Is There a Replication Crisis in Finance?

Theis Ingerslev Jensen, Bryan Kelly, and Lasse Heje Pedersen

Presented by: Long Zhen

Introduction – Backgrounds

- Two challenges to the financial replicability:
  - No internal validity
    - Cannot be replicated with the same data or not robust
  - No external validity
    - "p-hacking": significant without controlling the false discovery rate. → cannot be replicated in other samples or periods.

Summary

- Develop a Bayesian model
- Examine both challenges theoretically and empirically using global datasets
- Finding:
  - The majority of factors do replicate, survive joint modeling, hold up out-of-sample, strengthened by factors.
  - The number of factors can be understood as multiple version of a smaller number of themes.
  - Factors must be understood in light of economic theory and the Bayesian model offers a different interpretation.

Main results

- Replication rate: the pct of factors with a statistically significant average excess return.

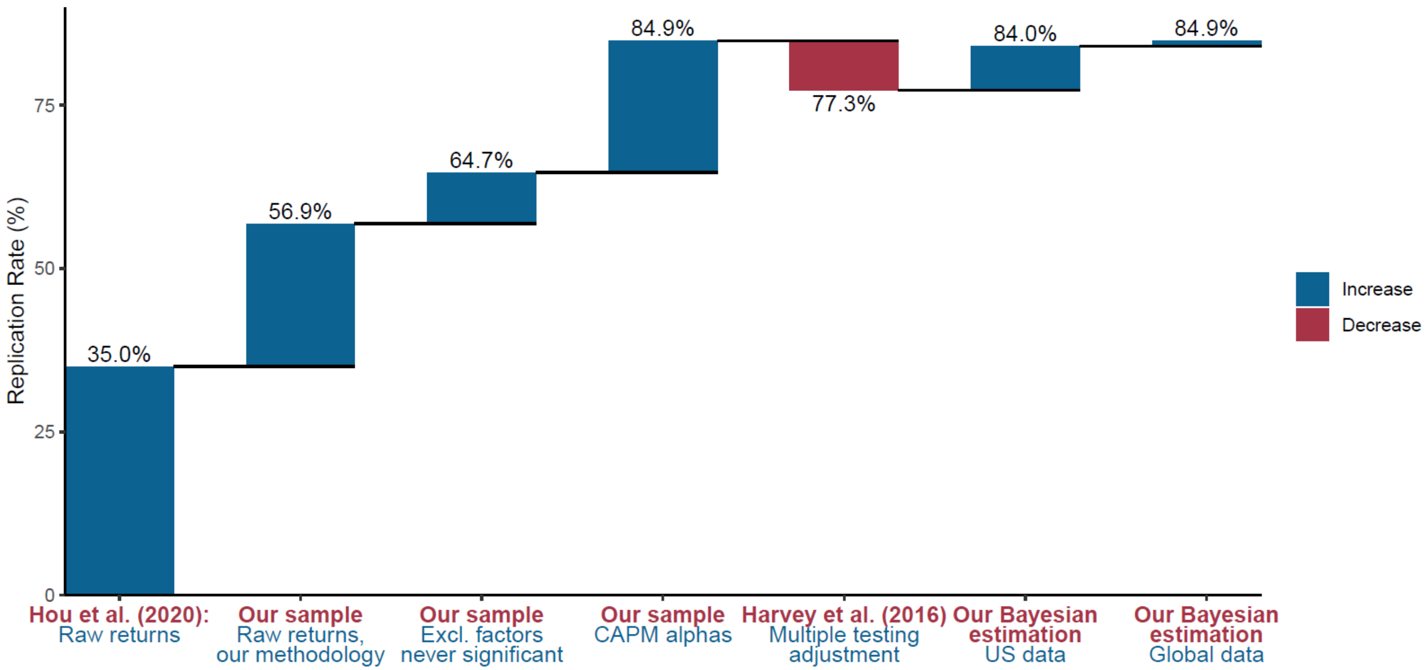


Figure 1: Replication Rates Versus the Literature

Note: This figure summarizes analyses throughout the paper. Refer to Section 3 for estimation details.

- 1: Hou et al.(2020) – 35% in 451 factors
- 2: Different sample and factor construction 56.9%
  - Longer sample +4.3%
  - Only 1-month holding period, instead of 1/6/12 +4.0%
  - Capped value-weighting +8.5%
    - Winsorize market caps at the NYSE 80% percentile.
  - Add 15 factors ignored by Hou et al. +2.4%
  - Other construction detail +2.7%

- 3: Exclude originally insignificant factors 64.7%
- 4: Alpha to the CAPM, not raw return 84.9%
  - When the raw return is insignificant, but the alpha is significant, then the factor's efficacy is masked by its risk exposure. Eg. Low-beta anomaly
- 5: Multiple testing and Bayesian modeling 77.3%
  - MT correction using a leading method(Benjamini and Yekutieli, 2001)
  - If the data are dependent, independent testing with an MT correction fails to make efficient use of the data.

- 6: Bayesian estimation 84.0%
  - Major benefits:
    - Impose a prior that all alphas are expected to be zero
      - Impose conservatism
    - Use a joint model of factors → infer all factors simultaneously
      - Draw more informative statistical inferences
- 7: Global replication 84.9%

→ 1-6: internal validity

→ 7: external validity

McLean and Pontiff(2016): US factors are 26% lower out-of-sample and 58% lower post-publication →

- Post-publication performance
  - A positive but attenuated post-publication alpha is the expected outcome based on Bayesian learning, rather than a sign of non-reproducibility.
  - Higher pre-publication alphas should be associated with higher post-publication alphas on average.

- The multidimensional challenge
  - Factors cluster into 13 highly correlated themes.
  - This property features prominently in Bayesian model.
- Why “factor zoo”?
  - The risk-return tradeoff is complex and difficult to measure.

