Big Data Asset Pricing

Lecture 4: The Factor Zoo and Replication

Lasse Heje Pedersen
AQR, Copenhagen Business School, CEPR

https://www.lhpedersen.com/big-data-asset-pricing

Overview of the Course: Big Data Asset Pricing

Lectures

- Quickly getting to the research frontier
  1. A primer on asset pricing
  2. A primer on empirical asset pricing
  3. Working with big asset pricing data (videos)
- Twenty-first-century topics
  4. **The factor zoo and replication**
  5. Machine learning in asset pricing
  6. Asset pricing with frictions

Exercises

1. Beta-dollar neutral portfolios
2. Construct value factors
3. Factor replication analysis
4. High-dimensional return prediction
5. Research proposal

Overview of the Lecture

- Factor zoo
  - Many factors: what are the concerns?
  - Many factors: potential benefits, an evolutionary perspective
- Replication
- Frequentist multiple testing adjustments
- Bayesian model
  - to interpret factor evidence
  - built-in multiple testing adjustments that preserves power
- Evidence on replication in finance
  - Is there a replication crisis in equity factor research?
  - Is there a replication crisis in corporate bond factor research?
  - Other areas of finance?

Factor Zoo

A Zoo of Many Factor Zoo

- Hundreds of factors claim to predict stock returns, see:
  - McLean and Pontiff (2016)
  - Harvey et al. (2016)
  - Jacobs and Müller (2020)
  - Hou et al. (2020)
  - Chen and Zimmermann (2022)
  - Jensen et al. (2023)

Taming the Factor Zoo: What are the Concerns?

- Sign of data mining or “p-hacking”
- Publication bias
- Factors should be economically motivated
- Many researchers have tried to find a simple factor model:
  - Based on only a few characteristics (e.g., 1, 3, 5, 6) that simultaneously prices many portfolios
  - E.g., Fama and French (1993), Fama and French (2015), ...
  - Several papers seek to “tame” factor zoo
- There is a single true pricing kernel, so shouldn't we be able to find a simple factor model?

Is there a Factor Zoo? Many "Real" Factors or Few?

- Chen and Zimmermann (2022), Jensen et al. (2023): Most factors are replicable
- Kozak et al. (2020): characteristics-sparse SDFs formed from a few such factors-e.g., the four- or five-factor models in the recent literature-cannot adequately summarize the cross-section of expected stock returns
- Even the critique by Harvey et al. (2016) estimates hundreds of *true* factors

Table 5
Estimation results: A model with correlations

True factors $=(1-p_0)M$	ρ	p_0	$\lambda(\%)$	M	t-statistic FWER(5%)	FWER(1%)	FDR(5%)	FDR(1%)
	Panel A: r = 1/2 (baseline)							
783	0	0.396	0.550	1,297	3.89	4.28	2.16	2.88
766	0.2	0.444	0.555	1,378	3.91	4.30	2.27	2.95
761	0.4	0.485	0.554	1,477	3.81	4.23	2.34	3.05
708	0.6	0.601	0.555	1,775	3.67	4.15	2.43	3.09
498	0.8	0.840	0.560	3,110	3.35	3.89	2.59	3.25
	Panel B: r = 2/3 (more unobserved tests)							
779	0	0.683	0.550	2,458	4.17	4.55	2.69	3.30
749	0.2	0.722	0.551	2,696	4.15	4.54	2.76	3.38
688	0.4	0.773	0.552	3,031	4.06	4.45	2.80	3.40
499	0.6	0.885	0.562	4,339	3.86	4.29	2.91	3.55
421	0.8	0.922	0.532	5,392	3.44	4.00	2.75	3.39

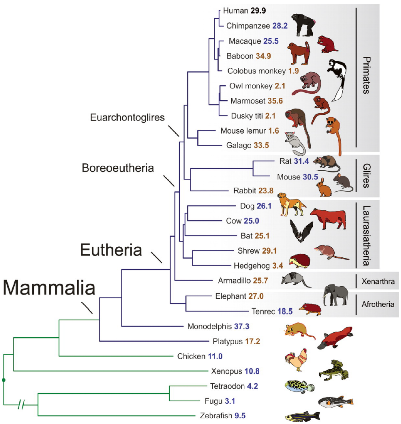
We estimate the model with correlations. r is the assumed proportion of missing factors with a t-statistic between 1.96 and 2.57. Panel A shows the results for the baseline case in which r=1/2, and panel B shows the results for the case in which r=2/3. ρ is the correlation coefficient between two strategy returns in the same period. p_0 is the probability of having a strategy that has a mean of zero. λ is the mean parameter of the exponential distribution for the monthly means of the true factors. M is the total number of trials.

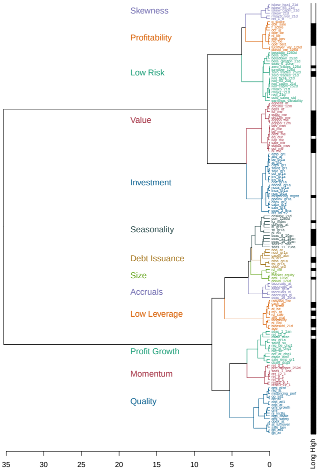
A Zoo of Factors: An Evolutionary Perspective

