中信期货
CITIC Futures

中信期货研究|金融工程专题报告

2023-01-31

截面回归与因子正交的二重奏

——组合优化专题报告（一）

投资咨询业务资格：
证监许可【2012】669 号

金融工程研究团队

研究员：
周通
021-80401733
从业资格号 F3078183
投资咨询号 Z0018055

期货多因子系列研究报告

专题报告五：不同频率视角下的选期因子——20221222
专题报告六：基于深度学习的期货组合优化——20221229

报告要点

本报告处理“因子正交”的问题。该专题将以综述的形式讨论了市面上常见的 4 种正交化方法、阐明了正交实践中遇到的问题与解决办法、设计了 1 种针对于底层资产池为商品期货组合的品种权重配置模式、配比了 1 种因子收益率估计的优化方式并进行回测总结。该报告对应第 1 种配置模式——作用于动态商品期货池、以截面回归方式来构建组合权重，是本专题的第一篇。

摘要：

本篇报告中，我们以综述的形式讨论了市面上常见的 4 种正交化方法、阐明了正交实践中遇到的问题与解决办法、设计了 1 种商品期货组合的配置模式、配比了 1 种因子收益率估计的优化方式。

回测效果来看，可作出以下结论：

（1）本报告提出的新配置方式——“作用于动态商品期货池、以截面回归方式来构建组合权重”的配置效果较好，其在本文入选的 6 个量价因子以—“未正交搭配简单移动平均 SMA（$n = 20$）对因子收益率估计”的优化方式回测时，取得了 11%左右的年化收益，夏普为 0.85，最大回撤为 0.2（见图表 11）；

（2）3 种正交化方式是把双刃剑，其实质来源于共线性的双刃性质。在本文入选的 6 个量价类因子的回测中，使用未正交因子的回测净值——无论是对因子收益率未优化时回测得到的年化收益 3.49%（见图表 7）、还是使用简单移动平均 SMA（$n = 20$）优化后回测得到的年化收益 10.95%（见图表 11），均优于相应的、基于正交化后因子的回测净值表现；

（3）对因子收益率估计的优化方式（简单移动平均）无论是针对非正交还是正交化后的因子均能提升策略有效性。

风险提示:本报告中所涉及的资产配比和模型应用仅为回溯举例，并不构成推荐建议。

目 录

图表目录

一、整体框架

共线性是指标的资产的因子之间存在某种程度的线性关系。共线性的存在不会影响回归系数的唯一性和无偏性，但它会使得回归系数估计值的方差变大，并且造成回归系数的置信区间变得很宽。它的缺点是使得估计量的精准度和 T 值都收到影响，从而导致一些因子通不过假设检验，尽管这些因子对收益率存在显著影响，即具有良好的解释力。

常见的判断共线性的方法涉及两个关键数：第一个是条件数，第二个是方差膨胀因子（variation inflation factor,简记作 vif)。具体而言：条件数是因子相关系数矩阵的最大特征值与最小特征值的比值，我们认为若条件数小于 100，则共线性程度较小；若条件数大于 1000，则存在共线性。而方差膨胀因子则是因子之间存在多重共线性是的方差与不存在多重共线性是的方差之比，我们认为当方差膨胀因子介于 0 到 10 之间，则不存在多重共线性；若介于 10 到 100 并包含为 10 的情况，则存在较强的多重共线性；当其值大于等于 100 时，则存在严重的多重共线性。

采用方差膨胀因子法检测共线性时，对于已经中心化、标准化的因子矩阵 F，计算其相关系数矩阵C，求得其逆为 $I=(i_{jk})\triangleq C^{-1}$，则其主对角元素 i_{jj} 为自变量 F_j 的方差膨胀因子，记 C_j^2 为自变量 f_j 对其余 $N-1$ 个因子回归的拟合度，可以证明 $vif_j=i_{jj}=\frac{1}{1-C_j^2}$，其相应的经济含义为某个因子 f_j 的方差膨胀因子越大，其与另外$N-1$ 因子的越线性相关，其解释力也就越容易被其他因子所替代。

共线性会影响回归效果，那么直接的方法就是把共线性显著的变量进行剔除，包括方差膨胀因子法、逐步 F 检验-t 检验、主成分回归等；但其缺点也是明显的，基于这些方法我们最后只能保留解释能力最强的几个因子，而有些因子则被剔除；此外也有相关研究表明这样的处理导致拟合度 C_j^2 偏低，出现“欠拟合”。