A Bayesian model of factor replication

- The Bayesian alpha:

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

  - The OLS $\hat{\alpha} = \frac{1}{T}\sum_t (f_t - \beta r_t^m) = \alpha + \frac{1}{T}\sum_t \epsilon_t, \hat{\alpha} \sim N(\alpha, \sigma^2/T)$
  - The posterior alpha $E(\alpha|\hat{\alpha}) = k\hat{\alpha}$

$$k = \frac{\tau^2}{\tau^2 + \sigma^2/T} \in (0,1)$$

  - → the Bayesian alpha is smaller than the OLS estimates
  - → the out-of-sample and post-publication “failure” confirm the Bayesian’s beliefs.

  - Then how to judge the replication failure?

- Alpha-hacking
  - Selectively reporting or manipulating data to artificially make the alpha seem larger
  - In the in-sample periods: $f_t = \alpha + \beta r_t^m + \widetilde{\epsilon_t} + u$
    - $u \sim N(\bar{\epsilon}, \sigma_u^2)$: return inflation due to alpha-hacking
    - $\epsilon_t = \widetilde{\epsilon_t} + u,\ \epsilon_t \sim N(\bar{\epsilon}, \bar{\sigma}^2)$
  - With alpha-hacking:

$$E(\alpha|\hat{\alpha}) = -k_0 + k^{hacking}\hat{\alpha}$$

$$k^{hacking} = \frac{1}{1 + \frac{\bar{\sigma}^2}{\tau^2 T}} \leq k$$

$$k_0 = k^{hacking}\bar{\epsilon} \geq 0$$

  - $k^{hacking} \rightarrow 0$: pure alpha-hacking

  - **→This setting accounts for alpha-hacking**

- Hierarchical Bayesian model
  - Factors are often correlated and conceptually related
  - Shared alpha:
    - Assume the global factor true alpha is the same as domestic

$$E(\alpha|\hat{\alpha},\hat{\alpha}^g) = k^g(\frac{1}{2}\hat{\alpha} + \frac{1}{2}\hat{\alpha}^g)$$

    - **→ the domestic and global alphas are shrunk both toward each other and toward zero.**

    - → factors may only partially shrunk towards each other

- Hierarchical alphas: partial pooling
- Factor i has a true alpha given by: $\alpha^i = c + w^i$
  - c is the common component, w^i is idiosyncratic

$$f_t^i = \alpha^i + \beta^i r_t^m + \epsilon_t^i$$

$$Cor\left(\epsilon_t^i, \epsilon_t^j\right) = \rho \geq 0$$

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

  - → having data on many factors is helpful for estimating any of them
  - →factors naturally belong to different themes.

- Multi-level hierarchical model
  - Partially shared alphas along factors in the same cluster
  - We have N signals across K regions, total of NK factors
    - Eg. US, developed, and emerging market version of B/M

$$E(\alpha|\hat{\alpha}) = \left(\Omega^{-1} + T\Sigma^{-1}\right)^{-1} \left(\Omega^{-1} 1_{NK} \alpha_0 + T\Sigma^{-1}\hat{\alpha}\right)$$

  - → each alpha is shrunk toward its cluster mean

Bayesian multiple testing and empirical bayes estimation

- Bayesian multiple testing (the “built-in” correction)
  - Model prior imposes conservatism
  - Hierarchical structure captures joint behavior of factors

- Typical MT corrections only inflates p-values, but this hierarchical model efficiently learn about all alphas simultaneously.

Empirical Bayes

- How to determine the prior? From the data itself
  - Compute each factor's $\hat{\alpha}$ in-sample
  - Set the over all alpha prior mean $\alpha^0 = 0$

