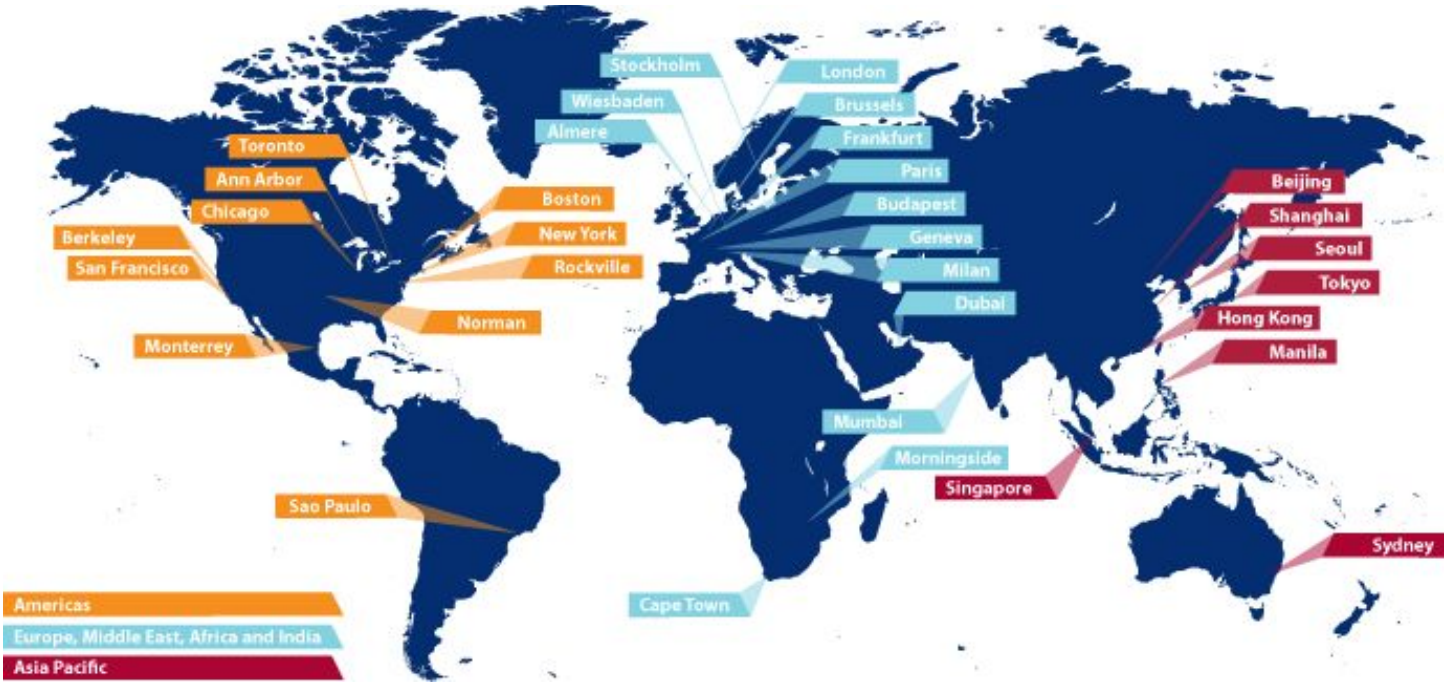

【多因子模型】Barra模型讲解（1）

文兄 · 8 天前

在文章开始之前，先帮原CMO集训队的T大数学系妹子的Team打个招募广告：

> 我们是一个主要由在校学生组成的量化团队，有着靠谱的IT团队（清华CS硕士*2 + 电子系phd + 姚班少年），也有优秀的策略开发人员（所有成员均来自清华\北大\北航\中科院 数学、物理、计算机、电子等相关专业，包括两枚CMO金牌以及即将去Princeton MFin, Stanford ICME, Baruch MFE, Columbia MSFN, Cornell MFE, Chicago MSFN 等Top金工项目读书的同学）。
>
> 目前我们有固定的Office场所进行活动，并已经基于wind的落地数据库，独立开发了一套类似worldquant websim的因子筛选、评测与策略回测平台，可以用于进行日内与日间的信号挖掘与策略研究。由于以在校学生居多，团队有着不错的交流和学习氛围，会兼顾每一个新来的同学的具体情况予以指导，同时每周都会定期组织讨论。
>
> 如果你对该我们团队感兴趣，并具有以下技能之一，请将简历发送至邮箱：thualphamax@163.com，成为我们的一员，与众大神一起学习进步，同时我们也会提供与市场行情相符的实习工资以及依据贡献程度的最终策略分成。如果你是非北京的优秀同学，可以考虑暑期前往（我们将提供食宿）。

