Index tracking strategy based on mixed-frequency financial data

Xiangyu Cui\textsuperscript{1,2,}\textsuperscript{*}, Xuan Zhang\textsuperscript{1}

1 School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China, 2 Shanghai Institute of International Finance and Economics, Shanghai University of Finance and Economics, Shanghai, China

\textsuperscript{*} cui.xiangyu@mail.shufe.edu.cn

Abstract

To obtain market average return, investment managers need to construct index tracking portfolio to replicate target index. Currently, most literatures use financial data that has homogenous frequency when constructing the index tracking portfolio. To make up for this limitation, we propose a methodology based on mixed-frequency financial data, called FACTOR-MIDAS-POET model. The proposed model can utilize the intraday return data, daily risk factors data and monthly or quarterly macro economy data, simultaneously. Meanwhile, the out-of-sample analysis demonstrates that our model can improve the tracking accuracy.

Introduction

The index tracking strategy, which aim at tracking the return of a given stock index when constructing the portfolio, is a major strategy adopted by fund managers. [1–7] theoretically and experimentally study the index tracking strategy under different constraints in reality, e.g. the number of stocks in the portfolio is limited.

The simplest and most widely used index tracking strategy is the global minimum variance strategy. Let $\mathbf{R}_t$ be the vector of daily excess returns of stocks over the target index and $\mathbf{\omega}_t$ be the global minimum variance portfolio weights. The difference between return of the index tracking strategy and return of the target index is $\mathbf{\omega}_t^\top \mathbf{R}_t$ on day $t$. Mathematically, the investors aim to minimize tracking error, which is measured by variance of the difference between return of the index tracking strategy and return of the target index, i.e.,

$$\min_{\mathbf{\omega}_t} \mathbf{\omega}_t^\top \Sigma_t \mathbf{\omega}_t \quad \text{s.t.} \quad \mathbf{\omega}_t^\top \mathbf{1} = 1,$$

where $\Sigma_t$ is the $N \times N$ conditional covariance matrix of $\mathbf{R}_t$. Obviously, the optimal index tracking strategy is

$$\mathbf{\omega}_t^{\text{min}} = \frac{\Sigma_t^{-1} \mathbf{1}}{\mathbf{1}^\top \Sigma_t^{-1} \mathbf{1}}.$$

We can see that the key part of global minimum variance strategy is to estimate the covariance matrix or inverse covariance matrix.
In the literature, methods for estimating the covariance matrix or inverse covariance matrix mainly focus on financial data with homogenous frequency. [8–10] try to estimate the covariance matrix based on quarterly, monthly and daily returns, respectively. [5, 11–18] aim to improve covariance matrix estimation using intraday data. Differently [19–23], focus on the estimation of inverse covariance matrix. With the improvement in high-speed computation and large amounts of storage, financial data streams become more and more real-time and complex, such as high-frequency data and ultra-high-frequency data. Besides the historical data in financial markets, monthly or quarterly macro-economic factors are also valuable information sources of stocks’ volatilities (see [24–29] further reveal that macro-economic factors, such as GDP growth, exchange rate and short-term interest rate, are important explanatory variables of the slow-moving component in volatilities.

However, due to the heterogeneous frequency of macro-economic factors and historical data, it is a great challenge to construct a unified econometric model. Among all proposed models, the mixed data sampling (MIDAS) method in [30] attracts great attentions, and induces several important extensions, such as GARCH mixed-data sampling (GARCH-MIDAS) model (see [31]), Factor-based mixed-data sampling (Factor-MIDAS) model (see [32]). Different from the MIDAS method, other economic models handling mix-frequency data, such as High-frequency-based volatility (HEAVY) model (see [33]); Factor GARCH-Itô model (see [34]), are mainly focusing on integrating the intraday financial data and daily financial data. Compared with homogenous-frequency models, mixed-frequency model contains more information in the original data, which can better capture market and have better accuracy in prediction. It provides a timely update on portfolio and helps fund managers achieve targeted index tracking performance.

In this paper, we propose a general framework for mixed-frequency financial data, called FACTOR-MIDAS-POET model, to estimate the covariance matrix. The proposed model combines monthly macro-economic factors, daily observable factors (market return and the innovation of VIX) and intraday returns to improve covariance matrix estimation. In empirical analysis, we compare our model with existing models in the literature, and find that the tracking accuracy of minimum variance tracking strategy is greatly improved by using the proposed model. The reason is that compared with other models, our model contains macro-economic information and option market information. We also find that when the amount of historical data decreases, performances of the index tracking portfolios based on different models all decrease, but our model is less affected. Moreover, when the estimation window is short (e.g., 3 months), integrating intraday return data into our model may yield better tracking performance than not using the intraday return data.

The remaining paper is organized as follows. In the estimations of the covariance and its inverse, we introduce existing methods for estimating the covariance matrix and inverse covariance matrix based on homogenous frequency data. In FACTOR-MIDAS-POET method and estimation, we introduce the FACTOR-MIDAS-POET model and its estimation method. In data and descriptive analysis, we explain the data used in this paper. In empirical study, we conduct the empirical study and compare performances of different models. In conclusions and discussion, we conclude our paper.

The estimations of the covariance and its inverse

To obtain the minimum variance index tracking strategy, there are two main ways. The first one is to estimate the covariance matrix and then obtain the inverse matrix; The second one is to estimate the inverse covariance matrix directly. The rest of this section summarizes the
important existing methods for estimating the covariance matrix and inverse covariance matrix.

