Taming the Factor Zoo

Guanhao Feng[1] Stefano Giglio[2] Dacheng Xiu[3]

[1]City University of Hong Kong

[2]Yale University

[3]University of Chicago

AFA, Jan 7th 2018

Motivation

- Hundreds of risk factors or "anomalies" in the last 30 years
  - data-snooping (Lo and MacKinlay (1990, RFS)), microcaps (Fama and French (2008, JF)), multiple testing (Harvey et al. (2015, RFS)), publication bias (McLean and Pontiff (2016, JF))

- We conjecture that most of these factors are redundant, due to **omitted variable bias** from the benchmark (i.e. **insufficient controls**).
  - A redundant factor may only be "useful" relative to the Fama-French 3- or 5-factor model.

- Suppose some economic theory/model/intuition/story suggests a "new" factor: is it truly new and useful in pricing the cross-section of assets?

Motivation

- We consider prominent 100 factors introduced in the last 30 years (database from Green et al. (2013, RAS))
- 14 factors on top finance journals in the last 5 years only
- How many are new? And new, **relative to what**?
- What is the right benchmark? FF3 (Market-Size-Value)? Others?

This paper

- How about controlling **all** factors proposed by the literature?
  - Standard analysis (like Fama-MacBeth) is inefficient or even infeasible! – **curse of dimensionality**
- Will need machine learning / model selection to **reduce the dimensionality** of the factor zoo
  - E.g., LASSO (Tibshirani (1996, JRSSB)), etc.
- Have to take into account potential **model-selection mistakes**
  - LASSO may **miss** some true factors! Omitted variable bias occurs again. (Belloni, Chernozhukov, and Hansen (2014, ReStud))

How **not** to check if a factor g_t is useful (I)

- **What about checking if LASSO selects it?**
- Our data (1825 assets, 99 factors, 1980-2016 monthly)
- Randomly draw a bootstrap subsample, regress average returns on factor covariances (univariate betas!) to check if a factor is selected

How **not** to check if a factor g_t is useful (I)

- LASSO can be very unstable; even the market factor is thrown away 25% of the samples; some non-prominent factors are selected.
- LASSO cannot always select the true model when covariates are correlated.
  - This is bad for parameter inference, but not so much for prediction.
- **Lesson**: LASSO cannot be trusted in picking the identities of factors.

How **not** to check if a factor g_t is useful (II)

- **Use LASSO to select the controls for g_t (benchmark)?**
- Do standard inference on g_t using the selected model as controls
- This is **single-selection**
- Major problem: in finite samples, LASSO will miss some factors and produce **omitted variable bias**

The solution: Double-Selection method

- Key idea is developed for treatment effect estimation by Belloni, Chernozhukov, and Hansen (2014, ReStud)

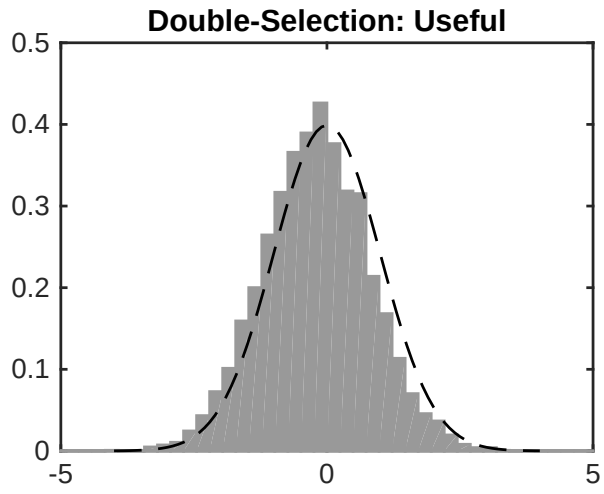
- We adapt it to the two-pass cross-sectional regression:
  - First LASSO: regress average returns $\bar{r}$ onto factor covariances $\widehat{C}_h$.
  - **Second LASSO**: regress the covariance of interest $\widehat{C}_g$ on to $\widehat{C}_h$.
  - Post-Selection: regress $\bar{r}$ onto $\widehat{C}_g$ and selected $\widehat{C}_h[I]$, where I is the union of selected variables in the previous two steps.

