华泰证券 HUATAI SECURITIES

金工研究/深度研究

2018年10月18日

林晓明 执业证书编号：S0570516010001
研究员 0755-82080134
linxiaoming@htsc.com

陈烨 执业证书编号：S0570518080004
研究员 010-56793927
chenye@htsc.com

相关研究

1《金工：周期视角下的因子投资时钟》2018.10
2《金工：财务质量因子在行业配置中的应用》2018.08
3《金工：人工智能选股之损失函数的改进》2018.08

基于回归法的基金持股仓位测算

基金仓位分析专题报告

仓位测算的基本思路是基于日频基金净值和一级行业指数的回归

本报告探讨了几种基于回归法的基金持股仓位测算模型，是以基金的日频收益率作为因变量，29个一级行业指数日收益率作为自变量，通过不同的回归方式对各行业变量前的回归系数进行拟合，再求所有回归系数之和，即得基金仓位预测值。我们对四种回归方式（主成分回归、逐步回归、岭回归、Lasso 回归）在普通股票型基金和偏股混合型基金的测试结果进行对比，发现 Lasso 回归和逐步回归的预测精度较高，主成分回归稍弱，岭回归存在系统性高估现象。

主成分回归、逐步回归、岭回归、Lasso 回归均能缓解自变量共线性问题

基金仓位测算回归模型中，自变量组（29个一级行业日收益率）存在明显的多重共线性，若直接采用普通最小二乘回归进行求解，则各行业变量前面的拟合系数会互相干扰，出现不合理的回归结果，并且共线性严重时回归方程无法通过数值方法求解。主成分回归可以将自变量组转化成互相正交的几个主成分；逐步回归可以选择一个自变量的子集进行回归；岭回归和 Lasso 回归则是在普通最小二乘的损失函数基础上添加正则化项，使原本较为病态的回归问题可以正常求解。四种回归方式均能缓解自变量的多重共线性问题。

仓位预测效果：Lasso 回归和逐步回归较好，岭回归相对较差

在普通股票型基金和偏股混合型基金中，主成分回归、逐步回归、Lasso 回归的预测误差大多落在[5%，15%]区间范围内，Lasso 回归和逐步回归的结果稍好于主成分回归，岭回归则存在系统性高估的现象。岭回归与另外三种回归方法最大的区别是不存在降维行为，主成分回归是通过主成分分析法将解释变量降维，逐步回归和 Lasso 回归的拟合结果中只有部分解释变量的回归系数不为零。岭回归的回归系数并不存在稀疏化特征，基本每个行业变量前面回归系数都不为零，我们猜测这可能是导致岭回归存在系统性高估现象的原因。

回归时间窗口长度敏感性：大于 30 天预测效果平稳，但也不宜长于 60 天

我们选取 2017 年四季度末、2018 年一季度末、二季度末三个横截面，在普通股票型基金和偏股混合型基金中对四种回归方法进行时间窗口长度敏感性测试，将窗口长度从 15 天到 59 天进行遍历，发现大部分情况下，各方法的预测误差均值在窗口长度大于 30 天之后比较平稳，趋于一个稳定的值，说明各方法的解已经收敛；在小于 30 天时没有明显规律。因为回归系数的实际含义是过去一段时间窗口内基金仓位的平均状况，并用这个值代表我们对当前时刻基金仓位的预测值，所以窗口长度也不宜取得太长（一般没有必要超过一个季度，约 60 个交易日），否则预测结果可能会滞后。

风险提示：本报告中所采用的基金仓位测算方法仅基于日频基金净值数据和行业数据，没有利用基金报告中公布的重仓股、行业分布等信息，存在一定局限性。本报告中所采用的基金仓位测算方法仅在普通股票型基金和偏股混合型基金中进行实证，在其它类别基金中可能不适用。本报告中采用的四种回归方法只能缓解自变量间的多重共线性，并不能完全解决这一问题，敬请注意。

正文目录

图表目录

研究背景

基金仓位是指基金持有的股票资产占基金资产的比例。公募基金作为 A 股市场中重要的机构投资者，其持股动向一直受到市场关注。一方面，出于对基金经理投资能力的认可，股票投资人一般认为公募基金的仓位变动反映了市场投资情绪变化等关键信息；另一方面，基金投资人也会随时关注其持有的基金的仓位变动情况，辅助自己的投资决策。然而，公募基金只在每个季度末对其资产配置情况进行披露，这就使得投资者与基金管理者之间存在一种相对的信息不对称性。所以对基金仓位的测算研究成为一项有意义的工作。

目前常见的基金仓位测算方法主要基于传统意义上的指数模拟法，即运用基金净值数据和指数点位数据进行回归计算。理论意义上的指数模拟法可以选取单一指数，也可以选取复合指数为基准。对于单一指数，主要选取市场覆盖性强、具有代表性的单个指数，而复合指数通常选取一组代表不同投资风格的指数进行加权合成。本篇报告也依循这一传统测算思路，采取几种不同的回归方式，对普通股票型基金和偏股混合型基金进行仓位测算，并与真实结果进行比对，评价几种方法的优劣。

基金仓位测算方法

数据选取

我们使用 29 个一级行业指数日收益率作为回归模型的自变量，基金的日频收益率作为因变量，通过几种不同的多元线性回归方式，进行基金仓位测算。本文用于测试的样本主要为 Wind 分类下的普通股票型基金和偏股混合型基金，剔除沪港深基金等非完全投资于 A 股市场的基金，剔除数据方面有缺失值或存在疑问的基金，剔除 2017 年以后成立的基金，共保留 190 只普通股票型基金及 414 只偏股混合型基金作为我们测试的样本。

行业指数的共线性及对回归方程的影响

以 29 个一级行业指数日收益率作为自变量、基金的日频收益率作为因变量的回归方程如下所示：

$$R_{f,t} = \sum_i \gamma_i R_{i,t} + \varepsilon$$

其中$R_{f,t}$为基金 f 在 t 日的收益率，$R_{i,t}$为一级行业指数 i 在 t 日的收益率，γ_i为待拟合回归系数，ε为残差项。我们认为γ_i代表基金投资于行业 i 的股票占比，那么$\sum_i \gamma_i$即为基金持有的股票资产占基金资产比例。

如果根据以上方程，直接使用 OLS 回归，会得到什么样的拟合结果呢？首先，我们不妨取一个例子进行简单试验。以某普通股票型基金为例（采样方式不重要，在大部分股票型基金中都能推出类似结论），采用 2018.5.18～2018.6.29（共 30 个交易日）内的数据进行回归，得到该支基金在这段时间内的持仓预测，分别为石油石化 12.93%，煤炭-7.31%，有色金属-7.14%，电力及公用事业-8.96%，钢铁-7.21%，基础化工-23.75%，建筑-38.94%，建材 22.59%，轻工制造-38.05%，机械-23.55%，电力设备 38.48%，国防军工 13.45%，汽车 1.07%，商贸零售-3.54%，餐饮旅游-1.25%，家电-32.85%，纺织服装 7.80%，医药 27.99%，食品饮料 30.11%，农林牧渔-23.13%，银行 9.68%，非银行金融 2.47%，房地产 30.88%，交通运输 35.41%，电子元器件 34.03%，通信 1.72%，计算机 21.66%，传媒-1.62%，综合-0.28%。许多行业的拟合权重是较大负值，结果欠缺合理性。

