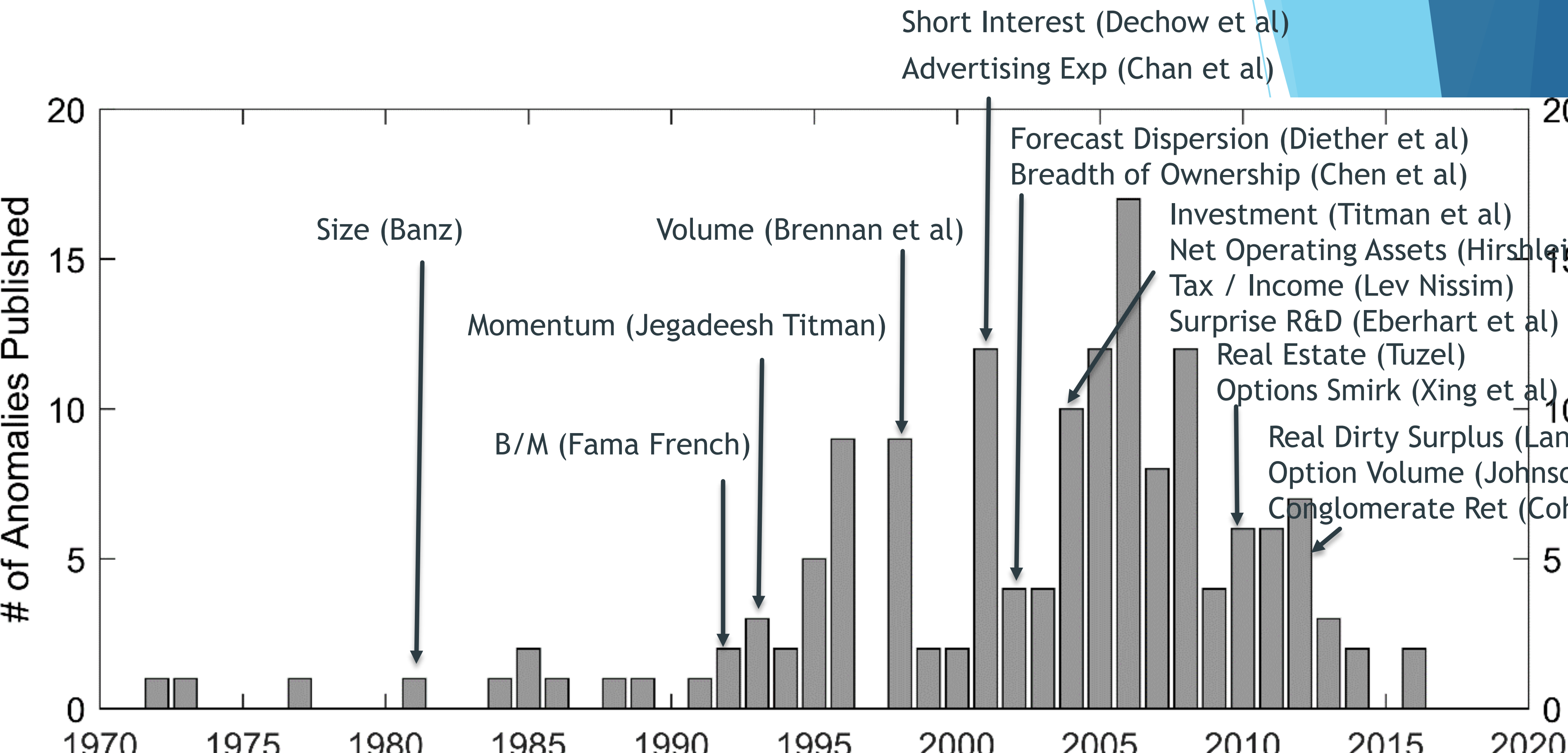
Week 7: Making Sense of the Anomaly Zoo

Professor Mihail Velikov

FIN 597 – Spring 2022

PennState
Smeal College of Business

From Chen and Velikov (2022)



How do we make sense of it?

- Clearly, they can't all be due to risk
  - What explains the zoo then?

- Several types of explanations
  - Lots of anomalies could just be manifestations of previous anomalies
    - Factor models (Fama and French, 1996; 2008; 2016; Hou, Xue, and Zhang, 2020)
  - Mispricing
    - If markets are efficient ⇒ predictability should go away post-publication (Schwert, 2003; Huang and Huang, 2013; McLean and Pontiff, 2016; Jacobs and Muller, 2020)
  - Data mining
    - Researchers try many things, some work (Harvey et al., 2016; Linnainmaa and Roberts, 2018; Chen and Zimmerman, 2020)
  - Transaction costs
    - Academic strategies not meant to be traded; real profitability much lower (Novy-Marx and Velikov, 2016; Chen and Velikov, 2022)

PennState
Smeal College of Business

Where do we go from here?

- Regardless of what explains the zoo, an emerging literature tries to combine anomalies:
  - Haugen and Baker (1997)
  - Brandt, Santa-Clara, and Valkanov (2009)
  - Lewellen (2015)
  - Light, Maslov, and Rytchkov (2017)
  - Freyberger, Neuhierl, and Weber (2020)

- … or to identify the independent characteristics that matter:
  - Green, Hand, and Zhang (2017)
  - Kelly, Pruitt, and Su (2019)
  - Kozak, Nagel, and Santosh (2020)
  - Feng, Giglio, and Xiu (2020)

PennState
Smeal College of Business

Factor models

- Fama and French
  - 1996: Multifactor explanations of asset pricing anomalies
    - Size, E/P, CF/P, B/M, Sales Growth, Long-run reversals explained by 3-factor model
    - Momentum anomalous
  - 2008: Dissecting anomalies
    - Splits into size groups (micro, small, and big)
    - Net issuance, accruals, and momentum anomalous
    - Asset growth and profitability less robust
  - 2016: Dissecting anomalies with a five-factor model
    - Share repurchases, low volatility explained by the five-factor model

- Hou, Xue, and Zhang (2020)
  - Value-weighting + NYSE breakpoints makes 65% of "452" anomalies disappear
  - 452 anomalies is a completely artificially inflated number
    - They do three versions of all monthly rebalanced strategies

How much of anomalies is due to mispricing?

- Schwert (2003)
  - Anomalies largely disappear post-publication
  - Looks at size, value, momentum, and the weird time-of-the-year anomalies (January, weekend effects)
  - Evidence points to a mix of mispricing and data mining

- Huang and Huang (2013)
  - A long-only strategy that rotates between the best out of 12 <u>published</u> anomaly strategies outperforms market after trading costs
  - Conclude that anomalies persist after controlling for data-snooping bias

