Discussion of:

Asset Embeddings

Xavier Gabaix, Ralph S.J. Koijen, Robert J. Richmond, and Motohiro Yogo

Kent Daniel†

†Columbia Business School & NBER

SQA/CFA-NY Joint Conference on Data Science in Finance
January 8, 2026

Inferring Asset Return Moments from Investor Holdings

$$\mathbf{w}_i^* = (\gamma_i \boldsymbol{\Sigma})^{-1} (\boldsymbol{\mu} - \mathbf{1} \cdot r_f)$$

- There is a long tradition in finance of using holdings to infer asset information.
- Markowitz (1952) showed, given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, how to find $\mathbf{w}_i^*$
- In the 1960s, Sharpe, Lintner, and Black argued that, if investors are, on average, smart, and hold MVE portfolios then:

$$\mathbf{w}_i^* \propto \mathbf{w}_m$$

- That is, we don't need to calculate $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$, we can hold the market.
- Similarly, if we know $\boldsymbol{\Sigma}$, and want to calculate $\boldsymbol{\mu}$, we can:

$$\boldsymbol{\mu} = \mathbf{1} \cdot r_f + \gamma_m \boldsymbol{\Sigma} \mathbf{w}_m \qquad \left(\text{or} \ \ \mathbb{E}[r_a] = r_f + \beta_a \left(\mathbb{E}[r_m] - r_f\right)\right)$$

Inferring Asset Return Moments from Investor Holdings

$$\mathbf{w}_i^* = (\gamma_i \boldsymbol{\Sigma})^{-1} (\boldsymbol{\mu} - \mathbf{1} \cdot r_f)$$

- There is a long tradition in finance of using holdings to infer asset information.
- Markowitz (1952) showed, given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, how to find $\mathbf{w}_i^*$
- In the 1960s, Sharpe, Lintner, and Black argued that, if investors are, on average, smart, and hold MVE portfolios then:

$$\mathbf{w}_i^* \propto \mathbf{w}_m$$

- That is, we don't need to calculate $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$, we can hold the market.
- Similarly, if we know $\boldsymbol{\Sigma}$, and want to calculate $\boldsymbol{\mu}$, we can:

$$\boldsymbol{\mu} = \mathbf{1} \cdot r_f + \gamma_m \boldsymbol{\Sigma} \mathbf{w}_m \qquad (\text{or} \quad \mathbb{E}[r_a] = r_f + \beta_a (\mathbb{E}[r_m] - r_f))$$

Inferring Asset Return Moments from Investor Holdings

$$\mathbf{w}_i^* = (\gamma_i \mathbf{\Sigma})^{-1} (\boldsymbol{\mu} - \mathbf{1} \cdot r_f)$$

- There is a long tradition in finance of using holdings to infer asset information.
- Markowitz (1952) showed, given $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$, how to find $\mathbf{w}_i^*$
- In the 1960s, Sharpe, Lintner, and Black argued that, if investors are, on average, smart, and hold MVE portfolios then:

$$\mathbf{w}_i^* \propto \mathbf{w}_m$$

- That is, we don't need to calculate $\boldsymbol{\mu}$ or $\mathbf{\Sigma}$, we can hold the market.
- Similarly, if we know $\mathbf{\Sigma}$, and want to calculate $\boldsymbol{\mu}$, we can:

$$\boldsymbol{\mu} = \mathbf{1} \cdot r_f + \gamma_m \mathbf{\Sigma} \mathbf{w}_m \qquad \left(\text{or} \quad \mathbb{E}[r_a] = r_f + \beta_a \left(\mathbb{E}[r_m] - r_f\right)\right)$$

Using Disaggregated Holding Data

- We are now in a world where we think the market is perhaps not so smart!
  - See, e.g., Daniel, Klos, and Rottke (2025)
- *What can we do in this world?*
  - One response is to use various data sources to estimate $\boldsymbol{\mu}$, use BARRA or Axioma to get $\boldsymbol{\Sigma}$, and build a portfolio with positive alpha.
- Starting with Koijen and Yogo (2019), these authors have started exploring how we might use holdings information in this post-EMH world.
- This paper explores the use of AI/LLM techniques to extract meaning from asset holdings by different funds/investors.

Embeddings

- The term "embedding" were coined in Bengio et al. (2003).
  - The idea was to develop a low-dimensional representation of words or "tokens".
  - The roots go back to the work of the linguist John Rupert Firth in the 1950s, who argued that "…a word is characterized by the company it keeps"
- In the 1980s, Latent Semantic Analysis used Singular Value Decomposition to reduce word-count tables into sparse low dimensional numerical representations (like recommender system here).
- In 2013 Word2Vec introduced modern embeddings, based on a 1-hidden-layer neural network.
  - A team at Google trained a 300 neuron network, based on the 100B tokens in a Google News dataset, generating the embeddings for 3 million tokens.
- Modern implementations use a transformer architecture (Vaswani et al., 2017) to generate contextual embeddings.
  - A (river) "bank" will have a different embedding than a (financial) "bank".

