均值—熵指数在投资组合风险分散中的应用研究
——基于主成分分析

刘遵雄，唐顺发
(华东交通大学 信息工程学院，南昌 330013)

[摘 要] 在高风险的股票债券投资领域，分散是减少投资风险的重要技术。然而，并非是越分散投资效益越好。笔者以马科维兹均值—方差模型为基础，首先通过主成分分析法引进信息熵，然后利用熵指数作为投资组合分散程度的度量方式，使得均值—熵指数有效前沿成为一定投资收益下分散最充分的投资组合，为投资者的分散投资提供了有效的权衡。

[关键词] 投资组合; 信息熵; 均值—方差模型; 分散; 投资风险
[中图分类号] F830.59 [文献标志码] A doi: 10.3969/j.issn.1009-1912.2017.01.005
[文章编号] 1009-1912(2017)01-0029-05

Research On the Application of Mean-entropy Index in the Risk Diversification of Portfolio
——Based on Principal Component Analysis

LIU Zun-xiong, TANG Shun-fa
(*School of Information Engineering*, *East China Jiaotong University*, *Nanchang* 330013, *China*)

Abstract: In the field of stock and bond investments that with high risk, diversification is a significant technology for reducing investment risk. However, it is not always true that the more diversification the more investment benefits. The author based on the Markowitz mean-variance model, the author firstly introduced the information entropy through the method of principal component analysis, and then taking the information entropy as the measure of diversification degree of portfolio, So that the mean-entropy index effective frontier becomes the most-diversified portfolio under certain investment return, the author provides the investors with an effective balance between diversification and return.

Key Words: portfolio; entropy; mean-variance model; diversification; investment risk

证券投资的目的在于获取最大的收益，然而，资本主义市场中存在的各种问题，如监管不严、财务作假、上市公司效率不高等，往往使收益发生较大波动。在不完全市场中，投资风险分为系统性风险和非系统性风险: 系统性风险是指由国家经济政策与宏观经济形势变动和财税改革以及银行利率

[收稿日期] 2016-09-26
[修回日期] 2016-10-25
[基金项目] 2016年国家自然科学基金“上市公司财务预警中正则化和贝叶斯变量选择技术研究”(Grant no. 71361009)
[作者简介] 刘遵雄(1967—)，男，江西南昌人，教授，博士。主要研究方向: 金融统计，投资风险管理。

等方面的变动而造成的整个市场的变动，这种变动将造成投资者的收益不在预期范围内，它又被称为不可分散风险，是一种不能通过资产组合而排除的风险；非系统性风险则指上市公司特有的风险，是由于公司的经营管理、市场销售、重大投资等因素影响公司股价走势而形成的风险。投资者可以分散投资，构建相关的组合降低甚至消除非系统风险，然而分散的效率却与投资组合的标的资产的相关性有很大的关系[1]。因此，对投资组合相关结构和风险因子进行分析，可以帮助投资者更好地掌握风险源，并获取更好的分散策略。

假设一个投资组合由 n 支股票构成，r_i 代表收益（i 代表 n 支股票中的任意一支），如果用这些股票构建一个投资组合，该组合收益就等于收益与权重之积的累加：

$$r_p = \sum_{i=1}^{n} w_i r_i \tag{1}$$

w_i 是资产 i 的权重，则每支股票的预期收益为 $E(r_i)$，那么投资组合的预期收益为：

$$E(r_p) = \sum_{i=1}^{n} w_i E(r_i) \tag{2}$$

组合收益的风险由方差表示：

$$Var(r_p) = \sum_{i=1}^{n} w_i w_j \Omega \tag{3}$$

其中 Ω 是收益的协方差矩阵，在不完全相关市场中，组合风险也可写作：

$$Var(r_p) = \sum_{i=1}^{n} w_i^2 \cdot Var(r_i) \tag{4}$$

由式（3）、式（4）说明，可以通过使风险权重均等化来实现分散最大化。

一、主成分分析（PCA）

在现代投资组合理论中，主成分分析是一个重要的概念，它是把有相互关联的变量的复杂关系进行简化分析的重要方法。主成分分析是在力求原始信息丢失最少的原则下，对高维的数据变量进行空间降维。将少数的几个线性组合组成的综合变量尽可能多地保留原始的信息，这些综合变量就称为主成分。

设原来的风险指标为 $x_1, x_2, \cdots, x_n$，它们的综合指标为 $f_1, f_2, \cdots, f_m$，m 是主成分的个数，通过计算 f_i 的贡献率和累积贡献率求得：主成分 f_i 的贡献率为 $\frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}$，其累积贡献率为 $\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$，当累积贡献率达到 90%，对应的 m 的值就是主成分的个数。

