多因素模型及其在沪深300中的实证

联系人：戴　军
分析师：黄志文 SAC执业证书编号：S0980206110185
分析师：葛新元 SAC执业证书编号：S0980200010107

国信证券经济研究所
Guosen Securities Economic Research Institute

目录

背景介绍

多因素模型中国内的一些不利条件

1、信息存在泄漏;
2、信息披露有一定的滞后性;
3、主动基金经理大多对此不认同;
4、样本过短，检验不充分。

为什么还要进行研究？

1、信息披露的规范，监管的加强;
2、规范研究方法的借鉴;
3、为进一步研究提供条件;
4、alpha策略的需要。

目录


获取数据
描述性变量的选择和标准化
风险指数的形成


风险因子报酬的估计
特定风险的预测
模型的更新

获取数据

首先搜集各种因子的数据，包括**市场行情**和**基本面**的数据。尤其要注意在选取数据期间公司发生的股利发放、资本重组等重大事件，要合理处理以保持数据的前后**一致性**和**可比性**。

描述性变量的选择和标准化

描述性变量（*Descriptor*）是描述各种因子的量化指标。描述性变量产生的渠道很多，可以只采用市场行情数据或只采用基本面数据，也可以是两类数据结合使用。从原则上来说，良好的描述性变量应该有实际经济意义，对市场的划分比较合理，对风险的分类描述比较全面。

描述性变量还要经过**标准化**，处理公式为：

$$[normalized\ descriptor] = \frac{[raw\ descriptor] - [mean]}{[standard\ deviation]}$$

风险指数的形成

描述性变量标准化之后，用资产的回报与行业因子和单个描述性变量做回归，并检验**显著性**。BARRA模型认为，只有通过显著性检验的描述性变量才可以用于合成风险指数。

BARRA对于描述行业的风险指数的处理方法有两种：

一是**单一行业的方法**，即把每个公司都列入某一类的行业，然后使用虚拟变量（Dummy）来描述行业因子；

二是**多行业的方法**，该方法把公司按照一定的指标（如总资产、销售或者营业利润等）把公司按比例归类到多种行业中去。如迪斯尼公司可以分为65%的传媒和35%的娱乐，即对传媒行业的风险因子暴露为0.65，对娱乐行业的风险因子暴露为0.35。

风险因子报酬的估计

风险因子报酬矩阵的计算　在对N期的截面数据做了广义最小二乘回归（GLS）之后，得到每期的因子报酬的估计值。如果假设这些因子报酬的分布具有强烈的稳定性，那么只需对N期的同类因子报酬求简单平均，则可以粗略地估计出其方差协方差矩阵（即风险因子报酬矩阵）。

由于金融变量的分布一般都不稳定，数据的相关性随着时间的间隔变长而变小，因此可以使用**指数衰减法**来估计协方差矩阵。

续：风险因子报酬的估计

波动率的估计　如果假设市场的波动是稳定的，那么这一步可以省略。如果考虑市场的波动有聚集的现象存在，可以用模型对变化的波动率进行模拟。在实证中有两种比较理想模型可以用来模拟，一种是DEWIV模型，一种是GARCH模型。

模型	表达式
DEWIV	$\sigma_{r,t}^2 = 21\cdot(1-\lambda)\sum_{s=1}^{\infty}\lambda^{s-1}(r_{t-s}-\bar{r})^2$
GARCH	$\sigma_{r,t}^2 = \varpi + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{r,t-1}^2$

续：风险因子报酬的估计

变量	含义
σ_s^2	用DEWIV或者GARCH模型模拟出来的波动率
σ_m^2	市场组合的波动率 $\sigma_m^2 = h_m^T XFX^T h_m + \sigma_{sp}^2$
F	未调整因子报酬矩阵（方差-协方差矩阵）
σ_{sp}^2	市场组合特定风险矩阵 $\sigma_{sp}^2 = h_m^T \Delta h_m$
h_m	市场指数组合的各种资产持有比例
Δ	市场组合特定风险的对角矩阵，BARRA假定是已知的