出于上述考量，我们本文将考虑另一种方式——因子的正交化。本文提供了市面上常见的 4 种因子正交化方法的综述；在此基础上，设计了 1 种期货组合的品种权重配置方式。

图表1：商品期货品种选择

类别	具体品种
黑色类	玻璃（FG）、热轧卷板（HC）、铁矿石（I）、焦炭（J）、焦煤（JM）、螺纹钢（RB）、不锈钢（SS）、纯碱（SA）
有色类	沪铝（AL）、沪铜（CU）、沪锌（ZN）、镍（NI）、锡（SN）
能源类	石油沥青（BU）、原油（SC）、燃料油（FU）、液化石油气（PG）、低硫燃料油（LU）
化工类	聚乙烯（L）、甲醇（MA）、聚丙烯（PP）、聚氯乙烯（V）、PTA(TA)、乙二醇（EG）、尿素（UR）、苯乙烯（EB）
软商品类	棉花（CF）、天然橡胶（RU）、白糖（SR）、纸浆（SP）、短纤(PF)
农产品类	豆一（A）、玉米（C）、玉米淀粉（CS）、鸡蛋（JD）、豆粕（M）、菜油（OI）、棕榈油（P）、菜粕（RM）、豆油（Y）、生猪（LH）

资料来源：中信期货研究所

这里的因子正交化方法具体为回归取残差、Schmidt 正交化、规范(Canonical)正交化、对称（Symmetric）正交化；而期货组合的品种权重配置方式为针对于动态商品期货池、以截面回归方式来构建组合权重,时间跨度为前后 5 个交易日。

本篇报告聚焦于量价类因子，具体使用到的 6 个因子为：分别是“3 日动量”（'mom_d3'），“243 日动量”（'mom_d243'），“10 日动量”（'mom_d10'），“243 日最小二乘回归”（'ols_d243'），“5 日量价相关性”（'cv_d5'）和“ 61 日振幅”（'amp_d61_g4'）。其相应的因子构造逻辑可参见本团队之前的相关研报。

二、正交化的方法与数学理论

我们这里提供市面上常见的正交化方法、相关的数学理论、在实际应用中遇到的问题和解决办法。

（一）问题背景

实际中拿到的因子往往具有共线性，导致因子重复暴露，我们希望通过正交化消除因子之间的相关性，并保持因子对于收益的解释度不变。

因子正交化有多种方式,目前市面上常用的有如下 4 种:回归取残差、Schmidt 正交化、规范（Canonical）正交化、对称（Symmetric）正交化。其中，我们对第 1 种可以给出证明，其实质上与 Schmidt 正交化是一致；而后 3 种都是通过矩阵乘法（又称线性变换、因子旋转）的方式来消除因子间的相关性。

（二）数学理论

上述提到的 3 种正交化方式共性与个性并存，我们这里先从其依据的线性代

数理论基础出发，了解一下其背后的简单逻辑。

1. 共性

设定某个时间截面上，市场上的期货数量为 N 个，入选的因子数量为 M 个，则该截面上的因子矩阵可以表示为

$$F_{N\times M} \triangleq [f^1, f^2, \dots, f^M] \triangleq \begin{pmatrix} f_1^1 & \cdots & f_1^M \\ \vdots & \ddots & \vdots \\ f_N^1 & \cdots & f_N^M \end{pmatrix},$$

其中，第 k 个因子在考虑的 N 个期货上的暴露值（简而言之即因子值）为 $f^i = (f_1^i, f_2^i, \dots, f_N^i)'$；我们的目标是找到一个“过渡矩阵” $S_{M\times M}$ 后再进行计算 $F_{N\times M}^{\perp} \triangleq F_{N\times M}\ S_{M\times M}$，而这里通过矩阵乘法得到的 $F_{N\times M}^{\perp}$ 就是正交化后的因子矩阵，也即一个正交阵。

因此我们可以发现，正交化的实质是对因子进行旋转，让旋转后的因子满足两两正交且整体方差不变（正交表明线性相关性为 0，方差可以刻画因子对于收益的解释度），即旋转后也为正交阵。用公式表达即为

$$\forall\, i,j \in \{1,2,\dots,M\}, i \neq j,\ (\tilde{f}^i)'\tilde{f}^j = 0,\ var(\tilde{f}^i) = var(\tilde{f}^j)\ ,$$