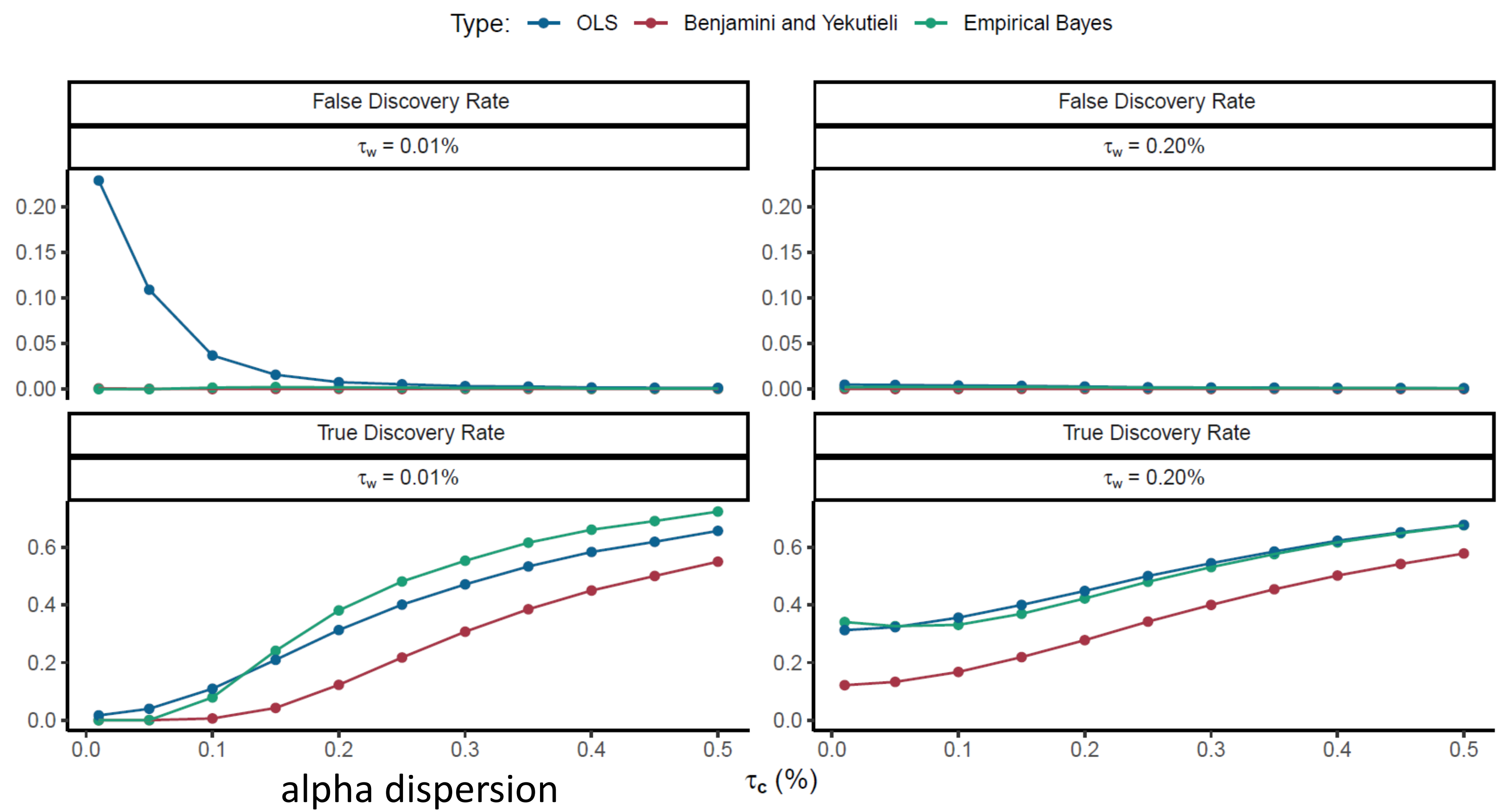


Figure 2: Simulation Comparison of False Discovery Rates

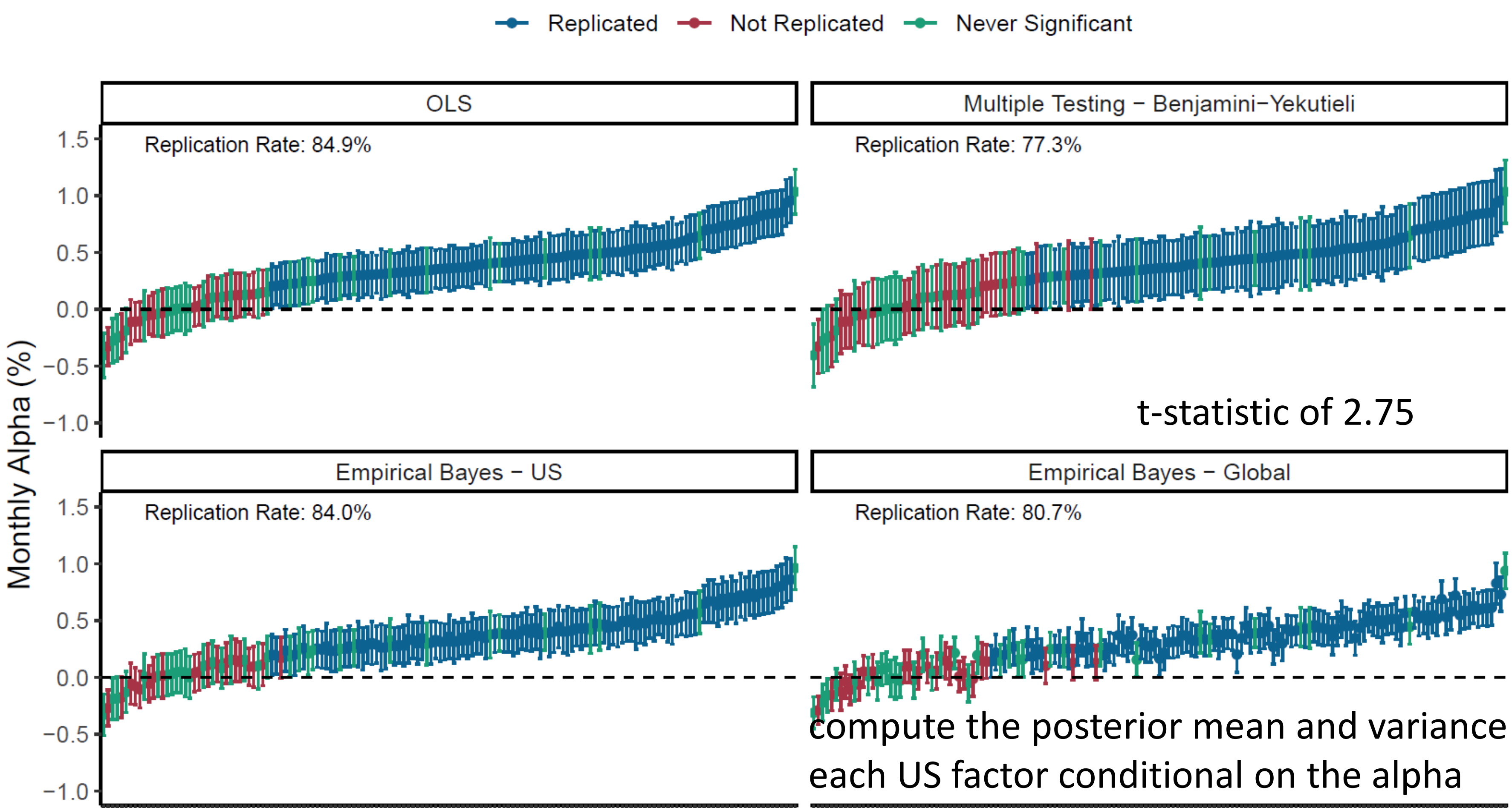
- Empirical Bayes accomplish a flexible MT correction

