts. In Section 4, we look at methodological issues of employing LASSO and PCA techniques to the task of sample optimisation, while Section 5 presents empirical results including the comparison of tracking efficiencies, algorithm performance on small-cap versus large-cap stocks, as well as on developed vs emerging markets. Finally, conclusions complete the paper.

2. Literature Review

2.1. Introduction

The problem of replicating an index has been tackled by researchers since 1980, before the birth of ETFs as an asset class, when index tracking portfolios were managed by mutual funds. As both the popularity of passive investment and availability of computing power grew, multiple algorithms have been proposed, as well as various tracking metrics, methods of estimating trading costs, and approaches to index rebalancing. In the article, we first outline the early development of index tracking methods, then provide an overview of main algorithms employed in sparse portfolio construction, and finally discuss machine learning approaches.

2.2. Early approaches

Rudd (1980) analysed an industry-developed method of stratified sampling and compared it to the Sharpe method of quadratic optimisation in a task of replicating the S&P 500 index using up to 350 stocks and testing both methods on 1977–1978 data. Using the residual standard deviation as a metric, Rudd found the performance of the optimisation method to be twice as effective as the stratification method, while at the same time allowing the manager to keep the portfolio beta relative to the index at exactly one, while the stratification portfolio beta varied from 0.97 to 1.02 across multiple trials.

A further, more in-depth comparison between stratification and optimisation was conducted by Meade and Salkin (1989), who tested four strategies on the 100 largest stocks from the Tokyo Stock Market New Index. The strategies tested were: unconstrained optimisation, stratified optimisation, capitalisation-weighted (iterative integer programming), and capitalisation--weighted-stratified. Unconstrained optimisation achieved the lowest tracking error, with each additional constraint diminishing the performance, and stratification constraint being more damaging to the result than the capitalisation constraint. Further research has therefore focused mostly on various optimisation methods, which were becoming increasingly complex as both algorithms and computing power developed.