PROJECT REPORT: FACTOR CLUSTERING MODEL IN EMPIRICAL ASSET PRICING

TIANFENG WANG

DEPARTMENT OF ECONOMICS

COLLEGE OF SOCIAL SCIENCES AND HUMANITIES

NORTHEASTERN UNIVERSITY, SEATTLE

BACKGROUND & MOTIVATION

- Prevalence in empirical pricing of Factor models as extensions from the Capital Asset Pricing Model (CAPM)
- Replication Crisis in finance
  - Internal Validity
  - External Validity
- This project focuses on the discovery of significant global factors using a Bayesian model of factor replication utilizing linear approach such as Principal Component Regression, and non-linear approach such as Hierarchical Agglomerative Clustering

FOUNDATIONS – FACTOR MODEL

- An extension from the Capital Asset Pricing Model:

$$ER_i = R_f + \beta_i(ER_m - R_f)$$

- Where:
  - ***ER_i***= Expected rate of return
  - ***R_f*** = Risk-free rate
  - ***ß*** = Factor's coefficient (sensitivity)
  - ***Market Factor: (r_m – r_f)*** = Market risk premium

FAMA-FRENCH 3-FACTOR MODEL

$$r = r_f + \beta_1(r_m - r_f) + \beta_2(SMB) + \beta_3(HML) + \varepsilon$$

- Where:
  - ***r*** = Expected rate of return
  - ***r_f*** = Risk-free rate
  - ***ß*** = Factor's coefficient (sensitivity)
  - ***Market Factor: (r_m – r_f)*** = Market risk premium
  - ***Size Factor: SMB (Small Minus Big)*** = Historic excess returns of small-cap companies over large-cap companies
  - ***Value Factor: HML (High Minus Low)*** = Historic excess returns of value stocks (high book-to-price ratio) over growth stocks (low book-to-price ratio)

HIERARCHICAL BAYESIAN MODEL

$$f_t = \alpha + \beta r_t^m + \varepsilon_t$$

- Where:
  - $\boldsymbol{f_t}$ = Factor's net performance
  - $\boldsymbol{r^m{}_t}$= Excess market factor
  - $\boldsymbol{\alpha}$ = Posterior $\boldsymbol{\alpha}$, (Prior $\boldsymbol{\alpha \sim N(0,\tau^2)}$)
  - $\boldsymbol{\varepsilon_t}$ = Error term, $\boldsymbol{\varepsilon_t \sim N(0,\sigma^2)}$
  - $\boldsymbol{\beta}$ = Factor's coefficient (sensitivity)

