中央财经大学 金融学院

School of Finance, Central University of Finance and Economics

资产定价横截面研究简介

朱一峰

中央财经大学金融学院

2024年1月11日

前言

- 资产定价研究中主要分时间序列研究和横截面研究。
- 时间序列研究主要是研究如何预测未来资产价格，对应的研究对象是时间序列数据。例如：预测未来的原油价格，我们需要用到现在的原油价格、原油库存、原油供需情况、成品油和原油价差等等。
- 预测某国GDP、股指、某支股票等时间序列未来的价格变化

前言

- 横截面研究主要研究不同资产在同一时间收益（或者风险等其他我们感兴趣的对象）的差别是因为那些资产的特征差异导致的。
- 比如两家上市公司A和B，A公司收益比B公司高，收益高是什么原因造成的？我们就会去对比两家公司各自的情况，可能是A公司的市值比B公司低，也可能是A公司的市盈率比B公司低。。。
- 和时间序列研究不同，横截面研究的对象是面板数据。

前言

- 如果知道那些特征差异导致了资产价格、收益的差别，我们就可以构造投资策略——量化投资策略。
  - 不同于价值投资、宏观投资等其他种类投资策略，量化投资策略基于横截面研究。
  - 根据交易频率分，量化投资策略分成高频和中低频量化策略。
  - 量化交易数据主要来源于公司财务报表和二级市场交易数据，所以基于公司财务报表的量化策略又称为基本面交易策略，基于二级市场交易数据的称为技术面投资策略。

提纲

- 第一节：股市数据预处理
- 第二节：描述性统计
- 第三节：相关性
- 第四节：持续性分析

第一节：股市数据预处理

一、股市数据预处理

- 横截面实证资产定价的研究数据包括股市、债市、商品期货、电子货币、外汇等等资产。在研究中，我们需要根据研究数据做先期预处理。下面以中美股市为例。
- 中国股市：股票数据来源包括万德数据库、国泰安数据库、锐思数据库等等。大多数研究从1997年(Nartea,Kong, and Wu, 2017)或者2000年开始(Liu,Stambaugh, and Yuan, 2019)，这是因为我国证券市场现行的涨跌停板制度是1996年12月16日开始实施。1996年12月24日，A股上市公司数量突破500家，直到2000年年9月19日，A股上市公司数量才突破1000家。

A股股市数据预处理

- A股股票数据包含所有A股上市公司，但会删去一些股票，删除股票的方法主要是两种。
- 第一种删除股票的方法(Gui and Zhu, 2021)：1)删除所有刚刚上市六个月公司数据（去除IPO效应）；2)删除账面价值市值比为负的公司；3)删除金融行业上市公司；4)删除ST、PT股票；5)删除过去12个月交易日不足120天的股票；6)删除过去一个月交易日不足15天的股票。这里需要注意，A股有些春节所在月份交易日不足15天，像这些不足15天交易日的月份数据还是需要保留的。比如1997年2月（10个交易日），1999年2月（7个交易日），2000年2月（12个交易日），2001年1月（14个交易日），2002年2月（10个交易日），2004年1月（13个交易日），2005年2月（13个交易日）。

美国股市数据预处理

- 第二种删除股票的方法(Liu,Stambaugh, and Yuan, 2019)：1)删除所有刚刚上市六个月公司数据（去除IPO效应）；2)删除市值最小的30%公司（小市值公司具有壳价值）；3)删除过去12个月交易日不足120天的股票或者过去一个月交易日不足15天的股票。

美国股市数据预处理

- 简单介绍一下如何预处理一下美国股市数据。美股数据来源于WRDS（沃顿）数据平台中的CRSP数据库（交易信息数据）和Compustat数据库（基本面数据库）。很多时候我们要求删除价格低于5美元或高于1000美元的股票（因为这类股票流动性差）。CRSP提供了1925年12月31日至今在NYSE（纽交所）、AMEX（美交所）、NASDAQ（纳斯达克）上市证券的数据。很多研究的起始时间是1963年，这是因为1962年7月以前的CRSP数据库不包含美交所股票。从总市值上看，美交所占比一直都很小，像2012年12月，美交所总市值只占三大交易所总市值的0.23%。

