Taming the factor zoo: A Test of New Factors

Guanhao Feng, Stefano Giglioz, Dacheng Xiu

The Journal of Finance, 2020.

雷印如 2020/12/12

Outline

- Introduction
- Research design
- Empirical result
- Robust Test
- Conclusion

Motivation

- A fundamental task facing the asset pricing field today is to bring more discipline to the proliferation of factors.
- When facing with potentially hundreds of factors, standard statistical methods become infeasible because of the curse of dimensionality.
- **How to judge whether a new factor adds explanatory power for asset pricing, relative to the hundreds of factors the literature has so far produced?**

Contribution

1. Marry the double-selection LASSO method with two-pass regressions to selects the factors in explaining the cross section of expected returns or in mitigating the omitted variable bias problem.
2. We focus on the evaluation of a new factor, rather than testing or estimating an entire reduced-form asset pricing model
3. Our procedure also helps alleviate the concern of data-snooping, which is another form of multiple testing

Methodology

Omitted variable:
$bias(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = \beta_2\hat{\delta}_1$

- Model Setup -- linear specification for the SDF

$$m_t := \gamma_0^{-1} - \gamma_0^{-1}\lambda_v^T v_t := \gamma_0^{-1}(1 - \lambda_g^T g_t - \lambda_h^T h_t)$$

- γ_0 is the zero-beta rate, we refer to λ_g and λ_h as the SDF loadings of the factors g_t and h_t. By risk return:

$$E(r_t) = \iota_n\gamma_0 + \beta_g\gamma_g + \beta_h\gamma_h$$

- β_g and β_h are the factor exposures and γ_g and γ_h are the risk premia
- SDF loadings λ and risk premia γ are directly related through the covariance matrix of the factors.

$$g_t = \eta h_t + zt, \qquad where\ Cov(z_t, h_t) = 0$$

Challenges in High-Dimensions

- Standard methodologies in the curse of dimensionality.
  - Ad hoc solutions to cherry-pick a handful of control factors, but selecting an incorrect model is problematic as it can lead to omitted variable bias.
- Sparsity assumption: a natural economic interpretation
  - Sparse models are easier to interpret and to link to economic theories, since no procedure does well in dense problems.
- To leverage sparsity, we use LASSO estimator.

Challenges in High-Dimensions

- LASSO and Model Selection Mistakes
  - If important factors are mistakenly excluded from the control, the SDF loadings will be affected by an omitted variable bias.
  - The omitted variable bias is exacerbated if the risk exposures to the omitted factors are highly correlated with the exposures to g_t
- Two-Pass Regression with Double-Selection LASSO
  - The first selection searches for factors in h_t in standard LASSO
  - A second selection is then added to search for factors in h_t that are potentially missed.

Two-pass variable selection

1. Run a cross-sectional LASSO regression of average returns on sample covariances between factors in h_t and returns

$$\min_{\gamma,\lambda}\left\{n^{-1}\left\|\bar{r}-\iota_n\gamma-\widehat{C}_h\lambda\right\|^2+\tau_0 n^{-1}\|\lambda\|_1\right\}$$

- Denote by $\{\hat{I}_1\}$ as the set of indices corresponding to the selected factors in this step.

2. For each factor $j\ in\ g_t\ (with\ j\ =\ 1,\cdots,d)$, run a cross-sectional LASSO regression of $\hat{C}_{g,\cdot,j}$ on $\hat{C}_h$

$$\min_{\xi_j,\chi_{j,\cdot}}\left\{n^{-1}\left\|\left(\widehat{C}_{g,\cdot,j}-\iota_n\xi_j-\widehat{C}_h\chi_{j,\cdot}^{\intercal}\right)\right\|^2+\tau_j n^{-1}\|\chi_{j,\cdot}^{\intercal}\|_1\right\}$$

- Denote by $\{\hat{I}_{2,j}\}$ the set of indices corresponding to the selected factors in the j_{th} regression, where $\hat{I}_{2,j}=\mathrm{U}_{j=1}^{d}\hat{I}_{2,j}$

Post-selection estimation

- Run an OLS cross-sectional regression using covariances between the selected factors from both steps and returns

$$(\widehat{\gamma}_0, \widehat{\lambda}_g, \widehat{\lambda}_h) = \arg \min_{\gamma_0, \lambda_g, \lambda_h} \left\{ \left\| \bar{r} - \iota_n \gamma_0 - \widehat{C}_g \lambda_g - \widehat{C}_h \lambda_h \right\|^2 : \right.$$
$$\left. \lambda_{h,j} = 0, \quad \forall j \notin \widehat{I} = \widehat{I}_1 \bigcup \widehat{I}_2 \right\}$$

- We refer to this procedure as the DS approach, as opposed to the single selection (SS) approach that involves only (1.a) and (2).