风险因子报酬矩阵的调整　　最终调整的风险因子报酬矩阵为

$$F_s = F + \frac{(\sigma_s^2 - \sigma_m^2)}{(\sigma_s^2 - \sigma_{sp}^2)} FX^T h_m h_m^T XF$$

特定风险的预测

特定风险 即为不同公司特定回报的标准差。BARRA模型设计了一个特定风险的预测方法，把公司的特定风险报酬分解为3个部分：绝对特定风险因素，相对特定风险因素和规模因素。

整体特定风险的估计式为

$$\hat{\sigma}_{it} = \kappa_{it} \cdot (1 + \hat{V}_{it}) \cdot \hat{S}_t$$

$\hat{S}_t$	**为绝对特定风险报酬因素，是一个滞后k期的时间序列**
$\hat{V}_{it}$	为相对特定风险报酬因素，计算公式为 $\hat{V}_{it} = \sum_{k=1}^{K} Z_{ikt}\gamma_k$
K	为一个给定的系数,代表获得每单位的特定风险回报必须承受的风险，它用于将特定风险回报转换成特定风险，可根据历史数据估计

模型的更新

当有新的市场行情数据和基本面数据公布时，可以重新计算每只股票的风险因子暴露，同时通过截面回归来估计最近一个月的因子报酬，更新风险因子报酬矩阵。同时更新特定风险报酬的3个因素：绝对特定风险报酬因素，相对特定风险报酬因素和规模因素，从而预测特定风险报酬。

将风险因子报酬和特定风险报酬两者结合就可以得到整个模型的风险报酬预测值。

目录

模型构建

$$R_i = \alpha_i + \beta_{i1} F_1 + \cdots + \beta_{ik} F_k + \varepsilon_i \quad i = 1, \ldots, n$$

R_i ：第i种股票的超额收益率

α_i ：回归方程的常数项，表示风险因子的公共效应

β_{ij} ：第i种股票对第j个公共因子的暴露度

$F_1, \ldots F_k$：k个公共因子

ε_i ：第i种股票的特殊回报

矩阵形式

$$R = \tilde{X}\tilde{F} + \varepsilon \tag{1}$$

即

$$\begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{pmatrix} = \begin{bmatrix} \alpha_1 & \beta_{11} & \cdots & \beta_{1k} \\ \alpha_2 & \beta_{21} & \ddots & \beta_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_n & \beta_{n1} & \cdots & \beta_{nk} \end{bmatrix} \begin{pmatrix} 1 \\ F_1 \\ \vdots \\ F_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

模型假设条件

1) $\varepsilon_1,\cdots,\varepsilon_n$ 相互独立

2) $\mathrm{E}(\varepsilon_i)=0$

3) ε_i与$F_1,\ldots F_k$独立

- $R=(R_1,\ldots,R_n)$ 的协方差矩阵为:

$$\begin{aligned}\Sigma &= \mathrm{cov}(R) = \mathrm{cov}(\tilde{X}\tilde{F}+e)\\ &= X\cdot\mathrm{cov}(F)\cdot X^T+\mathrm{cov}(\varepsilon)\\ &= X\cdot\Phi\cdot X^T+\Delta\end{aligned}$$

- 由*N*个风险性资产组成的投资组合*P*，其持有权重向量记作h_p

- 投资组合*P*的方差为:

$$\sigma_p^2 = x_p^T \cdot \Phi \cdot x_p + h_p^T \cdot \Delta \cdot h_p$$

$$= h_p^T \cdot \Sigma \cdot h_p$$

降维方法

- 等权重复合因子
- 逐步回归
- 主成分分析

等权重复合因子

- 利用各个指标自身的属性进行指标的合成，即可以把具有相同属性的指标归并成一个因子，从而达到降维的目的。
- 在指标合成时，可以采用等权重法或是加权权重法等具体的方法进行合成。

