Thousands of Alpha Tests

Stefano Giglio[1] Yuan Liao[2] Dacheng Xiu[3]

[1]Yale SOM

[2]Rutgers Econ

[3]Chicago Booth

IMS at NUS, Singapore
Aug 2019

Multiple Testing

- Multiple testing is pervasive in empirical finance.
  - Which factors, among hundreds, adds explanatory power for the cross-sectional variation of expected returns?
  - Which funds, among thousands, produce positive alpha (i.e., have skill)?
  - Which trading strategy, among millions, yields a desirable Sharpe ratio?
  - ...
- These are central questions in both academic finance and asset management industry.

Data Snooping and False Discoveries

- With multiple testing comes the concern of data snooping (i.e., data mining, data fishing, data dredging, p-hacking)
- Find patterns in the data that appear statistically significant even if there is no real underlying effect
- “If you torture the data enough, eventually they will surrender.”
- “I have never seen a bad backtest.”

Example: Identify Funds with Positive Alpha (Skill)

Although each positive alpha has a low probability of being due to chance alone (5%), a large fraction of positives may be “false discoveries.”

- We run 100 tests
- Suppose we know there are 10 true positives among these 100
- At 5% confidence level, we expect $90 \times 5\% = 4.5$ false discoveries
- **The false discovery rate is about** 31% **(4.5/14.5)**

Our Contribution

- We propose a rigorous framework to address testing for multiple alphas in linear asset pricing models.
- Derive a large-T, large-N extension of the FDR control procedure that solves the issues specific to asset pricing
- Illustration: **hedge fund performance evaluation**.

Existing Literature

- The academic finance literature have long been aware of the data snooping issue (Lo and MacKinlay (1990, RFS)).

- One typical approach avoids the multiple testing problem by formulating a single null hypothesis to test: for example, asking whether *any* fund beats the benchmark, (White (2000, ECMA)) or whether funds *on average* beat the benchmark, (Fama and French (2010, JF)).

- Harvey, Liu, and Zhu (2016, RFS) import statistical tools, e.g., FDR control by Benjamini and Hochberg (1995, JRSSB), to solve the multiple testing problem in asset pricing context. See also Barras, Scaillet, and Wermers (2010, JF).

Model Setup

- We assume excess returns, r_t, follow a linear factor model:

$$\underbrace{r_t}_{N\times 1} = \alpha + \beta\lambda + \underbrace{\beta}_{N\times K}\underbrace{(f_t - \mathbb{E}(f_t))}_{K\times 1} + u_t,$$

  where f_t is a vector of factors and u_t is the idiosyncratic component.

- λ are the factor risk premia.

- Note that $\lambda = \mathbb{E}(f_t)$ only if f_t contains exclusively tradable portfolios.

Multiple Hypotheses

We formulate a collection of null hypotheses, one for each fund:

$$\mathbb{H}_0^i : \alpha_i \leq 0, \quad i = 1, \ldots, N. \tag{1}$$

This is fundamentally different from the well-known GRS test (Gibbons, Ross, and Shanken (1989, ECMA)), in which the null hypothesis is a single statement that

$$\mathbb{H}_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0. \tag{2}$$

- GRS tests whether there exists (at least one) fund whose alpha is significantly different from zero.
- We want to find **all** those that have **positive** alphas, which involves two steps:
  1. estimate alpha
  2. control FDR

Controlling the FDR: Notation

- Suppose t_i is a test statistic for the null $\mathbb{H}_0^i$, which rejects the null whenever $t_i > c_i$ under a prespecified cutoff c_i.
- Let $\mathcal{H}_0 \subset \{1, ..., N\}$ denote the set of indices for which the corresponding null hypotheses are true.
- In addition, let $\mathcal{R}$ be the total number of rejections in a sample, and $\mathcal{F}$ be the number of *false* rejections in that sample:

$$\begin{aligned} \mathcal{F} &= \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i \text{ and } i \in \mathcal{H}_0\}, \\ \mathcal{R} &= \sum_{i=1}^{N} 1\{i \leq N : t_i > c_i\}. \end{aligned}$$

- We define *false discovery proportion* (FDP) and its expectation, FDR, as

$$\text{FDP} = \left(\frac{\mathcal{F}}{\max\{\mathcal{R}, 1\}}\right), \quad FDR = \mathbb{E}(\text{FDP}).$$