Empirical Analysis

A. Data

1. We download all workhorse factors in the U.S. equity market from Ken French's data library.
2. Published factors directly from the authors' websites
3. We also include factors from the AQR data library

- Our factor library contains 150 risk factors at the monthly frequency for the period from July 1976 to December 2017.
- Construct 135 long-short value-weighted portfolios as factor proxies

Empirical Analysis

A. Test Portfolios

- We use a total of 750 portfolios as test assets.

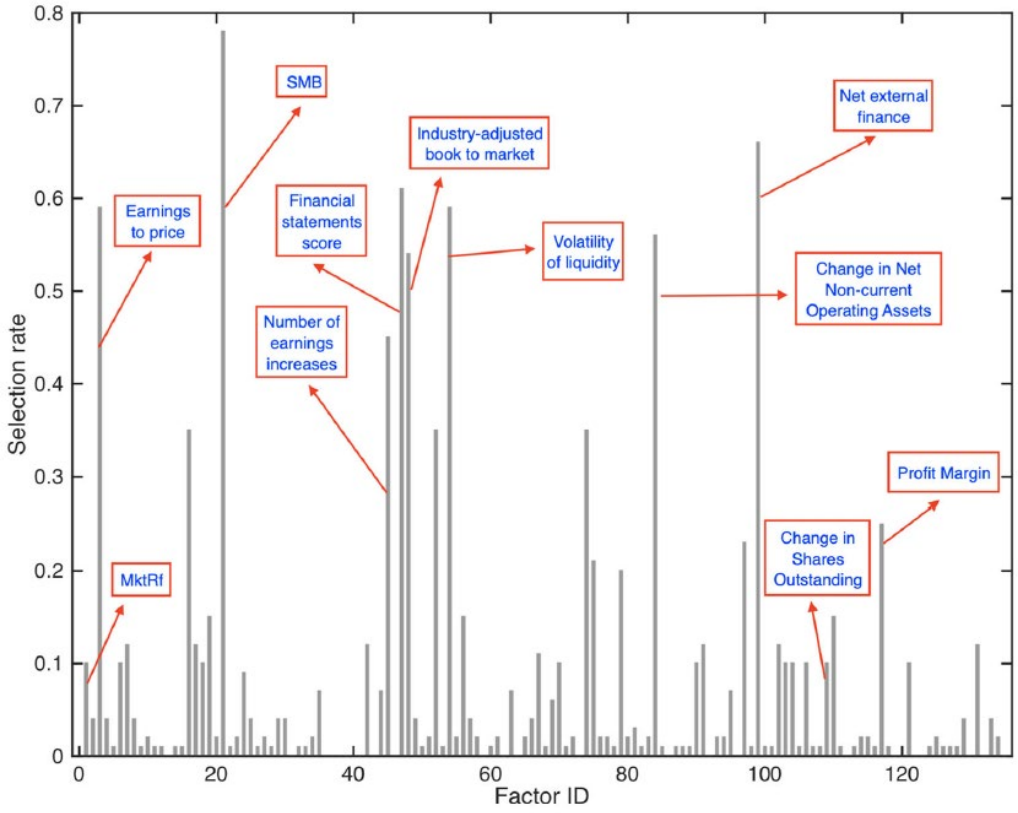
	bm	profitability	investment	reversal	momentum
size	3*2	3*2	3*2	3*2	3*2

- We add to these 36 portfolios 714 additional portfolios obtained from our factor zoo that cover additional characteristics.
- Six groups contains a sufficient number of stocks at least 10.(119 sets of 3 × 2 bivariate-sorted portfolios).
- we apply our methodology to factors proposed over the last five years (2012 to 2016)

Empirical Analysis

B. Evaluating New Factors

- We run 200 different 10-fold CV exercises using 200 different randomization seeds to pin down the two tuning parameters



Empirical Analysis

B. Evaluating New Factors

- If LASSO were able to perfectly select the true model, we should find that a small number of factors are selected 100% of the time, while the remaining factors are selected 0% of the time.
- Instead, Figure 1 shows that LASSO clearly has difficulty in pinning down which factors are the correct ones.
- This exercise thus cautions against using simple LASSO to decide whether a factor should be included in the SDF.