其中，$\tilde{f}^i$ 表示正交化后因子矩阵的列向量，即 $F_{N\times M}^{\perp} = [\tilde{f}^1, \tilde{f}^2, \dots, \tilde{f}^M]$。因子的旋转过程是通过过渡矩阵定义。

在此过程中，涉及到的线性代数知识包括：（L1）实对称矩阵一定可以正交对角化；（L2）实对称矩阵的特征值必为实数；（L3）属于实对称矩阵的不同特征值的特征向量是正交的；（L4）任何实对称矩阵都可以通过正交变换将其对角化。

具体的步骤如下：

第一步：对于两个 N 维随机向量 X 和 Y（其元素分别记作 X_i 和 Y_i，$i \in \{1, \dots, N\}$），其协方差矩阵 $\Sigma_{M\times M}$ 定义为

$$\Sigma_{M\times M} \triangleq cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - E(X))(Y_i - E(Y))}{N-1} = \frac{1}{N-1}(X. - E(X))'(Y. - E(Y))\ ;$$

应当注意的是，这里分母中无偏估计选取的 $N-1$ 是区别于最大似然估计选取的 N；此外在实际处理中，我们对每个因子都将进行针对截面的 z-score 标准化，满足均值为 0 和模为 1。标准化的意义在于，正交与不相关的概念本来是不等价的，正交不一定不相关，但加上 z-score 标准化后，正交等价于线性相关系数为 0。

对于任意的 $i \in \{1, \dots, M\}$，取 $X = Y = f^i$，则计算 $F_{N\times M}$ 的协方差矩阵

即为 $\Sigma_{M\times M} \triangleq cov(F_{N\times M})$ ，我们将 $(N-1)\Sigma_{M\times M}$ 整体记作

$$P_{M\times M} \triangleq F_{N\times M}{}'F_{N\times M} = \begin{pmatrix} f^{1\prime}f^{1} & \cdots & f^{1\prime}f^{M} \\ \vdots & \ddots & \vdots \\ f^{M\prime}f^{1} & \cdots & f^{M\prime}f^{M} \end{pmatrix},$$

此为对称矩阵；

第二步：根据正交阵定义，$I_{M\times M} = F^{\perp}_{N\times M}{}'F^{\perp}_{N\times M} = (F_{N\times M}\,S_{M\times M})'F_{N\times M}\,S_{M\times M} = S_{M\times M}{}'(F_{N\times M}{}'F_{N\times M})S_{M\times M} = S_{M\times M}{}'P_{M\times M}S_{M\times M}$，其中倒数第一个和第三个等号分别是利用了已于上方注明的 $P_{M\times M}$ 和 $F^{\perp}_{N\times M}$ 的定义，移项可得

$$S_{M\times M}S_{M\times M}{}' = P^{-1}_{M\times M} \ ;$$

第三步：上述方程的通解是

$$S_{M\times M} = P^{-\frac{1}{2}}_{M\times M}C_{M\times M} \ ,$$

其中 $C_{M\times M}$ 为任意正交阵；

第四步：根据前面提到的（L4），$P_{M\times M}$ 作为实对称矩阵可对角化，即存在正交阵 $O_{M\times M}$ 以及对角阵 $D_{M\times M}$ 使得下式成立：

$$P_{M\times M} = O_{M\times M}D_{M\times M}O^{-1}_{M\times M} \ ;$$

第五步：综上第三、四步，我们可以得到

$$S_{M\times M} = O_{M\times M}D^{-\frac{1}{2}}_{M\times M}O^{-1}_{M\times M}\ C_{M\times M} (= O_{M\times M}D^{-\frac{1}{2}}_{M\times M}O_{M\times M}{}'\ C_{M\times M}) \ 。$$

2. 个性

简要的概括一下我们上面提到的正交化的目标：

$$F^{\perp}_{N\times M} = F_{N\times M}S_{M\times M} \ ,$$

其中，$S_{M\times M} = O_{M\times M}D^{-\frac{1}{2}}_{M\times M}O_{M\times M}{}'\ C_{M\times M}$ 。

