Taming the Factor Zoo：A Test of New Factors

JF 2020(03)

汇报人：卫夏利

2020年12月9日

Taming the Factor Zoo：A Test of New Factors

JF 2020(03)

- Manuscript **received**: 2017.9.6
- Manuscript **accept**: 2019.6.20
- Accept online: 2020.1.24
- Record online: 2020.2.27
- **Issue** online: 2020.5.20

Authors

GUANHAO FENG

Employment

City University of Hong Kong College of Business

Assistant Professor of Statistics

Education

- Ph.D. - Business Administration (University of Chicago)
- M.B.A. - Economics and Finance (University of Chicago)
- B.S. - Mathematics (Penn State University)
- B.S. - Economics (Penn State University)

Research Interests

Financial Econometrics; Empirical Asset Pricing;

Machine Learning; Quantitative Finance

Publications:

1. FENG, Guanhao; GIGLIO, Stefano; XIU, Dacheng. **Taming the Factor Zoo: A Test of New Factors.** June 2020; In: The Journal of Finance. Vol. 75, No. 3, pp. 1327-1370.
2. Charoenwong, Ben; Feng, Guanhao. **Does Higher-Frequency Data Always Help to Predict Longer-Horizon Volatility?.** June 2017; In: Journal of Risk. Vol. 19, No. 5, pp. 55-75.
3. Feng, Guanhao; Polson, Nicholas; Xu, Jianeng. **The Market for English Premier League (EPL) Odds.** December 2016; In: Journal of Quantitative Analysis in Sports. Vol. 12, No. 4, pp. 167-178.

Authors

Stefano Giglio

☐ Employment

- Yale School of Management, Professor of Finance
- NBER, and CEPR

*Before joining Yale, Professor Giglio was an Associate Professor of Finance at the University of Chicago Booth School of Business.

☐ Education

- PhD, Harvard University, 2011
- M.Sc., Bocconi University, 2006
- BA, Bocconi University, 2004

☐ Research Interests

asset pricing, macroeconomics, real estate;

hedging macroeconomic risks using different financial instruments: crash risk, uncertainty risk, and climate risk

Publications:

1. Ian Dew-Becker, Stefano Giglio. Cross-Sectional Uncertainty and the Business Cycle: Evidence from 40 Years of Options Data[J]. **NBER Working Paper** No. 27864; Issued in September 2020.
2. Giglio S, Maggiori M , Stroebel J , et al. Inside the Mind of a Stock Market Crash[J]. **CESifo Working Paper Series**, 2020.
3. Stefano G , Yuan L , **Dacheng X**. Thousands of Alpha Tests[J]. **The Review of Financial Studies**, Published: 24 September 2020.
4. Ian Dew-Becker, Stefano Giglio, **Asset Pricing in the Frequency Domain: Theory and Empirics,** The Review of Financial Studies, Volume 29, Issue 8, August 2016, Pages 2029–2068.

Authors

DACHENG XIU*

☐ Appointment

University of Chicago, Booth School of Business, Professor of Econometrics and Statistics.

- Tsinghua University, PBC School of Finance, Special Term Professor, June 2019 - May 2022.
- Shanghai Jiao Tong University, SAIF, Special Term Professor, July 2019 - June 2021.

☐ Education

- Princeton University, Ph.D. Applied Mathematics, May 2011
- Princeton University, M.A. Applied Mathematics, June 2008
- University of Science and Technology of China, B.S. Mathematics, June 2006

☐ Research Interests

Financial Econometrics, Empirical Asset Pricing, Machine Learning in Finance, High-Dimensional Statistics, Quantitative Finance

Publications:

1. "Knowing Factors or Factor Loadings, or Neither? Evaluating Estimators of Large Covariance Matrices with Noisy and Asynchronous Data," with Chaoxing Dai and Kun Lu, ***Journal of Econometrics***, 208 (2019), 43-79.
2. "Efficient Estimation of Integrated Volatility Functionals via Multiscale Jackknife," with Jia Li and Yunxiao Liu, ***Annals of Statistics***, Vol. 47, No. 1 (2019), 156-176.
3. "Principal Component Analysis of High Frequency Data," with Yacine Aït-Sahalia, ***Journal of the American Statistical Association*** Vol. 114, No. 525, (2019), 287-303.

Abstract

- We **propose a model selection method** to systematically evaluate the contribution to asset pricing of any new factor, above and beyond what a high-dimensional set of existing factors explains.
- **Our methodology accounts for model selection mistakes** that produce a bias due to omitted variables, unlike standard approaches that assume perfect variable selection.
- We **apply our procedure to a set of factors** recently discovered in the literature. While most of these new factors are shown to be redundant relative to the existing factors, a few have statistically significant explanatory power beyond the hundreds of factors proposed in the past.

Introduction

The search for factors that explain the cross section of expected stock returns has produced **hundreds of potential candidates.**

A fundamental task facing the asset pricing field today is to bring more discipline to the proliferation of factors.

In particular, **a question** that remains open is: **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?

Introduction

This paper provides a **framework** for systematically evaluating the contribution of individual factors relative to existing factors as well as for conducting appropriate statistical inference in this high-dimensional setting.