**Estimators of covariance matrix**

Based on the daily financial data, we summarize three estimators as follows,

\[ S_{t,1} = \frac{1}{T - 1} \sum_{k=1}^{T} (R_{t-k} - \bar{R})(R_{t-k} - \bar{R})^T, \]  

**Equation 1**

\[ S_{t,2} = S_{t,1} + \frac{1}{T} \sum_{l=1}^{L} \sum_{k=1}^{T-l} \left( 1 - \frac{l}{L+1} \right) [(R_{t-k-l} - \bar{R})(R_{t-k} - \bar{R})]^T + (R_{t-k} - \bar{R})(R_{t-k-l} - \bar{R})^T], \]  

**Equation 2**

\[ S_{t,3} = \exp(-z)S_{t,1} + z \exp(-z)(R_{t-1} - \bar{R})(R_{t-1} - \bar{R})^T. \]  

**Equation 3**

\( S_{t,1} \) is the sample covariance matrix. \( S_{t,2} \) is the weighted lead and lag covariance matrix proposed by [35], which is designed to eliminate the non-synchronous trading effect. In this paper, \( L \) is set to 3 for daily return following [36]. \( S_{t,3} \) is the backward-looking rolling estimator proposed by [10].

When we have intraday financial data, these estimators are modified as follows,

\[ S_{t,1} = \frac{1}{T - 1} \sum_{k=1}^{T} \sum_{i=1}^{M} R_{i,t-k}R_{i,t-k}^T, \]  

**Equation 4**

\[ S_{t,2} = S_{t,1} + \frac{1}{T} \sum_{k=1}^{T} \sum_{l=1}^{L} \sum_{i=1}^{M} (R_{i,t-l-k}R_{i,t-k}^T + R_{i,t-k}R_{i,t-l-k}^T), \]  

**Equation 5**

\[ S_{t,3} = \exp(-z)S_{t,1} + z \exp(-z) \sum_{i=1}^{M} R_{i,t-1}R_{i,t-1}^T, \]  

**Equation 6**

where \( R_{i,t-k} \) is the vector of excess returns at time \( i \) in day \( t-k \). \( S_{t,1} \) is proposed by [12]. \( S_{t,2} \) is proposed by [5]. If overnight returns are involved in the estimators, we have

\[ S_{t,1} = \frac{1}{T - 1} \sum_{k=1}^{T} \sum_{i=1}^{M} R_{i,t-k}R_{i,t-k}^T + (R_{0,t-k} - \bar{R}_0)(R_{0,t-k} - \bar{R}_0)^T, \]  

**Equation 7**

\[ S_{t,3} = \exp(-z)S_{t,1} + z \exp(-z) \left( \sum_{i=1}^{M} R_{i,t-1}R_{i,t-1}^T + (R_{0,t-1} - \bar{R}_0)(R_{0,t-1} - \bar{R}_0)^T \right), \]  

**Equation 8**

where \( R_{0,t-k} \) is the overnight return in day \( t-k \).
Estimators of inverse covariance matrix

The estimators of inverse covariance matrix are often built upon the estimators of covariance matrix. We summarize these estimators in this subsection.

\[ S_{i,4} = (T - N - 1)S_{i,1}^{-1}, \]  
\[ S_{i,5} = \alpha (T - N - 1)S_{i,1}^{-1} + \alpha \frac{NT + N - 2}{tr(S_{i,1})} I + (1 - \alpha) \frac{NT - 2}{tr(S_{i,1})} I, \]  
\[ S_{i,6} = (T - N - 1)S_{i,1}^{-1} + \frac{NT - 2}{tr(S_{i,1})} u I, \]  
\[ S_{i,7} = (T - N - 1) \left( 1 - u^2 \right) S_{i,1}^{-1} + \frac{(T - N - 1)N}{tr(S_{i,1})} u^2 I, \]  
\[ S_{i,8} = (T - N - 1)S_{i,1}^{-1} + \frac{2(N - 1)}{tr(S_{i,1})} S_{i,1}, \]  
\[ S_{i,9} = HL^{-1} CH^T, \]

where \( tr(\cdot) \) is the trace of a matrix, \( u = N \text{det}(S_{i,1})^2 tr(S_{i,1})^{-1} \), \( \text{det}(\cdot) \) is the determinant of a matrix, \( H \) is an orthogonal matrix and \( L \) is a diagonal matrix such that \( S_{i,1} = HHL^T \), \( C \) is a diagonal matrix with elements \( c_i = T + N - 2i - 1 \) for \( i = 1, \ldots, N \). \( S_{i,1} \) is proposed by [19]. \( S_{i,1}^{inv} \) is the shrinkage estimator proposed by [19]. \( S_{i,6}^{inv} \) is proposed by [20]. \( S_{i,7}^{inv} \) is proposed by [21]. \( S_{i,8}^{inv} \) is proposed by [22], \( S_{i,9}^{inv} \) is proposed by [23].

FACTOR-MIDAS-POET method and estimation

There are two restrictions of the methods summarized in Section 2. The first one is that these classical methods require a homogenous sampling frequency, leading to a low usage of information contained in mixed-frequency financial data; The second is that these classical methods do not take monthly or quarterly macro-economic factors into consideration. To remedy these limitations, we propose a model involving multi-frequency financial data to better reflect financial market.

In our model, excess returns are driven by both observable and unobservable factors. Meanwhile, excess returns are influenced by the status of macro economy. More specifically, the model is shown as follows,

\[ r_{i,t,m} = a^T \tau_m F_{obs,t,m} + b^T F_{unobs,t,m} + \epsilon_{i,t,m}, \]  
\[ \tau_m = \beta_0 + \beta^T \sum_{k=1}^{K} B(k, \theta) X_{m-k}, \]  
\[ B(k, \theta) = \frac{\left( \frac{k}{K} \right)^{\theta_1 - 1} \left( 1 - \frac{k}{K} \right)^{\theta_2 - 1}}{\sum_{k=1}^{K} \left( \frac{k}{K} \right)^{\theta_1 - 1} \left( 1 - \frac{k}{K} \right)^{\theta_2 - 1}}, \]

where \( r_{i,t,m} \) is the vector of excess returns of \( N \) stocks at time \( i \) in day \( t \) and month \( m \), \( F_{obs,t,m} \) is
the vector of daily observable factors in day $t$ and month $m$, $F_{t,obs,t,m}$ is the vector of daily unobservable factors in day $t$ and month $m$, $\epsilon_{t,i,m}$ is the $N$-dimensional residual vector, $\tau_{m}$ is the long-run component associated with monthly observable macro-economic factors, $X_{t}$ is the vector of the macro-economic factors in month $k$, $B(k, \theta)$ is the widely used Beta weighted lag structure in MIDAS model. We need to mention that $\tau_{m}$ is known for any day $t$ in month $m$, and is updated in the beginning of the next month $m + 1$.

