Asset Embeddings

Xavier Gabaix   Ralph Koijen   Robert Richmond   Motohiro Yogo

Harvard - Chicago - NYU - Princeton

April 2024

Identifying similar firms

- In economics, we often try to find similar firms or assets.
  - E.g., in terms of growth rates, expected returns, risk, asset substitution, product markets, ...
- Common practice: Use observable characteristics.
  - E.g., industry definitions, accounting data, ...

Identifying similar firms

- In economics, we often try to find similar firms or assets.
  - E.g., in terms of growth rates, expected returns, risk, asset substitution, product markets, ...
- Common practice: Use observable characteristics.
  - E.g., industry definitions, accounting data, ...
- Those characteristics may be quite imperfect.
  - Standardized accounting data are an incomplete summary.
    - E.g., number of subscribers at Netflix, ...
  - New economic environments call for creative, new characteristics.
    - E.g., exposure to COVID-19, growth in intangibles, ...

Identifying similar firms

- In economics, we often try to find similar firms or assets.
  - E.g., in terms of growth rates, expected returns, risk, asset substitution, product markets, ...
- Common practice: Use observable characteristics.
  - E.g., industry definitions, accounting data, ...
- Those characteristics may be quite imperfect.
  - Standardized accounting data are an incomplete summary.
    - E.g., number of subscribers at Netflix, ...
  - New economic environments call for creative, new characteristics.
    - E.g., exposure to COVID-19, growth in intangibles, ...
- This paper: Use asset embeddings to measure firm similarity.

What are embeddings?

- Embeddings: Represent data (e.g., words) as continuous vectors in a potentially high-dimensional space: $x_a \in \mathbb{R}^N$.
- Embeddings play a central role in the development of large language models.
- In NLP, embeddings capture the similarity between words and it allows us to do “math with words:

$$x_{\text{Paris}} - x_{\text{France}} + x_{\text{Spain}} \simeq x_{\text{Madrid}}.$$

WHAT ARE EMBEDDINGS?

- Embeddings: Represent data (e.g., words) as continuous vectors in a potentially high-dimensional space: $x_a \in \mathbb{R}^N$.

- Embeddings play a central role in the development of large language models.

- In NLP, embeddings capture the similarity between words and it allows us to do "math with words:

$$x_{\text{Paris}} - x_{\text{France}} + x_{\text{Spain}} \simeq x_{\text{Madrid}}.$$

- The dense embedding vectors are learned from (lots of) data (not preselected).

- Despite the success of embedding techniques in these fields, their application in finance and economics largely unexplored.

Which data to use to learn embeddings?

- We introduce the concept of asset embeddings.
  - A vector representation per asset that we learn from data.
- Which data to use?

Which data to use to learn embeddings?

- We introduce the concept of asset embeddings.
  - A vector representation per asset that we learn from data.
- Which data to use?
- Our answer: Just like
  - documents organize words in NLP,
  - images organize pixels in vision,
  - songs organize notes in audio,

  investors organize assets in finance and economics.

Which data to use to learn embeddings?

- We introduce the concept of asset embeddings.
  - A vector representation per asset that we learn from data.
- Which data to use?
- Our answer: Just like
  - documents organize words in NLP,
  - images organize pixels in vision,
  - songs organize notes in audio,

  investors organize assets in finance and economics.
- Theoretically, we show how embeddings can be recovered by "inverting the asset demand system."

Which method to learn embeddings?

- Which method to use?

Which method to learn embeddings?

- Which method to use?
- Traditional approach: LSA (Latent Semantic Analysis), which is analogous to PCA/recommender systems.
- The recent ML/AI literature went way beyond that:
  - Context-invariant embeddings: E.g., GloVe and Word2Vec.
  - Embeddings with context: E.g., transformer models (e.g., BERT and GPT).
  - Parameters are estimated using masked language modeling.

Four main contributions

1. Uncover characteristics relevant to investors by "inverting" the asset demand system.
2. Five benchmarks to compare any type of asset embeddings.
   - Benchmarks play a key role in developing GenAI models.
3. Use various language model architectures to learn asset embeddings, including transformer models.
4. Implement the models using 13F and funds data.
   - Observed characteristics and LLM-based embeddings (Cohere and OpenAI) provide a reference point.

