利用信息协变量和错误发现率构建基金组合

韩雪菲

指导老师：邓璎函

西南证券金工组实习项目

版本：0.01

日期：2022 年 5 月 1 日

摘　要

Hsu et al. (2020) 提出了一种新的多重假设检验框架 $fFDR^+$，用于选择表现出色的基金组合。在估计错误发现率时结合了信息协变量。这九个信息协变量分别为 Carhart 四因子模型的 R 方、收益差距、活跃权重、基金规模和资金流量夏普比率、基于资本资产定价模型的贝塔和特雷诺比率以及 Carhart 四因子模型的的波动率。在给定的 $fFDR^+$ 目标下加入信息协变量，我们的投资组合可以表现出正向 alpha。

风险提示: 本文的研究是基于对历史数据的统计和分析，因子的历史收益率不代表未来收益率。若市场环境发生变化，因子的最终表现可能发生改变。

目录

1 前言

本文的目标是在中国基金市场复现基于 Hsu et al. (2020) 文章的多重假设检验框架下的基金组合策略。从主观方法来看，基金组合的构建比较复杂，要考虑宏观的经济形势，到中观的行业轮动，再到微观的基金研究，每个环节都需要深度的考虑。从量化角度，如果投资者仅根据他们过去的 alpha 选择组合持有基金是远远不够的，过去的高 alpha 可能仅仅是由于运气。

Hsu et al. (2020) 认为从量化角度出发, 使用传统的的多重假设检验框架 (FDR^+) 可以识别真正有能力的熟练型基金（skilled fund）-即由于优秀的管理能力及交易策略而具有真正正向 alpha 的基金。在此基础上，他们将传统的的 FDR^+ 与信息协变量结合提出 $fFDR^+$。他们指出通过将多个维度的信息协变量纳入考量，设定一个目标 $fFDR^+$，基金组合策略可以获得相当不错的表现。

我们在基金市场对 $fFDR^+$ 进行了相关实证, 第二部分进行信息协变量和 alpha 数据准备，第三部分我们解释如何构建 FDR^+ 和 $fFDR^+$ 来衡量由于运气的产生正 alpha 基金个数占全部熟练真正有正 alpha 基金的占比。第四部分通过设定目标 $fFDR^+$ 值和不同的信息协变量构建组合策略并进行回测。

2 数据准备

我们使用的基金池为同花顺开放式基金中的 647 个股票型基金（保留有相关信息协变量的数据的基金并剔除观察值小于 48）。数据为 2017.03-2022.03 的月度收益。

2.1 信息协变量的计算

我们将 $fFDR^+$ 定义为与基金业绩相关的协变量的函数。这样不仅利用基金业绩信息，同时也可以控制了 FDR^+。Hsu et al. (2020) 的实证结果突出了这些协变量的信息价值，因为基于它们，我们能够构建始终产生真正正 alpha 的投资组合。所以首先我们进行信息协变量的计算。

使用 2017-2020 月度数据计算不同基金的信息协变量如下:

1. R-square: 根据 Carhart 四因子模型估计并衡量基金的活跃度,R_square 越小，在基金往往表现更好. 在复现中，我们使用 2017-2020 年数据及 Carhart 四因子模型回归获得。
2. Fund Size: 基金规模反映了基金规模相对于整个活跃共同基金市场的增长。基金规模与基金业绩之间存在显着的负相关关系。在复现中，我们使用 2020 年 3 月的值作为 Fund Size。

$$FundSize_{i,t} = ln\frac{TNA_{i,t}}{IndustrySize_t} - ln\frac{TNA_{i,0}}{IndustrySize_0}$$