Based on principal orthogonal complement thresholding (POET) method in [37], the daily covariance matrix of excess returns, $\hat{\Sigma}_{t,m}$, is estimated as follows,

$$
\hat{\Sigma}_{t,m} = \tau_{m}^{2} a^{T} \hat{\Sigma}_{t,m,obs} a + \hat{R}_{t} + \hat{R}_{N-t-1},
$$

where $\hat{R}_{t} = \sum_{j=1}^{N} \hat{\lambda}_{j} \hat{u}_{j} \hat{u}_{j}^{T}$, $\hat{R}_{N-t-1} = \sum_{k=1}^{N} \hat{\lambda}_{k} \hat{u}_{k} \hat{u}_{k}^{T}$ is the eigenvector associated with $\hat{\lambda}_{j}$. It is unrealistic to assume that the matrix $\hat{R}_{t}$ is sparse because of the existence of common factors. But it is reasonable to assume that the matrix $\hat{R}_{N-t}$ is sparse. To guarantee sparsity, it is natural to set a threshold to shrink the non-diagonal elements of $\hat{R}_{N-t}$ into

$$
\hat{R}_{N-t} = (\hat{r}_{ij})_{N\times N}, \quad \hat{r}_{ij} = \begin{cases} 
\hat{r}_{ij}, i = j \\
\tau_{m}\hat{\lambda}_{i} I(\hat{\lambda}_{j} \geq \tau_{m}), i \neq j,
\end{cases}
$$

where $s(\cdot)$ is a generalized shrinkage function, $\tau$ is the threshold, $I(\cdot)$ is the indicator function. Then, we obtain the FACTOR-MIDAS-POET estimator of the covariance matrix as follows,

$$
\hat{\Sigma}_{t,m,FMP} = \tau_{m}^{2} a^{T} \hat{\Sigma}_{t,m,obs} a + \hat{R}_{t} + \hat{R}_{N-t-1}. 
$$

Based on [38], we apply the method of linear compression to obtain the shrinkage inverse covariance matrix estimator as follows:

$$
S_{t,m,FMP}^{inv} = c_{1} (S_{t,m})^{-1} + c_{2} I + c_{3} \hat{\Sigma}_{t,m,FMP}^{-1},
$$

where $c_{1}$, $c_{2}$ and $c_{3}$ are combination coefficients, $\hat{\Sigma}_{t,m,FMP}^{-1}$ is the target matrix. The optimized solution and expression can be found in [38].

### Data and descriptive analysis

#### Data

In empirical analysis, we use the stocks listed in Dow Jones Industrial Average (DJIA) to track the S&P 500 index. The stock tickers and full company names of 30 stocks listed in DJIA are available in Table 1. As there are too many missing values in the TAQ data files of TRV company, we remove this company and use the rest 29 DJIA stocks to track S&P 500 index. The sample period in our analysis is from Jan. 1st, 2006 to Dec. 31st, 2011. The daily and 5 minutes (5-min) data of S&P 500 index are obtained from Tick Data. The 5-min data of 30 DJIA stocks are collected from NYSE Trade and Quotations (TAQ) database. Considering the higher possibility of including biases and reporting errors in the first 30 minutes after opening, we discard the first 30 minutes data. Thus, there are 72 intraday 5-min returns and one overnight return in each trading day.

The daily value-weighted market return and daily VIX are considered as daily observable factors and obtained from CRSP database and Chicago Board Options Exchange (CBOE), respectively.

The monthly macro-economic factors includes eight important ones:

1. The daily value-weighted market return and daily VIX are considered as daily observable factors and obtained from CRSP database and Chicago Board Options Exchange (CBOE), respectively.
2. The monthly macro-economic factors includes eight important ones:
3. The monthly macro-economic factors includes eight important ones:
4. The monthly macro-economic factors includes eight important ones:
1. Short-term interest rate (Interest), which is measured by 3-month US treasury bill rate and obtained from the Federal Reserve Board’s H.10.

2. Exchange rate (Exch.), which is measured by the major currencies index collected from Federal Reserve Banks and obtained from the Federal Reserve Board’s H.15.

3. Inflation (Inflation), which is measured by the Consumer Price Index (CPI) and obtained from CRSP database.

4. Slope of the yield curve (Slope), which is measured by the spread between 10-year treasury rate and 3-month treasury rate and obtained from CRSP database.

5. Default rate (Default), which is measured by the difference between Moody’s Baa and Aaa corporate bond yields of the same maturity and obtained from Federal Reserve Board data files in WRDS database.

6. Consumer confidence (CC), which is measured by the Michigan Consumer Sentiment Index and obtained from Trading Economics.