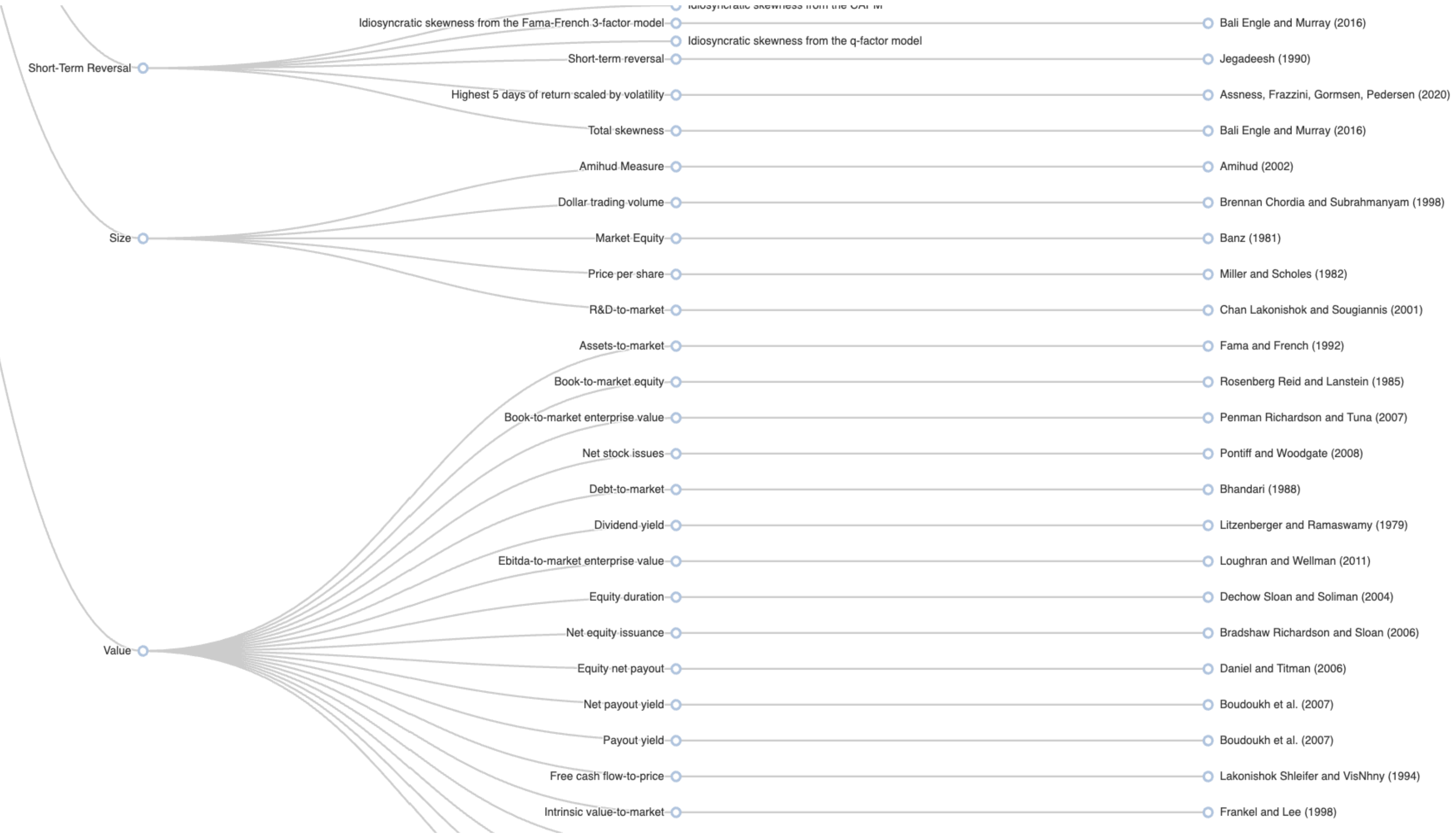
MULTI-LEVEL HIERARCHICAL BAYESIAN – POSTERIOR

$$\alpha^i = \alpha^o + c^j + s^n + w^i$$

- Where:
  - $\boldsymbol{\alpha^i}$ = Individual factor i
  - $\boldsymbol{\alpha^0}$ = Component common to all factors
  - $\boldsymbol{c^j}$ = Cluster specific component, $\boldsymbol{c^j \sim N(0,\tau_c^2)}$
  - $\boldsymbol{s^n}$= Signal specific component, $\boldsymbol{s^n \sim N(0,\tau_n^2)}$
  - $\boldsymbol{w^i}$= Idiosyncratic component, $\boldsymbol{w^i \sim N(0,\tau_w^2)}$

FACTOR ZOO – CLUSTERING

- Jensen, T. I., Kelly, B. T. & Pedersen, L. H. (n.d.). (2023). Is there a replication crisis in finance?, *The Journal of Finance (forthcoming)* **78** (5): 2465–2518.
- Divides factors into 13 clustered themes
- Source



FACTOR ZOO – CLUSTERING

- Jensen, T. I., Kelly, B. T. & Pedersen, L. H. (n.d.). (2023). Is there a replication crisis in finance?, *The Journal of Finance (forthcoming)* **78** (5): 2465–2518.
- Divides factors into 13 clustered themes
- Source

$$\alpha^i = \alpha^o + \boxed{c^j} + \boxed{s^n} + w^i$$

DATASET

- Global dataset with 153 factors in 93 countries (subset of Global Factor dataset, Jensen, Kelly, and Pedersen (2022)), with direction and magnitude
- Global Market Returns Data
- Country classification data for regional analysis(US, World, Frontier, Developed…)
- Factor Returns HML(High Minus Low) Data