股市数据预处理——缩尾和截尾

- 金融数据经常会有极端值，极端值经常对统计分析会有影响，而我们希望能降低极端值的影响。
- 比如，回归分析中的自变量前的回归系数因为是条件均值，容易受到极端值的影响。
- 通常有两个技术被用来处理资产定价中的极端值问题，第一个是**缩尾**技术，简单地将一个大于或者小于某个确定值的变量值设定为这个确定值。第二个是**截尾**技术，简单地将被认定为是极端值的变量值设定为缺失值。

股市数据预处理——缩尾和截尾

- 我们假设正在处理的变量X有n个不同的观测值，记为X1,X2,…,Xn。
- 缩尾处理是将大于X的所有观测值的上h百分位数的X 值设定为100-h百分位数。同样，将小于X的下l百分位数的值设定为X的l百分位数。例如，我们想在0.5%(h=0.5)水平处对X高的一端进行缩尾处理，就需要先计算X 值的99.5百分位数，记为Pctl(99.5,X)。然后，我们将所有大于Pctl(99.5,X)的X 值设定为Pctl(99.5,X) 。
- 现在，我们想在1%(l=1)水平处对X 低的一端进行缩尾处理，就需要先计算X 值的1百分位数，记为Pctl(1,X)，然后，我们将所有小于Pctl(1,X)的X值设定为Pctl(1,X)。在大部分情况下，h和l的值的选取是一样的，而且研究者在进行缩尾处理时常用的值是0.5%和1%。

股市数据预处理——缩尾和截尾

- 截尾处理和缩尾处理非常类似,不同于将大于Pctl(h,X)的X值设定为Pctl(h,X)我们将此值设定为缺失值或者不可利用的值。同样,低于Pctl(h,X)的X值也被设定为缺失值。因此,截尾处理和缩尾处理的不同之处在于,截尾处理会将一个确定变量的极端值从分析样本中移除,而缩尾处理会将极端值设定为更合适的水平。

第二节：描述性统计

二、描述性统计

- 对于一个实证研究者来说，为了理解研究结果并且评估研究范围之外的结果的实用性，对文章分析中所使用的数据有一个粗略的了解非常重要的。因此，大部分实证研究的文章会在讨论主要结果之前对数据进行描述性统计。通常，研究论文的第一个表格会展现这个描述性统计。

二、描述性统计

表 2.1　　β 每年的描述性统计 *

t	$Mean_t$	SD_t	$Skew_t$	$Kurt_t$	Min_t	$P5_t$	$P25_t$	$Median_t$	$P75_t$	$P95_t$	Max_t	n_t
1988	0.46	0.48	0.17	2.80	−4.29	−0.20	0.13	0.40	0.75	1.31	3.28	5 690
1989	0.46	0.53	0.15	1.88	−3.51	−0.27	0.11	0.40	0.79	1.38	3.63	5 519
1990	0.58	0.59	0.23	1.14	−3.15	−0.24	0.16	0.51	0.96	1.61	3.66	5 409
1991	0.57	0.61	0.23	1.96	−3.28	−0.29	0.17	0.52	0.95	1.62	5.29	5 303
1992	0.65	0.83	0.34	6.10	−5.21	−0.50	0.17	0.59	1.09	2.05	9.90	5 389
1993	0.62	0.77	−0.10	4.29	−4.70	−0.56	0.20	0.57	1.04	1.90	7.59	5 670
1994	0.70	0.71	−0.17	6.59	−6.92	−0.32	0.27	0.67	1.07	1.89	6.50	6 148
1995	0.64	0.84	0.30	5.17	−6.32	−0.49	0.19	0.56	1.02	2.15	8.77	6 288
1996	0.67	0.64	0.46	1.97	−4.32	−0.20	0.26	0.59	1.01	1.89	3.98	6 586

二、描述性统计

表 2.2　　β 的平均横截面描述性统计

Mean	*SD*	*Skew*	*Kurt*	*Min*	*P*5	*P*25	*Median*	*P*75	*P*95	*Max*	*n*
0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5 198