逐步回归

- 建立多元回归方程的过程;
- 按偏相关系数的大小次序将自变量逐个引入方程，对每个自变量偏相关系数进行统计检验，引入显著变量，如果变量不显著，不引入;
- 从方程中剔除非显著变量，只保留显著变量;
- 循此继续遴选下一个自变量。直至不再引入和剔除自变量为止，从而得到最优的回归方程。

主成分分析

- 将多个变量通过线性变换以选出较少个数重要变量的一种多元统计分析方法。
- 将原来众多具有一定相关性的*P*个指标，重新组合成一组新的互相无关的综合指标来代替原来的指标。

主成分分析步骤

- P个指标的协方差矩阵$\mathbf{A}$
- 求$\mathbf{A}$的P个特征值 $\lambda_1, \lambda_2, \ldots, \lambda_1$
- 求$\mathbf{A}$的P个特征向量 $u_1, u_2, \ldots, u_p$
- 第k个主成分的方差贡献率为 $\alpha_k = \lambda_k \Big/ \sum_{i=1}^{N} \lambda_i$

主成分表达式

$$\begin{cases} H_1 = u_{11}x_1 + u_{12}x_2 + \ldots + u_{1P}x_P \\ H_2 = u_{21}x_1 + u_{22}x_2 + \ldots + u_{2P}x_P \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ H_P = u_{P1}x_1 + u_{P2}x_2 + \ldots + u_{PP}x_P \end{cases}$$

参数估计-加权最小二乘法

- 同方差性

$$\mathrm{Var}(\varepsilon_i) = \sigma^2$$

- 异方差性

$$\mathrm{Var}(\varepsilon_i) = \sigma_{\varepsilon i}^2$$

假定模型存在异方差性

设

$$\begin{aligned}\operatorname{cov}(\varepsilon\varepsilon^T) &= \mathrm{E}(\varepsilon\varepsilon^T) \\ &= \operatorname{diag}(\sigma_{\varepsilon 1}^2, \cdots \sigma_{\varepsilon n}^2) = \sigma^2 W\end{aligned}$$

其中

$$W = \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{pmatrix}$$

$$W = DD^T$$

令

$$D = \begin{pmatrix} \sqrt{w_1} & & & \\ & \sqrt{w_2} & & \\ & & \ddots & \\ & & & \sqrt{w_n} \end{pmatrix}$$

用 D^{-1} 左乘 (1) 式两边，得到一个新的模型：

$$D^{-1}R = D^{-1}\tilde{X}\tilde{F} + D^{-1}\varepsilon \qquad (2)$$

- (2)式具有同方差性。
- 对(2)式进行一般最小二乘估计，可得

$$\hat{\tilde{F}} = ((\tilde{X}^*)^T \tilde{X}^*)^{-1}(\tilde{X}^*)^T R^*$$
$$= (\tilde{X}^T (D^{-1})^T D^{-1} \tilde{X})^{-1} \tilde{X}^T (D^{-1})^T D^{-1} R$$
$$= (\tilde{X}^T W^{-1} \tilde{X})^{-1} \tilde{X}^T W^{-1} R$$

目录

实证步骤

- 决定与计算描述性变量;
- 变量极值的处理;
- 描述性变量的标准化;
- 风险因子的选择（降维);
- 横截面回归模型估计因子报酬;
- 预测收益，选出股票。

因子选择：公司基本面数据、个股市场数据共29个：

Historical BETA（周、月）	B/P	Historical ALPHA（月）	负债比率（时价）
总市值（对数值）	买卖循环率（一个月）	Historical ALPHA（周）	销售额（营业收入）增长度
流通市值（对数值）	每日交易额的变动性	Total risk（月）	总资产的增长度
总资产（对数值）	买卖资金的变化（25日/120日）	Total risk（周）	销售额（营业收入）营业利润率
营业利润回报	买卖资金的变化（75日/250日）	Residual Risk（月）	销售额（营业收入）营业利润率Trend
营业收入回报	股价变动的平均偏离（25日）	Residual Risk（周）	总资产营业利润率
Specific RETURN(1个月)	股价变动的平均偏离（75日）	负债比率（账本价）	总资产营业利润率Trend

