**FOM EOA Project – Dataset information sheet**

Idea: Build a LASSO model for **price discovery in ETFs with international holdings**

Below is an outline of the **key datasets**, **features**, and **potential data sources** to build such a model:

**1. Key Datasets for Training a LASSO Model**

**a. Historical Price Data**

* **ETF price data**: This includes historical prices for the ETF itself, including daily or intraday data.
* **Underlying asset prices**: Prices of the international stocks, bonds, or commodities the ETF holds.
* **Index data**: The indices that the ETF tracks, if it's a passive fund. Examples include the **MSCI World Index** or regional indices like the **FTSE 100**, **DAX**, or **Nikkei 225**.

**Data Sources**:

* Yahoo Finance
* Bloomberg (paid)
* Alpha Vantage (free and paid tiers)
* Quandl

**b. Macroeconomic Data**

* **Global and regional economic indicators**: Inflation rates, GDP growth, employment numbers, and interest rates from the regions where the ETF holds assets.
* **Interest rates**: Domestic and international interest rates such as the **Federal Funds Rate**, **ECB rates**, and **LIBOR rates**.
* **Commodity prices**: Prices of commodities (e.g., oil, gold) if the ETF holds commodities or stocks in commodity-heavy industries.

**Data Sources**:

* World Bank Open Data
* FRED (Federal Reserve Economic Data)
* OECD Statistics
* IMF Data

**c. Exchange Rates and Currency Data**

* **Foreign exchange rates**: For ETFs holding international assets, exchange rate fluctuations can significantly impact the price. Include the relevant exchange rates between the ETF's base currency (e.g., USD) and the currencies of the markets where the ETF invests (e.g., EUR, JPY, GBP, etc.).

**Data Sources**:

* OANDA (free and paid tiers)
* FRED (FX rates)
* XE Currency

**d. Volatility Indices**

* **Global and regional volatility indices**: Include metrics like the **VIX** (CBOE Volatility Index) for general market sentiment and regional volatility indices for specific markets (e.g., **VSTOXX** for Europe, **VXJ** for Japan).

**Data Sources**:

* CBOE (for VIX data)
* Investing.com (global volatility indices)
* Bloomberg (paid)

**e. Sector and Industry Data**

* **Sector indices**: If the ETF focuses on specific sectors (e.g., technology, healthcare), use the corresponding sector indices or industry performance data.

**Data Sources**:

* S&P Dow Jones Indices
* MSCI Sector Indices
* FTSE Russell

**f. Fundamental Data**

* **Financial metrics**: This includes earnings per share (EPS), price-to-earnings (P/E) ratios, dividend yields, and other financial indicators of the underlying assets.
* **Earnings reports**: Quarterly or annual reports for companies held by the ETF.

**Data Sources**:

* Yahoo Finance (company-level financial data)
* Refinitiv (Thomson Reuters)
* FactSet (paid)
* Morningstar (ETF financials)

**g. ETF Flows and Holdings Data**

* **ETF holdings**: Data on the assets or securities held by the ETF, including sector allocations, country exposure, and top holdings. This is essential for tracking how changes in the portfolio affect price.
* **ETF fund flows**: Information on the inflows and outflows of capital in the ETF, as large flows can impact liquidity and price movements.

**Data Sources**:

* ETF.com (ETF holdings)
* Morningstar (ETF analysis)
* Bloomberg Terminal (for ETF flows)

**h. Market Sentiment Data**

* **Social sentiment data**: Social media sentiment, news sentiment, or analyst ratings on the ETF or its underlying assets.
* **News sentiment**: Daily or real-time news about macroeconomic developments or events that could impact global markets, such as geopolitical risks, trade wars, or regulations.

**Data Sources**:

* Twitter APIs (for social media sentiment)
* Refinitiv MarketPsych Indices (paid)
* Google Trends (for public interest in specific markets or ETFs)

**2. Key Features to Consider for the LASSO Model**

When selecting features for your LASSO model, you want to consider variables that are known to drive ETF prices or impact the value of the underlying international holdings. These include:

**a. Price Features**

* **Lagged prices** of the ETF and its holdings (e.g., 1-day, 5-day, 10-day lags).
* **Moving averages**: 50-day, 100-day, and 200-day moving averages for both the ETF and key underlying assets.
* **Daily returns**: Calculate returns for the ETF, underlying securities, and benchmark indices.
* **Price momentum**: Include momentum indicators like the **Relative Strength Index (RSI)** or **Moving Average Convergence Divergence (MACD)**.

**b. Macroeconomic and Currency Factors**

* **Interest rates**: Domestic and international rates, especially from regions where the ETF holds assets.
* **Exchange rates**: Track major currency pairs impacting the ETF (e.g., USD/EUR, USD/JPY).
* **Global economic growth**: GDP growth rates from countries where the ETF’s underlying assets are based.

**c. Volatility and Risk Features**

* **Volatility measures**: VIX index, historical volatility of the ETF and underlying indices.
* **Beta**: The ETF’s beta relative to major indices like the S&P 500 or MSCI World Index.

**d. Sector and Region-Specific Features**

* **Sector performance**: If the ETF focuses on a specific sector, include performance metrics for that sector (e.g., tech stocks in the Nasdaq, energy sector).
* **Country/region-specific performance**: Economic and financial data from countries where the ETF invests.

**e. Sentiment and Flows**

* **Sentiment scores**: Sentiment derived from news or social media regarding the ETF or its underlying holdings.
* **Fund flow data**: Track the inflows and outflows of capital into the ETF to measure investor sentiment and liquidity.

**3. Potential Data Sources**

Here’s a list of reliable sources where you can access the required data for building the model:

* **Yahoo Finance**: Free source for stock, ETF, and index prices, historical data, and financial statements.
* **Bloomberg Terminal**: Comprehensive but expensive option for ETF data, flows, macroeconomic indicators, and sentiment analysis.
* **Quandl**: Offers financial, macroeconomic, and alternative datasets. Some are free, but others require a subscription.
* **FRED (Federal Reserve Economic Data)**: Free source for macroeconomic data, including interest rates, GDP, inflation, and exchange rates.
* **Morningstar**: Provides in-depth ETF data, including holdings, financial ratios, and performance metrics (paid service).
* **Refinitiv**: Extensive paid data provider for financial markets, news sentiment, and global macroeconomic data.
* **OANDA or XE**: Sources for real-time and historical exchange rate data.
* **Investing.com**: Free and premium data on global indices, sector performance, and international markets.
* **CBOE**: For volatility indices like VIX and regional variants.

**Conclusion**

By using LASSO for **price discovery in ETFs with international holdings**, you can identify the most important factors (features) that drive the ETF’s price movements. You’ll need datasets ranging from **historical prices** of the ETF and its underlying assets to **macroeconomic indicators**, **exchange rates**, and **market sentiment** data. The most crucial aspect is to ensure your model includes relevant global and regional factors that can impact an international portfolio, such as **currency risks**, **interest rates**, and **sector performance**.

The variety of data sources listed above should give you a solid foundation to develop a LASSO model for your price discovery task.