北京大学量化交易协会2020级培训

多因子模型II

杨思源 姚雨薇 刘 畅 李琳雄

2020-10-19

多因子模型Ⅱ

多因子模型构建

大类因子分析和共线性分析

概 念

- 因子共线性分析和大类因子分析的本质目标都是一致的，都是避免最终的回归过程中出现多重共线性问题。多重共线性是指回归模型中的解释变量之间由于存在精确相关关系或高度相关性而使模型估计失真或者难以估计准确。
- 分别两个环节进行的理由是：
- 如果是经济含义类似的同类型因子，存在明显相关性，为尽可能多的保留因子信息，我们可以将因子进行合并;
- 如果是经济含义不同的因子，存在明显相关性，我们只能有所取舍，保留更加显著的因子，而舍弃相对不显著的因子， 因为多因子模型除了效果，最终还是要讲求因子本身的经济含义的。
- 注意：删除因子的方法会损失诸多有效因子信息。

大类因子分析

产生多重共线性的原因

- 多因子模型强调因子本身的经济含义和实证有效性两个方面。
- 在因子搜集的时候就会根据因子的具体经济含义对因子进行大类划分，但是同类型的因子可能存在较强的相关性，多元线性回归的时候会造成多重共线性。

解决方法

- 在有效因子筛选出来之后，我们首先需要根据大类对因子的相关性进行t检验，判断大类因子间是否有相关性。
- 对于相关性较高的因子，要么舍弃显著性较低的因子，要么进行因子合成。

大类因子分析

步骤一 同类因子的相关性检验

- 考虑同类型的K个候选因子，向前选取M个月的数据作为样本。
- 按月计算出因子载荷（因子暴露）之间的相关系数矩阵。
- 然后然后根据M个月的相关系数的绝对值进行检验t。

$$\rho^t = \begin{bmatrix} 1 & \rho_{12}^t & \dots & \rho_{1K}^t \\ \rho_{21}^t & 1 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{K1}^t & \cdots & \cdots & 1 \end{bmatrix}, t = 1, \dots M \quad \Longrightarrow \quad t = \frac{\overline{|\rho_{ij}^t|} - u}{\sigma / \sqrt{M-1}} \quad \text{where} \quad \overline{|\rho_{ij}^t|} = \sum_{t=1}^{M} |\rho_{ij}^t| / M$$

步骤二 因子取舍或者因子合成

- 对于相关性较高的因子集合，可以采取两种方式处理：
- 1. 根据因子本身的有效性进行排序，挑选最有效的因子进行保留，删除其他因子；
- 2. 对因子集合进行合成，尽可能多的保留有效因子信息；

因子共线性分析

例子：Fama-French 五因子模型

图表12： 因子间相互回归结果

被解释变量及对应统计量	截距项	R_M	SMB	HML	RMW	CMA	拟合优度
被解释变量——R_M							
回归系数	0.026		-0.38	-0.29	-1.63	-0.96	0.15
t 值	2.86		-1.37	-1.13	-3.73	-1.28	
被解释变量——SMB							
回归系数	0.02	-0.04		-0.33	-0.74	-0.31	0.37
t 值	6.48	-1.37		-4.58	-5.82	-1.33	
被解释变量——HML							
回归系数	0.02	-0.03	-0.41		0.04	1.33	0.47
t 值	5.62	-1.13	-4.58		0.26	5.68	
被解释变量——RMW							
回归系数	0.00	-0.06	-0.27	0.01		-1.28	0.80
t 值	1.98	-3.73	-5.82	0.26		-14.40	
被解释变量——CMA							
回归系数	0.00	-0.01	-0.04	0.14	-0.47		0.80
t 值	-1.80	-1.28	-1.33	5.68	-14.40		

- SMB:规模因子
- HML:估值因子
- RMW：盈利因子
- CMA：投资因子

因子共线性分析

解决办法：

- 观察上表我们发现，CMA 与 RMW 有很强的共线性，我们将投资因子中与其他因子共线的部分去除，令：

$$\mathrm{CMAO} = \mathrm{CMA} - \widehat{CMA}$$
$$\widehat{CMA} = \alpha + bR_M + sSMB + hHML + rRMW$$

- 针对修正之后的因子，得到了新的五因子模型：

$$\mathrm{R}_{\mathrm{it}} = \alpha_i + b_i R_{Mt} + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMAO_t + \epsilon_{it}$$

因子共线性分析

结果：

图表15： 美股规模-估值（BP）分组 25 宫格股票组合回归结果

BP	Small	2	3	4	Big
			Alpha		
Low	-0.29	-0.11	0.02	0.18	0.12
2	0.11	-0.10	-0.01	-0.23	-0.11
3	0.01	0.05	-0.07	-0.13	-0.10
4	0.12	0	-0.02	0.05	-0.15
High	0.12	-0.04	0.05	-0.09	-0.09

1.原模型下规模-估值分组25宫格股票组合回归结果

图表14： A 股规模-估值（BP）分组 25 宫格股票组合回归结果

BP	Small	2	3	4	Big
			Alpha		
Low	0.0028	-0.0081	-0.0057	-0.0018	0.0048
2	0.0006	-0.0050	-0.0055	-0.0029	-0.0031
3	0.0012	-0.0069	-0.0039	-0.0052	-0.0009
4	0.0016	-0.0039	-0.0041	-0.0044	-0.0002
High	0.0031	-0.0023	-0.0026	-0.0025	-0.0019

2.新模型下规模-估值分组25宫格股票组合回归结果

修正后的五因子模型基本解释了所有超额收益，即图表 2 中的 Alpha（就是回归的截距项）相对比较小，因此通过多元线性回归调整后的五因子模型有更好的结果

多因子模型构建

1 大类因子分析和共线性分析

2 因子合成

3 模型构建：排序打分法

4 模型构建：截面回归法

因子合成

因子合成的作用和方法

因子合成，尽可能保留有效因子的信息，因子合成主要目的是降低共线性以及生成大类风格因子

- 等权法
- IC加权法
- IC-IR加权法
- 最大化IC-IR加权法
- IC半衰加权法

因子合成

等权合成法

将所有相关性高的因子等权重进行合成，合成新的因子暴露，然后再进行标准化处理。

因子等权加权即赋予每个因子同样的权重，其权重向量

$$V=\left(\frac{1}{M},\frac{1}{M},\ldots,\frac{1}{M}\right)'$$

这种加权方式较为简单，但是没有考虑因子有效性的差异。

因子合成

IC加权法

IC值是每个时间截点上因子暴露值和股票下期收益的相关系数。IC值越高意味着该因子的暴露度与未来收益值存在较明显的相关关系。|IC|>0.03,认为因子有效，可以用于区分股票。