数据处理细节

动态数据：由于沪深300指数的成分股在不断变化调整

历史缺失：该支股票当期数据不计入计算范围

当期缺失：该支股票不计入计算范围

行情数据：

总股本（日数据）、流通股本（日数据）、日成交额、日收盘价、周收盘价、月收盘价、上证综指（日数据、周数据、月数据）

时间周期：2000年1月至2010年8月

财务数据：

总资产、总负债、所有者权益、营业收入、营业利润，以上皆为季度数据

时间周期：1999年12月至2010年8月

沪深300成分：

考虑沪深300样本调整，取每个月沪深300组成。

实证结果

表 2：多因素模型实证结果比较(2007.11.1～2010.8.31)

	等权复合因子	主成分分析	逐步回归	沪深 300
净值	0.735	0.483	0.619	0.512
标准差	41.10%	40.49%	40.20%	37.68%
日收益率	-0.044%	-0.105%	-0.069%	-0.096%
年化收益率	-10.49%	-23.07%	-15.86%	-21.40%
Sharpe	-0.255	-0.570	-0.394	-0.568

三种降维方法中，主成分分析没有跑赢基准，其他两种方法都有效的跑赢了基准。但经风险调整后的收益（Sharpe比率）逐步回归最高，等权复合因子年化收益率最高，超越基准10.91%。综合来说，等权重复合因子的降维方法是最优的，在实践中我们推荐使用等权重复合因子的降维方法来处理多因素模型。

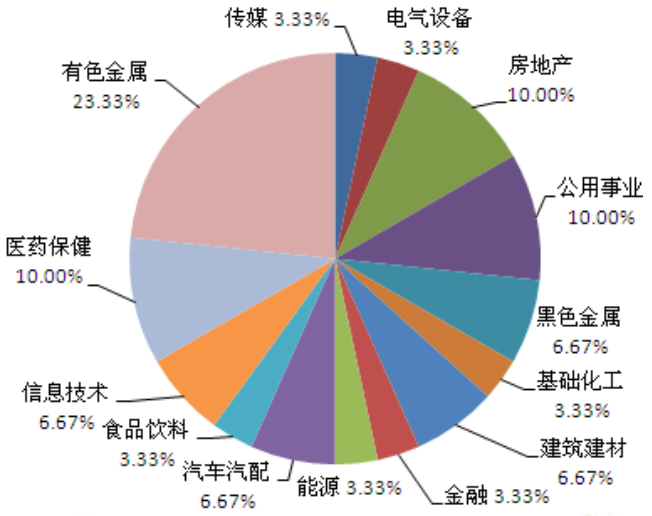
沪深300实证　股票配置建议

根据最新财报数据，使用等权复合因子方法，建议2010年9月配置以下30支股票

表3：多因素模型选股最新建议(2010.9.1～2010.9.30)

代码	名称	所属行业	代码	名称	所属行业
600547.sh	山东黄金	有色金属	000709.sz	河北钢铁	黑色金属
000895.sz	双汇发展	食品饮料	600282.sh	南钢股份	黑色金属
600497.sh	驰宏锌锗	有色金属	000060.sz	中金岭南	有色金属
600271.sh	航天信息	信息技术	600166.sh	福田汽车	汽车汽配
600143.sh	金发科技	基础化工	600583.sh	海油工程	能源
600432.sh	吉恩镍业	有色金属	600649.sh	城投控股	公用事业
600208.sh	新湖中宝	房地产	600219.sh	南山铝业	有色金属
600516.sh	方大炭素	有色金属	000690.sz	宝新能源	公用事业
600216.sh	浙江医药	医药保健	600170.sh	上海建工	建筑建材
600820.sh	隧道股份	建筑建材	600006.sh	东风汽车	汽车汽配
600550.sh	天威保变	电气设备	600804.sh	鹏博士	信息技术
600832.sh	东方明珠	传媒	600674.sh	川投能源	公用事业
000423.sz	东阿阿胶	医药保健	002001.sz	新和成	医药保健
600837.sh	海通证券	金融	000009.sz	中国宝安	房地产
600376.sh	首开股份	房地产	600489.sh	中金黄金	有色金属