More specifically, we provide a **methodology** for estimating and testing the marginal importance of <u>any factor</u> g_t in pricing the cross section of expected returns *beyond* what can be explained by a high-dimensional <u>set of potential factors</u> h_t.

Introduction

testing whether g_t is useful in explaining asset prices while controlling for the factors in h_t

◆ h_t **consists of a small number of factors**

estimating the loadings of the stochastic discount factor(SDF) on g_t and h_t **and testing** whether the loading of g_t is different from zero

- whether g_t is useful for pricing the cross section
- how shocks to g_t affect marginal utility(a direct economic interpretation)

◆ h_t **consists of potentially hundreds of factors**

- standard statistical methods to estimate and test the SDF loadings become infeasible
- result in poor estimates and invalid inference because of the curse of dimensionality

Introduction

◆ h_t consists of potentially hundreds of factors

The curse of dimensionality

↓

Reduce the dimensionality

↓

Variable selection techniques

↓

Oracle Property

Oracle Property:

An asymptotic property that guarantees that under certain assumptions, as the sample size goes to infinity, the procedures will eventually recover the true model.

↓

In practice(finite-sample), the oracle property never holds.

↓

- Any omission of relevant factors due to model selection errors
- distorts the asymptotic distribution of the estimator
- leading to incorrect inference on the significance of the loading(even the sign)

Introduction

double-selection (DS) estimation procedure:

Combines **cross-sectional asset pricing regressions** with the **DS LASSO** of Belloni, Chernozhukov, and Hansen (2014b)

- **(1)** starts by using a two-step selection method to select "control" factors from h_t (apply some dimension-reduction method (LASSO, Elastic Net, PCA, etc.))
  - **(1.a)** first-stage LASSO

    A first set of factors is selected from h_t based on their pricing ability for the cross section of returns.
  - **(1.b)** second LASSO ⟸ **The key contribution** ★

    the second step adds factors whose covariances with returns are highly correlated in the cross section with the covariance between returns and g_t.
- **(2)** then estimates the SDF loading of g_t from cross-sectional regressions that include g_t and the selected factors from h_t.

The key contribution

machine learning methods

↓

better *prediction*

↓

minimize out-of-sample prediction error

↓

Certain variables may exclude

(contribution to prediction<the cost of inclusion)

have small in-sample SDF loadings

(contribute little to pricing assets in the cross section)

whose covariance with returns(risk exposures) is **highly cross-sectionally correlated** with exposures to g_t

The key contribution of our paper is to show that despite the mistakes that LASSO inevitably makes in selecting the model, correct inference *can* be made about the contribution to asset pricing of a factor g_t.

Relation to the existing literature

- **Kozak, Nagel, and Santosh (2018)(first step)**

take a large set of factors (h_t), apply some dimension-reduction method, and interpret the resulting low-dimensional model as the SDF

- **Giglio and Xiu (2016)**

show how to make inference on risk premia in the presence of omitted factors (Importantly, only SDF loadings addressed in this paper can speak to the ability of factors to explain asset prices)

- **Belloni, Chernozhukov, and Hansen (2014b)**

the double-selection LASSO method(originally designed for linear treatment effect models)

- **Barillas and Shanken (2018) and Fama and French (2018))**

evaluate by estimating and testing the alpha of a regression of the new factor on existing factors

Content

I. Methodology

A. Model Setup

- We start from a linear specification for the SDF,

$$m_t := \gamma_0^{-1} - \gamma_0^{-1}\lambda_v^\top v_t := \gamma_0^{-1}(1 - \boxed{\lambda_g^\top} g_t - \boxed{\lambda_h^\top} h_t), \tag{1}$$

SDF loadings of the factors g_t *SDF loadings* of the factors h_t

γ_0 : the zero-beta rate
g_t : a $d \times 1$ vector of factors to be tested
h_t : a $p \times 1$ vector of potentially confounding factors

- We observe an $n \times 1$ vector of test asset returns, r_t. Because of (1), expected returns satisfy

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + C_v \lambda_v = \iota_n \gamma_0 + C_g \lambda_g + C_h \lambda_h, \tag{2}$$

where ι_n is an $n \times 1$ vector of 1s, $C_a = \mathrm{Cov}(r_t, a_t)$, for $a = g$, h, or v.

Equation (2) represents expected returns in terms of (univariate) covariances with the factors, multiplied by λ_g and λ_h.

A. Model Setup

- Furthermore, we assume that the dynamics of r_t follow a standard linear factor model,

$$r_t = \mathrm{E}(r_t) + \beta_g g_t + \beta_h h_t + u_t, \tag{3}$$

where β_g and β_h are $n \times d$ and $n \times p$ factor loading matrices and u_t is an $n \times 1$ vector of idiosyncratic components with $\mathrm{E}(u_t) = 0$ and $\mathrm{Cov}(u_t, v_t) = 0$.

- An equivalent representation of expected returns can be obtained in terms of multivariate betas,

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + \beta_g \gamma_g + \beta_h \gamma_h, \tag{4}$$

where β_g and β_h are the factor exposures (i.e., multivariate betas) and γ_g and γ_h are the *risk premia* of the factors.