实际上，这是由于股市的系统性风险及行业间联动效应，各一级行业指数之间存在较强的共线性，使回归问题变得比较病态，难以取得可靠的结果。

图表1： 一级行业间相关系数矩阵（2017.1.1～2018.8.10）

	石油石化	煤炭	有色金属	电力及公用事业	钢铁	基础化工	建筑	建材	轻工制造	机械	电力设备	国防军工	汽车	商贸零售	餐饮旅游	家电	纺织服装	医药	食品饮料	农林牧渔	银行	非银行金融	房地产	交通运输	电子元器件	通信	计算机	传媒	综合
石油石化	1.00	0.63	0.60	0.63	0.59	0.68	0.64	0.68	0.64	0.66	0.60	0.50	0.61	0.60	0.47	0.42	0.60	0.47	0.37	0.58	0.40	0.55	0.58	0.65	0.53	0.53	0.50	0.56	0.61
煤炭	0.63	1.00	0.73	0.52	0.82	0.58	0.56	0.64	0.56	0.53	0.47	0.37	0.52	0.46	0.37	0.34	0.47	0.29	0.29	0.43	0.35	0.45	0.51	0.56	0.41	0.39	0.34	0.42	0.54
有色金属	0.60	0.73	1.00	0.68	0.75	0.79	0.61	0.73	0.76	0.74	0.72	0.57	0.70	0.68	0.52	0.38	0.70	0.51	0.37	0.63	0.19	0.43	0.54	0.66	0.65	0.63	0.61	0.66	0.75
电力及公用事业	0.63	0.52	0.68	1.00	0.59	0.85	0.73	0.76	0.84	0.88	0.84	0.70	0.82	0.80	0.59	0.47	0.84	0.65	0.42	0.70	0.21	0.51	0.62	0.79	0.76	0.76	0.74	0.80	0.85
钢铁	0.59	0.82	0.75	0.59	1.00	0.66	0.61	0.75	0.63	0.62	0.57	0.45	0.57	0.53	0.42	0.35	0.54	0.35	0.30	0.46	0.28	0.45	0.57	0.60	0.48	0.48	0.44	0.50	0.63
基础化工	0.68	0.58	0.79	0.85	0.66	1.00	0.74	0.85	0.92	0.94	0.90	0.74	0.87	0.87	0.67	0.51	0.90	0.72	0.49	0.76	0.18	0.52	0.65	0.80	0.85	0.85	0.83	0.86	0.87
建筑	0.64	0.56	0.61	0.73	0.61	0.74	1.00	0.77	0.72	0.78	0.70	0.59	0.72	0.68	0.49	0.47	0.70	0.50	0.38	0.61	0.36	0.54	0.66	0.75	0.62	0.62	0.61	0.66	0.69
建材	0.68	0.64	0.73	0.76	0.75	0.85	0.77	1.00	0.81	0.82	0.77	0.60	0.77	0.74	0.59	0.53	0.76	0.61	0.46	0.68	0.28	0.53	0.72	0.77	0.70	0.69	0.64	0.69	0.79
轻工制造	0.64	0.56	0.76	0.84	0.63	0.92	0.72	0.81	1.00	0.92	0.88	0.71	0.87	0.89	0.70	0.54	0.91	0.76	0.52	0.77	0.20	0.53	0.68	0.80	0.82	0.83	0.79	0.85	0.86
机械	0.66	0.53	0.74	0.88	0.62	0.94	0.78	0.82	0.92	1.00	0.93	0.80	0.88	0.88	0.66	0.48	0.91	0.72	0.44	0.76	0.18	0.52	0.64	0.81	0.86	0.88	0.88	0.90	0.88
电力设备	0.60	0.47	0.72	0.84	0.57	0.90	0.70	0.77	0.88	0.93	1.00	0.74	0.87	0.86	0.66	0.51	0.87	0.74	0.50	0.78	0.13	0.49	0.57	0.76	0.87	0.87	0.83	0.87	0.84
国防军工	0.50	0.37	0.57	0.70	0.45	0.74	0.59	0.60	0.71	0.80	0.74	1.00	0.70	0.70	0.48	0.30	0.72	0.58	0.30	0.63	0.07	0.37	0.46	0.63	0.71	0.73	0.77	0.74	0.72
汽车	0.61	0.52	0.70	0.82	0.57	0.87	0.72	0.77	0.87	0.88	0.87	0.70	1.00	0.84	0.66	0.63	0.84	0.72	0.57	0.73	0.27	0.59	0.64	0.79	0.84	0.81	0.77	0.82	0.80
商贸零售	0.60	0.46	0.68	0.80	0.53	0.87	0.68	0.74	0.89	0.88	0.86	0.70	0.84	1.00	0.69	0.55	0.89	0.75	0.54	0.77	0.16	0.53	0.62	0.78	0.80	0.81	0.78	0.85	0.81
餐饮旅游	0.47	0.37	0.52	0.59	0.42	0.67	0.49	0.59	0.70	0.66	0.66	0.48	0.66	0.69	1.00	0.50	0.69	0.71	0.57	0.61	0.13	0.41	0.51	0.62	0.67	0.61	0.61	0.65	0.59
家电	0.42	0.34	0.38	0.47	0.35	0.51	0.47	0.53	0.54	0.48	0.51	0.30	0.63	0.55	0.50	1.00	0.48	0.57	0.78	0.45	0.34	0.58	0.56	0.52	0.58	0.50	0.38	0.45	0.42
纺织服装	0.60	0.47	0.70	0.84	0.54	0.90	0.70	0.76	0.91	0.91	0.87	0.72	0.84	0.89	0.69	0.48	1.00	0.72	0.48	0.76	0.14	0.46	0.61	0.77	0.80	0.81	0.80	0.86	0.84
医药	0.47	0.29	0.51	0.65	0.35	0.72	0.50	0.61	0.76	0.72	0.74	0.58	0.72	0.75	0.71	0.57	0.72	1.00	0.64	0.65	0.09	0.44	0.48	0.62	0.76	0.72	0.69	0.72	0.61
食品饮料	0.37	0.29	0.37	0.42	0.30	0.49	0.38	0.46	0.52	0.44	0.50	0.30	0.57	0.54	0.57	0.78	0.48	0.64	1.00	0.46	0.21	0.50	0.44	0.49	0.54	0.45	0.37	0.43	0.39
农林牧渔	0.58	0.43	0.63	0.70	0.46	0.76	0.61	0.68	0.77	0.76	0.78	0.63	0.73	0.77	0.61	0.45	0.76	0.65	0.46	1.00	0.12	0.43	0.60	0.69	0.69	0.70	0.67	0.72	0.72
银行	0.40	0.35	0.19	0.21	0.28	0.18	0.36	0.28	0.20	0.18	0.13	0.07	0.27	0.16	0.13	0.34	0.14	0.09	0.21	0.12	1.00	0.67	0.45	0.36	0.14	0.13	0.08	0.14	0.15
非银行金融	0.55	0.45	0.43	0.51	0.45	0.52	0.54	0.53	0.53	0.52	0.49	0.37	0.59	0.53	0.41	0.58	0.46	0.44	0.50	0.43	0.67	1.00	0.61	0.58	0.53	0.51	0.42	0.49	0.46
房地产	0.58	0.51	0.54	0.62	0.57	0.65	0.66	0.72	0.68	0.64	0.57	0.46	0.64	0.62	0.51	0.56	0.61	0.48	0.44	0.60	0.45	0.61	1.00	0.67	0.52	0.53	0.45	0.55	0.62
交通运输	0.65	0.56	0.66	0.79	0.60	0.80	0.75	0.77	0.80	0.81	0.76	0.63	0.79	0.78	0.62	0.52	0.77	0.62	0.49	0.69	0.36	0.58	0.67	1.00	0.70	0.70	0.65	0.72	0.76
电子元器件	0.53	0.41	0.65	0.76	0.48	0.85	0.62	0.70	0.82	0.86	0.87	0.71	0.84	0.80	0.67	0.58	0.80	0.76	0.54	0.69	0.14	0.53	0.52	0.70	1.00	0.91	0.86	0.84	0.75
通信	0.53	0.39	0.63	0.76	0.48	0.85	0.62	0.69	0.83	0.88	0.87	0.73	0.81	0.81	0.61	0.50	0.81	0.72	0.45	0.70	0.13	0.51	0.53	0.70	0.91	1.00	0.88	0.87	0.77
计算机	0.50	0.34	0.61	0.74	0.44	0.83	0.61	0.64	0.79	0.88	0.83	0.77	0.77	0.78	0.61	0.38	0.80	0.69	0.37	0.67	0.08	0.42	0.45	0.65	0.86	0.88	1.00	0.90	0.75
传媒	0.56	0.42	0.66	0.80	0.50	0.86	0.66	0.69	0.85	0.90	0.87	0.74	0.82	0.85	0.65	0.45	0.86	0.72	0.43	0.72	0.14	0.49	0.55	0.72	0.84	0.87	0.90	1.00	0.79
综合	0.61	0.54	0.75	0.85	0.63	0.87	0.69	0.79	0.86	0.88	0.84	0.72	0.80	0.81	0.59	0.42	0.84	0.61	0.39	0.72	0.15	0.46	0.62	0.76	0.75	0.77	0.75	0.79	1.00