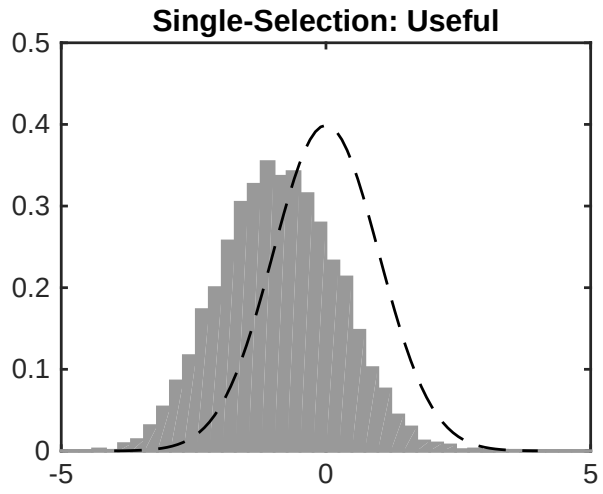
- Key intuition: an omitted variable bias would occur only if the discarded (true) variables in $\widehat{C}_h$ from the first LASSO are highly correlated with $\widehat{C}_g$.

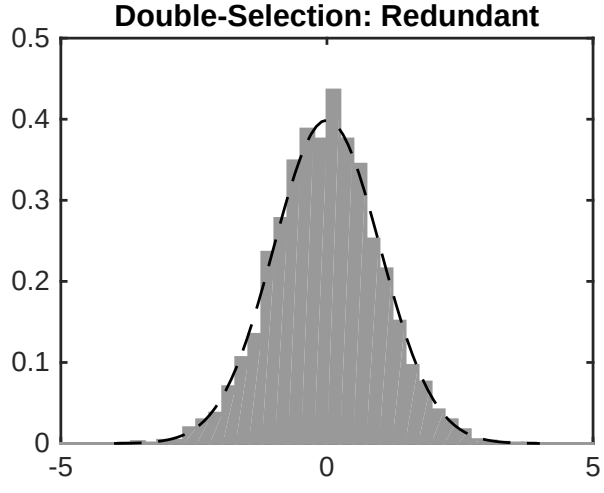
What happens if we don't account for LASSO mistakes

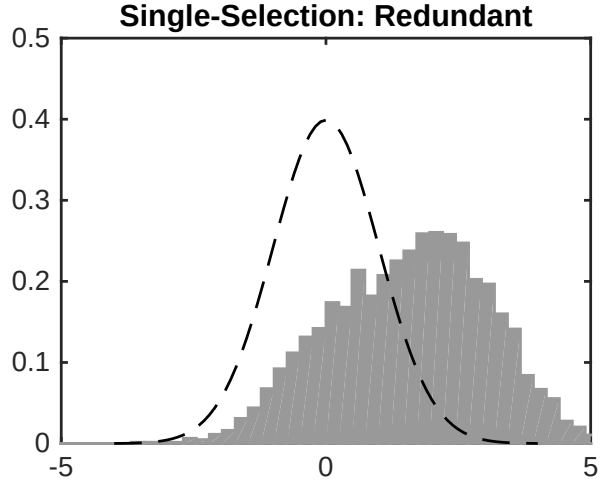
- Simulation exercise
- Take two factors, one **useful**, one **redundant**
- Simulate a world with 100 factors
- An ideal test would recognize that one of the factors is useful, the other is redundant
- Compare:
  - Single-Selection: ignore the potential LASSO mistakes
  - Double-Selection: account for the potential LASSO mistakes

What happens if we don't account for LASSO mistakes









Zoo of Factors: data

- 99 monthly factors covering from Jul. 1980 to Dec. 2016
- Covers all main anomaly categories: momentum, size, value/growth, investment, profitability, intangibles, and trading frictions.
- We also explored adding 197 nontradable factors (squared factors + interactions with SMB)
- We use a total of 1,825 portfolios as test assets
  - Portfolio sorts by all of our characteristics
- We ask whether each new factor helps explain the cross-section relative to the other factors

Are New Factors Useful?

- We first ask whether the recently introduced factors are redundant or outright useless in pricing the panel of returns.
  - Factors: all tradable factors introduced after 2011.
  - Controls: all tradable factors prior to 2011

Are New Factors Useful?