图表2：不同正交方法下的 $C_{M\times M}$ 与 $S_{M\times M}$ 定义方式

正交方法	$C_{M\times M}$ 与$S_{M\times M}$定义方式	前后关系	定序与否	其他优缺点
Schmidt 正交	$S_{M\times M}$为上三角矩阵，$S_{M\times M} = O_{M\times M}D^{-\frac{1}{2}}_{M\times M}O_{M\times M}{}'\ C_{M\times M}$	一一对应	是	初始因子不被正交
对称正交	$C_{M\times M} = I_{M\times M}$，$S_{M\times M} = O_{M\times M}D^{-\frac{1}{2}}_{M\times M}O_{M\times M}{}'$	一一对应	否	正交前后因子相似性最高
规范正交	$C_{M\times M} = O_{M\times M}$，$S_{M\times M} = O_{M\times M}D^{-\frac{1}{2}}_{M\times M}$	时序不稳定	是	-

资料来源：中信期货研究所

我们这里再具体补充几个细节：

（2.1）Schmidt 正交的问题在于需要确定因子的正交顺序，正交顺序不同，最终得到的因子不同，如果想要保持正交前、后因子的一一对应关系，正交顺序不能随时间变化，要保持一致；我们使用时代码里的正交顺序是直接按照输入因子矩阵的顺序，从左向右依次正交；

（2.2）对称正交的 $S_{M\times M}$ 是一个对角阵，因而叫作对称正交化；它有很多良好的性质，譬如：其前后因子的一一对应关系稳定，并且与因子的正交顺序无关，这从 $S_{M\times M}$ 的表达式就可以看出来，因子顺序变化后，对应的 $O_{M\times M}$ 和 $D_{M\times M}$ 列向量也调整，但最终各因子对应的列的乘积是不变的；对称正交是所有正交方法中，使得旋转前后因子间的距离

$$d \triangleq \sum_{i=1}^{M} \|f^i - \tilde{f}^i\|^2$$

最小的正交化方法，这就保证了正交化前后因子的相似性依然很高，信息损失小；

（2.3）规范正交中，上方图表中已清晰表明取 $C_{M\times M} = O_{M\times M}$ 和 $S_{M\times M} = O_{M\times M}D_{M\times M}^{-\frac{1}{2}}$，从而

$$F_{N\times M}^{1} = F_{N\times M}\, S_{M\times M} = F_{N\times M}O_{M\times M}D_{M\times M}^{-\frac{1}{2}},$$

其中，$O_{M\times M}$ 为 $F_{N\times M}$ 的协方差阵的特征向量阵，因此 $F_{N\times M}O_{M\times M}$ 实际上得到的是主成分分析中的所有主成分，主成分分析的第一主成分根据方差最大的方向确定，这就导致典型正交化前后因子的对应关系在时间上是变化的、不稳定，$D_{M\times M}$ 是对角阵，因子乘以 $D_{M\times M}^{-\frac{1}{2}}$ 可以视作对因子的缩放。

3. 回归取残差

这一部分的目的是为了证明前面提到的回归取残差这种正交方式其实质上与 Schmidt 正交化是一致。

首先我们简要地回顾一下 Schmidt 正交化方法：

Step 1:按照一定顺序把每个向量与之前所有向量进行正交；

Step 2:对于正交后的向量进行归一化，最终得到的所有向量两两正交且模为 1，正交后的因子暴露矩阵为正交阵。

其具体计算步骤为：

（S.1）$g_1 = f^1$ ；

（S.2）$g_2 = f^2 - \frac{\langle f^2,\ g_1\rangle}{\langle g_1\ g_1\rangle} g_1$ ；

（S.3）$g_3 = f^3 - \frac{\langle f^3,\ g_1\rangle}{\langle g_1,\ g_1\rangle} g_1 - \frac{\langle f^3,\ g_2\rangle}{\langle g_2, g_2\rangle} g_2$ ；……

（S.4）$g_M = f^M - \frac{\langle f^M,\ g_1\rangle}{\langle g_1,\ g_1\rangle} g_1 - \frac{\langle f^M,\ g_2\rangle}{\langle g_2,\ g_2\rangle} g_2 - \cdots - \frac{\langle f^M,\ g_{M-1}\rangle}{\langle g_{M-1},\ g_{M-1}\rangle} g_{M-1}$ ；

