•研究报告•

2016-5-23

金融工程(专题报告)

选股模型

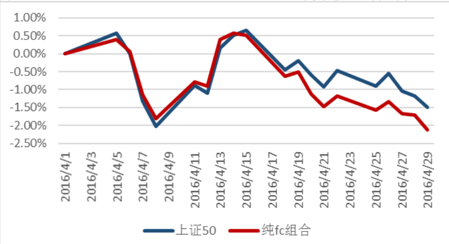
多因子模型系列报告之一
-----模型理论随想和纯因子组合构建

分析师 秦瑶
☎ (8627)65799830
✉ qinyao@cjsc.com.cn
执业证书编号：S0490513080002

联系人 邓越
☎ 02765799830
✉ dengyue1@cjsc.com.cn

联系人 杨靖凤
☎ （8621）68751636
✉ yangjf@cjsc.com.cn

市场表现对比图（近 12 个月）



资料来源：天软科技

相关研究

《Smart Money”策略 —— 基于资金流向选股》2016-5-20

《事件驱动策略的 Smart Beta 化》2016-5-6

《基差监控工具与 2016 年指数分红预测》2016-4-6

报告要点

■ **纯因子组合对模型中某个因子具有单位暴露度，对其他因子暴露度为 0**

纯因子组合对模型中某个因子具有单位暴露度，对其他因子暴露度为 0，对因子显著性检验、投资收益事后分析和多因子组合构建均有重要作用。

■ **从多因子模型可以构建纯因子组合**

在时间截面上，通过广义最小二乘法（GLS）就可以方便的求出因子收益的值，同时中间步骤给出了纯因子组合中个股的权重。

■ **基准组合对各个因子的因子暴露为 0，非零的因子暴露度就是投资组合偏离了基准组合多少倍标准差**

通过特定的因子值标准化方式可以使以流通市值为权重的投资组合(基准组合)是对所有风格因子 0 暴露的。流通市值加权的组合（基准组合）的收益就是模型的截距项。

■ **因子暴露度和可投资性之间通常需要妥协**

流通市值加权基准指数是容量最大的，对各因子 0 暴露；另类加权的策略，如等权重（加强市值因子暴露）、波动率加权（加强波动率因子暴露）等，持有基准指数所有股票，仅作权重的倾斜，可投资性次之。多空组合暴露度增大，可投资性减弱；因子暴露度最大、可投资性最低的则是纯因子组合。

■ **可投资的高暴露因子组合**

在追求纯因子暴露和可投资性之间妥协，可以适当放宽对其他因子的要求，构建出对某因子有较高暴露度，对其它因子有适当暴露度的可投资的高暴露因子组合。

目录

图表目录

多因子模型系列报告将从多因子理论探讨出发，介绍如何构建纯因子组合，如何从含有股票空头的纯因子组合构建可投资的纯多头组合，并阐述如何利用纯因子组合进行因子投资（因子配置和因子择时）。

本篇报告是该系列报告的第一篇，主要是对多因子模型理论基础的梳理，介绍从多因子模型出发构建纯因子组合的方法，并举了上证 50 指数成分股的例子，力求能把方法描述清楚。

纯因子组合对模型中某个因子具有单位暴露度，对其他因子暴露度为 0，对因子显著性检验、投资收益事后分析和多因子组合构建均有重要作用，这些在本系列的后续报告会陆续讨论。

由于关于多因子模型的研究报告已经汗牛充栋，我们不打算对背景信息太多赘述，下面直入主题。

一、模型的一般形式

理论上多因子模型应涵盖包含股票、债券、大宗商品、地产等所有资产，但我们的讨论仅局限在股票的二级市场。典型的股票多因子模型将 n 只股票的收益率分解为 m 个因子的线性组合和未被因子解释的残留部分，一般形式为：

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} f_1 + \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} f_2 + \cdots + \begin{bmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{bmatrix} f_m + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

其中，r_i 为对无风险收益率的超额部分，x_{ij} 为股票 *i* 对因子 *j* 的暴露度，f_i为因子收益。（x_{ij} 的名称有许多种，比如因子载荷、因子 beta、斜率、系数、因子敏感度等，此处及后文均采用 Barra 模型的语境，即因子暴露）。x_{ij}是随个股特征不同而变化的部分，而 f_i 为共同部分（共同因子）。

约定所有的向量均为列向量，且向量均用小写加粗字母表示、矩阵用大写加粗字体、标量用不加粗小写字体。上式也可写成向量形式：

$$\boldsymbol{r} = \boldsymbol{x_1} f_1 + \boldsymbol{x_2} f_2 + \cdots + \boldsymbol{x_m} f_m + \boldsymbol{u}$$

或矩阵形式：

$$\boldsymbol{r} = \boldsymbol{X}\boldsymbol{f} + \boldsymbol{u}$$

其中 $\boldsymbol{r} = [r_1, r_2, \ldots, r_n]^T, \mathbf{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_m}], \boldsymbol{f} = [f_1, f_2, \ldots, f_m]^T$ 。

后文将直接在这三种形式之间转换，而不再做说明。

二、组合的因子暴露度和纯因子组合

假设一个投资组合 P 中的个股权重为：

$$\boldsymbol{w_P} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

则这个投资组合的收益 r_P为：

$$r_P = \boldsymbol{w_P^T r} = \boldsymbol{w_P^T x_1} f_1 + \boldsymbol{w_P^T x_2} f_2 + \cdots + \boldsymbol{w_P^T x_m} f_m + \boldsymbol{w_P^T u}$$
$$= x_P^1 f_1 + x_P^2 f_2 + \cdots + x_P^m f_m + \boldsymbol{w_P^T u}$$

其中 $x_P^m = \boldsymbol{w_P^T x_m} = \sum w_i x_{im}$ 就是组合 P 对因子 m 的因子暴露度。

也就是说，一个投资组合的收益可以分解为 m 个因子收益的线性组合：

$$r_p = x_p^1 f_1 + x_p^2 f_2 + \cdots + x_p^m f_m + \boldsymbol{w_p^T u}$$

如果组合充分分散，应该有 $\boldsymbol{w_p^T u} \to 0$ 。

如果一个投资组合，对模型中某因子暴露度为 1，对其他风险因子暴露度为 0，则称该组合为这个因子的纯因子组合。假设因子 m 对应的纯因子组合 P_m 的权重为 $\boldsymbol{w_{P_m}}$，由定义，该组合对各个因子的暴露为$x_{P_m}^1 = 0, x_{P_m}^2 = 0, \ldots, x_{P_m}^{m-1} = 0, x_{P_m}^m = 1$,因此纯因子组合的收益为：

$$r_{P_m} = \boldsymbol{w_{P_m}^T r} = x_p^1 f_1 + x_p^2 f_2 + \cdots + x_p^m f_m + \boldsymbol{w_p^T u} = 1 \cdot f_m + \boldsymbol{w_p^T u}$$

可见，对于充分分散的组合，因子收益 f_m 就是因子 m 对应的纯因子组合 P_m 的收益。至此，将 f_m 称作是“收益”才有了很好的解释。通过广义最小二乘法（GLS）就可以方便的求出 f_m 的值，同时中间步骤还给出了该纯因子组合中个股的权重 $\boldsymbol{w_{P_m}}$,即是 $f_m = \boldsymbol{w_{P_m}^T r}$，具体计算方法见后文。

