第44卷第3期 中国科学技术大学学报 Vol. 44, No. 3
2014年3月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Mar. 2014

文章编号:0253-2778(2014)03-0244-04

聚类分组法修正协方差阵的最佳投资组合方案

潘 洋,程希骏

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:给出了Markowitz模型的解并分析了最优化投资组合不稳定的原因.在此基础上,提出了一个新的方法:用聚类分组法调整样本协方差阵从而得到一个更好的投资组合.为了证明该方法的合理性,运用来自中国股市的真实数据来模拟"真实的投资".事实证明,用这种方法所得到的投资组合较传统方法有更好的收益率和更低的风险,且可以用风险预测来进行事后验证.

关键词:协方差阵;投资组合;稳定性;聚类

中图分类号:F830.9 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2014.03.014

引用格式:Pan Yang, Cheng Xijun. The optimal portfolio with modified covariance matrix using clustering method [J]. Journal of University of Science and Technology of China, 2014,44(3):244-247,256.

潘洋,程希骏. 聚类分组法修正协方差阵的最佳投资组合方案[J]. 中国科学技术大学学报,2014,44(3):244-247,256.

The optimal portfolio with modified covariance matrix using clustering method

PAN Yang, CHENG Xijun

(*Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: The Markowitz optimal portfolio was introduced and the reason why the result was unstable was analyzed. Based on this analysis, a new method was presented: Using the clustering method to modify the sample covariance matrix to get a better investment option. To prove the new method's reasonableness, real data from the Chinese stock market were used to simulate "real investment". It was found that the portfolio obtained from this method was better in both mean return and stability than the traditional method, which can be further verified by using risk prediction.

Key words: covariance matrix; portfolio; stability; clustering method

0 引言

Markowitz的最优化投资组合模型[1],是关于M-V原则所描述的一种投资策略:对于一些收益率相同的投资组合方案来说,人们会选择投资风险最小的方案.这样,基于上面的原则,我们可以把Markowitz模型表示为

$$\left.\begin{aligned}&\min\ \sigma^2_{\text{port}} = X'\Sigma X\\ &\text{s.t.}\quad ER'X = r_{\text{port}}\\ &i'X = 1\end{aligned}\right\} \tag{1}$$

收稿日期:2012-10-30;**修回日期**:2013-01-13

基金项目:国家自然科学基金(11371340)资助.

作者简介:潘洋,女,1991年生,硕士. 研究方向:投资学. E-mail: yaya@mail.ustc.edu.cn

通讯作者:程希骏,副教授. E-mail: xjc@ustc.edu.cn

式中，σ^2_{port} 为资产组合方差，ER 为期望收益率向量，r_{port} 为目标收益率，$\Sigma=(\Sigma_{ij})_{N\times N}$ 为协方差矩阵. N 维列向量 $i=(1,1,\cdots,1)'$，而列向量 $X=(X_1,X_2,\cdots,X_N)'$，其中，X_i 表示资产 i 占总的资产组合的比例.

在这个由 Markowitz 提供的最佳资产组合模型中，对于给定的目标收益率 r_{port}，用 Lagrange 方法可以使其风险最小的投资组合解为

$$X=\Sigma^{-1}ER(ER'\Sigma ER)^{-1}r_{\text{port}} \tag{2}$$

然而，样本经验协方差阵 Σ 的特征很难分析：为了估计协方差阵，需要估计 $N(N-1)/2$ 个协方差值，因而需要足够长时间 T 的一个时间序列，但是市场的情况随着时间变化，如果 T 过大的话，资产的协方差性会不稳定. 但另一方便，若是 T 与 N 的比值太小，那么我们估计值会不够稳定，因此对应的经验协方差阵会含有一点随机性导致的结果，也就是会有“噪声”. 这些协方差含有的“噪声”会对最后的结果有很大的影响[2]，这是我们在使用协方差阵的时候所需要考虑的.

因此，怎样减少协方差阵的“噪声”及其影响是一个很重要的问题，前人对此做了很多的研究[3-4]. 常用的方法是对协方差阵做一些结构上的调整来减少在半径内需要估计的有效数据，如结合了企业因素和宏观微观因素考量的单指数模型和多指数模型[5]；也有一些纯粹的统计学的协方差估计方法，如主成分分析法和贝叶斯收缩估计法[6-7]；还有最近几年提出的利用 RMT 理论来减少“噪声”的方法[9-10]. 总的来说，通过对协方差阵的一些调整可以有效地减少“噪声”对最佳资产组合的影响.

