BACKTESTING

FINAL PRESENTATION

IEOR E4511 - UM4

AGENDA

1 INTRODUCTION

Project Objective

Current Timeline

Current Outputs

2 DATA PREPROCESS

Dataset and Problem

Fill Missing Value with Co-integration

3 BENCHMARK STRATEGY

Dataset Overview

Roadmap

4 FUTURE DIRECTION

Potential Improvements

Business Applications

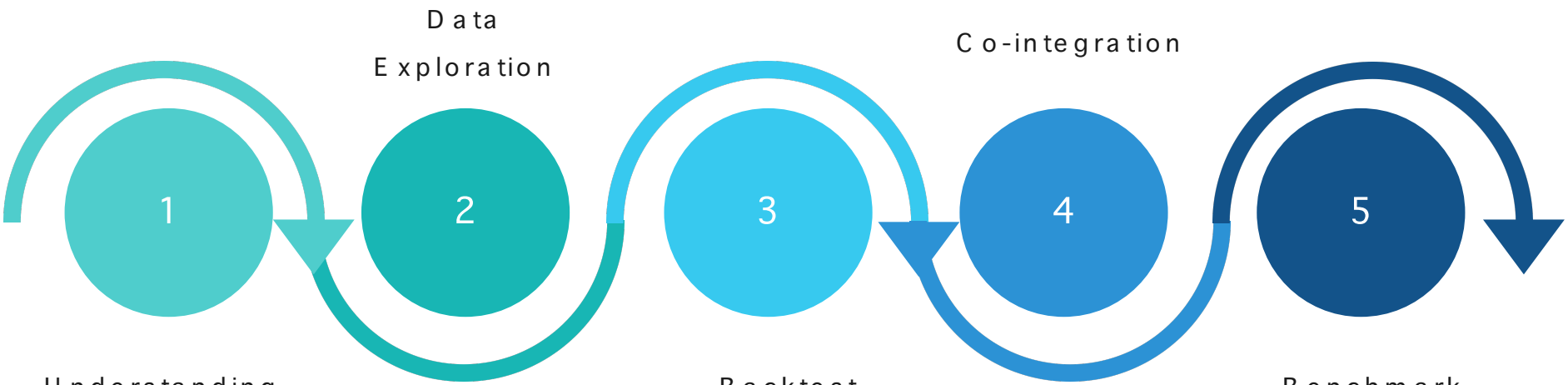
1

INTRODUCTION

OBJECTIVE

Assess the effectiveness of backtesting investment strategies using real-world data, with the goal of developing a user-friendly solution that enables clients to effectively evaluate their investment performance

TimeLine & Output



2

DATA PREPROCESS

datadate	0051B	0226B	0232B	0236B	0397B
2007-01-03	NaN	NaN	NaN	NaN	NaN
2007-01-04	NaN	NaN	NaN	NaN	NaN
2007-01-05	NaN	NaN	NaN	NaN	NaN
2007-01-08	NaN	NaN	NaN	NaN	NaN
2007-01-09	NaN	NaN	NaN	NaN	NaN
2007-01-10	NaN	NaN	NaN	NaN	NaN
2007-01-11	NaN	NaN	NaN	NaN	NaN
2007-01-12	NaN	NaN	NaN	NaN	NaN
2007-01-16	NaN	NaN	NaN	NaN	NaN
2007-01-17	NaN	NaN	NaN	NaN	NaN
2007-01-18	NaN	NaN	NaN	NaN	NaN
2007-01-19	NaN	NaN	NaN	NaN	NaN
2007-01-22	NaN	NaN	NaN	NaN	NaN
2007-01-23	NaN	NaN	NaN	NaN	NaN

datadate	1717B	2386B	3408B	3ACKH	3ADVB
2007-01-03	NaN	NaN	NaN	0.56943	0.045
2007-01-04	NaN	NaN	NaN	0.56943	0.04
2007-01-05	NaN	NaN	NaN	0.56943	0.05
2007-01-08	NaN	NaN	NaN	0.56943	0.05
2007-01-09	NaN	NaN	NaN	0.56943	0.05
2007-01-10	NaN	NaN	NaN	0.56943	0.0475
2007-01-11	NaN	NaN	NaN	0.56943	0.04
2007-01-12	NaN	NaN	NaN	0.56943	0.04
2007-01-16	NaN	NaN	NaN	0.56943	0.04
2007-01-17	NaN	NaN	NaN	0.55163	0.05
2007-01-18	NaN	NaN	NaN	0.55163	0.05
2007-01-19	NaN	NaN	NaN	0.56943	0.05