通过纯因子组合的线性组合可以构造出对每个因子具有指定暴露度的组合，如希望构建一个组合，使组合对规模因子（size）暴露为 0，对 E/P 因子暴露为 0.5，其他因子无要求，则可先由两因素模型解出纯因子组合的个股权重$\boldsymbol{w_{size}}$ 和 $\boldsymbol{w_{EP}}$，满足要求的投资组合权重为：

$$\boldsymbol{w_P} = 0 \cdot \boldsymbol{w_{size}} + 0.5 \cdot \boldsymbol{w_{EP}}$$

可见，纯因子组合可将各种风险收益精确的切割开来，求解出纯因子组合中的个股的权重，就可以精确控制组合对因子的暴露度，使组合只对希望暴露的风险因子暴露，从而让风险和收益精确匹配起来。

三、几种常见模型

这些模型都是总所周知的，无需赘言，这里列举几种（CAPM、Fama-French、Barra）来具体化上面的一般形式，后文重点展示从 Barra 模型构建纯因子组合的应用。

CAMP:

$$\begin{bmatrix} r_1 - r_f \\ r_2 - r_f \\ \vdots \\ r_n - r_f \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} (r_B - r_f) + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

可以其中为r_B 基准收益。$(r_B - r_f)$ 为因子收益，β_i 为个股的因子暴露。可以验证，基准组合就是 CAMP 的纯因子组合，因为基准组合对自身的$\beta = 1$ 。

Fama-French 三因素：

$$\begin{bmatrix} r_1 - r_f \\ r_2 - r_f \\ \vdots \\ r_n - r_f \end{bmatrix} = \begin{bmatrix} \beta_1^B \\ \beta_2^B \\ \vdots \\ \beta_n^B \end{bmatrix} (r_B - r_f) + \begin{bmatrix} \beta_1^{size} \\ \beta_2^{size} \\ \vdots \\ \beta_n^{size} \end{bmatrix} SMB + \begin{bmatrix} \beta_1^{value} \\ \beta_2^{value} \\ \vdots \\ \beta_n^{value} \end{bmatrix} HML + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Fama-French 的因子收益有市场基准收益、大小规模收益之差和高低估值收益之差。

Barra USE3 和 Barra CHE2：

$$\begin{bmatrix} r_1 - r_f \\ r_2 - r_f \\ \vdots \\ r_n - r_f \end{bmatrix} = \begin{bmatrix} x_1^{\mathrm{I}1} \\ x_2^{\mathrm{I}1} \\ \vdots \\ x_n^{\mathrm{I}1} \end{bmatrix} f_{\mathrm{I}1} + \cdots + \begin{bmatrix} x_1^{\mathrm{Ip}} \\ x_2^{\mathrm{Ip}} \\ \vdots \\ x_n^{\mathrm{Ip}} \end{bmatrix} f_{\mathrm{Ip}} + \begin{bmatrix} x_1^{S1} \\ x_2^{S1} \\ \vdots \\ x_n^{S1} \end{bmatrix} f_{S1} + \cdots + \begin{bmatrix} x_1^{Sq} \\ x_2^{Sq} \\ \vdots \\ x_n^{Sq} \end{bmatrix} f_{Sq} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

这是基于 APT 建立的模型，其中 $\mathrm{I}_1, \dots, \mathrm{I}_p$ 为p 个行业因子（Industry）,$f_{\mathrm{I}1}, \dots, f_{\mathrm{Ip}}$为对应的因子收益；$\mathrm{S}_1, \dots, \mathrm{S}_q$ 为q 个风格因子(Style),$f_{\mathrm{S}1}, \dots, f_{\mathrm{Sq}}$为对应的因子收益。$\mathbf{X}^{\mathrm{Ip}}$为行业因子暴露（0 或 1）,$\mathbf{X}^{\mathrm{Sq}}$ 为风格因子暴露。Barra 一直致力于对结构化风险模型（协方差）的研究，此处我们止步于它的中间步骤----即收益模型。

Barra USE4 和 Barra CHE5：

最新的 Barra USE4 和 Barra CNE5 在模型中加入了截距项，是我们这一多因子模型系列报告采用的形式。

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} f_c + \left\{ \begin{bmatrix} x_1^{\mathrm{I}1} \\ x_2^{\mathrm{I}1} \\ \vdots \\ x_n^{\mathrm{I}1} \end{bmatrix} f_{\mathrm{I}_1} + \cdots + \begin{bmatrix} x_1^{\mathrm{Ip}} \\ x_2^{\mathrm{Ip}} \\ \vdots \\ x_n^{\mathrm{Ip}} \end{bmatrix} f_{\mathrm{I_p}} \right\} + \left\{ \begin{bmatrix} x_1^{S1} \\ x_2^{S1} \\ \vdots \\ x_n^{S1} \end{bmatrix} f_{S_1} + \cdots + \begin{bmatrix} x_1^{Sq} \\ x_2^{Sq} \\ \vdots \\ x_n^{Sq} \end{bmatrix} f_{S_q} \right\} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

同时包含截距项和行业哑元变量将导致解不唯一，此处加入限制条件：

$$w_{\mathrm{I}_1} f_{\mathrm{I}_1} + w_{\mathrm{I}_2} f_{\mathrm{I}_2} + \cdots + w_{\mathrm{I_p}} f_{\mathrm{I_p}} = 0$$

其中，w_{Ip} 表示属于该行业的所有股票的流通市值之和。

以上是常见的 4 中形式的模型，前两种模型和 Barra 的方式不同之处在于，Barra 的因子收益 f_i 无法直接观察到，是通过在时间截面上回归得到的（后验），因子暴露度则是

β

由基本面、技术面等因子的值计算截面 z-score 直接得到的;而 CAMP 和 Fama-French 中的因子收益不需要截面回归,可以直接观察到(多空组合、后验),但因子暴露度(或称)需要通过历史时间序列回归来估计。

四、理解因子暴露度

因子暴露度是通过将因子值正态标准化得到:

$$x_i^S = \frac{x_{\mathrm{raw}_i}^S - m}{\mu}$$

但值得注意的是,m 是 $x_{\mathrm{raw}_i}^S$ 按流通市值加权的平均值 $\sum w_{cap}^i x_{\mathrm{raw}_i}^S$,而非算术平均,μ 是的样本标准差。这样的标准化方式可以使以流通市值为权重的投资组合(基准组合)对所有风格因子 0 暴露,因为投资组合对风格因子的暴露:

$$\boldsymbol{w}_{cap}^T \boldsymbol{x}^S = \sum w_{cap}^i x_i^S = \frac{\sum w_{cap}^i x_{\mathrm{raw}_i}^S - m}{std(\boldsymbol{x_{raw}})} = 0$$