资料来源：国信证券经济研究所，Wind资讯，国信一级行业分类



所占比重最大的行业是有色金属，其次是医药保健、公用事业、房地产、信息技术、黑色金属、汽车汽配以及建筑建材。

9月1日开始，截至9月9日，组合收益率1.68%，基准收益率0.80%；截至9月10日午盘，组合收益率2.43%，基准收益率1.35%。

资料来源：国信证券经济研究所，Wind资讯，国信一级行业分类

目录

Logistic模型原理 变量转换

一般多元回归：因变量连续

研究Logistic模型的动因？

Logistic回归：诸如0-1二分量的离散变量

因变量 $y=1$ 的概率记为 $\pi = P(y=1 \mid x_1, x_2, \cdots, x_n)$

$0 \le \pi \le 1$

将离散变量转换为连续变量

Logistic模型原理　Logit变换

$0 \le \pi \le 1$

$$\pi = \frac{\exp\left(\beta_0 + \sum_{i-1}^{n} \beta_i x_i\right)}{1 + \exp\left(\beta_0 + \sum_{i-1}^{n} \beta_i x_i\right)}$$

β_j 为常数

Logit变换

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i-1}^{n} \beta_i x_i$$

$$0 \le \frac{\pi}{1-\pi} < +\infty$$

即可对 $\log\left(\frac{\pi}{1-\pi}\right)$ 进行通常的多元线性回归

预测 $y=0$ 或 $y=1$ 的概率

沪深300实证　研究内容

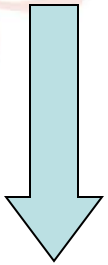
实证研究：沪深300指数成分股个股收益率超越该指数收益率的概率

0-1二分量（对于某支股票*k*）：

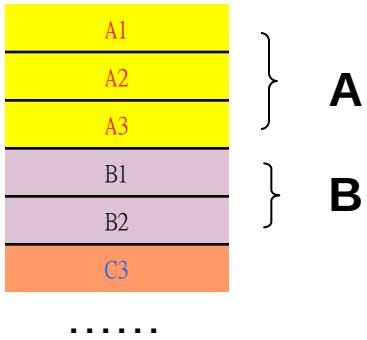
- 超越指数——　$y_k = 1$
- 跑输指数——　$y_k = 0$

沪深300实证　因子筛选与合并

29个因子——