> 1. 量化策略研究：对多因子、日内或是统计套利策略的研究感兴趣并有着一定coding经验与数学水平的同学
> 2. Python语言工程师：具有一定coding能力，熟悉Python语言（Pandas/Numpy/sklearn等库），可以长期参与系统编写与维护的同学
> 3. 新闻舆情抓取与分析：熟悉网络爬虫或是NLP的同学

Barra模型是MSCI公司开发的一个经典的金融风险控制模型，名声远播在全球有着大量的机构用户。本专栏旨在传播与讲解量化知识，故而决定从今天起，通过几篇文章陆续介绍Barra模型的构建的各步骤与相应的计算方法和细节，请大家持续关注~

本文作者：涂申昊（知乎ID：@感到不爽）

一、传统框架下的模型介绍

说到金融学的投资组合模型，那么不得不提就是马科维茨的投资组合模型。从1952年诞生至今，马老师的均值-方差基本就是金融学里面神一般的模型了。至今应用十分广泛，经久不衰，折磨了一代又一代的金融学子，让无数童鞋挂在了高高的模型上。

马老师的模型第一次将均值和方差来刻画股票的收益和风险，基于马老师的均值-方差模型，有了三种度量风险的方法，其实本质思想大同小异：

1.1 基本风险模型

即为马科维茨组合方差。在马科维茨的均值-方差理论中，投资组合的风险计算需要估计组合中每个资产的波动率及它们之间的相关系数。一般的，当组合中有N只股票的情况下，需要估计的波动率个数为N，而需要估计的相关系数的个数则为N*（N-1）/2。我们可以将所需要估计的参数总结到一个协方差矩阵V中。

问题在哪?

很明显，n*n维的协方差矩阵需要估计的参数太多；比如说A股市场，目前市场上超过3000只股票，也就是需要顾及大约450,0000个参数。

怎么办?

1.2 减少估计次数的模型

令

为第n和第m只股票的协方差，定义

其中p是股票之间的平均相关系数；虽然这样只要估计各个股票自身收益率的协方差，但是模型忽略了类似行业或者具有相似属性的股票之间的微妙联系。

既然结果又不精确了，怎么办?

1.3 历史数据估计

用一段时间的历史数据（比如说一年）来计算样本协方差矩阵，这也是目前各种论文中的通用方法。

有什么问题：

1.根据历史数据的协方差来计算股票的投资比例，也就是用纯粹的历史数据来预测未来，不言而喻，结果肯定会产生一定偏差。

2.从目前的学术研究成果来看，马科维茨投资组合模型的鲁棒性非常差，协方差矩阵稍微有一点变化，股票投资组合的比例变化极大。

3.如果估计月度数据，用T个时期的样本来估计一个N*N的协方差矩阵，并且要求

T>N（如果T<N，那么会导致协方差矩阵奇异，无法求其逆矩阵）这就意味着，如果要估计沪深 300 指数成分月度收益率的协方差矩阵，将需要至少超过 25 年的历史数据，这在应用中显然不切实际。

既然基于马科维茨的投资组合模型的协方差矩阵方法有着这样那样的问题，怎么办?

二、结构化多因子模型

接下来就要引入我们的主题了，结构化风险因子模型，也就是传说中的多因子模型。本文主要参考的是barra的risk handbook和Qian(2007)的Quantitative Equity Portfolio Management Modern Techniques，以及国内各大券商的研报。

结构化风险因子模型利用一组共同因子和一个仅与该股票有关的特质因子解释股票的收益率，并利用共同因子和特质因子的波动来解释股票收益率的波动。结构化多因子风险模型的优势在于，通过识别重要的因子，可以降低问题的规模，只要因子个数不变，即使股票组合的数量发生变化，处理问题的复杂度也不会发生变化。

结构化多因子风险模型首先对收益率进行简单的线性分解，分解方程中包含四个组成部分：股票收益率、因子暴露、因子收益率和特质因子收益率。那么，第j只股票的线性分解如下所示：

