指数复制

1 指数复制目的

构建一个组合，使得该组合与指数的未来走势差距（跟踪误差）最小。

2 指数复制方法

2.1 完全复制

是指按照指数中成分股的配比，构建组合。

优点：如果没有交易成本，该组合的跟踪误差为 0。

缺点：现实当中，有交易成本，完全复制交易的股票数目较多，导致交易成本比较高，所以跟踪误差增加。另外，现实当中，我们能够用来构造组合的股票会受到各种限制，不一定能够购买指数的所有成分股。

2.2 分层抽样复制

分层抽样复制，是将指数成分股按照某一属性（比如行业）划分成几类，每一类当中选取一只股票，然后将这些股票组合起来构建组合（该只股票在组合当中的权重设为该只股票所属类型的成分股在指数当中的权重）。其目的是相对于完全复制使用更少的成分股从而减小交易成本。但是抽样复制相对于完全复制，不考虑交易成本的跟踪误差要更大一些。

分层抽样复制的潜在假设是成分股的收益率结构为如下多因子结构

$$r_{i,t+1} = x_{i1t} f_{1,t+1} + \ldots + x_{iKt} f_{K,t+1} + \varepsilon_{i,t+1} \qquad (0.1)$$

，其中 K 为类型数目，$f_{k,t+1}$ 为第 k 类的因子收益率，x_{ikt} 为股票 i 在 t 时刻在因子 k 上的暴露，如果股票 i 属于第 k 类，则 $x_{ikt} = 1, x_{ijt} = 0 \ for \ j \neq k$。这样的话，我们构建的组合只要在共同因子上的暴露与指数在共同因子上的暴露相同，则我们的组合很可能对指数的跟踪误差较小[1]。

可以看到，上述假设与我们常用的风险模型有一些差异。对于行业因子上述模型与我们常用的风险模型相同。对于市值因子，常用风险模型其因子暴露是一个连续变量，而上述模型是将股票池按照市值分层几份（假设有三份），则会形成三个因子。

所以对于上述操作操作扩展的一个思路是，我们使用常用的风险模型，构造组合的时候约束组合在共同风险因子上的暴露与指数的接近，同时约束用到的股票数目不超过一定的数目。

2.3 优化抽样复制

优化抽样复制是通过最小化历史数据的跟踪误差来决定要选取的股票以及选取的股票的权重。

分层抽样和优化抽样都是选取部分股票，所以相对于完全复制法，交易成本要小。既然两种方法都选取了部分股票，那么在没有交易成本的情况下，两种方法相对于完全复制跟踪误差要较大。为了减小跟踪误差，分层抽样法是假设股票的收益率模型为(0.1)结构，基于这样的结构来选取股票以及权重；而优化抽样法是假设能够在历史数据上最小化跟踪误差的组合也能在接下来一段时间较好地跟踪指数[1,2]。

具体地[2]，假设我们当前处在T时刻，我们所具有的数据为0 到T 时刻的指数价格和0 到T 时刻的N 只股票的价格。我们的目标是选出最好的K 只股票，及其权重。

2.3.1 我们用到的参数和数据为：

ε_i 股票i 的持有比例下限，

δ_i 股票i 的持有比例上限，

X_i 当前持有的股票i 的股数

V_{it} 股票i 在t 时刻的价格

I_t 指数在t 时刻的价格

R_t 指数在$t-1$ 到t 时刻间的收益率，为$\log\left(I_t / I_{t-1}\right)$ 或$I_t / I_{t-1} - 1$

r_{it} 股票i 在$t-1$ 到t 时刻间的收益率，为$\log\left(V_{it} / V_{i,t-1}\right)$ 或$V_{it} / V_{i,t-1} - 1$

C T 时刻我们持有的总资产值，为当前持有的股票价值之和加上当前持有的现金

f_i^b 购买股票i 的成比例成本

f_i^s 出售股票i 的成比例成本

F_i^b 购买股票i 的固定成本

F_i^s 出售股票i 的固定成本

M_i^b 股票i 允许的最大购买量

M_i^s 股票i 允许的最大出售量

γ 总交易成本占当前总资产值的比例上限

2.3.2 **我们的决策变量为**：

x_i 我们将要持有的股票i 的股数，非负数

z_i 我们将是否持有股票i ，持有为 1，不持有为 0

α_i^b 我们将是否购买股票i ，购买为 1，不购买为 0

α_i^s 我们将是否出售股票i ，出售为 1，不购买为 0

y_i^b 我们将要购买的股票i 的股数，非负数

y_i^s 我们将要出售的股票i 的股数，非负数。

2.3.3 我们的**约束**为：

$$\sum_{i=1}^{N} z_{iT} = K \quad (0.2)$$

$$\varepsilon_i z_i \le V_{it} x_{it} / C \le \delta_i z_i, i = 1, \ldots, N \quad (0.3)$$