Table 1. Tickers and company full name of DJIA component stocks on June 8, 2009.

| Ticker | Company Name                          |
|--------|---------------------------------------|
| AA     | Alcoa Inc.                             |
| AXP    | American Express Company              |
| BA     | The Boeing Company                    |
| BAC    | Bank of America Corporation           |
| CAT    | Caterpillar Inc.                      |
| CSCO   | Cisco Systems, Inc.                   |
| CVX    | Chevron Corporation                   |
| DD     | E.I. du Pont de Nemours & Company     |
| DIS    | The Walt Disney Company               |
| GE     | General Electric Company              |
| HD     | The Home Depot, Inc.                  |
| HPQ    | Hewlett-Packard Company               |
| IBM    | International Business Machines Corp. |
| INTC   | Intel Corporation                     |
| JNJ    | Johnson & Johnson                     |
| JPM    | JPMorgan Chase & Co.                  |
| KFT    | Kraft Foods Inc.                      |
| KO     | The Coca-Cola Company                 |
| MCD    | McDonald’s Corporation                |
| MMM    | 3M Company                             |
| MRK    | Merck & Co., Inc.                     |
| MSFT   | Microsoft Corporation                 |
| PFE    | Pfizer Inc.                           |
| PG     | The Procter & Gamble Company          |
| T      | AT&T Inc.                             |
| TRV    | The Travelers Companies, Inc.         |
| UTX    | United Technologies Corporation       |
| VZ     | Verizon Communications Inc.           |
| WMT    | Wal-Mart Stores, Inc.                 |
| XOM    | Exxon Mobil Corporation               |

https://doi.org/10.1371/journal.pone.0249665.t001
7. Growth rate in the Industrial Production Index (IPI), which is obtained from Federal Reserve Board data files in WRDS database.

8. Unemployment rate (Unempl.), which is obtained from Trading Economics.

Utilizing the method in [28], we conduct the Principle Components Analysis (PCA) to these eight variables and select the first two principle components as macro-economic factors used in our model.

Descriptive analysis

Tables 2 and 3 present the descriptive statistics for daily returns and 5-min returns of 29 DJIA stocks and S&P index, respectively.

Although the mean of daily returns and 5-min returns for S&P 500 index are approximately zero, their standard deviations are significantly different. It indicates that when sampling frequency increase, there is a dramatic change in volatility. Moreover, the skewness switches from -0.1909 for daily return to 0.0440 for 5-min return, which is more right-skewed. The kurtosis increases from 7.4016 for daily return to 22.5620 for 5-min return.

Table 2. Descriptive statistics for daily returns.

| Stock | Mean × 1000 | 25th Percentile × 1000 | 75th Percentile × 1000 | SD × 1000 | Skewness | Kurtosis |
|-------|-------------|------------------------|------------------------|----------|----------|----------|
| AA    | -0.8021     | -14.7440               | 15.0712                | 33.2524  | -0.2453  | 5.8662   |
| AXP   | -0.0798     | -11.0745               | 11.5691                | 29.8404  | 0.1049   | 6.4575   |
| BA    | 0.0286      | -10.2271               | 10.7482                | 20.7952  | 0.1065   | 3.6824   |
| BAC   | -1.3867     | -13.5525               | 10.9586                | 43.5349  | -0.2921  | 13.7690  |
| CAT   | 0.3155      | -11.6219               | 12.8381                | 24.0225  | -0.0728  | 3.5591   |
| Cisco | 0.0425      | -9.8890                | 10.7245                | 21.5366  | -0.2270  | 6.6922   |
| CVX   | 0.4070      | -8.9778                | 10.5710                | 19.6762  | 0.1361   | 11.2978  |
| DD    | 0.0494      | -10.1948               | 10.6622                | 20.7331  | -0.4140  | 4.6500   |
| DIS   | 0.3111      | -9.2823                | 9.6864                 | 19.9378  | 0.3029   | 5.8237   |
| GE    | -0.4444     | -8.9738                | 8.8330                 | 23.1800  | -0.1263  | 6.7691   |
| HD    | 0.0271      | -10.1650               | 10.0027                | 20.0926  | 0.3543   | 3.3508   |
| HPQ   | 0.0150      | -9.4746                | 10.4959                | 20.4492  | 0.0518   | 4.1362   |
| IBM   | 0.5299      | -6.5984                | 8.3069                 | 15.1777  | 0.0212   | 4.2217   |
| INTC  | -0.0186     | -10.7499               | 11.1148                | 21.0361  | -0.1867  | 3.7357   |
| INJ   | 0.0490      | -4.7435                | 5.2749                 | 10.9609  | 0.1181   | 10.1205  |
| JPM   | -0.1124     | -12.3190               | 11.2466                | 32.5366  | 0.3622   | 9.9782   |
| KFT   | 0.1963      | -6.4856                | 7.1668                 | 13.5095  | 0.0097   | 3.8798   |
| KO    | 0.3546      | -5.7530                | 6.0818                 | 12.7197  | 0.4636   | 8.5198   |
| MCD   | 0.7355      | -6.5178                | 7.8164                 | 13.4517  | -0.0439  | 3.8460   |
| MMM   | 0.0280      | -6.8518                | 8.0256                 | 16.3772  | -0.3299  | 4.7732   |
| MRK   | 0.1455      | -8.4768                | 9.4536                 | 18.6671  | -0.2957  | 6.2243   |
| MSFT  | -0.0063     | -8.6208                | 8.7621                 | 18.9069  | 0.1018   | 8.1691   |
| PFE   | -0.0239     | -7.9898                | 8.3449                 | 16.1736  | -0.0495  | 4.5433   |
| PG    | 0.0897      | -4.8103                | 5.5626                 | 12.1929  | -0.1406  | 7.2973   |
| T     | 0.1492      | -7.7032                | 7.9086                 | 16.2232  | 0.5071   | 8.2793   |
| UTX   | 0.1827      | -7.6566                | 8.8274                 | 17.4147  | 0.2696   | 5.8673   |
| VZ    | 0.1781      | -7.0986                | 8.0542                 | 15.8020  | 0.2347   | 7.0804   |
| WMT   | 0.2052      | -6.7874                | 6.5878                 | 13.1521  | 0.1291   | 5.5092   |
| XOM   | 0.2658      | -8.1494                | 9.2605                 | 18.2490  | -0.0163  | 13.0925  |
| S&P 500 | 0.0485 | -5.4778                | 6.4014                 | 15.3059  | -0.1909  | 7.4016   |

https://doi.org/10.1371/journal.pone.0249665.t002
The standard deviations of individual stocks obtained from daily returns are significantly larger than those obtained from 5-min returns. Among the 29 individual stocks, 13 stocks have left-skewed distributions for daily return and 18 stocks have left-skewed distributions for 5-min return. Furthermore, the 29 stocks have larger kurtosis for 5-min return. The distributions of individual stock return are unsymmetrical with the long tail and sharp peak.