主成分分析相当于将风险指标转换成一个新坐标的正交线性变换，最大的风险指标投影在第一坐标，也就是第一主成分，第二大风险指标投影在第二坐标，以此类推。也就是说，第一主成分尽可能多的包含风险数据，第二主成分尽可能多的包含剩下的风险数据[2]。以二维空间数据为例，PCA 风险指标旋转变换如图1、图2。

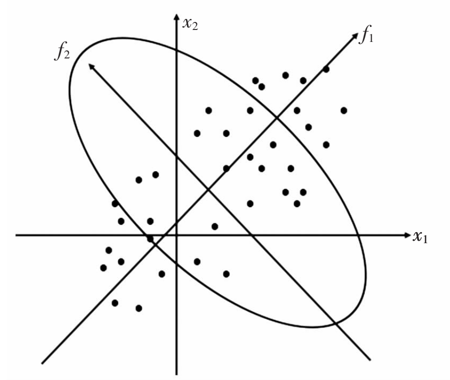

图1　旋转前

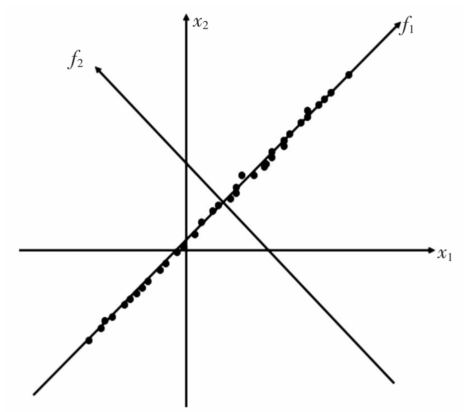

图2　旋转后

由收益率协方差矩阵有：

$$\boldsymbol{\Omega} = \boldsymbol{E\Lambda E}^{t} \text{ [3]} \tag{5}$$

其中，$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$和 $E = [\ell_1 \cdots, \ell_n]$ 分别为协方差矩阵的特征值和对应的特征向量，且 Λ 的主对角线的值是降序排列，对应的特征向量同样为降序排列。

由于收益率协方差矩阵 Ω 为对称正定矩阵，则有：

$$\Lambda = E^T \Omega E \quad (6)$$

对 E 进行旋转变换后的特征向量 $\bar{E} = [\bar{\ell}_1, \cdots, \bar{\ell}_n]$ 表示为 n 个不相关的投资组合，称为主成分投资组合。标的资产的投资权重 w 可以由主成分投资组合表示：

$$w = E\bar{w} \quad (7)$$

其转换后的投资权重为：

$$\bar{w} = E^{-1} w \quad (8)$$

容易证明特征向量 E 是正交对称矩阵，则投资权重也可写为：

$$\bar{w} = E^T w \quad (9)$$

组合收益 $\bar{r}_p = E^{-1} r_p = E^T r_p$，由上面论述可知，对应的收益也是逐渐减小的。若方差由旋转后对应的特征值表示，特征值与对应组合权重的乘积可定义为方差集群[4]：

$$\bar{v}_n = \bar{w}_n^2 \lambda_n^2, i = 1, \cdots, n \quad (10)$$

投资组合风险可表示为：

$$Var(r_p) = w^T \Omega w = w^T E^T \Lambda E w = \bar{w}^T \Lambda \bar{w} = \sum_{i=1}^{n} \bar{w}_n^2 \lambda_n^2 = \sum_{i=1}^{n} \bar{v}_n \quad (11)$$

则方差集群与组合风险的比值可以称为每个主成分投资组合占投资总风险的比重为：

$$p_n = \frac{\bar{v}_n}{Var(r_p)} = \frac{\bar{v}_n}{\sum_{i=1}^{n} \bar{v}_n}, i = 1, \cdots, n \quad (12)$$

由上式可知，由于 $\sum_{i=1}^{n} p_n = 1$，且恒为正数，可以 p_n 将视为概率，于是引入信息熵。

二、信息熵

信息熵最早由 Shannon 提出[5]。在信息理论中，熵越小，则表示不确定性越小，掌握的信息越多；反之，不确定性越大，掌握的信息越少，它利用了随机概率的不确定性，可用概率来描述。因此，一个投资组合的分散程度可以由 p_n 的分布情况知道：若该组合分散程度不高，那么随机概率聚集在其中几个主成分组合。也就是说，风险由其中几个主成分组合分担；反之，若投资组合已充分分散化，则随机概率均等分配，也就是投资风险均等分布于所有的主成分投资组合[6]。故引入熵和熵指数来量化投资组合分散水平：

$$H(p) = -\sum_{i=1}^{n} p_i \ln p_i, i = 1, \cdots, n \quad (13)$$

$$E_{H(p)} = \exp(H(p)) \quad (14)$$

由 $0 \leqslant p_n \leqslant 1$ 知，$H(p) \geqslant 0$。则对于一个含有 n 支股票的组合，熵指数 $E_{H(p)}$ 的取值为 $[1, n]$。投资者便可以从熵指数的大小来推断投资风险的的分散程度：当 $E_{H(p)} = 1$，则说明投资风险完全由一个主成分组合引起，此时的分散程度最低；当 $E_{H(p)} = n$，说明投资风险平均分布在 n 个主成分投资组合中，此时的分散化程度最高。