A. Model Setup

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + C_g \lambda_g + C_h \lambda_h, \tag{2}$$

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + \beta_g \gamma_g + \beta_h \gamma_h, \tag{4}$$

SDF loadings λ and risk premia γ are directly related through the covariance matrix of the factors, but they differ substantially in their interpretation.

The risk premium of a factor tells us whether investors are willing to pay to hedge a certain risk factor, but it does not tell us whether that factor is useful in pricing the cross section of returns.

As discussed extensively in Cochrane (2009), to understand whether a factor is useful in pricing the cross section of assets, we should look at its SDF loading instead of its risk premium.

A. Model Setup

- Because the link between SDF loadings and risk premia depends on the covariances among factors, write the projection of g_t on h_t as

$$g_t = \eta h_t + z_t, \quad \text{where} \quad \mathrm{Cov}(z_t, h_t) = 0. \tag{5}$$

- For the estimation of λ_g, characterize the cross-sectional dependence between C_g and C_h. So we write the cross-sectional projection of C_g onto C_h as

$$C_g = \iota_n \xi^\intercal + C_h \chi^\intercal + C_e, \tag{6}$$

where ξ is a $d \times 1$ vector, χ is a $d \times p$ matrix, and C_e is an $n \times d$ matrix of cross-sectional regression residuals.

B. Challenges with Standard Two-Pass Methods in High-Dimensional Settings

Two-Pass Methods(Jensen, Black, Scholes (1972) and Fama, MacBeth (1973))

The procedure involves **two steps**:

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + \beta_g \gamma_g + \beta_h \gamma_h,$$

I. an asset-by-asset time-series regression that yields estimates of the individual factor loadings βs

II. a cross-sectional regression of expected returns on the estimated factor loadings that yields estimates of the risk *premia* γ .

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + C_g \lambda_g + C_h \lambda_h,$$

- an asset-by-asset time-series regression that yields estimates of the covariances between returns and factors
- a cross-sectional regression of expected returns on the estimated the covariances between returns and factors that yields estimates of the *SDF loadings* of the factors λ.

B. Challenges with Standard Two-Pass Methods in High-Dimensional Settings

Challenges (Two-Pass Methods)

- In a low-dimensional setting, the method above should work smoothly
- hundreds of factors, the standard cross-sectional regression with all factor covariances included is at best highly inefficient.
- when $p > n$, standard Fama-MacBeth approach becomes infeasible.

Existing literature employs ad hoc solutions:

in testing for the contribution of a new factor, it is common to

- cherry-pick a handful of control factors, such as the prominent Fama-French three factors
- effectively imposing an assumption that the selected model is the true one and is not missing any additional factors.
- However, this assumption is clearly unrealistic.

C. Sparsity

- **impose a sparsity assumption in our setting**

  a relatively small number of factors exist in h_t, whose linear combinations along with g_t nest the SDF m_t
- **Does sparsity make sense in asset pricing?**

  Adopted the concept of sparsity without always explicitly acknowledging it
- **Compare with PCA**

  sparse models are easier to interpret and to link to economic theories
- **one should "bet on sparsity"**

  since no procedure does well in dense problems. (sparse versus dense)

  not means true model should always involve only a very small number of factors in absolute terms, say three or five. More nonzero coefficients can be identified given better conditions (e.g., larger sample size).

D. LASSO and Model Selection Mistakes

- **LASSO estimator**

incorporates into the least-squares optimization a penalty function on the L1 norm of parameters leads to an estimator that has many zero coefficients in the parameter vector.

- **"Post-LASSO" estimator**(Belloni, Chernozhukov (2013))

The Post-LASSO estimator runs LASSO as a model selector and then refits the least-squares problem without penalty, using only those variables that have nonzero coefficients in the first step.

Model Selection Mistakes

machine learning methods

↓

better *prediction*

↓

minimize out-of-sample prediction error

↓

Certain variables **may exclude**

(contribution to prediction<the cost of inclusion)

have small in-sample SDF loadings
(contribute little to pricing assets in the cross section)

whose covariance with returns(risk exposures) is **highly cross-sectionally correlated** with exposures to g_t

- In any finite sample, we can never be sure that LASSO or Post-LASSO will select the correct model, just like we cannot claim that the estimated parameter values in a given finite sample are equal to their population counterparts.
- We need to ensure that these factors are included in the set of controls *even if LASSO would suggest excluding them.*
- Note that this issue is not unique to high-dimensional problems, but it is arguably more severe in such a scenario because model selection is inevitable.