Global factors

- Factors
  - Based on Hou et al.(2020)
  - Finally get 153 global factors in 93 countries
- Clusters
  - Factor distance: 1 minus pairwise correlation
  - Cluster by dendrogram
  - 13 clusters: Accruals, Debt Issuance, Investment, Leverage, Low risk, Momentum, Profit Growth, Profitability, Quality, Seasonality, Size, Skewness, and Value
- Data
  - CRSP for the US (1926-), Compustat for others (1986-)

Empirical assessment of factor replicability

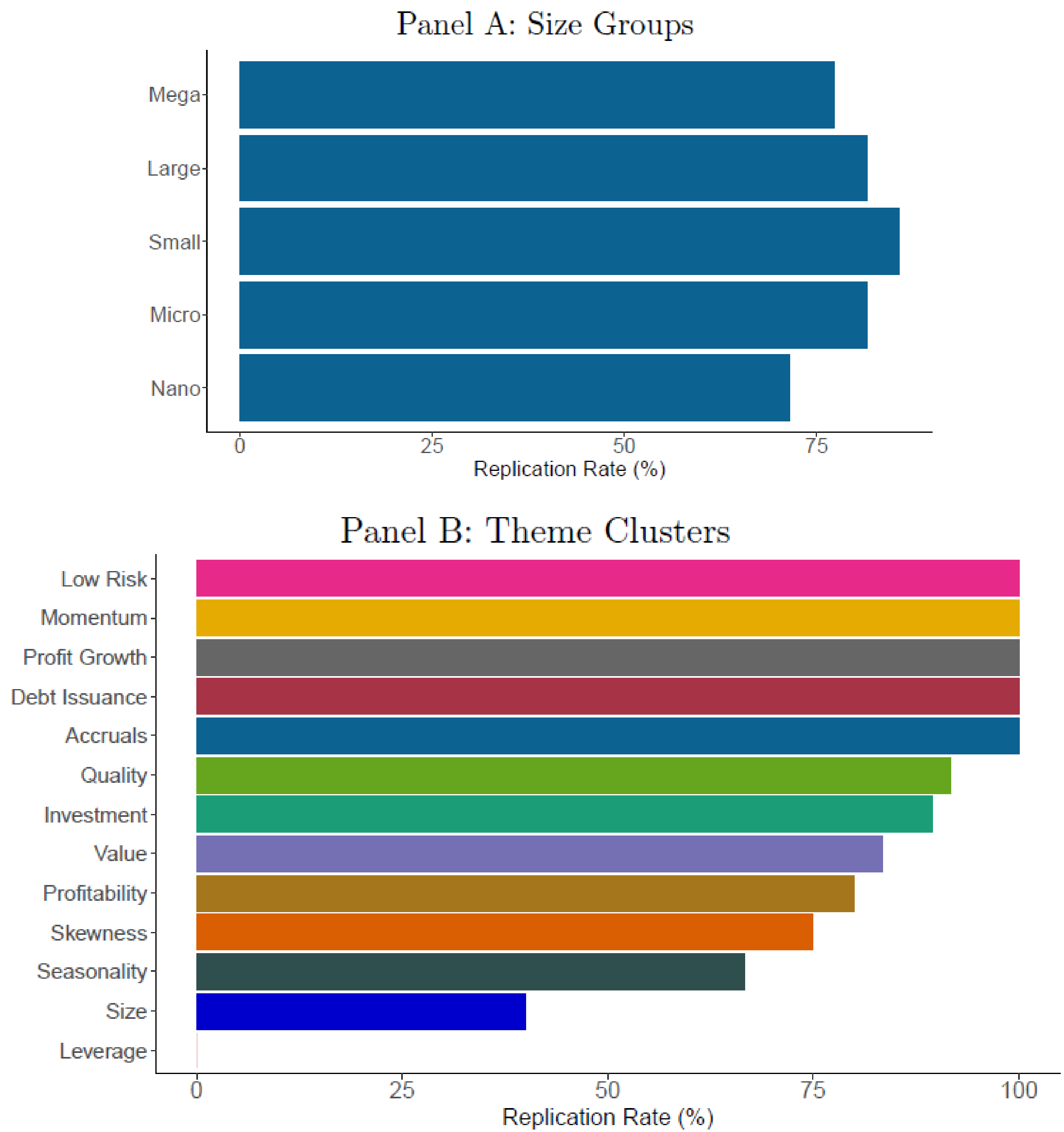
- Internal validity



t-statistic of 2.75

compute the posterior mean and variance for each US factor conditional on the alpha estimates for all factors in all regions.

Figure 3: Alpha Distributions for US Factors



Panel A: Size Groups
Mega
Large
Small
Micro
Nano
0
25
50
75
Replication Rate (%)
Panel B: Theme Clusters
Low Risk
Momentum
Profit Growth
Debt Issuance
Accruals
Quality
Investment
Value
Profitability
Skewness
Seasonality
Size
Leverage
0
25
50
75
100
Replication Rate (%)