Empirical Analysis

B. Double-Selection Estimator

		(1) DS		(2) SS		(3) FF3		(4) No Selection		(5) Avg. Ret.	
id	Factor Description	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (SS)	λ_s (bp)	tstat (OLS)	λ_s (bp)	tstat (OLS)	avg.ret. (bp)	tstat
136	Cash holdings	−34	−0.42	15	0.17	10	0.54	−18	−0.16	13	0.98
137	HML Devil	54	1.04	−13	−0.25	−100	−2.46**	68	0.84	23	1.46
138	Gross profitability	20	0.48	3	0.06	23	2.00**	13	0.26	15	1.45
139	Organizational Capital	28	0.92	−1	−0.03	20	1.91*	16	0.41	21	2.05**
140	Betting Against Beta	35	1.45	38	1.50	36	2.25**	49	1.49	91	5.98***
141	Quality Minus Junk	73	2.03**	4	0.11	39	3.10***	50	1.04	43	3.87***
142	Employee growth	43	1.36	−4	−0.12	−12	−0.89	18	0.37	8	0.83
143	Growth in advertising	−12	−1.18	0	0.03	12	1.32	−2	−0.13	7	0.84
144	Book Asset Liquidity	40	1.07	5	0.12	20	1.59	20	0.42	9	0.79
145	RMW	160	4.45***	15	0.41	20	1.80*	74	1.48	34	3.21***
146	CMA	38	1.10	0	0.01	3	0.28	7	0.14	26	3.02***
147	HXZ IA	51	2.11**	5	0.21	21	1.94*	40	1.08	34	4.17***
148	HXZ ROE	77	3.37***	23	0.83	33	2.92***	104	2.87***	57	4.99***
149	Intermediary Risk Factor	112	2.21**	60	1.19	4	0.08	22	0.32		
150	Convertible debt	−15	−1.36	−39	−3.22***	26	3.32***	17	1.01	11	1.70*

Empirical Analysis

C. Evaluating Factors Recursively

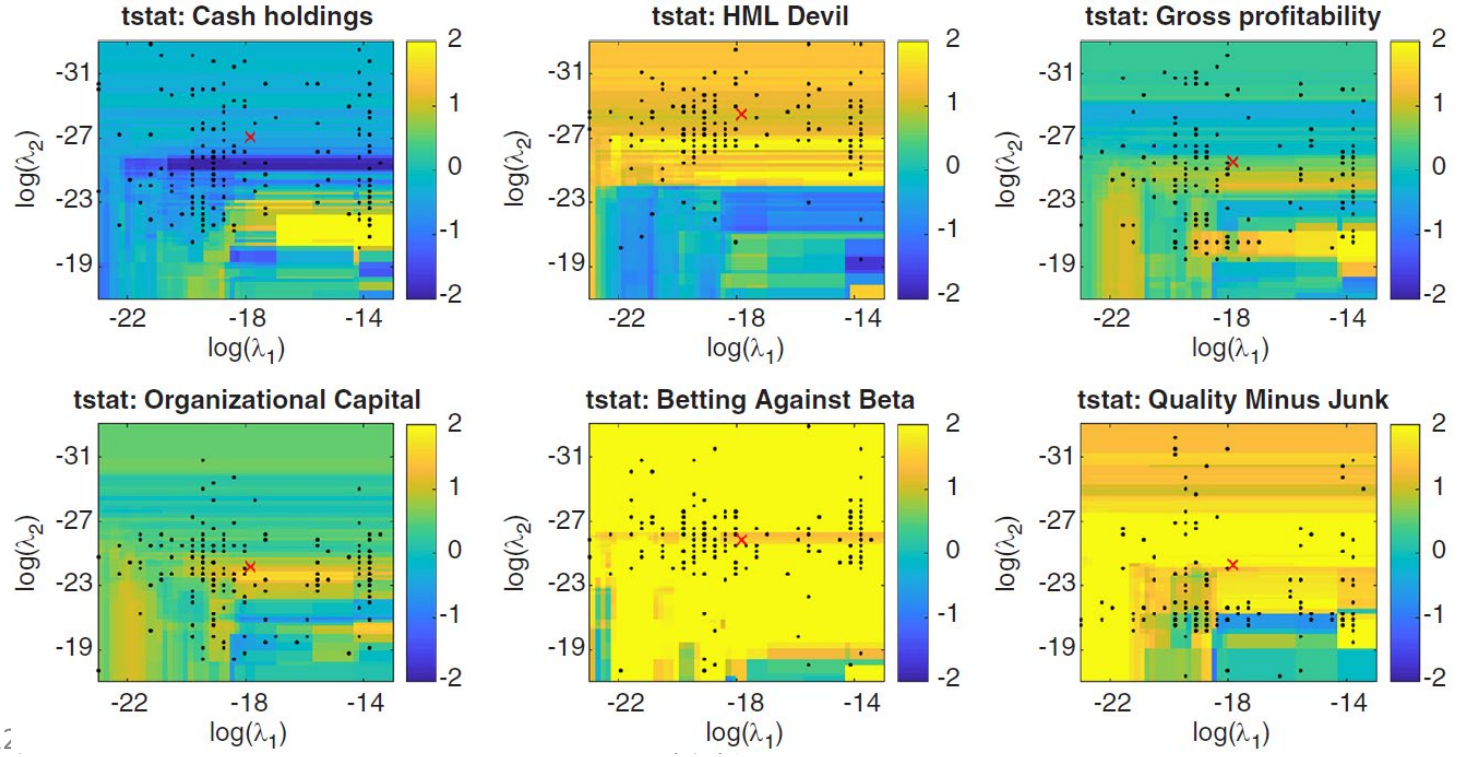
Year	(1) # Assets	(2) # Controls	(3) New factors (IDs)											
1994	138	25	26	27										
1995	150	27	28	29	30									
1996	150	30	31	32	33									
1997	168	33	34											
1998	174	34	35	36	37	38	39	40	41	42	43	44		
1999	228	44	45	46										
2000	234	46	47	48	49	50	51							
2001	252	51	52	53	54	55	56	57	58					
2002	294	58	59	60	61									
2003	312	61	62	63	64	65	66							
2004	336	66	67	68	69	70	71	72	73	74				
2005	372	74	75	76	77	78	79	80	81	82	83	84	85	86
			87	88	89	90								
2006	456	90	91	92	93	94	95	96	97	98	99	100	101	102
2007	516	102	103	104	105	106	107	108						
2008	552	108	109	110	111	112	113	114	115	116	117	118	119	120
2009	618	120	121	122	123	124								
2010	636	124	125	126	127	128	129							
2011	666	129	130	131	132	133	134	135						
2012	702	135	136											
2013	708	136	137	138	139									
2014	720	139	140	141	142	143	144							
2015	738	144	145	146	147	148								
2016	750	148	149	150										