IC 均值加权，是指用过去一段时间因子 IC 的均值作为权重，即权重向量

$$V = (\overline{IC_{f_1}}, \overline{IC_{f_2}}, \dots, \overline{IC_{f_M}})'$$

其中 $\overline{IC_{f_i}}$ 为因子 f_i 过去一段时间内 N 期 IC 的均值

$$\overline{IC_{f_i}} = \frac{1}{N}\sum_{t=1}^{N} IC_{f_i}^t$$

这种加权方式相较于等权加权，考虑了每个因子的长期有效性的差异，以因子的有效程度作为权重来加权。

因子合成

IC-IR加权法

IR即信息比率（Information Ratio），是超额收益的均值与标准差之比，可以根据 IC 近似计算。
IR= IC的多周期均值/IC的标准方差。

IC_IR 加权是指以因子过去一段时间的 IC 均值除以其标准差作为当期因子 f_i 的权重，即权重向量

$$V = (IR_{f_1}, IR_{f_2}, \dots, IR_{f_M})'$$

其中 IR_{f_i} 为因子 f_i 的 IC_IR

$$IR_{f_i} = \frac{\overline{IC_{f_i}}}{std(IC_{f_i})}$$

这种加权方式相较于 IC 均值加权，综合考虑了因子的区分度和稳定性。

因子合成

最优化复合IR加权

以历史一段时间的复合因子平均IC值作为对复合因子下一期IC值的估计，以历史IC值的协方差矩阵作为对复合因子下一期波动率的估计，根据IC_IR=IC的期望/Ic的标准差，可以得到最大化复合因子IC_IR的最优权重解。Python优化求解。

最优化复合 IR 加权是指以最优化复合因子 F 的 IR 后得到的最优化因子权重进行加权。给定各因子的 IC 均值向量 $\overline{IC} = (\overline{IC_{f_1}}, \overline{IC_{f_2}}, \dots, \overline{IC_{f_M}})'$，IC 的协方差矩阵 $\Sigma_{IC} = (\rho_{ij,IC})_{i,j=1}^{M}$，复合因子的 IR 为

$$IR_F = \frac{V' \cdot \overline{IC}}{\sqrt{V' \cdot \Sigma_{IC} \cdot V}}$$

对 V 求偏导数，令偏导为 0，可以解出最优化权重的解为

$$V^* = s \sum_{IC}^{-1} \overline{IC}$$

其中 s 是任意常数。这种加权方式考虑了因子的有效性，同时也考虑了加权的因子间的相关性。

因子合成

半衰期IC加权法

因子具有动量效应，因子近期的IC对于当期权重的影响比远期IC的影响更大，因而因子近期的IC要分配更大的权重。
半衰期：给定半衰期H，每隔H期的权重值以指数下降的方式降低一半。

我们的一种直观的想法是 IC 均值加权方式相当于为过去每期的因子 IC 等权分配权重，即给定因子 f_i 过去 N 期的因子 IC 向量 $IC_{f_i}=(IC_{f_i}^1,IC_{f_i}^2,\dots,IC_{f_i}^N)'$，因子 f_i 的权重为

$$v_{f_i}=w_1\cdot IC_{f_i}^1+w_2\cdot IC_{f_i}^2+\cdots+w_N\cdot IC_{f_i}^N=W_N\cdot IC_{f_i}$$

其中 $W_N=(w_1,w_2,\dots,w_N)=\left(\frac{1}{N},\frac{1}{N},\dots,\frac{1}{N}\right)$。

因子具有动量效应，因子近期的 IC 对于当期权重的影响比远期 IC 的影响更大，因而因子近期的 IC 要分配更大的权重，即右偏的 W_N 权重向量，如此能够更加适应市场短期的变化。

多因子模型构建

1 大类因子分析和共线性分析

2 因子合成

3 模型构建：排序打分法

4 模型构建：截面回归法

模型构建：打分法

1.因子打分

- 将某个因子标准化后的值进行排序，均分成n个等级，正向有效的因子从1到n赋分， 负向有效的因子从n到1赋分

- 优点:简单容易理解操作性强
- 缺点:相关因子重复

- 实证中的一些特殊处理:
  (1)PE等估值指标为负时，直接打0分
  (2)某些行业的某些股票不存在某些因子，直接给中档打分

模型构建：打分法

2. 因子权重分配

- 方法一：
- 【静态法】每一期模型都会为每个因子赋于同样的权重
- 优点：稳定、换手率低
- 缺点：缺少动态收益变化

- 方法二：
- 【动态法】按照因子收益率或者IC值调整权重，类似之前提的5种方法分配权重
- 3. 计算总分 挑选个股

3.计算总分 挑选个股

- 配权重的最终目的是选择超配组合和低配组合
- 根据对一个股票池里所有股票打分，我们把总分靠前的20%股票作为超配组合，得分靠后的20%组合作为低配组合。

多因子模型构建

模型构建：回归法

- 多元线性回归，我们的目的是通过T期的因子收益率和因子暴露，预测T+1期的股票收益率
- 满足经典线性模型的基本假设：线性模型、随机样本、无完全共线性、零条件均值、同方差、
- 残差独立服从正态分布

1.T期因子的收益率估计

$$\begin{bmatrix} r_1^{t+1} \\ r_2^{t+1} \\ \cdots \\ r_N^{t+1} \end{bmatrix} = \begin{bmatrix} x^t_{11}, x^t_{12,\ldots,} x^t_{1k} \\ x^t_{21}, x^t_{22,\ldots,} x^t_{2k} \\ \cdots\cdots \\ x^t_{N1}, x^t_{N2,\ldots,} x^t_{Nk} \end{bmatrix} \begin{bmatrix} \hat{f}_1^t \\ \hat{f}_2^t \\ \cdots \\ \hat{f}_k^t \end{bmatrix} + \begin{bmatrix} \hat{u}_1^t \\ \hat{u}_2^t \\ \cdots \\ \hat{u}_N^t \end{bmatrix}$$