E. Two-Pass Regression with Double-Selection LASSO

The regularized two-pass estimation proceeds as follows:

- **(1) Two-pass variable selection**
- **(1.a)** Run a cross-sectional LASSO regression of average returns on sample covariances between factors in h_t and returns,

best explain the cross section of expected returns

$$\min_{\gamma,\lambda}\left\{n^{-1}\left\|\bar{r}-\iota_n\gamma-\widehat{C}_h\lambda\right\|^2+\tau_0 n^{-1}\|\lambda\|_1\right\}, \tag{7}$$

- **(1.b)** For each factor j in g_t (with $j = 1, \cdots, d$), run a cross-sectional LASSO regression of $\hat{C}_{g,\cdot,j}$ (the covariance between returns and the jth factor of g_t) on C_h (the covariance between returns and all factors h_t)

whose exposures are highly correlated with the exposures to g_t in the cross section.

$$\min_{\xi_j,\chi_{j,\cdot}}\left\{n^{-1}\left\|\left(\widehat{C}_{g,\cdot,j}-\iota_n\xi_j-\widehat{C}_h\chi_{j,\cdot}^{\intercal}\right)\right\|^2+\tau_j n^{-1}\|\chi_{j,\cdot}^{\intercal}\|_1\right\}. \tag{8}$$

- **(2) Post-selection estimation**

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

$$(\widehat{\gamma}_0,\widehat{\lambda}_g,\widehat{\lambda}_h)=\arg\min_{\gamma_0,\lambda_g,\lambda_h}\left\{\left\|\bar{r}-\iota_n\gamma_0-\widehat{C}_g\lambda_g-\widehat{C}_h\lambda_h\right\|^2:\right.$$

$$\left.\lambda_{h,j}=0,\quad\forall j\notin\widehat{I}=\widehat{I}_1\bigcup\widehat{I}_2\right\}. \tag{9}$$

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

E. Two-Pass Regression with Double-Selection LASSO

- The LASSO estimators involve only convex optimizations, so that the implementation is quite fast. Statistical software such as R, Python, and Matlab have packages that implement LASSO using efficient algorithms.
- Double machine learning(Chernozhukov et al. (2018)): Either (1.a) or (1.b) can be replaced by other machine-learning methods such as regression tree, random forest, boosting, and neural network, or by subset selection, partial least squares, and PCA regressions.
- Double LASSO: the underlying asset pricing model is linear, the selected model is more interpretable, and its theoretical properties are more tractable.
- Harvey and Liu (2016): an algorithm that resembles the forward stepwise regression. Their algorithm evaluates the contribution of each factor relative to a preselected best model through model comparison and builds up the best model sequentially. It commits to certain variables too early, which prevents the algorithm from finding the best overall solution later.(robustness)
- Nonnegative regularization parameters to control the level of the penalty, we adopt the widely used CV procedure (Friedman, Hastie, and Tibshirani (2009)).

F. Statistical Inference

- We derive the asymptotic distribution of the estimator for λ_g under a jointly large n and T asymptotic design. d is fixed, s and p can be either fixed or increasing.

THEOREM 1: *Under Assumptions 1 to 6 in Internet Appendix B, if* $s^2 T^{1/2}(n^{-1} + T^{-1})\log(n \vee p \vee T) = o(1)$, *we have*

$$T^{1/2}(\widehat{\lambda}_g - \lambda_g) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Pi),$$

where the asymptotic variance is given by

$$\Pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathrm{E}\left((1 - \lambda^\intercal v_t)(1 - \lambda^\intercal v_s)\Sigma_z^{-1} z_t z_s^\intercal \Sigma_z^{-1}\right), \quad \Sigma_z = \mathrm{Var}(z_t).$$

F. Statistical Inference

provides a Newey-West-type estimator of the asymptotic variance

THEOREM 2: *Suppose the same assumptions as in Theorem 1 hold. In addition, Assumption 7 in the Internet Appendix holds. If* $qs^{3/2}(T^{-1/2}+n^{-1/2})\|V\|_{\text{MAX}}\|Z\|_{\text{MAX}} = o_p(1)$,[8] *we have*

$$\widehat{\Pi} \xrightarrow{p} \Pi,$$

where $\widehat{\lambda} = (\widehat{\lambda}_g : \widehat{\lambda}_h)$ *is given by (9),*

$$\widehat{\Pi} = \frac{1}{T}\sum_{t=1}^{T}(1-\widehat{\lambda}^\intercal v_t)^2 \widehat{\Sigma}_z^{-1}\widehat{z}_t\widehat{z}_t^\intercal \widehat{\Sigma}_z^{-1}$$

$$+\frac{1}{T}\sum_{k=1}^{q}\sum_{t=k+1}^{T}\left(1-\frac{k}{q+1}\right)\left((1-\widehat{\lambda}^\intercal v_t)(1-\widehat{\lambda}^\intercal v_{t-k})\widehat{\Sigma}_z^{-1}\left(\widehat{z}_t\widehat{z}_{t-k}^\intercal + \widehat{z}_{t-k}\widehat{z}_t^\intercal\right)\widehat{\Sigma}_z^{-1}\right),$$

$$\widehat{\Sigma}_z = \frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t^\intercal, \quad \widehat{z}_t = g_t - \widetilde{\eta}_{\widetilde{I}}h_t, \quad \widetilde{\eta}_{\widetilde{I}} = \arg\min_{\eta}\left\{\|G-\eta H\|^2 : \eta_{\cdot,j}=0, \quad j\notin \widetilde{I}\right\},$$

and $\widetilde{I}$ *is the union of selected variables using an LASSO regression of each factor in* g_t *on* h_t*:*

$$\min_{\eta_j}\left\{T^{-1}\|G_{j,\cdot}-\eta_j H\|^2 + \bar{\tau}_j T^{-1}\|\eta_j\|_1\right\}, \quad j = 1,2,\ldots,d. \qquad (10)$$

F. Statistical Inference

- Note that the asymptotic distribution of λ_g does not rely on covariances (C_g, C_h) or factor loadings (β_g ,β_h) of g_t and h_t because they appear in strictly higher order terms, which further facilitates inference.
- Using analysis similar to Belloni, Chernozhukov, and Hansen (2014b), the results can be strengthened to hold uniformly over a sequence of data-generating processes that may vary with the sample size and only under approximately sparse conditions.
- We stress that the inference procedure is valid even with imperfect model selection. our inference is valid without relying on perfect recovery of the correct model in finite sample.