也可以写成矩阵表达式：

其中，Rj表示第j只股票的收益率；Xj表示第j只股票在第k个因子上的暴露（也称为因子载荷，本质上说白了就是该股票的所对应的因子值）；Fj表示第j只股票第k个因子的因子收益率（即每单位因子暴露所承载的收益率）；u表示第j只股票的特质因子收益率。（一般情况下，我们都用N代表股票数，K代表因子数）

我们定义因子暴露（因子值）是在时刻t的结果，那么股票收益率、因子收益率和特质因子收益率均为t+1的结果。这就是一个很典型的，用因子当期值，来预测下一期因子收益率的问题了。

令投资组合的权重

那么投资组合的收益率为

现在我们假设每只股票的特质因子收益率与共同因子收益率不相关，并且每只股票的特质因子收益率也不相关（此假设后续模型一直能用到，非常关键）。那么在上述表达式的基础上，可以得到组合的风险结构为:

其中，X表示N只个股在K个风险因子上的因子载荷矩阵（N*K），F表示因子收益率的协方差矩阵（K*K），

delta表示因子的特异收益率方差矩阵（N*N的对角阵）。

三、多因子模型里因子的形式

在有了多因子模型框架之后，就是寻找因子的问题了。通常情况下，影响股票收益率的因子通常可以分为三类：宏观经济因子、基本面因子、统计面因子

（1）宏观经济因子，宏观经济因子通常只可观察的宏观经济序列数据，比如GDP、CPI、利率等。

但是宏观经济因子只有一个，股票却有3000个，这无疑会带来两个问题。首先，每一只股票的收益率都要和宏观经济因子做回归，这非常麻烦。第二，宏观经济指标滞后性明显，对股票收益率的预测效果并不显著。（研报的观点，个人觉得这个理论可以商榷）

（2）基本面因子，包括股票财务报表中的各种指标，以及各种K线指标也可以算作此类。

（3）统计面因子，个人理解，统计面因子主要是对因子做处理之后的新因子。比如说六个月动量，12个月之后的反转，或者可以从股票收益率协方差里面提取一些参数，作为统计面的因子。

一般而言，实际多因子模型中，用的最多的是基本面因子。

四、多因子模型的预处理流程

（1）去极值

目前去极值一般有三种方法：均值方差去极值、MAD方法去极值、分位数去极值、

1.均值方差去极值

求每一个因子的均值和方差，大于 $\mu + 3\sigma$ 和小于 $\mu - 3\sigma$ 的样本值转化为 $\mu + 3\sigma$ 和 $\mu - 3\sigma$ 。

2.MAD法去极值

MAD 法是针对均值标准差方法的改进，把均值和标准差替换成稳健统计量，样本均值用样本中位数代替，样本标准差用样本MAD代替：

通常把偏离中位数三倍MADe（如果样本满足正态分布，且数据量较大，可以证明 $\sigma \approx 1.483 * MAD$ ）以上的数据作为异常值。和均值标准差方法比，中位数和MAD的计算不受极端异常值的影响，结果更加稳健。

3.分位数去极值

分位数去极值是一种经验处理方法，假设Q1和Q3分别为数据从小到大排列的25%和 75%分位数，记IQR=Q3-Q1, 把区间

里的数据标识为异常点。

分位数是稳健统计量，因此分位数方法对极值不敏感，但如果样本数据正偏严重，且右尾分布明显偏厚时，分位数去极值方法会把过多的数据划分为异常数据。

所以，有了改进的分位数去极值法：

定义：

然后定义了调整的上下限:

在区间

上的数据被定义成了异常值。

（2）标准化

每个因子做完了去极值之后，就要消除各个因子之间的量纲影响，进行标准化。标准化的步骤通常都是zscore标准化法，非常基础，没什么可说的。

（3）中性化

中性化的内容barra框架中并未提及，但是在A股市场中，各家研报都认为中性化仍然很有必要的。

首先：A股票行业轮动明显，行业热点之间切换迅速，量化模型也很难有效预测轮动规律；其次，A股的小市值个股占比显著高于国外市场小市值个股具有高波动率、高收益率的特性，为了降低投资组合的波动性和回撤，需要进行行业中性化和市值中性化处理。

个人认为，如果说要博取更大的投资收益，并承担更高风险的话。市值中性化是可以不用去做的，因为A股市场最近十几年以来，小市值因子至少有20倍以上的收益。并且到2016年，小市值因子的有效性也没有消失。2017年初，小市值因子确实出现了失效的情况，但是未来会不会有效，这个就是玄学了。

