
# Barra 因子模型截面回归求解

石川
量化交易 话题的优秀回答者



## 摘要

Barra 因子模型求解采用了带权重和约束条件的最小二乘回归。本文解释这个回归求解的数学过程，并通过简单实证说明求解的正确性。

## 1 引言

我似乎对 Barra 的因子模型过分钟爱了？

That was a joke.

钟爱谈不上，Barra 的模型在中国市场有多大作用、在什么使用情景下有用（因为没有可投资性，它无法直接用来选股）也仍在摸索中。但是，这么多年一代代模型的推出和改进代表着 Barra 自身的思考；一步步的构建一个逐步完善的多因子投资体系。这个框架足以引发我们的思考并学习。

之前我们分三篇文章介绍了 Barra 的因子模型，它们分别是《正确理解 Barra 的纯因子模型》、《协方差矩阵的 Newey-West 调整》、《Barra 因子模型中的风险调整》。

文就来介绍截面回归的求解过程。

在那之前，我们再次来重申截面回归所用到的暴露和收益率数据在时间上的关系。截面回归的输入显然对求解至关重要。根据 Barra Risk Model Handbook 的说明，因子暴露和因子收益率数据的正确解读为：

> *... the previous steps have defined the exposures of each asset to the factors **at the beginning** of every period in the estimation window. The factor excess returns **over the period** are then obtained via a cross-sectional regression of asset excess returns on their associated factor exposures ...*

这意味着，对于给定某一期截面数据（记为 T 期），在截面回归时采用期初的因子暴露取值（等价于 T - 1 期期末的因子暴露取值）和股票在 T 期内的收益率进行截面回归。在 USE4 模型中，因子收益率是日频的，因此截面回归也应该是日频的，所以按照上述说明，在 T - 1 日结束后更新因子的暴露，并利用 T 日的股票收益率和因子暴露做截面回归。

下面就来介绍截面回归的求解。

## 2 数学推导

在下文中，粗体小写字母表示向量、粗体大写字母表示矩阵。使用矩阵和向量，多因子模型可以表示为：

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

其中 $\mathbf{X}$ 是期初因子暴露矩阵。假设一共有 1 + P + Q = K 个因子（包括 1 个国家因子、P 个行业因子以及 Q 个风格因子），则 $\mathbf{X}$ 是一个 N × K 阶矩阵（其中 N 为股票个数）。在行文中，我会不厌其烦的写明矩阵的阶数，这有助于编程复现这个求解过程。具体的，

$$\mathbf{X} = \begin{bmatrix} 1 & X_1^{I_1} & \cdots & X_1^{I_P} & X_1^{S_1} & \cdots & X_1^{S_Q} \\ 1 & X_2^{I_1} & \cdots & X_2^{I_P} & X_2^{S_1} & \cdots & X_2^{S_Q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_N^{I_1} & \cdots & X_N^{I_P} & X_N^{S_1} & \cdots & X_N^{S_Q} \end{bmatrix}$$

$\mathbf{r}$ （N × 1 阶）是当期个股超额收益率向量； $\mathbf{f}$ （K × 1 阶）是待求的当期因子收益率向量，即 $\mathbf{f}=[f_C,f_{I_1},\ldots,f_{I_P},f_{S_1},\ldots,f_{S_Q}]^T$ ； $\mathbf{u}$ 为 N × 1 阶个股特异性收益率向量。

令 $\mathbf{\Omega}$ 为待求解的纯因子投资组合权重矩阵。它是一个 K × N 阶矩阵，它的每一行对应某个因子的纯因子投资组合中所有 N 支股票的权重。 $\mathbf{\Omega}$ 具体可以表达为：

$$\mathbf{\Omega}=\begin{bmatrix}\omega_{C1} & \omega_{C2} & \cdots & \omega_{CN}\\ \omega_{I_1 1} & \omega_{I_1 2} & \cdots & \omega_{I_1 N}\\ \vdots & \vdots & \ddots & \vdots\\ \omega_{I_P 1} & \omega_{I_P 2} & \cdots & \omega_{I_P N}\\ \omega_{S_1 1} & \omega_{S_1 2} & \cdots & \omega_{S_1 N}\\ \vdots & \vdots & \ddots & \vdots\\ \omega_{S_Q 1} & \omega_{S_Q 2} & \cdots & \omega_{S_Q N}\end{bmatrix}$$