II. Empirical Analysis

II. Empirical Analysis

- **First,** we start by evaluating the marginal contribution of factors proposed over the last five years (2012 to 2016) to the large set of factors proposed before then.
- **Second,** we conduct a recursive exercise in which factors are tested as they are introduced against previously proposed factors. (result)
- **Third,** we explore an alternative application of our procedure(similar in spirit to forward stepwise selection).
- **Finally,** we study the robustness of our procedure from different angles.
  - using alternative methods to reduce the dimensionality of h_t, such as Elastic Net and principal component analysis (PCA), as well as using the stepwise procedure to select the benchmark.
  - alternative portfolio constructions.
  - the tuning parameters.

A. Data

- ***The zoo of factors***

150 risk factors(15+135); July 1976 - December 2017, Monthly frequency

- ***Test portfolios***
- A total of 750 portfolios as test assets(36+714)

  3 × 2 portfolios sorted by size and book-to-market ratio……
- Robustness check: the set of 202 portfolios employed by Giglio and Xiu (2016)

  25 portfolios sorted by size and book-to-market ratio……
- Second robustness check: 1,825 5×5 bivariate-sorted portfolios instead of the 750 3×2 portfolios(175+1650)

B. Evaluating New Factors

- All factors proposed in the 2012 to 2016 period are evaluated against the same benchmark, namely, the factors available up to 2012.

B.1. The First LASSO

the cross-sectional LASSO: select a parsimonious model that explains the cross section of expected returns

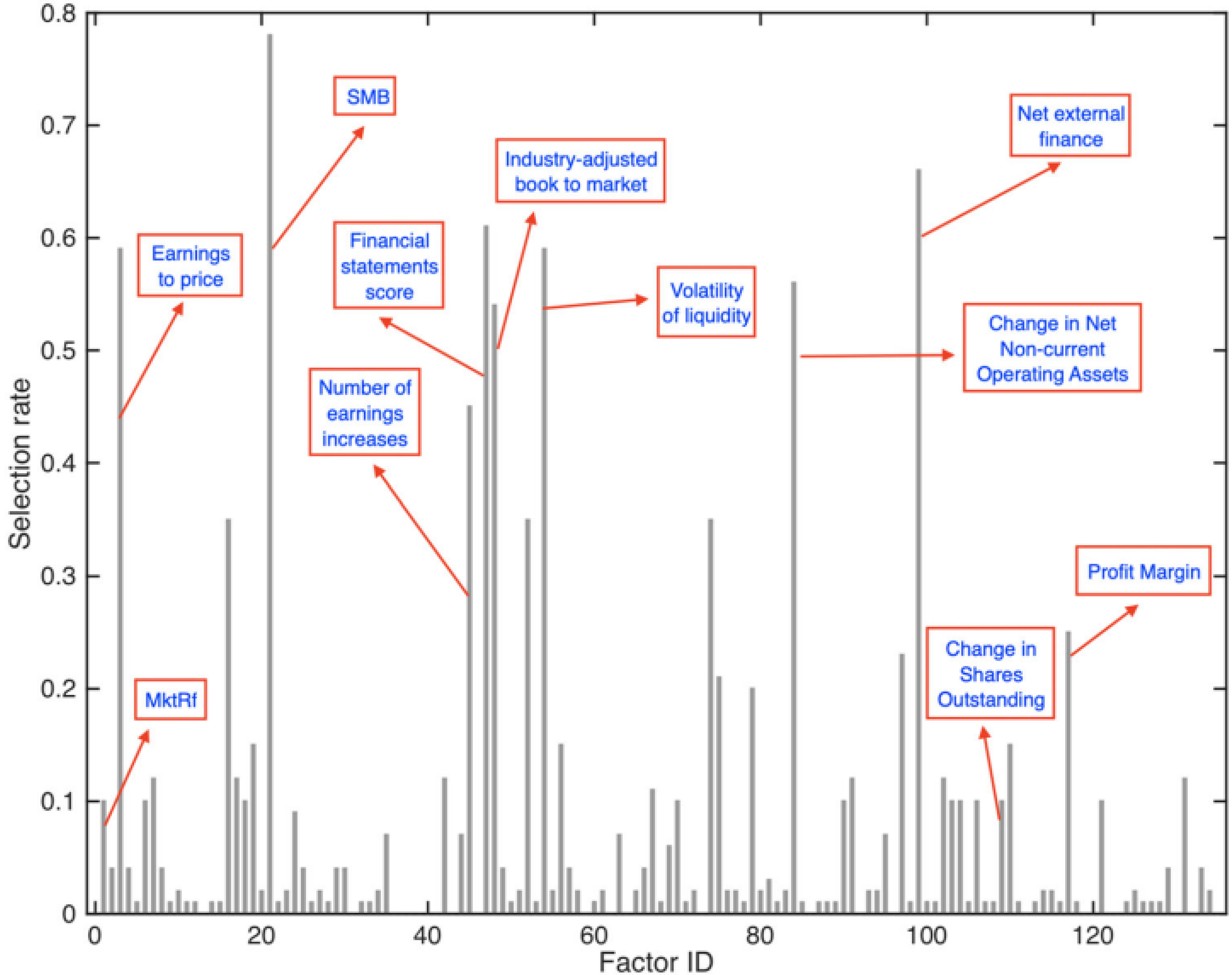
select a relatively small model of the SDF, with four factors: (21), (99), (109), (117).