Comparison with Alternative Approaches

- Naive approach that ignores multiple testing problem:

$$\mathbb{E}(\mathcal{F}) \leq N\tau.$$

- FDR control:

$$\mathbb{E}\left(\frac{\mathcal{F}}{\max\{\mathcal{R}, 1\}}\right) \leq \tau.$$

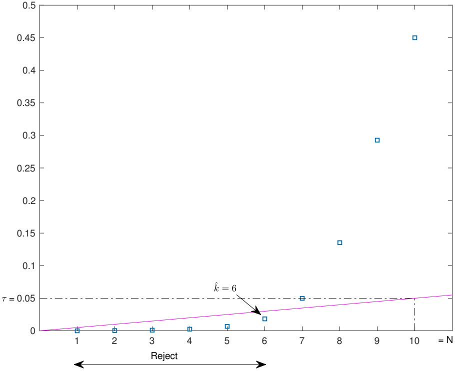
- Bonferroni procedure that controls the family-wise error rate:

$$\mathbb{P}(\mathcal{F} \geq 1) \leq \tau.$$

The B-H Procedure for FDR Control

S1. Sort the p-values of the individual tests, $\{p_i : i = 1, \ldots, N\}$. Denote $p_{(1)} \leq \ldots \leq p_{(N)}$ as the sorted p-values.

S2. Reject $\mathbb{H}_0^i$ if $p_i \leq p_{(\widehat{k})}$, where $\widehat{k} = \max\{i \leq N : p_{(i)} \leq \tau i/N\}$.



Assumptions behind the Classical B-H Procedure

1. Requires independence ⇒ but fund returns are correlated

2. Requires a small number of tests N ⇒ but number of funds N is large
   (Empirically we have 270 months and over 3,000 funds — big data)

3. Requires unbiased estimates for the p-values ⇒ but with fixed N alphas are in general estimated with bias

Issues with the Classical (N Small) Alpha Estimator

- Recall that FM procedure involves two steps to estimate this model:

$$\underbrace{r_t}_{N\times 1} = \underbrace{\alpha + \beta\lambda}_{\mathbb{E}(r_t)} + \underbrace{\beta}_{N\times K} \underbrace{(f_t - \mathbb{E}(f_t))}_{K\times 1} + u_t,$$

  1. estimate β using TS regressions
  2. estimate α as residuals from a CS regression of $\bar{r} = \frac{1}{T}\sum_t r_t$ on $\widehat{\beta}$.

- Two problems:
  1. $\widehat{\alpha}$ is inconsistent, even if the model is correctly specified (regressor β and residual α are not necessarily orthogonal).
  2. Another concern is the possibility of missing some common factors:

$$r_t = \underbrace{\alpha + \beta_l\lambda_l}_{\text{“alpha”}} + \beta_o f_{o,t} + \underbrace{\beta_l(f_{l,t} - \mathbb{E}f_{l,t}) + u_t}_{\text{“idiosyncratic” error}}$$

  where f_o contains observable portfolios, like FF3 or FF4.

The “Blessings” of Dimensionality

- If all factors are observable, then only when N is large, as $N, T \to \infty$ for each $i \leq N$, the FM estimator of α is consistent:

$$\begin{aligned}
\sigma_{i,NT}^{-1}(\widehat{\alpha}_i - \alpha_i) &\xrightarrow{d} \mathcal{N}(0,1), \quad \text{where} \\
\sigma_{i,NT}^{2} &= \frac{1}{T}\mathrm{Var}(u_{it}(1 - v_t\Sigma_f^{-2}\lambda)) + \frac{1}{N}\mathrm{Var}(\alpha_i)\frac{1}{N}\beta_i' S_\beta^{-1}\beta_i.
\end{aligned}$$

Allowing for Latent Factors

- If all factors are latent, we can use PCA to recover them, and then apply FM to estimate λ (Giglio and Xiu (2017)) and α (Connor and Korajczyk (1986, JFE))!
- Recall that

$$r_t = \underbrace{\alpha + \beta H H^{-1} \lambda}_{\mathbb{E}(r_t)} + \beta H H^{-1} (f_t - \mathbb{E}(f_t)) + u_t,$$

  and suppose f_t is latent.

- The SVD of the matrix of (demeaned) returns can recover β and f_t up to some invertible matrix H and H^{-1}, (Bai and Ng (2002, ECMA), Bai (2003, ECMA)).
- It is clear that α is identifiable without being affected by H.

The Mixture Case

- The same identification result holds for general models, where asset returns follow:

$$r_t = \alpha + \left[\begin{array}{cc} \beta_o & \beta_l \end{array}\right] \left[\begin{array}{c} \lambda_o \\ \lambda_l \end{array}\right] + \left[\begin{array}{cc} \beta_o & \beta_l \end{array}\right] \left[\begin{array}{c} f_{o,t} - \mathbb{E} f_{o,t} \\ f_{l,t} - \mathbb{E} f_{l,t} \end{array}\right] + u_t,$$

  where $f_{o,t}$ is a $K_o \times 1$ vector of **observable** factors, and $f_{l,t}$ is a $K_l \times 1$ vector of **latent** factors, respectively.