Embeddings

- The term “embedding” were coined in Bengio et al. (2003).
  - The idea was to develop a low-dimensional representation of words or “tokens”.
  - The roots go back to the work of the linguist John Rupert Firth in the 1950s, who argued that “. . . a word is characterized by the company it keeps”
- In the 1980s, Latent Semantic Analysis used Singular Value Decomposition to reduce word-count tables into sparse low dimensional numerical representations (like recommender system here).
- In 2013 Word2Vec introduced modern embeddings, based on a 1-hidden-layer neural network.
  - A team at Google trained a 300 neuron network, based on the 100B tokens in a Google News dataset, generating the embeddings for 3 million tokens.
- Modern implementations use a transformer architecture (Vaswani et al., 2017) to generate contextual embeddings.
  - A (river) “bank” will have a different embedding than a (financial) “bank”.

How Embeddings Learn "Meaning"

- The network learns that ice cream and gelato are similar by processing billions of sentences.
    - In training, it figures out that both words frequently appear near context words like *sweet*, *dessert*, *frozen*, and *scoop*.
- Because they "keep the same company," their mathematical representations are pulled together in the 300-dimensional vector space.
    - As a result, the cosine similarity of their embeddings is close to 1.

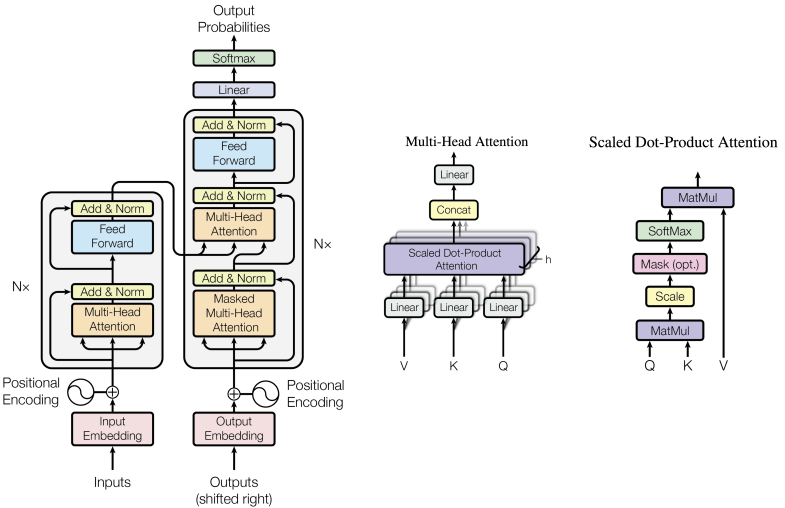
Language embedding examples (Word2Vec)

```
> import gensim; import gensim.downloader as api
> wv = api.load('word2vec-google-news-300')
> wv['ice_cream']

[0.125977 0.029785 0.008606 0.013964 ....  -0.279297 -0.085937 0.091308 0.251953]

> wv.similarity('ice_cream','gelato')

0.6252224

> wv.most_similar(positive=['ice_cream','Italy'],negative=['US'])

[('gelato', 0.579361), ...  ('zeppole', 0.492295), ...  ('cannoli', 0.485007)]

> wv.most_similar(positive=['grilled_cheese','France'],negative=['US'])

[('jambon', 0.529636), ('croque_monsieur', 0.513646)]

> wv.most_similar(positive=['king','woman'],negative=['man'])

[('queen', 0.711819), ('monarch', 0.618967)]
```

Embeddings

- In the 1980s, Latent Semantic Analysis used Singular Value Decomposition to reduce word-count tables into sparse low dimensional numerical representations.
- In 2013 Word2Vec introduced modern embeddings, based on a 1-hidden-layer neural network.
  - A team at Google trained a 300 neuron network, based on the 100B tokens in a Google News dataset, generating the embeddings for 3 million tokens.
- Modern implementations use a transformer architecture (Vaswani et al., 2017) to generate contextual embeddings.
  - A (river) "bank" will have a different embedding than a (financial) "bank".
  - Current Transformers (ChatGPT, Claude, Gemini, ...) are trained on "tens of trillions" of tokens, and use flexible embedding vectors with dimensions of up to 12,288 ($= 2^{12} \times 3$).

