Brooks Li, Gayathree Gopi, Mason Shu, Vannie Sung

Optimization Project 2 - Integer Programming

**Index Mutual Fund Analysis: Portfolio Composition and Optimization**

**Overview**

This report provides a detailed analysis for constructing an indexed mutual fund that aims to mirror the NASDAQ-100 (NDX) index very closely. The goal is to identify an optimal set of component stocks and determine their respective weights to minimize tracking error, which is the divergence between the portfolio’s returns and the index’s returns over time. Our analysis is based on historical return data from the years 2023 and 2024, allowing us to validate the model by comparing in-sample tracking error (using 2023 data) and out-of-sample tracking error (using 2024 data).

Note the following objective functions and constraints for our analysis:

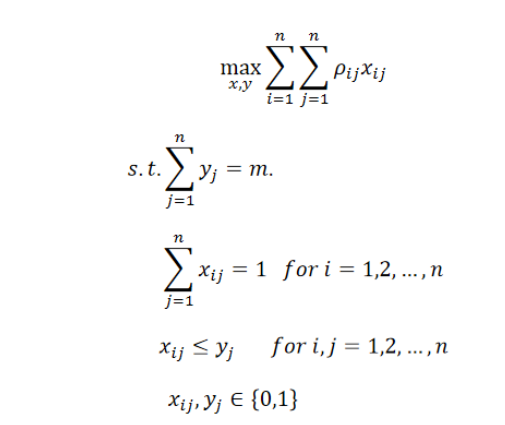
**Stock Selection through Integer Programming**

To determine the ideal subset of stocks, we use a Linear programming (LP) model that maximizes similarity between the NASDAQ-100 index and the selected subset (our portfolio). Here, the model leverages a similarity matrix (𝜌), which is constructed using the historical correlations of returns between each pair of NASDAQ-100 stocks. The matrix element, ρij ​, quantifies the similarity between stock i and stock j based on their return correlations, assuming that highly correlated stocks exhibit similar price behavior. For the optimization, we benchmark two optimization models to make recommendations, namely

* Integer Programming (IP)
* Mixed Integer Programming (MIP)

The optimization is conducted through a simulation approach consisting of different allocation of stocks (m) in our portfolio using the above models. This is done through solving -

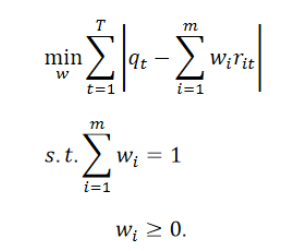
1. Below optimization function to select “m” stocks based on optimizing the similarity model maximizes the similarity between the n stocks and their representatives in the fund (“m” stocks).



Where

* The binary variable ​ represents whether stock j is included in the subset.
* The binary variable represents whether stock j is the most similar stock to stock i in the fund.

For each chosen “m” above, the tracking error is minimized by calculating optimal set of weights which indicates investment (individual stock allocation within the selected portfolio).

****

**1. Data Preparation and Return Calculation**

The project uses two datasets:

1. **In-sample data (2023)**: Historical returns for NASDAQ-100 stocks, used to construct the similarity matrix and optimize initial weights.
2. **Out-of-sample data (2024)**: Future data used for validation to test tracking accuracy under varying market conditions.

To calculate the daily percentage returns for each stock in the NASDAQ-100 stocks and the index, we apply the following:

1. For missing values of raw data, we apply forward-fill and backward-fill methods (.ffill() and .bfill()) for imputation to ensure non-null values.
2. We use historical closing price data for each component stock and the index, calculating the percentage changes from one trading day to the next. The percentage return captures the price movement in a standardized form, allowing for an apples-to-apples comparison across stocks with different price levels.
3. Once the returns are calculated, we ensure that both index returns and selected stock returns have compatible lengths by aligning the stock and index return data based on common dates. This also ensures that all comparisons between the portfolio and the index are made on consistent trading days, avoiding biases introduced by non-trading days or missing data points.
4. We used the 2023 stocks returns to estimate a correlation matrix 𝜌. Here, higher correlation implies two stocks (returns) are similar. This matrix is a key input for the integer programming (IP) model used for selecting stocks. The correlation matrix and indexed returns will be input to the Optimization Model (subject to constraints).

