因子研究之分组法．第27讲

渔阳 2018-10-15 20:28

字数 4,611 阅读需 12分钟

“欢迎来到量化小学”

▲ 加入“量化小学”校友圈儿提问交流

投资全球更要投资自己

详细内容请在wifi环境下观看视频

<本期课程4909字，视频16分48秒，请合理安排学习时间>

来自特辑

[redacted] 量化小学

内容阅读

大家好，欢迎来到量化小学。

今天我们继续来讲因子研究，介绍一个最基本的方法，也就是**分组法**。主要是两部分内容，首先，我们再一次理清因子研究的步骤，然后讲一讲最基本的分组法或者是分层的方法。

Alpha策略选股的一般策略

首先因子研究有这么几个大的步骤。我们的目标是在选股空间内，通过一种横向的比较，来挑选做多或者做空的目标，英文叫cross-sectional study。比如说在中证500的空间以内，哪些是好股票，哪些是坏股票。那怎么来做呢？会通过一系列的因子来描述每只股票的特征。

解放你的投资动手能力

结合前面讲过的内容，这里面有风险因子，包括行业市值等，也有一些是我们即将寻找的阿尔法因子，也就是对未来的收益率具有一定的预测能力的因子。那么涉及到因子，又分**单因子、多因子组合**两个步骤。

我们首先是研究单个因子，它需要对未来一段时间的股票的相对表现有一定区分能力，例如说前面我们讲到了一个股票的市盈率，或者一个股票的成长性等等。

有了单个因子之后，在选股之前，我们还要把单个因子组合起来，**用一个线性或者非线性的组合来对股票做一个总体的评价。**这就好比决定大学录取哪个学生的时候，你不能光看他的数学成绩，也不能光看外语成绩，要把很多的性质给综合起来，挑选一个总分最高的一个学生，这也就是多因子模型的概念。当然，这里面可以有线性的组合，也可以有非线性的组合，在后面都会有所涉及。

对股票未来的收益率或者未来的排名有了一个认识之后，就可以最终进入到选股和交易的步骤。所以从大的角度来说，我们单因子研究，然后多因子研究，再过渡到组合构建，最

后是交易执行，这可以说是阿尔法选股策略的基本的业务流程。

分组法：单因子研究的基本方法

因子研究最基本的方法，就是我们今天要讲的分组法。它分为几大步骤，目的还是比较显然的。首先，你要在每一期，比如说每个每天或者每周或者每月，对选股空间的所有股票计算一个因子值。以市盈率为例，它的市盈率究竟是10倍、20倍还是30倍，要计算一下。

第二，**要对因子值进行加工。**以市盈率为例，它有时候会有一些比较极端的情况，说一个公司不赚钱，它的市盈率是负的；或者一个公司市盈率非常高等等，因此就需要进行一些去极值的处理，否则会影响你比较的效果。还可能会需要把极端的值进行一些标准化或者归一化的处理。

具体的方法在我们课后推荐的材料里面都有，是华泰证券的一篇关于多因子的报告。怎么样做标准化，怎么样做归一化，它都有详细的步骤解说，其实跟我们前面讲过的那些正态分布、标准差都很有关系，大家可以具体地去看。我们量化小学的目的也是给大家一个提纲挈领的讲解，先知道我们要去干什么，然后你再去看那些数学公式就会变得比较简单。

所以第二步是对因子进行加工，**去极值、标准化、归一化**，就像炒菜一样，把这个鸡给它切成鸡丁，然后就可以进入下一步了。

那么分组法其实是非常直接的。按照因子值，直接把选股空间的股票进行分组，通常来说是分成五组或者十组。然后进行观察，看看在未来一段时间相对表现究竟有没有差异。

在这我要特别指出，因子都有几种，相当于有几种这个模式。可以不做中性化的处理，也可以做行业中性的处理，也可以做市值中性的处理，也可以做行业和市值都中性的处理，

为什么要这样?