- The economic case for a simple factor model is weak
  - If the CAPM works, then fine, but, once we go beyond, no
  - E.g., if expected future profitability matters, there are many ways to predict it
    - An average of these might be better than any one of them
  - If covariance with economic conditions (growth, investment opportunities, liquidity, etc) matters, then the most correlated portfolio might not be simple
  - If we observe many characteristics, s_t^t, and have $E(r_{t+1}^i | s_t^i) = f(s_t^i)$, then f might not be simple
- Evolutionary perspective for having many factors
  - Several types of information matter for predicting stock returns
  - Each can be measured in various ways
  - Researchers have gradually learned about these, building on earlier research

Structuring the Factor Zoo: Evolutionary Perspective

- Jensen et al. (2023) provide “phylogenetic factor tree” via clustering
  - “phylogenetic tree”=a diagram containing a hypothesis of relationships of organisms that reflects their evolutionary history





Replication

Replication in Science

- Replication is important to establish true knowledge
  - Researchers should not mislead readers, make up data, etc.
  - Researchers should behave according to the code of professional conduct and ethics, e.g.

    https://afajof.org/wp-content/uploads/files/afa_code_of_professional_con.pdf

    https://www.aeaweb.org/about-aea/code-of-conduct
  - Many journals now have rules regarding code sharing, e.g. *JF*
- Forms of replication, Hamermesh (2007) (see also Welch (2019))
  - **pure replication**: "checking on others' published papers using their data"
  - **statistical replication**: "different sample, but identical model and population."
  - **scientific replication**: "different sample, different population and perhaps similar, but not identical model."
- External validity
  - **Out-of-sample evidence**: does the finding hold up in other countries, asset classes, or time periods
    - If not: original finding is (a) spurious or (b) specific to the time period or sample (e.g., effect arbitraged away once known)?

Replication Crises

Several fields face replication crises (or credibility crises):

- Medicine (Ioannidis 2005), psychology (Nosek et al. 2012), ...

Now finance? Two main challenges to equity factor research:

1. **No internal validity**

   Results cannot be reproduced with slightly different methodology or data

   "Most anomalies fail to hold up to currently acceptable standards"

   – Hou et al. (2020)

2. **No external validity**

   Results replicate in-sample, but are spurious due to "p-hacking"

   "most claimed research findings in financial economics are likely false"

   – Harvey et al. (2016)

- **Most factors are in fact replicable:**
  - Chen and Zimmermann (2022) consider "pure replication," reproducing 98% of factors
  - Jensen et al. (2023) "scientific replication" of 82%, with robustness and external validity (out-of-sample evidence) – **later in lecture**

Frequentist Multiple-Testing Adjustments

The Issue with Multiple Tests

- **Standard workflow:**
  1. Create a factor and estimate its alpha, $\hat{\alpha}$
  2. Compute p-value $= \Pr(\hat{\alpha}\,|H_0\colon\ \alpha = 0)$
  3. Reject H_0 if p-value $< a$
- **Question**: Suppose we test K uncorrelated factors where H_0 in true (i.e., $\alpha = 0$) using a critical value $a = 5\%$. What's the probability that we reject the null hypothesis for at least 1 factor?
- **Answer:** $1 - (1 - a)^K$

The Issue with Multiple Tests

- **Standard workflow:**
  1. Create a factor and estimate its alpha, $\hat{\alpha}$
  2. Compute p-value $= \Pr(\hat{\alpha}\,|H_0\colon\ \alpha = 0)$
  3. Reject H_0 if p-value $< a$
- **Question**: Suppose we test K uncorrelated factors where H_0 in true (i.e., $\alpha = 0$) using a critical value $a = 5\%$. What's the probability that we reject the null hypothesis for at least 1 factor?
- **Answer:** $1 - (1 - a)^K$

Number of tests, K	$\Pr(\#Reject \geq 1 \| a = 0.05)$
1	5.0%
2	9.8%
3	14.3%
10	40.1%
100	99.4%

The Issue with Multiple Tests

- **Standard workflow:**
  1. Create a factor and estimate its alpha, $\hat{\alpha}$
  2. Compute p-value $= \Pr(\hat{\alpha}\,|H_0\colon\ \alpha = 0)$
  3. Reject H_0 if p-value $< a$
- **Question**: Suppose we test K uncorrelated factors where H_0 in true (i.e., $\alpha = 0$) using a critical value $a = 5\%$. What's the probability that we reject the null hypothesis for at least 1 factor?
- **Answer:** $1 - (1 - a)^K$

Number of tests, K	$\Pr(\#Reject \geq 1 \vert a = 0.05)$
1	5.0%
2	9.8%
3	14.3%
10	40.1%
100	99.4%

- **Solution**: Multiple testing adjustments
  - E.g., Bonferroni adjustment: reject if p-value$^i \leq a/K$

The Issue with Multiple Tests

- **Standard workflow:**
  1. Create a factor and estimate its alpha, $\hat{\alpha}$
  2. Compute p-value $= \Pr(\hat{\alpha}\,|H_0\colon\ \alpha = 0)$
  3. Reject H_0 if p-value $< a$
- **Question**: Suppose we test K uncorrelated factors where H_0 in true (i.e., $\alpha = 0$) using a critical value $a = 5\%$. What's the probability that we reject the null hypothesis for at least 1 factor?
- **Answer:** $1 - (1 - a)^K$

Number of tests, K	$\Pr(\#Reject \geq 1 \vert a = 0.05)$	$\Pr(\#Reject \geq 1 \vert a = 0.05/K)$
1	5.0%	5.0%
2	9.8%	4.9%
3	14.3%	4.9%
10	40.1%	4.9%
100	99.4%	4.9%

- **Solution**: Multiple testing adjustments
  - E.g., Bonferroni adjustment: reject if p-value$^i \leq a/K$

