BARRA 风格因子及协方差矩阵的估计

章瑞麒 (zhangruiqi_hep@163.com)

1 代码说明

代码由 C++ 完成，依赖 boost-1.70 以及 gsl-2.5。将输入数据放入 data 文件夹中，修改代码中的输入文件路径然后编译 Makefile（相关库的路径需要作对应修改）生成可执行文件即可运行。代码内容包括了数据的读入，因子载荷，因子收益率的计算，数据清洗，因子收益率协方差矩阵的估计及其逆矩阵的计算。可以在 main.cpp 中选择将相关指标输出到 csv 文件中，会保存在 output 文件夹中。

2 简介

BARRA 中定义了了十个不同的风格因子以及国家因子，行业因子，利用这些因子来解释股票的收益来源。目前可用的数据可以计算出其中的六个风格因子，分别是 Size, Beta, Momentum, Residual Volatility, Non-linear Size 和 Liquidity。在计算中使用了 A 股中的 1000 支股票在 2007∼2013 的数据。在逐日计算得到了这六个风格因子的载荷后，与国家因子一起对下一个交易日的股票收益率作截面回归，得到因子的收益率序列后分别用样本估计和 Newey-West 方法估计了因子的协方差矩阵并利用 LU 分解的方法计算了协方差矩阵的逆矩阵。计算中得到的结果均为逐日计算的日频结果。

3 数据整理

由于不同的风格因子的量级不同，在将它们放到一起分析之前需要对它们作标准化处理，另外数据中的极值与缺失值也需要妥善处理。对于极值使用中位数去极值法：

$$\tilde{x} = \left\{ \begin{array}{ll} x_M + n * D_{MAD}, & x_j > x_M + n * D_{MAD} \\ x_M - n * D_{MAD}, & x_j < x_M - n * D_{MAD} \\ x_i, & else \end{array} \right\}, \tag{1}$$

式中的 x_M 是序列 x 的中位数，D_{MAD} 是序列 $|x - x_M|$ 的中位数，这里选取的 n 为 3.5。为了将不同因子放在相同的量级里进行分析，利用市值加权平均和算术标准差进行标准化处理：$x = \frac{x_i - \mu}{\sigma}$，其中 μ 为市值加权平均值，σ 为算术标准差。由于每日的数据会有存在缺失值的问题，对于每个交易日内的股票数据，如果缺失率低于 30%，利用均值进行填充，如果缺失率过高则放弃这个交易日的计算。

4 因子载荷及因子收益

BARRA 中对于 Size, Beta, Momentum, Residual Volatility, Non-linear Size 和 Liquidity 这六个风格因子的具体定义如下：

- Size 是股票市值的对数。

- Beta 是利用市值加权的市场收益率对过去 252 个交易日的股票净收益作 WLS 得到的回归系数，回归公式为：

$$r_t - r_{ft} = \alpha + \beta * R_t + e_t, \tag{2}$$

其中 r_t 为股票收益率，r_{ft} 为无风险收益率，R_t 为市场收益，e_t 为股票的特异收益率。

- Momentum 是过去的净收益的对数的累加，定义为：

$$\sum_{t=L}^{T+L} w_t[ln(1+r_t) - ln(1+r_{ft})], \tag{3}$$

其中 T 为 504，L 为 21，w_t 是 half life 为 126 的权重。

- Residual Volatility 是由三个因子组合而成：0.74 * DASTD + 0.16 * CMRA + 0.10 * HSIGMA。DASTD 是过去 252 个交易日的日净收益率的波动率；CMRA 定义为：

$$\begin{aligned} &ln(1+Z_{max}) - ln(1+Z_{min}), \\ &Z(T) = \sum_{\tau=1}^{T} [ln(1+r_\tau) - ln(1+r_{f\tau})], \end{aligned} \tag{4}$$