- External validity
  - Global replication

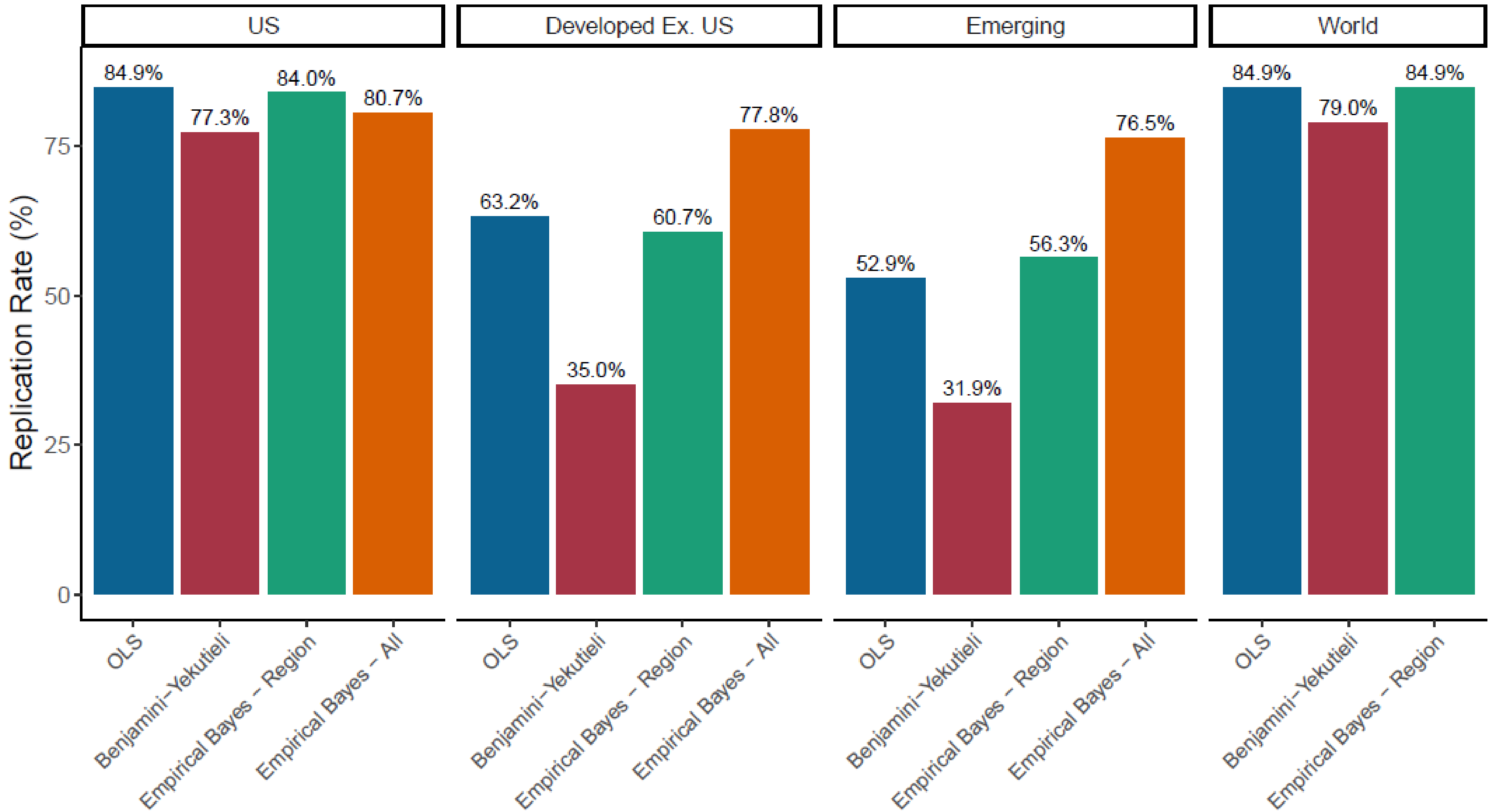


Figure 5: Replication Rates in Global Data

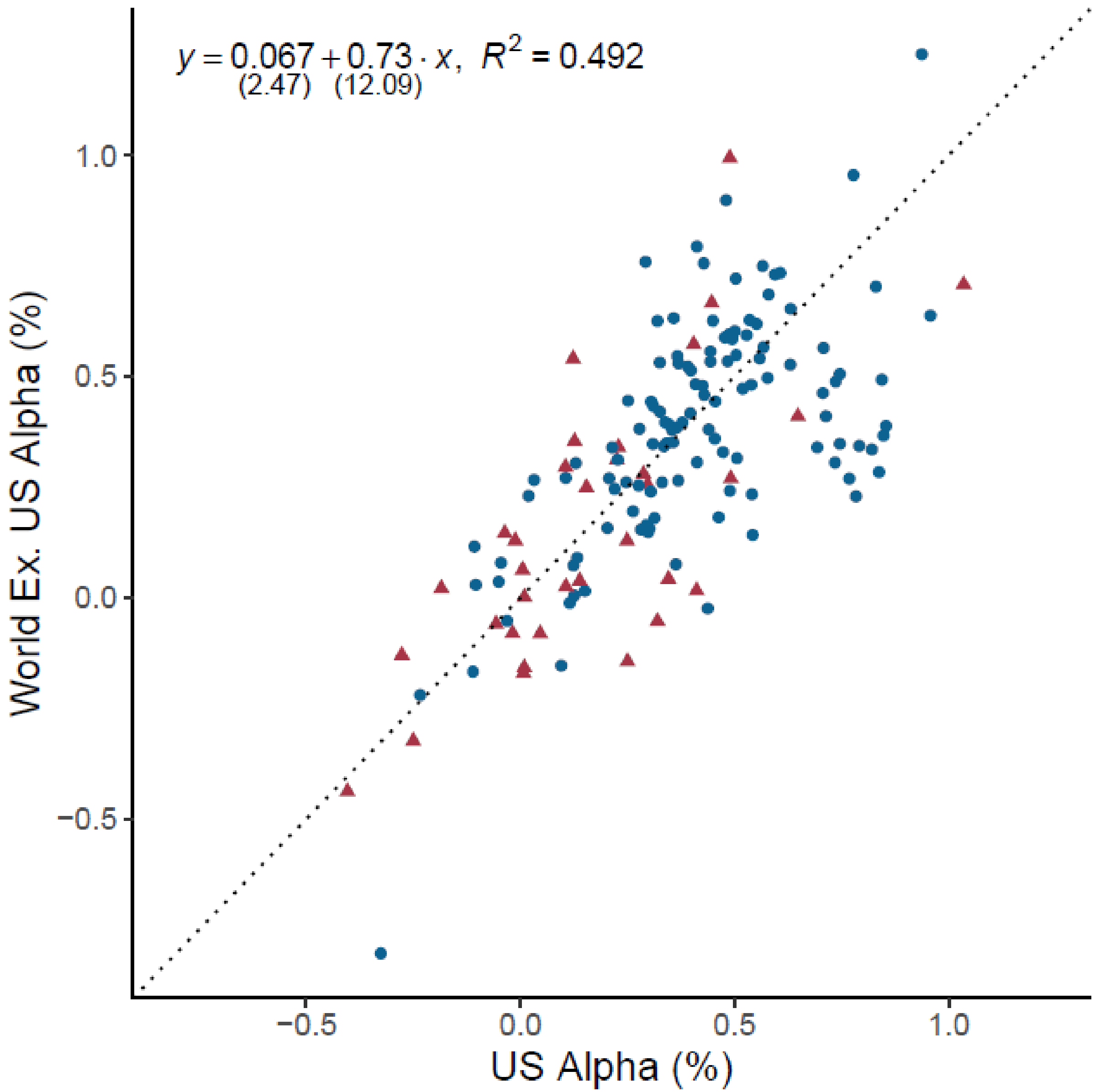


Figure 6: US Factor Alphas Versus World Ex. US

- Time series out-of-sample evidence