- Only 17 factors would have been considered useful, with a large majority identified as redundant or useless

Robustness

A. Robustness to the Choice of Tuning Parameters

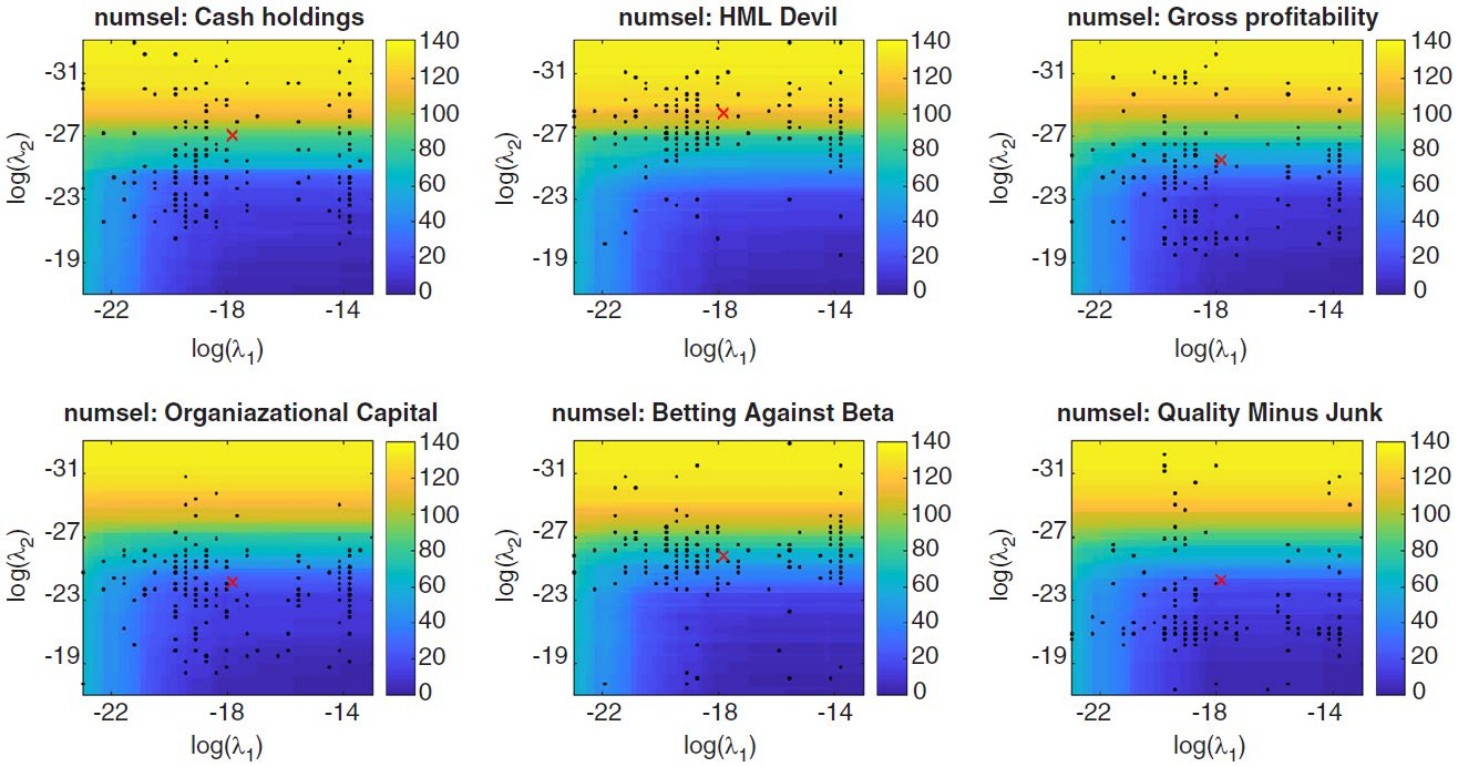
- We then look at how each λ_g's t-statistic varies across choices of the tuning parameters.