资料来源：Wind，华泰证券研究所

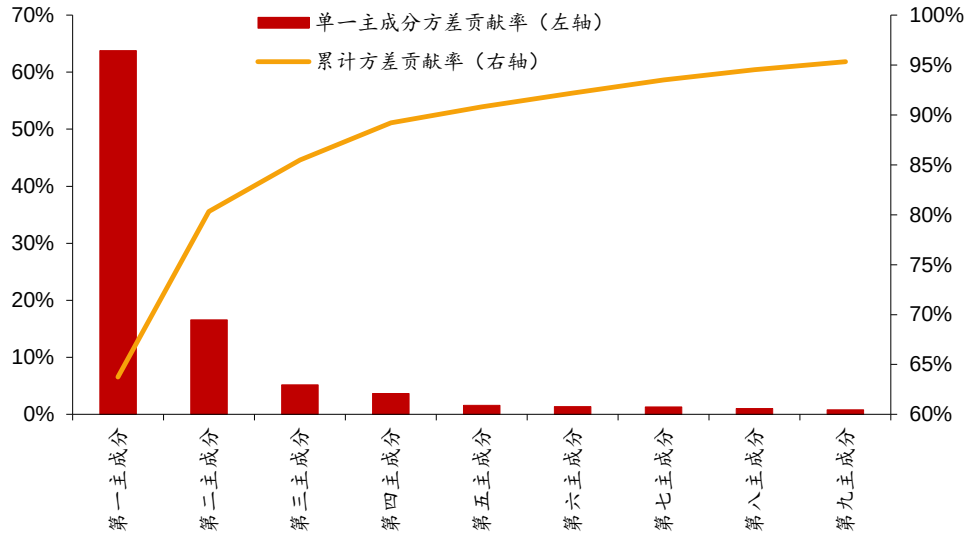
上面图表展示了各个行业日收益率序列的相关系数（以 2017.1.1～2018.8.10 的数据进行计算）。我们发现，相关系数在 0.5 以上的比例超过了 70%，各一级行业指数日收益率之间确实存在明显的共线性。为了解决这一问题，我们尝试了几种方法对 OLS 回归进行改进，下面将进行详细介绍。

主成分回归

主成分回归的基本原理是通过正交变换将一组可能存在相关性的变量进行压缩，转换为一组线性不相关的变量，转换后的这组变量叫主成分。但是用主成分得到的回归关系不像用原自变量建立的回归关系那样容易解释，因此常见的处理方法是用主成分分析法对回归模型的自变量进行处理，将得到的主成分变量作为自变量进行回归分析，然后根据转换矩阵将原自变量代回模型，得到原自变量的拟合系数。

在实际操作中，首先利用主成分分析法对 29 个一级行业日收益率数据进行主成分提取。我们选取累计方差贡献率达到 95%的前几个主成分构成自变量组，以基金的日收益率作为因变量进行回归，可以拟合得到各主成分前面的回归系数。下图展示了一个小例子，以 2017.1.1～2018.8.10 的数据进行计算，29 个行业日收益率变量的前九个主成分就达到了累计方差贡献率 95%以上。因为每个主成分变量都可以表达为原 29 个行业日收益率变量的线性组合，所以可将主成分前面的回归系数还原成 29 个行业日收益率变量的系数，从而得到各个行业的权重。这里需要注意的是，线性回归采样时间段内包含的交易日个数需大于提取的主成分的个数，否则无法求解出结果，后面将要讲述的回归模型也都会面临这个问题，就不再一一赘述了。

图表2： 一级行业日收益率变量组主成分分析的累计方差贡献率（2017.1.1～2018.8.10）



资料来源：Wind，华泰证券研究所

逐步回归

主成分回归为了解决共线性，构造了一组新的线性无关的主成分变量，但主成分变量欠缺经济学意义，且在信息解读方面比较困难。下面我们将探讨一种新的回归方式——逐步回归，一定程度上可以缓解以上两个问题。其基本思想是有进有出，将变量一个一个引入，并对已选入的变量进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，则将其剔除。每引入一个变量或剔除一个变量都要进行 F 检验，以确保每次引入新的变量前回归方程只包含显著的变量，直到不再有变量被选入或剔除为止，保证最后所得回归子集是最优子集。对于变量引入的顺序，本篇报告通过对 29 个行业与因变量（单只基金）的相关系数进行排序，按照相关性从大到小的顺序依次引入。变量被保留的 P 值为 0.05，被剔除的 P 值为 0.1。

也就是说，在逐步回归过程中，我们逐渐剔除掉共线性较强的行业，保留相对独立的剩余板块，假设基金只在这些板块进行配置，从而得到一个相对有解释效力的回归系数，加和便得股票仓位预测值。

下面举一个小例子说明逐步回归法的结果存在一定合理性。我们为了预测基金在第 T 个交易日收盘时的持仓权重，取 T-29～T 交易日的基金收益率数据和 29 个一级行业指数收益率数据，采用逐步回归法，得到逐步回归中入选自变量前面的回归系数。下表展示了某只军工指数基金在每个季度末的回归结果。因为中证军工指数成份股在我们所使用的 29 个一级行业分类下，大部分被归入国防军工行业，少量被归入通信行业，所以下表回归结果基本合理，并且在基金半年报、年报上会披露详细持仓数据，由此计算的真实行业权重也与我们的回归结果大致相符。

图表3： 对于某军工指数基金进行逐步回归法季末仓位预测

	20170630	20170929	20171229	20180330	20180629
国防军工	79.98%	74.80%	80.47%	77.72%	86.47%
通信	18.61%	15.06%	11.28%	24.28%	17.63%
电力及公用事业	0	0	10.85%	0	0
综合	0	0	0	-22.51%	-14.95%
钢铁	-4.18%	0	0	0	0
非银行金融	-3.64%	0	0	0	0
其它行业	0	0	0	0	0

资料来源：Wind，华泰证券研究所

逐步回归方法用于行业配置较为集中的基金仓位预测，效果还是比较有保障的，但上表中仍然出现了一些负的回归系数，结果称不上完美。而且，该方法对行业配置较为分散的基金预测准确性可能受限。

岭回归

在逐步回归之外，我们还尝试使用岭回归和 Lasso 回归对基金仓位进行预测。

岭回归是一种适用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法。对于本章第二小节中提到的回归方程

$$R_{f,t} = \sum_i \gamma_i R_{i,t} + \varepsilon,$$

假设我们在预测基金交易日 T 收盘的持仓权重时，取交易日 T-K～T 的数据进行回归，记回归方程中因变量$Y^{(T)} = (R_{f,T-K}, R_{f,T-K+1}, \cdots, R_{f,T})^T$，自变量$X^{(T)} = (\mathrm{X}_1^{(T)}, \mathrm{X}_2^{(T)}, \cdots, \mathrm{X}_{29}^{(T)})$，其中$\mathrm{X}_i^{(T)} = (R_{i,T-K}, R_{i,T-K+1}, \cdots, R_{i,T})^T$，待拟合系数$\gamma^{(T)} = (\gamma_1, \gamma_2, \cdots, \gamma_{29})^T$，则该 OLS 回归的损失函数为

$$J(\gamma^{(T)}) = \left\| Y^{(T)} - X^{(T)}\gamma^{(T)} \right\|_2^2$$

损失函数达到最小值时的系数$\gamma^{(T)}$就是上面回归方程的解，此时

$$\gamma^{(T)} = ((X^{(T)})^T X^{(T)})^{-1}(X^{(T)})^T Y^{(T)}$$

$X^{(T)}$的列向量之间存在多重共线性，即$(X^{(T)})^T X^{(T)}$可能是一个病态矩阵，造成该回归问题求解困难或数值解不稳定。在岭回归中，对损失函数引入一个惩罚项

$$J(\gamma^{(T)}) = \|Y^{(T)} - X^{(T)}\gamma^{(T)}\|_2^2 + \lambda\|\gamma^{(T)}\|_2^2$$

则该问题的解变成

$$\gamma_{ridge}^{(T)} = ((X^{(T)})^T X^{(T)} + \lambda I)^{-1}(X^{(T)})^T Y^{(T)}$$

其中，λ为一个可调参数，称为岭参数。岭回归相较于普通的 OLS 回归，对病态问题的容忍度提升很多。病态回归问题的数值解容易出现很大或很小的异常解，而岭回归的惩罚项起到了限制数值解的范数的作用，减轻过拟合风险。但与此同时，岭回归得到的拟合系数是有偏的。本文考虑到损失函数$J(\gamma^{(T)})$中相加的两项的量级，λ宜选择10^{-3}左右，此处我们直接取λ=0.002（实际上在10^{-3}这个量级上，λ变动不会对预测结果产生太大影响）。最后用所有自变量前拟合系数之和当作本期仓位预测值。