$$x_i = X_i + y_i^b - y_i^s, i = 1, \ldots, N \quad (0.4)$$

$$\alpha_i^b + \alpha_i^s \le 1, i = 1,\ldots,N \qquad (0.5)$$

$$y_i^b \le M_i^b \alpha_i^b, i = 1,\ldots,N \qquad (0.6)$$

$$y_i^s \le \min\left(X_i, M_i^s\right)\alpha_i^s, i = 1,\ldots,N \qquad (0.7)$$

$$\sum_{i=1}^{N} V_{iT} x_i = C - \sum_{i=1}^{N}\left(f_i^b V_{iT} y_i^b + f_i^s V_{iT} y_i^s + F_i^b \alpha_i^b + F_i^s \alpha_i^s\right) \qquad (0.8)$$

$$\sum_{i=1}^{N}\left(f_i^b V_{iT} y_i^b + f_i^s V_{iT} y_i^s + F_i^b \alpha_i^b + F_i^s \alpha_i^s\right) \le \gamma C \qquad (0.9)$$

$$y_i^b, y_i^s, x_i \ge 0, i = 1,\ldots,N \qquad (0.10)$$

$$\alpha_i^b, \alpha_i^s, z_i \in \{0,1\}, i = 1,\ldots,N \qquad (0.11)$$

然后，我们定义一个目标函数，通过受上述约束最优化该目标函数来求解决策变量。目标函数一般是跟踪误差的函数，所以先定义跟踪误差。

2.3.4 **跟踪误差**为[3]：

2.3.4.1 Return-based

t 时刻跟踪误差为t 时刻组合的收益率减去指数的收益率，即

$$TE_t = r_{pt} - R_t \qquad (0.12)$$

，其中组合收益率为

$$r_{pt} = \sum_{i=1}^{N} w_{it-1} r_{it} \qquad (0.13)$$

，其中

$$w_{it} = \frac{x_i V_{it}}{\sum_{j=1}^{N} x_j V_{jt}} \qquad (0.14)$$

。

可以看到，跟踪误差TE_t是决策变量$x_i, i=1,\ldots,N$的非线性函数。

如果不做简化，最优化会比较难求解，因为目标函数为自变量的非凸函数。所以一般做假设常数权重，即w_{it}不随时间变化，是常数w_i，所以其也等于w_{iT}，而w_{iT}为

$$w_{iT}=\frac{x_i V_{iT}}{\sum_{j=1}^{N} x_j V_{jT}} \quad (0.15)$$

，如果没有交易成本，根据(0.8)式，(0.15)中的分母为T 的总资产量C 。所以假设常数权重，在无交易成本的情况下，TE_t 为决策变量$x_i, i=1,\ldots,N$的线性函数。

而在有交易成本的情况下，要做进一步假设，即假设(0.9)式等号近似成立，则(0.15)的分母为$V_{pT}\equiv\sum_{j=1}^{N} x_j V_{jT}=C-\gamma C$ 也是一个常数。所以，

$$w_{it}=w_{iT}=\frac{x_i V_{iT}}{C-\gamma C} \quad (0.16)$$

。

另外，上述常数权重的假设有可能比较牵强，我们可以做其他假设。非线性主要来自于(0.14)的分母$V_{pt}\equiv\sum_{j=1}^{N} x_j V_{jt}$，我们可以假设组合归一化的市值与指数归一化的市值近似相同，即

$$\frac{V_{pt}}{V_{pT}}=\frac{I_t}{I_T} \quad (0.17)$$

，则

$$w_{it}=\frac{x_iV_{it}}{\sum_{j=1}^{N}x_jV_{jt}}=\frac{x_iV_{it}}{\frac{I_t}{I_T}V_{pT}}=x_i\frac{V_{it}I_T}{I_tV_{pT}}$$
$$=\left(\text{如果要以权重}w_{iT}\text{为决策变量}\right)\frac{V_{pT}\cdot w_{iT}}{V_{iT}}\frac{V_{it}I_T}{I_tV_{pT}}=w_{iT}\frac{V_{it}I_T}{V_{iT}I_t}$$