r_N^{t+1} 股票N在t+1期的股票收益率

x_{Nk}^t 股票N在t期在因子k上的因子暴露值

$\hat{f}_k^t$ 估计出的因子k的在t期的因子收益率

模型构建：回归法

2.通过前T期因子收益率，估计出T+1期的因子收益率

- 方法一：历史均值法
- 用前N期因子历史收益率的均值作为$T+1$期因子的预期收益率， N通常=36 or 60

$$\widetilde{f_k^{T+1}} = \frac{\sum_{t=T-N+1}^{T} \widetilde{f_k^t}}{N}$$

- 方法二：指数加权移动平均法 EWMA
- 因子收益率包含的信息可能也存在衰减，离当前越近的观测值权重越重

$$EWMA(t) = \lambda * \widetilde{f_k^t} + (1-\lambda) * EWMA(t-1), \quad t = 1, 2, 3, \cdots, n$$

$$\widetilde{f_k^{T+1}} = EWMA(t)$$

$EWMA(t)$: t时刻的修正估计量

$\widetilde{f_k^t}$: t时刻因子收益率观察值

λ: 权重因子

$0 < \lambda < 1$，λ越接近1，则当前观察值权重越大，之前的历史值权重越小。

模型构建：回归法

3.T+2期股票收益率预测

估计出T+1期的因子收益率（$\hat{f}_1^{t+1}, \hat{f}_2^{t+1}, \dots, \hat{f}_k^{t+1}$），并得到T+1期的因子暴露矩阵，计算得到T+2期限的股票收益率($\hat{r}_1^{t+2}, \hat{r}_2^{t+2}, \dots, \hat{r}_k^{t+2}$)

$$\begin{bmatrix} \hat{r}_1^{t+2} \\ \hat{r}_2^{t+2} \\ \cdots \\ \hat{r}_N^{t+2} \end{bmatrix} = \begin{bmatrix} x_{11}^{t+1}, x_{12}^{t+1}, \dots, x_{1k}^{t+1} \\ x_{21}^{t+1}, x_{22}^{t+1}, \dots, x_{2k}^{t+1} \\ \cdots\cdots \\ x_{N1}^{t+1}, x_{N2}^{t+1}, \dots, x_{Nk}^{t+1} \end{bmatrix} \begin{bmatrix} \hat{f}_1^{t+1} \\ \hat{f}_2^{t+1} \\ \cdots \\ \hat{f}_k^{t+1} \end{bmatrix}$$

模型构建：回归法

例子

1.【目前的时刻是2019年9月底】2019年8月底的因子暴露值作为自变量，9月的股票收益率作为因变量回归，计算8月因子收益率（之前的时刻也如此操作，假设从2019年1月开始， 那么一共是8期）
2. 根据前面的2种方法预测9月的因子收益率，乘以9月底的因子暴露值，得到10月的 股票收益率预测值
3. 挑选收益率最高的前20%作为超配组合
4.【目前的时刻是2019年10月底】2019年9月底的因子暴露值作为自变量，10月的股票收益率作为因变量，如此循环

多因子模型Ⅱ

选股有效性评价

- 评价方法
  - Normal IC
  - 基于回归系数T的Barra评价方法
  - 基于随机选股的策略评价体系

选股有效性评价

Normal IC

- 多因子模型常使用因子截面序列与截面超额收益率序列的向量相关系数 IC 来检验因子的有效程度
- 并通常认定该 IC 序列的绝对值均值 大于某一阈值则将因子认定为阿尔法源

IC（Information coefficient 信息系数）的定义：t期（这里的期一般指的是调仓周期）的因子载荷（因子值）对t+1期的收益预测值和实际收益之间的相关系数：

$$IC_A=correlation(f_A,r)$$

IC_A:因子A在改期的IC值，f_A：t期因子A对T+1期收益率的预测值向量，r：t+1期股票实际收益率向量

很多文献中是直接用t期的因子值来当作这个对t+1期的收益预测值，也就是直接计算t期因子值与t+1期的收益率之间的相关系数，作为IC值。

缺陷：单纯的 IC 序列绝对值均值无法判定因子的稳定性，且单纯的 IC 法无法检验因子之间的共线性问题

选股有效性评价

回归法

- 回归法是最常用于检验因子有效性的方法
- 将T期因子的暴露度与T+1期的股票收益率进行回归，所得的回归系数即为T期的因子收益率

缺失值处理：

- 在进行回归法分析时，除了在对数据进行标准化及离群值处理，还需要对因子的缺失值进行填补，从而提升回归结果的可信度
- 常用的方法是设为0、均值、上下数据、插值法，和算法拟合进行填充

基于回归系数T，BARRA模型从因子对收益率影响的显著程度、稳定性以及因子之间共线性问 题的角度着手，给出了更为完整的因子有效性检验的标准：

1） 单因子回归方程系数T检验值的绝对值均值，通常该值大于2认为是理想的结果，表明因子对收益率的影响显著性程度较高

2） 单因子回归方程系数T检验值绝对值序列大于2的占比，该值用以解释在测试时间范围内，因子显著性程度的分布特征

选股有效性评价

3） 年化因子收益率，该值表明因子对收益率的贡献程度，取年化的原 因在于可以与策略年化收益率有个较为直观的比较

4） 年化因子收益的波动率，该值表明因子对收益率贡献的波动程度， 取年化的原因同样在于可以与策略年化收益的波动率有个较为直观的比较

5） 因子收益率比因子收益波动率，该值衡量了经因子收益波动调整后 的因子收益率情况，是考察因子稳定性的指标

6） 因子收益率与基准收益率的相关性，该值检验因子收益是否与基准 收益率有较高相关性，原则上相关性越低的结果越为理想（我们一 般采用沪深 300 指数的收益率作为比较基准）

7） 因子自稳定性系数（Factor Stability Coeff），该值检验因子收益 率的稳定性，计算公式为：

$$\rho_{kt}=\frac{\sum_n v_n^t(x_{nk}^t-\overline{x}_{nk}^t)(x_{nk}^{t+1}-\overline{x}_{nk}^{t+1})}{\sqrt{\sum_n v_n^t(x_{nk}^t-\overline{x}_{nk}^t)^2}\sqrt{\sum_n v_n^t(x_{nk}^{t+1}-\overline{x}_{nk}^{t+1})^2}},$$

其中Vn为股票的回归加权权重

选股有效性评价

8） 因子方差膨胀系数VIF 值（Variance Inflation Factor），VFI 是利用需要检验的因子作为应变量，已通过检验的因子作为自变量，（每个因子和其他所有因子回归）

构建多元方程，并计算回归方程的R2：

$$X_{nk} = \sum_{k' \neq k} X_{nk'} b_{k'} + \varepsilon_{nk}$$

然后，VIF 的计算公式为：

$$VIF_k = \frac{1}{1 - R_k^2}$$

通常意义上，VIF 值越大则表明被检验因子与其他因子的共线性程度越 高。经验表明，当VIF 值大于 3 时，该因子的共线性程度较高，应该拒绝纳入风险因子范围。

“运气”还是“能力”