Multiple-Testing Adjustments: Frequentist and Bayesian

Frequentist:

- **Controlling the family-wise error rate (FWER):**
  - Simple method based on just the number of tests:
    - Bonferroni (1936) (very conservative - few "discoveries")
  - Better method using all the *p*-values:
    - Holm (1979)
- **Controlling the false discovery rate (FDR)**
  - Based on all the *p*-values
  - Most commonly used method:
    - Benjamini and Hochberg (1995)
  - More conservative method
    - Benjamini and Yekutieli (2001)

Bayesian:

- **Based on all the underlying data**
  - Hierarchical Bayesian model, e.g., Gelman et al. (2012)
  - Empirical Bayes: e.g., ch. 15, Efron and Hastie (2021)
  - Applied to finance: Jensen et al. (2023), **later in lecture**

Multiple-Testing Adjustments vs. Publication Bias

- Note that *all* the multiple-testing adjustments
  - require that we know the number of tests
- Sometimes the researcher does not know what tests have been done
  - E.g., publication bias
  - We later discuss ways to address this

Bonferroni Adjustment

- You start with K tests with $p\text{-value}^1, \ldots, p\text{-value}^K$
- **Example** (used throughout):
  - each $p\text{-value}^j$ refers to a test that a specific factor has $\alpha^j = 0$
  - "rejection of the null, H_0^j" = "discovery of a factor, j"
- **Significance level without Bonferroni adjustment,** a:
  Reject the null, H_0^i , iff

$$p\text{-value}^i < a$$

  E.g., $a = 5\%$ corresponding to $|t\text{-stat}| > 1.96$ (two-sided test)
- **Bonferroni adjustment**: reject instead if

$$p\text{-value}^i < \frac{a}{K}$$

  E.g., $\frac{a}{K} = \frac{5\%}{100} = 0.05\%$, corresponding to $|t\text{-stat}| > 3.48$

Bonferroni Adjustment: Family-wise Error Rate (FWER)

- To understand the motivation for Bonferroni, recall significance level, a, is

$$a = \Pr(\text{reject true } H_0) = \Pr(\text{false discovery})$$

- Define the family-wise error rate as

$$FWER = \Pr(\text{reject any true } H_0^i, i = 1, ..., K) = \Pr(\#\text{false discoveries} \geq 1)$$

- Let $I_0 \subset \{1, ..., K\}$ be the (unknown) set of true H_0^i and $K_0 = \# I_0$
- **Bonferroni adjustment controls FWER at level a:**

$$\begin{aligned} FWER =& \Pr(p\text{-value}^i < \frac{a}{K}, \text{ for any } i \in I_0) \\ \leq& \sum_{i \in I_0} \Pr(p\text{-value}^i < \frac{a}{K} | H_0^i) \\ \leq& K_0 \frac{a}{K} \leq a \end{aligned}$$

- What is the downside?
  - Trying to avoid a *single* false discovery, possibly out of 100s of tests
  - Inequality can be crude (only based on the number K)
  - Loss of power – fewer discoveries
  - No trade-off between errors of type I vs. II

False Discovery Rate (FDR)

- FWER too conservative in most applications
- More standard objective: control the false-discovery rate (FDR)
- Definition

$$FDR = \mathrm{E}\left(\frac{\#\text{false discoveries}}{\#\text{discoveries}}\right)$$

  where the ratio is taken to be zero when #discoveries=0.
- Typical goal: ensure that $FDR \leq 5\%$ or $FDR \leq 10\%$
  - With 100s or 1000s of tests, it may be OK to have several false discoveries (depending on the application)
  - as long as they are a small proportion of all discoveries
  - Controlling FDR leads to more discoveries than FWER
- Note
  - Researcher obviously does not know #false discoveries
  - But some clever and simple methods nevertheless control FDR

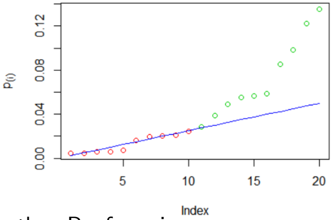
Benjamini and Hochberg (1995) Adjustment (BH)

BH at level a:

- Order the p-values: $p\text{-value}^1 \leq p\text{-value}^1 \leq ... \leq p\text{-value}^K$
- Define k as the largest i for which $p\text{-value}^i \leq \frac{i}{K} a$
- Reject all H_0^i for $i = 1, ..., k$, that is, reject if $p\text{-value}^i \leq \frac{k}{K} a$

Result (BH controls FDR)

If the p-values corresponding to valid null hypotheses are independent of each other, then the BH procedure implies $FDR_{BH} = \frac{\#\text{true null hypotheses}}{K} a \leq a$.