Robustness

A. Robustness to the Choice of Tuning Parameters

- The heat maps display numbers of controls selected for each factor.



Robustness

B. Robustness to Test Assets and Regularization Method

- Sorting the test assets into 5 × 5 instead of 3 × 2 portfolios and use different dimension-reduction procedures.

		(1) Bivariate 3 × 2		(2) Bivariate 5 × 5		(3) 202 Portfolios		(4) Elastic Net		(5) PCA		(6) Stepwise	
id	Factor Description	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)	λ_s (bp)	tstat (DS)
136	Cash holdings	−34	−0.42	34	0.40	131	0.89	−13	−0.14	−65	−0.62	−73	−0.87
137	HML Devil	54	1.04	15	0.29	56	0.57	62	1.23	−27	−0.51	49	1.01
138	Gross profitability	20	0.48	28	0.66	88	1.42	−11	−0.26	16	0.35	16	0.47
139	Organizational Capital	28	0.92	23	0.75	6	0.16	12	0.38	21	0.57	0	0.01
140	Betting Against Beta	35	1.45	43	1.94*	31	1.03	28	1.12	59	2.56***	62	2.57***
141	Quality Minus Junk	73	2.03**	58	1.67	123	2.45**	74	2.13**	71	1.89*	40	1.16
142	Employee growth	43	1.36	12	0.34	54	1.34	51	1.49	−4	−0.09	33	0.98
143	Growth in advertising	−12	−1.18	6	0.57	17	1.30	9	0.74	−6	−0.57	3	0.27
144	Book Asset Liquidity	40	1.07	−24	−0.61	37	0.77	26	0.68	24	0.63	33	1.00
145	RMW	160	4.45***	104	3.13***	112	1.98**	125	3.43***	88	2.11**	96	2.71***
146	CMA	38	1.10	19	0.59	33	0.52	32	0.85	18	0.44	23	0.67
147	HXZ IA	51	2.11**	44	1.87*	−45	−1.42	69	2.77***	36	1.31	49	1.92*
148	HXZ ROE	77	3.37***	72	2.62***	116	2.22***	103	3.85***	41	1.46	101	3.87***
149	Intermediary Risk Factor	112	2.21**	38	0.73	−16	−0.33	−16	−0.33	103	1.92*	−10	−0.17
150	Convertible debt	−15	−1.36	−6	−0.56	68	5.13***	−12	−1.08	−9	−0.88	0	−0.02

Conclusion

- We propose a regularized two-pass cross-sectional regression approach to systematically select the best control model out of the large set of factors, taking into account model selection mistakes.
- We apply our methodology to a large set of factors proposed in the literature over the last 30 years
    1. Several newly proposed factors are useful in explaining asset prices
    2. The SDF loadings' estimates for several factors are robust to changes in the tuning parameters.
    3. Studying the marginal contribution of new factors is a conservative and productive way to screen new factors and bring discipline to the "zoo of factors."

Conclusion

- 将计量中的问题引入到了机器学习和资产定价结合中
- 本文的两步LASSO是遗漏变量的解决方法之一，核心是利用了遗漏变量和控制变量的相关性；Giglioz et al.(2020)使用PCA旋转不变性来解决遗漏变量的问题，利用了工具变量的核心思想