基于随机选股的策略评价体系

1.设某个策略组合的收益率序列为$r_t(P)$，$t=1,\dots,T$。S是包含N个股票的股票池。

2. 从 S 中随机选取 k（通常等于策略组合中的股票个数）个股票形成组合，并重 复 M 次。

3. 计算第i（i=1,…,M）个随机组合在第t个持有期上的收益，记为$r_t(P_i)$。

4. 记α为一个很小的概率（通常取0.1，0.05），$r_t(P_{1-\alpha})$为这M个随机组合收益率序 列的 1-α分位数。

5. 若$r_t(P) > r_t(P_{1-\alpha})$，认为“选股策略有效”

- 策略是否“有效”的结论是通过比较策略组合的收益和随机组合收 益的 1-α分位数得到的，而非简单的均值或中位数。 $r_t(P_{1-\alpha})$代表的是一个可以被认为“有 效”的策略应当满足的最低收益水平。

选股有效性评价

在每一个持有期t内，上述方法能够方便地判断在某一个α下，策略是否有效。但是， 对策略的信任与否仅凭一次检验是远远不够的。更多时候，策略开发者面对的是一个较长回测期上的评价工作。为了给出更为全面而准确的认识，此处提供几个更为直观的指标：

“有效”率（PO）

回测期上，策略组合的收益超越随机组合收益率序列的 1-α 分位数的比例 或者说，策略组合的收益所对应的分位数大于 1-α 的比例：

$$PO = \frac{\sum_{t=1}^{T} I(r_t(P) > r_t(P_{1-\alpha}))}{T} \times 100$$

I(x)为阶跃函数，x>0时I(x)取值为1，x<0时，I(x)取值为0

分位数跟踪误差（QTE）

$$QTE = \frac{\sum_{t=1}^{T} (r_t(P) - r_t(P_{1-\alpha}))}{T}$$

选股有效性评价

该指标反映的是，平均意义上，策略组合的收益高于 1-α 分位数的大小。 可以期望，一个比较有效的选股策略，其 QTE 应当大于零。

分位数跟踪误差的波动率（QTEV）

在极端情况下，仅依赖平均跟踪误差不足以得到可靠的结论，提供其波动率的大小也是非常重要的。分位数跟踪误差的波动率定义为：

$$QTEV = \frac{\sum_{t=1}^{T}(r_t(P) - r_t(P_{1-\alpha}))^2}{T}$$

‘运气’还是‘能力’?

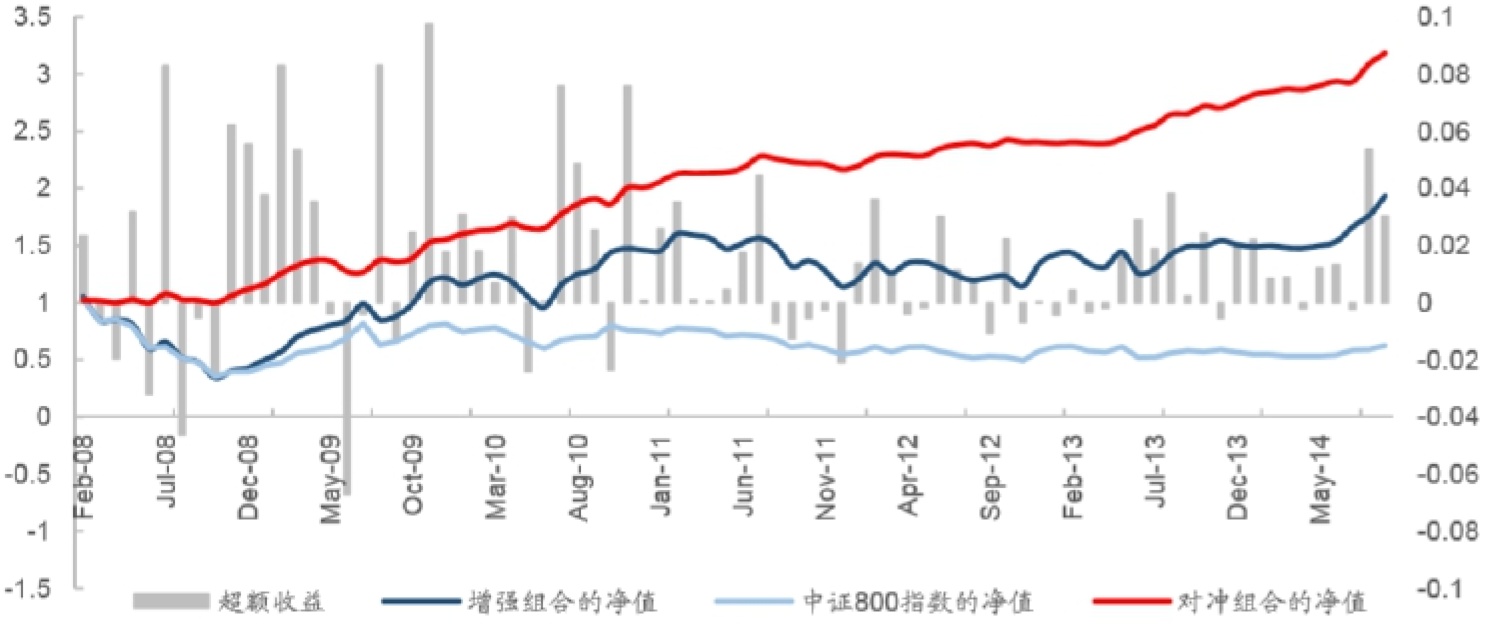
实证分析

- 以一个中证 800 对冲策略为例，通过多因子打分排序在 800 个指数成分股中挑选 100 个形成组合，每个自然月换仓。

表 1 中证 800 对冲策略的收益-风险特征

胜率	信息比	年化收益	最大回撤	盈亏比
66.25%	1.733	12.41%	7.51	4.31

图 1 中证 800 对冲策略的收益净值曲线



- 从表 1 中策略的收益-风险指标来看，轻易否认其有效性似乎也没有十足的根据。 但是，这66%的胜率和优秀的业绩中，是否存在运气成分？比例是多少？策略和随机选股相比又有多大优势？