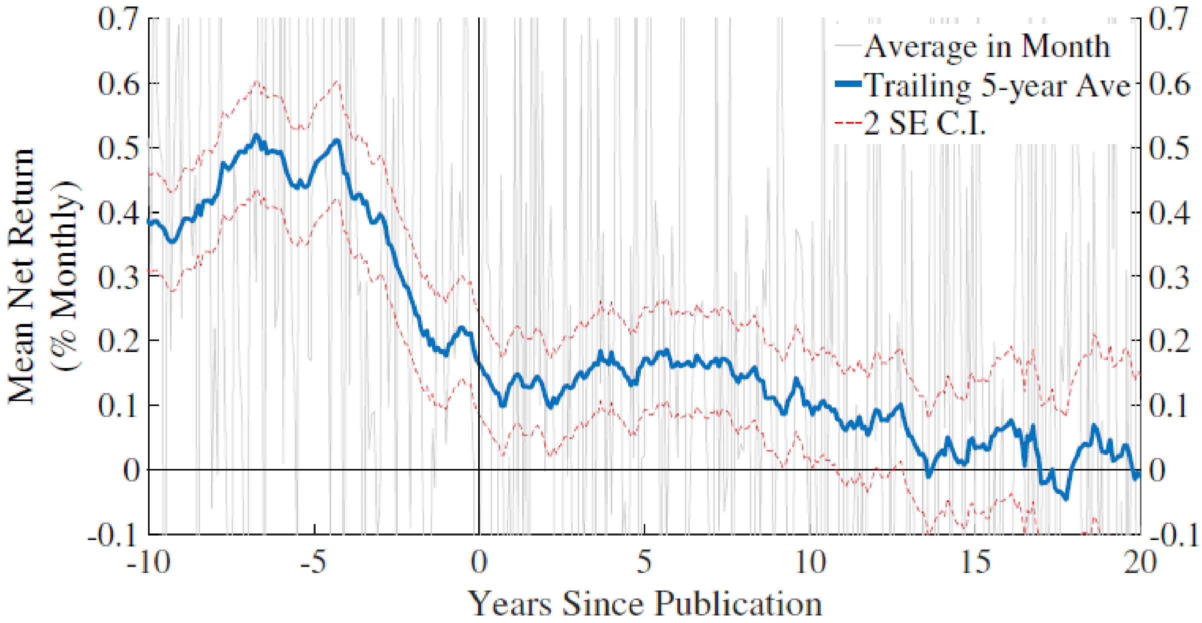
- McLean and Pontiff (2016)
  - Look at out-of-sample and post-publication returns for 97 predictive variables
  - Find portfolio returns decline 26% out-of-sample (after sample in paper ends and before publication) and 58% post-publication
  - Out-of-sample decline is upper bound on data mining, so 32% (58%-26%) lower bound on publication-induced trading. We matter!

PennState
Smeal College of Business

How much of anomalies is due to mispricing?

- McLean and Pontiff (2016) result from Chen and Velikov (2022):

Figure 5: Event-Time Net Returns for Cost-Mitigated Implementations. For a given month relative to publication, light lines plot the mean net return across all anomalies. Dark lines show the trailing 5-year moving average of mean returns, and dashed lines show 2 standard error confidence bounds. Cost mitigation is effective before publication, but net returns become tiny afterwards.



PennState
Smeal College of Business

How much of anomalies is due to mispricing?

- Jacobs and Muller (2020)
  - Internationally, we don’t observe the same effects as in McLean and Pontiff (2016)
  - Look at 241 cross-sectional anomalies in 39 stock markets
  - US is the only country with reliable post-publication decline in returns
  - Interpret this as evidence of barriers to arbitrage creating segmented markets

PennState
Smeal College of Business

How much is data mining?

- Some old literature on statistical biases in finance research
  - Leamer (1978), Ross (1989), Lo and Mackinlay (1990), and Fama (2003)

- Linnainmaa and Roberts (2017)
  - They look at 36 accounting anomalies for the 1918-1960s period using Moody's manuals
    - Focus is on profitability and investment
  - Combine pre- and post-publication samples
  - Conclude that:

> … most accounting-based return anomalies are spurious. When examined out-of-sample by moving either backward or forward in time, anomalies' average returns decrease, and volatilities and correlations with other anomalies increase

PennState
Smeal College of Business

How much is data mining?

- Harvey, Liu, and Zhu (2016)
  - Starts an entire agenda against data mining in finance academia
    - You should really see Harvey's slides (e.g., here and here)
  - Argues we should account for multiple testing bias
    - You only report the positive results, without telling us how many you tested
  - Claims there are more than 316 published papers that "study cross-sectional return patterns"
    - However!
      - They include theory papers (e.g., Lucas, 1978; Breeden, 1979)
      - They include 63 working papers
  - Conclude that "most claimed research findings in financial economics are likely false."

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

Table 2
Contingency table in testing M hypotheses

Panel A: An example

	Unpublished	Published	Total
Truly insignificant	500	50	550
Truly significant	100	50	150
Total	600	100(R)	700(M)

Panel B: The testing framework

	H_0 not rejected	H_0 rejected	Total
H_0 true	$N_{0\|a}$	$N_{0\|r}$	M_0
H_0 false	$N_{1\|a}$	$N_{1\|r}$	M_1
Total	$M-R$	R	M

Panel A shows a hypothetical example for factor testing. Panel B presents the corresponding notation in a standard multiple testing framework.

- In multiple hypothesis testing, Type I error should account for joint occurrence of false discoveries ($N_{0|r}$)

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Two ways of summarizing the "joint occurrence"
  - Family-wise error rate (FWER)

$$FWER = \Pr(N_{0|r} \geq 1)$$

    - Measures probability of even a single false discovery
    - We assign a significance or threshold level α, and have methods to ensure FWER does not exceed it

  - False discovery rate (FDR):

$$FDR = E[FDP]$$

    - the expected proportion of false discoveries among all discoveries FDR

$$FDP = \begin{cases} \frac{N_{0|r}}{R}, & R > 0 \\ 0, & R = 0 \end{cases}$$

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- FWER is more stringent than FDR
  - Especially for large M

- Remedy typically prescribed by FDR and FWER adjustment procedures is to lower p-value thresholds for the individual hypotheses

- Usually this leads to an increase in Type II errors as well
  - I.e., the number of times we would not reject a false null

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Three approaches to adjust for multiple testing bias
  - Suppose we have M tests and we set a FWER at α_w and FDR at α_d
  - Example: $M = 10, \alpha_w = \alpha_D = 5\%$

Table 3
A summary of *p*-value adjustments

Adjustment type	Single/Sequential	Multiple test
Bonferroni	single	family-wise error rate
Holm	sequential	family-wise error rate
Benjamini, Hochberg, and Yekutieli (BHY)	sequential	false discovery rate

- Bonferroni's adjustment
  - Single-step correction, controls for FWER
  - Reject any hypothesis with p-value $\leq \frac{\alpha_w}{M}$
  - Alternatively, multiply the individual test p-values by M:

$$p_i^{Bonferroni} = \min[M \times p_i, 1]$$

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Holm's adjustment:
  - Sequential method, controls for FWER
  - Procedure:
    - Order the original *p*-values such that $p_{(1)} \leq p_{(2)} \leq \ldots\ p_{(b)} \leq \ldots\ p_{(M)}$, and let the associated null hypotheses be $H_{(1)}, H_{(2)}, \ldots, H_{(b)} \ldots, H_{(M)}$
    - Let *k* be the minimum index such that $p_{(b)} > \frac{\alpha_w}{M+1-b}$
    - Reject the null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$ (i.e., declare these factors significant), but do not reject $H_{(k)}, \ldots, H_{(M)}$
  - Intuition:
    - Instead of comparing the p-values to $\frac{\alpha_w}{M}$, compare to $\frac{\alpha_w}{M}, \frac{\alpha_w}{M-1}, \ldots, \frac{\alpha_w}{2}, \frac{\alpha_w}{1}$
    - Think of *k* as the first *p*-value that is not low enough to validate rejection
  - Adjusted p-value given by:

$$p_i^{Holm} = \min[\max_{j \leq i}\{(M-j+1)p_j\}, 1]$$

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Benjamini, Hochberg, and Yekutieli's (BHY) adjustment:
  - Sequential method, controls for FDR
  - Procedure:
    - Order the original p-values such that $p_{(1)} \leq p_{(2)} \leq \ldots\ p_{(b)} \leq \ldots\ p_{(M)}$, and let the associated null hypotheses be $H_{(1)}, H_{(2)}, \ldots, H_{(b)} \ldots, H_{(M)}$
    - Let k be the maximum index such that $p_{(b)} \leq \frac{b}{M \times c(M)} \alpha_d$
      - where $c(M)$ is a function that controls for the generality of the test, e.g.,

$$c(M) = \sum_{j=1}^{M} \frac{1}{j}$$

    - Reject the null hypotheses $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ (i.e., declare these factors significant), but do not reject $H_{(k+1)}, \ldots, H_{(M)}$
    - Adjusted p-value given by:

$$p_{(i)}^{BHY} = \begin{cases} p_{(M)}, \ if\ i = M \\ \min[p_{(i+1)}^{BHY}, \frac{M \times c(M)}{i} p_{(i)}],\ if\ \ i \leq M-1 \end{cases}$$

PennState Smeal College of Business

Harvey, Liu, and Zhu (2016)

Table 4
An example of multiple testing

Panel A: Single tests and "significant" factors

Test →	1	2	3	4	5	6	7	8	9	10	# of discoveries
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	10
p-value (%)	**4.66**	**0.85**	**2.71**	**0.05**	**3.00**	**0.84**	**0.00**	**0.00**	**0.60**	**1.28**	

Panel B: Bonferroni "significant" factors

Test →	1	2	3	4	5	6	7	8	9	10	
t-statistic	1.99	2.63	2.21	3.43	2.17	2.64	4.56	5.34	2.75	2.49	3
p-value (%)	4.66	0.85	2.71	**0.05**	3.00	0.84	**0.00**	**0.00**	0.60	1.28	

Panel C: Holm adjusted p-values and "significant" factors

Reordered tests b	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Old order	8	7	4	9	6	2	10	3	5	1	4
p-value (%)	**0.00**	**0.00**	**0.05**	**0.60**	0.84	0.85	1.28	2.71	3.00	4.66	
$\alpha_w/(M+1-b)$ $\alpha_w = 5\%$	0.50	0.56	0.63	0.71	0.83	1.00	1.25	1.67	2.50	5.00	

Panel D: BHY adjusted p-values and "significant" factors

Reordered tests b	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	
Old order	8	7	4	9	6	2	10	3	5	1	6
p-value (%)	**0.00**	**0.00**	**0.05**	**0.60**	**0.84**	**0.85**	1.28	2.71	3.00	4.66	
$(b \cdot \alpha_d)/(M \times c(M))$ $\alpha_d = 5\%$	0.17	0.34	0.51	0.68	0.85	1.02	1.19	1.37	1.54	1.71	

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

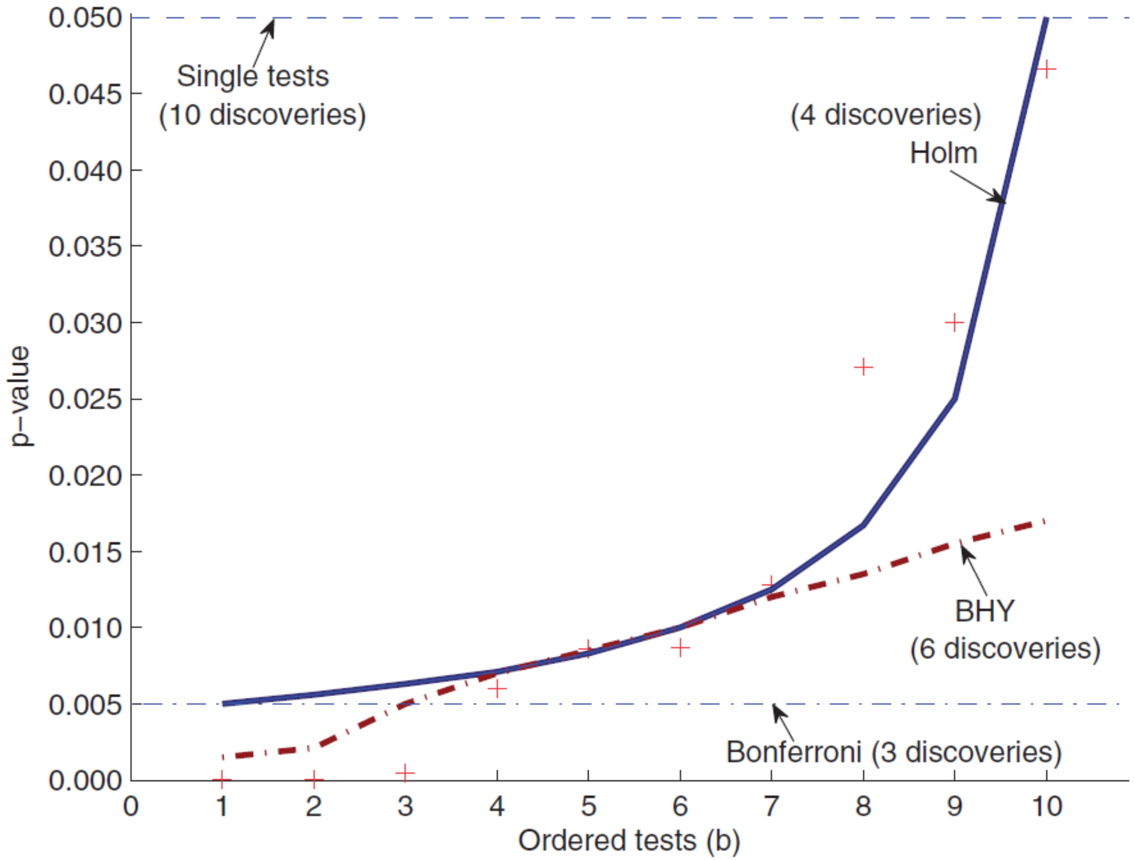


Figure 1
Multiple test thresholds for example A,
The ten *p*-values for the example in Table 4 and the adjusted *p*-value lines for various adjustment procedures. All ten factors are discovered using the standard criteria for single tests, three under Bonferroni, four under Holm, and six under BHY. The significance level is set at 5% for each adjustment method.

