因子处理与回测模型

乔冠卓　中意资产实习生 2020/01/13

本文主要描述构建因子模型和回测的主要流程，并且带入美股道琼斯工业指数成分股以简单说明。本文主要分为三个部分，第一个部分是模型构建，第二部分是模型提升，第三部分是例子与结果。参照天风证券《**基于自适应风险控制的指数增强策略**》构造模型，R代码可以在GitHub中查询：https://github.com/guanzhuo-qiao/multi_factor。

模型构建

由于缺乏基础数据，我们尝试用市场量价信息构建一系列的技术指标以代替因子值。对于因子值我们首先应该进行预处理。

数据处理

处理步骤主要为：去缺失值，极端值处理，标准化和市值及行业中性化。

对于缺失值我们视情况补0或者直接舍弃，对于极端值我们采用天风报告中所做，运用MAD方法，将极端值均匀压缩在3-3.5个"标准差"范围内，这样做的好处是保持了原始数据的秩信息，压缩了绝对距离。标准化我们采用普通的减均值，除标准差的做法。而市值及行业中性化需要做回归并取残差。残差由于与被回归变量有正交特性，故我们认为其不带有被回归变量的属性。由于无法获得市值或者流通股数的信息，这一步在代码中并无体现。

事实上在我们真正做因子模型之前，我们还需考虑单个因子的有效性，或者说我们希望单个因子有对收益率有一定的解释力度，常用的做法是单因子回归，单因子IC检测等等，我们这里假设我们的因子都通过了检验。为了将多个因子值结合，我们还需要考虑多重共线性的问题。多重共线性使得模型参数不具有统计意义，或者说单个属性重复暴露，一般的处理方法是将多个共线因子结合。天风的做法是将原有因子做正交化处理，正交后的因子不再共线，但也不与以前的因子完全相同，但具有相同的解释力度（如何判别？）我们采用对称正交方法以保持原有向量与正交后向量的相关性，同时由于不需要决定正交顺序，故只需要一个截面数据即可完成，对计算效率有很高的提升。

$$\tilde{F}_{N\times K} = F_{N\times K} S_{K\times K}$$
$$S_{K\times K} = U_{K\times K} D_{K\times K}^{-1/2} U'_{K\times K}$$

其中，记矩阵 $M = F'_{N\times K} F_{N\times K}$，$D_{K\times K}$ 为 M 的特征根构成的对角阵，$U_{K\times K}$为每一列由 M 的特征向量构成的矩阵。$S_{K\times K}$是一个对称矩阵，由此得名对称正交。

因子整合

一只股票往往在一个时间截面上对应多个因子值，而这些因子值需要最终被压缩为一个值以表示对其的判断。我们采用的是线性加和，而权数根据历史上ICIR的值决定。具体来说，我们回看一个时期，比如4个月，在这段时间内我们计算每个时间截面上，因子值与下一期收益的相关系数IC，并将IC的时间序列汇总成为ICIR，再以此加权因子值，形成本期的最终因子。

此外还有其他的因子加权方法，例如IC加权，半衰ICIR加权（给予更多权重给近期数据），回归系数加权，ICIR最大化加权等等，其本质上都是希望获得一个加权方式以使得最终结果和收益率有一个显著的单调关系。

风险模型

在我们得到最终因子值后，我们根据这一因子值将股票分为数个层次，理论上因子最大的那一层股票池总体表现应该最好。至此我们基本上完成了因子选股的工作，接下来需要考虑的是如何将股票组合。这就是我们考虑风险模型的时候，一般的方法是采用规划模型以决定组合中个股的占比。而在估计风险时因子模型也有一定优势，我们如果需要估计股票的方差协方差矩阵，那么其实可以只估计因子的方差协方差矩阵和残差项的方差协方差矩阵。这样我们将N个股票的估计转为K个因子的估计和N个残差的估计，这使得计算量大大减小。