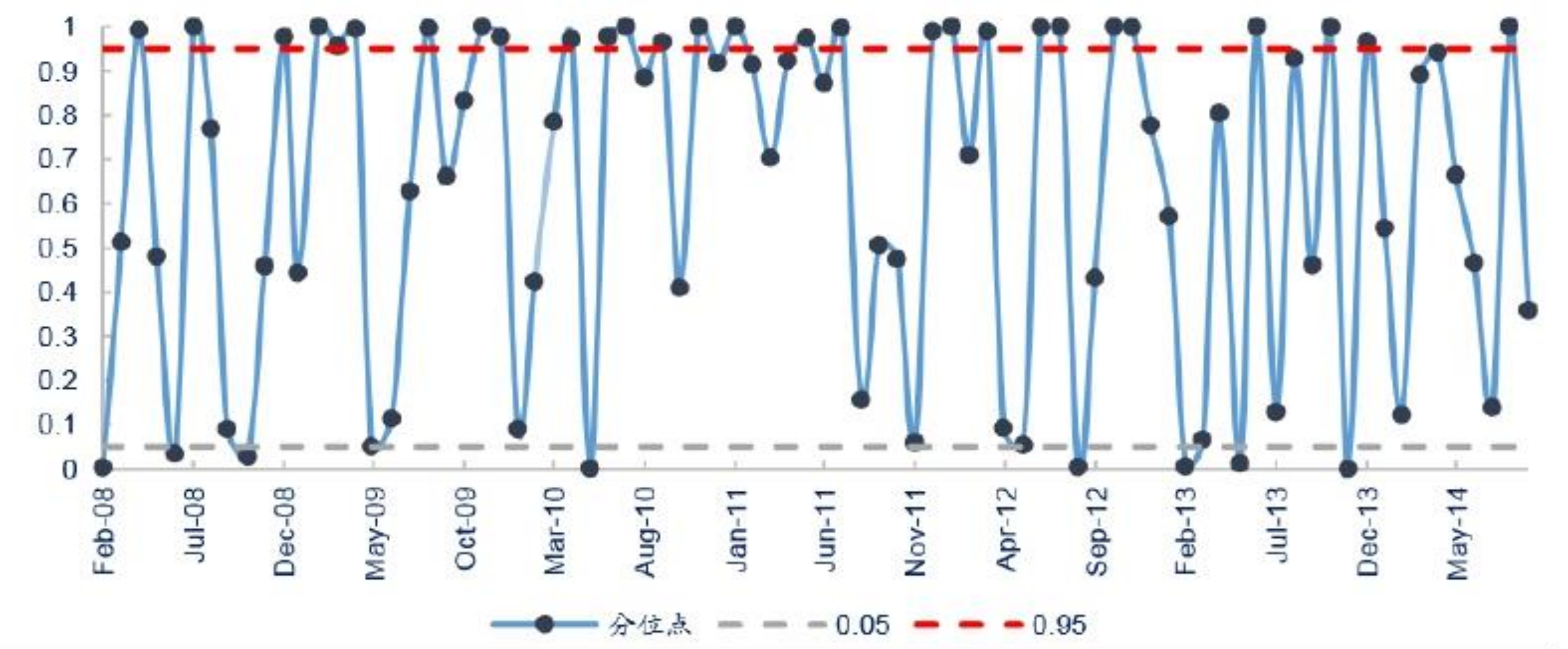
数据来源：海通证券专题报告_选股有效性评价

‘运气’还是‘能力’？

实证分析

先确定随机组合的大小 k 及重复次数 M。为了保证比较的一致性，选择 k=100。 对于 M 的取值，经验表明，中等程度的重复足以保证良好的精度。为方便计算，此处取 M=1000。每个随机组合的加权方式与对冲组合相同。

图 2 中证 800 对冲策略月度收益率的分位点



- 左图2展示的是每个月的对冲组合在随机组合收益率序列中对应的分位点。红色虚线 对应 95%（α=0.05）的阈值，在其之上的点表示策略在该月份是有效的。

数据来源：海通证券研究所

“运气”还是“能力”？

实证分析

下表 2 展示的是两个不同的阈值水平下，各指标的取值。

阈值	PO	QTE	QTEV
95%（α=0.05）	35%	-0.709%	0.0408%
90%（α=0.10）	45%	-0.366%	0.0363%

注：***表示在 **0.001** 的水平下显著，**表示在 **0.01** 的水平下显著

上表可以进一步说明该对冲策略中运气成分的占比情况。当选择 95%作为阈值时，“有效”率（PO）为 35%，即，在回测的 80 个月中，策略有 28 期展现了选股能力。但从平均的跟踪误差来看，对冲策略的收益显著地小于随机组合 90%和 95%的分位数。此策略的优异表现有较大运气成分。

避免‘唯指标论’：和经典的夏普比率一样，我们通常需要将 **PO**、**QTE** 作为对策略、模型甚至是选股因子的一个排序指标，以相互比较的方式来 获得客观的评价

数据来源：海通证券研究所

多因子模型Ⅱ

组合优化

组合构建的主流方法

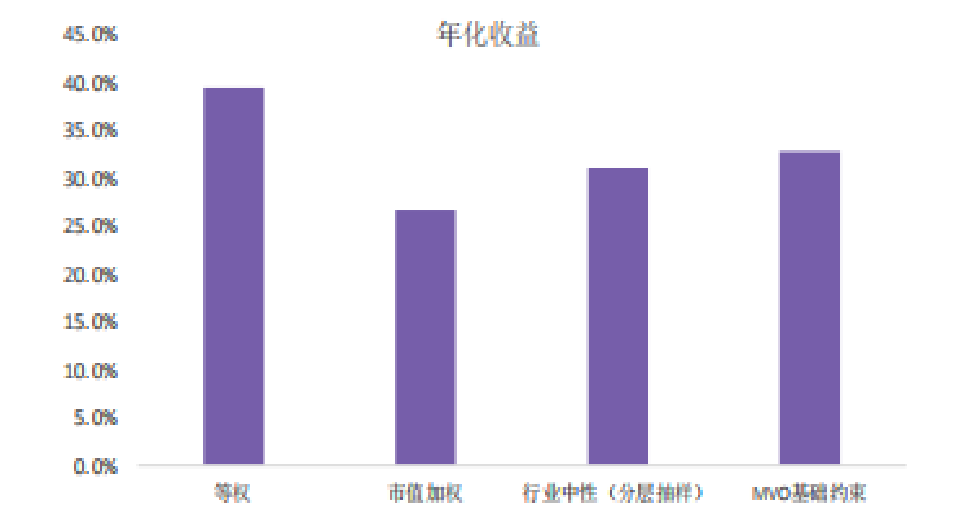
- 等权配置
- 等权持有组合内的所有股票
- 由于小市值股票的权重与大市值股票保持了一致，等权组合会暴露小市值因子
- 仅适用于纯多头产品，对于指数增强或者要求限制跟踪误差的产品不适用