为了求解 $\mathbf{\Omega}$ ，我们还需要用到另外两个矩阵，即回归权重矩阵 $\mathbf{V}$ 和约束矩阵 $\mathbf{R}$ 。约束矩阵对应的是下面这个因为国家和行业共线性造成的约束条件（不考虑这个约束的话，截面回归的求解不唯一）：

$$s_{I_1}f_{I_1}+s_{I_2}f_{I_2}+\cdots+s_{I_P}f_{I_P}=0$$

先来看看这个回归权重矩阵 $\mathbf{V}$ 是什么。

回归权重矩阵 $\mathbf{V}$ 是一个 N × N 阶对角阵，第 n 个对角元素代表着股票 n 的权重 $v_n$ 。 $v_n$ 和股票 n 的市值 $s_n$ （在本文第三节的实证中考虑流通市值）的平方根成正比，并满足权重值和为 1。因此可得：

$$v_n=\frac{\sqrt{s_n}}{\sum_{i=1}^{N}\sqrt{s_i}}$$

而 $\mathbf{V}$ 的表达式为：

$$\mathbf{V}=\begin{bmatrix}v_1 & 0 & \cdots & 0\\ 0 & v_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & v_N\end{bmatrix}$$

明：

> *Factor returns in USE4 are estimated using **weighted least-squares regression**, assuming that the variance of specific returns is **inversely proportional to the square root of total market capitalization**. This regression-weighting scheme reflects the empirical observation that the idiosyncratic risk of a stock decreases as the market capitalization of the firm increases.*

这段话的意思是，股票的特异性收益率的风险是不同的。然而，股票的特异性风险是不可测的。**经验表明，股票的特异性风险与它的总市值平方根成反比。**在构建纯因子投资组合时，应该加以考虑这一点。这在数学上可以通过在回归时，给股票加上基于特异性风险的回归权重，即带权重的最小二乘回归。

基于上述考虑，Menchero (2010) 指出**回归权重应该和市值的平方根成正比**：

> *In order to reduce estimation error in the factor returns, regression weights are used so that "noisy" stocks (i.e., those with high specific risk) are down-weighted. In practice, **regression weights are often taken as proportional to the square root of market capitalization**, although other weighting schemes are possible.*

这就是使用回归权重矩阵 $\mathbf{V}$ 的意义。

再来看看约束矩阵 $\mathbf{R}$ 。约束矩阵 $\mathbf{R}$ 是代表上文提到的约束条件（即所有行业的因子组合收益率线性相关）在求解时对行业因子收益率的限制条件。根据 Ruud (2000) 提出的理论，K 个因子收益率之间的约束条件（在此我们仅有一个约束条件）可以由以下等式表达：

$$\begin{bmatrix} f_C \\ f_{I_1} \\ \vdots \\ f_{I_P} \\ f_{S_1} \\ \vdots \\ f_{S_Q} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{s_{I_1}}{s_{I_P}} & -\frac{s_{I_2}}{s_{I_P}} & \cdots & -\frac{s_{I_{P-1}}}{s_{I_P}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} f_C \\ f_{I_1} \\ \vdots \\ f_{I_{P-1}} \\ f_{S_1} \\ \vdots \\ f_{S_Q} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

上式中，等号右边的矩阵就是约束矩阵 $\mathbf{R}$ ，它是一个 K × K - 1 阶矩阵，这是因为所有 K 个收益率变量之间有一个约束条件，因此它们的自由度为 K - 1。不失一般性，在构造 $\mathbf{R}$ 时，我们将行业 P 的因子组合收益率 $f_{I_P}$ 用其他行业的收益率的线性组合来表达。

式，感兴趣的朋友可进一步参考。

$$\mathbf{\Omega} = \mathbf{R}\left(\mathbf{R}^T\mathbf{X}^T\mathbf{V}\mathbf{X}\mathbf{R}\right)^{-1}\mathbf{R}^T\mathbf{X}^T\mathbf{V}$$