（S.5）归一化：$e_i = \frac{g_i}{|g_i|}, i \in \{1, \ldots, M\}$ 。

回归取残差的方法其过程类似于 Schmidt 正交化，按照一定的顺序将每个因子向量同之前的所有向量回归取残差代替原值。接下来证明， Schmidt 正交化与最小二乘下的回归取残差是一致的。差别仅在于，Schmidt 正交化多了一步归一化。

Step 1：对于因子 f^1 ，之前没有别的因子，不需要进行回归，保持不变：$\tilde{f}^1 \triangleq f^1$；

Step 2：对于因子 f^2 ，做回归模型： $f^2 = \alpha + \beta f^1 + \varepsilon$；由于因子进行了标准化，均值为 0，所以参数的最小二乘估计量可以化简为

$$\begin{cases} \beta = \frac{cov(\tilde{f}^1, f^2)}{var(\tilde{f}^1)} = \frac{\langle \tilde{f}^1, f^2\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \\ \alpha = 0 \end{cases} ;$$

用残差代替原值：$\tilde{f}^2 \triangleq \varepsilon = f^2 - (\alpha + \beta \tilde{f}^1) = f^2 - \frac{\langle \tilde{f}^1, f^2\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \tilde{f}^1$，并且此时 $\tilde{f}^2$ 的均值仍为 0；

Step 3： 对于因子 f^3 ，

（RR3.1）先跟 $\tilde{f}^1$ 进行回归，同上结果为

$$\tilde{f}^3_{tmp} \triangleq \varepsilon = f^3 - (\alpha + \beta \tilde{f}^1) = f^3 - \frac{\langle \tilde{f}^1, f^3\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \tilde{f}^1;$$

（RR3.2）再跟 $\tilde{f}^2$ 回归，结果为

$$\tilde{f}^3 \triangleq \tilde{f}^3_{tmp} - (\alpha + \beta \tilde{f}^2) = \tilde{f}^3_{tmp} - \frac{\langle \tilde{f}^2, \tilde{f}^3_{tmp}\rangle}{\langle \tilde{f}^2, \tilde{f}^2\rangle} \tilde{f}^2;$$

（RR3.3）把上式中 $\tilde{f}^3_{tmp}$ 用第一步回归的结果替换：

$$\tilde{f}^3 \triangleq \tilde{f}^3_{tmp} - \frac{\langle \tilde{f}^2, \tilde{f}^3_{tmp}\rangle}{\langle \tilde{f}^2, \tilde{f}^2\rangle} \tilde{f}^2 = f^3 - \frac{\langle \tilde{f}^1,\ f^3\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \tilde{f}^1 - \frac{\left\langle \tilde{f}^2, f^3 - \frac{\langle \tilde{f}^1,\ f^3\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \tilde{f}^1 \right\rangle}{\langle \tilde{f}^2, \tilde{f}^2\rangle} \tilde{f}^2$$

$$= f^3 - \frac{\langle \tilde{f}^1, f^3\rangle}{\langle \tilde{f}^1, \tilde{f}^1\rangle} \tilde{f}^1 - \frac{\langle \tilde{f}^2, f^3\rangle}{\langle \tilde{f}^2, \tilde{f}^2\rangle} \tilde{f}^2,$$

最后一个等号用到了 $\langle \tilde{f}^1, \tilde{f}^2 \rangle = 0$ ；

Step 4：以此类推，对于任意的 $i > 1$ ，都有

$$\tilde{f}^i = f^i - \frac{\langle \tilde{f}^1, f^i \rangle}{\langle \tilde{f}^1, \tilde{f}^1 \rangle} \tilde{f}^1 - \frac{\langle \tilde{f}^2, f^i \rangle}{\langle \tilde{f}^2, \tilde{f}^2 \rangle} \tilde{f}^2 - \cdots - \frac{\langle \tilde{f}^{i-1}, f^i \rangle}{\langle \tilde{f}^{i-1}, \tilde{f}^{i-1} \rangle} \tilde{f}^{i-1},$$

跟 Schmidt 正交化归一化前的结果一模一样。

三、正交化实践中遇到的问题

从正交化的“线性代数理论基础”具体落地到 python 的“自定义函数”，一般而言，数学与代码的对照实现没有难度。然而在实践中，还是较频繁出现了一些典型问题，因此有必要在这里进行阐明我们的发现、总结与解决办法。