Methods to extract embeddings

- We consider the following embedding models:
  1. (Supervised) PCA (recommender systems).
  2. Word2Vec.
  3. Models with attention: Transformer models.
     - We build on the BERT architecture and specialize it to holdings data.

From Word Embeddings to Asset Embeddings

- General approach to estimate language models, such as Word2Vec,[2]
  - Task: Guess masked words.
    - E.g. “Please pass me the _____ and pepper”.
  - Use a context window to maximize the probability of a missing word given the context info:

$$\mathbb{P}(w_a \mid w_c) = \frac{\exp(x_a' x_c)}{\sum_b \exp(x_b' x_c)}.$$

[2] Mikolov, Sutskever, Chen, Corrado, Dean (2013a, b).

From Word Embeddings to Asset Embeddings

- General approach to estimate language models, such as Word2Vec,[2]
  - Task: Guess masked words.
    - E.g. "Please pass me the _____ and pepper".
  - Use a context window to maximize the probability of a missing word given the context info:

$$\mathbb{P}(w_a \mid w_c) = \frac{\exp(x_a' x_c)}{\sum_b \exp(x_b' x_c)}.$$

- Using holdings data:
  - Sentences $\Rightarrow$ Investors.
  - Words $\Rightarrow$ Assets.
  - Task: Guess masked assets.

[2] Mikolov, Sutskever, Chen, Corrado, Dean (2013a, b).

Masked asset modeling

- Example: The ARKK ETF in July 2023:

Holdings Data - ARKK
As of 07/07/2023

ARK INVEST

ARKK
ARK Innovation ETF

	Company	Ticker	CUSIP	Shares	Market Value ($)	Weight (%)
1	TESLA INC	**TSLA**	88160R101	3,496,872	$967,024,982.88	12.43%
2	COINBASE GLOBAL INC -CLASS A	**COIN**	19260Q107	7,945,138	$620,515,277.80	7.98%
3	ROKU INC	**ROKU**	77543R102	8,865,426	$546,110,241.60	7.02%
4	~~ZOOM VIDEO COMMUNICATIONS-A~~	~~**ZM**~~	~~98980L101~~	~~8,258,591~~	~~$534,248,251.79~~	~~6.87%~~
5	UIPATH INC - CLASS A	**PATH**	90364P105	28,152,366	$463,106,420.70	5.95%
6	BLOCK INC	**SQ**	852234103	7,069,493	$456,759,942.73	5.87%
7	EXACT SCIENCES CORP	**EXAS**	30063P105	4,031,264	$368,739,718.08	4.74%
8	UNITY SOFTWARE INC	**U**	91332U101	8,350,868	$338,627,697.40	4.35%
9	SHOPIFY INC - CLASS A	**SHOP**	82509L107	5,430,238	$335,751,615.54	4.32%
10	DRAFTKINGS INC-CL A	**DKNG UW**	26142V105	12,035,607	$303,658,364.61	3.90%

Data

- Holdings data from FactSet:
  - 13F filings.
  - Mutual funds, ETFs, closed-end funds, variable annuity funds.
- Sample construction:
  - 2000.Q1 - 2022.Q4.
  - Remove nano and micro caps,.
  - Keep investors (stocks) with at least 20 positions (investors).
- Accounting data and stock returns from CRSP / Compustat, using the Jensen, Kelly, and Pedersen (2023) construction.