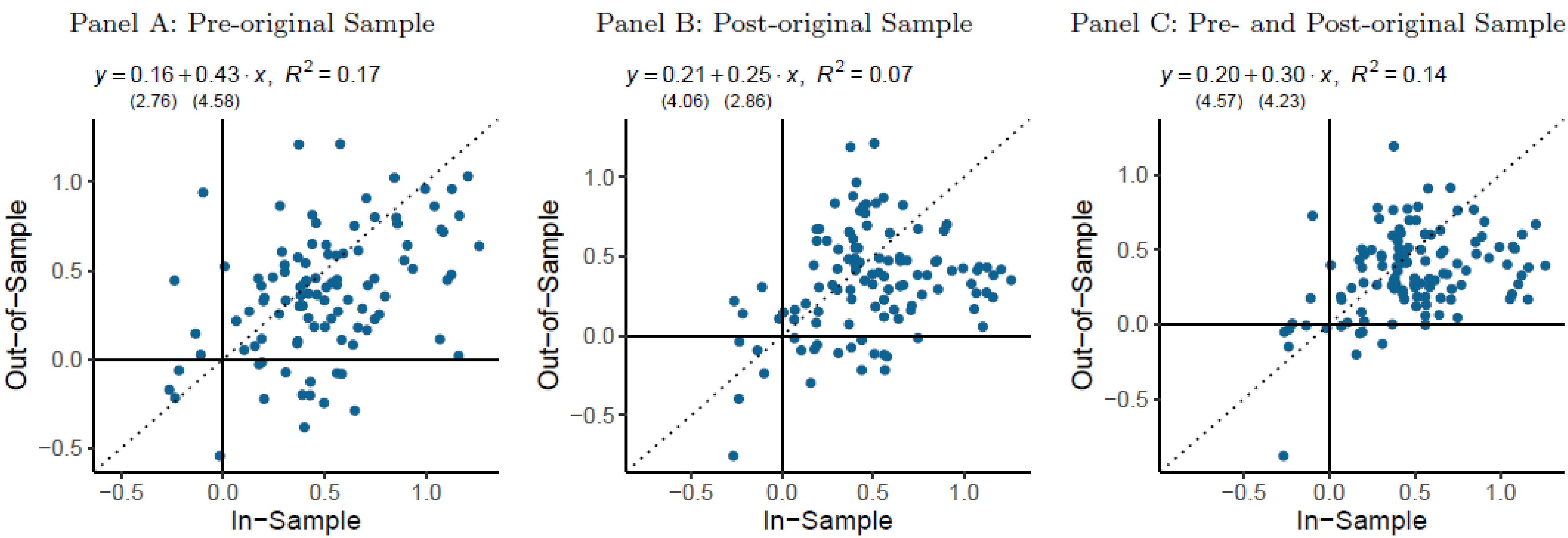


Figure 7: In-Sample versus Out-of-Sample Alphas for US Factors

- in-sample alphas contain something “real” rather than being the outcome of pure data mining, as factors that performed better in-sample also tend to perform better out-of-sample.