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Academic factors sampled by them:

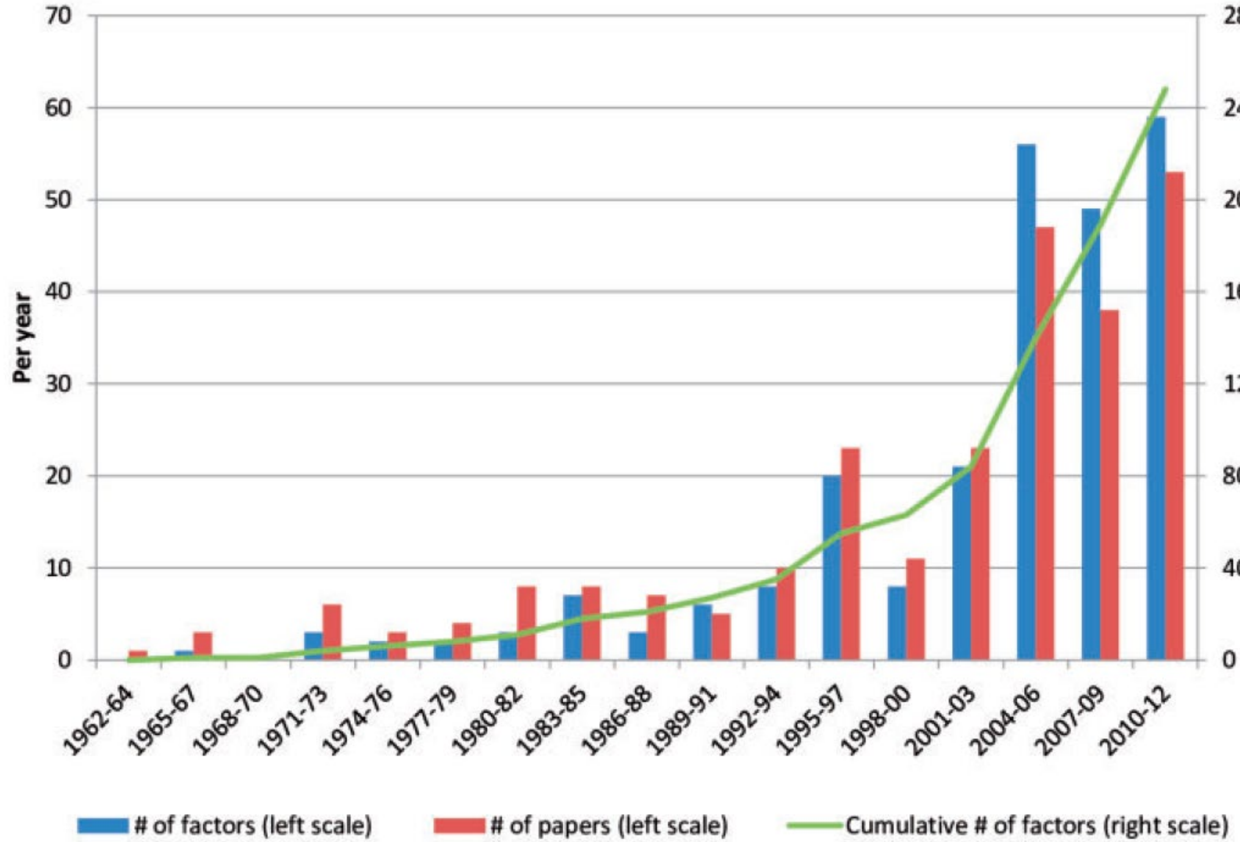
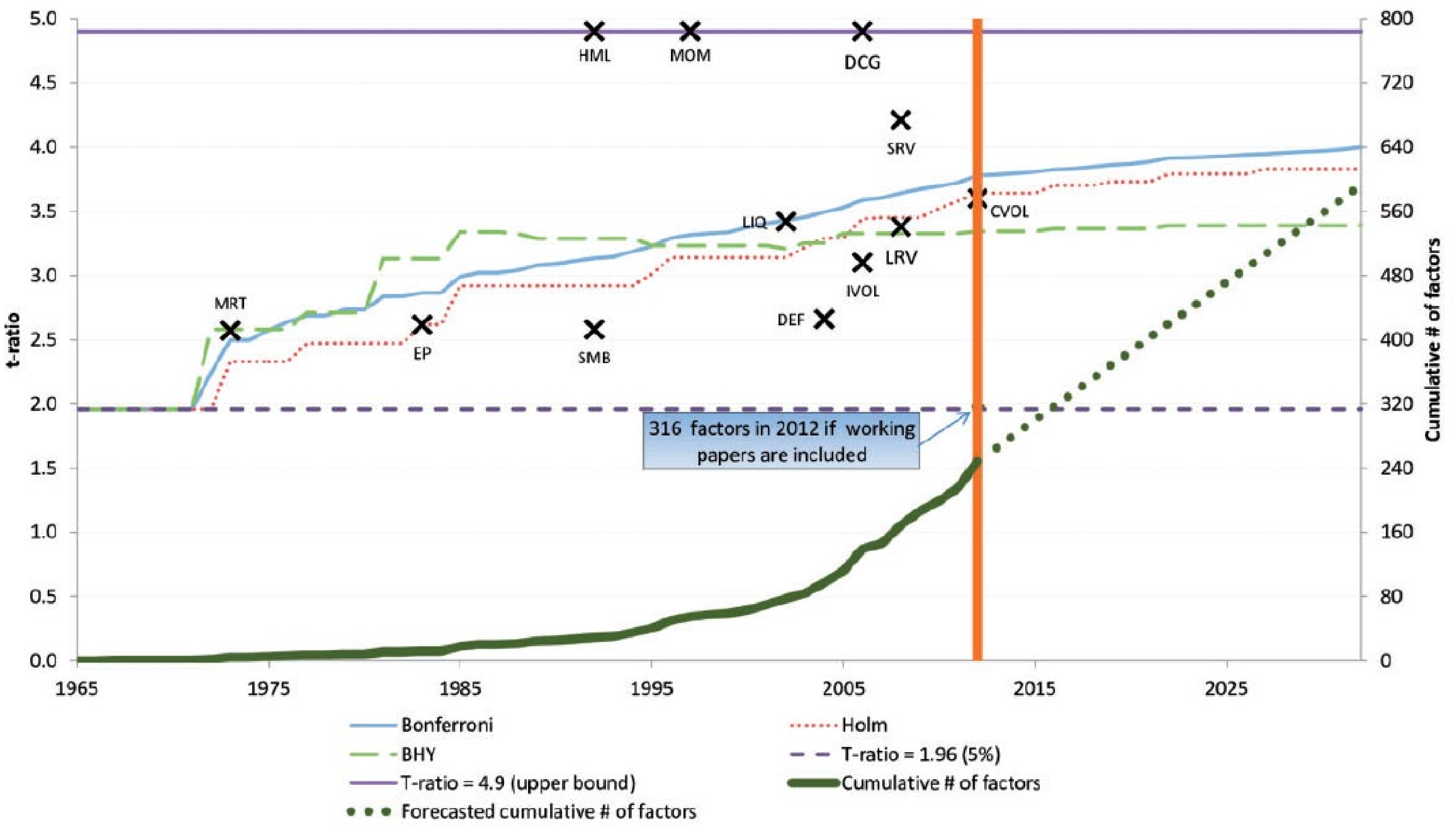


Figure 2
Factors and publications.

PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- T-stat threshold adjusted for multiple testing (assuming $M = R$):



PennState
Smeal College of Business

Harvey, Liu, and Zhu (2016)

- Here's them effectively admitting data mining in paper criticizing data mining:

One concern with our results is that factors are discovered at different times and tests are conducted using different methods. This heterogeneity in the time of discovery and testing methods may blur the interpretation of our results. Ideally, we want updated factor tests that are based on the most recent sample and the same testing method.[31] To alleviate this concern, we focus on the group of factors that are published no earlier than 2000 and rely on Fama-MacBeth tests. Additionally, we require that factor tests cover at least the 1970–1995 period and have as controls at least the Fama-French three factors (Fama and French 1993). This leaves us with 124 factors. Based on this factor group, the Bonferroni and Holm implied threshold *t*-statistics are 3.54 and 3.20 (5% significance), respectively, and the BHY implied thresholds are 3.23 (1% significance) and 2.67 (5% significance) by 2012. Not surprisingly, these statistics are smaller than the corresponding thresholds based on the full sample.

PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Formal model accounting for what they call publication bias (i.e., data mining)
  - Replicate 156 predictors from papers published in selective journals
  - Conclude that nearly all predictors were real!

- In their model, there is tension between:
  - Data mining
    - Too many predictors
    - Some were large in-sample by pure chance
    - We don't know how many signals people tried
    - ⇒ must have a lot of data mining
  - Publication process
    - You need more than just a tstat to convince the referees and editor
    - The publication process corrects for data mining

PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Motivating story for their model (from Andrew's AFA slides):

"The Lord of the p-value"

p-hacking
- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

Journal Review
- robustness tests
- theoretical motivations
- supporting results
- a scientific, ethical culture

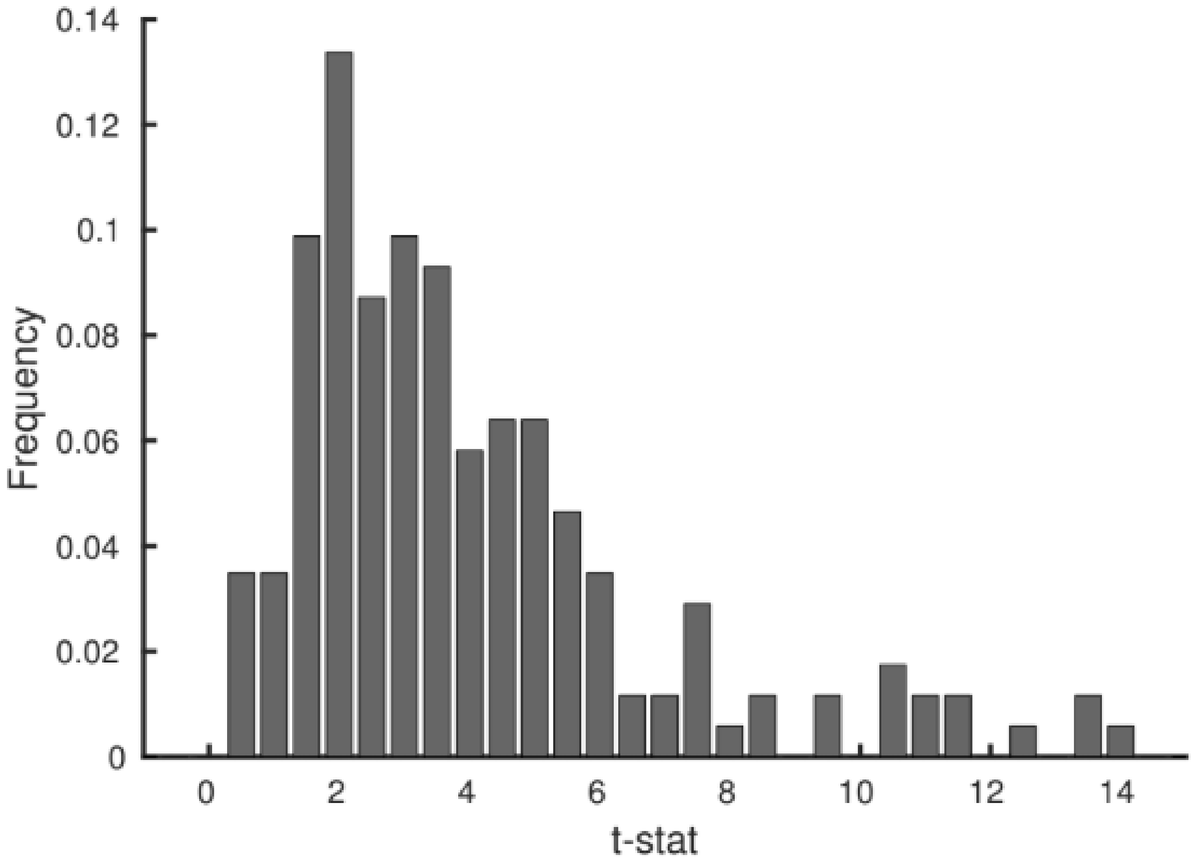
The Cross-Sectional Asset Pricing Lit

Our Question: Which Side is Winning?

PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Basic intuition is that if all anomalies were data-mined, you wouldn't see the long right tail in anomaly returns



PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Simple model:
  - Assume:

$$r_i = \mu_i + \sigma_i \epsilon_i$$
$$\mu_i \sim N(0, \sigma_\mu)$$
$$\epsilon_i \sim N(0,1)$$

    - where r_i is the sample mean return for predictor i, μ_i is the true mean return, σ_i is the standard error, and σ_μ is the dispersion of true returns.

  - We would like to know the distribution of μ_i
    - However, we can only observe r_i conditioned on signal i being published:

$$E(r_i|published_i) = E(\mu_i|published_i) + \underbrace{E(\epsilon_i|published_i)}_{>0}$$

    - we may be observing signals that got lucky in-sample

PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Main insight:

$$E(\mu_i | r_i, \sigma_i, \sigma_\mu, published_i) = E(\mu_i | r_i, \sigma_i, \sigma_\mu)$$

  - Bayes' rule is "immune" to selection bias (Efron, 2011)
  - Assuming you know $r_i, \sigma_i, \sigma_\mu$, you can make inference regarding the distribution of μ_i

- Assuming Bayesian normal-normal updating:

$$E(\mu_i | r_i, \sigma_i, \sigma_\mu) = (1 - s_i) r_i$$

  - where $s_i = \frac{\sigma_i^2}{\sigma_\mu^2 + \sigma_i^2}$ is the "shrinkage" or bias in the observed mean return

- They actually estimate a richer model using MLE, but intuition is similar
  - Main result is that the average shrinkage is about 12.3%

PennState
Smeal College of Business

Chen and Zimmerman (2020)

- Identification comes from the right tail
  - If no dispersion in μ_i, then published predictors are those with t-stats > 2
  - Then they would follow a standard normal distribution truncated at 2