其实在前面风险模型的章节当中，我们也曾经反复地提到过，因为行业和市值是A股市场两个最大的风险因子。为了让阿尔法更纯粹一些，你可能要先把这两个风险的因子给控制一下。

如果要进行行业中性，方法也很简单，就是在行业内部进行分组比较就可以了。如果要进行市值中性，也是类似的方法，你可以先按照市值，将股票分成若干组，就大股票在一起，小股票在一起，在组内再按照因子进行分组。

最后你也可以不做中性化。为什么只有三种你也可以不做中心化呢？因为有些阿尔法因子，它其实自动的带有一定选行业或者选市值的能力，既然它能够通过行业或者市值来赚钱，你也不妨就让他去赚这个钱。至于在实际中，你到底要不要做中性，要做哪些中性，这就是研究员自己的解读了。

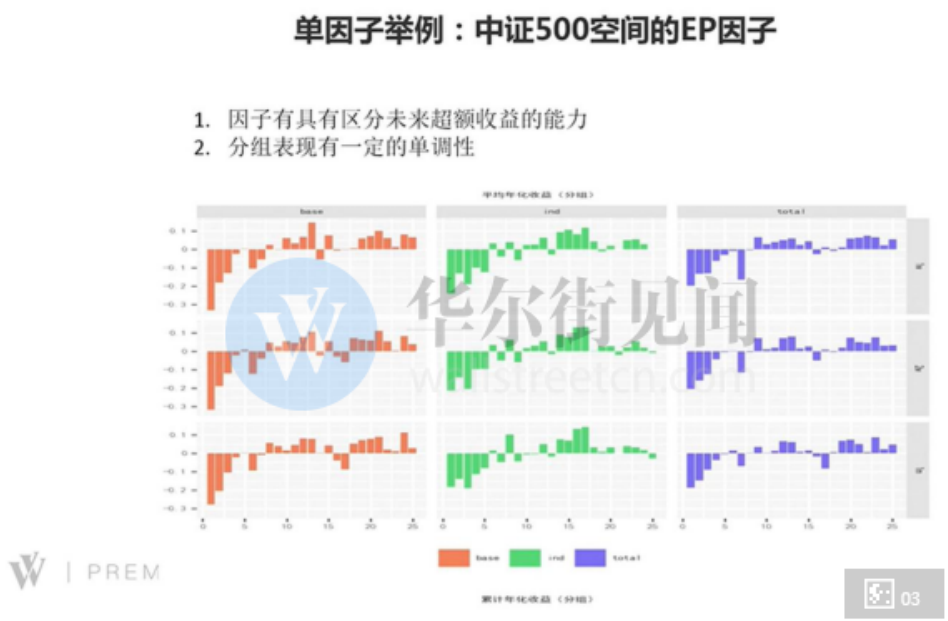
根据实际经验来说，这几种情况你都会遇到。有些因子中性化之后会好一些，有些因子不做中性化会好一些，这也和你做这件事情的目的有关。比如有些管理人，他是比较保守的，他不要这个市场、这个行业或者是市值的暴露，那么它更加趋向于中性化；也有些人他的目的是多赚钱，可能不做中性化的时候就会多一些。

最后一步就比较简单。首先观测一下第一组和最后一组未来收益是否有明显的区分，第二个就是各组的表现是否有单调性。

单因子举例：中证500空间的EP因子

我们来看一个例子，用jack系统实际生成的一个最基本的因子报告。它是EP就是市盈率的因子。它把市盈率取了一个倒数，因为这样就好比较。EP值越高的股票相对来讲越便宜。

在这简单的复习一下，假如说市盈率是10倍，就意味着如果股价是10块钱的话，这个公司一年的盈利是1块钱，如果我取倒数的话，它的EP就变成0.1，也就是说，EP值越高，股票相对而言越便宜。



这个EP的因子，对于未来的超额收益是不是有任何的区分能力呢？在这张图上可以看一下。

首先这三种颜色就是刚才我们讲到的那三种最基本的因子的形式，红色的是不做任何中心化处理的，绿色的是做了行业中性化处理的，蓝色是做了市值中性化处理。，那么每一个那三列，实际上是过去一周、两周和一个月的表现，集中看第一行就可以了。