Lasso 回归

Lasso 回归的原理与岭回归有相似之处，岭回归的损失函数相较于普通 OLS 回归添加了一个 L2 惩罚项，而 Lasso 使用的是 L1 惩罚项。Lasso 回归的损失函数具体表达式为

$$J(\gamma^{(T)}) = \|Y^{(T)} - X^{(T)}\gamma^{(T)}\|_2^2 + \lambda\|\gamma^{(T)}\|_1$$

Lasso 回归主要的作用是使回归系数稀疏化，即寻找有用的解释变量，减少冗余，提高回归预测准确性。实际上，稀疏约束最直观的形式应该是采用 L0 惩罚项，亦即用回归系数中非零元个数之和当作惩罚项，但 L0 范数是不连续且非凸的，这是一个 NP 难问题，难以求解。L1 范数是 L0 范数的最优凸近似，在一定条件下，用 L1 范数替代 L0 范数也可以达到稀疏约束的效果。L1 范数易于求解，所以大部分用到稀疏约束的场景都是在使用 L1 范数。

Lasso 回归从逻辑上来讲，是比较适合本文中提出的基金仓位预测问题的。因为各行业指数的日收益率向量间存在多重共线性，Lasso 回归可以将某些行业前面的回归系数压缩成 0，提取出一组“回归效果最好”的行业组作为解释变量组，而不依赖于解释变量的预设排序或人工选择过程，更为科学，且不会陷入局部解。对于同样具有稀疏化选取作用的逐步回归而言，以上是 Lasso 回归的相对优势。

Lasso 回归也具有一个可调参数λ，此处与岭回归相同，仍取λ=0.002。因岭回归与 Lasso 回归属于机器学习算法，感兴趣的投资人可以参阅华泰金工研报《人工智能选股之广义线性模型》（2017.6.22）了解更多详情。

基金仓位测算方法效果对比

在普通股票型基金中测试效果对比

本篇报告对 190 只普通股票型基金（数据选取详见上一大章第一小节），分别在 2017 年二季度末至 2018 年二季度末，共计 5 个季末横截面进行仓位测算。我们为了预测基金在第 T 个交易日收盘时的持仓权重，取 T-29～T 交易日的基金收益率数据和 29 个一级行业指数收益率数据，采用主成分回归、逐步回归、岭回归、Lasso 回归法（这四种方法在下面图表中依次缩写为 PCA、Step_wise、Ridge、Lasso），得到仓位预测值，对所有普通股票型基金的仓位预测值取均值，并与每个季度末的真实值进行对比（基金季报上会公布仓位）。根据证监会规定，普通股票型基金持股仓位下限是 80%，因此我们设置仓位预测值的范围为[0.8,1]，若回归法计算出的预测值超出了这一范围则将预测值取为相近的边界值。对比结果如下表所示：

图表4： 普通股票型基金中各仓位测算方法效果对比

	实际仓位均值	预测值	预测值	预测值	预测值
		PCA	Step_wise	Ridge	Lasso
20170630	85.14%	89.98%	88.89%	88.63%	87.35%
20170929	87.87%	92.31%	89.49%	90.77%	88.37%
20171229	87.16%	94.62%	92.97%	96.40%	93.79%
20180330	86.33%	95.43%	93.28%	98.33%	93.15%
20180629	85.85%	86.07%	89.18%	99.17%	88.59%

资料来源：Wind，华泰证券研究所

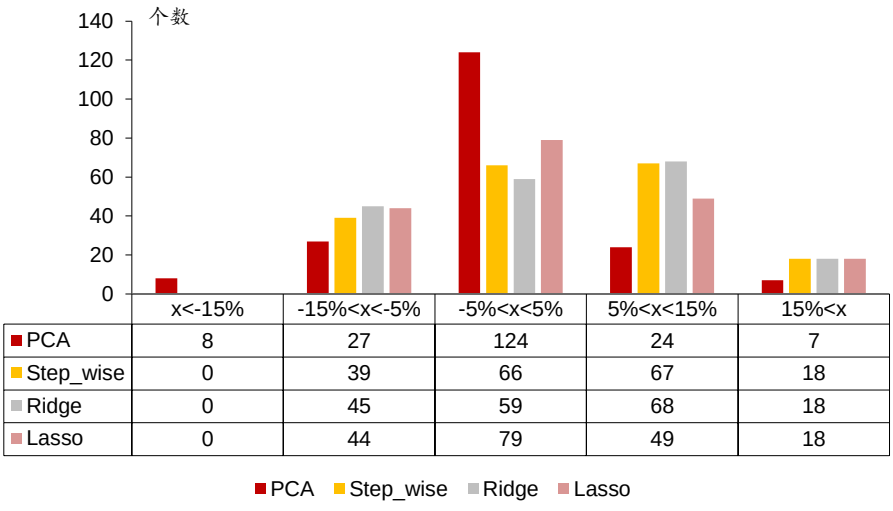
图表5： 普通股票型基金中各仓位测算方法统计数据

	PCA		Step_wise		Ridge		Lasso	
预测值减实际值	均值	方差	均值	方差	均值	方差	均值	方差
20170630	4.84%	1.25%	3.75%	1.24%	3.49%	1.32%	2.02%	1.36%
20170929	4.44%	0.91%	1.62%	0.99%	2.90%	1.31%	0.84%	1.07%
20171229	7.46%	0.93%	5.81%	1.07%	9.24%	0.92%	1.49%	1.11%
20180330	9.10%	0.70%	6.95%	0.93%	12.00%	0.87%	5.19%	0.98%
20180629	0.23%	0.58%	3.33%	0.65%	13.33%	0.41%	5.12%	0.83%

资料来源：Wind，华泰证券研究所

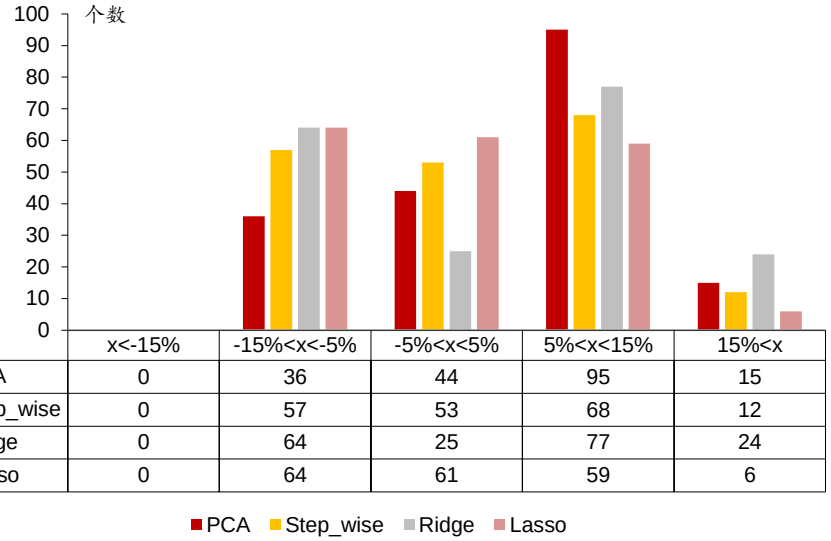
令 x 为仓位预测值减实际值，则其分布规律为：

图表6： 2017 年二季度末普通股票型基金中各仓位测算方法误差范围对比



资料来源：Wind，华泰证券研究所

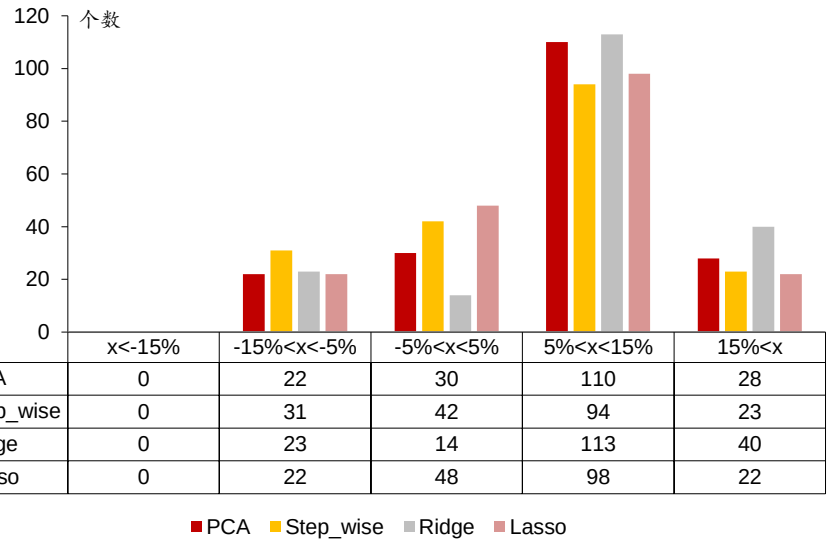
图表7： 2017 年三季度末普通股票型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	0	36	44	95	15
Step_wise	0	57	53	68	12
Ridge	0	64	25	77	24
Lasso	0	64	61	59	6