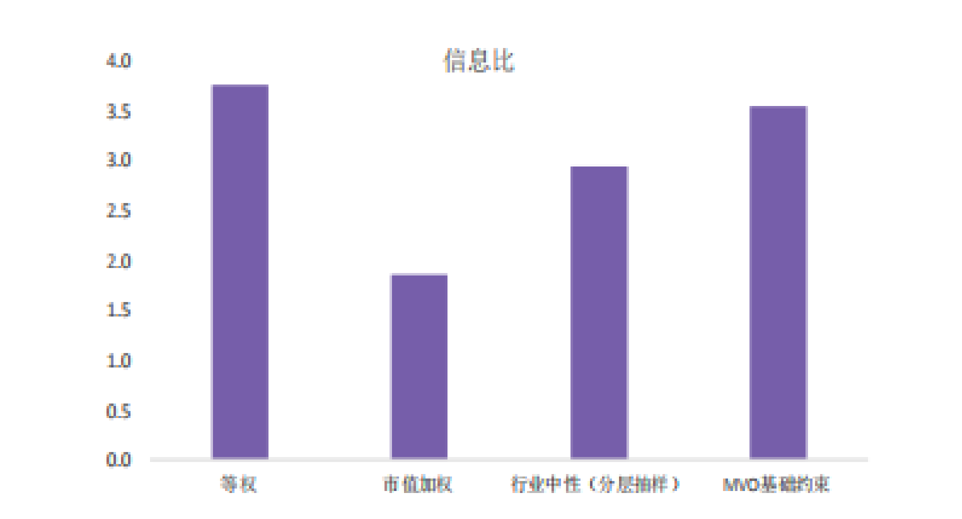
- 市值加权
- 弥补等权组合对于小市值风格暴露的问题
- 可以在一定程度上减少小市值因子暴露

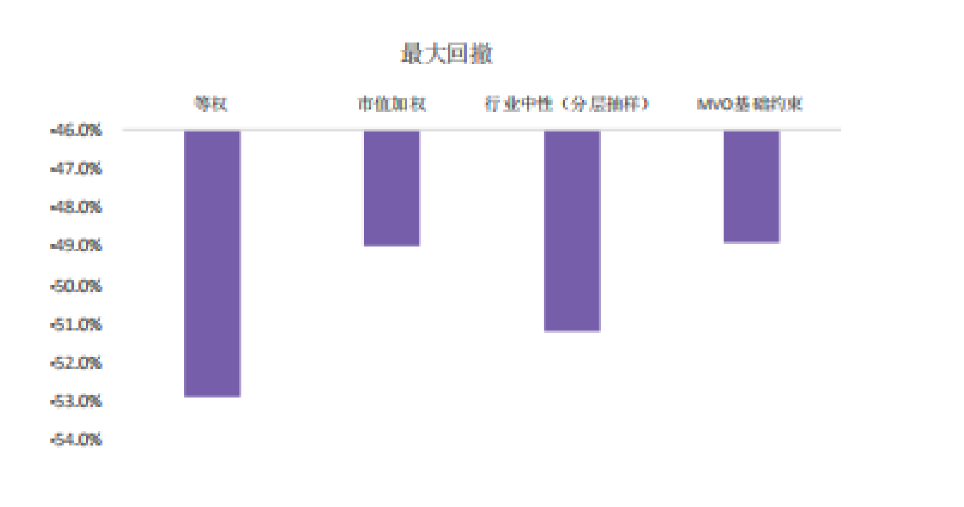
- 分层法
- 常见分层法为行业分层法，在各个一级行业内分别按照股票的多因子得分排序，选取前10%的股票，股票合集作为最终持仓标的，并等权持有
- 可以减少投资组合在某一风格上的暴露

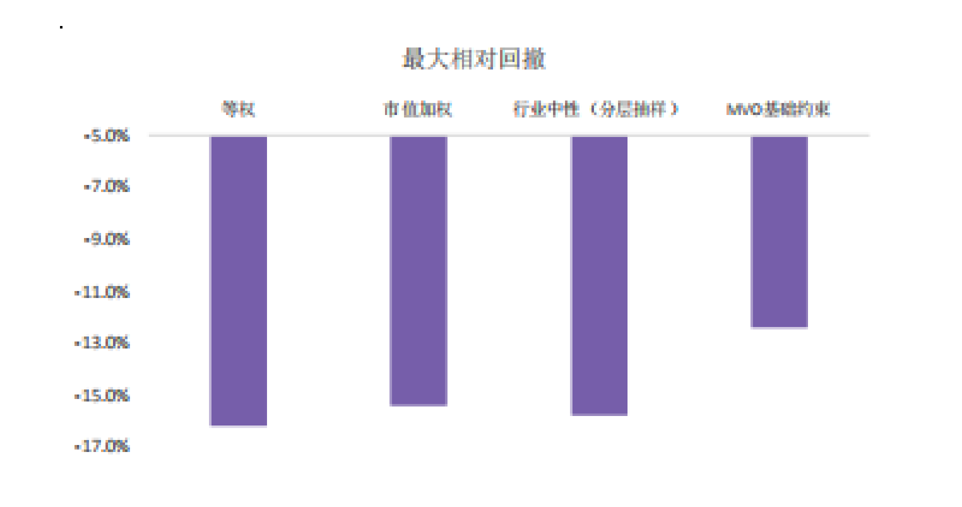
组合优化

几种组合构建方法的比较









组合优化

组合优化的作用

- 较为系统和准确得控制投资组合中的预期收益、风险暴露、个股权重等
- 优化模型：二次规划

优化模型构建

- 目标函数
- 收益目标：将预期风险控制在一定水平之下，选择投资组合使得期望收益最大
- 风险目标：在预期收益不低于某一特定水平之下，选择投资组合使得预期风险最小
- 同时考虑收益和风险：使用风险对收益进行调整

- 约束条件
- 个股最高权重限制
- 行业中性
- 因子暴露约束
- 换手率限制

组合优化——目标函数

投资组合预期收益率

- Step 1: 通过多因子模型回归得到所有因子的历史收益率序列
- Step 2: 估计因子在T+1时期的预期收益（历史平均法、指数加权移动平均法等）
- Step 3: 计算股票预期收益 （因子收益率向量*因子载荷矩阵）
- Step 4: 表示组合预期收益率