其中 r_τ 与 $r_{f\tau}$ 表示的是月度的复合收益率。HSIGMA 是计算 Beta 时所得的残差收益率 e_t 的标准差。

- Non-linear Size 是 Size 因子载荷标准化之后取立方之后再对 Size 作回归后得到的残差序列。

- Liquidity 是由过去一个月，一个季度以及一年的换手率组合而成：0.35 * STOM + 0.35 * STOQ + 0.30 * STOA。

$$\begin{aligned} STOM &= ln(\sum_{t=1}^{21} \frac{V_t}{S_t}) \\ STOQ &= ln[\frac{1}{T}\sum_{\tau=1}^{T} exp(STOM_\tau)] \quad T = 3 \\ STOA &= ln[\frac{1}{T}\sum_{\tau=1}^{T} exp(STOM_\tau)] \quad T = 12, \end{aligned} \tag{5}$$

式中的 V 和 S 表示交易量和流通股。

将市场数据经过去极值，标准化处理之后，按照上述的定义计算出每个交易日每支股票的因子载荷，对于某个交易日内缺失的因子载荷，使用市值加权平均值进行填充。在计算得到每日的因子载荷后，对因子载荷再做一次标准化处理。之后利用标准化的六个风格因子载荷以及国家因子一起对下一个交易日的股票收益率作 WLS 拟合，权重为市值的平方根，如式 6所示。

$$r_n = f_c + \sum_s X_{ns} f_s + \mu_n, \tag{6}$$

式中的 r_n 代表第 n 支股票次一交易日的净收益，f_c 代表国家因子收益率也即市场收益率，X_{ns} 代表第 n 支股票在第 s 个风格因子上的载荷，f_s 代表风格因子收益率。回归得到的回归系数整理为一个长度为六的因子收益率向量记作 F_t，逐日重复这一过程便可以得到 T 期的因子收益率数据以便进行更进一步的分析。六个风格因子的复合收益率如图1所示。

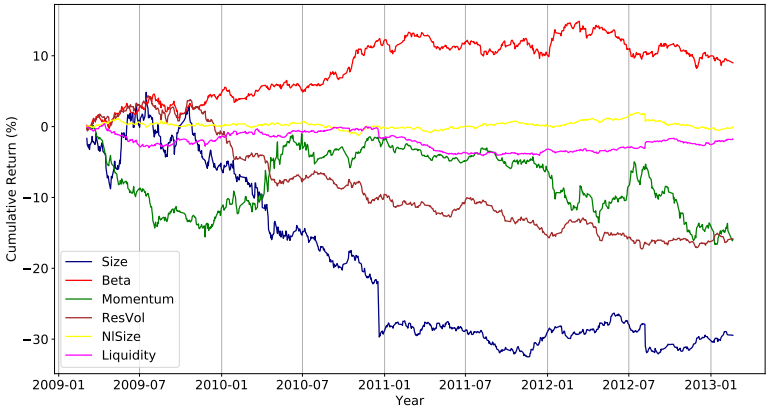

图 1: BARRA 风格因子 2009~2013 的复合收益率。

5 协方差矩阵

由于股票的数目过于庞大，直接利用股票收益率计算其协方差矩阵以进行风险分析是不可行的，但是股票的收益率协方差矩阵也可以通过因子收益率的协方差矩阵和股票的因子载荷矩阵来计算。