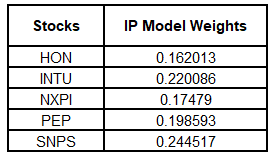
**2. Results**

When constructing a passively managed index fund that tracks the NASDAQ-100 with a reduced number of stocks, the primary objectives are to minimize tracking error and maintain practical fund management. The charts included in the analysis provide visual insights that are essential for evaluating the effectiveness of the fund in terms of stock selection, weighting strategy, and overall performance compared to the NASDAQ-100 index. Each chart offers unique insights, together forming a comprehensive basis for recommending the number of stocks (m) to include in the portfolio and how to allocate weights effectively.

**2.1. Preliminary Optimization Results with m=5**

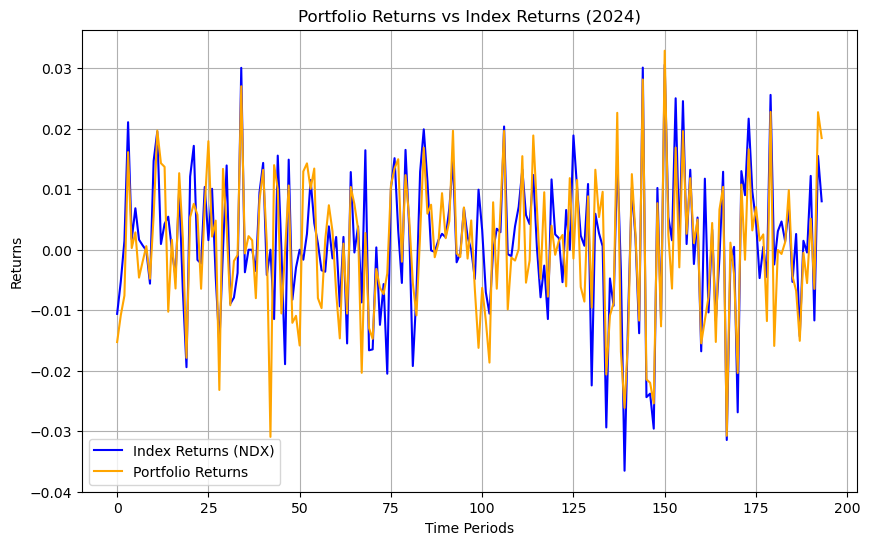
Through bipartite matching and weighting, we determined the optimal selection of five stocks, HON, INTU, NXPI, PEP, and SNPS.

**Table 1. Preliminary results with m = 5 portfolio**

****

To assess how well the optimized portfolio performs on new data, we use the 2023 weights to calculate the portfolio returns for 2024. The out-of-sample tracking error is calculated from the predictions as seen in *Figure 1* by measuring the absolute deviations between the portfolio’s returns and the NDX index returns over 2023 (in-sample)/2024(out-of-sample) samples. Note that the absolute deviation between the portfolio and index returns for 2024 is computed using the weights derived from 2023. The total tracking error for 2024 is printed as a measure of out-of-sample performance. This analysis is crucial for evaluating the robustness of the model and determining whether the portfolio requires regular rebalancing.

*Figure 1* below shows m=5 portfolio returns vs actual index returns over selected daily records.

**Figure 1: Preliminary m=5 portfolio returns vs Actual Index (Market) returns**

**2.2. Full-scale Optimization Results - Portfolio Size and Weight Allocation Strategy**

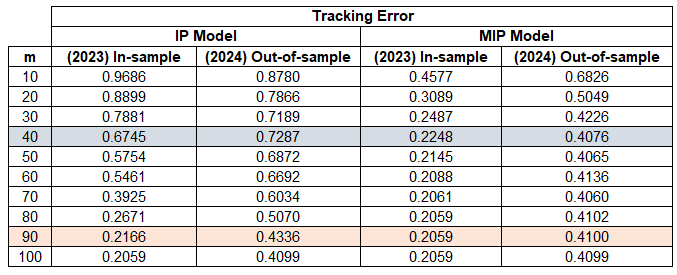
With the optimization model in place, the next step is to evaluate how different portfolio sizes affect tracking error. Specifically, we examine portfolios of varying sizes (from 10 to 100 stocks) to understand the relationship between portfolio size and tracking performance.

To do this, we follow the following steps using a simulation approach.

1. Select a subset of stocks (m = 10, 20,..<100) in our portfolio (See Table 1 for results)
2. For this subset of stocks in (1), we calculate the optimal weights of each stock in our portfolio that minimizes the tracking error.
3. Tracking errors (absolute error of portfolio vs index returns) were computed for evaluation.
4. Steps (1-3) were repeated by IP and MIP models.