- Economic significance of factors
  - By region and by size

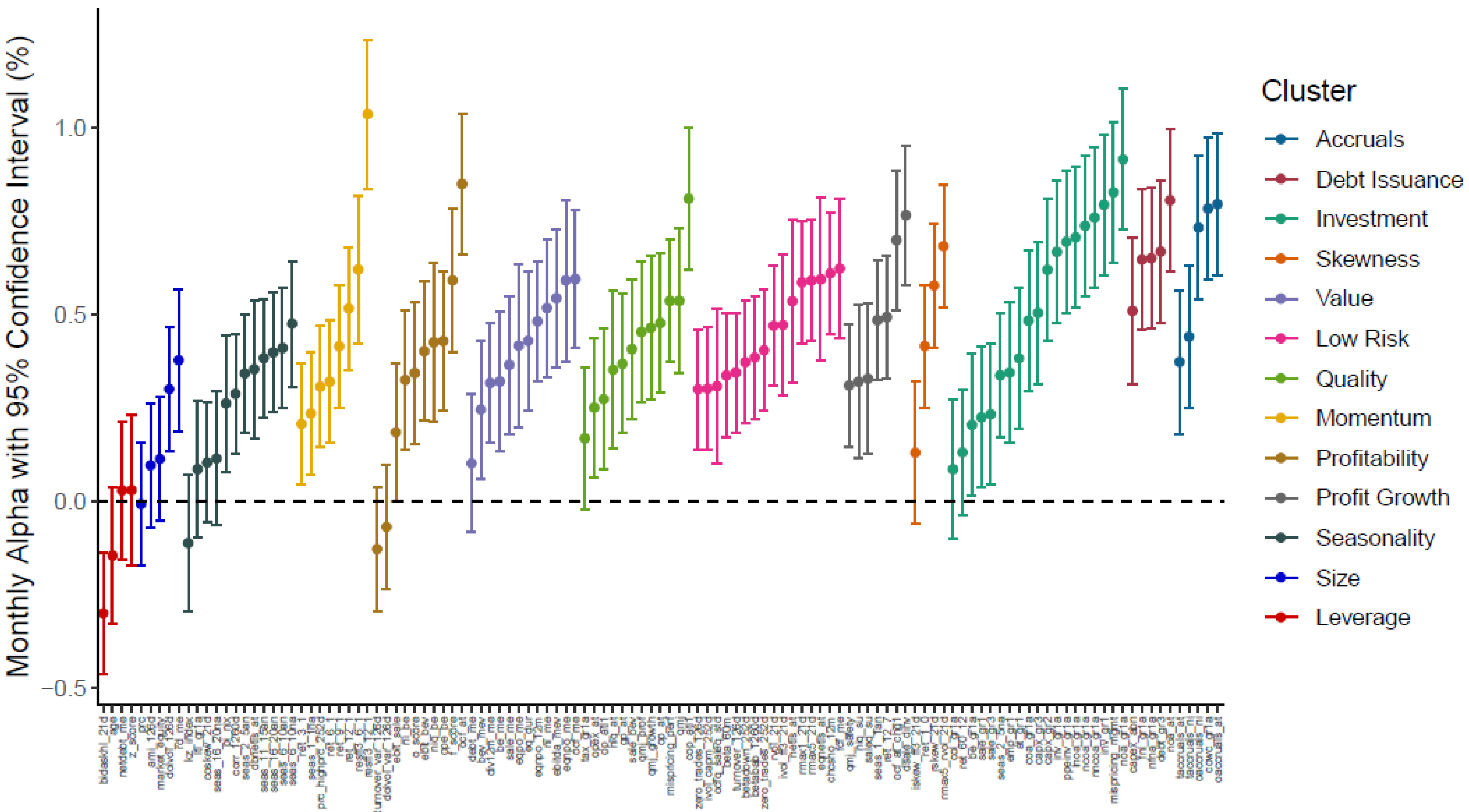
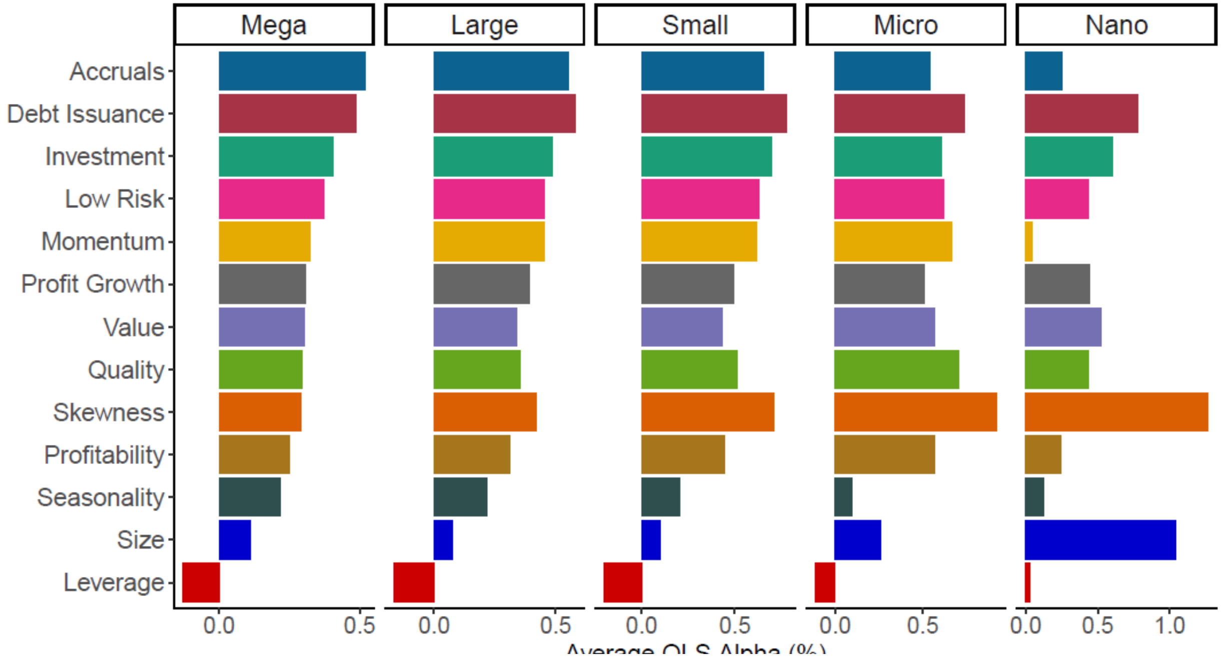
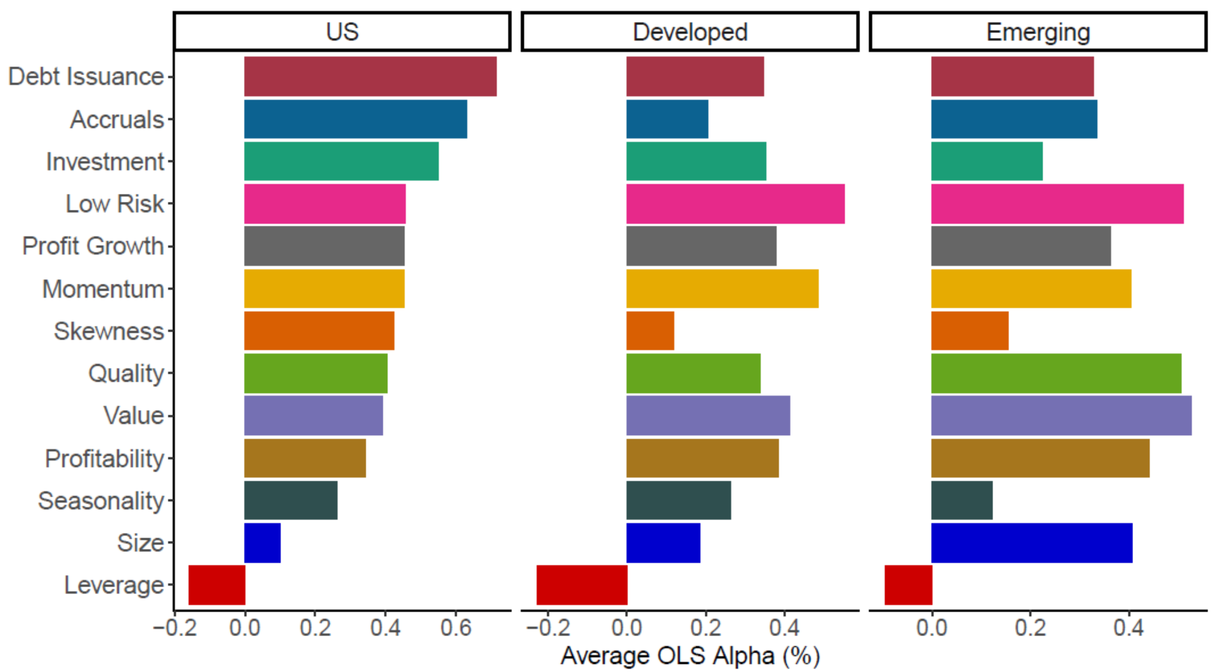