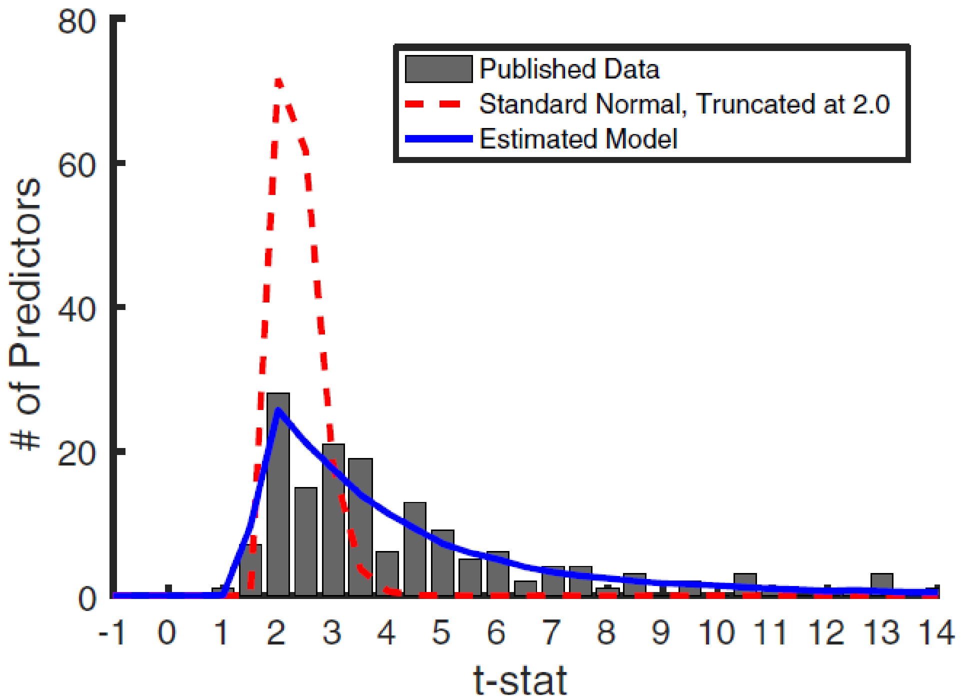


Figure 1
Published data versus models

PennState
Smeal College of Business

Shrinking the cross-section

- Emerging literature trying to do two things:
  - Figure out ways to combine anomalies
  - More efficiently distinguish independent effects in the cross-section

- Haugen and Baker (1997) & Lewellen (2015)
  - Sort stocks based on fitted values of backward-looking Fama-MacBeth regressions
  - Not surprisingly, combination strategies work quite well.

- Related paper: Clarke (2022)
  - See discussion on Cochrane's blog

PennState
Smeal College of Business

Brandt, Santa-Clara, and Valkanov (2009)

- Pretty neat idea - parametric portfolio policies

- Methodology
  - Investors choose weights to maximize the conditional expected utility of portfolio return:

$$\max_{\{w_{i,t}\}_{i=1}^{N_t}} E_t[u(r_{p,t+1})] = E_t\left[u\left(\sum_{i=1}^{N_t} w_{i,t} r_{i,t+1}\right)\right]$$

  - Portfolio weights are a function of the stocks' characteristics:

$$w_{i,t} = \bar{w}_{i,t} + \frac{1}{N_t}\theta^T \hat{x}_{i,t}$$

  - Assuming a CRRA functional form for $u(\cdot)$, they estimate θ using GMM:

$$\max_{\theta} \frac{1}{T}\sum_{t=0}^{T} u\left(\sum_{i=1}^{N_t}\left(\bar{w}_{i,t} + \frac{1}{N_t}\theta' \hat{x}_{i,t}\right) r_{i,t+1}\right)$$

PennState
Smeal College of Business

Brandt, Santa-Clara, and Valkanov (2009)

Table 1
Simple linear portfolio policy

Variable	VW	EW	In Sample PPP	Out of Sample PPP
θ_{me}	–	–	−1.451	−1.124
Std. err.	–	–	(0.548)	(0.709)
θ_{btm}	–	–	3.606	3.611
Std. err.	–	–	(0.921)	(1.110)
θ_{mom}	–	–	1.772	3.057
Std. err.	–	–	(0.743)	(0.914)
LRT p-value	–	–	0.000	0.005
$\lvert w_i \rvert \times 100$	0.023	0.023	0.083	0.133
$\max\ w_i \times 100$	3.678	0.023	3.485	4.391
$\min\ w_i \times 100$	0.000	0.023	−0.216	−0.386
$\sum w_i I(w_i < 0)$	0.000	0.000	−1.279	−1.447
$\sum I(w_i \leq 0)/N_t$	0.000	0.000	0.472	0.472
$\sum \lvert w_{i,t} - w_{i,t-1} \rvert$	0.097	0.142	0.990	1.341
CE	0.064	0.069	0.175	0.118
$\bar{r}$	0.139	0.180	0.262	0.262
$\sigma(r)$	0.169	0.205	0.188	0.223
SR	0.438	0.564	1.048	0.941
α	–	–	0.174	0.177
β	–	–	0.311	0.411
$\sigma(\epsilon)$	–	–	0.181	0.214
IR	–	–	0.960	0.829
me	2.118	−0.504	−0.337	−0.029
btm	−0.418	0.607	3.553	3.355
mom	0.016	0.479	1.623	2.924

This table shows estimates of the portfolio policy with three characteristics: size (me), book-to-market ratio (btm), and momentum (mom), as specified in Equation (3) and optimized for a power utility function with a relative risk aversion of five. We use data from the merged CRSP–Compustat database from January 1964 to December 2002. In the "out-of-sample" results, we use data until December 1973 to estimate the coefficients of the portfolio policy and then form out-of-sample monthly portfolios using those coefficients in the next year. Every subsequent year, we reestimate the portfolio policy by enlarging the sample. All statistics are reported for the period January 1974 to December 2002. The columns labeled "VW," "EW," and "PPP" display statistics of the market-capitalization-weighted portfolio, the equally weighted portfolio, and the optimal parametric portfolio policy, respectively. The first set of rows shows the estimated coefficients of the portfolio policy with bootstrapped standard errors in parentheses. The bootstrapped p-value of the Wald test under the null hypothesis that the parameter estimates are jointly equal to zero is also displayed. The "out-of-sample" results display time-series averages of coefficients, standard errors, and p-values. The second set of rows shows statistics of the portfolio weights averaged across time. These statistics include the average absolute portfolio weight, the average minimum and maximum portfolio weights, the average sum of negative weights in the portfolio, the average fraction of negative weights in the portfolio, and the turnover in the portfolio. The third set of rows displays average portfolio return statistics: certainty-equivalent return, average return, standard deviation, and Sharpe ratio of returns; the alpha, beta, and volatility of idiosyncratic shocks of a market model regression; and the information ratio. The final set of rows displays the average normalized characteristics of the portfolio. The average risk-free rate in the sample is 0.061 (annualized).

PennState
Smeal College of

Green, Hand, and Zhang (2017)

- Summary
  - Fama-MacBeth regressions of returns on 94 characteristics
    - Avoid overweighting microcaps (vwls & all-but-microcaps) & adjust p-values for data snooping (BHY)
  - Conclude only 12 matter
  - Green (former PSU accounting Professor) has SAS code on constructing characteristics on his website

PennState
Smeal College of Business

Light, Maslov, and Rytchkov (2017)

- Summary
  - Partial least squares estimator that filters out expected returns from 26 characteristics that predict returns
  - Horse race with other methods (factor analysis, PCA, Fama-MacBeth, sort on average rank)

- Methodology – two-step estimation of expected returns in each month, t:
  - Step 1: For each characteristic, a, estimate the following cross-sectional regression (observation is a stock in that month):
$$R_{it} = \lambda_t^a X_{it-1}^a + \varepsilon_i$$
  - Step 2: For each firm, i, estimate the following regression (observation is a characteristic in that month):
$$X_{it}^a = \mu_{it} \hat{\lambda}_t^a + \varepsilon_a$$
- Use $\hat{\mu}_{it}$ as an expected return proxy

PennState
Smeal College of Business

Shrinking the cross-section

- Kelly, Pruitt, and Su (2019)
  - Instrumented Principal Component Analysis (IPCA)
    - Latent factors ($f_{i,t}$) and time-varying loadings ($\beta_{i,t}$) by using observable characteristics ($z_{i,t}$) that instrument for the unobservable dynamic loadings