天风报告中并没有使用方差项，而是用因子暴露矩阵直接控制因子的暴露，从而模型从二次规划转为线性规划。这样做有利有弊，好处在于计算量小，对优化器的要求也不高，但坏处在于无法细致刻画风险或者对参数正确度要求高。天风认为好处大于坏处，且二次项较难刻画，限制的传导也不如直接控制个股占比更直接。

$$
\begin{gathered}
max \ \ r^T w \\
s.t. s_l \leq X(w - w_b) \leq s_h \\
h_l \leq H(w - w_b) \leq h_h \\
w_l \leq w - w_b \leq w_h \\
b_l \leq B_b w \leq b_h \\
\mathbf{0} \leq w \leq l \\
\mathbf{1}^T w = 1
\end{gathered}
$$

X和H分别为风格暴露和行业暴露矩阵。w_b为基准中个股占指数比重，B_b 为个股是否属于基准指数成分股的 0-1 向量。通过以上优化，我们可以将权重最终决定并进行进一步的组合表现回测。

模型提升

这一节中，我们尝试运用一种自适应的方法来动态的决定个股比例与成分股的偏离程度问题。具体过程请结合天风报告。这一方法希望通过对近期的历史数据进行回测从而确定一个上下限以完成组合不偏离基准太多的目标。换句话说，算法通过迭代，决定组合回测跟踪误差最接近目标跟踪误差的那个限制条件，以此限制条件得到最近的个股权重。

这种自适应的方法的基础在于股票波动率具有聚集效应，波动率大的数据总是在时间上更加聚集。故我们希望对近期数据做出回测并得到最佳的限制参数。"具体来说：

1. 在 T 月底建仓时，首先计算 [T-2,T] 月时间内以个股权重偏离度 w_i ($i \in [1, n]$) 优化得到的组合的年化跟踪误差 TE_i；
2. 对于给定的目标跟踪误差 TE_{target}，找到满足 $TE_k \leq TE_{target}$ 的个股权重偏离度的最大值 w_k ($k \in [1, n]$) 作为 T 月底的个股权重偏离度约束条件。

以中证 500 指数增强组合为例，我们设置了个股相对于成分股权重最大偏离 0.1%-0.5%，分别对 0.1%,0.2%,0.3%,0.4%,0.5%共 5 种约束方式下进行组合优化得到 5 个组合的滚动 3 个月跟踪误差，如图 16。可以看到，在每一期，个股权重偏离约束和组合实际跟踪误差满足单调关系，权重偏离约束越宽，则组合的实际跟踪误差越大。给定预期跟踪误差，在每个截面上，以能达到目标跟踪误差的最大个

股偏离度作为下一期组合优化时的约束条件。例如给定预期跟踪误差为 3.5%，在 20131231，以最大偏离 0.2%的组合过去 3 个月的年化跟踪误差为 3.06%，以 0.3%为约束的组合过去 3 个月的年化跟踪误差为 3.88%，因此在当期约束跟踪误差时，我们以 0.2%作为个股权重最大偏离的约束来求解下一期组合；在 20150731，以最大偏离 0.1%的组合过去 3 个月的年化跟踪误差为 3.76%，其他约束下的跟踪误差都高于 4%，因此在当期我们以 0.1%作为个股权重最大偏离约束；在 20170630,期，以 0.5%为约束的组合过去 3 个月的年化跟踪误差为 3.39%，其他约束下的跟踪误差都低于 3%，因此在当期我们以 0.5%作为个股权重最大偏离约束。"（天风证券基于自适应风险控制的指数增强策略）

简单的例子和结果展示

数据：美国道琼斯工业指数部分成分股（29支）。

因子：技术指标，例如：MACD，PPO，PVO，RSI等。

区间：2017/1/1至2019/12/31

基准：道琼斯工业指数

调仓频率：月度调仓（月末决定，月初建仓）

我们将数据清洗并且正交化处理后得到一个三维array。再对每一个时间截面上的因子进行整合加权，得到一个因子值。最终的结果是一个二维array，行为股票，列为时间，元素为整合后的因子值。下面这个例子是29支股票在29个月的因子值。