The Transformer Architecture



From Vaswani et.al.(2017). See also `https://www.youtube.com/watch?v=sznZ78HquPc`

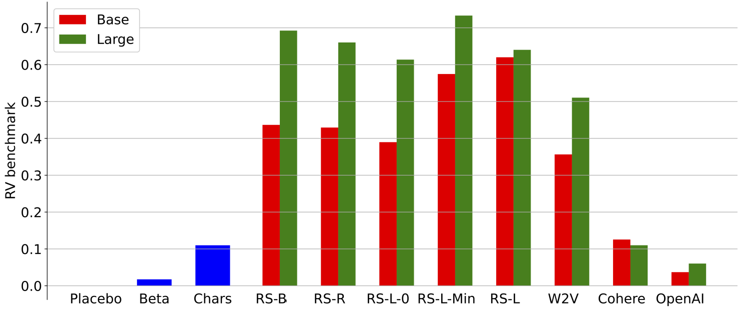
Asset Embeddings

- In this paper, asset and investor embeddings are calculated which are based on:
  - which assets are similar, based on the ordering of assets in funds (PS-BERT)
  - which funds/owners are similar, based on the ownership shares (OS-BERT)
- Currently, PS-BERT and OS-BERT are just trained separately, on the cross-section.
  - The appendix proposes an integrated model of asset- and investor-embeddings.

Model: Embedding Based Asset Pricing Model

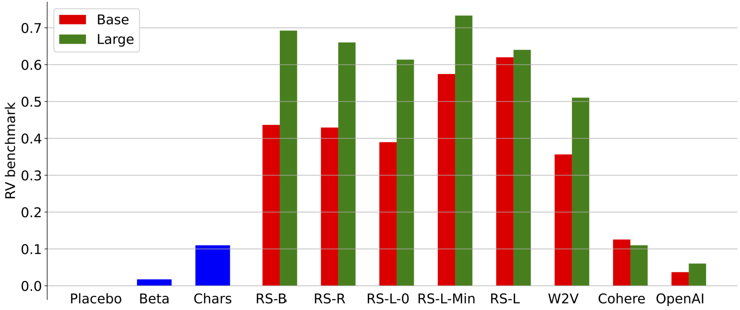
- Argues that embeddings contain all relevant info about firms
- Model premise is that:
  1. log dollar holding of an asset as the dot product of the investor embedding and the asset embedding.
  2. asset embeddings are latent characteristics that capture differences in expected profitability or risk exposure across assets.
  3. investor embeddings capture heterogeneity in preferences for the asset embeddings across investors.

Valuation Ratios



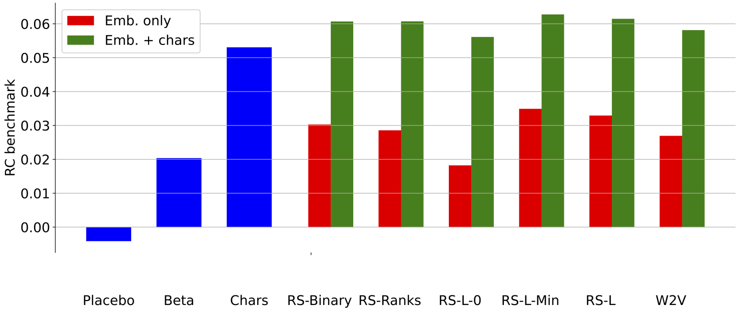
- Chars here are β, asset growth, profitability, and Div/AT.
- Valuation ratios are calculated as $p_{at} = \gamma_t b_{at} + \alpha_t + p_{at}^{\perp}$, where RV is the oos R^2 from a ridge-regression of $p_{at}^{\perp}$ on betas, chars, or embeddings.
- Funds can and do select stocks based on P/B, ...
- *Can embeddings forecast future $\Delta P/Bs$?*

Valuation Ratios



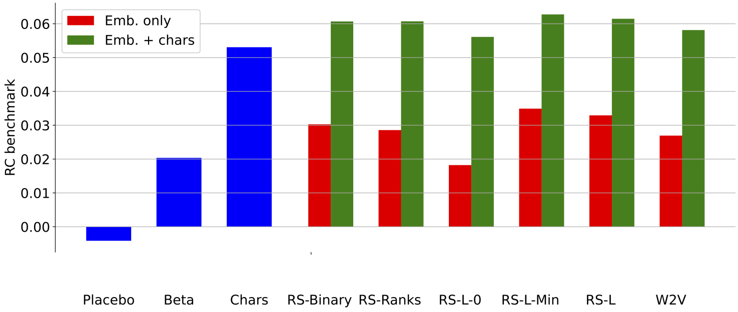
- Chars here are β, asset growth, profitability, and Div/AT.
- Valuation ratios are calculated as $p_{at} = \gamma_t b_{at} + \alpha_t + p_{at}^{\perp}$, where RV is the oos R^2 from a ridge-regression of $p_{at}^{\perp}$ on betas, chars, or embeddings.
- Funds can and do select stocks based on P/B, ...
- *Can embeddings forecast future $\Delta P/Bs$?*