（一）协方差阵的特征值分解中遇到复数和负值（针对 python 用户）

针对复数的处理：

（CPLX 1）弃 np.linalg.eig()，使用 sympy 包的 Matrix().diagonalize()；

（CPLX 2）只取实部：np.real_if_close()；

针对负数的处理：

（NEG 1）负特征值给 0 并删除，并删除相应特征向量；

（NEG 2）特征值全体取正；

这里有必要多解释几句：从线性代数的理论出发，多因子的协方差矩阵是半正定的而不必正定，这样就导致了特征值为 0 的情形。而 Python 的正交分解实践中会虽然遇到复值或负值，但复值的虚数部分和负值都极小接近于 0，其实质就是 0，只是计算机在运算过程中对较长小数位做了截断后再做正交分解是造成的后果，也因此上述的（CPLX 2）、（NEG 1）和（NEG2）在这里是合理的解决办法。

四、组合加权和因子预期收益率的估计

这一小节具体介绍一下加权方式（针对动态期货池、以截面回归方式来构建组合权重）以及其中涉及到的“因子预期收益率估计”这一环节的优化思路。

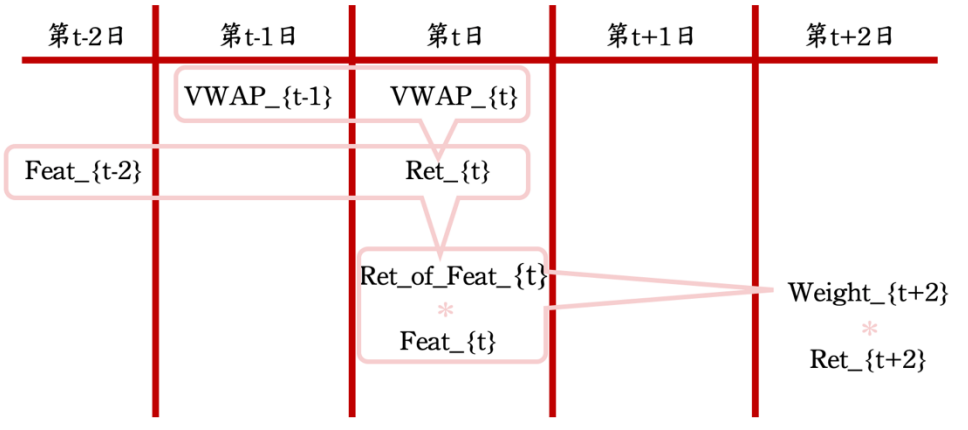
（一）组合加权方式

该方法的思路是受到期货样本池中期货合约上市日期的启发。一般而言，Barra 模型是以其作为风险模型的定位受到市场的熟知和广泛应用，譬如其针对中国 A 股市场推出的 CNE5 多因子模型等，我们利用它主要是为了得到其在相应因子上的因子收益率作为风险暴露，便于将收益与模型的多因子对照起来以考察相应的因子贡献。而我们这里则是将因子收益率作为一个中间值用于我们之后预测收益率。

具体而言，我们考虑了 41 种商品期货合约，不同合约指数的上市日期有先后之分，最早的日期是 2005 年 1 月 4 日，如 A、AL、C、CU、M 和 RU，最晚的日期为 2021 年 1 月 8 日，如 LH；最早与最晚的上市日期相隔 16 年之久。如果我们考虑一个静态的期货池，即保持存在行情数据的每个截面的期货品种不变，那么势必会极大地缩短回测的时间序列长度，从而错失对之前数据的考量以及误判整个市场的全貌。

我们对此问题的解决办法是使用定长的 5 日窗口来在整个时间序列上滚动完成截面回归从而构建组合权重，相应的图示如下。具体而言，我们首先定位在回看时间序列中的第 t 日来开始我们的讨论，基于此我们可以获得第$t-2$的多因子数据 $feature_{t-2}$和第 t 日的收益率 r_t ，而后者是基于第 t 日 $VWAP_t$ 与和第 $t-1$ 日成交量加权价 $VWAP_{t-1}$ 得到；其次，我们使用线性回归来求得第 $t-2$ 的多因子数据 $feature_{t-2}$和第 t 日的收益率 r_t 的权重系数 β_t ，部分文献也将其称之为因子收益率；再者，我们将第 t 期的因子收益率 β_t 与这一期的因子值 $feature_t$ 以点乘的形式来得到第 $t+2$ 日的预测权重向量 $\tilde{r}_{t+2}$ ，经过归一化处理后其与第 $t+2$ 日的期货实际收益率向量的乘积和即作为我们对整个期货组合的预测收益率。