我们知道这样一种趋势：收益率相关性较低的两个资产，其样本相关系数中“噪声”相对“非噪声”的比例要大. 而线性相关性几乎为零的两个资产，因为“噪声”的存在，其样本估计协方差也不为零. 所以过滤掉协方差阵中这些较弱的相关关系所代表的协方差可以减少“噪声”对结果的影响. 在多元统计学中有一种聚类分组的方法，可以将点集按照点间距离分组，这里我们运用类似的思想，将点集分组，使得组内点的距离小于组间点的距离. 节 1 我们将介绍该方法及其算法，并将用聚类分组法将资产按其收益率相关性的强弱来分组，使得组内资产收益率的相关系数大于组间资产收益率的相关系数，然后通过过滤掉组间资产收益率的协方差来调整协方差阵，从而达到通过减少需要估计的数值来减少噪声影响的效果.

我们将用中国沪深股市的数据来证明上述方法的有效性. 首先通过模拟“真实的投资”，比较修正前和修正后的协方差阵对应最佳资产组合，证明本文方法得出的资产组合的性质更好. 然后用修正前和修正后的协方差阵来估计投资风险，发现本文方法估计的风险更贴近真实风险.

1　聚类分组法

1.1　聚类分组法及其应用

在多元统计分析中，聚类分组法是一种将 N 维欧式空间中的点集根据距离的远近来进行分组的方法. 这里我们首先以两个资产之间的收益率相关系数来定义资产间的距离，当然相关系数绝对值越小，对应的资产之间的距离越大，然后用类似的思想分组使得组内资产的距离小于组间资产的距离.

于是，我们令点的集合 $\Omega=\{H_1,H_2,\cdots,H_N\}$ 代表资产 $1,2,\cdots,N$. 定义距离矩阵 $D=(D_{ij})_{N\times N}$，$D_{ij}=1-|C_{ij}|$，其中，$|C_{ij}|$ 为相关系数矩阵 C 的元素 C_{ij} 的绝对值.

$$D_{ij}=\begin{cases}1+C_{ij}, C_{ij}\leqslant 0;\\ 1-C_{ij}, C_{ij}>0\end{cases} \tag{3}$$

式中，D_{ij} 为资产 H_i 与资产 H_j 的距离. 将 Ω 分割为若干组 $\Omega_1,\Omega_2,\cdots,\Omega_M$（$M>1$），使得组内的资产的距离小于组间的资产的距离（即组内的资产间的相关系数的绝对值大于组间的资产间的相关系数的绝对值），具体可表示为

$$\forall H_i,\ H_j\in\forall\Omega_P,\ \forall H_g\in\Omega_q,\ p\neq q,$$

有

$$D_{ij}<D_{ig}, D_{ij}<D_{jg}.$$

这里取使 M 值最小的分组.

为了实现上述目标，我们按照以下算法将资产分组：

第一步：找出集合中距离最远的两个资产（即找出相关系数绝对值最小的两个资产），分别标记为 P_1,P_2. 比较集合内其他的资产与这两个资产的距离. 将 P_1 和与 P_1 距离更近的资产分到第一组 G_1，剩余资产分到第二组 G_2.

第二步：分别逐一查看所有组的资产：若某一资产与它所在组的其他资产的距离不小于该资产与其他组的资产的距离，将该资产标记；若所有的资产与它所在组的其他资产的距离均小于该资产与其他组的资产的距离，则分组结束.

第三步：将所有标记的资产单独分出来，算为另

一组 G_3.

第四步:重复第二步.比较连续两轮查看中标记的资产的数目:若数目减少,则重复第三步分出另一组;若数目不变或增加,则将所有标记的资产按第一步的方法分成两组.

第五步:重复第四步,直至标记的资产的数目为零(即所有组内资产的距离均小于组间资产的距离)为止.

按上面的算法,可以得到满足所需条件的但 M 值最小的分组.

之前提到,当资产间的相关性较弱时,其中噪声占的比例会较大,对结果的影响也较大,而有时资产间线性相关性几乎为零,但因为噪声的存在,估算出的经验样本协方差值较小但是并不为零,所以这里不妨令

$$\Sigma_{ij}^{m}=\begin{cases}\Sigma_{ij},\text{若 } H_i,H_j \text{ 包含于同一组};\\ 0,\text{其他}\end{cases} \tag{4}$$

从而过滤掉较弱的资产间相关关系.