股票预期收益率

$$(\widetilde{f_1^{T+1}}, \widetilde{f_2^{T+1}}, \cdots, \widetilde{f_k^{T+1}}) \quad X^{T+1} = \begin{bmatrix} X_{11}^{T+1} & X_{12}^{T+1} & \cdots & X_{1K}^{T+1} \\ X_{21}^{T+1} & X_{22}^{T+1} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1}^{T+1} & \cdots & \cdots & X_{NK}^{T+1} \end{bmatrix}$$

$$\widetilde{r_j^{T+1}} = \sum_{k=1}^{K} X_{jk}^{T+1} * \widetilde{f_k^{T+1}}$$

组合预期收益率

$$\widetilde{r_P^{T+1}} = \sum_{j=1}^{N} \widetilde{r_j^{T+1}} * h_j$$

组合优化——目标函数

投资组合预期风险

- Step 1: 使用每期因子收益的历史序列值，计算因子协方差矩阵
- Step 2: 估计残差风险
- Step 3: $\sigma_P^2 = x_P^T * F * x_P + h_P^T * \Delta * h_P = h_P^T * V * h_P$

多因子模型的本质是将对于N只股票的收益-风险预测转变成对于K个因子的收益-风险预测

$$\begin{bmatrix} \tilde{r}_1 \\ \tilde{r}_2 \\ \cdots \\ \tilde{r}_N \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1K} \\ X_{21} & X_{22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ X_{N1} & \cdots & \cdots & X_{NK} \end{bmatrix} \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \vdots \\ \tilde{f}_K \end{bmatrix} + \begin{bmatrix} \tilde{u}_1 \\ \tilde{u}_2 \\ \vdots \\ \tilde{u}_N \end{bmatrix}$$

X_{nk}: 股票n对因子k的风险暴露（因子载荷）

假设残差收益率与因子收益率不相关，不同股票之间收益率也互不相关，市场风险结构为

$$V_{i,j} = \sum_{k1,k2=1}^{K} X_{i,k1} * F_{k1,k2} * X_{j,k2} + \Delta_{i,j}$$

$V_{i,j}$: 股票i和股票j的协方差

$X_{i,k1}$: 股票i对因子$k1$的暴露度（因子载荷）

$F_{k1,k2}$: 因子$k1$和因子$k2$之间的收益率协方差

$\Delta_{i,j}$: 股票i和股票j之间残差的协方差，$i \neq j$时为 0

组合优化：目标函数

控制风险，最大化收益模型

$$\max_{h_j} \sum_{j=1}^{N} \widetilde{r_j^{T+1}} * h_j$$

$$s.t. \quad h_p^T * V * h_p \leq \sigma^2$$

$$\sum_{j=1}^{N} h_j = 1, h_j \geq 0 \quad j = 1,2,\cdots,N$$

保证收益，最小化风险模型

$$\min h_p^T * V * h_p$$

$$s.t. \quad \sum_{j=1}^{N} \widetilde{r_j^{T+1}} * h_j \geq r$$

$$\sum_{j=1}^{N} h_j = 1, h_j \geq 0 \quad j = 1,2,\cdots,N$$

Q：其他种类的目标函数
最大化组合信息比例等等

组合优化：均值-方差优化模型

均值方差优化

- 马科维茨均值方差优化模型的目标函数为最大化投资者效用函数
- 经典算法中风险度量由个股收益的协方差矩阵表示，由于个股数量多，协方差矩阵会出现奇异矩阵
- 使用因子进行降维

$$\text{Max } F(w) = w^T\mu - \lambda w^T\Sigma w/2$$

其中，

λ代表风险厌恶系数；

w = (w, w_2, … , w_n)代表待求解的各标的权重向量；

μ代表各标的预期收益率向量；

Σ代表各标的收益率的协方差矩阵。

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} * f_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} * f_2 + \cdots + \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix} * f_k + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

$$R_p = \sum_{i=1}^{n} w_i \left(\sum_{k=1}^{K} x_k f_{ik} + u_i \right)$$

$$\sigma_p = \sqrt{w^T (X \Sigma X^T + \Delta) w}$$

组合优化：约束条件

个股最高权重限制

- 上限：多因子模型本质是统计套利模型，并且强调投资的宽度（通过多个不同维度的因子），因此必须对个股的权重上限进行约束，避免风险在单只股票上分配过多的权重
- 下限：个股权重的下限约束是0，及不允许卖空

如果组合P存在业绩基准B，业绩基准B的个股权重向量h_B，那么组合P的主动权重暴露h_{PA}是：

$$h_{PA} = h_P - h_B$$

对于h_{PA}而言，个股权重的下限约束向量则变成$-h_B$。假设股票j在基准B中的权重为h_{Bj}，由于在组合P中的权重下限是0，其主动权重暴露就是$-h_{Bj}$。

$$\sum_{j=1}^{N} h_{PAj} = 0, h_j^{upper} \geq h_{PAj} \geq -h_{Bj} \quad j = 1,2,\cdots,N$$