影响计算速度！ 多！



因子筛选与合并
29个 → 8个！

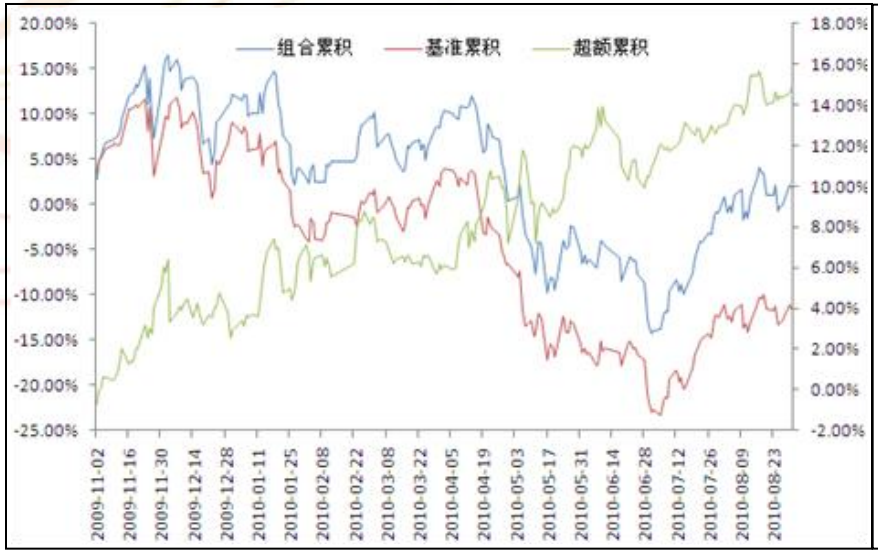


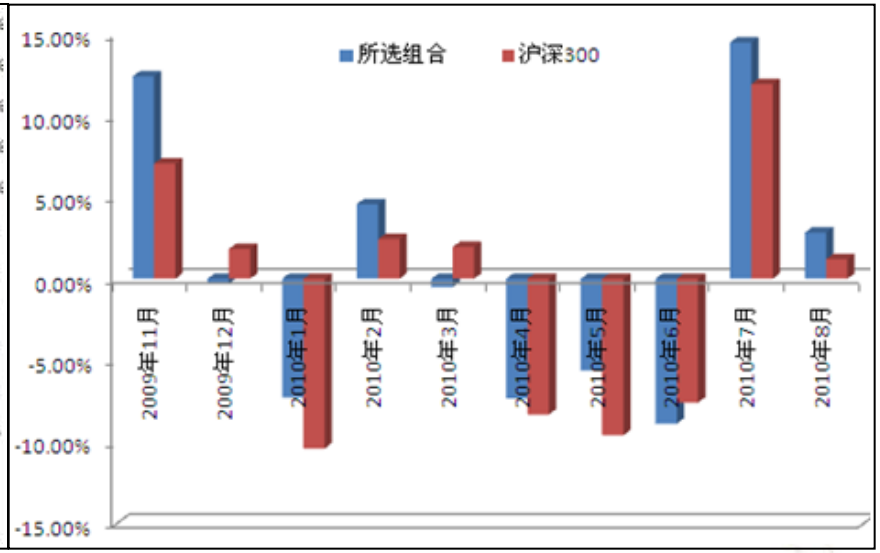
BETA、企业规模、相对低估性、流动性、投资成果、波动性、杠杆性、成长性

筛选与合并方法：
- 定性分组
- 标准化
- 简单算数平均

沪深300实证　实证结果

自2009年11月起统计，样本内年化收益率1.69%，沪深300指数年化收益率为-13.64%，超额收益率年化17.53%





Logistic选股模型实证结果比较(2009.11～2010.8)

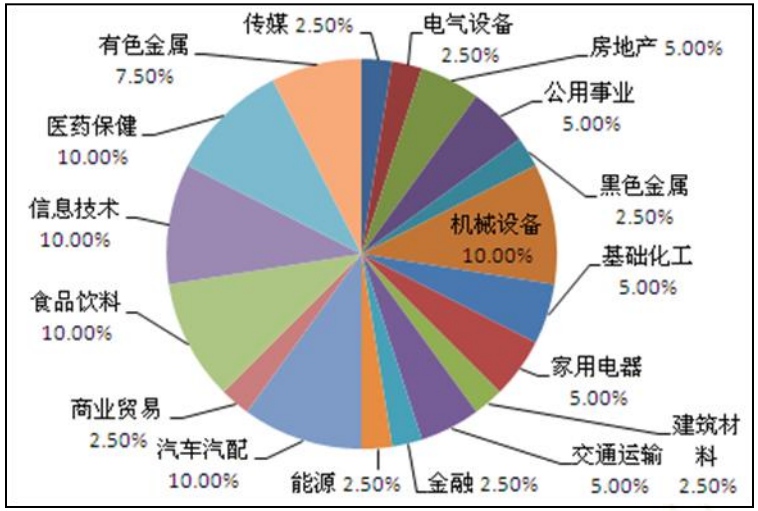
时间	所选组合	沪深300	超额收益率	组合累积	基准累积	超额累积
2009年11月	12.40%	7.05%	5.35%	12.40%	7.05%	5.35%
2009年12月	-0.31%	1.82%	-2.13%	12.05%	9.00%	3.11%
2010年1月	-7.27%	-10.39%	3.12%	3.91%	-2.33%	6.32%
2010年2月	4.57%	2.42%	2.15%	8.65%	0.04%	8.61%
2010年3月	-0.51%	1.95%	-2.46%	8.10%	1.99%	5.94%
2010年4月	-7.33%	-8.32%	0.99%	0.18%	-6.50%	6.99%
2010年5月	-5.62%	-9.59%	3.97%	-5.45%	-15.46%	11.23%
2010年6月	-8.88%	-7.58%	-1.30%	-13.85%	-21.87%	9.79%
2010年7月	14.48%	11.93%	2.55%	-1.37%	-12.55%	12.59%
2010年8月	2.82%	1.20%	1.62%	1.41%	-11.50%	14.41%
平均值	0.44%	-0.95%	1.39%	1.69%*	-13.64%*	17.53%*
标准差	28.47%	26.43%	9.09%			
Sharpe	0.059	-0.516	1.930			