三、均值—熵指数有效前沿

最大熵模型的建立：

$$\max H(p) = -\sum_{i=1}^{n} p_n \ln p_n \text{[7]} \quad (15)$$

$$(1)\ \text{s.t} \begin{cases} \sum_{i=1}^{n} E(r_i) \geqslant \bar{u} \\ -0.1 \leqslant w_i \leqslant 1.0 \\ \sum_{i=1}^{n} w_i = 1 \\ i = 1, \cdots, n \end{cases} \qquad (2)\ \text{s.t} \begin{cases} \sum_{i=1}^{n} E(r_i) \geqslant \bar{u} \\ 0 \leqslant w_i \leqslant 1.0 \\ \sum_{i=1}^{n} w_i = 1 \\ i = 1, \cdots, n \end{cases} \quad (16)$$

其中，$\bar{u}$ 是每支股票的周平均收益率，第一个约束模

型允许卖空，则投资权重可以最低取负数，此处设为 -0.1; 第二个约束模型不允许卖空，则投资权重不能为负数[10]。

均值—熵指数模型建立:

$$\hat{\omega} = \arg\max\{\theta w^T \bar{u} + (1-\theta) H(p)\{w\}\} \quad (17)$$

其中，参数 $\theta \in [0,1]$ 是对在预期收益与分散之间的权衡参数。

通过建立二次规划和线性规划，联立以上两个模型便可得到均值—熵指数有效前沿。

四、实证分析

从上证 100 指数中选取包括能源、交通、金融、房地产、饮食、医疗、通信、化工、汽车等行业 30 只股票进行研究。为了保证数据的平稳有效性，本文选取其从 2013 年 1 月 1 日到 2013 年 11 月 17 日的日收益率来研究。

通过计算收益率协方差矩阵的特征向量、特征值和累积贡献率，并取前面累计贡献率达 90% 以上的 16 个主成分，计算结果见表 2。

采用最大方差法对收益率协方差矩阵特征向量进行旋转处理，得到 16 个主成分模型:

$$\begin{cases} f_1 = 0.2181x_1 + 0.2138x_2 + \cdots 0.1003x_{30} \\ f_1 = 0.3831x_1 - 0.3089x_2 + \cdots 0.0991x_{30} \\ \cdots\cdots \\ f_{16} = -0.0301x_1 - 0.1045x_2 + \cdots - 0.2213x_{30} \end{cases} \quad (18)$$

验证投资组合是否已经充分分散的各个主成分的方差占比 p_n，如图 3 所示。

由图 3 可知，投资风险均等分布于每个主成分，说明该模型已经实现充分分散，达到预期效果。

表 2　特征值、方差贡献率和方差累积贡献率

主成分	1	2	3	4	5	6	7	8
特征值	0.006 1	0.001 1	0.000 9	0.000 8	0.000 7	0.000 6	0.000 6	0.000 5
方差贡献率	0.403 1	0.069 5	0.061 7	0.051 8	0.049 1	0.042 8	0.040 2	0.032 9
方差累积贡献率	0.403 1	0.472 6	0.534 3	0.586 1	0.635 2	0.678 0	0.718 1	0.751 0
主成分	9	10	11	12	13	14	15	16
特征值	0.000 4	0.000 4	0.000 3	0.000 3	0.000 3	0.000 3	0.000 2	0.000 2
方差贡献率	0.027 4	0.024 7	0.020 8	0.020 6	0.017 5	0.016 7	0.013 4	0.012 8
方差累积贡献率	0.778 4	0.803 1	0.823 8	0.844 4	0.861 9	0.878 5	0.891 9	0.904 8