DATA

THE DATA IS EXPORTED FROM WRDS

General Introduction of Data

- One Identifier: Ticker, Permno…
- 1st Column: Datetime
- Other Columns: Stocks
- Cell Value: Total Price, Adjusted Return

Problem: Missing Value

- Total 31905 stocks
- Remove columns with all missing values
- 15793 stocks left

MISSING VALUE

Why need to handle

- Models might not be able to work with missing values
- Stock with too much missing value might turn into bias
- Affect Model robustness and accuracy

	Pros	Cons
Remove	• Simplicity & Efficiency • Integrity	• Lose Information-non-random • Reduce Sample size
Fill	• No Data Loss & Completeness • Consistency in Time Series	• Potential Bias • Overfitting

Principal

- All Cells are Missing: Remove
- Pre-Start and Post-End Dates: Fill in missing values with 0.
- Within Start and End Dates: Fill by Co-integration

FILL IN MISSING DATA METHOD

1. Edge Case Treatment

2. Linear Interpolation for Missing Data

3. Conservative Approach for Sparse Data

4. Engle-Granger Two-step Cointegration Test for Filling Data

```
function pricesTT = fillEdgeNanWithZero(pricesTT)
    % Get the number of rows and columns in the timetable
    [numRows, numColumns] = size(pricesTT);

    % Iterate over each column (skipping the 'Date' column if it is the first)
    columnsToRemove = []; % Initialize an array to keep track of columns to remove
    for i = 2:numColumns % Assuming the first column is 'Date'
        % Get the data for the current column
        data = pricesTT{:, i};

        % Check if the entire column is NaN
        if all(isnan(data))
            columnsToRemove = [columnsToRemove i]; % Add this column to the list of columns to remov
            continue; % Skip further processing for this column
        end

        % Find indices of leading and trailing NaNs
        leadingNans = find(~isnan(data), 1, 'first') - 1;
        trailingNans = find(~isnan(data), 1, 'last');

        % Fill leading NaNs with zero if they exist
        if leadingNans > 0
            data(1:leadingNans) = 0;
        end

        % Fill trailing NaNs with zero if they exist
        if trailingNans < numRows
            data((trailingNans + 1):end) = 0;
        end

        % Update the timetable with the filled data
        pricesTT{:, i} = data;
    end

    % Remove columns where all data are NaN
    pricesTT(:, columnsToRemove) = [];
end
```

```
function pricesTT = linearInterpolation(pricesTT)
    for i = 1:width(pricesTT)
        if any(isnan(pricesTT{:,i}))
            pricesTT{:,i} = fillmissing(pricesTT{:,i}, 'linear');
        end
    end
end

% Skip filling if there are not enough data points
function pricesTT = skipInsufficientData(pricesTT, minDataPointsRequired)
    pricesTT = pricesTT(:, arrayfun(@(col) sum(~isnan(pricesTT{:,col})) >= minDataPointsRequired, 2:wid
end

function pricesTT = cointegrationFilling(pricesTT)
    % Assume the first column of pricesTT is 'Date' and should be skipped.
    nCols = width(pricesTT);

    % Iterate over all pairs of price series
    for i = 2:nCols-1
        for j = i+1:nCols
            % Extract the two time series
            ts1 = table2array(pricesTT(:, i));
            ts2 = table2array(pricesTT(:, j));

            % Ensure there are no missing values for the cointegration test
            validIndices = ~isnan(ts1) & ~isnan(ts2);
            ts1Clean = ts1(validIndices);
            ts2Clean = ts2(validIndices);

            % Perform the Engle-Granger two-step cointegration test
            if length(ts1Clean) > 1 % Ensuring there is more than one data point
                [hypothesis,~,~,~,reg] = egcitest([ts1Clean, ts2Clean], 'alpha', 0.05);

                % If the series are cointegrated, use the relationship to fill missing values
                if hypothesis == 0 % A cointegration relationship exists
                    beta = reg.coeff(2); % The coefficient from the regression

                    % Fill missing values in ts1 using ts2
                    missingValuesTs1 = isnan(ts1) & ~isnan(ts2);
                    ts1(missingValuesTs1) = (ts2(missingValuesTs1) - reg.coeff(1)) / beta;

                    % Fill missing values in ts2 using ts1
```