简言之,基准组合对各个因子的因子暴露为 0,非零的因子暴露度就是投资组合偏离了基准组合多少倍标准差。

在此种限制条件下,流通市值加权的组合(基准组合)的收益就是截距项,f_c 因为流通市值加权的组合(基准组合)的收益:

$$\begin{aligned}
\mathrm{r_p} &= \boldsymbol{w}_{cap}^T \boldsymbol{r} \\
&= \boldsymbol{w}_{cap}^T \left[\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} f_c + \left\{ \begin{bmatrix} x_1^{\mathrm{I}1} \\ x_2^{\mathrm{I}1} \\ \vdots \\ x_n^{\mathrm{I}1} \end{bmatrix} f_{\mathrm{I}_1} + \cdots + \begin{bmatrix} x_1^{\mathrm{Ip}} \\ x_2^{\mathrm{Ip}} \\ \vdots \\ x_n^{\mathrm{Ip}} \end{bmatrix} f_{\mathrm{I}_{28}} \right\} + \left\{ \begin{bmatrix} x_1^{S1} \\ x_2^{S1} \\ \vdots \\ x_n^{S1} \end{bmatrix} f_{S_1} + \cdots + \begin{bmatrix} x_1^{Sq} \\ x_2^{Sq} \\ \vdots \\ x_n^{Sq} \end{bmatrix} f_{S_q} \right\} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \right] \\
&= f_c + \left(w_{\mathrm{I}_1} f_{\mathrm{I}_1} + w_{\mathrm{I}_2} f_{\mathrm{I}_2} + \cdots + w_{\mathrm{I_p}} f_{\mathrm{I_p}} \right) + \left(\boldsymbol{w}_{cap}^T \boldsymbol{x}^{S1} f_{S_1} + \cdots + \boldsymbol{w}_{cap}^T \boldsymbol{x}^{Sq} f_{S_q} \right) + \boldsymbol{w}_{cap}^T \boldsymbol{u} \\
&= f_c + \boldsymbol{w}_{cap}^T \boldsymbol{u}
\end{aligned}$$

五、鸿沟:从解释到预测

上文介绍模型时我们总是将“分解”二字加粗,想要强调的是,不论 Barra 还是 CAMP 和 Fama-French,都只是收益的解释模型,并不包含任何预测信息,套用陆游的名句,汝果欲做预测,功夫还在模型外。

前文均在截面上讨论,忽略了时间下标,将模型加入时间下标后的表示应该是:

$$\boldsymbol{r}^{t1} = \boldsymbol{X}^{t0} \boldsymbol{f}^{t1} + \boldsymbol{u}^{t1}$$

其中,$\boldsymbol{r}^{t1}$表示 t0 到 t1 区间上的股票超额无风险收益率的收益,$\boldsymbol{X}^{t0}$ 则是在期初对因子暴露度的估计,f^{t1} 表示 t0 到 t1 区间的因子收益。

要预测未来的股票收益,就需要知道未来的因子收益,从未来到未来,可见,模型本身并不提供预测。但模型的功用至少有三点:

一是，事后对取得的收益做分解和评价，毕竟如果不知道收益从何处而来，也会不知道今后收益为何而去；

二是，起降维的作用，从预测 n 只股票到只需预测 m 个因子的收益$(m \ll n)$

三是，找到合适的因子，考察因子的对应的纯因子组合的收益序列，避免了简单排序导致对其他因子的暴露，更能反映该因子的历史收益情况。

如果仅从形式上来说，将模型右侧的 $\boldsymbol{f}^{t1}$ 改为 $\boldsymbol{f}^{t0}$ 即构成一个看起来不错的预测模型：

$$\boldsymbol{r}^{t1} = \boldsymbol{X}^{t0}\boldsymbol{f}^{t0} + \boldsymbol{u}^{t1}$$

这其实隐含了预测 $\boldsymbol{f}^{t1} = \boldsymbol{f}^{t0}$，我们称这是一个朴素预测（Naïve Forecast），即“过去即未来”。在没有任何信息的情况下，这也是一种选择，但因子收益一般是轮动的，我们有理由追求更好的轮动模型，这是本系列报告将来会讨论的。

在“预测”的语境下，模型通常被写作：

$$\hat{\boldsymbol{r}}^{t1} = \boldsymbol{X}^{t0}\boldsymbol{w}^{t0}$$

其中，$\boldsymbol{X}^{t0}$仍然是因子值或因子暴露，$\boldsymbol{w}^{t0}$ 是需要在期初确定的各因子权重。可见，决定因子权重，就是预测未来因子收益。若令 $\boldsymbol{w}^{t0} = \boldsymbol{f}^{t0}$,则是给上期表现好的因子更高的权重（动量），也可以做反转预测 $\boldsymbol{w}^{t0} = -\boldsymbol{f}^{t0}$，这是因子的动态配置的例子；或者还可以静态的给各因子等权重 $\boldsymbol{w}^{t0} = \boldsymbol{c}$。

另外，模型默认了组合是充分分散的，即组合的特异收益率$\mathbf{w}_{\mathbf{P}}^{T}\boldsymbol{u} \to 0$,如果股票过少（例如只选择 $\hat{\boldsymbol{r}}^{t1}$ 最大的 10 只股票），只能寄希望于模型无法解释的 $\boldsymbol{u}$ 碰巧和同 $\hat{\boldsymbol{r}}^{t1}$方向了，即多因子持股不适用于集中持股的策略。

此篇报告并不会讨论因子轮动或因子择时，我们将这个最难的问题放在这一多因子系列报告的最后一篇。

由于模型和预测之间的鸿沟如此巨大，报告中提到模型时都避免使用“预测”的字样。

六、模型的解

线性模型

$$\boldsymbol{r} = \boldsymbol{X}\boldsymbol{f} + \boldsymbol{u}$$

的解十分简单：

$$\boldsymbol{f}=\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{W}\boldsymbol{r}$$

其中 $\boldsymbol{W}$ 为回归权重，理想取值是残差收益率的方差的倒数 $1/\mathrm{Var}(\boldsymbol{u})$，但解出该模型前残差收益率未知，一般使用市值的平方根作为权重。

此解不仅给出了因子收益率 $\boldsymbol{f}$ 的值，更是给出了纯因子组合的权重。因为，解的右边部分 $\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{W}$，一定可以写为行向量的形式：

$$\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{W}=\begin{bmatrix}\boldsymbol{w}_{P_1}^{T}\\ \boldsymbol{w}_{P_2}^{T}\\ \vdots\\ \boldsymbol{w}_{P_m}^{T}\end{bmatrix}$$

所以

$$\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_m\end{bmatrix}=\boldsymbol{f}=\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{W}\boldsymbol{r}=\begin{bmatrix}\boldsymbol{w}_{P_1}^{T}\boldsymbol{r}\\ \boldsymbol{w}_{P_2}^{T}\boldsymbol{r}\\ \vdots\\ \boldsymbol{w}_{P_m}^{T}\boldsymbol{r}\end{bmatrix}$$

即

$$f_1=\boldsymbol{w}_{P_1}^{T}\boldsymbol{r},f_2=\boldsymbol{w}_{P_2}^{T}\boldsymbol{r},\dots,f_m=\boldsymbol{w}_{P_m}^{T}\boldsymbol{r}$$

其中，$\boldsymbol{w}_{P_1}^{T},\boldsymbol{w}_{P_2}^{T},\dots,\boldsymbol{w}_{P_m}^{T}$ 就是相应因子的纯因子投资组合的个股权重。