不做市值中性化的话，完全可以建立一个市值轮动模型进行替代。行业中性化也是同理。但是如果想找到稳定的alpha因子，那么市值中性化和行业中性化还是要做的。

行业中性化通常有两种办法：

1.简单的标准化法

利用申万行业指数，将各个行业内股票的因子进行标准化处理，即减均值除标准差。

2.回归取残差法

将因子值作为y，行业哑变量作为x，进行线性回归，然后回归模型的残差即为行业中性化后的因子值。

市值中性化因为市值因子是连续的，所以采用的是回归取残差法，因子值作为y，市值作为x。

五、因子收益率向量的估计

因子找好之后，就进入了估计因子收益率的部分了。在barra的框架中，因子收益率通常是日数据。

利用第二部分的公式

因子暴露X已知，因子收益率R已知，所以针对每一天的截面数据进行回归，X取当天的因子值，R取下一天因子的收益率。就可以估算出当天的因子收益率了。

当然，有个非常关键的问题，又扯到之前的假设上来了。

Barra模型认为，每只股票的特质收益率u不相关，这个假设在计量经济学的理论框架里，造成了一个非常明显的问题，异方差性。

那么怎么估计因子收益率呢，只能用WLS方法了（加权最小二乘）。问题又来了，加权最小二乘的算法中，权重怎么取？通常的计量经济学方法是取残差平方的倒数，但是barra模型中，这个权重取了根号市值，这个问题是个玄学。

利用WLS方法，可以得到因子收益率的最终表达式：

其中W是加权最小二乘法的权矩阵。

参考文献：

1. 国泰君安，数量化专题之五十七：基于组合权重优化的风格中性多因子选股策略
2. 爱建证券，多因子系列之一：多因子模型梳理探索
3. 华泰证券，多因子系列之一：华泰多因子模型体系初探
4. 东方证券，选股因子数据的异常值处理和正态转换——《金工磨刀石系列之二》
5. Barra，USE4
6. Qian, Quantitative Equity Portfolio Management, modern techniques and applications

量化交易　宽客 (Quant)　金融工程学

10 条评论

veyyiey

麻烦注明出处，参考**券商研报，谢谢！

8 天前　4 赞

腾宇James

` X取当天的因子值 ', 应该是 f 吧 :)

8 天前

littlestar 回复 **腾宇James** 查看对话

f是要回归出来的系数，代表因子收益率。

7 天前 1 赞

文兄（作者） 回复 **veyyiey** 查看对话

已经补充了参考文献~

7 天前

greedisgood

天下文章一堆抄→原来文献是这样引用的。

7 天前 2 赞

肖剑飞

这基本是copy的华泰多因子1呀，兄弟

7 天前 2 赞

文兄（作者） 回复 **肖剑飞** 查看对话

是吗，我问问…

7 天前

胖子

很不错，作为在校学生，这个水平超过了大多数同龄人，鼓励一下，再接再厉

7 天前 2 赞

非凡猫

请教一下，因子负荷是如何算出的？比如市场平均市值100亿，那500亿的股票，市值负荷就是5吗

7 天前

子非末

还有一步因子正交化处理，有比较多的争议，是否能把您的观点也分享出来呢?

7 天前 1 赞

文章被以下专栏收录

清华大学量化投资协会成果集萃

清华量化协会关于量化交易模型、策略的学术探讨

进入专栏

推荐阅读

【多因子模型】Barra模型讲解（2）

接着第一部分继续，第一部分地址：【多因子模型】Barra模型讲解（1） - 知乎专栏第一部分主... 查看全文 >

感到不爽 · 5 天前

【统计套利】找到配对交易里的最优threshold

今天我们讨论的论文是利用随机控制的方法计算配对交易中的最优threshold，所涉及的paper下载... 查看全文 >

文兄 · 1 个月前 · 发表于 清华大学量化投资协会成果集萃

旅途——外交部淘来的大宝贝奔驰E级旅行车 s210

人生即是一场旅行一台来自中国外交部交通司的“旅行家”——Mercedes-benz e class wagon (s... 查看全文 >

王煊 · 7 天前 · 编辑精选 · 发表于 车艺志

风吹草低见牛羊，威少哈登谁更强 | 饭特稀实验室Vol.1

在作者写下这篇文章之前，薛定谔的威少很可能已经成为了这赛季的常规赛MVP。当然了，常规赛... 查看全文 >

Brad Zeng · 1 个月前 · 编辑精选 · 发表于 饭特稀实验室