The main drawback: make mistakes in any finite sample

evaluate the robustness: the LASSO tuning parameter τ_0 (10-fold CV)

in the case of 10-fold CV, we divide the full sample period into 10 disjoint and random subsamples.

we run 200 different 10-fold CV exercises using 200 different randomization seeds.


Selection rate
Factor ID
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
20
40
60
80
100
120
SMB
Net external finance
Industry-adjusted book to market
Earnings to price
Financial statements score
Volatility of liquidity
Change in Net Non-current Operating Assets
Number of earnings increases
Profit Margin
Change in Shares Outstanding
MktRf

- ***B.2. The Second LASSO***
- identify the factors most likely to cause omitted variable bias

The first LASSO

↓

selects a very parsimonious model, with four factors.
(a high τ_0)

The second-stage LASSO

↓

tends to select between 20 and 80 control factors.
(Any factor that could potentially bias the estimate of λ_g should be retained)

↓

Many factors are close cousins

B. Evaluating New Factors

B.3. The Double-Selection Estimator

Table I
Testing for Factors Introduced in the 2012 to 2016 Period

Average excess returns(risk premia)

		(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*

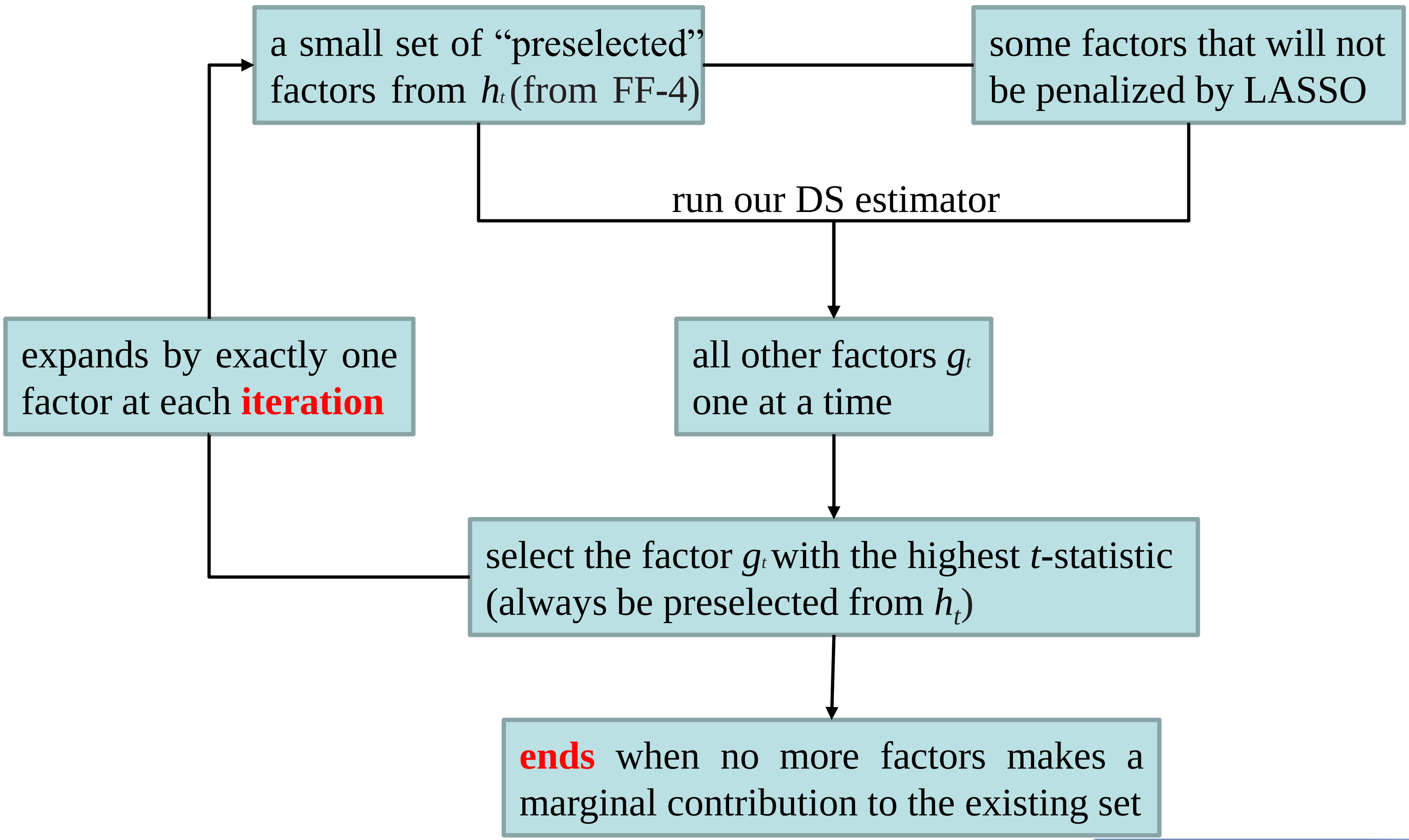
C. Evaluating Factors Recursively

Table II
Testing Factors Recursively by Year of Publication

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										

17 factors

D. A Forward Stepwise Procedure



D. A Forward Stepwise Procedure

Selection results

- **D:** A Forward Stepwise Procedure

148, 88, 51, 62, 74, 61, 49, 122, 6, 55, 72, 53, 119, 140, 44, 147, 65, 32, 31, 87, 123, 5

Introduced in 2012 to 2016
140: Betting Against Beta
147: HXZ investment
148: HXZ profitability

- **C:** Evaluating Factors Recursively

34,38,41,44,50,51,53,64,66,72,95,99,123,140, 145,147,148

D vs C:

caveat

- C:mimics the discovery process over time(2012-2016)
- D:researchers with different *priors* on the correct benchmark model

E. Robustness

E.1. Robustness to the Choice of Tuning Parameters

E.2. Robustness to Test Assets and Regularization Method

Figure 2. Factors introduced in the 2012 to 2016 period: robustness to tuning parameters (t-statistics).

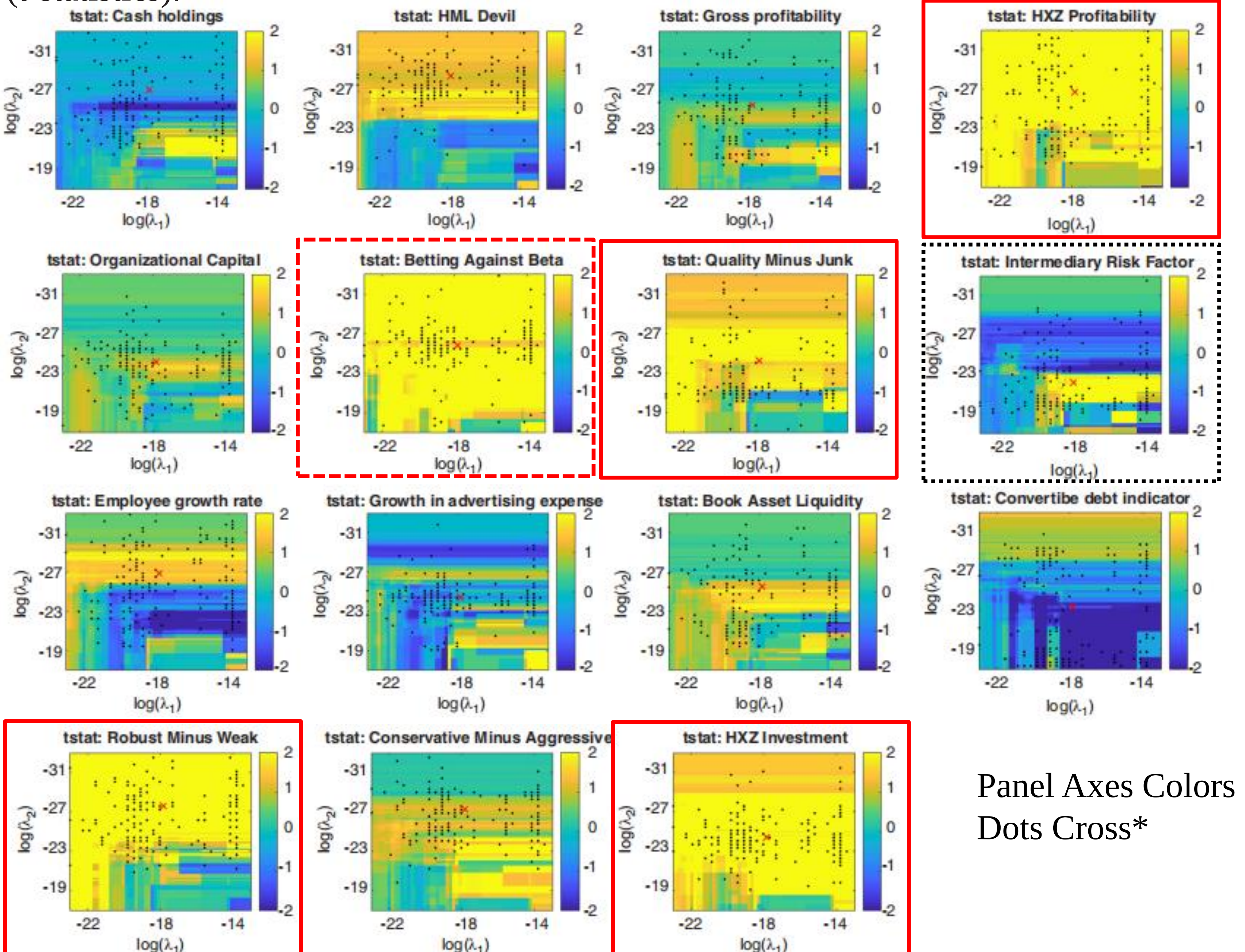


Figure 3. Factors introduced in the 2012 to 2016 period: robustness to tuning parameters (# selected controls).