- α is identifiable even if f_o and f_l are correlated in time series, and/or β_o and β_l are correlated in the cross section.

Algorithm to Estimate α

S1. a. Run time series regressions and obtain the OLS estimator $\widehat{\beta}_o$ and residual matrix Z:

$$\widehat{\beta}_o = (R\mathbb{M}_{1_T} F_o')(F_o\mathbb{M}_{1_T} F_o')^{-1}, \quad Z = \bar{R} - \widehat{\beta}_o \bar{F}_o, \tag{3}$$

where $F_o = (f_{o,1}, f_{o,2}, \ldots, f_{o,T})$.

b. Let $S_Z = \frac{1}{T} ZZ'$ be the $N \times N$ sample covariance matrix of Z. Let

$$\widehat{\beta}_l = \sqrt{N}(b_1, \ldots, b_{K_l}),$$

where $b_1, \ldots, b_{K_l}$ are the K_l eigenvectors of S_Z, corresponding to its largest K_l eigenvalues.

The resulting $\widehat{\beta}$ is given by

$$\widehat{\beta} = (\widehat{\beta}_o, \widehat{\beta}_l).$$

S2. Run CS regression using $\widehat{\beta}$ to obtain $\widehat{\alpha}$.

Constructing Test Statistics for the B-H Procedure

- In all cases, the asymptotic expansion of α is, if $T \log N = o(N)$,

$$\sqrt{T}(\widehat{\alpha}_i - \alpha_i) \quad = \quad \frac{1}{\sqrt{T}} \sum_{t=1}^{T} u_{it}(1 - v_t' \Sigma_f^{-1} \lambda) + o_P(1/\sqrt{\log N}).$$

Algorithm

S3. Calculate the standard error as

$$\mathrm{se}(\widehat{\alpha}_i) \quad = \quad \frac{1}{\sqrt{T}} \widehat{\sigma}_i, \quad \widehat{\sigma}_i^2 = \frac{1}{T} \sum_{t=1}^{T} \widehat{u}_{it}^2 (1 - \widehat{v}_t' \widehat{\Sigma}_f^{-1} \widehat{\lambda})^2,$$

where $\widehat{u}_{it} = z_{it} - \widehat{\beta}_{I,i}' \widehat{v}_{I,t}$ is the residual, and

$$\widehat{v}_t = \begin{pmatrix} f_{o,t} - \bar{f}_{o,t} \\ \widehat{v}_{I,t} \end{pmatrix}, \quad \widehat{v}_{I,t} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\beta}_{I,i}(z_{it} - \bar{z}_i), \quad \widehat{\Sigma}_f = \frac{1}{T} \sum_{t=1}^{T} \widehat{v}_t \widehat{v}_t'.$$

The B-H Procedure with Alpha-Screening

- Because we test (many) inequalities, the procedure might suffer from a power loss if many test statistics are overwhelmingly negative, i.e., "deep in the null."
- We show we can safely reduce the set of funds if their *t*-stats are too negative (i.e., very bad funds),

  $$\widehat{\mathcal{I}} = \left\{ i \leq N : t_i > -\sqrt{\log(\log N)} \right\}.$$

  That is, we accept the null for tests in $\widehat{I}^c$.
- A well known problem with hedge fund datasets is that "bad" funds will simply not report to the dataset.
- This is an important issue for understanding *the average* alpha, but it increases our power to identify good funds.

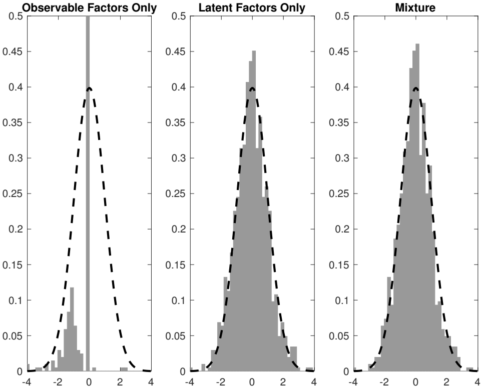
FDR in Simulations

- N $=$ 3,000, T$=$ 300.
- True number of factors $=$ 7, observed $=$ 4, and missing $=$ 3.
- α is a mixture of two Gaussians $\mathcal{N}(-2\sigma, \sigma^2)$ and $\mathcal{N}(2\sigma, \sigma^2)$, plus a point mass at zero, with mixture probabilities: p_1, p_2, $1 - p_1 - p_2$.