Data > hml.csv

```
excntry,characteristic,eom,signal,n_stocks,n_stocks_min,ret_ew,ret_vw,ret_vw_cap
ARE,age,2006-02-28,228,56,1,0.132139149458182,0.102877466792561,0.0974309061180571
ARE,age,2006-03-31,228,57,1,-0.136900240376786,-0.0995254206112998,-0.113379301723879
ARE,age,2006-04-30,228,58,1,-0.172193079338597,-0.208551649897568,-0.191348312820803
ARE,age,2006-05-31,228,62,1,-0.0394808395295082,-0.0292303563844853,-0.0328852185130773
ARE,age,2006-06-30,228,59,1,-0.0482795156913793,-0.0695617148570919,-0.0621085549182469
ARE,age,2006-07-31,228,58,1,0.0176952353929825,0.0325142731802064,0.0228269559911925
ARE,age,2006-08-31,228,61,1,0.015887547515,0.0577189425654402,0.037522349065535
ARE,age,2006-09-30,228,61,1,0.156368506431667,0.135737765458901,0.139498594475562
ARE,age,2006-10-31,228,60,1,0.0227192690745763,0.0210401600497816,0.0147622522547573
```

- CRSP for the United States (beginning in 1926) and from Compustat for all other countries

DATA CONSTRUCTION

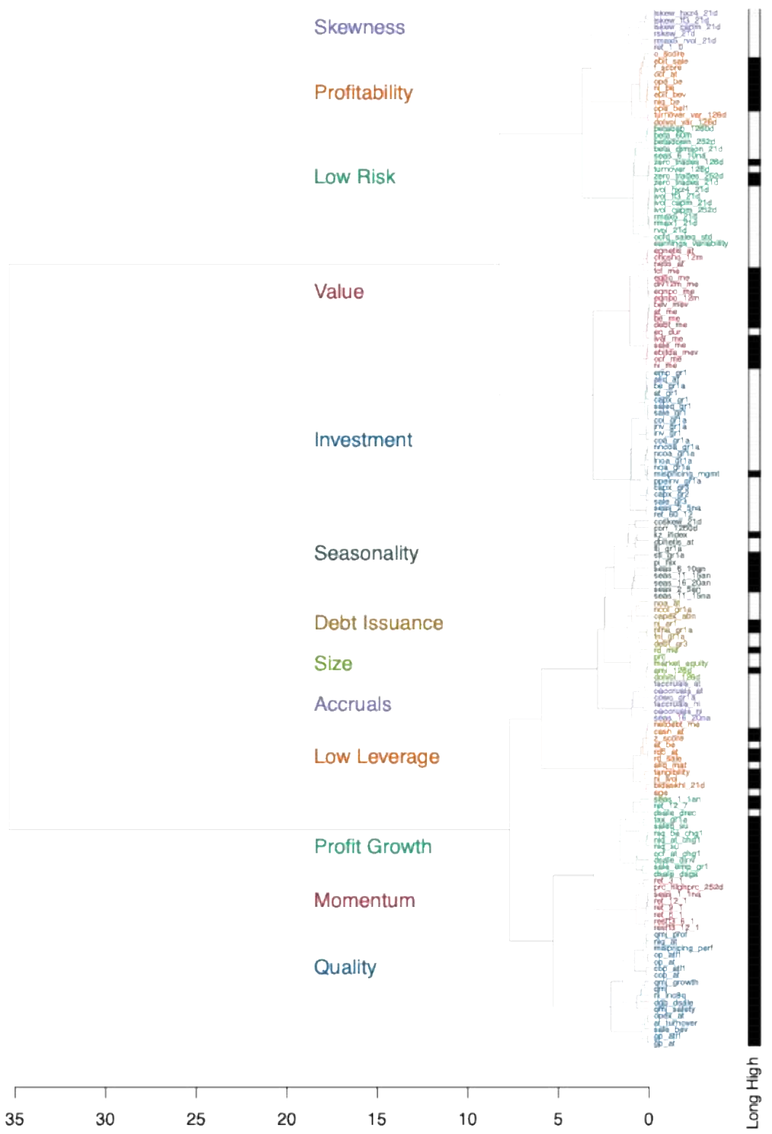
- focus on a one-month holding period for all factors, and only include the version that updates with the most recent accounting data
- in each country and month, sort stocks into characteristic terciles (top/middle/bottom third) with breakpoints based on non-micro stocks in that country
- for each tercile, compute its "capped value weight" return, meaning that stocks are weighted by their market equity winsorized at the NYSE 80th percentile
- factor is then defined as the high-tercile return minus the low-tercile return, corresponding to the excess return of a long-short zero-net-investment strategy