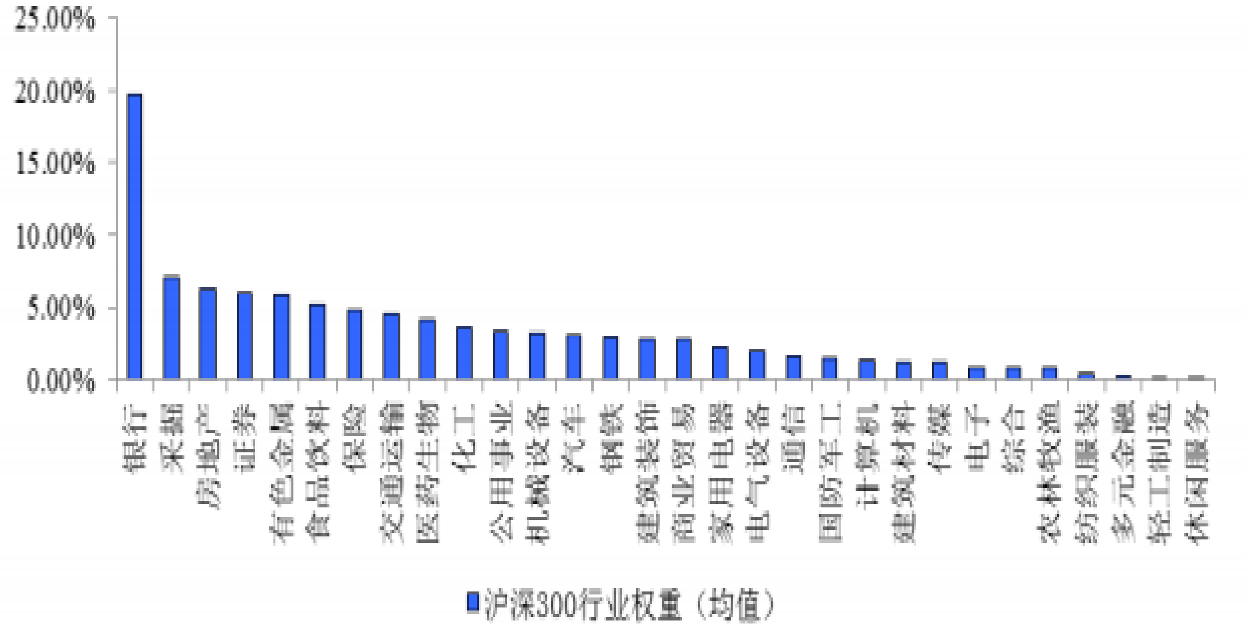
组合优化：约束条件

行业中性限制

- 多头组合的行业配置与对冲基准的行业配置相一致
- 行业中性配置的目的在于剔除行业因子对策略收益的影响
- 与传统观念不同，传统行业配置试图找到在未来某一段时间内强势因子予以超配，弱势因子予以低配，而行业中性的特点在于剔除行业层面的影响，仅考虑行业内部个股的超额收益
- 假设S为样本股票的行业哑变量矩阵，那么行业中性的约束为

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1S} \\ s_{21} & s_{22} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ s_{N1} & \cdots & \cdots & s_{NS} \end{bmatrix}$$

$$\sum_{j=1}^{N} h_{PAj} * s_{ji} = 0$$



组合优化：约束条件

因子暴露约束

- 避免在单个因子上暴露过多的风险
- 组合P在因子k上的暴露度计算公式为

$$\sum_{j=1}^{N} h_{PAj} * X_{jk}$$

- 如果对于因子K的暴露上限为X_k

$$\left|\sum_{j=1}^{N} h_{PAj} * X_{jk}\right| \leq x_k$$

换手率约束

- 出于增强投资策略稳定性，会对组合换手率有一定限制：

$$|h_{PA} - h_{PA}{}^{o}| \leq h$$

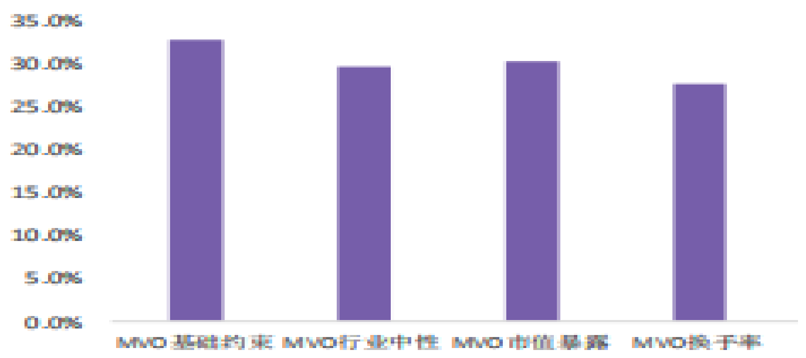
Q：其他约束条件

风格因子中性（风格特征匹配，追求稳健的阿尔法收益，而非市场某种风格的收益）等

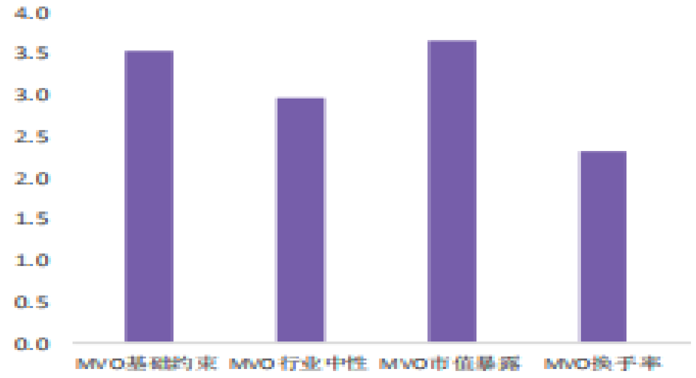
组合优化

不同约束条件的比较

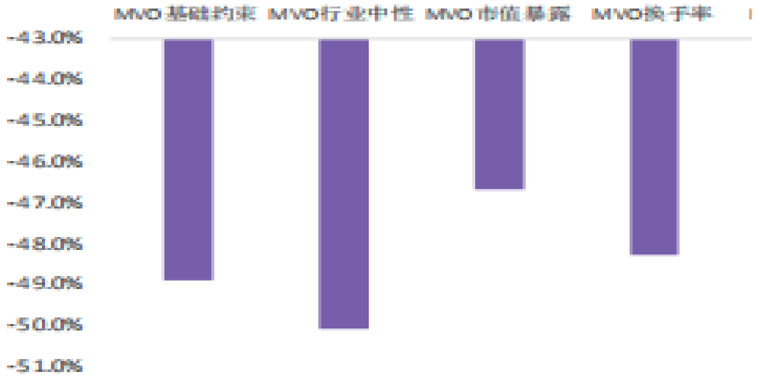
年化收益



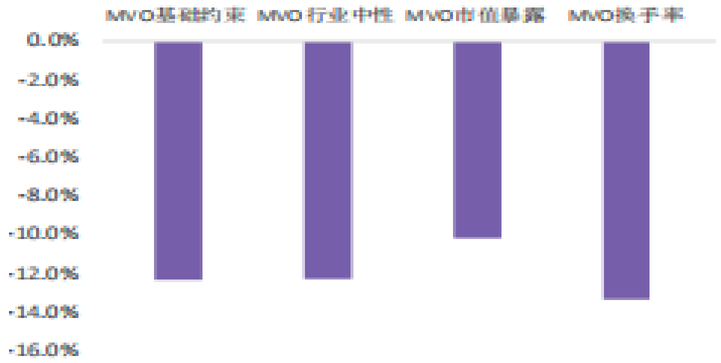
信息比



最大回撤



最大相对回撤



参考资料

组合优化

- 华泰证券-多因子系列之一：华泰多因子模型体系初探
- 光大证券：组合优化算法探析及指数增强实证
- 国泰君安：基于组合权重优化的风格中性多因子选股策略
- 广发证券：多因子Alpha模型探究：沪深300成份股的应用分析

北京大学量化交易协会