(0.18)。

2.3.4.2 Value-based

t 时刻跟踪误差为t 时刻组合的归一化的市值减去指数的归一化的市值，即

$$TE_t=\frac{V_{pt}}{V_{pT}}-\frac{I_t}{I_T}\quad(0.19)$$

其中，

$$V_{pt}=\sum_{i=1}^{N}x_iV_{it}\qquad(0.20)$$

，所以，假设没有交易成本，则根据(0.8)$V_{pT}=C$，所以TE_t 为决策变量的线性函数。

而如果有交易成本，我们需要做假设，使得V_{pT}为一个常数，可以同样假设$V_{pT}=C-\gamma C$。

2.3.4.3 **总结一下**

目标函数是跟踪误差TE_t的函数，而TE_t是决策变量的函数，如果TE_t是决策变量的线性函数，那么最优化会比较容易求解。在无交易成本的时候，value-based 不需要做假设，而 return-base 需要假设常数权重；在有交易成本的时候，value-based 需要假设$V_{pT}=C-\gamma C$，而 return-base 需要假设$V_{pT}=C-\gamma C$且（常数权重或组合归一化的市值与指数归一化的市值近似相同）。

2.3.5 **目标函数**[2]：

目标函数一般是跟踪误差的函数。

2.3.5.1 二次函数（MSE）

$$\frac{1}{T}\sum_{t=1}^{T}\left(TE_t\right)^2\qquad(0.21)$$

由于TE_t是决策变量的线性函数，所以目标函数是决策变量的二次函数。最终问题为：

最小化(0.21)

Subject to 约束(0.2)到(0.11)。

该问题为二次规划问题，但是决策变量一部分为连续变量，一部分为 0,1 变量。

2.3.5.2 平均绝对值（MAD）

$$\frac{1}{T}\sum_{t=1}^{T}\left|TE_t\right| \qquad (0.22)$$

对于(0.22)，可以引入额外的决策变量$d_t \geq 0, t=1,\ldots,T$，则原问题变为

$$\min \frac{1}{T}\sum_{t=1}^{T} d_t \qquad (0.23)$$

Subject to:

$$\begin{aligned} &TE_t \leq d_t, t=1,\ldots,T \\ &-TE_t \leq d_t, t=1,\ldots,T \qquad (0.24) \\ &d_t \geq 0, t=1,\ldots,T \end{aligned}$$

以及约束(0.2)到(0.11)。

该问题为线性规划问题，只不过有一部分自变量为连续变量，一部分为 01 变量，该问题称为 mixed integer linear programming，对于该问题有比较成熟的模块求解，比如 CPLEX 和 Gurobi。

2.3.5.3 最大绝对值

$$\max_{t\in\{1,\ldots,T\}}\left|TE_t\right| \qquad (0.25)$$

对于(0.25)，可以引入额外的决策变量$d \geq 0$，则原问题变为

$$\min d \quad (0.26)$$

Subject to:

$$\begin{aligned} & TE_t \le d, t = 1, \ldots, T \\ & -TE_t \le d, t=1, \ldots, T \qquad (0.27) \\ & d \ge 0 \end{aligned}$$

以及约束(0.2)到(0.11)。

同样，该问题称为 mixed integer linear programming，对于该问题有比较成熟的模块求解。

2.3.5.4 一般形式为

$$\frac{1}{T}\left(\sum_{t=1}^{T}\left|TE_t\right|^p\right)^{1/p} \quad (0.28)$$

$p \ge 1$ 。我们甚至可以加上不同期的权重，为

$$\frac{1}{T}\left(\sum_{t=1}^{T}\Delta_t\left|TE_t\right|^p\right)^{1/p} \qquad (0.29)$$

。

2.3.6 评论

上述约束中，要用到每只股票的交易成本，即需要交易成本模型来估计交易成本。一种早期的做法是，只要我构造组合的时候选用较小的股票数目 K，则交易成本相对于完全复制来说就会比较小，至于不同的股票的交易成本有多大，就不建模了，这样的话上述约束中交易成本相关的地方都为 0。

上述提到的 mixed integer linear programming，主要是因为使用了表示股票i 是否在组合里的 01 决策变量。一种早期的做法是，先通过一些 heuristic 来确定哪些股票要在组合中用到，然后只对这些股票的权重做最优化。一些常用的 heuristic 如下[4]：

按市值：即选取市值最大的 K 只股票来做最优化。

Forward selection:使用所有股票做最优化，把权重最大的股票加入待使用 list，然后不带这只股票，再做最优化，这样直到待使用 list 里面有 K 只股票。最后，使用这 K 只股票做最优化。

Backward selection:使用所有股票最最优化，将权重最小的股票去掉，用剩下的股票做最优化，直到只剩下 K 只股票。最后，用这 K 只股票做最优化。

2.4 优化抽样复制扩展

优化抽样复制的假设是在当前时间点构造一个组合，如果这个组合在最近一段历史数据上跟踪误差较小，则该组合在接下来一段时间上跟踪误差较小。我们可以多加入一些我们认为有用的假设。比如，我们可以将分层抽样的假设（风险模型）用在优化抽样上。至少有 2 种结合方式。