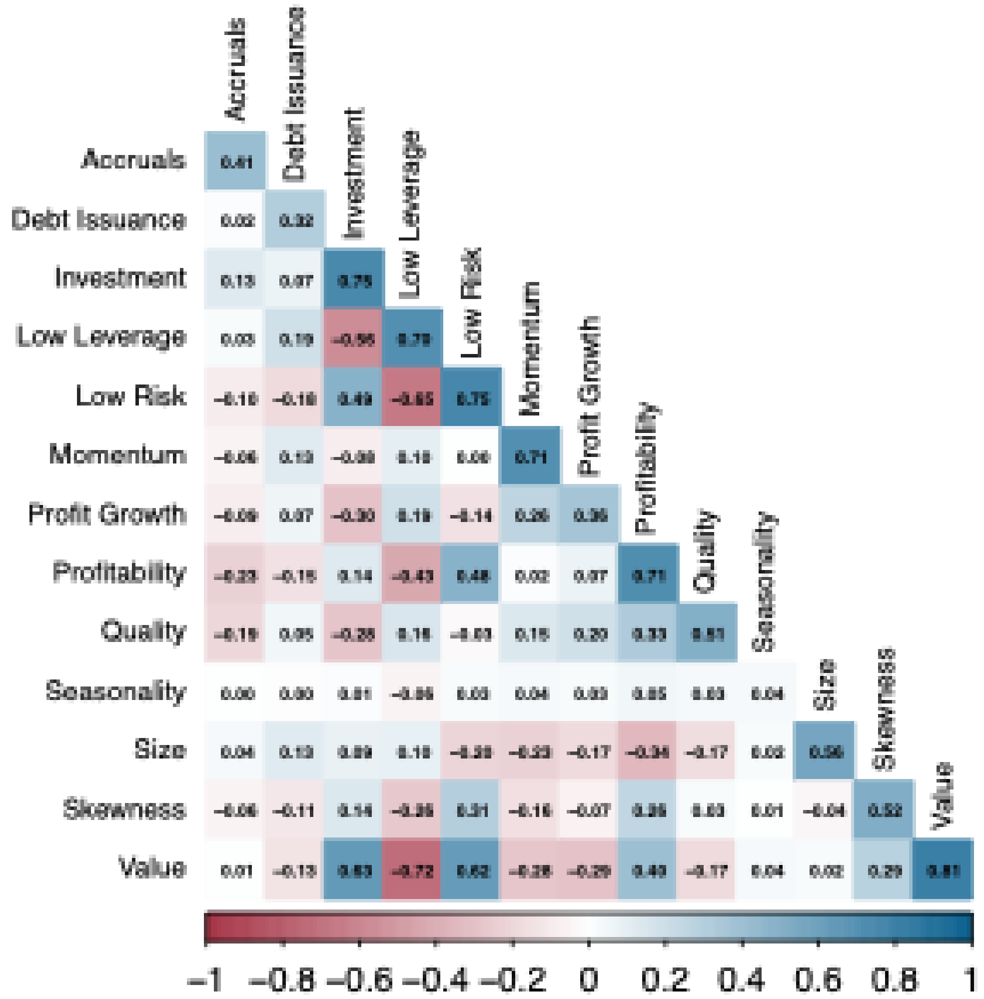
FACTOR GROUPING

- group factors into clusters using Hierarchical Agglomerative Clustering (HAC)
- define the distance between factors as one minus their pairwise correlation and use the linkage criterion of Ward (1963).

