Université Laval

Index Tracking under Cardinality Constraints

Internship Report

Author :	Jérémy Aubé-Lévesque
Academic Supervisor :	Prof. Pierre Miasnikof
Host Organization :	CIMMUL
Internship Period :	May 2024 – August 2024

Québec, Canada

August 2024

Acknowledgments

I would like to thank Pierre Miasnikof for supervising my internship and for his valuable mentorship throughout the project. His guidance, advice, and support have been greatly appreciated, and I am especially thankful for the opportunity he gave me to meet inspiring people during this experience.

I am also grateful to Mohammad Bagherbeik for his helpful advice and assistance on the computational aspects of the project.

Finally, I would like to acknowledge the CIMMUL for their financial support, which made this work possible.

1 Introduction

Portfolio management is typically divided into two broad categories : active and passive strategies. Active management assumes that markets are not fully efficient and that skilled managers, through superior forecasting techniques and security selection, can generate excess returns. Passive management, on the other hand, rests on the belief that beating the market consistently over the long term is highly unlikely. The role of passive managers is therefore not to outperform the market, but to replicate a benchmark as closely as possible, usually following a predetermined set of rules. Evidence from historical data suggests that most actively managed funds do not deliver persistent outperformance [1]. This fact has strengthened investor interest in passive investment approaches.

The most direct way to track an index is through full replication, which involves holding all its constituent assets in their exact proportions. This approach ensures perfect tracking accuracy but comes with significant drawbacks, such as exposure to illiquid securities, elevated transaction costs, and the operational burden of managing a large number of positions. Exchange-traded funds (ETFs) provide an alternative, with some funds relying on full replication and others applying different forms of sampling to approximate the index.

This report focuses on index-tracking portfolios designed under a cardinality constraint. Instead of investing in every single component of the benchmark, the idea is to restrict the portfolio to a fixed number of securities. This limitation reduces costs and simplifies portfolio management, while still aiming to achieve close tracking of the index. In practice, the cardinality constraint highlights a common challenge for institutional investors : finding the right balance between replication accuracy and operational efficiency.

A natural way to formulate the index tracking problem with a cardinality constraint is to directly penalize the number of assets included in the portfolio. This can be expressed as follows :

$$\min_{w\in\mathbb{R}^N} \frac{1}{T}\|Xw-y\|_2^2 \quad \text{subject to} \quad \sum_{i=1}^{N} w_i = 1,\ w\geq 0,\ \|w\|_0 = K, \tag{1}$$

where $X \in \mathbb{R}^{T\times N}$ is the matrix of asset returns, $y \in \mathbb{R}^T$ is the index return vector, $w \in \mathbb{R}^N$ is the portfolio weight vector, and $\|w\|_0$ denotes the number of nonzero weights. The constraint $\|w\|_0 = K$ ensures that the portfolio contains K assets.

This problem can equivalently be written in terms of binary selection variables $s \in \{0,1\}^N$, where $s_i = 1$ if the i-th stock is included in the portfolio and $s_i = 0$ otherwise. In this case, the formulation becomes

$$\begin{aligned} \min_{w,\,s}\ & \frac{1}{T}\left\|X\left(w\odot s\right)-y\right\|_2^2 \\ \text{subject to}\ & (w\odot s)^\top \mathbf{1} = 1, \\ & w \geq 0, \\ & s^\top \mathbf{1} = K, \end{aligned} \tag{2}$$

where $\odot$ denotes the Hadamard product. This formulation highlights that the cardinality-constrained index tracking problem is a *mixed-integer quadratic program (MIQP)*, which is known to be NP-hard.

Although solving the joint optimization over both discrete (s) and continuous (w) variables may appear natural, it quickly becomes computationally prohibitive. A more practical and widely used approach is to decompose the problem into two stages : (i) **asset selection**, where a subset of K assets is identified, and (ii) **capital allocation**, where the optimal weights are computed on the chosen subset. This two-step strategy leverages the convexity of the continuous subproblem.

2 Materials and Methods

To determine the binary decision variables $s \in \{0,1\}^N$, we adopt a graph-based clustering approach. The stock universe is represented as a complete weighted graph $G = (V, E)$, where each vertex $i \in V$ corresponds to a stock and each edge $(i, j) \in E$ is assigned a weight reflecting the dissimilarity between the returns of assets i and j. Two dependence measures are employed to define these dissimilarities.