资料来源：Wind，华泰证券研究所

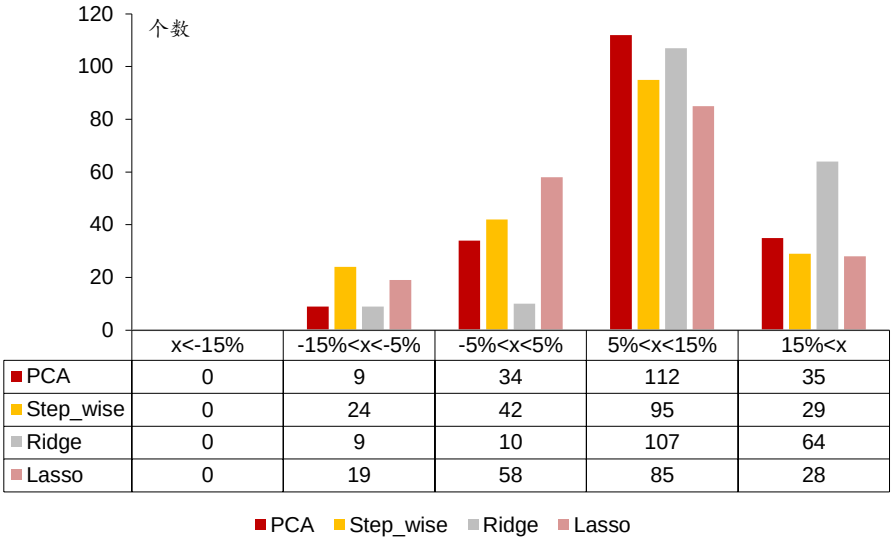
图表8： 2017 年四季度末普通股票型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	0	22	30	110	28
Step_wise	0	31	42	94	23
Ridge	0	23	14	113	40
Lasso	0	22	48	98	22

资料来源：Wind，华泰证券研究所

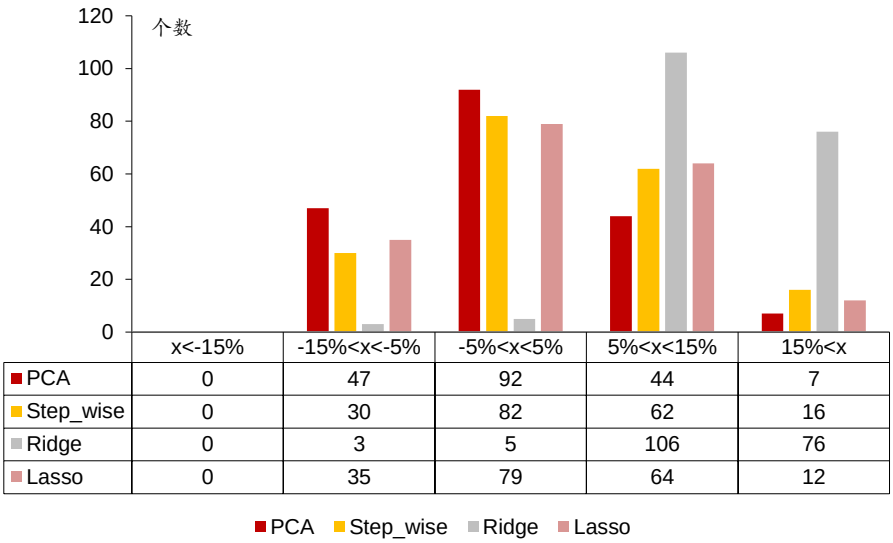
图表9： 2018 年一季度末普通股票型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	0	9	34	112	35
Step_wise	0	24	42	95	29
Ridge	0	9	10	107	64
Lasso	0	19	58	85	28

资料来源：Wind，华泰证券研究所

图表10： 2018 年二季度末普通股票型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	0	47	92	44	7
Step_wise	0	30	82	62	16
Ridge	0	3	5	106	76
Lasso	0	35	79	64	12

资料来源：Wind，华泰证券研究所

通过以上图表可知，除岭回归外，其余方法在不同横截面的误差大多落在[5%，15%]的区间范围内，岭回归存在系统性高估的现象，误差要稍大一些。

在偏股混合型基金中测试效果对比

接下来，我们对 414 只偏股混合型基金（数据选取详见上一大章第一小节）进行仓位测算，各项设置基本与上一小节相同，这里不再赘述。偏股混合型基金持股仓位下限是 60%，因此我们设置仓位预测值的范围为[0.6,1]，若回归法计算出的预测值超出了这一范围则将预测值取为相近的边界值。对比结果如下表所示：

图表11： 偏股混合型基金中各仓位测算方法效果对比

	实际仓位均值	预测值	预测值	预测值	预测值
		PCA	Step_wise	Ridge	Lasso
20170630	80.94%	80.94%	83.78%	82.93%	82.57%
20170929	83.16%	83.16%	87.19%	80.89%	80.78%
20171229	81.58%	81.58%	90.94%	84.75%	94.67%
20180330	81.30%	81.30%	92.76%	89.24%	96.84%
20180629	78.57%	78.57%	77.78%	80.83%	97.45%

资料来源：Wind，华泰证券研究所

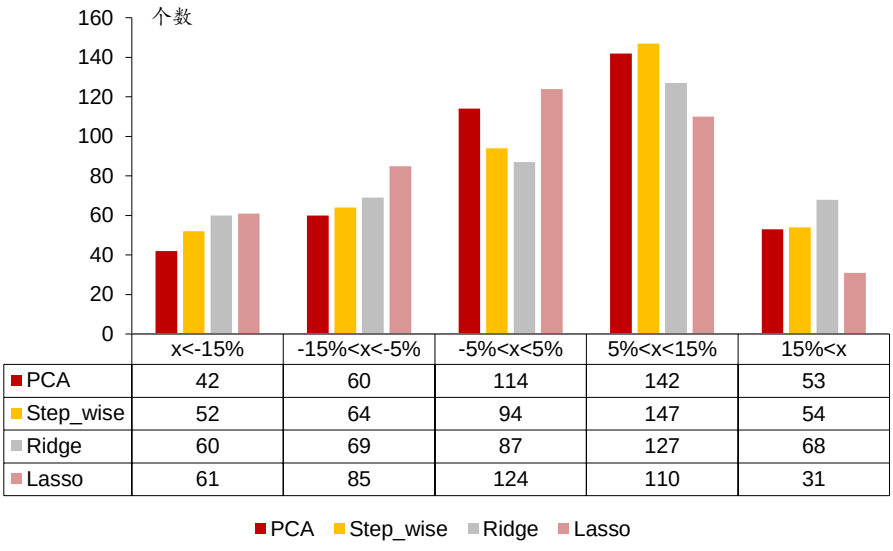
图表12： 偏股混合型基金中各仓位测算方法统计数据

	PCA		Step_wise		Ridge		Lasso	
预测值减实际值	均值	方差	均值	方差	均值	方差	均值	方差
20170630	2.84%	1.64%	1.96%	1.93%	1.63%	2.07%	-1.71%	2.62%
20170929	4.03%	2.09%	-2.17%	2.78%	-2.38%	3.83%	-3.32%	3.17%
20171229	9.36%	1.73%	3.14%	2.64%	13.09%	1.81%	-2.21%	2.77%
20180330	11.46%	0.86%	7.99%	1.30%	15.55%	1.46%	5.42%	1.42%
20180629	0.78%	1.38%	2.36%	1.62%	18.92%	1.59%	5.00%	2.04%