Table 4 further displays the correlation coefficients of eight macro-economic factors. It reveals the strong correlations between short-term interest rate and slope of the yield curve, between exchange rate and default rate, between growth rate in the Industrial Production Index and unemployment rate. Introducing all macroeconomic variables in covariance estimation model at one time will lead to a lack of statistical significance due to the presence of multicollinearity. Thus, we use principal component analysis (PCA) to create new independent variables and estimate inverse covariance matrix based on principle components in empirical study.

Table 5 shows the principle component matrix using PCA. The first principal component and second principal component we use explain 96.4945% of the total variance. The first
component has significantly correlations with short-term interest rate and slope of the yield curve, which is called monetary factor. It is viewed as an indicator of monetary policy and bond market. The second component named as economic factor has positive correlations with default rate and exchange rate, which mostly captures the periodic fluctuations in economic activity.

**Empirical study**

In our model, excess returns, which could be daily returns or 5-min returns, are driven by two daily observable factors: (i) the value-weighted stock market index; (ii) the innovation of the VIX index. The VIX index is constructed by the implied volatilities of S&P 500 index options. It indicates the investors’ expectation for future 30-day volatility of S&P500 index. We also include two monthly observable principle components of the eight macro-economic factors. We fit the proposed model, estimate the inverse covariance matrix, $S_{inv}^{t,1}$, and compute the minimum variance index tracking strategy with a rolling window scheme. And then we apply the derived strategy for the next day.

**Comparison of current different estimations**

Here, we choose a rolling window of one year (252 trading days). Table 6 compares out-of-sample performances of minimum variance index tracking strategies, which are derived according to different covariance matrix or inverse covariance matrix estimators based on daily return, intraday return with overnight return and intraday return without overnight return, respectively.

Several conclusions can be obtained from Table 6. First, tracking error achieved by $S_{inv}^{t,1}$ with daily return is 0.043973, which is smaller than the tracking errors of using intraday return.

**Table 4. Correlation coefficients between macro-economic variables.**

|        | Interest | Exch. | Inflation | Slope | Default | CC | IPI | Unempl. |
|--------|----------|-------|-----------|-------|---------|----|-----|---------|
| Interest | 1        |       |           |       |         |    |     |         |
| Exch.   | 0.056    | 1     |           |       |         |    |     |         |
| Inflation | 0.397   | -0.309| 1         |       |         |    |     |         |
| Slope   | -0.521   | -0.039| -0.333    | 1     |         |    |     |         |
| Default | -0.316   | 0.512 | -0.371    | 0.163 | 1       |    |     |         |
| CC      | -0.074   | -0.077| -0.092    | -0.040| -0.171  | 1  |     |         |
| IPI     | 0.175    | -0.046| 0.182     | -0.059| -0.025  | -0.317| 1   |         |
| Unempl. | -0.053   | 0.157 | -0.142    | 0.003 | 0.087   | -0.010| -0.417| 1       |

https://doi.org/10.1371/journal.pone.0249665.t004

**Table 5. Principle component matrix.**

| Macro-economic variables | Principle component 1 | Principle component 2 |
|--------------------------|------------------------|-----------------------|
| IPI                      | 0.0016                 | 0.0007                |
| Default                  | -0.0361                | 0.1003                |
| Exch.                    | 0.0006                 | 0.0078                |
| Inflation                | 0.0019                 | -0.0011               |
| Interest                 | 0.4148                 | 0.0085                |
| Slope                    | -0.0257                | -0.0002               |
| Unempl.                  | -0.0016                | 0.0024                |
| CC                       | -0.0046                | -0.0212               |

https://doi.org/10.1371/journal.pone.0249665.t005
Second, the shrinkage inverse covariance matrix estimator, $S^\text{inv}_t$, has the best out-of-sample performance (i.e., smallest tracking error). Third, when including intraday returns, introducing overnight returns often yields better tracking performance, which implies that the overnight returns are useful. Fourth, compared with covariance matrix estimators, most inverse covariance matrix estimators do not have significantly better performances. Furthermore, Table 6 provides turnover rates of applying different strategies. The turnover rates of strategies based on intraday return are smaller, which indicates the value of intraday returns. The tracking strategy with the estimator $S^\text{inv}_{t,7}$ has the lowest turnover rate.