3. Return Gap：基金的回报差距等于基金本期实际披露的回报与基金根据其上一次披露信息的理论回报之间的差异。过去回报差距越大的共同基金在未来往往表现更好。在复现中，我们使用在 2019 年 3 月到 2020 年 3 月的平均回报差距。
4. Active Weight：活跃权重定义为股票价值权重与基金分配给其投资组合中股票的实际权重的绝对差值之和。活跃权重较高的基金往往表现更好。(同花顺无相关数据权限，本文未讨论)。
5. Fund Flow：资金流入的基金比流出的基金表现更好，在复现中，我们使用 2019 年 3 月到 2020 年 3 月资金流动作为 Fund Flow。

$$FundFlow_t = \frac{TNA_t - (1+r_t)TNA_{t-1}}{(1+r_t)TNA_{t-1}}$$

6. Sharpe ratio, Beta，Treynor ratio: 由使用 2017-2020 年数据和 CAPM 模型回归得到。
7. Idiosyncratic volatility (Sigma): Carhart 四因子模型的 alpha 的波动率 (Sigma)

2.2 根据 Carhart 模型的 alpha 分布

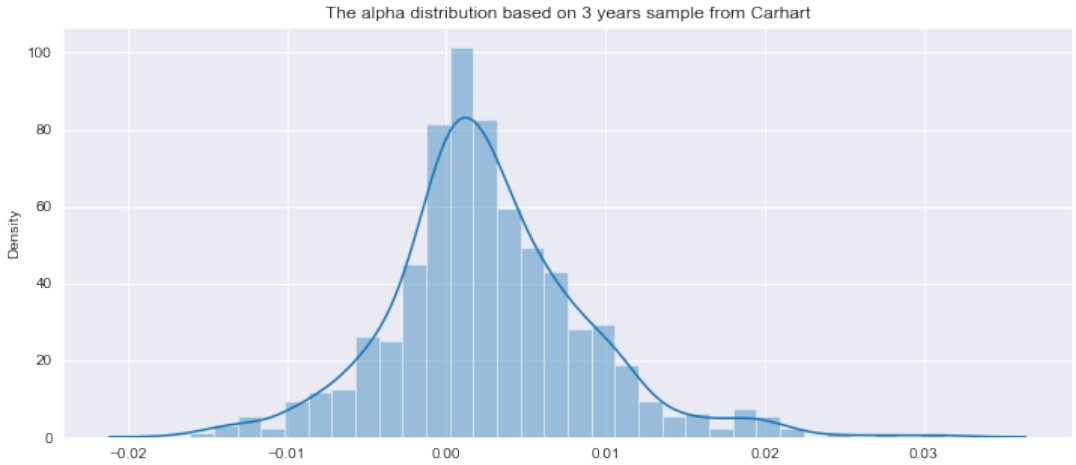
我们使用 Carhart (1997) 的四因子模型来计算基金月度业绩:

$$r_{i,t} = \alpha_i + b_i r_{m,t} + s_i r_{smb,t} + h_i r_{hml,t} + m_i r_{mom,t} + \varepsilon_{i,t}, i = 1, ..., m,$$

$r_{i,t}$ 是基金 i 超过无风险利率（即月度银行一年固定利息）的超额净收益
$r_{m,t}$ 为市场风险溢价因子：考虑现金红利再投资的月市场回报率 (流通市值加权平均法) 与月度化无风险利率之差
$r_{smb,t}$ 市值因子: 小盘股组合和大盘股组合的月收益率之差，组合划分基于 FAMA 2*3 组合划分方法。组合月收益率的计算采用流通市值加权计算
$r_{hml,t}$ 账面市值比因子: 高账面市值比组合和低账面市值比组合的月收益率之差，组合划分基于 FAMA 2*3 组合划分方法。组合投资收益率的计算采用流通市值加权
$r_{mom,t}$ 动量因子:Carhart 四因子计算方法：形成期为 11 个月，累积收益最高的前 30% 的投资组合收益率-累积收益最低的 30% 的投资组合收益率（流通市值加权）
数据来源：国泰安数据库
根据估计的 alpha 我们定义:

1. 非熟练基金/零 alpha 基金（unskilled fund 或者 zero alpha fund)：基金经理的选股技能足以收回其交易成本和费用，但是 alpha 为 0 或者负值
2. 熟练基金（skilled fund)：拥有足够选股能力的基金，可以提供超额 alpha，而不仅仅是收回交易成本和费用（alpha > 0)。