(i) Pearson correlation. Let ρ_{ij} denote the Pearson correlation between the return series of assets i and j, computed over a window of T observations. The corresponding distance is defined as

$$d_{ij} = \sqrt{2\Big(1-\rho_{ij}\Big)} \in [0,2], \tag{3}$$

so that pairs of assets with strong positive linear dependence are assigned smaller distances [3].

(ii) Distance–correlation. To account for potential nonlinear dependence, we also employ the nonparametric distance correlation statistic [5]. For two assets i and j, the distance is defined as

$$d_{ij} = 1 - R(X, Y) \in [0, 1], \tag{4}$$

where $R(X, Y)$ denotes the sample distance correlation between their return series. So that pairs of assets with strong dependence are assigned smaller distances.

QUBO Model

Once the dissimilarities between all pairs of assets have been computed, the next step is to extract a representative subset of stocks. This can be understood as a clustering problem on the graph G, where the goal is to partition the market into groups of assets that behave similarly. Within each cluster, one stock is chosen as a "centroid" or exemplar, serving as a proxy for the dynamics of the entire cluster. By doing so, we reduce the dimensionality of the index while preserving its main sources of variation. In other words, rather than investing in all assets, we select a small number of central representatives whose returns are strongly correlated with those of their respective clusters. To formalize this idea, we adopt a K-medoids clustering approach formulated as a QUBO problem [2].

We construct the similarity matrix $\Delta = [\delta_{ij}]$,

$$\delta_{ij} = 1 - \exp\Big(- \tfrac{1}{2} d_{ij} \Big), \tag{5}$$

which provides a smoothed and more robust measure of association across all asset pairs [2].

We follow the method developed by Bauckhage et al. (2019) [2]. Our K-medoids formulation makes use of a tradeoff parameter α - β to find an optimal combination of dispersed and central stocks on the market graph.

Model (2.1) seeks to find K nodes that are considered the most central within the graph,

$$\min_{s} \; s^\top \Delta \mathbf{1} \quad \text{s.t. } s^\top \mathbf{1} = K, \; s_i \in \{0, 1\}, \; \forall i \in V. \tag{6}$$

Meanwhile, model (2.2) seeks to find K nodes that are considered most dispersed within the graph,

$$\max_{s} \; \frac{1}{2} s^\top \Delta s \quad \text{s.t. } s^\top \mathbf{1} = K, \; s_i \in \{0, 1\}, \; \forall i \in V. \tag{7}$$

We then apply the $\alpha - \beta$ tradeoff parameters to weigh the contributions of Model (2.1) and Model (2.2). Our ultimate goal is to find portfolios consisting of K exemplars that best replicate the returns of each index in our study. To achieve this goal, a quadratic penalty is added and the problem becomes :

$$\min_{s} \; \beta \, s^\top \Delta \mathbf{1} - \frac{\alpha}{2} \, s^\top \Delta s + \lambda \, (s^\top \mathbf{1} - K)^2, \quad s_i \in \{0, 1\}, \; \forall i \in V, \tag{8}$$

with $\alpha = \frac{1}{K}$ and $\beta = \frac{1}{N}$.

Boltzman machine

To solve the QUBO formulation described above, we rely on a Boltzmann machine as proposed by Miasnikof et al. (2022) [4]. The key idea is to encode the QUBO objective as an energy function, so that feasible solutions naturally correspond to low-energy states. Unlike traditional optimization methods that require explicit penalty terms to enforce constraints, the Boltzmann machine integrates them directly through its probabilistic structure.

Practically, the algorithm explores the solution space by sampling from the distribution induced by the energy landscape. This stochastic mechanism makes it possible to avoid exhaustive enumeration while still identifying high-quality solutions to the NP-hard clustering problem. In our setting, the Boltzmann machine thus serves as an efficient heuristic for selecting the subset of K representative assets.

For a more detailed description of the Boltzmann machine implementation, we refer the reader to Miasnikof et al. (2022) [4].

Capital Allocation

Once the optimal subset of assets s^* has been identified in the first stage, we proceed to the second stage of the two-step approach : the allocation of capital. Here, the objective is to determine the portfolio weights w restricted to the selected assets, so as to minimize the tracking error with respect to the benchmark index.

Formally, given s^*, we solve the following quadratic optimization problem :

$$\begin{aligned} \min_{w \in \mathbb{R}^N} \ & \frac{1}{T} \left\| X(w \odot s^*) - y \right\|_2^2 \\ \text{subject to } & (w \odot s^*)^\top \mathbf{1} = 1, \\ & w \geq 0, \end{aligned} \tag{9}$$

This problem is convex and can be efficiently solved using standard quadratic programming methods. The constraint $(w \odot s^*)^\top \mathbf{1} = 1$ ensures that the portfolio is fully invested, while the non-negativity constraint avoids short positions.