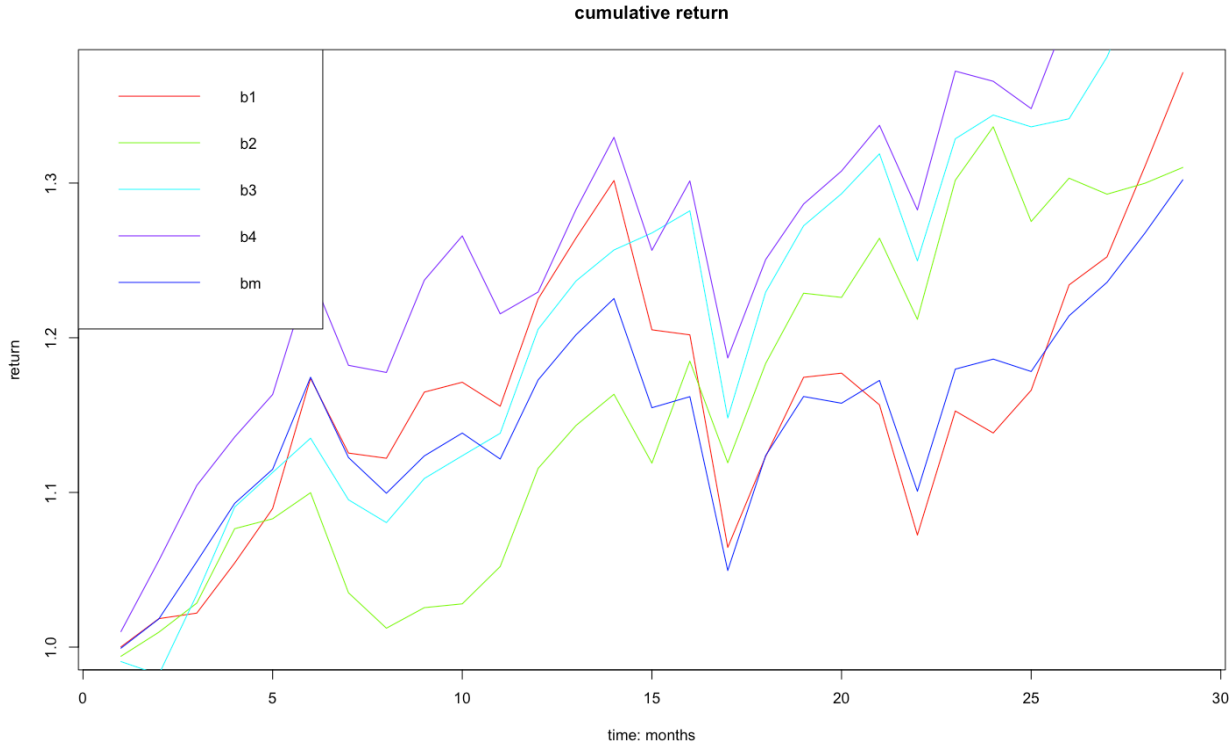
```
> head(bt_features)
            [,1]       [,2]       [,3]       [,4]        [,5]        [,6]
  [,7]        [,8]
[1,]  0.43598134 -0.5750812 -0.1058110  0.5053070 -0.11832659 -0.37091667
 0.3543298  0.28102932
[2,]  0.08372542  0.4599505 -0.4591310  0.1495429 -0.32159339 -0.12337529
-0.1200087 -0.23193981
[3,]  0.13438791  0.2503357 -0.2640883 -0.1354622  0.26518895 -0.27071438
-0.3243023  0.27039123
[4,] -0.70687629 -0.5934455 -0.1957683  0.6396821 -0.09935977  0.58053270
-0.4071547  0.96075376
[5,] -0.17369295  0.3054716  0.1905802 -0.3483065 -0.05433360 -0.09594862
 0.1887733 -0.15582574
[6,]  0.13524096  0.1675407  0.5525201  0.1626873  0.39510130  0.16057637
 0.7418007  0.03234227
            [,9]       [,10]      [,11]      [,12]       [,13]       [,14]
  [,15]      [,16]
[1,]  0.83070309  1.14911186 -0.3962300  0.3781016  0.19247831  0.35635552
 0.01748212  0.3246255
[2,] -0.28888875 -0.57402939  0.4976762  0.1906381  0.19456817  0.27173860
-0.13751911 -0.5949686
[3,]  0.08238604 -0.06851622  0.3168511 -0.2070851 -0.09320662  0.15762307
 0.42933176  0.2747580
[4,]  0.93117735  0.55472605  0.2011126  0.3318564  0.05993579  0.04343523
 0.36802666 -0.2566583
```

```
[5,]  0.02685606 -0.86175433 -1.0417464 -0.7884720 -0.20694020 -0.16024021
-0.40284032  0.1130009
[6,]  0.44912192 -0.15848117 -0.7329819 -0.4734821 -0.63666255 -0.01802672
 0.03002440  0.1883757
           [,17]       [,18]      [,19]        [,20]       [,21]      [,22]
      [,23]
[1,] -0.40965043  0.96810902  0.3811665 -0.005115914  0.25356436  0.2451652
 0.2917801415
[2,]  0.75346243 -0.12460955 -0.4926893  0.063939634 -0.75028841 -0.8135961
-0.0004792547
[3,] -0.50675154 -0.48298162 -0.5007111  0.073956522 -0.30329833 -0.1280207
-0.2448001210
[4,] -0.06973566  0.07597387 -0.3735699 -0.577429712 -0.16238844 -0.1308649
-0.1157356414
[5,]  0.41774394  0.25231699 -0.2292901 -0.038407991 -0.05887342 -0.4136213
 0.2729978323
[6,] -0.30062836 -0.78136316 -0.2715915  0.308837751  0.54720352 -0.3951290
 0.0707896790
           [,24]        [,25]       [,26]       [,27]      [,28]        [,29]
[1,] -0.08681787 -0.099129994 -0.31054138 -0.15953853  0.2213028 -0.001094287
[2,]  0.76128850 -0.009527014  0.44563172  0.51818076 -0.1705808  0.162288725
[3,]  0.15560287 -0.124576727  0.66879377 -0.41414596 -0.3827018 -0.020974086
[4,] -0.18387129  0.496765245 -0.12727276  0.53042203 -0.0163231  0.204298633
[5,]  0.12851549  0.270503564 -0.02116984  0.06092551  0.1806420 -0.297538104
[6,]  0.39881084  0.058126584 -0.02493874 -0.63688062 -0.1966412 -0.425293876
```