其中 -1 表示矩阵的逆矩阵。

由前文可知， $\mathbf{\Omega}$ 的每一行是一个 1 × N 向量；它就代表着第 k 个因子的纯因子投资组合中所有股票的权重。得到 $\mathbf{\Omega}$ 之后，可通过下式计算出所有因子在当期的因子收益率：

$$f_k = \sum_{n=1}^{N} \omega_{kn} r_n$$

以上就是 Barra 因子模型截面回归的求解。

## 3 简单实证

本节对上述求解过程做一个简单的实证，最主要的目的是检验 $\mathbf{\Omega}$ 求解公式是否正确。此外，通过构建的纯因子组合，我们也可以验证在《正确理解 Barra 的纯因子模型》谈到的三类因子（国家因子、行业因子、风格因子）的特性是否成立。

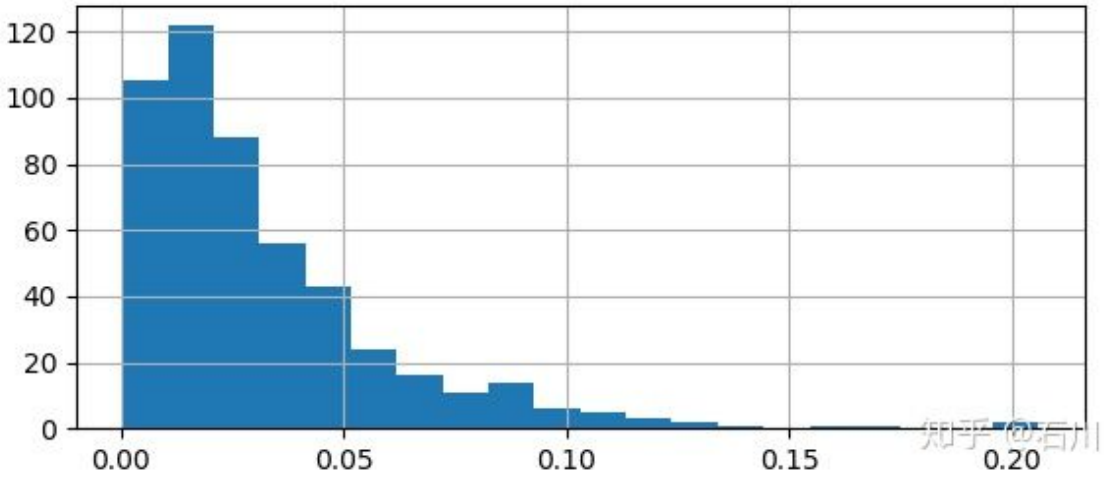
我们选用中证 500 指数的成分股在 2016 年 5 月 31 日的截面数据和这些股票在 2016 年 6 月 1 日的收益率作为回归的输入。除国家因子外，行业因子考虑了 27 个申万行业，并考虑以下 11 种风格因子（再次重申，本实证的目的是为了验证 $\mathbf{\Omega}$ 的求解，因此对于如何构建这些风格因子不做描述）：GROWTH，EP，BP，LIQ，SCALE，SCALENL，BETA，RESIDSTD，MOM，REV 以及 LIB2ASSET。

根据上一节的求解方法，得到这 39 个因子（1 个国家 + 27 个行业 + 11 个风格）的投资组合在 2016 年 6 月 1 日的因子收益率如下。

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 801010 | -0.053 | 801150 | -0.294 | 801740 | 0.269 | LIQ | 0.117 |
| 801020 | 0.419 | 801160 | 0.105 | 801750 | -0.666 | SCALE | 0.137 |
| 801030 | -0.148 | 801170 | 0.490 | 801760 | -0.649 | SCALENL | -0.062 |
| 801040 | -0.135 | 801180 | 0.105 | 801770 | -0.521 | BETA | 0.318 |
| 801050 | 0.809 | 801200 | 0.275 | 801790 | -1.331 | RESIDSTD | -0.116 |
| 801080 | 0.662 | 801210 | -0.041 | 801880 | 0.595 | MOM | 0.070 |
| 801110 | -0.333 | 801230 | -0.168 | 801890 | -0.252 | REV | 0.185 |
| 801120 | -0.438 | 801710 | -0.300 | GROWTH | 0.067 | LIB2ASSET | -0.012 |
| 801130 | -0.295 | 801720 | 0.114 | EP | -0.153 | | |



观察不同因子的收益率可知，它们的数量级大致相当。结果显示，国家因子的收益率为 0.429%，当日中证 500 的收益率是 0.44%。这两个数字满足《正确理解 Barra 的纯因子模型》提到的**国家因子组合近似的等于市场组合。**比较国家因子组合中个股的权重和中证 500 指数中个股权重，权重差别的均值为 3.2%，权重差别的分布如下图所示（提醒，这仅仅是当期的结果）：