在经典的统计教材中，通常用样本协方差矩阵作为对真实协方差矩阵的估计，样本协方差矩阵的计算公式为：

$$\tilde{\Sigma} = \frac{1}{T-1}(\sum_{t=1}^{T} F_t F_t^T) \tag{7}$$

当数据量 T 足够大时，公式中的 $\frac{1}{T-1}$ 也可以用 $\frac{1}{T}$ 代替。使用公式 7来估计的协方差矩阵如下所示：

$$\begin{bmatrix}
2.538e-05 & -4.562e-06 & -9.169e-06 & -3.687e-06 & -4.911e-06 & -3.055e-06 \\
-4.562e-06 & 4.718e-06 & -2.017e-06 & 1.395e-06 & 4.288e-07 & 3.656e-08 \\
-9.169e-06 & -2.017e-06 & 1.535e-05 & -4.121e-06 & 8.681e-08 & -1.264e-07 \\
-3.687e-06 & 1.395e-06 & -4.121e-06 & 5.447e-06 & 7.597e-07 & 2.056e-07 \\
-4.911e-06 & 4.288e-07 & 8.681e-08 & 7.597e-07 & 2.500e-06 & 1.135e-06 \\
-3.055e-06 & 3.656e-08 & -1.264e-07 & 2.056e-07 & 1.135e-06 & 1.804e-06
\end{bmatrix}$$

这一方法估计协方差矩阵的优点在于方法简单，矩阵半正定等性质得以保证，但是这一方法并没有考虑到因子收益率在时序上的相关性，因而并不能保证当数据量足够大时估计的协方差矩阵会收敛到真实协方差矩阵，会带来一定的偏

差。BARRA 中所采用的 Newey-west 方法便是针对因子收益率的时序自相关的修正，通过对自相关的刻画来保证估计的准确性。Newey-west 的具体做法为用 MA(q) 来刻画因子收益率序列的相关性，并引入 Bartlett 权重系数 $1-\frac{i}{1+q}$ 来保证保证矩阵的半正定性：

$$
\begin{aligned}
V_f &= \Gamma_0 + \sum_{i=1}^{q} \omega_i (\Gamma_i + \Gamma_i^T) \\
\Gamma_i &= \frac{1}{T} \sum_{t=1}^{T-i} F_t F_{t+i}^T \\
\omega_i &= 1 - \frac{i}{1+q}
\end{aligned}
\tag{8}
$$

参考 BARRA 中的 Newey-west lags，这里采用的 q 为 5，由此按照 Newey-west 方法得到的协方差矩阵为：

$$
\begin{bmatrix}
2.524e-05 & -5.129e-06 & -8.943e-06 & -3.715e-06 & -4.574e-06 & -2.880e-06 \\
-4.395e-06 & 5.140e-06 & -2.448e-06 & 1.386e-06 & 3.908e-07 & -7.302e-08 \\
-8.763e-06 & -2.811e-06 & 1.680e-05 & -5.018e-06 & -6.364e-08 & -1.468e-07 \\
-3.062e-06 & 1.744e-06 & -4.689e-06 & 6.042e-06 & 2.783e-07 & -3.126e-07 \\
-5.399e-06 & 7.182e-07 & -2.000e-07 & 9.087e-07 & 2.623e-06 & 1.348e-06 \\
-3.622e-06 & 3.412e-07 & -5.248e-07 & 3.967e-07 & 1.345e-06 & 2.063e-06
\end{bmatrix}
$$

6 协方差矩阵的逆

当矩阵是三角阵或对角阵时，无论是求解方程组或是对矩阵求逆都会变得十分方便，因而当面对复杂的矩阵时，矩阵分解是非常常用的一种方法。因子收益率的协方差矩阵是实对称的半正定矩阵，当因子之间的共线性不强时，它必定正定，因此可以考虑对它作 LU 分解，将矩阵分解为一个上三角阵和一个下三角阵从而得到矩阵的逆。LU 分解可以通过高斯消元的方法来做，下三角阵实际就是消元矩阵的逆。代码中使用 boost::numeric::ublas 来完成。

$$
A = LU = (E_n E_{n-1} \cdots E_2 E_1)^{-1} U \tag{9}
$$

除了 boost 中自带的函数，我也自己实现了一遍基于 LU 分解的矩阵求逆，首先将矩阵分解为上下三角阵，由于 LU 分解可能存在多个解，因此限定下三角阵的对角元为 1 以保证解的唯一，求解方程组可得上下三角阵为 10

$$
\begin{aligned}
U_{i,k} &= A_{i,k} - \Sigma_{j=0}^{i} L_{i,j} * U_{j,k} \quad i = 0, 1, 2..., n-1; k = i, i+1...n-1; \\
L_{j,k} &= (A_{i,k} - \Sigma_{j=0}^{i} L_{i,j} * U_{j,k}) / U_{k,k} \quad i = k+1, k+2...n-1;
\end{aligned}
\tag{10}
$$