可以验证：

$$\begin{aligned}f_m=\ \boldsymbol{w}_{P_m}^{T}\boldsymbol{r}&=\boldsymbol{w}_{P}^{T}\boldsymbol{x}_1f_1+\boldsymbol{w}_{P}^{T}\boldsymbol{x}_2f_2+\cdots+\boldsymbol{w}_{P}^{T}\boldsymbol{x}_mf_m+\boldsymbol{w}_{P}^{T}\boldsymbol{u}\\&=x_P^1f_1+x_P^2f_2+\cdots+x_P^mf_m+\boldsymbol{w}_{P}^{T}\boldsymbol{u}\end{aligned}$$

要使等式左右相等，需要有 $x_P^1=0,x_P^2=0,\dots,x_P^m=1$ 。

虽然回归需要用到当期收益率，但计算纯因子组合的组合权重只需要期初的因子暴露度数据即可：

$$\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{W}=\begin{bmatrix}\boldsymbol{w}_{P_1}^{T}\\ \boldsymbol{w}_{P_2}^{T}\\ \vdots\\ \boldsymbol{w}_{P_m}^{T}\end{bmatrix}$$

即计算因子收益 f_m 需要股票收益 $\boldsymbol{r}$，但计算纯因子组合权重不需要。

对于带有等式约束的模型，只需引入拉格朗日乘子，而后如法炮制，此处不再赘述。

七、共线性问题和 t 统计量

共线性问题

要想顺利求解，必须要求 $\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\in\mathbf{R}^{m\times m}$ 可逆。如果两个因子存在完全的线性关系，将使

$$\mathrm{rank}\left(\mathrm{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\right)<m$$

该矩阵减少一个自由度，即　　　　　，矩阵完全不可逆。通常这种完全的线性关系不会发生，但存在相关关系的两个因子导致矩阵近乎不可逆，勉强求出的解数值误差极大，即便因子本身多么优秀，之后的工作也是徒劳。

这里列举两例常见共线性的陷阱来说明问题。

行业因子和常数项的共线性(假设 4 只股票，分属两个行业)，因子模型为：

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

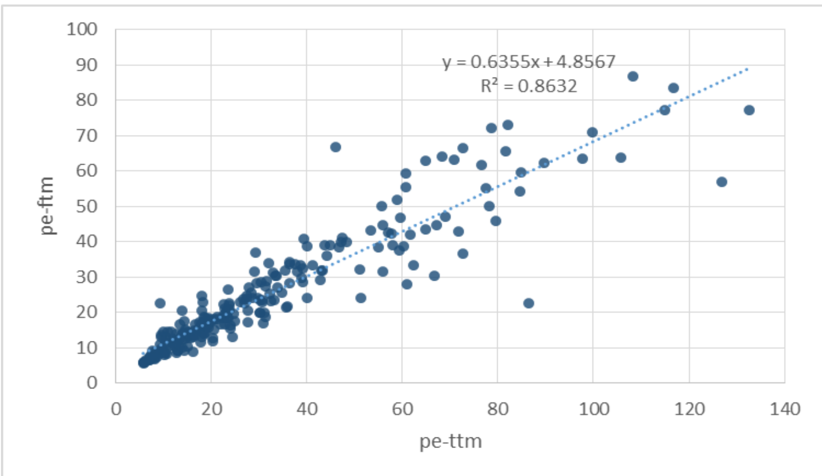
$$\mathrm{rank}(\mathbf{X}^{\mathbf{T}}\boldsymbol{X}) = \mathrm{rank}\left(\begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}\right) = 2$$

此时还需要加入一个条件或去掉常数项，才能使矩阵满秩。

同时包含 PE-ttm 因子和一致预期 PE 因子的模型

沪深 300 指数成分股 2016 年 4 月 1 日提取的 PE-ttm 和未来 12 月一致预期 PE-ftm 散点图：

图 1：PE-ttm 因子和一致预期 PE 因子相关性



资料来源：Wind，长江证券研究所

因子模型：

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{300} \end{bmatrix} = \begin{bmatrix} 1 & pe_{ttm}^1 & pe_{ftm}^1 \\ 1 & pe_{ttm}^2 & pe_{ftm}^2 \\ \vdots & \vdots & \vdots \\ 1 & pe_{ttm}^{300} & pe_{ftm}^{300} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{300} \end{bmatrix}$$

信息矩阵 $\mathbf{X}^{\mathbf{T}}\boldsymbol{X}$ 虽然可逆，但其特征值是 $[6.1 \times 10^5; 7.8 \times 10^3; 76.8]$ ，条件数

是 8.0×10^3，引入不小的数值计算误差。

过小的特征值也会增大因子估计的方差，导致 t 值变小。该截面回归上，pe-ttm 和 pe-ftm 的 t 值分别为-0.15 和-0.44；如果分别构建单因子模型，pe-ttm 的 t 值为-1.53，pe-ftm 的 t 值为-1.58。

所以在构建模型时，特别是加入新的因子时，对共线性的检查尤为重要，否则一个共线性坏了一锅模型。通常检查的方法是计算方差膨胀因子（VIF）值，通常 VIF 大于 5 就认为该因子和其它某因子存在较强共线性（但无法给出是和哪个因子共线性）。因子的 VIF 的计算方法是：

首先，构建该因子和其他因子的线性回归模型

$$\boldsymbol{x}_i = a_0 + a_1\boldsymbol{x}_1 + \cdots + a_{i-1}\boldsymbol{x}_{i-1} + a_{i+1}\boldsymbol{x}_{i+1} + \cdots + a_m\boldsymbol{x}_m + \boldsymbol{\varepsilon}$$

在截面上回归，并计算该模型的 R_i^2，则该因子对应的 VIF 值就是

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

对各个因子重复上面步骤，求出所有因子的 VIF 值。由上述步骤可以看出，VIF 值不需要当期股票收益，也是先验的。

如果发现存在共线性因子，还可通过回归时的信息矩阵的特征值和条件指数（condition index）对照方差分解比例（variance decomposition proportions）来确定哪些因子存在共线性。

t 统计量

从因子收益的角度理解 t 值更直观。 计算出 f_i 后通常还会计算每个因子的 t 统计量，如果 $|t| > 2$ 则认为因子显著。但是，“因子显著”是在说因子的什么显著？回归本源，此处 t 检验的完整表达是：

$$\mathrm{H}_0: f_i \neq 0$$
$$\mathrm{H}_1: f_i = 0$$

$$若|t| = \left|\frac{f_i - 0}{s.e(f_i)}\right| > 2, 则在0.05的显著性水平下，拒绝\mathrm{H}_0；否则，无法拒绝\mathrm{H}_0$$

所以“因子显著”的意义是，如果 $|t| > 2$，我们观察到的因子收益 $f_i \neq 0$ 是可以相信的；当 $|t| < 2$ 时，我们观察到的因子收益 $f_i \neq 0$ 可能仅仅是由随机波动导致的。

既然是因子收益，就不能希望期期都有显著收益（特别是行业因子），所以应当考察因子收益的持续性，并坦然接受因子有可能（较长时间）失效。Barra 的指标是计算历史上发生的 $|t| > 2$次数占比。

八、纯因子组合----以上证 50 为例

前文已经阐明，截距项因子收益 f_c 对应的组合是基准组合，基准组合对各风格因子的暴露均为 0，对自身暴露为 1（全额投资），回归得到的纯 f_c 因子组合的权重应该和市值加权十分接近。