我们可以看到 5 年平均 alpha 的分布情况如图:



3 将 $fFDR$ 运用到 FDR^+ 策略得到 $fFDR^+$ 基金选择策略

3.1 传统 FDR^+ 策略

当我们得到每个基金的 alpha，我们对每个基金进行假设性检验:

$$H_0 : \alpha = 0, \qquad H_1 : \alpha \neq 0$$

由此得到 p 值，我们认为 p 值是均匀分布的，那么对于某个基金，零 alpha 基金的占比: $\pi_0(\lambda)$ 可以由下面公式获得:

$$\pi_0(\lambda) = \frac{\#(p_i | p_i > \lambda, i = 1, ...m)}{(1-\lambda)m}$$

这里的 λ 等于 λ^* 最优，此时 λ^* 是一个足够大的 p 值的阈值由优化求解使 $\pi_0(\lambda)$ 的均方误差 (MSE) 最小获得.

然后我们利用定义的 FDR^+ 来衡量选定的基金组合（alpha 为正而且 p 值显著的基金组合）中的 false discovery rate（错误发现率），

$$FDR^+ = \mathbb{E}(\frac{V^+}{max\{R^+, 1\}})$$

R^+ 是实证中熟练基金个数，V^+ 为错误地选择了零阿尔法或不熟练的基金个数

在给定的一个显著性的阈值 γ，我们认为当一个基金 p 值 $\leq \gamma$ 时，他有大于 0 的估计 alpha ，也就是认为他是熟练基金，所以此基金组合的 FDR^+ 为

$$FDR^+ = \frac{\pi_0 \gamma / 2}{R^+ / m}$$

R^+ 是 alpha 大于 0 且 p 值 $\leq \gamma$ 的基金个数, 在复现过程中，我们的基金池中，只有 444 个基金的 alpha 为正，在 10% FDR^+ 目标下，只有以下基金的 p 值小于 γ，会被选中为熟练基金。

time	Hypothesis Correction Result
re000751OF	True
re002236OF	True
re003359OF	True
re003646OF	True
re003647OF	True
re167702OF	True

3.2 利用 $fFDR$ 将 FDR^+ 升级 $fFDR^+$

将信息协变量纳入考量后，在已知给定的信息协变量，单个基金的假设性检验如下：

$$H_0 : \alpha = 0, \qquad H_1 : \alpha \neq 0$$

我们定义 h 为原假设的状态，即如果假设 α = 0 为真，则 h = 0，否则为 h = 1。此外，P 是检验 p 值的随机变量表示，Z 代表某个信息协变量，所以单个基金的信息集为 T = (P, Z)，我们假设 (h|Z = z) 服从 Bernoulli$(1 - \pi_0(\lambda))$，以 Z = z 为条件，该基金具有零 alpha 的概率为 $\pi_0(\lambda)$.

以之前同样的方式求解 λ 的最优值（我们没有对单个基金求 λ 而是在多重检验时，将基金依据相同的信息协变量分为小组，对每组以最小化平均积分平方误差（MISE）求最优化的 λ.

接着我们对实证中已选中的熟练基金组合（alpha 大于 0 且 p 值 $\leq \gamma$ 的 m 个基金）进行评估，进行 multiple hypothesis testing 多重假设检验

$$H_{0,i} : \alpha_i = 0, \qquad H_{1,i}\alpha_i \neq 0, \qquad i = 1, ...m$$

我们假设所有基金都是独立的，并且每个基金的信息集 T =(P,Z) 都具有相同的分布。最后，我们用 f(p,z) 表示 (P,Z) 的联合密度函数。我们可得：在观察到某个 T = (p, z) 时 null hypothesis 零假设为真的后验概率为

$$\mathbb{P}(h = 0|T = (p, z)) = \frac{\pi_0(z)}{f(p, z)} \doteq r(p, z)$$

这里的 $\pi_0(z)$ 是将上述每组的 $\pi_0(\lambda)$ 使用平滑样条法 smoothing spline method 所得；

f(p,z) 是 probit 转换后的 local likelihood kernel density estimation (KDE) 方法求得. 详细参考 Hsu et al. (2020) 附录 A.1。目前可以使用 R package locfit 来实现。