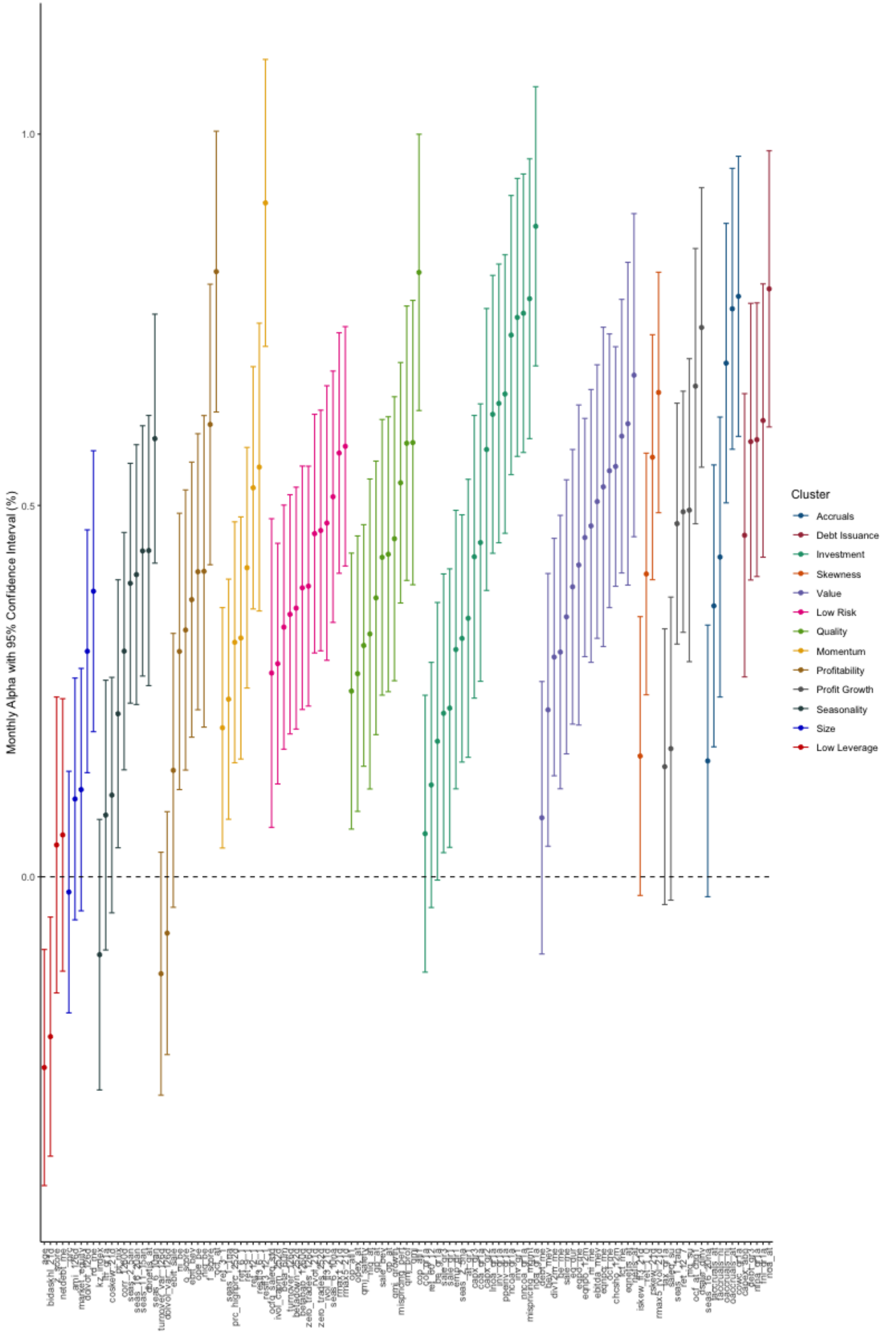
HIERARCHICAL CLUSTERING

- Distance Calculation – Cophenetic correlation between dendrogram and distance(var-cov) matrix





VARIANCE-COVARIANCE MATRIX



MONTHLY ALPHA, BY FACTOR

JOINT-FACTOR BAYESIAN APPROACH TO THE MT PROBLEM

- Multiple Testing problem to the frequentist approach
- Allows simultaneous inference of factor alphas
- Zero-alpha prior shrinks alpha estimates of all factors, thereby leading to fewer discoveries (i.e., a lower replication rate)
- Allows knowledge about the alpha of any individual factor, borrowing estimation strength across all factors (i.e., a higher replication rate)