风格因子收益 f_{S_n} 对应的纯因子组合，对截距项暴露为 0（零额投资），即该组合一定为多空组合，且多头/空头为 100/100，对各行业因子暴露为 0（在各行业中也是 100/100 零额投资），对其他风格因子暴露为 0，对自身风格因子偏离基准组合 1 个标准差。

行业因子收益 f_{I_n} 对应的纯因子组合对本行业暴露为 1（即权重 100%），对截距项暴露为 0（零额投资），所以需要 100%自身行业多头，同时 100%基准组合空头，即行业因子收益本身就是一种超额收益，并且该组合对各个风格因子暴露为 0。

因为上证 50 股票数量少，适合作为例子展示结果，我们以上证 50 为例来说明纯因子组合的上述特性比较容易。同时我们将模型简化为包含截距项、银行行业、非银金融行业、其他行业、规模、BP 等 6 个因子的模型，模型的形式为：

$$\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{50} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} f_c + \begin{bmatrix} x_1^{\text{银行}} \\ x_2^{\text{银行}} \\ \vdots \\ x_{50}^{\text{银行}} \end{bmatrix} f_{\text{银行}} + \begin{bmatrix} x_1^{\text{非银}} \\ x_2^{\text{非银}} \\ \vdots \\ x_{50}^{\text{非银}} \end{bmatrix} f_{\text{非银}} + \begin{bmatrix} x_1^{\text{其它}} \\ x_2^{\text{其它}} \\ \vdots \\ x_{50}^{\text{其它}} \end{bmatrix} f_{\text{其它}} + \begin{bmatrix} x_1^{\text{size}} \\ x_2^{\text{size}} \\ \vdots \\ x_{50}^{\text{size}} \end{bmatrix} f_{\text{size}} + \begin{bmatrix} x_1^{\text{bp}} \\ x_2^{\text{bp}} \\ \vdots \\ x_{50}^{\text{bp}} \end{bmatrix} f_{\text{bp}} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{50} \end{bmatrix}$$

并附带前文所述限制条件以保证解唯一。

在 2016 年 4 月 1 日截面上，提取因子暴露度数据，如下表：

表 1：上证 50 成分股的因子暴露值

股票名称	银行行业因子暴露	非银行业因子暴露	其他行业因子暴露	BP 因子暴露	规模因子暴露	流通市值权重	回归权重
工商银行	1	0	0	1.00	1.07	10.59%	5.26%
农业银行	1	0	0	0.99	1.07	8.60%	4.74%
中国银行	1	0	0	1.33	0.78	6.57%	4.14%
招商银行	1	0	0	0.06	-0.03	3.07%	2.83%
浦发银行	1	0	0	-0.07	-0.02	3.10%	2.84%
兴业银行	1	0	0	0.51	-0.16	2.72%	2.66%
民生银行	1	0	0	0.10	-0.25	2.50%	2.55%
交通银行	1	0	0	1.33	-0.47	2.02%	2.30%
中信银行	1	0	0	0.67	-0.61	1.78%	2.15%
光大银行	1	0	0	1.33	-0.88	1.38%	1.89%
北京银行	1	0	0	-0.05	-1.06	1.16%	1.74%

华夏银行	1	0	0	0.54	-1.56	0.73%	1.38%
中国人寿	0	1	0	-1.27	0.40	4.56%	3.45%
中国平安	0	1	0	-0.95	0.01	3.16%	2.87%
中国太保	0	1	0	-1.00	-0.78	1.51%	1.99%
中信证券	0	1	0	-0.70	-0.73	1.59%	2.04%
CodeX	0	1	0	-0.81	-2.60	0.27%	0.84%
海通证券	0	1	0	-0.70	-1.15	1.06%	1.67%
招商证券	0	1	0	-1.31	-1.47	0.79%	1.43%
CodeXX	0	1	0	-1.74	-1.81	0.57%	1.22%
华泰证券	0	1	0	-0.67	-1.39	0.85%	1.49%
新华保险	0	1	0	-1.32	-1.50	0.77%	1.41%
方正证券	0	1	0	-1.09	-2.07	0.45%	1.09%
国金证券	0	1	0	-1.57	-2.29	0.37%	0.98%
中国石油	0	0	1	-0.07	1.07	10.59%	5.44%
中国神华	0	0	1	0.57	-0.41	2.14%	2.36%
上汽集团	0	0	1	-0.41	-0.45	2.05%	2.31%
中国中车	0	0	1	-1.70	-0.40	2.15%	2.37%
中国建筑	0	0	1	0.22	-0.74	1.57%	2.03%
中国中铁	0	0	1	-0.55	-0.96	1.28%	1.83%
中国交建	0	0	1	-0.39	-0.93	1.31%	1.85%
中国铁建	0	0	1	-0.48	-1.14	1.08%	1.68%
中国重工	0	0	1	-1.35	-1.06	1.17%	1.75%
上港集团	0	0	1	-1.28	-1.09	1.13%	1.72%
中国核电	0	0	1	-1.82	-2.60	0.27%	0.85%
包钢股份	0	0	1	-1.18	-2.11	0.43%	1.06%
大秦铁路	0	0	1	-0.04	-1.28	0.95%	1.57%
保利地产	0	0	1	-0.68	-1.31	0.92%	1.55%
伊利股份	0	0	1	-2.09	-1.46	0.80%	1.44%
中国联通	0	0	1	-0.15	-1.36	0.88%	1.51%
中国电建	0	0	0	0.00	0.00	0.00%	0.00%
中航动力	0	0	1	-2.16	-1.99	0.48%	1.12%
东方明珠	0	0	1	-1.80	-2.14	0.42%	1.05%
康美药业	0	0	1	-1.97	-1.72	0.63%	1.28%
海螺水泥	0	0	1	-0.26	-1.73	0.62%	1.27%
国电电力	0	0	1	-0.06	-1.85	0.55%	1.20%
北方稀土	0	0	1	-2.16	-2.60	0.26%	0.83%
中国船舶	0	0	1	-1.15	-2.45	0.31%	0.91%
中国神华	0	0	1	0.57	-0.41	2.14%	2.36%

资料来源：长江证券研究所，天软科技

依据前面“模型的解”一节内容，由截面数据便可计算出各因子对应的纯因子组合的权重，即虽然组合的收益是未知，但纯因子组合的权重是先验可知的。求解出的各因子对应的纯因子组合的个股权重见下表。