- BH more discoveries than Bonferroni:
  - $p\text{-value}^i \leq \frac{k}{K} a$ is easier than $p\text{-value}^i \leq \frac{1}{K} a$
- BH works even in cases with correlated tests (Benjamini and Yekutieli (2001))

Benjamini and Yekutieli (2001) Adjustment (BY)

BY at level a:

- Order the p-values: $p\text{-value}^1 \leq p\text{-value}^1 \leq \ldots \leq p\text{-value}^K$
- Define k as largest i for which $p\text{-value}^i \leq \frac{i}{Kc(K)}a$, where $c(K) = \sum_{j=1}^{K} \frac{1}{j}$
- Reject all H_0^i for $i = 1, \ldots, k$, i.e., reject if $p\text{-value}^i \leq \frac{k}{Kc(K)}a$

Result (BY controls FDR)

Under arbitrary dependence of the tests, BY implies FDR $\leq a$

- More conservative then BH, which replaces $c(K)$ by 1

$$c(1) = 1, c(10) = 2.9, c(100) = 5.2, c(1000) = 7.5$$

A Bayesian Model

Based on Jensen et al. (2023)

Bayesian Models and Multiple Testing

Why Bayesians don't worry about multiple testing (Gelman et al., 2012)

- Classical hypothesis testing is designed for testing one parameter
  - Hence, need to make adjustment when testing multiple parameters
- A Bayesian analysis should model all parameters jointly
  - Hence, multiplicity is built into the posterior

Related point: Bayesian immune to selection bias if prior is right

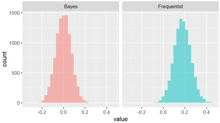
- Assume factors are iid generated from
  $\hat{\alpha}^i = \alpha^i + \epsilon^i, \quad \alpha^i \sim N(0, 0.1^2), \quad \epsilon^i \sim N(0, 0.1^2)$
- A frequentist takes α^i to be $\hat{\alpha}^i$ while a Bayesian uses $E[\alpha^i | \hat{\alpha}^i] = \frac{1}{2}\hat{\alpha}^i$
- Suppose you test 100 factors and select the best one. What is difference between the true alpha and the Bayes/frequentist estimate?

Bayesian Inference is Immune to Selection Bias (iff the prior is correct)

Simulate the following process 10,000 times:

1. Draw 100 $\alpha^i \sim N(0, 0.1^2)$ and 100 $\epsilon^i \sim N(0, 0.1^2)$ such that $\hat{\alpha}^i = \alpha^i + \epsilon^i$
2. Select factor with highest $\hat{\alpha}^i$ (called $\hat{\alpha}^{max}$)
3. Record $\text{b}^{Bayes} := \frac{1}{2}\hat{\alpha}^{max} - \alpha^{max}$ and $\text{b}^{freq} := \hat{\alpha}^{max} - \alpha^{max}$

Figure shows the histogram of b^{Bayes} (left) and b^{freq} (right)



Bayesian Model: A Single Factor

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

Data $\hat{\alpha} = \frac{1}{T}\sum_t (f_t - \beta r_t^m) = \alpha + \frac{1}{T}\sum_t \varepsilon_t$

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

Looking OOS, what is a successful replication vs. replication failure?

- A positive, but lower, alpha sometimes interpreted as sign of failure
- But it is **expected** outcome from Bayesian perspective

Bayesian Model: Alpha Hacking

- In-sample time period, $t = 1, \dots, T$, with alpha hacking:

$$f_t = \alpha + \beta r_t^m + \underbrace{\tilde{\varepsilon}_t + u}_{\varepsilon_t}$$

- $\tilde{\varepsilon}_t \sim N(0, \sigma^2)$ captures usual return shocks
- $u \sim N(\bar{\varepsilon}, \sigma_u^2)$ represents return inflation due to alpha-hacking
- $\varepsilon_t \sim N(\bar{\varepsilon}, \bar{\sigma}^2)$, where $\bar{\varepsilon} \geq 0$ is the alpha-hacking bias, and the variance $\bar{\sigma}^2 = \sigma^2 + \sigma_u^2 \geq \sigma^2$ is elevated

Proposition (Alpha-hacking)

The posterior alpha with alpha-hacking is given by

$$E(\alpha|\hat{\alpha}) = -\kappa_0 + \kappa^{hacking}\hat{\alpha}$$

where $\kappa^{hacking} = \frac{1}{1+\frac{\bar{\sigma}^2}{\tau^2 T}} \leq \kappa$ *and* $\kappa_0 = \kappa^{hacking}\bar{\varepsilon} \geq 0$. *Further,* $\kappa^{hacking} \rightarrow 0$ *in the limit of "pure alpha-hacking,"* $\tau \rightarrow 0$ *or* $\bar{\sigma} \rightarrow \infty$.

Bayesian Model: Out-of-Sample Alpha with Alpha Hacking

Proposition (Out-of-sample alpha)

The posterior alpha based on an in-sample data from time 1 to T with alpha-hacking, and an out-of-sample period from $T+1$ to $T+T^{oos}$ is given by

$$E(\alpha|\hat{\alpha},\hat{\alpha}^{oos}) = \kappa^{oos}\left(w(\hat{\alpha}-\bar{\varepsilon})+(1-w)\alpha^{oos}\right)$$

where $w=\frac{\sigma^2/T^{oos}}{\bar{\sigma}^2/T+\sigma^2/T^{oos}}\in(0,1)$ is the relative weight on the in-sample period relative to the out-of-sample period, and $\kappa^{oos}=\frac{1}{1+1/(\tau^2([\bar{\sigma}^2/T]^{-1}+[\sigma^2/T^{oos}]^{-1}))}$ is a shrinkage parameter.