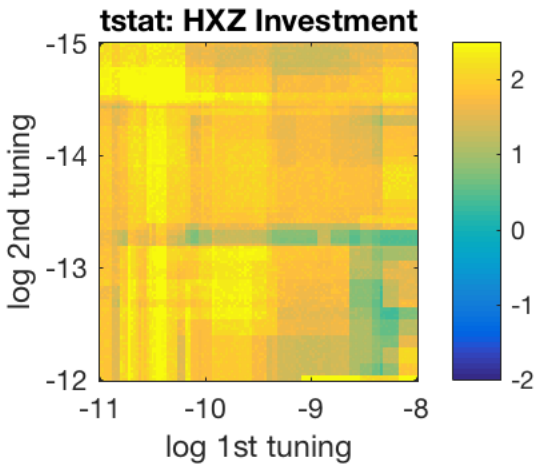
id	Factor Description	λ_s (bp)	tstat
84	Maximum Return	74	0.60
85	Percent Accruals	-38	-1.47
86	Cash Holdings	59	0.73
87	HML Devil	5	0.07
88	Gross profitability	-55	-1.03
89	Organizational Capital	-67	-1.52
90	Betting Against Beta	16	0.52
91	Quality Minus Junk	31	0.81
92	Investment (HXZ)	48	2.10**
93	Profitability (HXZ)	80	2.43**
94	Employee Growth	10	0.26
95	Profitability (FF)	94	2.56***
96	Investment (FF)	56	2.02**
98	Intermediary Investment	-96	-1.17
99	Convertible Debt	14	1.14

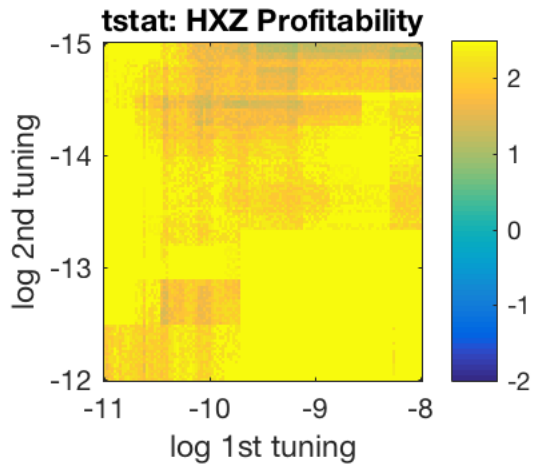
Evaluating Factors Recursively

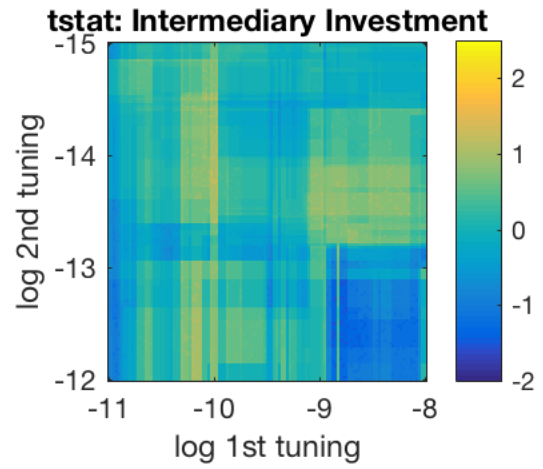
- One of the motivations for using our methodology is that it can help distinguish useful from redundant factors as they are introduced in the literature.
- Over time, can this limit the proliferation of factors?
- We test new factors as they are introduced against previously-existing factors

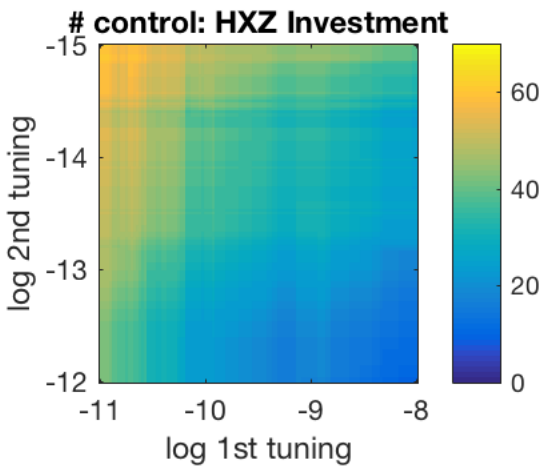
	(1)	(2)	(3)						
Year	# Assets	# Controls	New factors (IDs)						
1994	450	22	23	24					
1995	500	24	25	26	27				
1996	500	27	28	29					
1997	550	29	30						
1998	575	30	31	32	33	34	35	36	
1999	725	36	37	38					
2000	750	38	39	40	41	42			
2001	800	42	43	44	45				
2002	825	45	46	47	48				
2003	875	48	49	50	51				
2004	925	51	52	53	54	55	56		
2005	1025	56	57	58	59	60	61		
2006	1100	61	62	63	64	65	66	67	68
2007	1275	68	69	70	71				
2008	1350	71	72	73	74	75			
2009	1450	75	76	77	78	79			
2010	1525	79	80	81	82	83			
2011	1625	83	84	85					
2012	1675	85	86						
2013	1700	86	87	88	89				
2014	1750	89	90	91	92	93	94		
2015	1825	94	95	96					
2016	1825	96	98	99					