Data

For the empirical analysis, we rely on daily stock market data obtained from the Wharton Research Data Services (WRDS) CRSP database. We use adjusted returns that account for both dividends and stock splits, ensuring that the data accurately reflects the performance of each security over time. Missing values were imputed with zeros.

The investment universe corresponds to the constituents of the S&P 500 index. For each year between 2014 and 2024, we take the composition of the index as of the beginning of the year. Based on this cross-section of securities, we construct portfolios using three years of historical return data to solve the optimization problem. Portfolios are then rebalanced annually according to the updated index composition and the most recent return history.

All computations were performed on a cloud-based server running Linux (AWS EC2, instance type `r5.4xlarge`). This instance provides 16 virtual CPUs and 128 GiB of memory, offering the computational resources necessary to efficiently handle large-scale optimization problems.

3 Results

We evaluate index-tracking portfolios with a cardinality constraint of $K = 30$ assets. Portfolios are rebalanced annually from 2014 to 2024, and each optimization is given a **3-hour** compute-time budget. Two solvers are compared : `Gurobi` and a *Boltzmann machine* . For graph construction, we test both *Pearson correlation* and *distance correlation.*

Performance measure. Out-of-sample performance is assessed with the **tracking error** (TE), defined as

$$\mathrm{TE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(R_{p,t} - R_{b,t}\right)^2}, \tag{10}$$

where $R_{p,t}$ and $R_{b,t}$ are the tracking portfolio and benchmark returns over one year of out-of-sample observations.

Empirical findings. Figure 1 shows cumulative portfolio returns compared with the S&P 500. All methods track the index closely. Results consistently indicate that **Pearson correlation yields lower deviations from the benchmark than distance correlation**. By contrast, the choice of solver (**Gurobi vs. Boltzmann**) makes little difference : their outcomes are highly similar, with no persistent advantage.

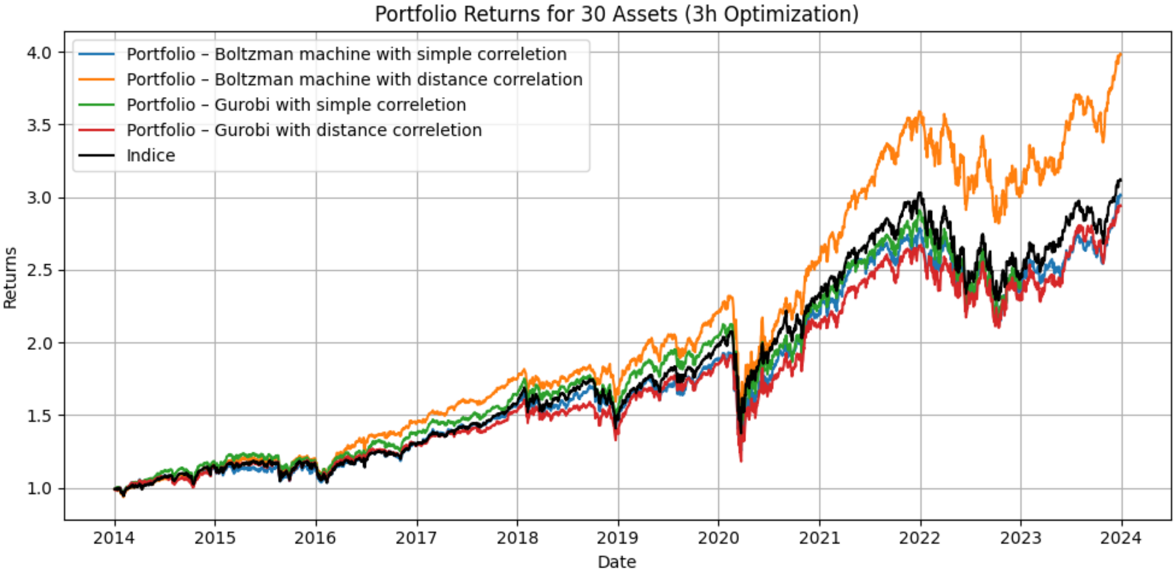

Figure 1 – Cumulative returns for $K = 30$ portfolios under annual rebalancing (3-hour budget per optimization). Pearson-based portfolios track the index more closely, while Gurobi and Boltzmann deliver nearly identical results.