- Note: even if $T=T^{oos}$, alpha-hacking means that we give more weight to OOS data

Bayesian Model: Related Factors

The power of related factors: domestic+global or BAB1+BAB2

- "Domestic" evidence: $f_t = \alpha + \beta r_t^m + \varepsilon_t$
- "Global" evidence: $f_t^g = \alpha + \beta^g r_t^g + \varepsilon_t^g$

Proposition (The Power of Shared Evidence)

The posterior alpha given domestic ($\hat{\alpha}$) and global ($\hat{\alpha}^g$) evidence:

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

Less shrinkage

$$\kappa^g = \frac{1}{1 + \frac{\sigma^2}{\tau^2 T} \frac{1+\rho}{2}} \in [\kappa, 1]$$

More conviction

$$Var(\alpha|\hat{\alpha}) \geq Var(\alpha|\hat{\alpha}, \hat{\alpha}^g)$$

Bayesian Model: Hierarchical Model

Many factors: $f_t = \alpha^i + \varepsilon_t^i, \qquad \alpha^i = \alpha^o + c^j + \omega^i$
Propositions: see Jensen et al. (2023)

- Common alpha: $\alpha^o = 0$
- Cluster alpha: $c^j \sim N(0, \tau_c^2)$
- Factor specific alpha: $\omega^i \sim N(0, \tau_\omega^2)$
- Global analysis adds another tier to hierarchy

Estimation

- Joint estimation of all factors
- Empirical Bayes
  - First estimate OLS alphas and noise variance-covariance, $\mathrm{Var}(\varepsilon)$
  - Then estimate hyper-parameters, τ_c^2 and τ_ω^2, using MLE
    - Intuition: realized dispersion in $\hat{\alpha}^i$'s can inform prior
  - Then compute the posterior distribution of alphas

Empirical Bayes

Bayesian model

- Problem: Prior+data determine posterior, but prior is subjective
- Empirical Bayes: Choose prior most consistent with the data

Estimation (given assumptions from previous slide)

- Marginal distribution of observed alphas

$$f_t \sim N(0, \Omega + \Sigma)$$
$$\hat{\alpha} \sim N(0, \Omega + \Sigma / T)$$

where $\Sigma = \mathrm{Var}(\varepsilon)$ and $\Omega = \mathrm{Var}(\alpha)$

- The variance-covariance matrix for alpha, $\Omega = \mathrm{Var}(\alpha)$, is

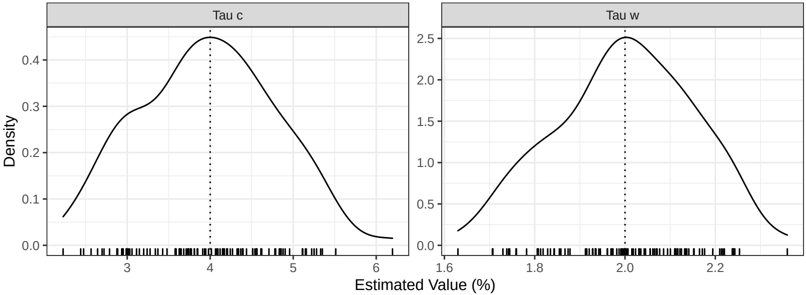
$$\Omega_{i,k} = \mathrm{Cov}(\alpha^i, \alpha^k) = \begin{cases} \tau_c^2 + \tau_\omega^2 & \text{if } i = k \\ \tau_c^2 & \text{if } i \text{ and } k \text{ from same cluster} \\ 0 & \text{otherwise} \end{cases}$$

- Find τ_c and τ_ω to maximize the likelihood of observed $\hat{\alpha}$

Empirical Bayes - Example

Setup

- Data: 150 factors from 15 clusters, observed over 70 years
- Σ: Annual volatility of 10%, correlation 50% if same cluster, and 0% otherwise
- True hyper-parameters: $\tau_c = 4\%$, $\tau_\omega = 2\%$
- For simulation $s = 1, \ldots 100$
  1. Simulate true alphas by drawing 15 cluster alphas, c^j, from $N(0, \tau_c^2)$ and 150 idisyncratic alphas, ω^i, from $N(0, \tau_\omega^2)$
  2. Simulate observed alphas by adding noise from $N(0, \Sigma/T)$ to true alphas
  3. Estimate hyper-parameters via maximum likelihood
  4. Save $\hat{\tau}_c^s$ and $\hat{\tau}_\omega^s$
- Result of simulation



Bayesian Multiple Testing

- A factor "discovered" by the Bayesian
  - if its z-score is greater than $\bar{z} = 1.96$:

$$\frac{E(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau)}{\sqrt{\mathrm{Var}(\alpha^i|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau)}} > 1.96$$

  - Equivalently, factor i is discovered if $p\text{-null}_i < 2.5\%$, where

$$p\text{-null}_i = Pr(\alpha^i < 0|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau)$$

    - Similar to frequentist p-value
    - Posterior probability of a "false discovery"

- Bayesian FDR:

$$\mathrm{FDR}^{\mathrm{Bayes}} = E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}}\middle|\hat{\alpha}^1,\ldots,\hat{\alpha}^N,\tau\right)$$

where we condition on denominator$\neq 0$, otherwise FDR=0

Bayesian Multiple Testing

Proposition (Bayesian FDR)

Conditional on the parameters of the prior distribution and data with at least one discovery, the **Bayesian false discovery rate** *is*

$$FDR^{Bayes} = \frac{1}{\#discoveries} \sum_{i \text{ discovery}} p\text{-}null_i \leq 2.5\%.$$

Bayesian multiple testing

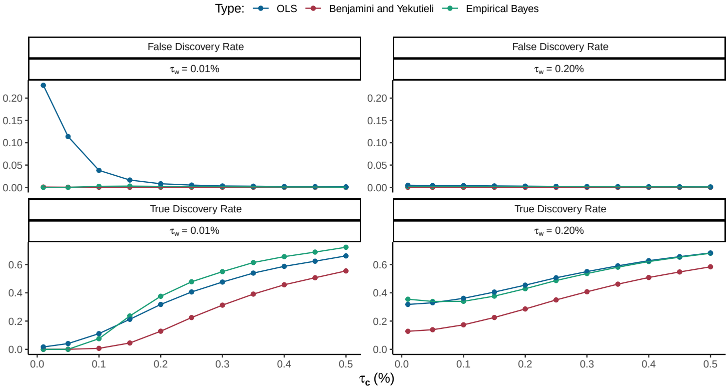
- Controls false discoveries, yet preserves power (cf. frequentist corrections)
- Note: an "oracle" result, conditional on knowing the prior parameters
- From posterior, can make *any* inference calculation (p-value, FDR, FWER, ...)