图表3：使用以截面回归方式来构建组合权重的计算示意图



资料来源：中信期货研究所

（二）因子预期收益率的估计

在每一个截面上进行线性回归，我们可以得到当期所有因子的收益率；将同一因子不同截面的收益率依时间先后拼接汇总起来，则可以得到该因子的历史收益率时间序列。对于下一期因子预期收益率，除了使用上一小节直接计算得到的数值，我们还尝试了以下优化方式——简单移动平均：用过去 n 个交易日的历史收益率均值作为下一期的因子预期收益率。

五、回测

我们将考虑多种组合方式："未正交/正交"以及搭配"动态商品期货池的以截面回归方式来构建组合权重"，其中使用的因子收益率除了直接使用原始值，还考虑了以下几种情形：过去 n 日简单移动平均线（5 日和 20 日）。

我们下面进行的论述，更多的是从在量化回测中遇到的实际问题作为切入点，以服务于量化实务为目标，来铺陈上述提到的多种组合方式。

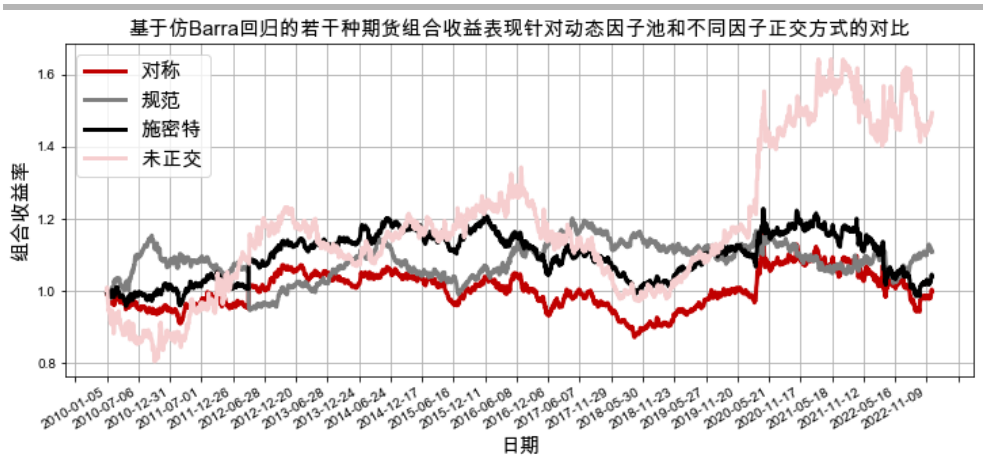
（一）截面正交时，期货品种少于因子数的相应处理

商品期货市场上不同期货品种合约上市日期不一致，而不同因子计算时的回看期也不尽相同。本篇研报在计算 5 日 Barra 回归的因子收益率时，这两者引起的实际问题则是在极少数交易日中"（数值不为空）期货品种数目少于全体因子个数"。从线性代数理论上讲，这种情况下经由三种正交化方式得到的正交后的正交化（线性无关的）因子向量的个数应该小于等于期货品种数，但 python 实际运行中并非如此，譬如正交化后出现比理论规定的更多数目的向量。

针对这个问题，我们尝试了两种方式：第一种是直接跳过"期货品种数目少于全体因子数目"的交易日、不对其进行考虑，这样做的前提是这类型的交易日数目非常少，不会对总体回测区间有较大影响；第二中则是不跳过、但人为地截取"理论个数"个正交向量。