表 2：各因子对应的纯因子组合的个股权重

股票名称	纯 fc 因子组合	纯银行因子组合	纯非银因子组合	纯其他行业因子	纯 BP 因子组合	纯规模因子组合
工商银行	9.09%	5.42%	-3.74%	-4.51%	-0.11%	5.96%
农业银行	8.19%	4.96%	-3.46%	-4.12%	-0.22%	5.41%
中国银行	6.59%	2.30%	-0.52%	-2.35%	3.96%	1.98%
招商银行	3.40%	7.60%	-7.09%	-5.60%	-4.05%	0.72%
浦发银行	3.43%	8.22%	-7.79%	-5.99%	-4.95%	1.04%
兴业银行	3.04%	5.30%	-4.47%	-4.09%	-0.73%	-0.71%
民生银行	2.80%	6.75%	-6.20%	-5.01%	-2.95%	-0.25%
交通银行	2.28%	1.69%	-0.42%	-1.71%	4.34%	-3.06%
中信银行	1.99%	3.87%	-3.02%	-3.09%	0.96%	-2.24%
光大银行	1.51%	1.51%	-0.39%	-1.52%	4.15%	-3.65%
北京银行	1.23%	5.22%	-4.77%	-3.88%	-1.57%	-2.02%
华夏银行	0.65%	2.95%	-2.33%	-2.35%	1.16%	-3.19%
中国人寿	4.85%	-2.26%	13.11%	-2.75%	-4.97%	7.04%
中国平安	3.50%	-3.16%	12.54%	-1.51%	-1.16%	3.56%
中国太保	1.68%	-1.81%	8.43%	-1.37%	0.13%	0.29%
中信证券	1.77%	-2.83%	9.75%	-0.76%	1.48%	-0.01%
CodeX	-0.02%	-0.81%	3.81%	-0.63%	1.59%	-2.22%
海通证券	1.11%	-2.23%	7.97%	-0.72%	1.76%	-1.03%
招商证券	0.73%	-0.46%	5.22%	-1.58%	-0.21%	-0.90%
CodeXX	0.43%	0.50%	3.47%	-1.94%	-1.11%	-0.97%
华泰证券	0.83%	-2.01%	7.19%	-0.65%	1.93%	-1.46%
新华保险	0.70%	-0.43%	5.12%	-1.57%	-0.20%	-0.94%
方正证券	0.25%	-0.64%	4.36%	-1.04%	0.89%	-1.79%
国金证券	0.12%	0.20%	3.06%	-1.44%	-0.14%	-1.58%
中国石油	9.63%	-11.52%	1.78%	12.07%	0.34%	12.05%

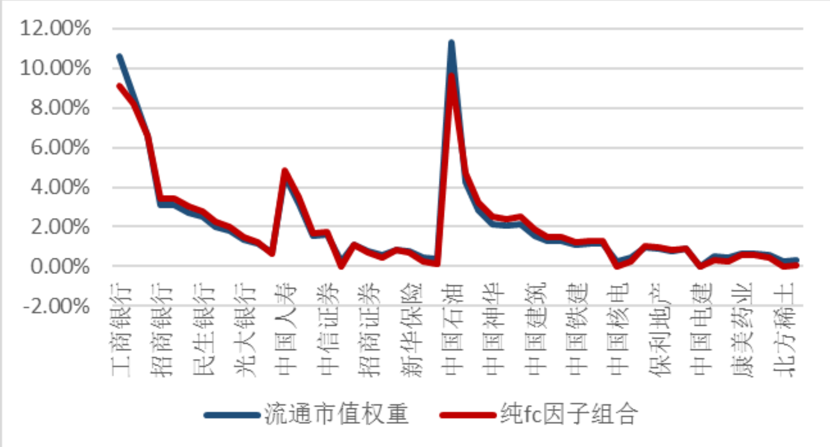
中国石化	4.73%	-11.88%	6.97%	10.39%	9.70%	1.39%
贵州茅台	3.25%	3.66%	-9.65%	-0.20%	-10.70%	5.55%
中国神华	2.51%	-6.90%	3.39%	6.29%	6.30%	-0.94%
上汽集团	2.41%	-3.15%	-0.86%	3.84%	0.92%	0.62%
中国中车	2.52%	1.61%	-6.51%	0.83%	-6.35%	3.09%
中国建筑	1.84%	-4.71%	1.58%	4.59%	4.24%	-1.24%
中国中铁	1.47%	-1.96%	-1.18%	2.65%	0.81%	-0.65%
中国交建	1.51%	-2.44%	-0.67%	2.98%	1.46%	-0.79%
中国铁建	1.19%	-1.92%	-0.90%	2.49%	1.24%	-1.12%
中国重工	1.31%	0.38%	-3.72%	1.07%	-2.39%	0.17%
上港集团	1.26%	0.18%	-3.43%	1.17%	-2.01%	-0.01%
中国核电	0.01%	0.99%	-2.58%	-0.06%	-1.11%	-1.51%
包钢股份	0.27%	0.11%	-1.99%	0.68%	-0.20%	-1.64%
大秦铁路	1.02%	-2.86%	0.41%	3.01%	2.93%	-1.88%
保利地产	0.97%	-1.25%	-1.40%	1.95%	0.62%	-1.19%
伊利股份	0.80%	2.08%	-5.06%	-0.28%	-4.04%	0.09%
中国联通	0.92%	-2.49%	0.10%	2.72%	2.54%	-1.86%
中国电建	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
中航动力	0.34%	1.82%	-4.10%	-0.38%	-2.88%	-0.74%
东方明珠	0.25%	1.13%	-3.15%	0.01%	-1.69%	-1.18%
康美药业	0.56%	1.64%	-4.21%	-0.14%	-2.97%	-0.51%
海螺水泥	0.55%	-1.80%	-0.20%	2.08%	2.16%	-2.14%
国电电力	0.45%	-2.05%	0.24%	2.18%	2.70%	-2.41%
北方稀土	0.01%	1.42%	-3.05%	-0.35%	-1.74%	-1.27%
中国船舶	0.08%	0.08%	-1.65%	0.57%	0.14%	-1.87%

资料来源：天软科技，长江证券研究所

下面来验证上述论断。

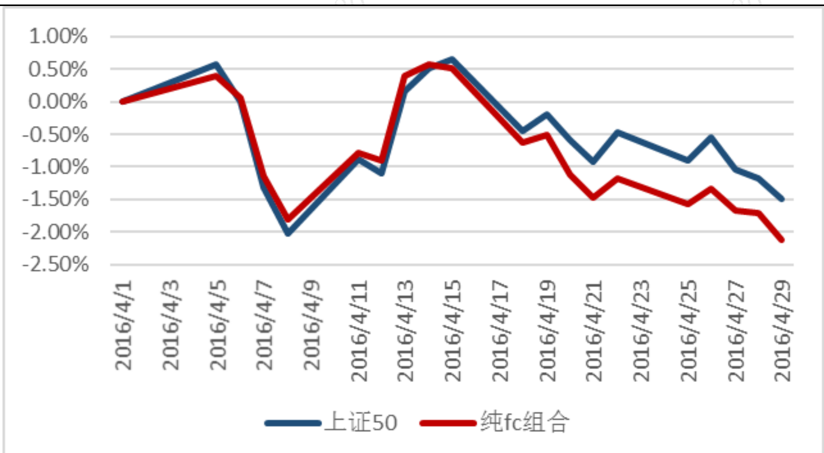
验证纯 fc 因子组合接近市值加权组合。

图 2：纯 fc 因子组合接近市值加权组合



资料来源：天软科技，长江证券研究所

图 3：纯 fc 因子组合接近指数走势



资料来源：天软科技，长江证券研究所

与上证 50 指数的偏差来源主要是流通市值权重和上证 50 指数权重不完全相同。

计算各个纯因子组合的权重之和、在各行业中的权重，以及对风格因子的暴露度。

表 3：纯因子组合的对各种因子的暴露

	对 fc 因子暴露度 $\sum w_i$	对银行因子暴露度 $\sum w_i x_i^{银行}$	对非银因子暴露度 $\sum w_i x_i^{非银}$	对其他行业因子暴露度 $\sum w_i x_i^{其它}$	对 BP 因子暴露度 $\sum w_i x_i^{bp}$	对规模因子暴露度 $\sum w_i x_i^{size}$
纯 fc 因子组合	100.00%	44.20%	15.96%	39.85%	-1.07E-15	1.16E-15
纯银行因子组合	0.00%	55.80%	-15.96%	-39.85%	-1.10E-15	-8.89E-16
纯非银因子组合	0.00%	-44.20%	84.04%	-39.85%	2.12E-15	-4.86E-17
纯其他行业因子组合	0.00%	-44.20%	-15.96%	60.15%	-6.94E-18	-1.41E-16
纯 BP 因子组合	0.00%	0.00%	0.00%	0.00%	1.00E+00	-1.91E-17
纯规模因子组合	0.00%	0.00%	0.00%	0.00%	6.25E-17	1.00E+00