在正 alpha 基金的基金池中我们得到 $fFDR^+$：

$$fFDR^+(\Gamma) = \mathbb{P}(h = 0|T \in \Gamma) = \int_\Gamma r(p, z)dpdz$$

Γ 是所有的信息集组合。

我们定义在给定的目标 τ 下的 q 值为:

$$q(p,z) = \inf_{\{\Gamma_\tau | (p,z) \in \Gamma_\tau\}} fFDR^+(\Gamma_\tau)$$

实践中在给定的 (p_i, z_i) 下的 q 值计算公式为:

$$q(p_i,z_i) = \frac{1}{S_i} \sum_{k \in S_i} r(p,z) \qquad S_i = \{j | r(p_j,z_j) \leq r(p_i,z_i)\}$$

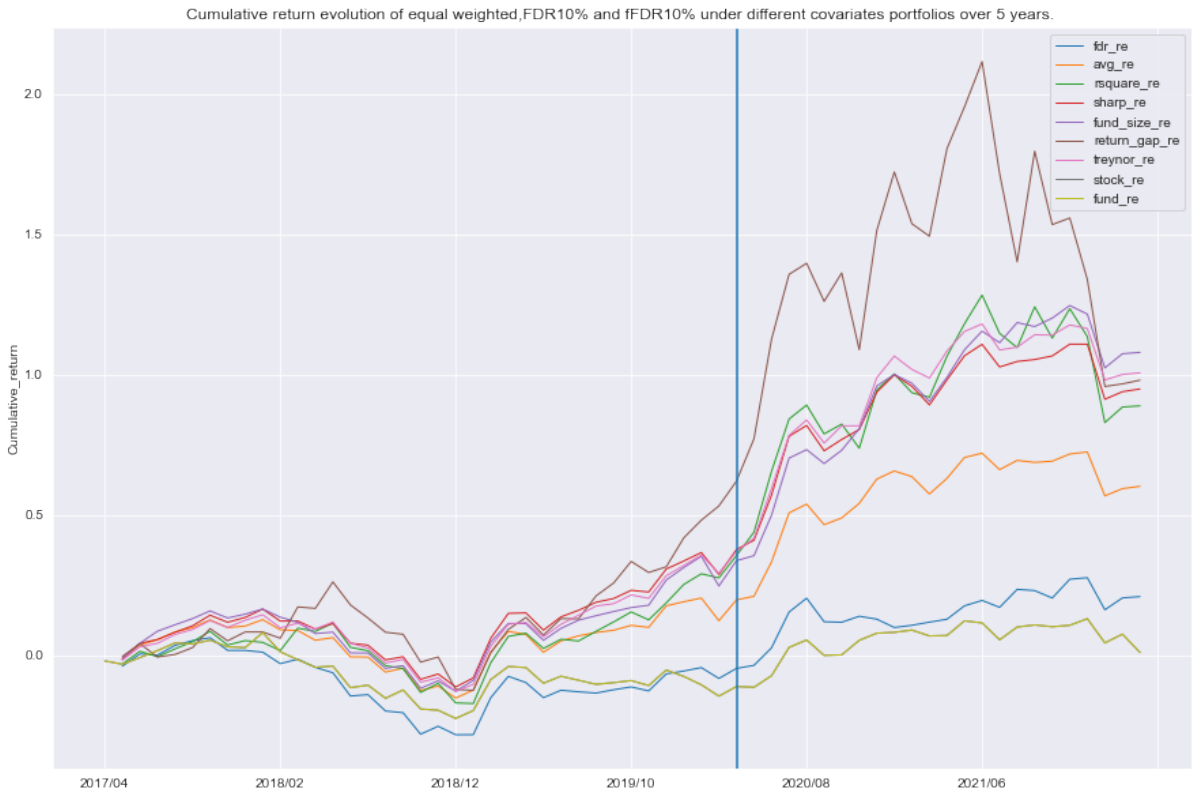
p_i 是第 i 次假设检验的 p-value, $z_i = r_i/m$, 这样我们将信息协变量 Z_i 的观测值转换为满足假设 Z_i 服从 Uniform(0, 1) 的形式.r_i 是观测值 Z_i 在的样本中的排名。