该表展示了 β 每年横截面描述性统计的时间序列均值。表中展示了均值（*Mean*）、标准差（*SD*）、偏度（*Skew*）、超额峰度（*Kurt*）、最小值（*Min*）、第 5 百分位数（*P*5）、第 25 百分位数（*P*25）、中位数（*Median*）、第 75 百分位数（*P*75）、第 95 百分位数（*P*95）和最大值（*Max*），这里均值由样本中所有年份计算得出。标记为 n 的列显示的是具有有效 β 观测值的个数均值。

二、描述性统计

表 2.3　　β、*MktCap* 和 *BM* 的描述性统计

	Mean	*SD*	*Skew*	*Kurt*	*Min*	5%	25%	*Median*	75%	95%	*Max*	*n*
β	0.75	0.62	0.34	1.78	−2.78	−0.13	0.33	0.71	1.12	1.85	4.40	5 198
MktCap	2 030	10 230	14.20	282.85	0	9	48	188	802	7 524	287 033	5 550
Size	5.04	2.07	0.32	−0.07	−1.19	1.89	3.56	4.91	6.39	8.70	12.33	5 550
BM	0.71	2.90	−9.49	1 226.68	−124.31	0.05	0.29	0.57	0.97	2.11	44.87	4 273
r_{t+1}	12.40	80.83	5.94	125.33	−97.86	−67.46	−26.87	0.90	31.84	124.54	1 841.43	5 381

该表展示了样本的描述性统计。样本期间包括 1988—2012 年，而且包含 CRSP 数据库中所有的美国普通股。每一年，计算每个变量的横截面分布均值（*Mean*）、标准差（*SD*）、偏度（*Skew*）、峰度（*Kurt*）、第 5 百分位数（5%）、第 25 百分位数（25%）、中位数（*Median*）、第 75 百分位数（75%）、第 95 百分位数（95%）和最大值（*Max*）。这个表展示了每个横截面分布的时间序列均值。标记为 n 的列表示有效变量的股票的平均只数。β 是通过在 t 年用日度数据的超额收益率对市场超额收益率回归计算出来的股票的贝塔值。*MktCap* 以百万美元为单位，是在 t 年最后一个交易日计算出来的股票市值。*Size* 是 *MktCap* 的自然对数。*BM* 是股票的账面价值对市值的比。r_{t+1} 是未来第一年的超额收益率。

二、描述性统计——分组展示

Panel A: Summary statistics for decile portfolios of stocks sorted by *VaR*1

Decile	*VaR*1	*ES*1	*SIZE*	*BM*	*MOM*	*TURN*	β	*ISKEW*	*MAX*	*IVOL*
1(lowest) *VaR*1	4.953	5.781	8.145	6.565	13.773	0.208	0.879	0.337	4.110	1.492
2	5.923	6.833	7.730	6.552	11.540	0.270	1.008	0.355	4.544	1.650
3	6.368	7.303	7.579	6.538	11.573	0.307	1.062	0.378	4.772	1.741
4	6.702	7.632	7.474	6.506	11.622	0.342	1.104	0.386	4.959	1.825
5	7.000	7.935	7.388	6.493	11.667	0.370	1.128	0.391	5.011	1.877
6	7.296	8.221	7.352	6.462	13.958	0.389	1.153	0.395	5.148	1.929
7	7.613	8.512	7.300	6.431	13.117	0.408	1.170	0.387	5.241	1.982
8	7.991	8.835	7.272	6.404	14.091	0.437	1.194	0.389	5.371	2.056
9	8.478	9.244	7.201	6.385	13.806	0.473	1.223	0.389	5.546	2.143
10(highest)*VaR*1	9.272	9.803	7.198	6.322	16.502	0.540	1.237	0.387	5.796	2.299

第三节：相关性

三、相关性

- 描述性统计对研究中使用的变量的单变量分布提供一个直观了解。然而，描述性统计对变量之间的关系没有给出任何明示。在绝大多数情况下，了解变量之间的关系比了解变量的单变量属性通常来说更重要——关系才是研究的重点。因此，除了展现单变量描述性统计，研究者通常还会展示主要变量之间的相关性。
- 现在，我们介绍两种不同的测量相关性系数的方法。第一种是皮尔逊积矩相关系数，它用来测度两个变量之间线性关系的强度。第二种是斯皮尔曼等级相关系数，它检测了两个变量之间的单调性关系。