Figure 2 reports the annual out-of-sample tracking errors. The ranking matches the return analysis : **Pearson consistently outperforms distance correlation**, while **Gurobi and Boltzmann remain comparable**, with minor year-to-year variations but no significant gap.

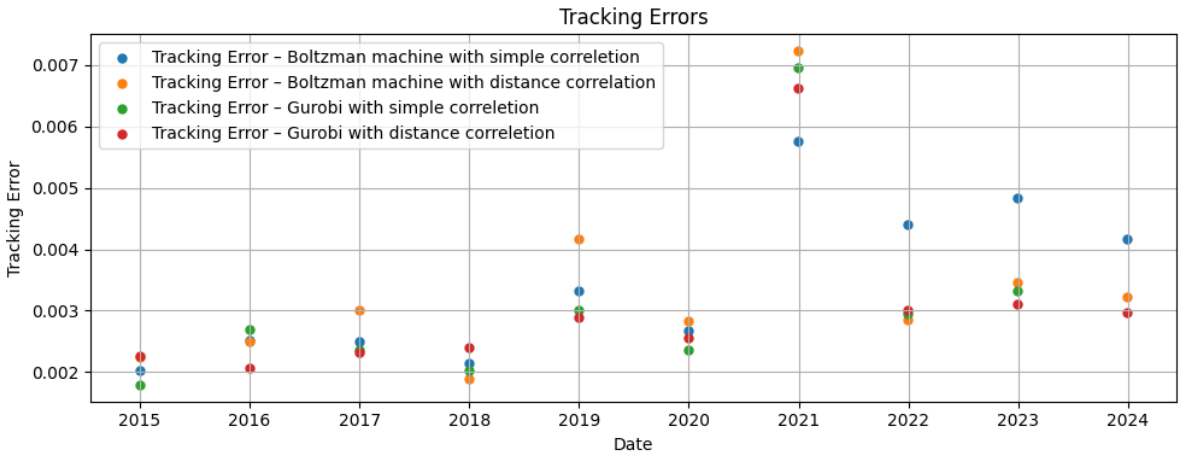

FIGURE 2 – Annual out-of-sample tracking errors for $K = 30$ portfolios (2014–2024, 3-hour optimization). Pearson leads to lower tracking errors than distance correlation, while solver choice has little effect.

4 Future Work and Conclusion

This work has presented a methodology for index tracking under a cardinality constraint, formulated as a two-stage optimization problem. In the first stage, a graph-based clustering approach was used to select a subset of representative assets, leveraging both Pearson correlation and distance correlation as measures of dependence. In the second stage, a convex quadratic program was solved to allocate portfolio weights, ensuring a close replication of the benchmark index.

Future research will aim to extend this approach to larger and more complex benchmarks, such as the MSCI World index, which involves a much broader set of securities across multiple countries. This setting introduces additional layers of complexity, including currency exposure, heterogeneous market structures, and the need to account for transaction costs. Incorporating transaction cost constraints directly into the optimization model would enhance its practical relevance, as real-world index tracking strategies must balance tracking accuracy against transaction costs and rebalancing costs.

From a methodological perspective, the use of graph-based models remains a promising direction. The study of multi-layer or multiplex graphs—where layers correspond to different markets, sectors, or regions—offers a compelling framework for capturing the complexity of global indices.

In conclusion, the results obtained on the S&P 500 highlight the potential of combining graph clustering methods with convex allocation techniques to tackle cardinality-constrained index tracking. Extending this framework to global benchmarks with transaction cost considerations represents a natural and impactful direction for future research.

Références

[1] Brad M. Barber et Terrance Odean. “Trading Is Hazardous to Your Wealth : The Common Stock Investment Performance of Individual Investors”. In : *The Journal of Finance* (2000).

[2] Christian Bauckhage et al. “A QUBO Formulation of the k-Medoids Problem”. In : *Proceedings of the Conference on Lernen, Wissen, Daten, Analysen (LWDA)*. 2019.

[3] Rosario N. Mantegna. “Hierarchical Structure in Financial Markets”. In : *The European Physical Journal B* (1999).

[4] Pierre Miasnikof, Mohammad Bagherbeik et Ali Sheikholeslami. “Graph Clustering with Boltzmann Machines”. In : *Discrete Applied Mathematics* 343 (2024). Preprint available at arXiv :2203.02471, p. 206-218.

[5] Gábor J. Székely, Maria L. Rizzo et Nail K. Bakirov. “Measuring and Testing Dependence by Correlation of Distances”. In : *The Annals of Statistics* (2007).