由此我们得出调整后的协方差阵:

$$\Sigma^{m}=(\Sigma_{ij}^{m})_{N\times N}.$$

1.2 该方法的不足及补充

在实际的应用中,如果仅凭样本相关系数来聚类的话,过于武断,因为样本相关系数不一定能真正反映实际的资产间的相关性,所以在实际的聚类中,可以同时参照其他的标准:例如两个股票在主观考虑上是否相关,如果都是能源股或者都是重金属股,这样的两只股票,即使样本相关系数显示可能相关性很弱,但是由主观可以判断出其相关性较高,应该分到一组.还有例如政府新出台了一些政策,据分析该政策会对某些股票的价格都起到可能的抬高或是打压作用,那么这些股票也应该被分到一组.

此外,在定义聚类所需的距离矩阵时,可以不用相关系数,而是用其他的参数,例如相关系数检验的 p-value.

聚类法适合于资产间明显存在可以分组的情况.此外,我们不得不承认的一点是:即使是收益率相关性较低的两个资产,其样本相关系数中也存在一定的"非噪声"的比例.所以这里在过滤样本协方差阵的一些元素时.可以进一步细化:

$$\Sigma_{ij}^{m}=\begin{cases}\Sigma_{ij}, & \text{若 } H_i,H_j \text{ 包含于同一组或}\\ & \text{其相关系数检验不为零};\\ 0, & \text{其他}\end{cases} \tag{5}$$

这里的相关系数检验可以随着实际情况来调整.

2 实证分析

现在我们把这种方法运用到中国股票市场,假设股票市场是可以卖空的.

我们选择沪深 300 指数的 178 只股票,根据 2010～2011 年的数据,总共有 $T=485$ 天.为了保留有效信息,减少噪声影响,如前所述,先将股票分组,然后用

$$\Sigma_{ij}^{m}=\begin{cases}\Sigma_{ij},\text{若 } H_i,H_j \text{ 包含于同一组};\\ 0,\text{其他}\end{cases} \tag{6}$$

得到调整后的协方差阵 $\Sigma^{m}=(\Sigma_{ij}^{m})_{N\times N}$.

为了证明该方法的有效性,我们用 2010 年即第 1 天($T=1$)到第 242 天($T=242$)的数据得到原始的协方差阵,然后根据式(2)计算出最佳资产组合 X_{11},再根据第 2011 年第 1 天($T=243$)的日收益

$$R(243)=(R_1(243),\cdots,R_{174}(243))$$

计算出 $T=243$ 的收益率

$$R_p1_1=X'_{11}R(243).$$

同样的,我们用第 2 天($T=2$)到第 244 天($T=244$)的数据计算出最佳资产组合 X_{12} 和对应日收益率 R_p1_2.然后重复上面的过程 100 次,最后可以得到对应的日收益率序列

$$R_p1=(R_p1_1,R_p1_2,\cdots,R_p1_{100}).$$

同时,还是用一样的数据和思路,但在计算最佳资产组合时用我们之前所定义的调整过的协方差阵 Σ^{m},然后得到了另一个类似的日收益率序列 R_p2.其结果如图 1 所示.

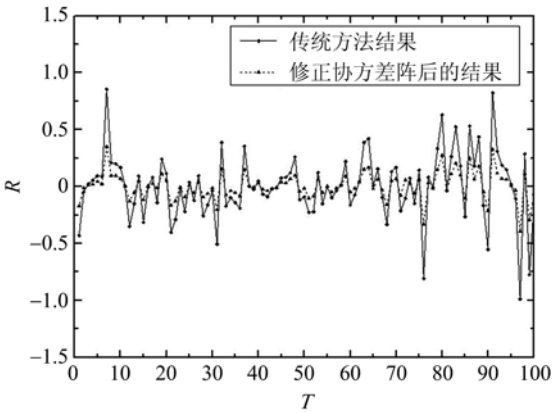

图 1 当 $R_{\text{port}}=5/10\ 000$ 时得到的结果

Fig. 1 The result when $R_{\text{port}}=5/10\ 000$

经计算得

$$E(R_p1) = 3.188523\mathrm{E}-5,$$
$$E(R_p2) = 5.150821\mathrm{E}-5,$$
$$\mathrm{Var}(R_p1) = 8.481028\mathrm{E}-6,$$
$$\mathrm{Var}(R_p2) = 1.397432\mathrm{E}-6.$$

比较这两个结果，我们可以发现 R_p2 比 R_p1 有更好的期望和更小的方差. 也就是说当用新的方法时投资会有相对更高的收益及更小的风险. 很明显我们这种方法比传统的方法有更高的效率和更好的稳定性.

我们已经证明了所提出的方法在求最佳投资组合时的有效性，现在用该方法来预测投资风险.