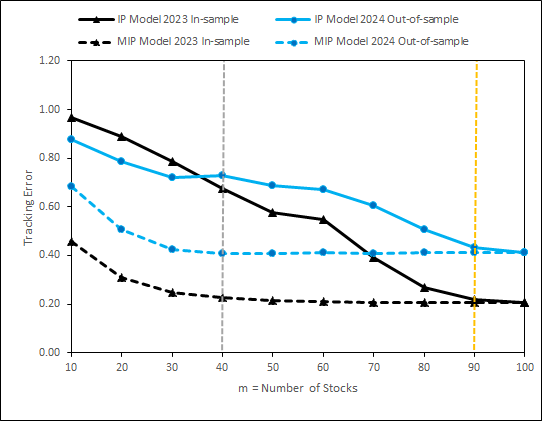
The data in *Table 1* summarizes the tracking errors for different portfolio sizes across both in-sample (2023) and out-of-sample (2024) periods as obtained by the IP model.

**Table 1: Number of stocks (m) selected in our portfolio vs tracking error (in-sample) and out-of-sample (IP Model)**

**

*Figure 2*shows a plot of tracking errors for the IP and MIP models, respectively. For the MIP model’s maximum possible weight (Big M) was defined to be 1 (Upper Bound).

**Figure 2. IP/MIP Model Tracking Errors for different allocations of portfolio sizes (m)**



Based on the results, there are 3 key observations:

1. Ideally, we want to obtain a range of m where the tracking error stabilizes (elbow-type point). Beyond this ‘m’, we do not expect tracking error to substantially decrease mostly due to the selection of weakly correlated stocks to index returns.
2. *Tracking error for IP model in Figure 2* does not show any intuitive trend indicating an optimization has occurred. This is clear in that if we increase ‘m’, tracking error would obviously decrease. However, the level of decrease is strongly monotonic - indicating an optimal m would be around 90 stocks..
3. *Tracking error for MIP model in Figure 2* clearly indicates an optimization has occurred in that m=40 is an ideal elbow point beyond which tracking error does not substantially decrease in in-sample data and almost flat meaning the optimization has been reached. The diminishing returns in tracking error reduction as m increases. Initially, with smaller m values, we observe a steep decline in tracking error, indicating that adding more stocks to the portfolio significantly enhances the ability to track the index. However, as m continues to increase, particularly beyondm = 40, the tracking error stabilizes for both in-sample and out-of-sample data. This plateau suggests that adding more stocks beyond this point provides minimal benefit in terms of reducing tracking error.

We can view the 40 stocks selected by the MIP model and their correlation with the index.

**3. Recommendations**

Based on the result of our analysis, we recommend constructing a portfolio with **40 stocks based on the results of the MIP model**. This portfolio size provides a good trade-off between minimizing tracking error and reducing the complexity and cost associated with a larger portfolio. Increasing the portfolio size beyond 40 stocks seems to result in diminishing returns regarding tracking error reduction.

Monthly or quarterly rebalancing is recommended to ensure weights remain effective in capturing market movements while minimizing trading costs. Regarding the selection of weights for the selected stocks, it should be determined by setting up a linear program as demonstrated in Task 2 and running the linear optimization model to find the optimal weights. The optimization problem is formulated to minimize the absolute difference between the index returns and the weighted sum of stock returns. In addition, the weights must satisfy two constraints: the sum of the portfolio weights must equal 1, and all weights must be non-negative. Using optimized weights based on minimizing tracking error between portfolio returns and the index, weights should be dynamically recalibrated to keep the portfolio aligned with index movements.

**4. Conclusion**

To conclude, our analyses utilized an optimized approach to constructing an indexed mutual fund tracking the NASDAQ-100. Through bipartite matching and weighting, we determined the optimal selection of five stocks, HON, INTU, NXPI, PEP, and SNPS, that minimized tracking error for both in-sample (2023) and out-of-sample (2024) periods. Our findings suggest that a portfolio size of around 40 stocks achieves the best balance between tracking precision and management efficiency. During periods of market volatility, however, regular rebalancing will be beneficial for reinforcing alignment with the index.

[1] Index refers to market index (NDX) whereas portfolio is a subset of stocks that we identify to track the index.