再来看看行业因子收益率。行业因子投资 100% 做多该行业，100% 做空市场，因此它表示行业相对市场的超额收益。然而，**行业因子的投资组合收益率并不等于申万这些行业指数和中证 500 指数收益率的差值。**这是因为行业纯因子投资组合对所有风格因子的暴露为零，而申万行业指数无法满足这个限制，所以二者中个股的权重是不同的，因此它们的收益率也会有出入。

使用因子投资组合的权重矩阵 $\Omega$ （K × N 阶）乘以当期的因子暴露矩阵 $\mathbf{X}$ （N × K 阶），就得到一个 K × K 阶的矩阵，**该矩阵的每一行都是其对应的因子投资组合在其他因子上的暴露。**检查这个矩阵的结果可以帮助我们检验 Barra 纯因子组合的性质。下图就是 $\Omega$ 乘以 $\mathbf{X}$ 得到的矩阵。

图中（看的不是太清楚，我尽量解释），排除列名所在的最上面一行不考虑，第一行是国家因子；蓝色长方形框出来的部分是行业因子；红色长方形框出来的部分是风格因子。白色的单元格表示的数字是 0 —— 因此我们很容易看出，**国家因子和任一个行业因子组合在所有风格因子上的暴露都是 0；而任何一个风格因子纯因子组合在国家、所有行业以及其他风格因子上的暴露也都是 0。**

下面再来具体看看不是零的单元格（我们从图中分别针对国家和行业因子、以及风格因子截取一部分解释）。

下图显示了该矩阵左上角的部分，包括国家因子和几个行业因子。**第一行**（除了列名外）为国家因子，每一列对应的单元格中的数字是国家因子在相应因子上的暴露。可见，国家因子对自身的暴露为 1，因为它近似的等于市场，而市场包含了所有行业，因此它在每个行业上都有一定程度的暴露（比如，国家因子在 801010 行业上的暴露为 0.033，在 801020 行业上的暴露为 0.020）。

| | COUNTRY | IND_801010 | IND_801020 | IND_801030 | IND_801040 | IND_801050 | IND_801080 | IND_801110 | IND_801120 | IND_801130 |
|---|---|---|---|---|---|---|---|---|---|---|
| COUNTRY | 1.000 | 0.033 | 0.020 | 0.059 | 0.010 | 0.060 | 0.070 | 0.019 | 0.035 | 0.015 |
| IND_801010 | 0.000 | 0.967 | -0.020 | -0.058 | -0.010 | -0.060 | -0.069 | -0.019 | -0.035 | -0.015 |
| IND_801020 | 0.000 | -0.033 | 0.980 | -0.058 | -0.010 | -0.060 | -0.069 | -0.019 | -0.035 | -0.015 |
| IND_801030 | 0.000 | -0.033 | -0.020 | 0.941 | -0.010 | -0.060 | -0.070 | -0.020 | -0.035 | -0.015 |
| IND_801040 | 0.000 | -0.033 | -0.020 | -0.058 | 0.990 | -0.060 | -0.069 | -0.019 | -0.035 | -0.015 |
| IND_801050 | 0.000 | -0.033 | -0.020 | -0.059 | -0.010 | 0.940 | -0.070 | -0.019 | -0.035 | -0.015 |
| IND_801080 | 0.000 | -0.034 | -0.020 | -0.059 | -0.010 | -0.061 | 0.930 | -0.020 | -0.036 | -0.015 |
| IND_801110 | 0.000 | -0.033 | -0.020 | -0.059 | -0.010 | -0.060 | -0.070 | 0.980 | -0.035 | -0.015 |
| IND_801120 | 0.000 | -0.034 | -0.020 | -0.059 | -0.010 | -0.061 | -0.070 | -0.020 | -0.036 | -0.015 |
| IND_801130 | 0.000 | -0.033 | -0.020 | -0.059 | -0.010 | -0.060 | -0.070 | -0.020 | -0.035 | 0.985 |