2 KEY MODEL FEATURES

- Feature 1. Model prior: anchors the researcher's beliefs to a sensible default (e.g., all alphas are zero)

$$f_t = \alpha + \beta r_t^m + \varepsilon_t\,,\ \ \alpha \sim N(0, \tau^2)$$

  - Derive the posterior alpha distribution via Bayes' rule, posterior alpha is normal with mean

$$E(\alpha|\hat{\alpha}) = \ \kappa\hat{\alpha}\,, \qquad\qquad \kappa = \frac{\tau^2}{\tau^2 + \sigma^2/T} = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T}} \in (0,1)$$

2 KEY MODEL FEATURES

- Hierarchical (alpha) structure: each alpha is shrunk toward its posterior cluster mean (i.e., toward related factors)

$$E(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) = \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{\tau_c^2 N}}\,\hat{\alpha}^{\cdot} + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \frac{1}{1+\frac{\tau_w^2+(1-\rho)\sigma^2/T}{(\tau_c^2+\rho\sigma^2/T)N}}\,\hat{\alpha}^{\cdot}\right)$$

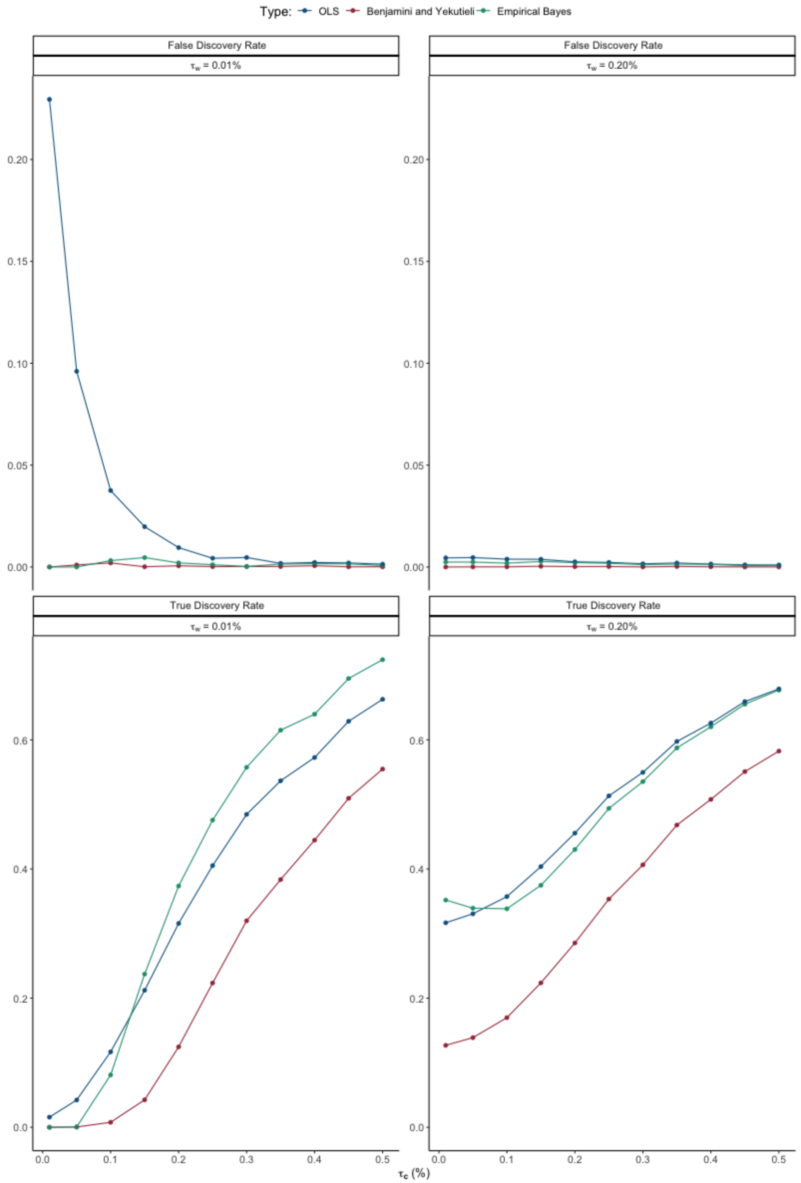
where $\hat{\alpha}^{\cdot} = \frac{1}{N}\sum_j \hat{\alpha}^j$ *is average alpha. When the number of factors* N *grows, the limit is*

$$\lim_{N\to\infty} E(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N) = \frac{1}{1+\frac{\rho\sigma^2}{\tau_c^2 T}}\,\hat{\alpha}^{\cdot} + \frac{1}{1+\frac{(1-\rho)\sigma^2}{\tau_w^2 T}}\left(\hat{\alpha}^i - \hat{\alpha}^{\cdot}\right)$$

  - Intuitively, the posterior for any individual alpha depends on all of the other observed alphas because they are all informative about the common alpha component

FDR CONTROL – EMPIRICAL BAYES

- Compared to Benjamini & Yukutieli and ordinary OLS(Harvey et al. (2016).)