首先我们注意观察几点，就是**因子它对未来的收益率是有一定的区分能力的。**体现在红色的线有的是向上的，有的是向下的，向上的意味着是超额收益，年化的向下意味着是负的超额收益。

我们以左上角红色的这张图为例。它是不做任何中性化的因子的分组表现。好的那些股票，它可能一年超额收益能达到将近10%，但是坏的这部分，它的超额收益就负的特别厉害，能够有负百分之二十几。这也是由于过去一段时间，A股整个的跌幅也比较大，那些垃圾股跌得尤其厉害，所以这张图也反映了这一点。

所以说，**因子对于未来的超额收益是有一定区分能力的。**第二个我们也看到一定的单调性，就是，不光是第一组好最后一组不好，如果你看前面几组和后面几组的对比，尽管不是那么的线性，但还是有这样一定的特征的。

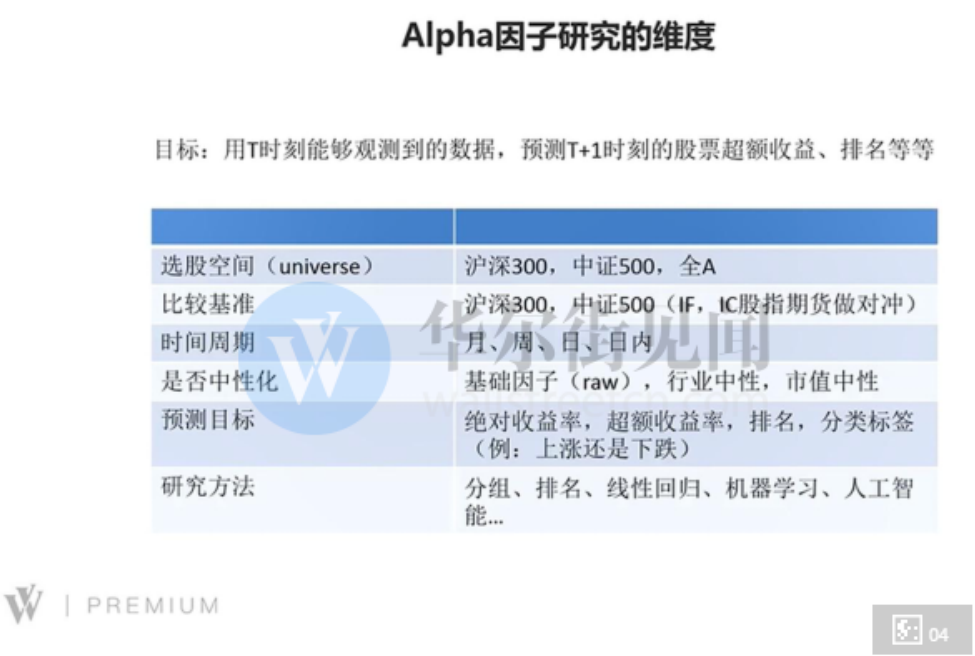
如果往右看，可以看到做过中性化处理之后，整个这个样子还是在那儿。但我们也看到，亏钱的那些不好的股票，相对的收益也小了，赚钱的那些，相对的收益也变小了。这就是刚才我们讲过的概念，你做了中性化处理之后，确实会对风险控制是有帮助的，但有时候也会影响因子的收益。

那究竟怎样权衡？看你的目的是什么。跟前面我们讲的夏普比率也有异曲同工之妙。就是，我到底是愿意承担更多的上下波动，还是我就是为了极大化收益率。比如说，中间绿色的做了行业中性处理之后，各组之间表现相对来说还是更均匀一些。

后面我会讲到，你可以用一个类似夏普比率的数值来量化，这一点叫做SEIR。可能在下一讲当中会涉及，那么在这儿我们就先有一个基本的印象。反正分组法就是，我把选股空间里的股票分成组，然后直接来观察它的收益率是不是具有一定什么样的特点。