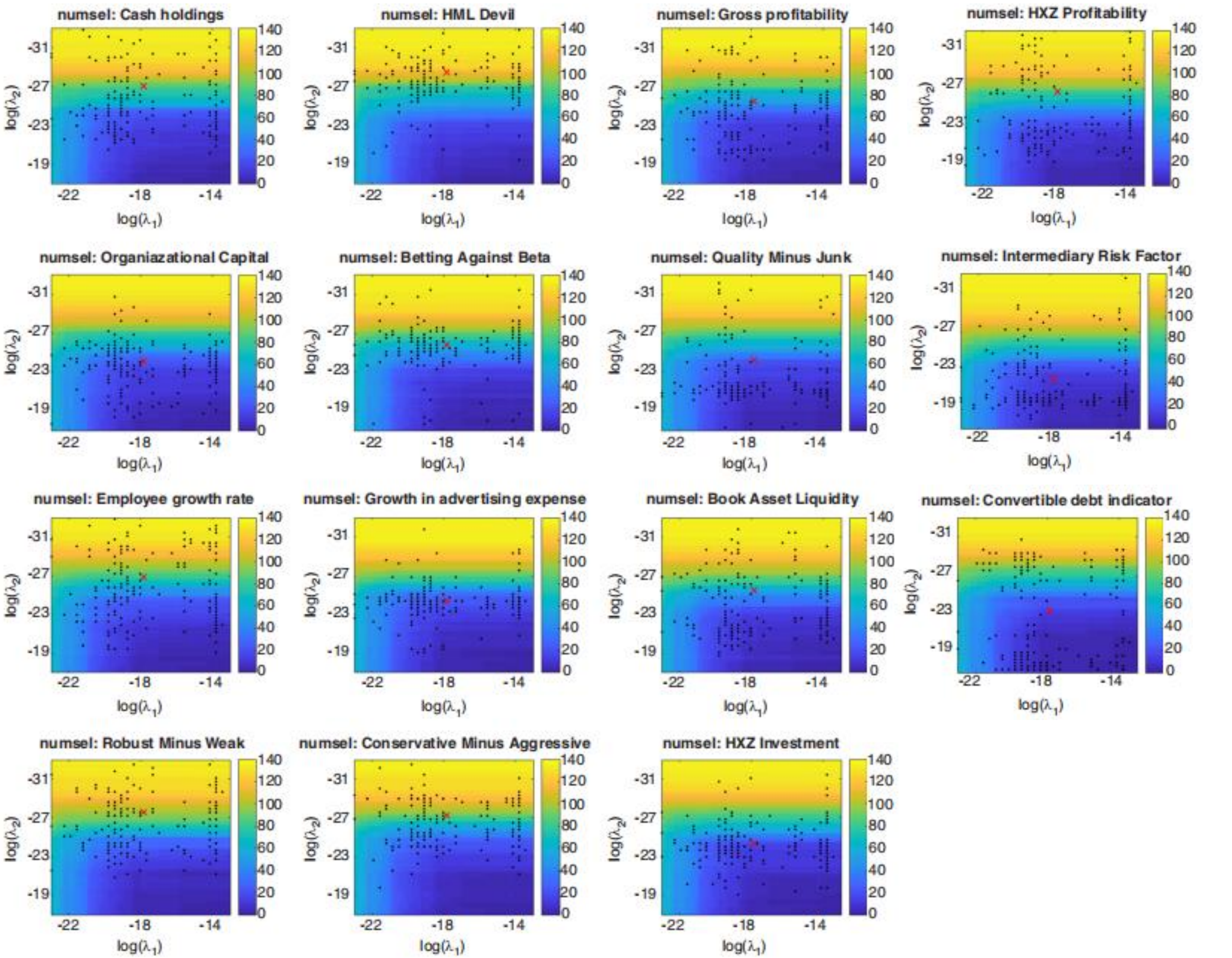


Table III

Robustness for Factors Introduced in the 2012 to 2016 Period

		(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

III. Conclusion

Conclusion

Methodology

- propose a regularized two-pass cross-sectional regression approach
- the DS procedure

Empirical findings

- several newly proposed factors are useful in explaining asset prices
- the SDF loadings' estimates for several factors are robust to changes in the tuning parameters
- only a small number of factors proposed in the literature significant(recursively)
- obtain simply by using the risk premia of the factors or the standard Fama-French three factor model as a control

bring discipline to the "zoo of factors"

Thanks!