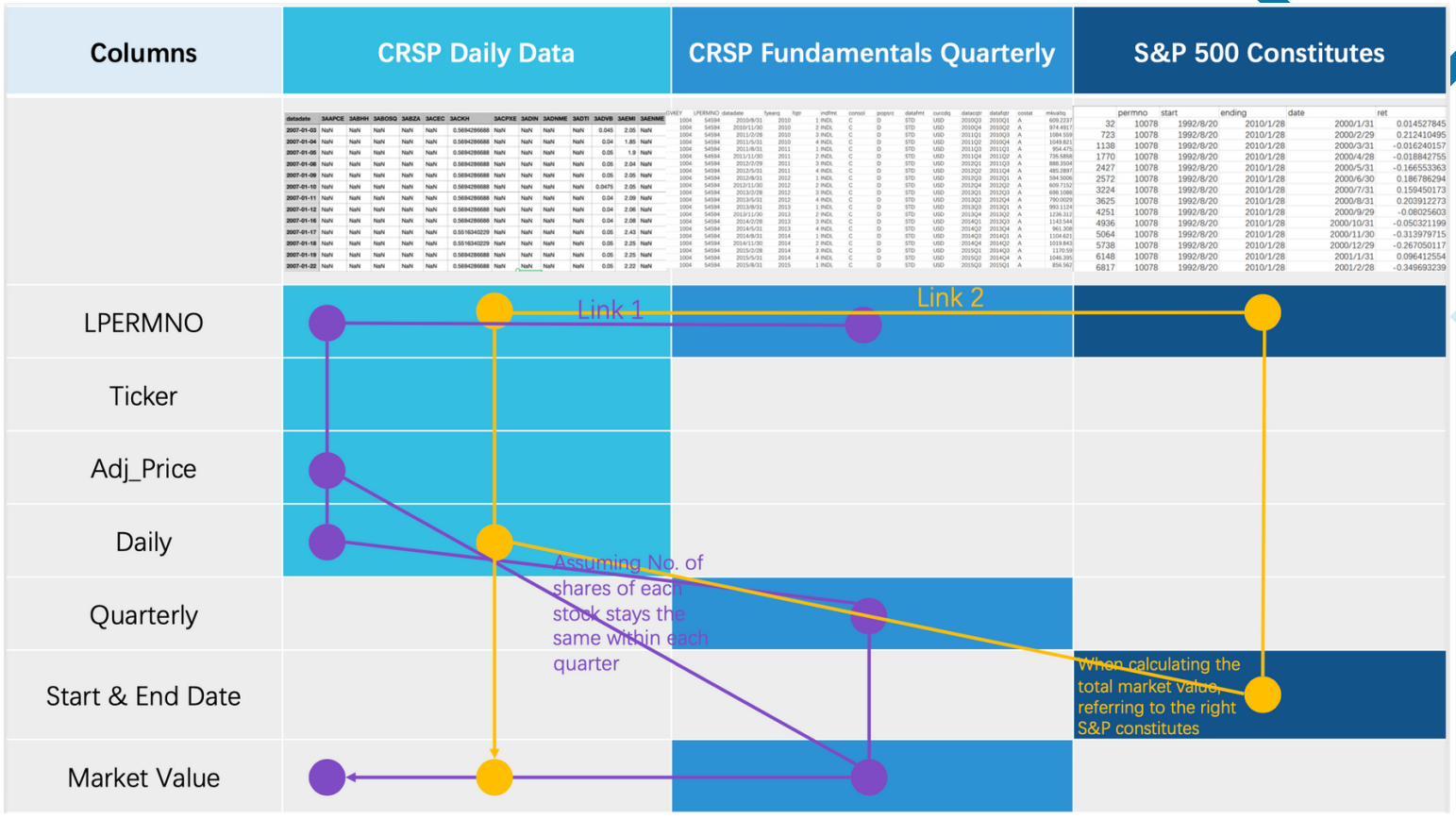
3

BENCHMARK STRATEGY

Equal Weighted
Benchmark Strategy

Market Capitalization
Benchmark Strategy

Overview



USED FOR CALCULATING BENCHMARK WEIGHT

Roadmap

1. Reprocess CRSP Daily Dataset, link with CRSP Quarterly Data to access daily information of stock price and market value

2. Fill in missing data rows, approximate daily market value for each stock

3. Merge S&P 500 Consitutes dataset with Daily dataset to access the daily total market value of each stock in S&P 500 Consitutes

4. Calculate daily total catagorical market value within S&P 500 Constitutes to assign weight score respectively

5. Apply the categorical weight score back to CRSP Daily Dataset to create Benchmark Strategy

1 datadate	2 LPERMNO	3 APx	4 mkvaltq
2010-09-16	10001	35.5920	NaN
2010-09-17	10001	35.2699	NaN
2010-09-20	10001	35.3987	NaN
2010-09-21	10001	35.7079	NaN
2010-09-22	10001	34.5613	NaN
2010-09-23	10001	35.2377	NaN
2010-09-24	10001	35.6564	NaN
2010-09-27	10001	35.7530	NaN
2010-09-28	10001	35.7530	NaN
2010-09-29	10001	35.1088	NaN
2010-09-30	10001	35.8174	67.5429
2010-10-01	10001	35.4309	NaN

Current Stage

Notes for the next group

1. Date: Due to missing cells, the date for total market value might not be in the date list of CRSP daily dataset. Therefore, the better way to merge the monthly market value dataset and CRSP daily dataset could be 1. Filling and turning market value into daily dataset 2. Filling the missing value in CRSP daily dataset.

2. When calculating the daily categorical weight using the market value, it might take very long in locating the adjusted price among all of the stocks using “for” loop. The next group can think of more efficient approach in achieving the value.