三、相关性

- 对每一个时段t，我们计算X和Y之间的皮尔逊积矩相关系数和斯皮尔曼等级相关系数。t时刻X和Y之间的皮尔逊积矩相关系数被定义为

$$\rho_t(X,Y) = \frac{\sum_{i=1}^{n_t}[(X_{i,t}-\bar{X}_t)(Y_{i,t}-\bar{Y}_t)]}{\sqrt{\sum_{i=1}^{n_t}(X_{i,t}-\bar{X}_t)^2}\sqrt{\sum_{i=1}^{n_t}(Y_{i,t}-\bar{Y}_t)^2}}$$

- 其中每一个求和是对t时刻样本中X和Y的第i个有效的元素同时求和，$\bar{X}_t$ 和 $\bar{Y}_t$分别为 $X_{i,t}$和 $Y_{i,t}$的样本均值，选取的是同样的元素集。这里，n_t是在给定t时刻X和Y同时有效的元素的个数。

三、相关性

- 为了计算斯皮尔曼等级相关系数，我们首先必须计算X和Y每一个元素的等级。我们记$X_{i,t}$的等级为$x_{i,t}$，它是由t时刻X和Y同时有效的元素计算所得。因此，如果元素i是X的最小值，那么$x_{i,t}$就是1。如果元素i是X的最大值，那么$x_{i,t}$就是n_t。如果X中多个元素的值是一样的，那么这些元素就会被分配为一个等级，即X所有元素排序之后的这些具有相同值元素的平均位置。Y的等级计算方法相同，记为$y_{i,t}$。值得注意的是，在计算斯皮尔曼等级相关系数时，数据不可以进行缩尾处理。

三、相关性

- 对每一个元素i，X中元素的等级和Y中元素的等级之间的差值被定义为$d_{i,t} = x_{i,t} - y_{i,t}$。最后，t时刻X和Y的斯皮尔曼等级相关系数计算公式如下

$$\rho_t^S(X, Y) = 1 - \frac{6\sum_{i=1}^{n_t} d_{i,t}^2}{n_t(n_t^2 - 1)}.$$

三、相关性

表 3.3　　β、*Size*、*BM* 和 r_{t+1} 的相关系数

	β	*Size*	*BM*	r_{t+1}
β	—	0.42	−0.23	−0.04
Size	0.39	—	−0.22	0.06
BM	−0.18	−0.23	—	0.08
r_{t+1}	−0.04	−0.02	0.06	—

该表展示了变量 β、*Size*、*BM* 和 r_{t+1} 两两之间的年横截面皮尔逊积矩相关系数和斯皮尔曼等级相关系数的时间序列均值。下三角的元素代表皮尔逊积矩相关系数的均值。上三角的元素代表斯皮尔曼等级相关系数的均值。

三、相关性

Panel B: Correlations of different measures of VaRs and other stock characteristics

Decile	1% ES	5% ES	10% ES	1% VaR	5% VaR	10% VaR	SIZE	BM	MOM	IVOL
1% ES	1.000									
5% ES	0.937	1.000								
10% ES	0.894	0.990	1.000							
1% VaR	0.901	0.970	0.957	1.000						
5% VaR	0.792	0.943	0.975	0.904	1.000					
10% VaR	0.763	0.910	0.954	0.872	0.967	1.000				
SIZE	−0.400	−0.464	−0.476	−0.454	−0.471	−0.462	1.000			
BM	−0.101	−0.119	−0.120	−0.119	−0.121	−0.111	−0.218	1.000		
MOM	−0.040	−0.006	0.003	−0.000	0.022	0.013	0.011	−0.244	1.000	
IVOL	0.604	0.671	0.680	0.646	0.665	0.659	−0.402	−0.086	0.0187	1.000

第四节：持续性分析

四、持续性分析

- 实证资产定价研究当中很多变量被用来捕获样本元素的持续性特征。这意味着通过给定变量所捕获的元素特征在一定时期内被合理假定是稳定的。这样的变量通常由历史数据估计而得，并且由历史数据计算出来的值被认为是对元素未来给定特征的一个很好估计。
- 横截面资产定价是假定同一时刻下，资产收益率的差异是由哪些特征造成的，时间必须一致。但在实际市场解释或者预测当期或未来资产的收益时，我们没办法知道同一时间的特征值，所以必须用过去的某一时刻的特征值来作为当期特征值的指标。如果持续性不好，我们就不能用历史特征值来作为当期特征值的合理指标。我们以资产市值预测月度收益率为例。