在对矩阵进行 LU 分解之后，分别对下三角阵和上三角阵求逆，之后将逆矩阵相乘即得到原矩阵的逆矩阵。通过我自己的代码所计算的逆矩阵与通过 boost::numeric::ublas 求得的逆矩阵是完全一致的。

由于在实际计算中，常常会出现矩阵接近奇异的情况，为了解决这个问题，我令 $A = A + \lambda * I$，其中 I 为单位矩阵，λ 根据 A 中的取值范围取一个小数如 1e-6，如此便可以顺利计算出逆矩阵。使用 LU 分解得到的因子协方差矩阵的逆

矩阵为：

$$\begin{bmatrix} 8.366e+06 & 8.339e+06 & 8.339e+06 & 8.331e+06 & 8.349e+06 & 8.288e+06 \\ 8.331e+06 & 8.524e+06 & 8.330e+06 & 8.27143e+06 & 8.266e+06 & 8.291e+06 \\ 8.33e+06 & 8.333e+06 & 8.391e+06 & 8.364e+06 & 8.295e+06 & 8.291e+06 \\ 8.322e+06 & 8.253e+06 & 8.354e+06 & 8.513e+06 & 8.254e+06 & 8.315e+06 \\ 8.361e+06 & 8.287e+06 & 8.299e+06 & 8.244e+06 & 8.943e+06 & 7.879e+06 \\ 8.295e+06 & 8.278e+06 & 8.297e+06 & 8.290e+06 & 7.904e+06 & 8.948e+06 \end{bmatrix}$$

协方差矩阵与其逆矩阵的乘积为：

$$\begin{bmatrix} 0.832672 & -0.166781 & -0.166798 & -0.166614 & -0.166997 & -0.165772 \\ -0.166622 & 0.829516 & -0.166613 & -0.165429 & -0.165325 & -0.165817 \\ -0.166738 & -0.166678 & 0.832163 & -0.167294 & -0.165917 & -0.165828 \\ -0.166455 & -0.165071 & -0.167098 & 0.829732 & -0.165087 & -0.166311 \\ -0.167228 & -0.165715 & -0.165994 & -0.164886 & 0.821128 & -0.157594 \\ -0.165919 & -0.165561 & -0.165951 & -0.1658 & -0.158092 & 0.821037 \end{bmatrix}$$

从协方差矩阵的逆矩阵来看，仅经过 Newey-west 调整之后得到的协方差矩阵是 ill-conditioned，若要得到对真实协方差矩阵的准确估计，还需要进一步将 BARRA 中所提到的修正一一加上。

7 小结

按照 BARRA CNE5 文档中描述的风格因子，利用 A 股中 1000 支股票在 2007∼2013 的市场数据，逐日计算了六个风格因子的因子收益率，并用样本协方差方法和 Newey-west 方法估计因子收益率的协方差矩阵，最后使用 LU 分解的方法求得协方差矩阵的逆矩阵。

由于时间仓促的关系，未能完全按照 BARRA 中的描述来计算因子和协方差矩阵。比如在 BARRA 中对于因子收益率的缺失值是通过市值加权回归来填充的，协方差矩阵也是由相关系数矩阵得来，对于协方差矩阵的修正也远不止 Newey-west 这一项。对于多因子模型和 BARRA 的风险模型还有许多需要学习的地方。

逆矩阵的求解也远不止 LU 分解这一种方法，LU 分解的优点是快速，计算消耗小但是代价便是准确性与稳定性，还有许多其他的矩阵求逆的方法比如基于 Cholesky 分解或 SVD 等。