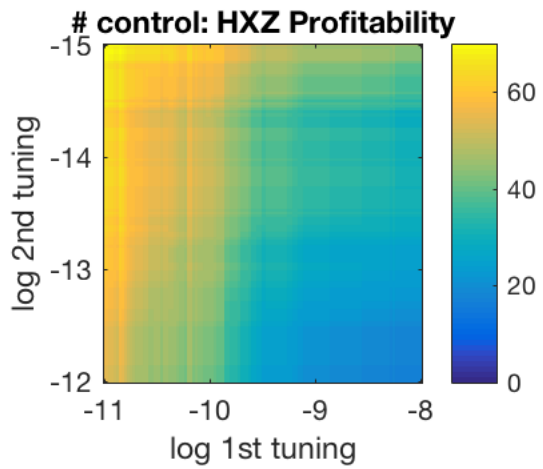
Robustness

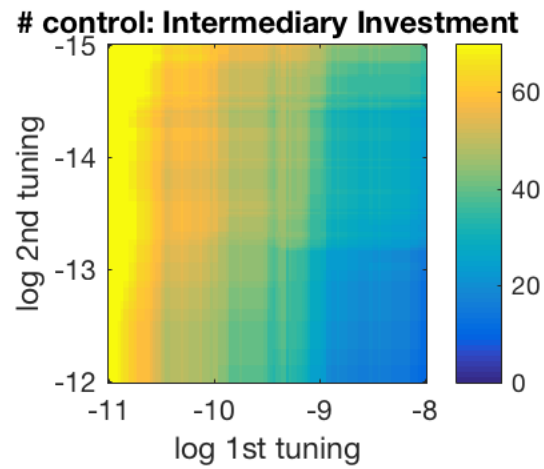

tstat: HXZ Investment
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
2
1
0
-1
-2


tstat: HXZ Profitability
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
2
1
0
-1
-2


tstat: Intermediary Investment
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
2
1
0
-1
-2


control: HXZ Investment
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
60
40
20
0


control: HXZ Profitability
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
60
40
20
0


control: Intermediary Investment
log 2nd tuning
log 1st tuning
-15
-14
-13
-12
-11
-10
-9
-8
60
40
20
0

Robustness Checks

		(1) Bivariate 5 × 5		(2) Sequential 5 × 5		(3) Pre-1994		(4) Elastic Net	
id	Factor Description	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)
84	Maximum Return	74	0.60	35	0.29	-149	-1.12	-24	-0.19
85	Percent Accruals	-38	-1.47	-40	-1.57	-36	-1.41	-20	-0.78
86	Cash Holdings	59	0.73	-12	-0.20	83	1.20	-142	-1.73
87	HML Devil	5	0.07	-35	-0.48	26	0.37	7	0.10
88	Gross profitability	-55	-1.03	-23	-0.47	-26	-0.47	-44	-0.83
89	Organizational Capital	-67	-1.52	-84	-1.94*	-92	-2.16**	-45	-1.02
90	Betting Against Beta	16	0.52	-20	-0.68	-60	-1.98**	-25	-0.87
91	Quality Minus Junk	31	0.81	31	0.84	76	1.97**	88	2.32**
92	Investment (HXZ)	48	2.10**	36	1.66*	65	2.96***	42	1.91*
93	Profitability (HXZ)	80	2.43**	78	2.45**	114	3.38***	78	2.51**
94	Employee Growth	10	0.26	45	1.19	-7	-0.20	8	0.23
95	Profitability (FF)	94	2.56***	89	2.48**	64	1.76*	116	3.12***
96	Investment (FF)	56	2.02**	41	1.55	75	2.73***	40	1.45
98	Intermediary Investment	-96	-1.17	-74	-0.91	-75	-0.92	-32	-0.40
99	Convertible Debt	14	1.14	37	2.40**	8	0.53	3	0.20

Conclusion

1. **Omitted variable bias** (wrong benchmark) may contribute to the proliferation of factors.
2. Be mindful of what machine learning can and cannot do (e.g. which factors are in the SDF?)
3. Machine learning and model selection can make important mistakes, but these can be accounted for for some purposes!
4. Strongest new factors: profitability and investments
5. Many others are redundant