再来看看行业因子。以 801010 这个行业为例（即排除列名外的第二行）。前文反复强调过，**行业的纯因子组合等价于 100% 做多该行业，100% 做空国家因子。**因此，对于 801010 这个行业来说，它在所有行业（包括它自己）上的暴露应该是行向量 [1, 0, 0, ... ,0]（第一个 1 代表对它自己的 100% 多头）和国家因子在这些行业上的暴露 —— 即向量 [0.033, 0.020, 0.059, ... ] —— 的**差**（做差就相当于做空国家因子）：

$$[1,0,0,\ldots,0]-[0.033,0.020,0.059,\ldots]=[0.967,-0.020,-0.059,\ldots]$$

而如果我们考察 801010 所在的第二行的数值，则上面计算得到的这个向量 [0.967, -0.020, -0.059, ...]（忽略计算误差）中的数值正是对应 801010 在不同行业（包括它自己）上的暴露！

扩展一下上述结论，**对于给定的行业，它在其他行业的暴露等于向量** [0, 0, ..., 0, 1, 0, ..., 0] **—— 假设该行业在所有行业中的位置为 p，则这个向量中的位置 p 为 1，其他位置为 0 —— 与国家因子在这些行业上的暴露向量之差。**这也解释了为什么在上图中我们观察到，任何其他行业在行业上的暴露都相等（在误差范围内），且等于国家因子在行业 p 上暴露加个负号。

| | GROWTH | EP | BP | LIQ | SCALE | SCALE_NL | BETA | RESID_STD | MOM | REV | LIBILITYTOA |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GROWTH | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| EP | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| BP | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| LIQ | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SCALE | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| SCALE_NL | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| BETA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| RESID_STD | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| MOM | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 |
| REV | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 |
| LIBILITYTOA | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |

以上我们从多个角度检验了截面回归的求解结果。得到的数据和 Barra 对于纯因子组合的构建相符合，这说明了 $\Omega$ 求解过程的正确性。

## 4 结语

本文介绍了截面回归的求解。结合之前的几篇文章，对 Barra 模型的介绍基本比较完整了。然而，我们对它的思考和实践应远不止于此。

在国内的一些优秀券商金工报告中，已经开始使用最优化的思想，加上各种可投资性的限制，利用 Barra 的这套纯因子模型来构建投资组合了。这无疑是一种很好的尝试。

另外，有朋友反馈说，使用了 Newey-West 调整后，协方差矩阵的 bias statistic 反而变差。还有其他各种各样的问题。在我自己的实践中，尚未遇到所有小伙伴们遇到的问题，因此暂时无法对所有问题都给出靠谱的评论。

**无论我们是否使用 Barra 模型，最重要的是理解它内在的含义和它使用的各种统计手段。切莫把 Barra 当作多因子投资的“标准姿势”，误以为把它套用到 A 股数据上就会产生什么神奇的化学反应。**那无疑是本末倒置。正确的做法是理解其含义，并针对 A 股数据的特点有的放矢、灵活应用。让我们在践行多因子选股的道路上，为找到收益风险比更佳的投资组合而努力。

### 参考文献


- Menchero, J. (2010). *Characteristics of Factor Portfolios*. MSCI Barra Research Notes.
- Menchero, J., D. J. Orr, and J. Wang (2011). *The Barra US Equity Model (USE4)*. MSCI Barra Research Notes.
- Menchero, J. and J.-H. Lee (2015). Efficiently combining multiple sources of alpha. *Journal of Investment Management*, Vol. 13(4), 71 - 86.
- Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. New York, NY: Oxford University Press.
- Barra Risk Model Handbook (2007). MSCI.

编辑于 2019-07-03

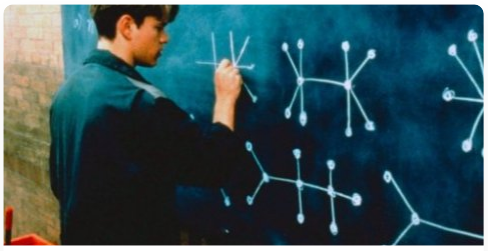
BARRA模型 多因子模型 量化交易



## 文章被以下专栏收录



## 推荐阅读



**多因子量化选股模型的筛选和评价：打分法与回归法**

JD-Quant

**Black-Litterman模型**

Black-Litterman模型是基于MPT基础上的资产配置理论。BL模型在隐含市场收益率和分析师主观预测信息的基础上，成功解决了MPT模型中假设条件不成立，参数敏感等问题。 我们先简单回顾一下Mark...

llleafer

**为什么不能阵来估计**

窦福成

## 还没有评论

因作者设置，评论已关闭