Figure 10: World Alpha Posterior By Factor and Cluster

Panel A: Size Groups



Panel B: Regions



- Controlling for other themes
    - estimate cluster weights in a tangency portfolio that invests jointly in all cluster-level portfolios.

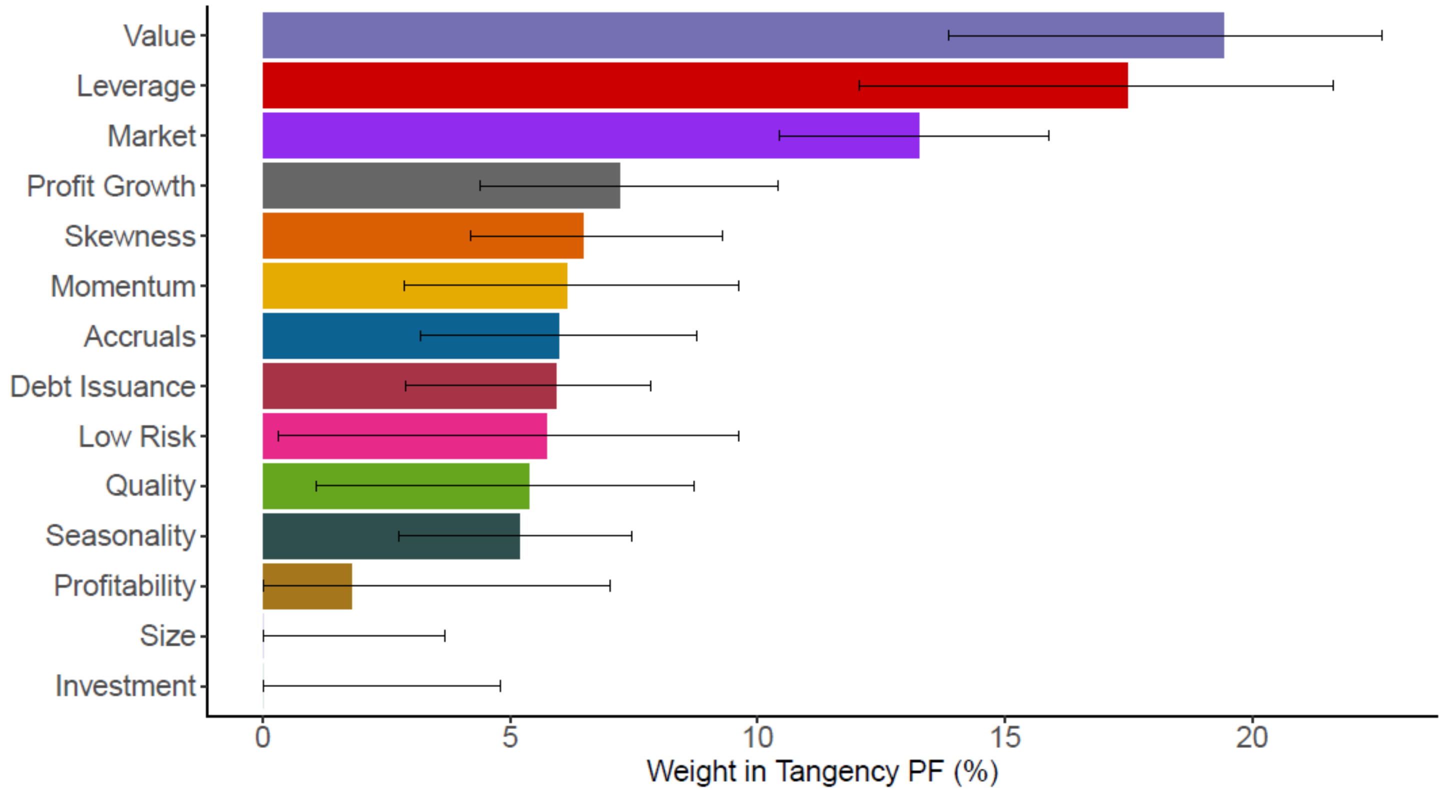


Figure 12: Tangency Portfolio Weights

- When a factor has a significant weight in the tangency portfolio, it means that it matters for an investor, even controlling for all the other factors.

Conclusion

- This paper introduce a hierarchical Bayesian model of alphas that emphasizes the joint behavior of factors.
- They find that equity factors have a high degree of internal and external validity.
- The Bayesian framework has additional applications, like correctly interpret out-of-sample evidence, alpha-hacking.