资料来源：长江证券研究所，天软科技

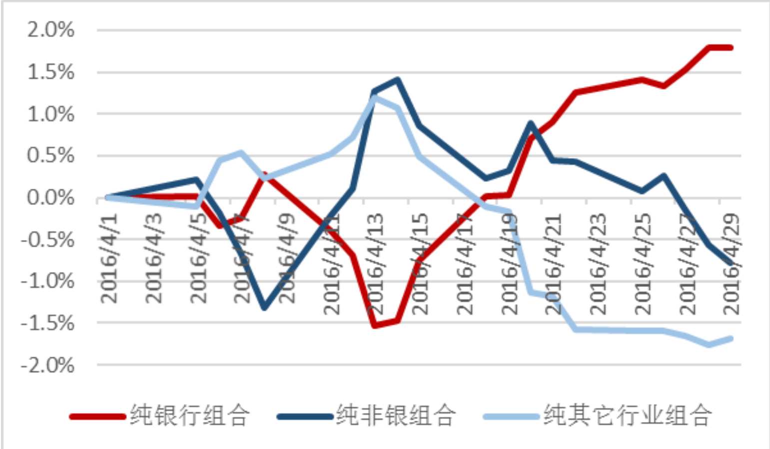
可见：

1. 纯 fc 因子组合就是基准组合（全额投资，行业中性，对各风格因子 0 暴露）。
2. 纯银行因子组合持有银行行业 100%多头，持有 fc 组合 100% 空头（银行 44.2%、非银 15.96%、其它 39.85%），最终权重为 55.8%银行，-15.96%非银，-39.85%其他行业，是一个零额投资的多空组合，并且对各种风格因子暴露为 0。
3. 非银行业和其他行业的纯因子组合类似。
4. 纯风格因子组合也是零额投资的多空组合，并且在各行业中也是这样，同时对模型中其他风格因子暴露为 0，对自身暴露度为 1。

九、纯因子组合的用武之地

一个自然的延伸是通过纯因子组合的表现来衡量因子的显著性。既知组合权重，就可观察各纯因子组合在当月的表现，如图 4、图 5。

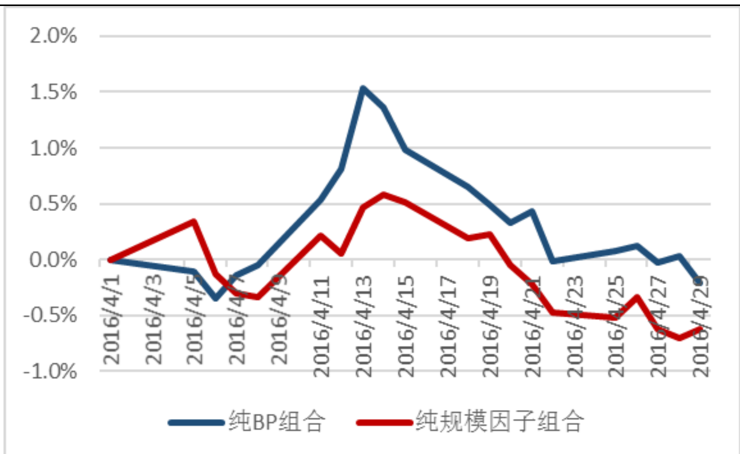
图 4：纯行业因子组合走势



资料来源：天软科技 长江证券研究所

在剔除风格暴露影响后，银行组合从 4 月中旬开始发力。

图 5：纯风格因子组合走势



资料来源：天软科技，长江证券研究所

此处作为例子，仅在一个时间截面上做了回归，通过在一系列历史截面上回归，可以分析纯因子组合的历史收益。由于纯因子组合对模型中其他因子的暴露为 0，相比排序进行多空组合，更能检验因子的作用力。

再者，通过分析纯因子组合在各个行业中的收益率，还可以了解因子在不同行业中的特性，以便因行业而制宜。分析纯因子组合在各个行业中的收益率：

表 4：纯因子组合的在各个行业的表现

	在银行业中获得的收益	在非银金融行业中获得的收益	在其他行业中获得的收益
纯银行因子组合	0.4%	0.2%	1.2%
纯非银因子组合	-0.5%	-2.4%	2.2%
纯其他行业因子组合	-0.2%	0.7%	-2.1%
纯 BP 因子组合	-0.8%	0.5%	0.1%
纯规模因子组合	-0.1%	-0.2%	0.7%

资料来源：Wind，长江证券研究所

纯银行、纯非银因子组合的收益主要来自卖空其他行业。低 PB 股票在银行业中收益为负；小市值股票在银行和非银中收益为负。不过，这只是在 2016 年 4 月这个截面上的例子，主要展示如何通过纯因子组合分析因子的效用，并不代表一般结论。

除了检验因子效果、分析收益来源外，纯因子组合还可以用于构建投资组合，使投资组合只暴露在想要暴露的风险因子上，实现精确地风险收益匹配。已经注意到，纯行业、纯风格因子组合是一个多空组合，在存在卖空限制的市场上没有办法实现，下面我们还是依上文上证 50 的例子构建一个在全额投资和不允许卖空的限制条件的可以投资的组合。

纯 fc 因子组合是一个全额投资组合，纯行业、纯风格因子都是零额投资组合，所以纯 fc 因子组合和任意行业、风格因子的组合是一个全额投资组合。因为，

$$\boldsymbol{w}_p = \boldsymbol{w}_{fc} + a_1\boldsymbol{w}_{银行} + a_2\boldsymbol{w}_{非银} + a_3\boldsymbol{w}_{其他} + a_4\boldsymbol{w}_{size} + a_5\boldsymbol{w}_{bp}$$

有：

$$\begin{aligned} & sum(\boldsymbol{w}_p) \\ & = sum(\boldsymbol{w}_{fc}) + a_1 sum(\boldsymbol{w}_{银行}) + a_2 sum(\boldsymbol{w}_{非银}) + a_3 sum(\boldsymbol{w}_{其他}) + a_4 sum(\boldsymbol{w}_{size}) + a_5 sum(\boldsymbol{w}_{bp}) \\ & = 1 + a_1 \cdot 0 + a_2 \cdot 0 + a_3 \cdot 0 + a_4 \cdot 0 + a_5 \cdot 0 = 1 \end{aligned}$$