再通过调用规划函数，得到每个股票在每一期的权重向量，并以此进行回测。值得注意的是，我们将整个股票分为四个股票池，依据的是股票因子值的大小排序。

最终的回测结果如下：

```
> result_report_table
         col_names
row_names  return_p  return_b         te        mdd    sharpe
       b1 1.371392 1.302109 0.01067095 0.23717453 0.3821377
       b2 1.310100 1.302109 0.01411259 0.08759393 0.1367888
       b3 1.490173 1.302109 0.01177096 0.13395185 0.4594947
       b4 1.551115 1.302109 0.01083887 0.14265854 0.2455361
```



我们看到不同层没有太大的分层效果，这可能是由于技术指标衰减过快导致。另外本文在测算ICIR时利用的本月数据，故在回测时必须再向后推迟一个月以避免使用未来数据（收益率），所以b1代表的第一层组合效果并不算很好，而反而是第三层的夏普率非常高。

对于自适应的改进，由于缺乏数据并没有形成结果，但是其函数已经搭建好，并且在带入随机生成的数据后可以工作，如果今后获得数据，我们可以进一将这一功能加入其中。

附录

代码方面，factor_download.R用以下载数据，数据来源为雅虎财经的R语言接口，且数据样本已经下载好存入项目文件地址中，所以不需运行此文件。用户可直接运行factor_test.R以观察中间处理因子的过程，或者运行back_test.R以获得最终回测结果。

本文借鉴天风证券2018年7月5日发布的**基于自适应风险控制的指数增强策略**。

本文仅供学习参考交流，回测模型建立在历史数据基础上，存在失效的可能。