4

FUTURE DIRECTION

Business Applications

Potential Improvements

BUSINESS APPLICATION

Cointegration Analysis:

- Provide a better approximation when handling missing values in datasets that comprise an important number of missing values
    - Deleting missing values can lead to bias
    - Estimation of a dynamic relationship between pairs of companies within the same industry

BUSINESS APPLICATION

Benchmark Solution Analysis:

- Approach the actual market performance with reference to market capitalization strategies
- Match the industry category proportion to our client investment information (Match the client's investment portfolio to the industry proportions of the S&P 500)
- Provide clients with a pre-processed solution-- to have raw data processed inside the model according to dataset format and directly apply the benchmark strategy

Improvement Parts

1 COINTEGRATION PAIRS MISSING HANDLING TACTIC

- For better accuracy of identification of cointegrated pairs, we can Identify multiple cointegrating relationships at the same time and thus find the best cointegration pairs

2 FILLING THE COINTEGRATED PAIRS

- Make an assumption of Linear Relationships when filling the cointegrated pairs, and the next stage will be the determinant of coefficient or vector based function's inclusion

3 THE BENCHMARK STRATEGIES

- Filter out or filling the missing category information for potential other companies that want to be included by financial institutions
- High reliance on S&P500 and CRSPAP quarterly dataset in WRDS to propossess with the benchmark strategies

THANK YOU

Q & A