所以全额投资是自动满足的。

不允许卖空是一个线性不等约束条件：

$$\boldsymbol{w}_{fc} + a_1\boldsymbol{w}_{银行} + a_2\boldsymbol{w}_{非银} + a_3\boldsymbol{w}_{其他} + a_4\boldsymbol{w}_{size} + a_5\boldsymbol{w}_{bp} \geq 0$$

再给定目标函数，即是一个完备的最优化问题。

假设目标是最大化对 BP 因子的暴露，但要求对其他因子 0 暴露，则问题的完整表述是：

$$a_5 = \operatorname{argmax}(a_5)$$

$$
\text{s.t}\ \begin{cases} \boldsymbol{w}_{fc}+a_1\boldsymbol{w}_{\text{银行}}+a_2\boldsymbol{w}_{\text{非银}}+a_3\boldsymbol{w}_{\text{其他}}+a_4\boldsymbol{w}_{size}+a_5\boldsymbol{w}_{bp}\geq -10^3 \\ -10^{-5}\leq a_1,a_2,a_3,a_4\leq 10^{-5} \end{cases}
$$

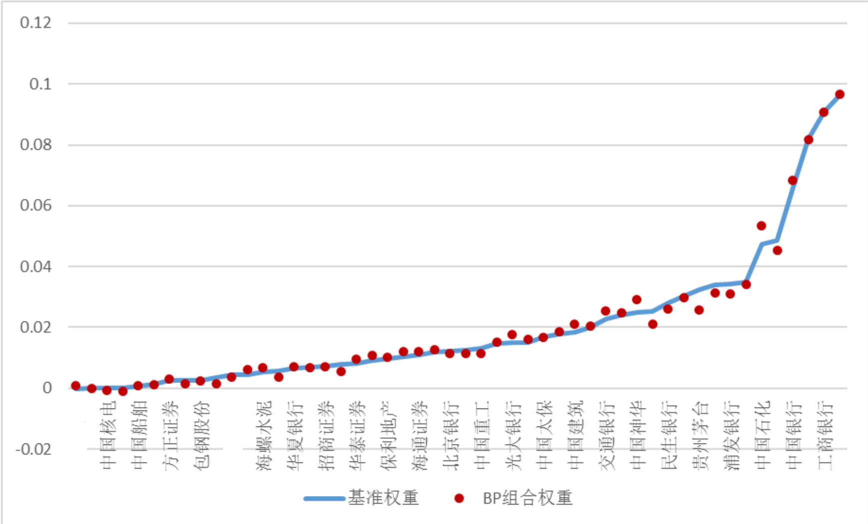
这是一个线性规划问题，丢给任何一个优化器（如 matlab: linprog），就可求得在限制条件下能取得的最大暴露度 a_5。（此处采取了更宽松的限制条件，以期解存在。）

将上证 50 例子的数据带入，求得

$$
a_5=0.0633, a_1=-10^{-5}, a_2=10^{-5}, a_3=-10^5, a_4=-10^5
$$

在全额投资和不允许卖空的限制下，保证对其它因子 0 暴露时，能达到最大的 BP 因子暴露仅为 0.0633。$\boldsymbol{w}_p$和基准组合的权重差别如图。

图 6：限制条件下的纯因子组合权重



资料来源：天软科技，长江证券研究所

因子暴露度和可投资性之间通常需要妥协。流通市值加权基准指数是容量最大的，对各因子 0 暴露，可投资性最强；另类加权的策略，如等权重（加强市值因子暴露）、波动率加权（加强波动率因子暴露）等，持有基准指数所有股票，仅作权重的倾斜，可投资性次之；再增加因子暴露度则是只持有一部分基准指数成分股，同时权重倾斜向愿意暴露的因子，此时可投资性再下一等；再继续则是多空组合（如 130/30）；因子暴露度最大、可投资性最低的则是纯因子组合。

在追求纯因子暴露和可投资性之间妥协，可以适当放宽对其他因子的要求，构建出对某因子有较高暴露度，对其它因子有适当暴露度的可投资的高暴露因子组合。

十、另一个角度理解纯因子组合的权重

从因子暴露度的角度来说，某个因子对应的纯因子组合的权重对模型中的行业因子暴露为零，对其他风格因子风格暴露为零。另一方面，从数学语言上来说，这个组合权重与其他因子暴露度正交：

$$\begin{cases} \boldsymbol{w}_{\mathrm{P}_m}^T \boldsymbol{x}_i = 0, \forall\, i \neq m \\ \boldsymbol{w}_{\mathrm{P}_m}^T \boldsymbol{x}_m = \mathbf{1} \end{cases}$$

即求解纯因子组合权重的过程，就是正交化的过程，得到的值 $\boldsymbol{w}_{\mathrm{P}_m}$ 可以理解为原因子值$\boldsymbol{x}_m$对行业因子和其他风格因子正交化后的新的因子值，而且所有因子值的正交化可以通过前述求解过程一步求出，十分方便。

在这种视角下，上节中对各纯因子组合的线性规划，就是将因子值正交化后为各因子分配权重：

$$\boldsymbol{w}_p = \boldsymbol{w}_{fc} + a_1 \boldsymbol{w}_{\text{银行}} + a_2 \boldsymbol{w}_{\text{非银}} + a_3 \boldsymbol{w}_{\text{其他}} + a_4 \boldsymbol{w}_{size} + a_5 \boldsymbol{w}_{bp}$$

同样，计算纯因子组合的权重和未来股票收益率的相关数，可认为是正交化后的因子的IC。

十一、系列报告后续计划

本片报告是多因子系列报告的第一篇，主要目的是在梳理模型理论和探讨纯因子组合的运用，但并没有涉及任何实际的投资策略。

接下来的计划是：

1. 利用纯因子组合来检验常见因子的显著性，筛选出几个比较能持续产生因子收益的常见因子。
2. 纯因子组合无法直接投资，利用线性规划方法将纯因子组合合成为对几个常见因子的高暴露的因子组合。
3. 因子收益具有轮动特性，后续会探讨通过多种高暴露因子的配置来降低单个因子带来的波动性，提高组合夏普比率。
4. 因子择时是最难的部分，放在这一系列报告的最后，我们希望找到大概有效的模型来预测因子的涨跌方向。

风险提示：模型风险、过去收益不保证未来业绩、模拟回溯测试不代表真实交易

投资评级说明

行业评级	报告发布日后的 12 个月内行业股票指数的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	看　　好：	相对表现优于市场
	中　　性：	相对表现与市场持平
	看　　淡：	相对表现弱于市场
公司评级	报告发布日后的 12 个月内公司的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	买　　入：	相对大盘涨幅大于 10%
	增　　持：	相对大盘涨幅在 5%～10%之间
	中　　性：	相对大盘涨幅在-5%～5%之间
	减　　持：	相对大盘涨幅小于-5%
	无投资评级：	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

联系我们

上海

浦东新区世纪大道 1589 号长泰国际金融大厦 21 楼（200122）
电话：021-68751100　　传真：021-68751151

武汉

武汉市新华路特 8 号长江证券大厦 11 楼（430015）
传真：027-65799501

北京

西城区金融大街 17 号中国人寿中心 606 室（100032）
传真：021-68751791

深圳

深圳市福田区福华一路 6 号免税商务大厦 18 楼（518000）
传真：0755-82750808、0755-82724740

重要声明