资料来源：国信证券经济研究所，*代表年化收益率

沪深300实证　股票配置建议

根据预测数据及取胜概率排名，建议2010年9月配置以下40支股票

代码	名称	所属行业	代码	名称	所属行业
000039.SZ	中集集团	机械设备	600376.SH	首开股份	房地产
600066.SH	宇通客车	汽车汽配	600516.SH	方大炭素	有色金属
600456.SH	宝钛股份	有色金属	600518.SH	康美药业	医药保健
600597.SH	光明乳业	食品饮料	600779.SH	水井坊	食品饮料
600118.SH	中国卫星	信息技术	600316.SH	洪都航空	机械设备
600216.SH	浙江医药	医药保健	600839.SH	四川长虹	家用电器
600643.SH	爱建股份	金融	000538.SZ	云南白药	医药保健
000900.SZ	现代投资	交通运输	600085.SH	同仁堂	医药保健
600006.SH	东风汽车	汽车汽配	600741.SH	华域汽车	交通运输
600236.SH	桂冠电力	公用事业	000858.SZ	五粮液	食品饮料
000061.SZ	农产品	房地产	600500.SH	中化国际	商业贸易
000528.SZ	柳工	机械设备	600688.SH	S上石化	能源
600596.SH	新安股份	基础化工	000568.SZ	泸州老窖	食品饮料
600832.SH	东方明珠	传媒	600718.SH	东软集团	信息技术
600660.SH	福耀玻璃	汽车汽配	600089.SH	特变电工	电气设备
600879.SH	航天电子	信息技术	000060.SZ	中金岭南	有色金属
000959.SZ	首钢股份	黑色金属	000951.SZ	中国重汽	汽车汽配
600863.SH	内蒙华电	公用事业	600820.SH	隧道股份	建筑材料
600183.SH	生益科技	信息技术	000059.SZ	辽通化工	基础化工
000768.SZ	西飞国际	机械设备	000527.SZ	美的电器	家用电器



所占比重最大的行业是机械设备、医药保健、汽车汽配、信息技术、食品饮料，其次是有色金属。

9月1日开始，截至9月9日，组合收益率1.94%，基准收益率0.80%；截至9月9日午盘，组合收益率2.44%，基准收益率1.35%。

资料来源：国信证券经济研究所，Wind资讯，国信一级行业分类

结果分析

- 分类归并的方法是最好的，但是该方法需要计算全部的因子，因子的处理比较简单，只是简单平均归并；
- 逐步回归的预测效果更好些，但是，由于逐步回归每次得到的显著因子都不同，所以难以从时间序列角度考虑风险因子的变化；
- 而用主成分分析方法提取第一主成分，第二主成分，……，可以分析主成分之间的协方差，看出风险因子的关联性。

可能存在的问题

1、信息存在一定程度的泄漏，各个股票财报公布的时间为一个区间
2、对于行业、主题概念等属性很难加入到模型中
3、因子是否全面的问题
4、缺少因子之间的关联与逻辑的研究
5、国内股市的时间过短，数据的可靠性、完整性得不到保证
6、选股绩效的稳定性与可靠性还需要大样本验证

国信证券2010年指数基金与量化投资论坛

谢谢！