Table: False Discovery Rates in Simulations

p_1	p_2	A-S B-H	B-H	B-H	B-H	no FDR
		Mixture	Mixture	Latent	Observable	Mixture
0.1	0.1	5.49	4.63	8.04	23.46	31.77
0.1	0.3	3.92	3.38	5.92	12.27	10.55
0.1	0.5	2.55	2.24	3.79	6.84	4.57
0.3	0.1	4.97	3.42	7.91	18.65	25.29
0.3	0.3	3.16	2.20	4.75	8.49	7.18
0.5	0.1	4.33	2.23	5.99	13.38	18.08

Histograms of the Test Statistics



Illustrative Analysis using Hedge Fund Alphas

- We apply our methodology to the Lipper TASS hedge funds dataset, covering the time period 1994 - 2016.
- Panel of returns and asset under management (AUM)
- The Lipper TASS dataset is subject to a number of potential biases.
  - Reporting to the dataset is voluntary, which induces a selection issue in the funds that appear in the dataset.
  - Funds are able to backfill returns for periods before they start reporting to TASS.
  - Funds can also revise returns they had previously entered in the dataset.

More Cleanning Steps

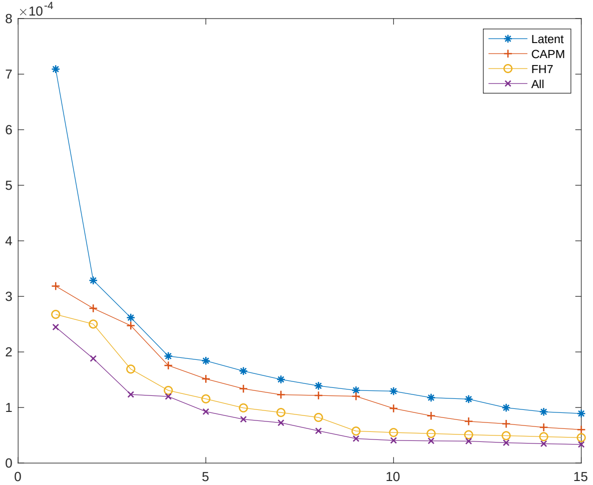
- We compute when possible the returns using the changes in NAV; when NAV is not available, we use reported returns.
- We remove funds that do not report monthly, and funds that do not report net-of-fee returns.
- We also remove funds that do not consistently report AUM: we require funds to report AUM at least 95% of the time.
- We also remove suspicious returns: monthly returns above 200% or below -100%, stale returns (equal to the past two monthly returns), cases where AUM is reported as zero, and cases where within 6 months funds display a 10,000% increase immediately reversed, as these are likely just data entry errors.
- We remove duplicate series by screening for cases in which two funds have return correlations of 99% or more for periods of at least 12 months.
- We require funds to have reported returns and AUM to the dataset for at least 12 months.
- We require funds to have at least $20m of AUM, and drop them after they fall below this amount.

Benchmark Models

- We consider several alternative benchmark models.

  1. the CAPM and the Fama-French 4-factor model (market, size, value and momentum factors).
  2. the Fung and Hsieh 7-factor model (Fung and Hsieh (2004)), that includes market, size, a bond factor, a credit risk factor, and three trend-following factors (related to bonds, currencies, and commodities).
  3. two option-based factors (an out-of-the-money call and an out-of-the-money put factor) from Agarwal and Naik (2004).