Representing firms: The competitors

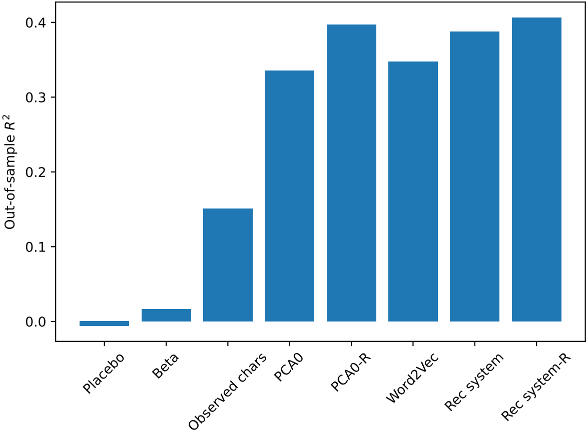
- Observed characteristics:
  - Market cap, book-to-market, asset growth, profitability, beta, momentum.
- Holdings-based embeddings.
- LLM-based embeddings from Cohere and OpenAI.
  - Cohere:
    - Model: `embed-english-v3.0`.
    - Reduce the dimensionality using UMAP.
  - OpenAI:
    - Model: `text-embedding-3-large`.
    - Download the embeddings for the appropriate size.

Evaluating asset embeddings: Benchmarks

- In ML: Benchmark competitions identify the best performing models, and give metrics for success.
  - E.g. ImageNet to measure improvement in performance in vision tasks.
- We could do the same in finance.
- We consider five benchmarks
  1. Explaining valuations.
  2. ETF similarity.
  3. Predicting announcement returns.
  4. Missing characteristics.
  5. Predicting demand.

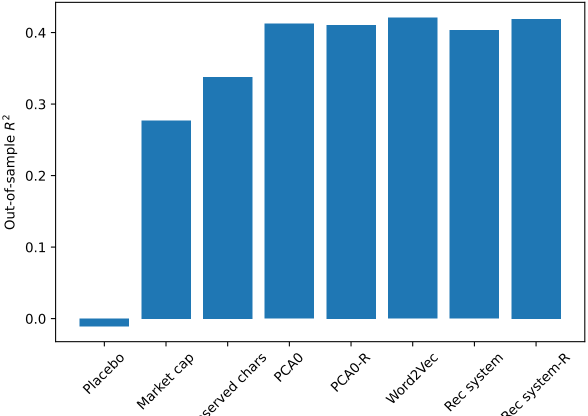
BM 1: Explaining valuations

- Regress $m_{at} = \beta_0 + \beta_1 b_{at} + m_{at}^{\perp}$.
- Fit the valuation residual, $m_{at}^{\perp}$, on x_{at} for 80% of the sample and evaluate, out of sample (OOS), on the remaining 20% using the R^2.



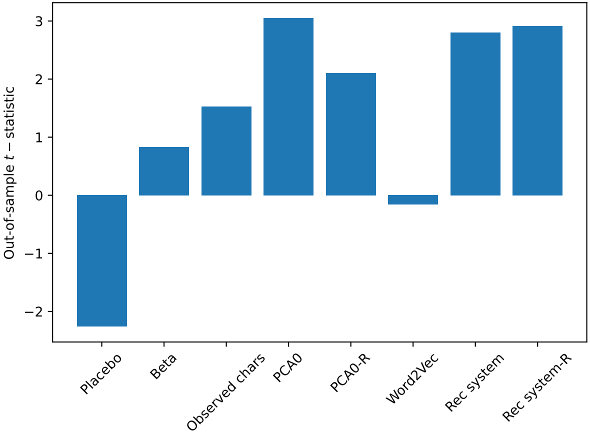
BM 2: ETF SIMILARITY

- We estimate a logit model to predict whether a stock is in a given focused ETF (between 100 and 250 stocks), and compute average performance across ETFs.
- Use 80% of the data (positive and negative samples) to estimate the model and compute the pseudo R2 for the remaining 20% of the data OOS.



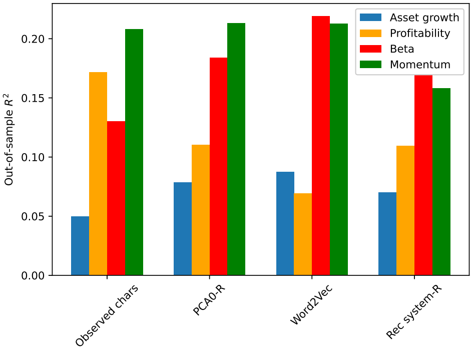
BM 3: Predicting announcement returns

- Regress $CAR3_{at}$ on $x_{a,t-1}$ for the first 80% of announcement days in an earnings quarters and predict the sign of the returns for the remaining 20% OOS. We report the t-stat on slope.