| Covariance | Panel A: daily return | Panel B: intraday return (with overnight) | Panel C: intraday return (without overnight) |
|------------|----------------------|----------------------------------------|----------------------------------------|
|            | Covariance | SD $\times$ 100 | One-tail t test | Turnover (%) | 25th Turnover (%) | 75th Turnover (%) | Covariance | SD $\times$ 100 | One-tail t test | Turnover (%) | 25th Turnover (%) | 75th Turnover (%) | Covariance | SD $\times$ 100 | One-tail t test | Turnover (%) | 25th Turnover (%) | 75th Turnover (%) |
| $S_{t,1}$  | 4.3973     | -0.7453       | 3.3174         | 1.6897       | 4.2155         |                  | $S_{t,1}$  | 4.5334     | -1.4928       | 1.9859         | 1.1308         | 2.2552         | $S_{t,1}$  | 4.7817     | -3.8486       | 1.8787         | 1.1561         | 2.0094         |
| $S_{t,2}$  | 4.9495     | -3.2205       | 6.6824         | 3.7051       | 8.2612         |                  | $S_{t,2}$  | 4.4037     | 1.5562        | 2.1551         | 1.2385         | 2.4591         | $S_{t,2}$  | 4.4680     | 0.1306        | 2.2587         | 1.3622         | 2.4046         |
| $S_{t,3}$  | 4.3569     | 0.5755        | 1.5611         | 0.7266       | 1.9814         |                  | $S_{t,3}$  | 4.7988     | -3.1698       | 1.1020         | 0.6804         | 1.2085         | $S_{t,3}$  | 4.7988     | -3.1698       | 1.1020         | 0.6804         | 1.2085         |
| $S^\text{inv}_{t,4}$ | 4.3973     | -0.7453       | 3.3174         | 1.6897       | 4.2155         |                  | $S^\text{inv}_{t,4}$ | 4.5334     | -1.4928       | 1.9859         | 1.1308         | 2.2552         | $S^\text{inv}_{t,4}$ | 4.7817     | -3.8486       | 1.8787         | 1.1561         | 2.0094         |
| $S^\text{inv}_{t,5}$ | 4.2786     | 0.9673        | 2.5055         | 1.2826       | 3.1920         |                  | $S^\text{inv}_{t,5}$ | 4.4053     | 0.6865        | 1.1572         | 0.7627         | 1.3270         | $S^\text{inv}_{t,5}$ | 4.4603     | 0.3487        | 1.0710         | 0.7222         | 1.1658         |
| $S^\text{inv}_{t,6}$ | 4.3805     | -0.5302       | 3.2310         | 1.6396       | 4.1386         |                  | $S^\text{inv}_{t,6}$ | 4.4553     | -0.7483       | 1.7850         | 1.0500         | 2.0424         | $S^\text{inv}_{t,6}$ | 4.6069     | -2.9174       | 1.6364         | 1.0127         | 1.7634         |
| $S^\text{inv}_{t,7}$ | 4.3837     | 0.1312        | 1.4594         | 0.7839       | 1.7540         |                  | $S^\text{inv}_{t,7}$ | 4.5583     | 1.0620        | 0.9477         | 0.6554         | 1.0710         | $S^\text{inv}_{t,7}$ | 4.5661     | 0.4126        | 0.9483         | 0.6399         | 1.0783         |
| $S^\text{inv}_{t,8}$ | 4.3943     | -0.6026       | 3.3065         | 1.6867       | 4.2125         |                  | $S^\text{inv}_{t,8}$ | 4.4782     | -0.6905       | 1.8969         | 1.0914         | 2.1464         | $S^\text{inv}_{t,8}$ | 4.6563     | -3.3159       | 1.7443         | 1.0761         | 1.8579         |
| $S^\text{inv}_{t,9}$ | 4.3483     | -0.4707       | 1.4594         | 0.7839       | 1.7540         |                  | $S^\text{inv}_{t,9}$ | 4.5042     | -2.0110       | 1.9724         | 1.1696         | 2.3179         | $S^\text{inv}_{t,9}$ | 4.7317     | -4.1228       | 1.8818         | 1.1474         | 2.1217         |

Notes: standard deviations are annualized.

https://doi.org/10.1371/journal.pone.0249665.t006
Surprisingly, involving intraday returns may lower the tracking performance but improve the turnover rates for these estimators. What’s more, overnight returns are meaningful to improve the tracking performance.

Performance of FACTOR-MIDAS-POET method

We consider two models. The first one is based on daily returns of stocks, two daily observable factors, and two monthly macro-economic principle components. The second one is based on 5-min returns of stocks, two daily observable factors, and two monthly macro-economic principle components. Similar to [30], we estimate the models with slow weights ($\theta_1 = 1$ and $\theta_2 = 4$). We also try other forms of weights and obtain the similar conclusions.

Table 7 presents out-of-sample tracking performances of strategies based on FACTOR-MIDAS-POET model. The lag period column reports how many monthly macro-economic principle components are included in the models. Panel A presents the results based on daily returns, while Panel B and Panel C present the results based on 5-min returns. We report the one-tailed t test of tracking errors of different portfolios based on Newey-West standard deviations with six lags in Tables 6 and 7. Following [39], we examine the average squared excess returns over the target index of various estimates with the average squared excess returns of FACTOR-MIDAS-POET method ($K = 3$). To test the average of $s_{1,F-M-P}^2, s_{2,F-M-P}^2, \ldots, s_{N,F-M-P}^2$, we test whether the mean of the sequence $\log(s_{1,F-M-P}^2/s_{1,other}^2), \log(s_{2,F-M-P}^2/s_{2,other}^2), \ldots, \log(s_{N,F-M-P}^2/s_{N,other}^2)$ is significantly smaller than zero using a one-tailed test.

Some conclusions can be obtained from Table 7. First, the tracking performance and turnover ratio based on daily returns are both better than those based on 5-min returns. Second, the tracking performance of our model depends on how many monthly macro-economic principle components are included. Third, except $s_{3,3}$, $s_{1,5}^{inv}$, and $s_{1,7}^{inv}$, the tracking performance and turnover ratio of proposed model are both better than other covariance estimation models based on daily returns. Meanwhile, the one-tail t value of $s_{2,3}$, $s_{1,5}^{inv}$ and $s_{1,7}^{inv}$ are not great.
Similarly, compared with most models reported in Table 6, the proposed FACTOR-MIDAS-POET model has better out-of-sample tracking performance using intraday return. Because our method utilizes financial data with different resources and frequencies. However, the turnover rate of our model is high. This indicates that investment strategies constructed by our method are more active. The high turnover rate is not always a negative indicator. When trying to minimize the index tracking errors for investors, the index investment strategy constructed by our model can reach the targeted performance.

Robust analysis

In this robust analysis, we change the rolling window into three months (3 m), six months (6 m) and nine months (9 m), in order to check whether the length of rolling windows affects our main results.

Tables 8–10 summarize out-of-sample performances of strategies based on different models under three different lengths of rolling windows. When the length of estimation window decreases, tracking errors and turnover rates of all models increase. Because less information is used. Meanwhile, as the length of estimation window decreases, the value of 5-min return data becomes more and more important. Therefore, the intraday information can help us estimate covariance matrix or inverse covariance matrix and construct investment strategies when there is a lack of other information. Most importantly, the proposed FACTOR-MIDAD-POET model has better tracking performance than other models in general, especially when the rolling window is three months. Thus, when the length of estimation window is relatively short, introducing macro-economic information and option market information can greatly improve index tracking strategy.