Comparison with standard approach

	Literature: OLS+MT correction	**Our hierarchical model**
Joint estimation	no	yes
Point estimate	OLS (no MT correction)	shrunk$\longrightarrow$conservative prior, cluster mean
Confidence interval	widens	contracts
False discoveries	controlled	controlled
#discoveries vs. OLS	lower	lower or higher, *depending on data*
Power	sacrificed	preserved

Bayesian Multiple Testing: Simulation

$$\alpha^i = \alpha^o + c^j + \omega^i, \quad c^j \sim N(0, \tau_c^2), w^i \sim N(0, \tau_w^2)$$



Upper panels: the realized FDR, i.e., proportion of discovered factors for which the true alpha is negative, averaged over 10,000 simulations.
Lower panels: the true discovery rate, i.e., number of discoveries where the true alpha is positive divided by the total number of factors where the true alpha is positive.

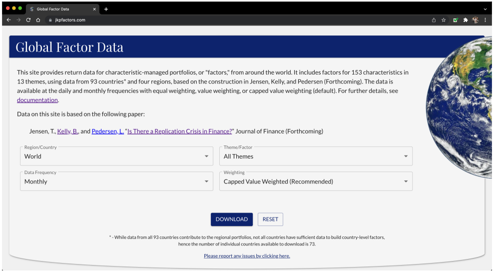
Left and right panels: use low and high values of idiosyncratic variation in alphas (τ_w), respectively. The x-axis varies cluster alpha dispersion, τ_c.

Evidence on Replication of Equity Factors

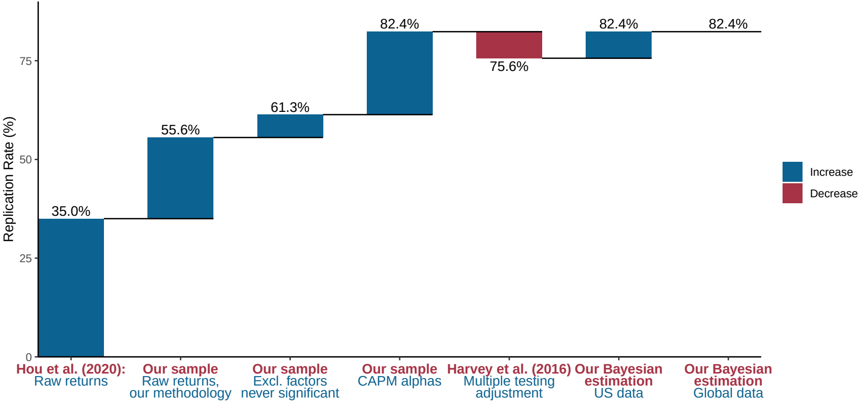
Based on Jensen et al. (2023)

A New Public Data Set of Global Factors

- Data:
  - US: CRSP (price and return) and COMPUSTAT (accounting)
  - Rest of the world: COMPUSTAT
- Coverage
  - Countries: 93
  - Factors: 153 of which 119 originally significant
  - Return time period: 1926-2022 (US) and 1986-2022 (rest)
- Factors
  - (Capped-)value-weighted within top/bottom terciles
  - Clustered via hierarchical clustering
- Code and data publicly available https://JKPfactors.com



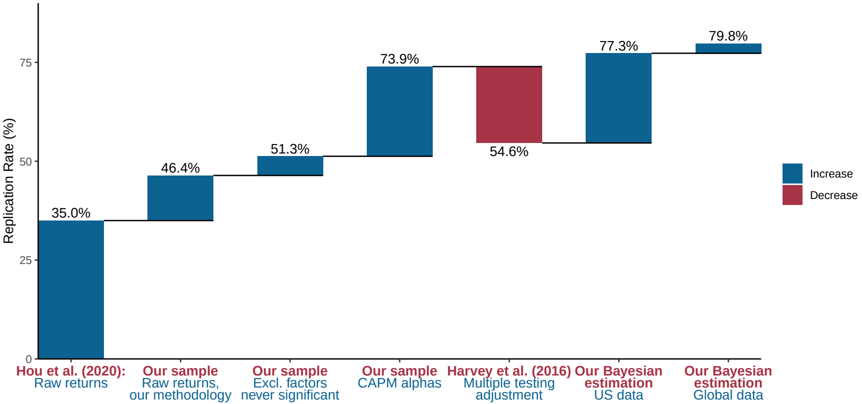
Jensen et al. (2023) Results vs. Literature