REPLICATION RATE IDENTICAL TO THE OLS-BASED RATE

- Jensen, T. I., Kelly, B. T. & Pedersen, L. H. (n.d.). (2023). Is there a replication crisis in finance?, *The Journal of Finance (forthcoming)* **78** (5): 2465–2518.

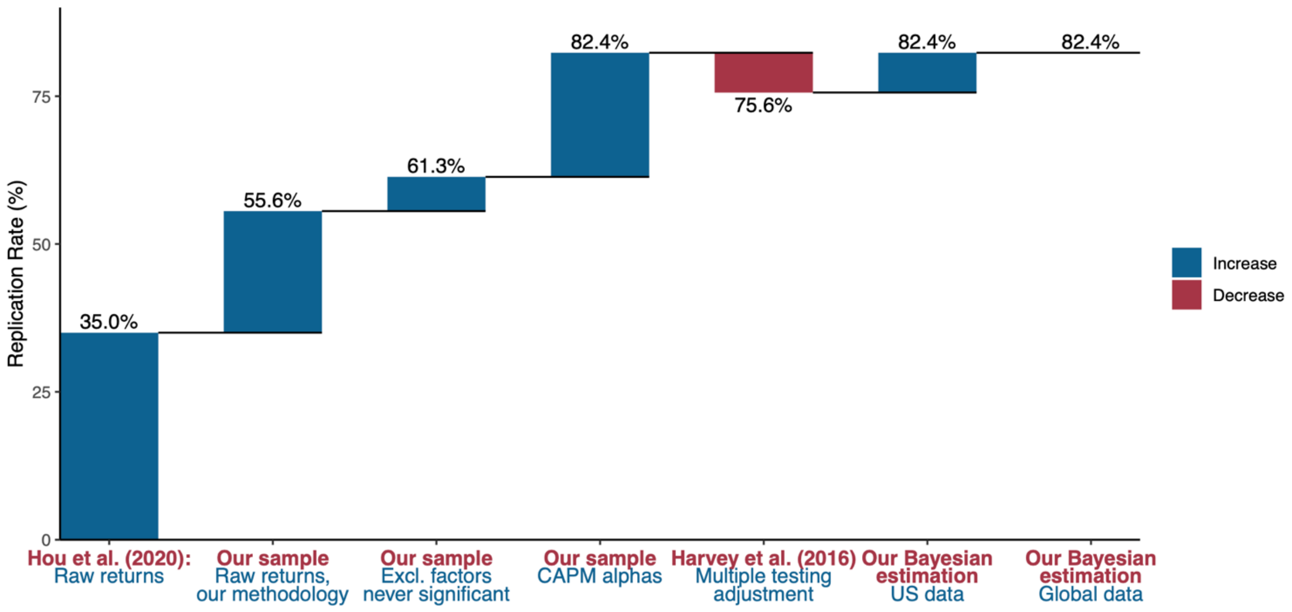


Figure 1: Replication Rates Versus the Literature

CONCLUSION & OUTLOOK

- In summary, the joint model with hierarchical alphas has the dual benefits of identifying the common component in alphas and tightening confidence intervals by sharing information among factors.
- The Empirical Bayes model help establish stable and replicable discovery rate, regionally and globally.
- Implementing CNN for more precise distance calculation for clustering correlation matrix
- Testing the out-of-sample replication rate globally – country specific idiosyncratic factor component
- Additional/end use for model:
  - look for evidence of alpha-hacking
  - compute the expected number of false discoveries based on the posterior
  - analyze portfolio choice taking into account both estimation uncertainty and return volatility
  - evaluate asset pricing models