资料来源：Wind，华泰证券研究所

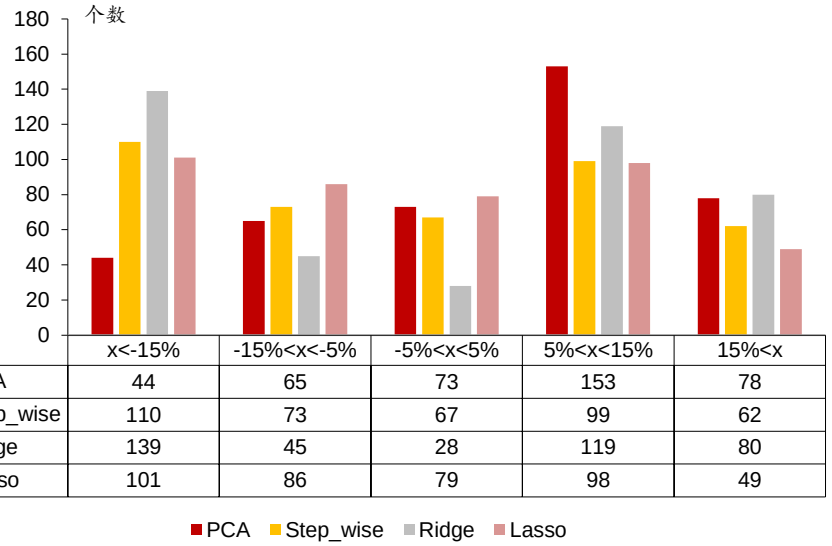
令 x 为仓位预测值减实际值，则其分布规律为：

图表13： 2017 年二季度末偏股混合型基金中各仓位测算方法误差范围对比



资料来源：Wind，华泰证券研究所

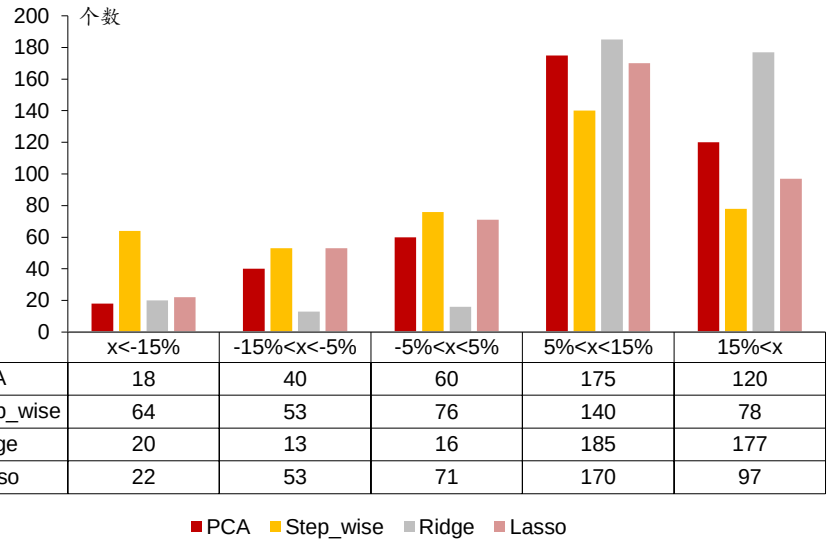
图表14： 2017 年三季度末偏股混合型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	44	65	73	153	78
Step_wise	110	73	67	99	62
Ridge	139	45	28	119	80
Lasso	101	86	79	98	49

资料来源：Wind，华泰证券研究所

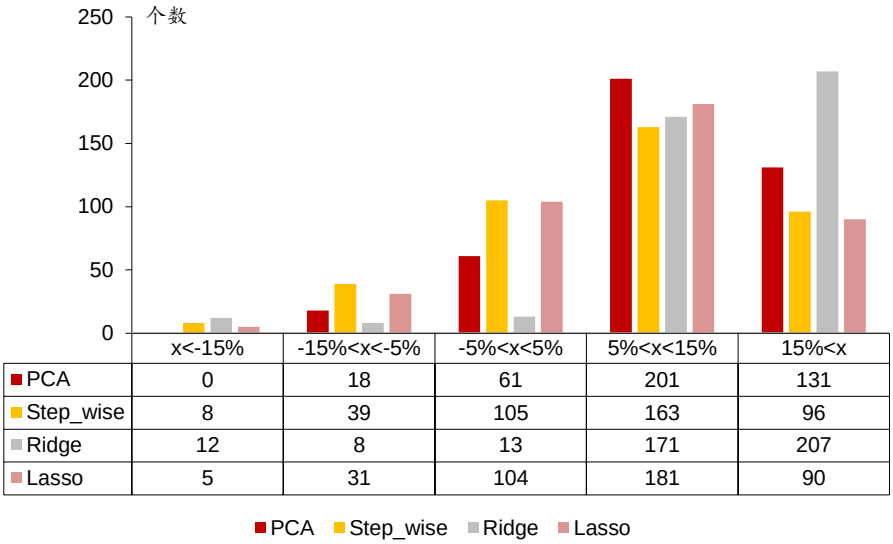
图表15： 2017 年四季度末偏股混合型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	18	40	60	175	120
Step_wise	64	53	76	140	78
Ridge	20	13	16	185	177
Lasso	22	53	71	170	97

资料来源：Wind，华泰证券研究所

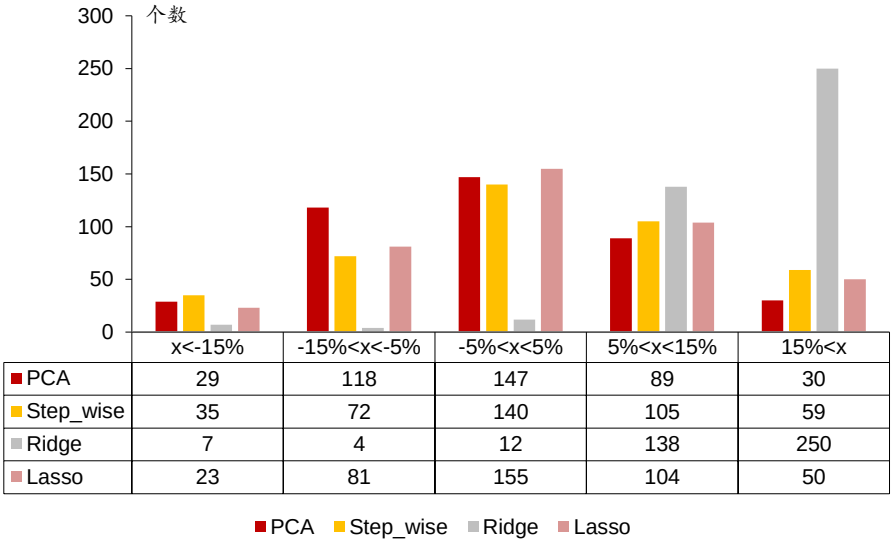
图表16： 2018 年一季度末偏股混合型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	0	18	61	201	131
Step_wise	8	39	105	163	96
Ridge	12	8	13	171	207
Lasso	5	31	104	181	90

资料来源：Wind，华泰证券研究所

图表17： 2018 年二季度末偏股混合型基金中各仓位测算方法误差范围对比



	x<-15%	-15%<x<-5%	-5%<x<5%	5%<x<15%	15%<x
PCA	29	118	147	89	30
Step_wise	35	72	140	105	59
Ridge	7	4	12	138	250
Lasso	23	81	155	104	50