Covariances



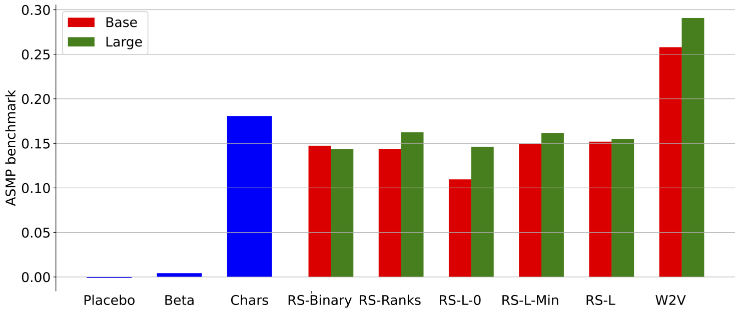
- Chars are β, $\log(ME)$, $\log(B/M)$, asset growth, profitability, and momentum.
- Can the embeddings beat risk models/historical covariance structure?
  - BARRA, Axioma, or Daniel, Mota, Rottke, and Santos (2020).

Covariances



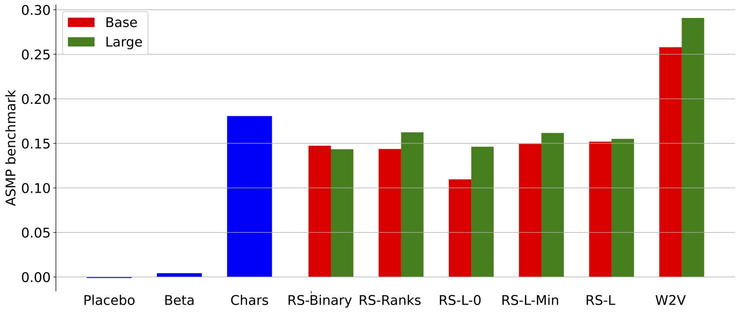
- Chars are β, $\log(ME)$, $\log(B/M)$, asset growth, profitability, and momentum.
- Can the embeddings beat risk models/historical covariance structure?
  - BARRA, Axioma, or Daniel, Mota, Rottke, and Santos (2020).

Masked Portfolio Holdings



- Measures ability to predict masked assets in MFs, ETFs, HFs?
- Chars: β, $\log(ME)$, $\log(B/M)$, asset growth, profitability, and momentum.
- Can the embeddings can forecast future weight changes/returns?
    - Do embeddings forecast the future?

Masked Portfolio Holdings



- Measures ability to predict masked assets in MFs, ETFs, HFs?
- Chars: β, $\log(ME)$, $\log(B/M)$, asset growth, profitability, and momentum.
- Can the embeddings can forecast future weight changes/returns?
  - Do embeddings forecast the future?

Conclusions

- These embedding techniques explored here can potentially really help to extract useful information about the cross-section of risk and expected returns, and other attributes (e.g., liquidity)
- Relation to quant overlay strategies in firms like Point72, etc.
- This paper has made big strides in developing these techniques.
- Could more data be fed into these systems?
  - Prospectuses (Abis, 2020) (Sec 8.3), 10-Ks (Cohen, Malloy, and Nguyen, 2020), earnings calls, news stories (Sec. 9.5)
  - N-PORT data (incl. derivatives and short positions)
- The paper's stance that holdings subsume all other information seems misguided.
  - Section 9 discusses approaches to expanding data, which is great.
- It would be good to challenge the models more powerful tests.

References I

Abis, Simona, 2020, Man vs. machine: Quantitative and discretionary equity management, Available at SSRN, Abstract 3717371.

Bengio, Yoshua, Jean-Françcois Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Roux, and Marie Ouimet, 2003, Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering, *Advances in neural information processing systems* 16.

Black, Fischer, 1972, Capital market equilibrium with restricted borrowing, *Journal of Business* 45, 444–455.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2020, Lazy prices, *The Journal of Finance* 75, 1371–1415.

Daniel, Kent, Alexander Klos, and Simon Rottke, 2025, Inefficiencies in the securities lending market, Columbia Business School working paper.

Daniel, Kent, Lira Mota, Simon Rottke, and Tano Santos, 2020, The cross section of risk and return, *The Review of Financial Studies* 33, 1927–1979.

Koijen, Ralph S.J., and Motohiro Yogo, 2019, A demand system approach to asset pricing, *Journal of Political Economy* 127, 1475–1515.

Lintner, John, 1965, Security prices, risk and maximal gains from diversification, *Journal of Finance* 20, 587–616.

References II

Markowitz, Harry M., 1952, Portfolio selection, *Journal of Finance* 7, 77–91.

Sharpe, William F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance* 19, 425–442.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, 2017, Attention is all you need, *Advances in Neural Information Processing Systems* .