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{i,t+1} + \varepsilon_{i,t+1}$$

where:

$$\alpha_{i,t} = z'_{i,t}\Gamma_\alpha + \nu_{\alpha,i,t}$$
$$\beta_{i,t} = z'_{i,t}\Gamma_\beta + \nu_{\beta,i,t}$$

  - Γ_β is the mapping from a potentially large number of characteristics $z_{i,t}$ to a small number of risk factor exposures $\beta_{i,t}$
    - Neat dimension reduction of the characteristics space

  - After plugging in, we get:

$$r_{i,t+1} = z'_{i,t}(\Gamma_\alpha + \Gamma_\beta f_{i,t+1}) + \varepsilon^*_{i,t+1}$$

where:

$$\varepsilon^*_{i,t+1} = \nu_{\alpha,i,t} + \nu_{\beta,i,t} f_{i,t+1} + \varepsilon_{i,t+1}$$

PennState
Smeal College of Business

Shrinking the cross-section

- Kelly, Pruitt, and Su (2019)
  - In vector form, the specification becomes:

$$\underbrace{r_{t+1}}_{N\times 1} = \underbrace{Z_t}_{N\times L}\,(\underbrace{\Gamma_\alpha}_{L\times 1} + \underbrace{\Gamma_\beta}_{L\times K}\,\underbrace{f_{t+1}}_{K\times 1}) + \underbrace{\varepsilon^*_{i,t+1}}_{N\times 1}$$

  - Restricted model ($\Gamma_\alpha = \mathbf{0}_{L\times 1}$)

$$\min_{\Gamma_\beta, F} \sum_{t=1}^{T-1} \left(r_{t+1} - Z_t\Gamma_\beta f_{t+1}\right)' \left(r_{t+1} - Z_t\Gamma_\beta f_{t+1}\right)\text{s}$$

FOC:

$$\hat{f}_{t+1} = \Big(\underbrace{\hat{\Gamma}'_\beta}_{K\times L}\,\underbrace{Z'_t}_{L\times N}\,\underbrace{Z_t}_{N\times L}\,\underbrace{\hat{\Gamma}'_\beta}_{L\times K}\Big)^{-1}\,\underbrace{\hat{\Gamma}'_\beta}_{K\times L}\,\underbrace{Z'_t}_{L\times N}\,\underbrace{r_{t+1}}_{N\times 1}\ \forall t$$

$$vec\left(\hat{\Gamma}'_\beta\right) = \left(\sum_{t=1}^{T-1} \underbrace{Z'_t}_{L\times N}\,\underbrace{Z_t}_{N\times L} \otimes \underbrace{\hat{f}_{t+1}}_{K\times 1}\,\underbrace{\hat{f}'_{t+1}}_{1\times K}\right)^{-1} \left(\sum_{t=1}^{T-1} \Big[\underbrace{Z_t}_{N\times L} \otimes \underbrace{\hat{f}'_{t+1}}_{1\times K}\Big]' \underbrace{r_{t+1}}_{N\times 1}\right)$$

  - Estimate $\hat{\Gamma}_\beta$ and $\hat{f}_{t+1}$ using alternating least squares

PennState
Smeal College of Business

Freyberger, Neuhierl, and Weber (2020)

- Summary
  - Adaptive group LASSO method on 62 characteristics to screen & combine them
  - Conclude only 13 have incremental explanatory power
  - Report some crazy Sharpe ratios

- Methodology:

$$R_{it} = \sum_{s=1}^{S} \tilde{m}_{ts}\left(\tilde{C}_{s,it-1}\right) + \varepsilon_{it}$$

  - $\tilde{C}_{s,it-1}$ are rank-transformed characteristics and $\tilde{m}_s()$ are given by:

$$\tilde{m}_{ts}\left(\tilde{C}_{s,it-1}\right) = \sum_{k=1}^{L+2} \beta_{tsk} p_k(\tilde{c})$$

  - where $p_k(\tilde{c})$ are known functions and β_{tsk} are estimated $\beta_{tsk} p_k(\tilde{c})$

PennState
Smeal College of Business

Freyberger, Neuhierl, and Weber (2020)

- Adaptive LASSO:
  - Step 1 (Pure LASSO):

$$\hat{\beta}_t = \underset{b_{sk}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k \left(\tilde{C}_{s,it-1} \right) \right)^2 + \lambda_1 \sum_{s=1}^{S} \left(\sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}$$

  - Step 2 (Adaptive LASSO):

$$\hat{\beta}_t = \underset{b_{sk}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(R_{it} - \sum_{s=1}^{S} \sum_{k=1}^{L+2} b_{sk} p_k \left(\tilde{C}_{s,it-1} \right) \right)^2 + \lambda_2 \sum_{s=1}^{S} \left(w_{ts} \sum_{k=1}^{L+2} b_{sk}^2 \right)^{\frac{1}{2}}$$

    - where $w_{ts} = \begin{cases} \left(\sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2\right)^{-\frac{1}{2}} & if\ \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 \neq 0 \\ \infty & if\ \sum_{k=1}^{L+2} \tilde{\beta}_{sk}^2 \neq 0 \end{cases}$

PennState
Smeal College of Business

Shrinking the cross-section

- Partial list of ever more sophisticated machine-learning techniques:
  - Feng, Polson, and Xu (WP, 2019)
  - Feng, Giglio, and Xiu (JF, 2020)
  - Gu, Kelly, and Xiu (RFS, 2020)
  - Lettau and Pelger (RFS, 2020)
  - Giglio, Liao, and Xiu (RFS, 2021)
  - Giglio and Xiu (JPE, 2021)
  - Chen, Pelger, and Zhu (WP, 2021)
  - Cong, Tang, Wang, and Zhang (WP, 2022)

- See review paper by Giglio, Kelly, and Xiu (2022)

PennState
Smeal College of Business

Issues with combining characteristics

- Novy-Marx (2016)
  - Backtests on signals that combine characteristics are severely biased
  - His paper was motivated by some of the earlier papers that combine characteristics
    - Piotroski's F-score (2000)
      - 9 signals
    - Asness et. al. Quality Score (2014)
      - 21 signals
    - Stambaugh-Yuan "mispricing factor" (2016)
      - 11 signals
  - But is even more relevant now with the literature we just discussed

PennState
Smeal College of Business

Novy-Marx (2016)