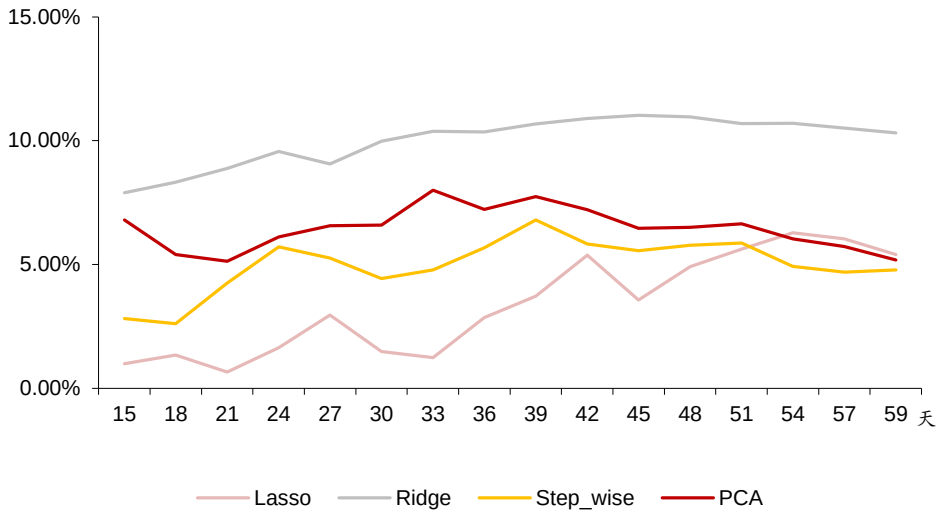
资料来源：Wind，华泰证券研究所

偏股混合型基金的测算结果与普通股票型基金相差不多，预测误差大多落在[5%，15%]区间范围内，岭回归存在系统性高估的现象。

回归时间窗口长度敏感性分析

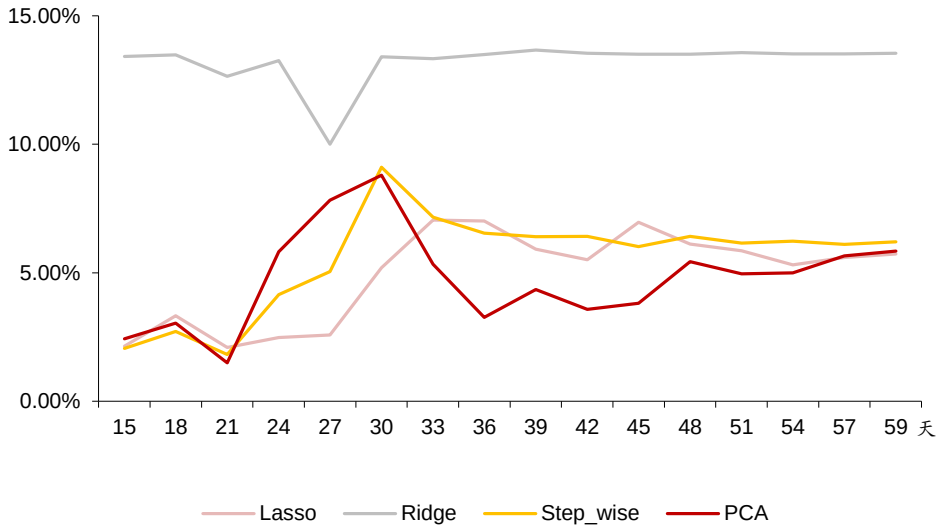
在前面的测算过程中，我们采用过去 30 个交易日的数据进行回归拟合，来预测当前基金仓位。考虑到回归窗口长度会对预测结果产生一定的影响，本报告对窗口长度在一定范围内进行调整，同时对普通股票型基金和偏股混合型基金在 2017 年四季度末、2018 年一季度末、二季度末三个横截面进行测试，结果如以下图表所示：