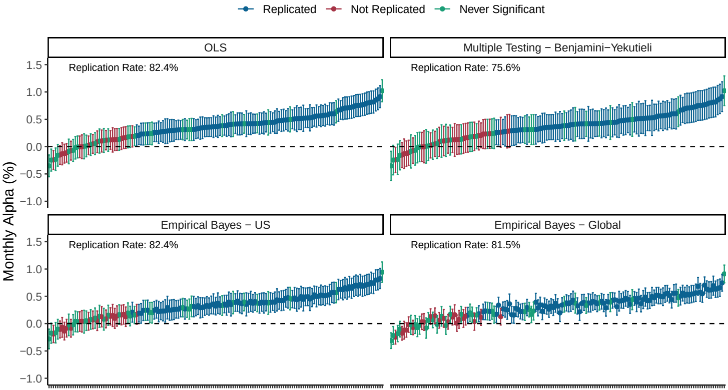
Details on differences in sample and factor construction:

- Capped value weights (+9.2%)
- 1 month holding period (vs. 1, 6, and 12 month) (+5.0%)
- Longer time series (+8.3%)
- Top/bottom 33% (rather than 10%) and other breakpoints (-6.0%)
- Consistent factor construction and other robustness (4.0%)

Jensen et al. (2023) vs. Literature: Straight Value Weights

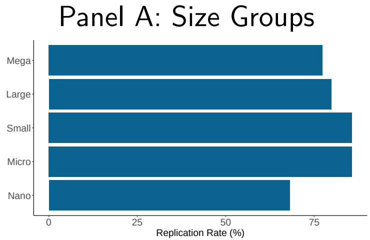


Internal Validity: Replication of US Factors

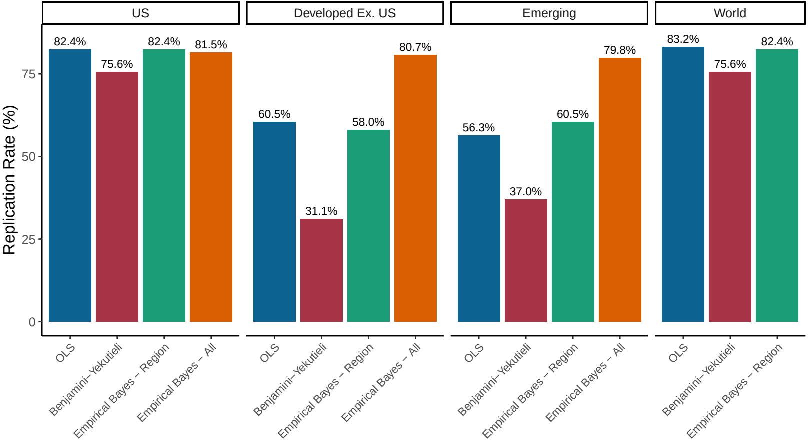


	Literature: OLS+MT correction	Our hierarchical model
Joint estimation	no	yes
Point estimate	OLS (no MT correction)	shrunk⟶conservative prior, cluster mean
Confidence interval	widens	contracts
False discoveries	controlled	controlled
#discoveries vs. OLS	lower	lower or higher, *depending on data*
Power	sacrificed	preserved

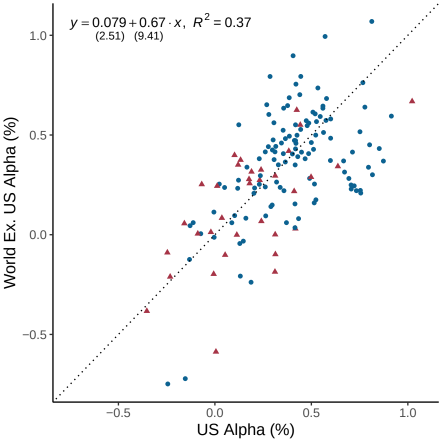
Internal Validity across Size Groups and Themes



External Validity: Global

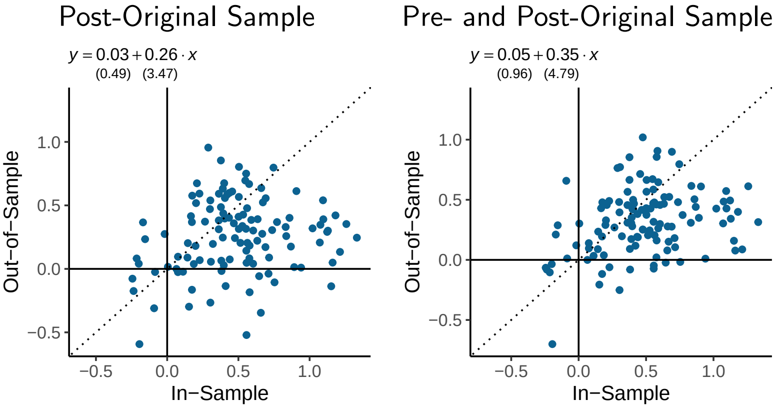


External Validity: Global



- Blue dots: factors that were significant in the original study
- Red triangles: factors not significant in the original paper
- Dotted line: 45° line.