四、持续性分析

- 这一节，我们将讨论一个技术，我们将之称为持续性分析。我们用持续性分析检验元素的一个给定特性是否真的具有持续性。持续性分析也能够被用来检验问题变量捕获期望特征的能力。基本的方法是检验在两个不同时点测量的给定变量的横截面相关系数。如果这个相关性系数是高的，就意味着这个变量是持续性的；低的相关性显示较少或者没有持续性。

四、持续性分析

表 4.1　　β 持续性分析

t	$\rho_{t,t+1}(\beta)$	$\rho_{t,t+2}(\beta)$	$\rho_{t,t+3}(\beta)$	$\rho_{t,t+4}(\beta)$	$\rho_{t,t+5}(\beta)$
1988	0.50	0.48	0.47	0.39	0.34
1989	0.52	0.45	0.38	0.35	0.36
1990	0.55	0.45	0.42	0.40	0.37
1991	0.46	0.43	0.41	0.37	0.40
1992	0.39	0.37	0.36	0.41	0.38
1993	0.40	0.33	0.38	0.39	0.39
1994	0.38	0.39	0.38	0.37	0.33
1995	0.46	0.44	0.38	0.40	0.48

四、持续性分析

表 4.2　　β 的平均持续性

$\rho_1(\beta)$	$\rho_2(\beta)$	$\rho_3(\beta)$	$\rho_4(\beta)$	$\rho_5(\beta)$
0.61	0.53	0.48	0.46	0.42

该表展示了在年份 t 测量的 β 与在年份 $t+\tau$ 测量的 β 之间的横截面皮尔逊积矩相关系数的时间序列均值，其中 $\tau \in \{1,2,3,4,5\}$。

表 4.3　　β、*Size* 和 *BM* 的持续性

τ	β	*Size*	*BM*
1	0.61	0.96	0.74
2	0.53	0.92	0.59
3	0.48	0.90	0.50
4	0.46	0.89	0.46
5	0.42	0.88	0.43

该表展示了 β、*Size* 和 *BM* 的持续性结果。对每一年 t，计算时段 t 测量的给定变量和 $t+\tau$ 时刻测量的给定变量之间的横截面相关系数。这个表展示了每年的横截面相关系数的均值。标记为 τ 的列表示持续性的滞后期数。

四、持续性分析

- 除了看相关系数，我们还可以检测转移矩阵(Transition Matrix)来分析测度的持续性。(下图来自Atilgan,Bali,Demirtas, and Gunaydin, 2019)

Table 7
Transition matrix.

This table presents transition probabilities for VaR1 at a lag of 12 months between 1962 and 2014. At each month t, all stocks are sorted into deciles based on an ascending ordering of VaR1. The procedure is repeated in month $t+12$. Portfolio 1 is the portfolio of stocks with the lowest value-at-risk and Portfolio 10 is the portfolio of stocks with the highest value-at-risk. For each VaR1 decile in month t, the percentage of stocks that fall into each of the month $t+12$ VaR1 decile is calculated. Table presents the time-series averages of these transition probabilities. Each row corresponds to a different month t VaR1 portfolio and each column corresponds to a different month $t+12$ VaR1 portfolio.

	Port1 (%)	Port2 (%)	Port3 (%)	Port4 (%)	Port5 (%)	Port6 (%)	Port7 (%)	Port8 (%)	Port9 (%)	Port10 (%)
Port1	52	22	11	6	4	2	1	1	1	1
Port2	23	26	19	12	8	5	3	2	1	1
Port3	12	21	20	16	12	8	5	3	2	1
Port4	6	14	18	17	14	11	8	5	4	2
Port5	4	9	14	16	16	14	11	8	6	3
Port6	2	6	9	13	16	15	14	11	8	6
Port7	1	3	6	10	13	15	16	15	12	9
Port8	1	2	4	6	10	14	17	18	16	13
Port9	0	1	2	4	7	11	15	19	21	20
Port10	0	1	1	2	4	7	11	17	24	33