图表18： 2017 年四季度末普通股票型基金中各仓位测算方法误差均值随时间窗口变化曲线



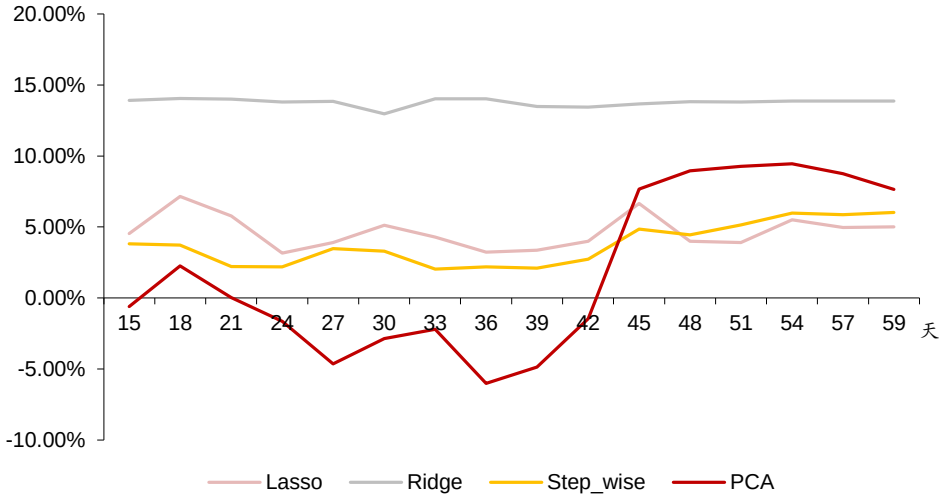
资料来源：Wind，华泰证券研究所

图表19： 2018 年一季度末普通股票型基金中各仓位测算方法误差均值随时间窗口变化曲线



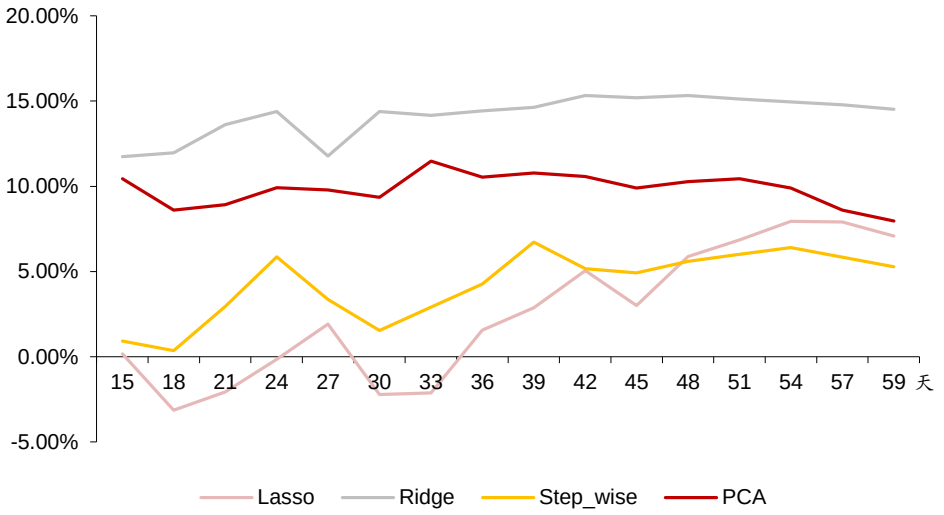
资料来源：Wind，华泰证券研究所

图表20： 2018 年二季度末普通股票型基金中各仓位测算方法误差均值随时间窗口变化曲线



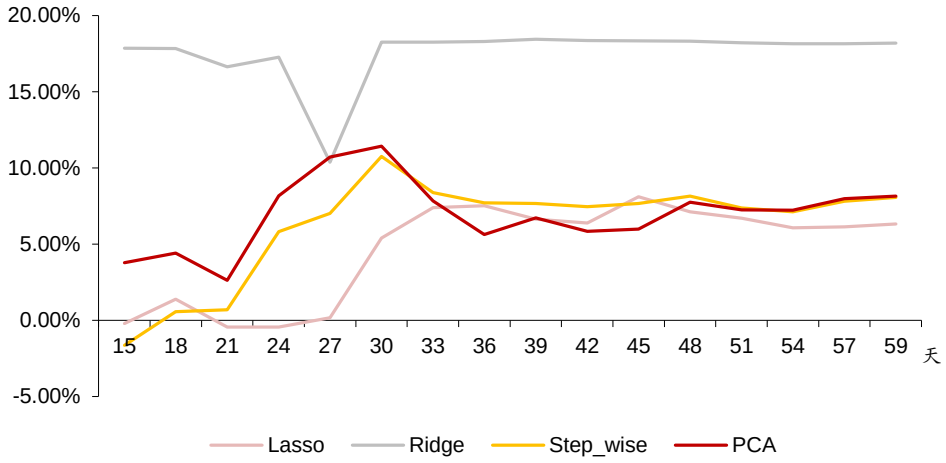
资料来源：Wind，华泰证券研究所

图表21： 2017 年四季度末偏股混合型基金中各仓位测算方法误差均值随时间窗口变化曲线



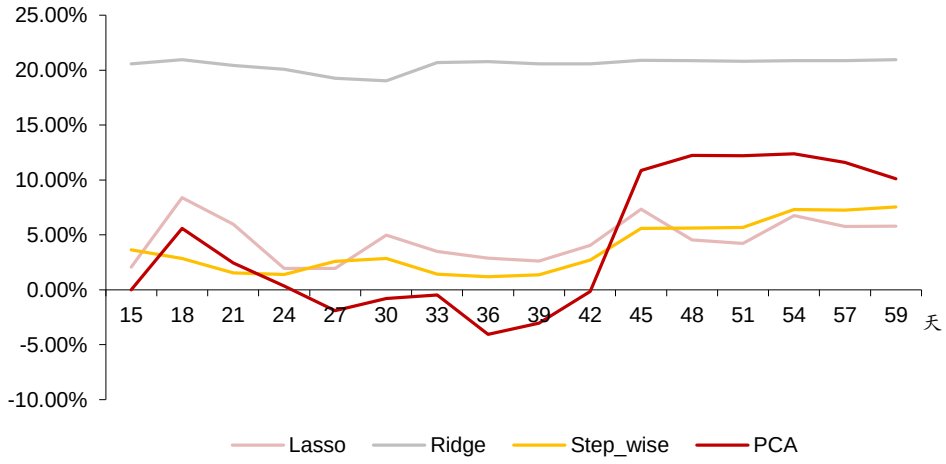
资料来源：Wind，华泰证券研究所

图表22： 2018 年一季度末偏股混合型基金中各仓位测算方法误差均值随时间窗口变化曲线



资料来源：Wind，华泰证券研究所

图表23： 2018 年二季度末偏股混合型基金中各仓位测算方法误差均值随时间窗口变化曲线



资料来源：Wind，华泰证券研究所

我们将回归时间窗口长度从 15 天到 59 天进行遍历，发现大部分情况下，各方法的预测误差均值在窗口长度大于 30 天之后比较平稳，趋于一个稳定的值，说明各方法的解已经收敛；在小于 30 天时没有明显规律。因为回归系数的实际含义是过去一段时间窗口内基金仓位的平均状况，并用这个值代表我们对当前时刻基金仓位的预测值，所以窗口长度也不宜取得太长（一般没有必要超过一个季度，约 60 个交易日），否则预测结果可能会滞后。本报告中我们选择使用 30 天还是比较合理的。

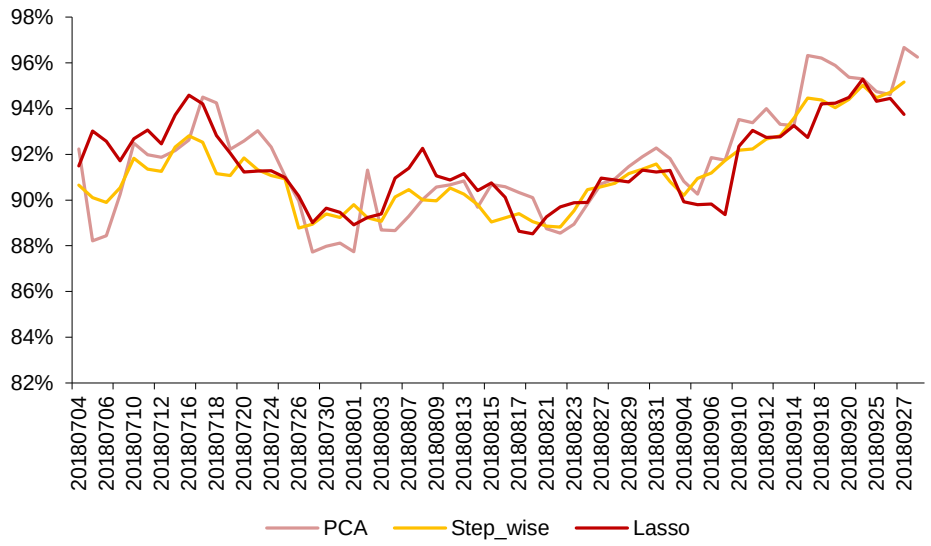
小结

以上四种基金仓位测算方法，都是基于过去一段时间的基金日频净值和一级行业指数数据，利用不同的回归模型，将各行业的拟合系数相加，得到最终仓位预测结果。四种方法本质上还是比较相近的，所以呈现的结果也并无太大差异，在本报告选择的几个静态截面上，Lasso 回归和逐步回归的预测精度稍好于主成分回归，岭回归效果最差。岭回归与另外三种回归方法最大的区别是不存在降维行为，主成分回归是通过主成分分析法将解释变量降维，逐步回归和 Lasso 回归则是通过某种方式剔除冗余变量，只有部分解释变量的回归系数不为零。我们猜测这可能是导致岭回归存在系统性高估现象的原因，有待进一步证实。

近期基金仓位测算观察

目前基金 2018 年半年报都已发布完毕，离三季报发布还有若干个工作日的时间，我们对基金三季度以来的仓位变化情况进行了测算。我们仍然使用过去 30 个交易日的数据进行回归计算，按日滚动回归，得到 2018 年 7 月 4 日至 2018 年 9 月 28 日的普通股票型基金和偏股混合型基金仓位预测值均值变化曲线，如以下两图所示：

图表24： 普通股票型基金仓位预测值均值曲线（2018.7.4~2018.9.28）



资料来源：Wind，华泰证券研究所

图表25： 偏股混合型基金仓位预测值均值曲线（2018.7.4~2018.9.28）

资料来源：Wind，华泰证券研究所

通过测算结果可知，从二季度末以来，基金仓位整体经历了一个先微升、后降、再缓慢攀升的过程。2018 年 7 月下旬至 8 月，基金仓位处于一个阶段性的底部，正好对应 A 股大盘震荡下行的一段行情。自 8 月底基金仓位开始缓慢攀升，也能对应于当时市场上曾出现的一些择时观点——认为大盘已经筑底完毕、即将开始反弹（现在来看该观点不一定对）。从某种程度来讲，公募基金的仓位变化确实能与市场情绪形成一定的对照。

风险提示

本报告中所采用的基金仓位测算方法仅基于日频基金净值数据和行业数据，没有利用基金报告中公布的重仓股、行业分布等信息，存在一定局限性。本报告中所采用的基金仓位测算方法仅在普通股票型基金和偏股混合型基金中进行实证，在其它类别基金中可能不适用。本报告中采用的四种回归方法只能缓解自变量间的多重共线性，并不能完全解决这一问题，敬请注意。

免责申明

收到本报告而视其为客户。

本报告基于本公司认为可靠的、已公开的信息编制，但本公司对该等信息的准确性及完整性不作任何保证。本报告所载的意见、评估及预测仅反映报告发布当日的观点和判断。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。同时，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。本公司不保证本报告所含信息保持在最新状态。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，不构成所述证券的买卖出价或征价。该等观点、建议并未考虑到个别投资者的具体投资目的、财务状况以及特定需求，在任何时候均不构成对客户私人投资建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。对依据或者使用本报告所造成的一切后果，本公司及作者均不承担任何法律责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

本公司及作者在自身所知情的范围内，与本报告所指的证券或投资标的不存在法律禁止的利害关系。在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为之提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本公司具有中国证监会核准的“证券投资咨询”业务资格，经营许可证编号为：91320000704041011J。
全资子公司华泰金融控股（香港）有限公司具有香港证监会核准的“就证券提供意见”业务资格，经营许可证编号为：AOK809

评级说明

行业评级体系

一报告发布日后的 6 个月内的行业涨跌幅相对同期的沪深 300 指数的涨跌幅为基准；

一投资建议的评级标准

增持行业股票指数超越基准

中性行业股票指数基本与基准持平

减持行业股票指数明显弱于基准

公司评级体系

一报告发布日后的 6 个月内的公司涨跌幅相对同期的沪深 300 指数的涨跌幅为基准；

一投资建议的评级标准

买入股价超越基准 20%以上

增持股价超越基准 5%-20%

中性股价相对基准波动在-5%~5%之间

减持股价弱于基准 5%-20%

卖出股价弱于基准 20%以上

华泰证券研究

南京

南京市建邺区江东中路 228 号华泰证券广场 1 号楼/邮政编码：210019

电话：86 25 83389999 /传真：86 25 83387521

电子邮件：ht-rd@htsc.com

北京

北京市西城区太平桥大街丰盛胡同 28 号太平洋保险大厦 A 座 18 层

邮政编码：100032

电话：86 10 63211166/传真：86 10 63211275

电子邮件：ht-rd@htsc.com

深圳

深圳市福田区深南大道 4011 号香港中旅大厦 24 层/邮政编码：518048

电话：86 755 82493932 /传真：86 755 82492062

电子邮件：ht-rd@htsc.com

上海

上海市浦东新区东方路 18 号保利广场 E 栋 23 楼/邮政编码：200120

电话：86 21 28972098 /传真：86 21 28972068

电子邮件：ht-rd@htsc.com