External Validity: US Time Series



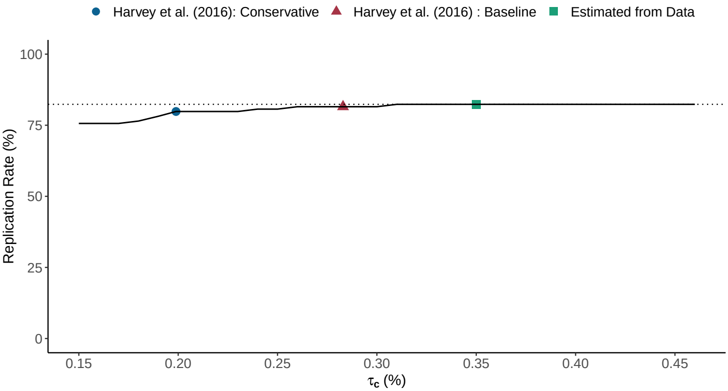
Our Bayesian Multiple Testing: Economic Benefits

- Factors
  - *discovered* by our EB framework
  - *but rejected* by Harvey et al. (2016)
- performance out-of-sample relative to original publication

	Full sample	Post-Harvey et al.
IR: US	0.93	1.00
	(5.16)	(2.83)
IR: World ex. US	1.10	1.60
	(6.13)	(4.52)

Addressing Publication Bias

- Factor more likely to be published if it appears to work
  - Researchers have tried more factors →unobserved
  - Unobserved factors affect distribution of alphas
- Addressing the issue in our Bayesian framework:
  - Adjust distribution of alphas to have more mass around 0
  - E.g., use distribution of unob. factors from Harvey et al. (2016)



Estimate of FDR: Another Benefit of Our Model

Recall:

Proposition (Bayesian FDR)

Conditional on the parameters of the prior distribution and data with at least one discovery, the **Bayesian false discovery rate** *is*

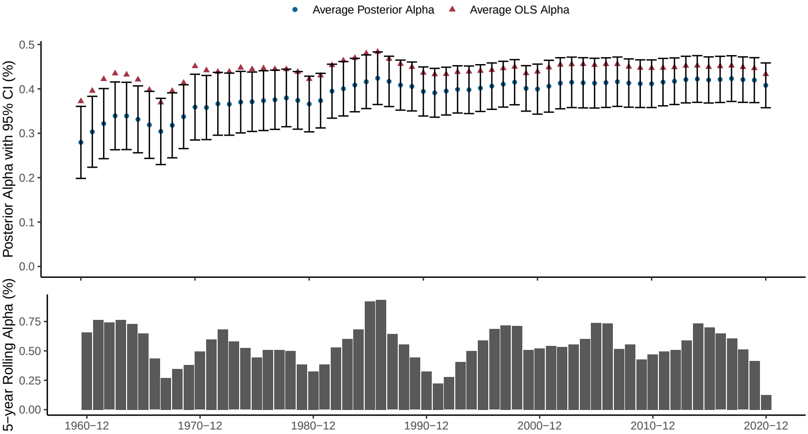
$$FDR^{Bayes} = \frac{1}{\#discoveries} \sum_{i \text{ discovery}} p\text{-}null_i \leq 2.5\%.$$

Empirically:

$$\text{FDR}^{\text{Bayes}} = E\left(\frac{\sum_i 1_{\{i \text{ false discovery}\}}}{\sum_i 1_{\{i \text{ discovery}\}}} \middle| data \right) = 0.1\%$$

$$\text{FWER}^{\text{Bayes}} = Pr\left(\sum_i 1_{\{i \text{ false discovery}\}} \geq 1 \middle| data \right) = 5.5\%$$

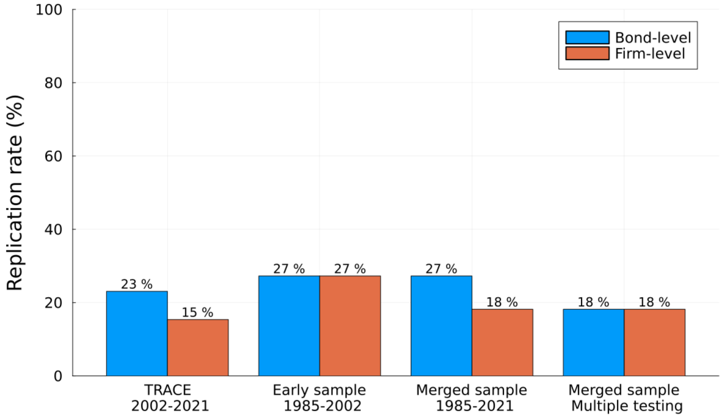
Bayesian Posterior: Evolution over Time



Evidence on Replication of Corporate Bond Factors

Based on Dick-Nielsen et al. (2023)

Replication Rate for Corporate Bond Factors



- Dick-Nielsen et al. (2023) make a new clean data set of
  - individual corporate bond returns
  - corporate bond factors

Other Replication Problems in Finance or Economics?

References Cited in Slides I

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze 8*, 3–62.

Chen, A. Y. and T. Zimmermann (2022). Open source cross-sectional asset pricing. *Critical Finance Review, forthcoming*.

Dick-Nielsen, J., P. Feldhütter, L. H. Pedersen, and C. Stolborg (2023). Corporate bond factors: Replication failures and a new framework. *Available at SSRN*.

Efron, B. and T. Hastie (2021). *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, Volume 6. Cambridge University Press.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Gelman, A., J. Hill, and M. Yajima (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness 5*(2), 189–211.

Hamermesh, D. S. (2007). Replication in economics. *Canadian Journal of Economics/Revue canadienne d'économique 40*(3), 715–733.

Harvey, C. R., Y. Liu, and H. Zhu (2016). . . . and the cross-section of expected returns. *The Review of Financial Studies 29*(1), 5–68.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *The Review of Financial Studies 33*(5), 2019–2133.

Jacobs, H. and S. Müller (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics 135*(1), 213–230.

References Cited in Slides II

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

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

McLean, R. D. and J. Pontiff (2016). Does academic research destroy stock return predictability? *The Journal of Finance 71*(1), 5–32.

Welch, I. (2019). Reproducing, extending, updating, replicating, reexamining, and reconciling. *Critical Finance Review 8*(1-2), 301–304.