2.4.1 将风险模型只用在约束上

我们可以约束组合在一些共同风险因子上的暴露与指数相应暴露的差处于一定范围。

$$\begin{aligned}&\sum_{i=1}^{N} w_{iT} e_{iTk} - e_{ITk} \le thres_1, k = 1, \ldots, K \\ &-\left(\sum_{i=1}^{N} w_{iT} e_{iTk} - e_{ITk}\right) \le thres_1, k = 1, \ldots, K\end{aligned} \quad (0.30)$$

，其中，e_{iTk} 为第 i 只股票在 T 时刻在第 k 个共同因子上的暴露，e_{ITk} 为指数在相应因子上的暴露。其中 w_{iT} 如(0.15)。

2.4.2 用在目标函数上

也可以将目标函数设为下期组合收益率与指数收益率差的方差

$$Var_T\left(r_{p,T+1} - R_{T+1}\right) \quad (0.31)$$

。

而

$$\begin{aligned}&Var_T\left(r_{p,T+1} - R_{T+1}\right) = \\ &\left(w_{1T} - w_{1IT}, \ldots, w_{NT} - w_{NIT}\right) Var_T\left(r_{t+1}\right)\left(w_{1T} - w_{IT}, \ldots, w_{NT} - w_{IT}\right)'\end{aligned} \quad (0.32)$$

，其中，$Var_T\left(r_{T+1}\right)$ 为风险模型得到的 N 只股票收益率 $r_{T+1} = \left(r_{1,T+1}, \ldots, r_{N,T+1}\right)'$ 的方差矩阵。同样，其中 w_{iT} 如(0.15)。所以，该问题目标函数为决策变量的二次函数。

将(0.31)与 return-based 的(0.21)比较，可以看到(0.21)相当于对 $E_T\left(r_{p,T+1}-R_{T+1}\right)^2$ 的样本内估计，而 $E_T\left(r_{p,T+1}-R_{T+1}\right)^2$ 与 $Var_T\left(r_{p,T+1}-R_{T+1}\right)$ 相差 $\left(E_T\left(r_{p,T+1}-R_{T+1}\right)\right)^2$。从指数跟踪的角度看，最小化 $E_T\left(r_{p,T+1}-R_{T+1}\right)^2$ 要比最小化 $Var_T\left(r_{p,T+1}-R_{T+1}\right)$ 要好，因为 $r_{p,T+1}$ 和 R_{T+1} 如果只相差一个常数，则前者不是 0 而后者是 0，所以前者更能体现跟踪误差。但是，以 $E_T\left(r_{p,T+1}-R_{T+1}\right)^2$ 作为目标函数，我们就只能使用历史数据做最优化，而使用 $Var_T\left(r_{p,T+1}-R_{T+1}\right)$ 作为目标函数，我们还能利用到收益率的方差信息（估计方差的时候对收益率结构做一些假设，比如多因子风险模型），类似于正则化的作用，这样找出来的组合可能样本外表现跟稳健。

3 现有程序即改进建议

目前现有的程序是目标函数近似为(0.21)，只不过减去了 TE_t 的均值。即为 return-based，但是现有程序又没有做常数权重或组合归一化的市值与指数归一化的市值近似相同的假设，所以，现有程序的目标函数是决策变量的非线性函数（也非二次），所以最优化可能会比较耗时。

另外，目前程序中当某个股票的最优权重小于一定阈值时，把它的权重设为 0。这也属于前面介绍的 heuristic 的一种。

现有的程序不对交易成本建模。

综上，目前的程序改进：

1) 做常数权重或组合归一化的市值与指数归一化的市值近似相同的假设，使得最优化已于求解。
2) 可以尝试 mixed integer linear programming，但是不使用成本模型（因为没有成本模型）。
3) 跟踪误差既可以尝试 return-base 也可以尝试 value-based。
4) 约束中可以加入共同因子暴露约束。

4 参考文献

[1] Montfort K V , Draat L F V , Visser E . Index Tracking by Means of Optimized Sampling[J]. The Journal of Portfolio Management, 2008, 34(2):143-151.
[2] Mezali, H., Beasley, J.E. Index tracking with fixed and variable transaction costs. Optim Lett 8, 61–80 (2014)

[3] O. Strub,P. Baumann,Optimal construction and rebalancing of index-tracking portfolios, European Journal of Operational Research (2017)
[4]Yu Zheng & Bowei Chen & Timothy M. Hospedales & Yongxin Yang, 2019. "Index Tracking with Cardinality Constraints: A Stochastic Neural Networks Approach," Papers 1911.05052, arXiv.org, revised Nov 2019