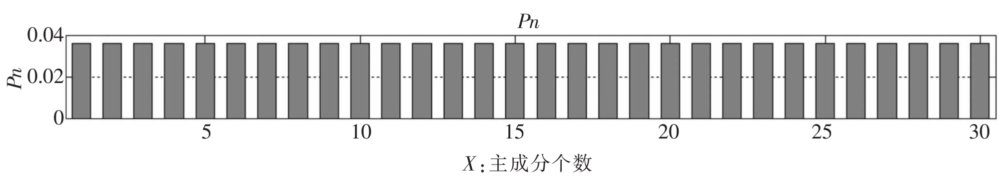

图 3　各主成分方差占比

为了了解投资分散程度与收益的关系，笔者构建了均值—熵指数有效前沿，如图 4、图 5 所示。

由图 4、图 5 可以清楚地看出: 允许卖空时，在一定范围内，随着熵指数的增大，预期收益逐渐减少，在熵指数为 17 以后，收益几乎直线下跌。同样的情况在不允许卖空的模型中发生，当熵指数小于 17 以前，收益虽然一直下跌，但在熵指数为 17 左右以后便出现了断点。然而 2 个有效前沿的不同点是: 可做空的有效前沿轮廓呈凸形，在一定范围内预期收益下降不明显; 而不允许做空的有效前沿轮廓是直线下降的，可以解释为，做多投资组合可以利用的基金对冲的优势，抵抗风险的方法更多。

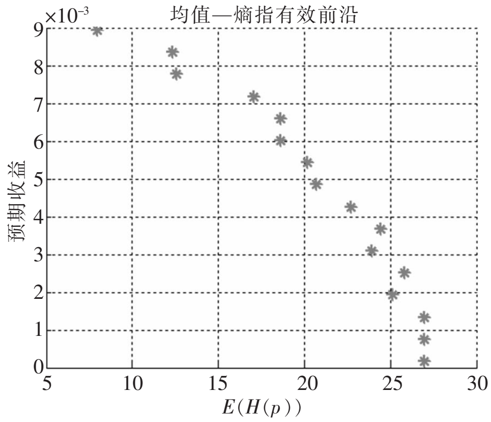

图 4　均值—熵指数有效前沿（允许卖空）

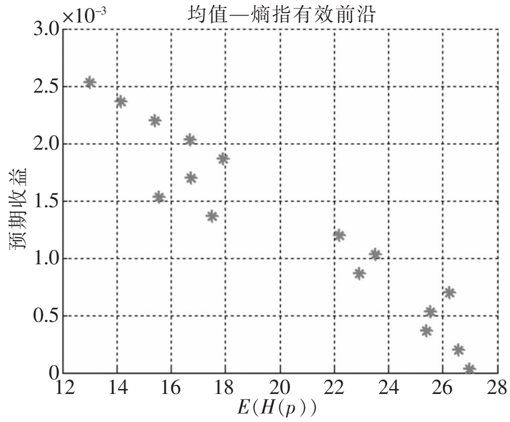

图 5　均值—熵指数有效前沿（不允许卖空）

结　语

投资的最终收益往往会受到各种影响，从而使之与预期收益存在一定的偏差，也就是存在投资风险。而风险的本质是具有不确定性，可理解为未来投资收益的不确定性及其发生的概率[11]。熵的内涵是信源不确定性的度量，用熵作为风险的度量的同时，它的指数可用于度量组合分散的程度。受马科维兹均值—方差模型启发，笔者引入均值—熵指数有效前沿的概念，为投资组合的分散结构构建了一个直观的显示图，投资者可以通过控制投资组合的分散结构，利用对冲或利用杠杆来抵抗或降低组合风险。这也为投资者的分散投资提供了一个新的方法。

【参考文献】

[1] 曾建华，汪寿阳. 一个基于模糊决策理论的投资组合模型[J]. 系统工程理论与实践，2003（1）：81.

[2] 李文泓. 关于宏观审慎监管框架下逆周期政策的探讨[J]. 金融研究，2009（7）：5.

[3] 刘红忠，赵玉洁，周冬华. 公允价值会计能否放大银行体系的系统性风险[J]. 金融研究，2011（4）：82.

[4] 宋晓秋. 模糊数学原理与方法[M]. 徐州：中国矿业大学出版社，1998：25.

[5] 荣喜民，张世英. 组合证券资产选择的模糊最优化模型和有效边界的研究[J]. 管理工程学报，1998（3）：24.

[6] Wilson A G. *The use of the concept of entropy in eystem modeling*[J]. Operational Quarterly，1970（2）：20.

[7] Almgren R C，Thum E. *Equity market impact*[J]. Risk Magazine，2005（18）：51.

[8] Duffie D，Andreas E，Guillaume H，et al. *Frailty correlated default*[J]. Journal of Finance，2009（5）：90.

[9] Huang X，Zhou H，Zhu H，et al. *A framework for assessing the systemic risk of major financial institutions*[J]. Journal of Banking and Finance，2009（5）：33.

[10] Lehar A，Alfred W. *Measuring systemic risk: a risk management approach*[J]. Journal of Banking & Finance，2005（9）：29.

[11] Tarashev N，Claudio B，Kostas T. *Attributing systemic risk to individual institutions*[J]. BIS Working Papers，2009（18）：29.