Alpha因子研究的维度

那么刚才我们是看了一个具体的例子，涉及到Alpha因子，它的研究是有多种维度的。任何一个因子你都要先把一些维度描述清楚，那么这个因子才有意义。主要是以下几个维度。



首先是**选股空间**。你是从沪深300当中选，中证500里面选，还是从全A里面选，这是最常用的几个选股空间。

第二，既然是谈论超额收益或者是相对表现，你需要有一个**比较基准**，最常用的就是沪深300或者中证500，最大的原因是他们两个有与之对应的股指期货来作为对冲。我们马上也会看一下这两个指数有什么异同。

第三，是**时间周期**。月、周、日，当然也有日内，通常涉及到Alpha研究，日内应该说不是太常见，至少在国内。那么你如果看院商、券商的研报的话，月级别的会稍微多一点，从实际的交易的角度来说，可能周级别和日级别会更稍微多一点。

第四点就是刚才我们讲过的，这个因子到底**中性化**过没有。可以是基础因子，也可以是行业中性过的，也可以是市值中性过的，或者是行业和市值都中性的。

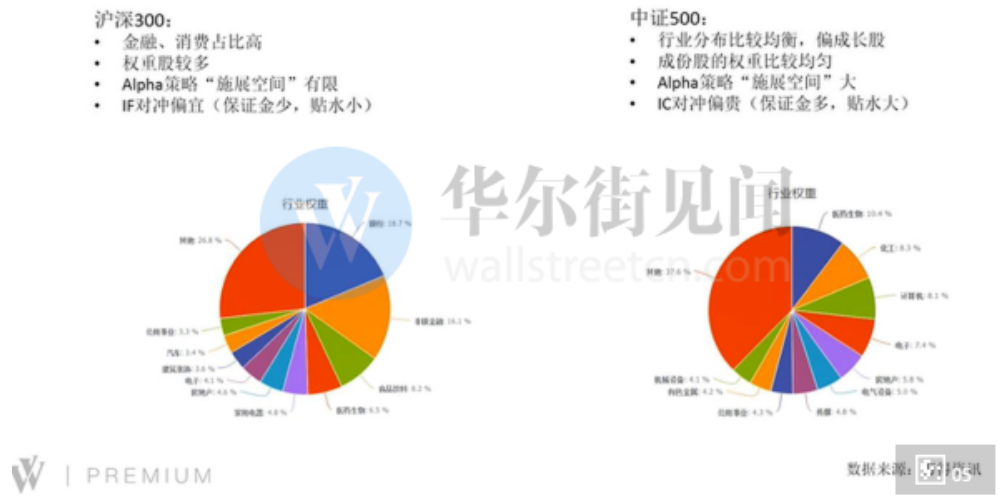
下一个涉及到因子就是你到底是**预测什么东西**。你可以是预测绝对收益率，这个不是太常见，通常还是说是超额收益率。另外一个是在选股空间内的排名，那么我们后面会涉及到人工智能或者机器学习。还有一个常用的就是分类标签，它究竟是上涨还是下跌的，是大涨还是大跌的。那么这是常见的一些预测的目标。

那么最后一点，Alpha因子的**研究方法**，这可能是大家比较关心的也听说比较多的。我们今天涉及的是最简单的分组法，接下来还会讲线性回归，还会讲秩回归就是排名的方法，然后慢慢地就会过渡到机器学习、人工智能。你会发现这些方法本身并不是分立的关系，它是一个逐步演进的关系。

这些是阿尔法因子研究的维度。那么刚才我们讲了两个选股空间和基准，沪深300和中证500，这是一个实证当中你会经常遇到的事情，所以我们简单的讲一下这两个有什么不一样。

选股空间和基准的比较

沪深300底下这张图是我从万德上截的，它讲的是沪深300的权重。我们看到银行和非银金融占据了很大的比例，将近35%。然后接下来是食品饮料、生物医药、家用电器等等，所以它是金融和消费的占比比较高。