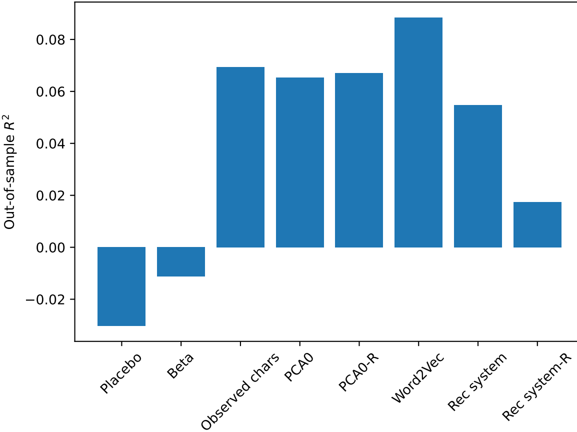
BM 4: Missing characteristics

- Similar to explaining valuations but now with characteristics for asset growth, profitability, momentum, and beta.
  - Use 80% to estimate the link between the characteristic and embeddings to explain 20% OOS.
- To explain missing characteristics, we use other characteristics + size and book/market or large embedding models.
- In progress: Use supervised, regularized recommender systems.



BM 5: Predicting demand

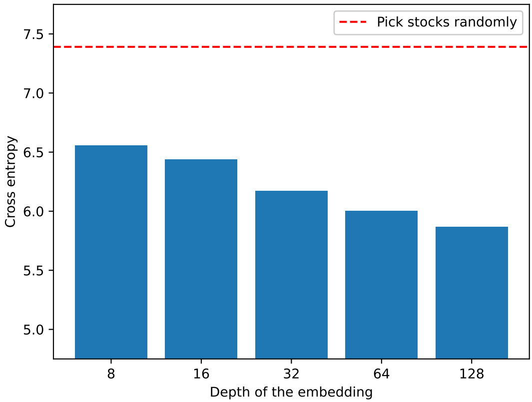
- For investors with more than 250 stocks, we compute their rebalancing (excluding price effects).
- Using 80% of the sample, explain their rebalancing for the remaining 20% OOS.



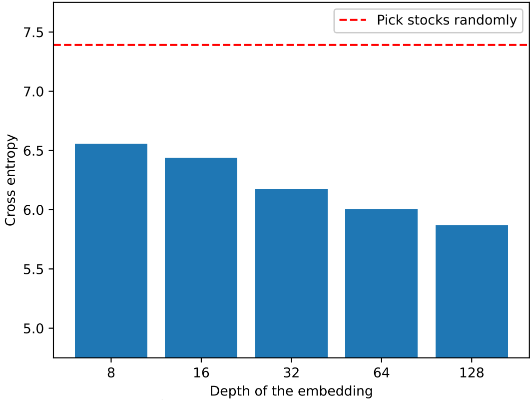
Evaluating transformer models

- AssetBERT generates a distribution over masked assets.
- We consider an initial estimate of the model for a single quarter, 2019.Q4.
- We evaluate the model relative to observed embeddings and the asset embeddings recovered from the recommender system.
- Draw 1,000 managers (with replacement) and, for each manager, mask a stock that we try to predict.

Out-of-sample results AssetBERT



Out-of-sample results AssetBERT



- Relative entropy of
  - Observable characteristics: -0.35 $\Rightarrow$ Likelihood ratio $= 1.41$
  - AssetBERT: -1.67 $\Rightarrow$ Likelihood ratio $= 5.31$
- AssetBERT is 3.71 times more accurate than observable characteristics.

Conclusions

- Recent advances in AI/ML can be applied to economics and finance via asset embeddings.
- We provide a micro foundation for using holdings data.
- We adjust methods that have been successful in related areas (e.g., NLP, vision, ...) to economics:
  - LSA, Word2Vec, Supervised PCA, and Transformer models.
- We show that asset embeddings outperform observable characteristics across a range of benchmarks.