这样在给定目标 $\tau \in$ [0,1]，当且仅当其 q 值 $\leq \tau$ 我们才会拒绝零假设，这样可以保证才可以保证 FDR 控制在 τ。目前可以用我们使用了 r package fFDR 实现。

4 构建投资组合

通过 $fFDR^+$ 来控制因为运气选入的非熟练资金的比例，目标 $fFDR^+$=10%
我们利用 2017-2020 年数据得到基金的信息协变量的观测值，估计的 alpha 和假设检验的 p 值。按照 $fFDR^+$ 目标 10% 和滚动的 5 年样本构建我们的投资组合。从 2021 年开始，在 t 年末，我们根据过去五年（t-4 到 t）的历史信息，选择一组基金在 t+1 年进行投资。我们算出了滚动样本下不同的信息协方差和 alpha 值，（见 Jupyter Notebook 2) 但是由于 KDE 和 fFDR 是由 R 代码实现的。Python 代码实现仍在调试，我们没有在第二年进行 rebalance 重平衡。
由于数据量不足，我们在利用基金流量，波动性，贝塔比率作为信息协变量时，没有满足条件的基金入选我们的投资组合. 以。
在 Carhart 四因子模型的 R 方下 $fFDR+$ 策略有 2 个基金入选；在收益差距 (return gap) 下有 1 个基金入选；在基金规模 (fund size) 下有 19 个基金入选；在夏普比率 (sharp ratio) 下有 100 个入选；基于资本资产定价模型的特雷诺比率（treynor ratio) 下有 7 个入选。

目前我们将已有的五个信息协变量下的 $fFDR+$ 策略，$FDR+$ 策略，等权重策略，与股票指数和基金指数的回报进行对比, 得到如 Hsu et al(2020) 中的不同的信息协变量下该策略的回报在样本内和样本外的时间序列:



下表对比了不同的策略的表现：

	FDR+	avg	Rsquare	Sharp	Fund_size	Return_gap	Treynor
Annual return	5,2%	10,8%	15,0%	14,9%	16,3%	17,6%	15,5%
Sigma	4,8%	4,5%	5,8%	4,6%	4,7%	7,9%	4,6%
Sharp	8,9%	19,7%	21,3%	26,5%	28,8%	18,5%	27,8%
alpha_Carhart	0,1%	0,3%	0,7%	0,6%	0,7%	0,8%	0,7%
alpha_FF	-0,1%	0,8%	1,1%	1,2%	1,7%	2,0%	1,5%

加入 $fFDR^+$ 我们可以看到这五个信息协变量下 $fFDR^+$ 策略的表现，五个信息协变量下 $fFDR^+$ 策略均有正向 alpha，而且可以产生 15% 及以上的年化收益。目前来看，相比于等权重策略和传统 FDR^+，加入信息协变量的 $fFDR^+$ 表现更为优秀。

5 总结与展望

本文利用开放型基金市场股票型基金，复现了 Hsu et al. (2020) 多假设检验框架的 $fFDR^+$ 策略，来选择表现出色的基金组合。在估计错误发现率时创新性的结合了信息协变量。在不同的信息协变量和给定的 $fFDR^+$ 目标下，构建的投资组合可以有较好收益。

6 风险提示

本文的研究是基于对历史数据的统计和分析，因子的历史收益率不代表未来收益率。若市场环境发生变化，因子的最终表现可能发生改变。

7 参考文献

Hsu, Po-Hsuan and Kyriakou, Ioannis and Ma, Tren and Sermpinis, Georgios, Informative Covariates, False Discoveries and Mutual Fund Performance (November 25, 2020).