Merton[8]提到的资产组合有效边界函数为

$$\sigma_p^2 = f(R_p) = \frac{CR_p - 2AR_p + B}{D},$$

其中，

$$W = \Sigma^{-1};\ A = \sum\sum w_{ij}R_j;\ B = \sum\sum R_iR_j;$$
$$C = \sum\sum w_{ij};\ D = BC - A^2;$$
$$\Sigma = \frac{1}{T-1}\sum_{k=1}^{T}(R_k - E[R_k])(R_k - E[R_k])'.$$

R_i 表示第 i 天的资产池中所有股票的估计收益率的向量.

这里，我们将之前提到的 78 种沪深 300 指数的股票在 2010～2011 年总共有 $T=485$ 天的数据分为两部分(2010 年 242 天的数据和 2011 年 243 天的数据). 我们用第一部分的数据得到原始的和调整后的样本协方差阵，然后得到最佳投资组合集和对应的有效边界. 这里假设投资者对未来(2011 年)的收益率有完美的估计，也就是 R_i 取 2011 年中第 i 天的实际收益率.

我们用由 2010～2011 年的数据得到的原始样本协方差阵 Σ_{10} 和由 2008～2009 年的数据得到平均收益率 R_i 计算出最佳资产组合集并画出有效边界曲线($\sigma_p^2 = f(R_p)$). 然后用调整后的协方差阵 Σ_{10}^m 替代 Σ_{10} 从而得到另一组最佳资产组合集的另一条有效边界曲线($\sigma_{pm}^2 = f_m(R_p)$). 最后用 2008 年实际的协方差阵 Σ_{11} 得到实际的有效边界曲线 $\sigma_{\mathrm{preal}}^2 = f_{\mathrm{real}}(R_p)$. 结果如图 2 所示.

通过比较，我们发现，函数曲线 $\sigma_{pm}^2 = f_m(R_p)$ 比函数曲线 $\sigma_p^2 = f(R_p)$ 离真实的有效边界曲线 $\sigma_{\mathrm{preal}}^2 = f_{\mathrm{real}}(R_p)$ 更近，也就是说用我们的方法可以更好地预测风险. 而且由此可以解释之前的结果(图 1)：既然用我们的方法预测的风险更准确，得出的日收益当然更稳定.

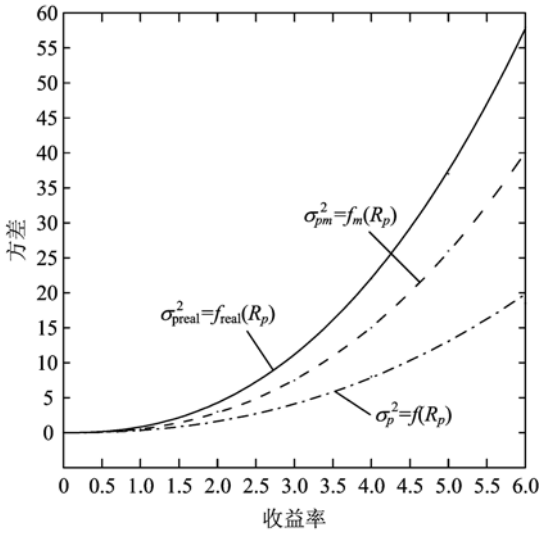

图 2　有效边界曲线

Fig. 2　Efficient boundary curve

3　结论

在传统的 Markowitz 模型中，计算出的最佳资产组合因为“噪声”的存在而稳定性不强. 这里我们用聚类分组法分析股市的数据，通过过滤调整协方差阵，从而得出一种新的计算最佳资产组合的方法. 虽然这种方法会使我们忽略不少信息，但是更能减少噪声对结果的影响，保留了协方差阵中“优质的元素”. 我们用这种方法分析中国股市，模拟真实的投资，发现与传统的方法相比，该方法得到的投资组合性质更好，而且用该方法还可以更好地预测风险，更有利于风险控制.

该方法还有一个作用，那就是当 $T<N$ 时资产收益率之间的协方差阵不可逆，而通过聚类法将资产分组，当每组的资产数目均满足 $T\geqslant N$ 时，对应的用式(4)调整后的协方差阵就可逆了，我们将在以后的文章中对该方法进行实证.

参考文献(References)

[1] Markowitz H. Portfolio Selection: Efficient Diversification of Investment [M]. New York: Wiley, 1959.

[2] Pafka S, Kondor I. Noisy covariance matrices and portfolio optimization [J]. The European Physical Journal B, 2002, 27(2): 277-280.

(下转第 256 页)