另外一个就是**权重股占比很高**。比如说中国平安一只股票大概就占据了5%到6%的权重，就导致了几个事情，有好也有不好。从不好的方面讲，Alpha策略的施展空间是有限的，因为它的行业分布比较偏，股票从权重的角度来说分配的也比较偏，你就不得不花很多的时间精力来控制风险。

我究竟要我的银行股大概占比多少？也不能不买是吧？而银行股本身动力又不是很厉害，会影响你赚钱的效率。但是这个也有好处，就是我们刚才讲到对冲的原因。IF股指期货对冲它是比较便宜的，只要保证金是15%，现在它股指期货的贴水也比较小，就意味着你的对冲成本是比较低的。

另外一个比较常见的**选股空间和比较基准就是中证500**。可能在真正做Alpha的基金中，占比应该是比沪深300还要更高一些。为什么呢？它虽然有不好，比较贵，做起来成本比较高，但是它施展的空间也大。

比如说从这张行业的分布图上就可以看到它比较匀。占比高的那几个行业，像生物医药大概占10%，化工、计算机、电子都将近10%。不但比较匀，而且这几个行业当中又有比较多的股票可以供给你选。

另外**这些股票也有一定波动性**。我们前面讲投资基本原理的时候，大家可以回去复习一下那个公式，你赚钱的能力是跟股票自身的波动性相关的。它要是不动的话你是没有可能赚钱的对吧？中证500的成分股本身分布比较均匀，同时也有一定波动率，对于Alpha策略是很有帮助的，相对而言成长股的比例也会大一些。

但最后不好的地方是，它的Alpha对冲偏贵。它保证金大概现在是30%，另外它的贴水是比较大的，也就相当于，如果用IC股指期货来进行对冲的话，我一年可能光在这上面就要损失大概5到10个百分点。所以还是比较很贵的。这是两个最常见的基准。

因子研究方法树状图

今天我们简单的讲了一下最基本的方法就是分组法。那么接下来我们也为马上要讲的东西做一个预览。接下来我们会涉及到什么？这些方法之间又有什么一个逻辑关系？

今天是分组法，那么数学上更严谨的一些方法如图所示。接下来这三个框框可以用来做一些线性回归。你可以用排名来做一个线性回归，这应该叫Spearman correlation 或者是rank correlation；也可以做普通的线性回归，这个是直接预测收益率的。最后有一种叫逻辑回归，就是把要研究的当成一个分类问题，去预测这个股票是属于涨的还是属于跌的，这都可以统称为线性方法。我们下一讲讲这些。

这些是基础的方法，那么之后是一些进阶的方法。也就是我们现在比较火的机器学习或者叫统计学习。它可以有KNN、SVM、随机森林还有增强算法等等，现在如果不太明白没关系，我们很快就会看到这些都是线性方法的一种自然的延伸。

再继续学习之后，我们现在又有一些更偏人工智能的方法，包括神经网络、深度学习和强化学习等等。但说实在的，现在一般涉及到金融的话，这些可能应用的场景还相对来讲比较有限，或者说它的效果也未必比前面的方法好，所以我把它的颜色稍微淡化了一些。当然事物也在发展，也许随着数据的增长，这些更进阶的方法就变得更加重要了。

那么在接下来几讲当中，我们会简单地给大家介绍一下这些方法。让大家明白这些方法是在干什么的，然后你们再去看具体的技术方面的材料，就会有一个事半功倍的效果。所以下一次我们就会讲到线性的方法。谢谢大家。

-END-

量化小学

渔生

小学而大不遗，量化师生联谊会

感谢大家订阅《量化小学》，这里是学校见识社群，你可以随时提问、随时互动，我们一起投资，一起分享！

风险提示及免责条款

市场有风险，投资需谨慎。本文不构成个人投资建议，也未考虑到个别用户特殊的投资目标、财务状况或需要。用户应考虑本文中的任何意见、观点或结论是否符合其特定状况。据此投资，责任自负。

写评论

请发表您的评论

表情 图片 发布评论

违法和不良信息

举报电话: 021-60675200（周一到周五9:30-11:30，13:00-18:30）
举报邮箱: contact@wallstreetcn.com
网站举报: 点击这里