Conclusions and discussion

In this paper, we propose the FACTOR-MIDAS-POET model, which integrates the intraday return data, daily risk factors data and monthly or quarterly macro economy data, simultaneously. In empirical analysis, we show that the FACTOR-MIDAS-POET model with macro-
economic factors has better out-of-sample tracking performance than most current used models in the literature. The proposed model can fully utilize the financial data with different resources and frequencies. However, the better tracking performance often accompanies with higher turnover rates. Meanwhile, we find that our model has better performance as the length of the estimation window decreases.

Our work is a preliminarily study on the index tracking with mix-frequency data. There are still a lot of aspects we do not cover. For instance, we do not consider the noise of intraday data in the model. Also, when studying index tracking problems, the transaction cost is not explicitly appeared in the model. These problems are worth to be further explored in future studies.

Table 9. Out-of-sample performances of portfolios with $T = 6$ m.

|                      | daily return | intraday return (with overnight) | intraday return (without overnight) |
|----------------------|--------------|---------------------------------|-----------------------------------|
|                      | SD × 100     | One-tail t test | Turnover (%) | SD × 100 | One-tail t test | Turnover (%) | SD × 100 | One-tail t test | Turnover (%) |
| $S_{t,1}$            | 4.5932       | -1.5837          | 7.1564       | 4.5772 | -1.3815          | 3.4341       | 4.9135 | -3.0659          | 2.9827       |
| $S_{t,2}$            | 5.6487       | -5.9653          | 17.1580      | 4.3574 | 0.0090           | 3.7923       | 4.5930 | -0.9154          | 4.1070       |
| $S_{t,3}$            | 4.3676       | 0.3467           | 1.5805       | 4.6044 | -2.5585          | 1.1412       | 4.7986 | -2.8346          | 1.1020       |
| $S_{t,4}^{\text{sw}}$| 4.5932       | -1.5837          | 7.1564       | 4.5772 | -1.3815          | 3.4341       | 4.9135 | -3.0659          | 2.9827       |
| $S_{t,5}^{\text{sw}}$| 4.3597       | -0.5670          | 4.8490       | 4.4063 | 1.0346           | 1.3557       | 4.4801 | 0.8104           | 1.1356       |
| $S_{t,6}^{\text{sw}}$| 4.5404       | -1.7153          | 6.7657       | 4.3990 | -0.3520          | 2.6126       | 4.5555 | -1.5450          | 2.0895       |
| $S_{t,7}^{\text{sw}}$| 4.3750       | 1.1381           | 2.9238       | 4.5292 | 0.2777           | 0.9860       | 4.5485 | 0.0680           | 0.9686       |
| $S_{t,8}^{\text{sw}}$| 4.5829       | -2.1251          | 7.1017       | 4.4431 | -0.5275          | 3.0179       | 4.6282 | -2.0229          | 2.4208       |
| $S_{t,9}^{\text{sw}}$| 4.4795       | -0.4644          | 7.6952       | 4.5050 | -1.6695          | 3.2569       | 4.8155 | -3.2856          | 3.0981       |

FACTOR-MIDAS-POET Method

| $K = 3$ m          | 4.3536       | 4.1701          | 4.2871       | 4.3858 | 13.4000          | 4.3858       | 13.8514 |
| $K = 6$ m          | 4.3482       | 1.3738          | 4.2596       | 4.4069 | 12.9060          | 0.1706       | 13.3304 |
| $K = 9$ m          | 4.3475       | 1.5734          | 4.1916       | 4.4064 | 12.5355          | 0.0551       | 13.0920 |
| $K = 1$ y          | 4.3467       | 1.6594          | 4.2397       | 4.4041 | 12.2162          | -0.0793      | 12.8645 |

https://doi.org/10.1371/journal.pone.0249665.t009

Table 10. Out-of-sample performances of portfolios with $T = 3$ m.

|                      | daily return | intraday return (with overnight) | intraday return (without overnight) |
|----------------------|--------------|---------------------------------|-----------------------------------|
|                      | SD × 100     | One-tail t test | Turnover (%) | SD × 100 | One-tail t test | Turnover (%) | SD × 100 | One-tail t test | Turnover (%) |
| $S_{t,1}$            | 7.5510       | -9.9487          | 54.5114      | 4.7384 | -0.1167          | 9.3569       | 5.1382 | 0.2119          | 10.7838      |
| $S_{t,2}$            | 4.3720       | 1.3574          | 1.6026       | 4.6092 | -2.6760          | 1.1474       | 4.7990 | -3.1201          | 1.1023       |
| $S_{t,3}^{\text{sw}}$| 5.7823       | -4.8004         | 24.0316      | 4.7317 | -1.4858          | 6.9206       | 4.9736 | -3.0404          | 5.5949       |
| $S_{t,4}^{\text{sw}}$| 4.6812       | -1.1423         | 13.8312      | 4.4119 | 1.1202           | 1.4502       | 4.4645 | -0.1240          | 1.1311       |
| $S_{t,5}^{\text{sw}}$| 5.4748       | -3.7443         | 21.6904      | 4.2959 | 2.1670           | 3.4687       | 4.3263 | 1.3167           | 2.3255       |
| $S_{t,6}^{\text{sw}}$| 4.6957       | -0.6223         | 13.1463      | 4.3981 | 0.6004           | 1.1569       | 4.4206 | -0.7299          | 1.0495       |
| $S_{t,7}^{\text{sw}}$| 5.7178       | -4.4180         | 23.6444      | 4.3432 | 1.8964           | 4.7818       | 4.3408 | 0.8790           | 3.0793       |
| $S_{t,8}^{\text{sw}}$| 5.2228       | -3.1534         | 24.9688      | 4.5047 | -0.6196          | 6.0299       | 4.7502 | -3.4748          | 5.3643       |