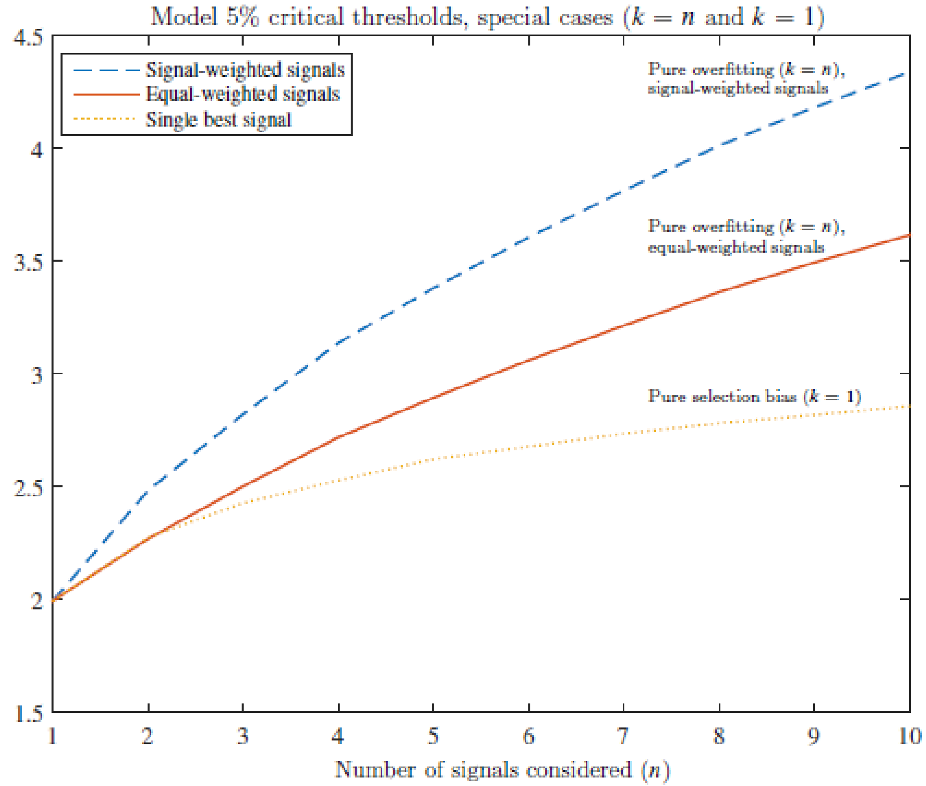


Fig. 1. Five percent critical t-statistics for the active return on best 1-of-n and best n-of-n strategies. The bottom, dotted line shows 5% critical thresholds for the pure selection bias case, when the investigator presents the strongest result from a set of n random strategies, where $n \in \{1, 2, ..., 10\}$. The middle, solid line and the top, dashed line, shows 5% critical thresholds when there is pure overfitting bias. In these cases stocks are selected by combining all n random signals, but the underlying signals are individually signed so that each predicts positive in-sample active returns. In the top, dashed line signals are signal-weighted, while in the middle, solid line signals are equal-weighted. Critical values come from generating 100,000 sets of n randomly generated signals. Returns are signal-weighted, with stocks held in proportion to the signal used for strategy construction, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

PennState
Smeal College of Business

Novy-Marx (2016)

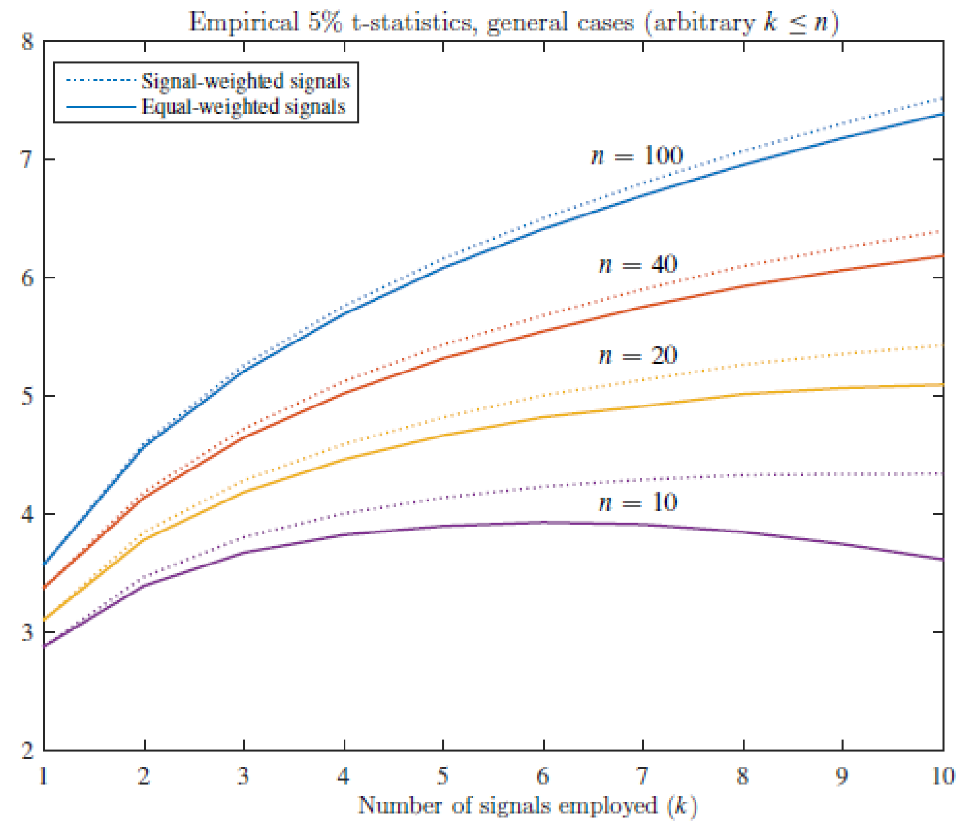


Fig. 2. Five percent critical t-statistics for best k-of-n strategies. The figure shows 5% critical thresholds for strategies selected using a signal constructed by combining the best $k = 1, 2, ..., 10$ performing signals, when the investigator considered $n \in \{10, 20, 40, 100\}$ candidate signals. Solid lines show the cases when the composite signal is constructed by equal-weighting the k best performing candidate signals, and dotted lines the cases when the composite signal is constructed by signal-weighting the signals. Critical values come from generating 100,000 sets of n randomly generated signals. Returns are signal-weighted, with stocks held in proportion to the signal used for strategy construction, and rebalanced annually at the start of each year. Return data come from CRSP, and the sample covers January 1995 through December 2014.

PennState
Smeal College of Business

Issues with combining characteristics

- Kozak, Nagel, and Santosh (JFE, 2020)
  - Reverse engineer a stochastic discount factor using Bayesian methods
    - Beware!

Table 1. Largest SDF factors (50 anomaly portfolios).

Coefficient estimates and absolute *t*-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the original 50 anomaly portfolios. Panel (b) pre-rotates returns into PC space and shows coefficient estimates corresponding to these PCs. Coefficients are sorted descending on their absolute *t*-statistic values. The sample is daily from November 1973to December 2017.

(a) Raw 50 anomaly portfolios			(b) PCs of 50 anomaly portfolios		
	b	*t*-stat		*b*	*t*-stat
Industry rel. rev. (L.V.)	−0.88	3.53	PC 4	1.01	4.25
Ind. mom-reversals	0.48	1.94	PC 1	−0.54	3.08
Industry rel. reversals	−0.43	1.70	PC 2	−0.56	2.65
Seasonality	0.32	1.29	PC 9	−0.63	2.51
Earnings surprises	0.32	1.29	PC 15	0.32	1.27
Value-profitablity	0.30	1.18	PC 17	−0.30	1.18
Return on market equity	0.30	1.18	PC 6	−0.29	1.18
Investment/Assets	−0.24	0.95	PC 11	−0.19	0.74
Return on equity	0.24	0.95	PC 13	−0.17	0.65
Composite issuance	−0.24	0.95	PC 23	0.15	0.56
Momentum (12m)	0.23	0.91	PC 7	0.14	0.56

PennState
Smeal College of Business

Issues with combining characteristics

- Avramov, Cheng, and Metzker (MS, 2021)
  - Shows that strategies based on deep learning signals derive their profitability from difficult-to-arbitrage stocks and during high limits-to-arbitrage market states.
  - Excluding microcaps, distressed stocks, or episodes of high market volatility significantly reduces profitability.
  - Machine learning-based performance further deteriorates in the presence of reasonable trading costs.

PennState
Smeal College of Business