Scree Plot of Eigenvalues

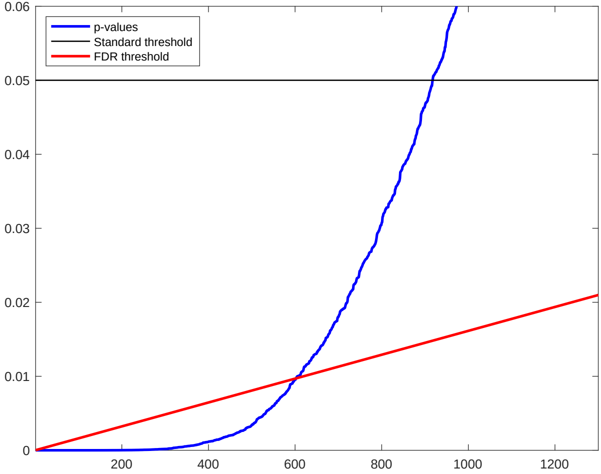


In-Sample Results

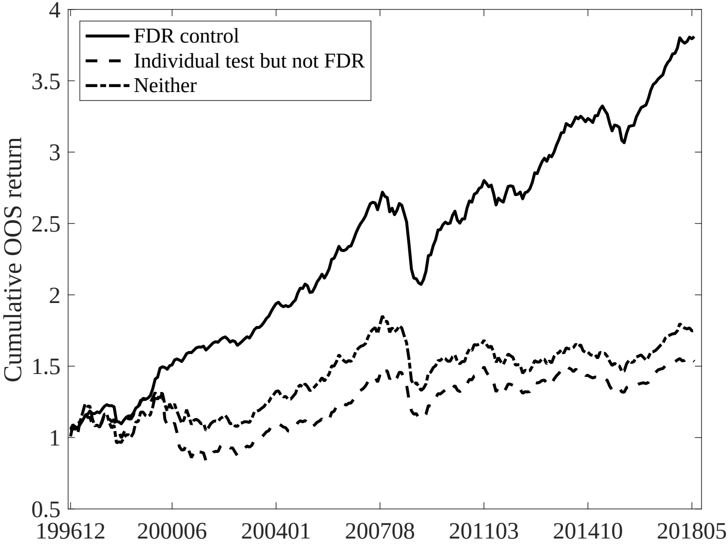
Table: Standard multiple tests vs. FDR control test: in-sample results

	# factors	Individual test	FDR	% Excluded
CAPM	1	1342	1079	20%
FF4	4	1367	1121	18%
FH	7	1200	927	23%
FH + Option factors	9	1041	731	30%
All observable	11	1112	817	27%
5 Latent factors	5	1296	1043	20%
10 Latent factors	10	1137	835	27%
All observable + 5 Latent	16	917	602	34%

Illustration of B-H Procedure (All observable + 5 Latent)



Out-of-sample Portfolio Return



Out of Sample Results

Model: FH7		Panel 3: Closure rate			Panel 4: α (bp/month)			Panel 5: T-stats		
Cutoff	# Funds	FDR	Indiv.	None	FDR	Indiv.	None	FDR	Indiv.	None
200112	225	0.13	0.29	0.19	29	40	43	1.90	2.23	1.08
200212	410	0.11	0.15	0.13	46	35	44	3.51	1.54	1.64
200312	481	0.08	0.19	0.16	39	24	38	2.47	1.17	1.41
200412	582	0.13	0.15	0.24	39	28	43	2.98	1.49	1.48
200512	673	0.13	0.03	0.23	38	47	46	2.92	1.40	1.97
200612	870	0.16	0.20	0.22	34	71	30	2.08	1.59	1.08
200712	929	0.31	0.61	0.48	-12	-40	-66	0.77	-0.37	-0.40
200812	805	0.19	0.33	0.31	34	11	16	1.56	0.57	0.76
200912	897	0.13	0.23	0.28	26	15	9	1.76	1.05	0.83
201012	730	0.13	0.23	0.32	32	14	12	2.09	1.28	1.22
201112	827	0.12	0.10	0.26	32	22	14	2.19	1.41	0.89
201212	786	0.11	0.20	0.20	43	19	21	2.47	0.85	0.77
Model: 5 latent										
200112	225	0.13	0.14	0.21	23	34	48	1.81	0.60	1.35
200212	410	0.09	0.17	0.15	42	38	31	3.04	1.82	1.27
200312	481	0.07	0.19	0.16	31	30	29	2.27	1.45	1.28
200412	582	0.12	0.11	0.25	33	22	32	2.58	1.15	1.20
200512	673	0.11	0.19	0.23	41	48	48	2.78	1.72	1.66
200612	870	0.15	0.13	0.22	39	58	45	2.21	1.69	1.26
200712	929	0.28	0.30	0.46	37	16	13	1.65	0.80	0.63
200812	805	0.19	0.37	0.31	30	24	12	1.49	0.66	0.80
200912	897	0.14	0.13	0.28	30	18	12	2.01	1.15	0.81
201012	730	0.13	0.26	0.29	39	24	14	2.50	1.26	1.19
201112	827	0.12	0.16	0.24	43	37	20	2.65	1.50	0.87
201212	786	0.10	0.12	0.20	46	29	26	2.50	1.31	0.92

Conclusion

- When applying machine learning techniques to high dimensional problems in asset pricing, data snooping may lead to potentially numerous false discoveries.
- We present a rigorous framework to address the data-snooping concerns that arise when applying multiple testing in the asset pricing context.
- The methodology exploits the "blessing of dimensionality" to build a valid FDR control procedure
  - when the benchmark includes non-tradable factors
  - robust to the presence of omitted factors
- We illustrate this procedure by applying it to the evaluation of hedge fund performance.