FACTOR-MIDAS-POET Method

| $K = 3$ m          | 4.5092       | 7.5565          | 4.3852       | 14.8179 | 4.4330          | 14.2007      |
| $K = 6$ m          | 4.5023       | 0.0629          | 7.6250       | 4.3960 | -0.1793         | 14.8875      | 4.4467 | -0.3520          | 13.9836      |
| $K = 9$ m          | 4.4996       | -0.3466         | 7.6462       | 4.4041 | -0.9606         | 14.6158      | 4.4541 | -0.6767          | 13.7511      |
| $K = 1$ y          | 4.4993       | -0.7947         | 7.6341       | 4.4032 | -0.7387         | 14.0725      | 4.4540 | -0.6693          | 13.1841      |

https://doi.org/10.1371/journal.pone.0249665.t010
Supporting information

S1 Dataset.
(XLSX)

Author Contributions

Conceptualization: Xiangyu Cui.
Formal analysis: Xiangyu Cui, Xuan Zhang.
Funding acquisition: Xiangyu Cui, Xuan Zhang.
Methodology: Xiangyu Cui, Xuan Zhang.
Validation: Xiangyu Cui.
Writing – original draft: Xiangyu Cui, Xuan Zhang.
Writing – review & editing: Xiangyu Cui, Xuan Zhang.

References

1. Rudolf M, Wolter H, and Zimmermann H. A linear model for tracking error minimization. Journal of Banking and Finance, 23(1):85–103, 1999. https://doi.org/10.1016/S0378-4266(98)00076-4
2. Jansen R and Dijk RV. Optimal benchmark tracking with small portfolios. The journal of portfolio management, 28(2):33–39, 2002. https://doi.org/10.3905/jpm.2002.319830
3. Coleman TF, Li Y, and Henniger J. Minimizing tracking error while restricting the number of assets. Journal of Risk, 8(4):33, 2006. https://doi.org/10.21314/JOR.2006.134
4. Corielli F and Marcellino M. Factor based index tracking. Journal of Banking & Finance, 30(8):2215–2233, 2006. https://doi.org/10.1016/j.jbankfin.2005.07.012
5. Liu Q. On portfolio optimization: How and when do we benefit from high-frequency data? Journal of Applied Econometrics, 24(4):560–582, 2009. https://doi.org/10.1002/jae.1062
6. Canakgoz NA and Beasley JE. Mixed-integer programming approaches for index tracking and enhanced indexation. European Journal of Operational Research, 196(1):384–399, 2009. https://doi.org/10.1016/j.ejor.2008.03.015
7. Guastaroba G and Speranza MG. Kernel search: An application to the index tracking problem. European Journal of Operational Research, 217(1):54–68, 2012. https://doi.org/10.1016/j.ejor.2011.09.004
8. Bollerslev T, Engle RF, and Wooldridge JM. A capital asset pricing model with time-varying covariances. Journal of Political Economy, 96(1):116–131, 1988. https://doi.org/10.1086/261527
9. Ledoit O and Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of empirical finance, 10(5):603–621, 2003. https://doi.org/10.1016/S0927-5398(03)00007-0
10. Fleming J, Kirby C, and Ostdiek B. The economic value of volatility timing using realized volatility. Journal of Financial Economics, 67(3):473–509, 2003. https://doi.org/10.1016/S0304-405X(02)00259-3
11. Merton RC. On estimating the expected return on the market: An exploratory investigation. Journal of financial economics, 8(4):323–361, 1980. https://doi.org/10.1016/0304-405X(80)90007-0
12. Andersen TG and Bollerslev T. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International Economic Review, 39(4):885–905, 1998. https://doi.org/10.2307/2527343
13. Andersen TG, Bollerslev T, Diebold FX, and Ebens H. The distribution of realized stock return volatility. Journal of financial economics, 61(1):43–76, 2001. https://doi.org/10.1016/S0304-405X(01)00055-1
14. Andersen TG, Bollerslev T, Diebold FX, and Labys P. Modelling and forecasting realized volatility. Econometrica, 71(2):579–625, 2003. https://doi.org/10.1111/1468-0288.00418
15. Bandi FM, Russell JR, and Zhu Y. Using high-frequency data in dynamic portfolio choice. Econometric Reviews, 27(1-3):163–196, 2008. https://doi.org/10.1080/07474930701870461
16. Pooter MD, Martens M, and Dijk DV. Predicting the daily covariance matrix for s&p 100 stocks using intraday data but which frequency to use? Econometric Reviews, 27(1-3):199–229, 2008. https://doi.org/10.1080/07474930701873333
26. Baele L, Bekaert G, and Inghelbrecht K. The determinants of stock and bond return comovements.

27. Cohen KJ, Hawawini GA, Maier SF, Schwartz RA, and Whitcomb DK. Friction in the trading process and the identity.

28. Efron B, Morris C. Multivariate empirical bayes and estimation of covariance matrices.

29. Hastie TJ, Tibshirani R, and Friedman J. The elements of statistical learning: Data mining, inference, and prediction.

30. Newey WK and West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

31. Shephard N and Sheppard K. Realising the future: Forecasting with high-frequency-based volatility models.

32. Engle RF, Ghysels E, and Sohn B. Stock market volatility and macroeconomic fundamentals.

33. Marcellino M and Schumacher C. Factor midas for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP.

34. Kim D and Fan J. Factor garch-ito models for high-frequency data with application to volatility matrix prediction.

35. Newey WK and West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.

36. Cohen KJ, Hawawini GA, Maier SF, Schwartz RA, and Whitcomb DK. Friction in the trading process and the estimation of systematic risk.

37. Fan J, Liao Y, and Mincheva M. Large covariance estimation by thresholding principal orthogonal complements.

38. Kouris A, Dotis G, and Markellos RN. Parameter uncertainty in portfolio selection: Shrinking the inverse covariance matrix.

39. Basak GK, Ma T, and Jagannathan R. Assessing the risk in sample minimum risk portfolios.