以下是第一种情形的回测净值统计：

图表4：直接跳过“期货品种数目少于全体因子数目”的交易日的回测净值图



资料来源：同花顺 iFind、中信期货研究所

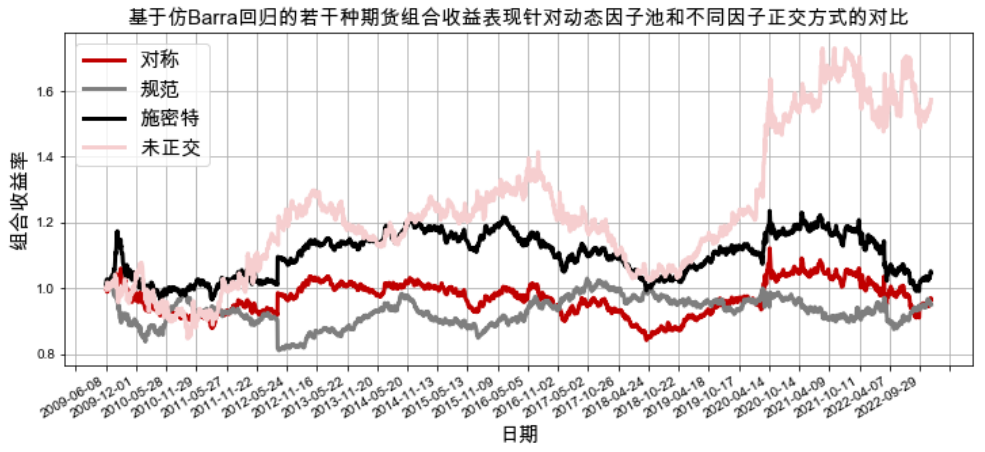
对应的净值统计如下：

图表5：直接跳过“期货品种数目少于全体因子数目”的交易日的回测净值统计

正交与否	年化收益%	年化波动%	最大回撤	夏普	卡玛
未正交	3.17	12.27	0.28	0.26	0.11
对称	-0.22	6.9	0.21	-0.03	-0.01
规范	-0.4	6.88	0.2	-0.06	-0.02
Schmidt	0.26	7.1	0.2	0.04	0.01

以下是第二种情形的回测净值统计：

图表6：不跳过、但人为截取“理论个数”个正交向量的回测



资料来源：同花顺 iFind、中信期货研究所

图表7：不跳过、但人为截取“理论个数”个正交向量的回测净值统计

正交与否	年化收益%	年化波动%	最大回撤	夏普	卡玛
未正交	3.49	12.33	0.28	0.28	0.13
对称	-0.25	6.9	0.21	-0.04	-0.01
规范	-0.43	6.88	0.2	-0.06	-0.02
Schmidt	0.31	7.1	0.2	0.04	0.02

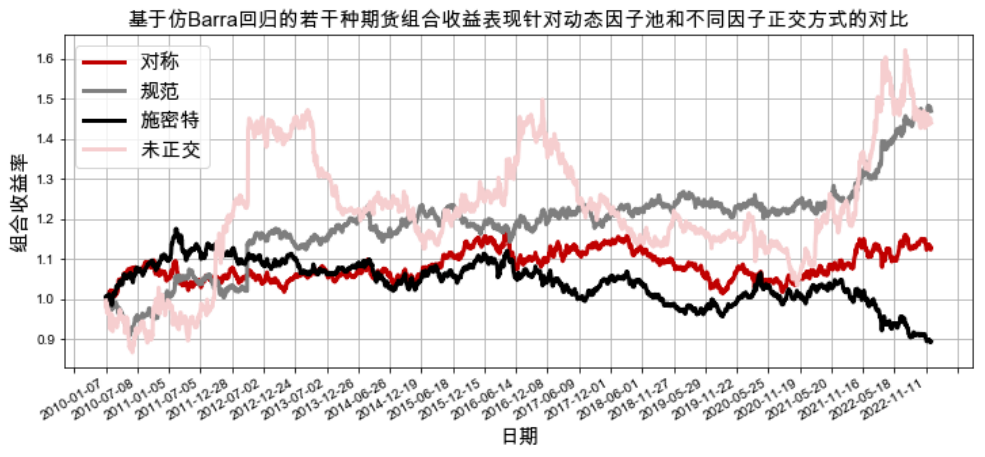
从上方两张图表可以看到第二种方式略微优于于第一种方式，譬如在“未正交”和“Schmidt 正交”这两种方式的尝试中均有年化收益率的提升，此外前者的夏普也得到提升。那么下面的讨论，我们都将采用第二种方式，即不跳过、但人为截取“理论个数”个正交向量进行回测。

（二）若干种因子收益率估计方式的对比

下面是针对非正交化因子的优化方式：简单移动平均 SMA。

首先是 $n=5$ 的简单移动平均 SMA：

图表8：针对非正交化/正交化因子的使用 SMA 优化方式的回测净值图(不跳过、但人为截取“理论个数”个正交向量的回测，$n=5$)



资料来源：同花顺 iFind、中信期货研究所

以下是净值统计结果：

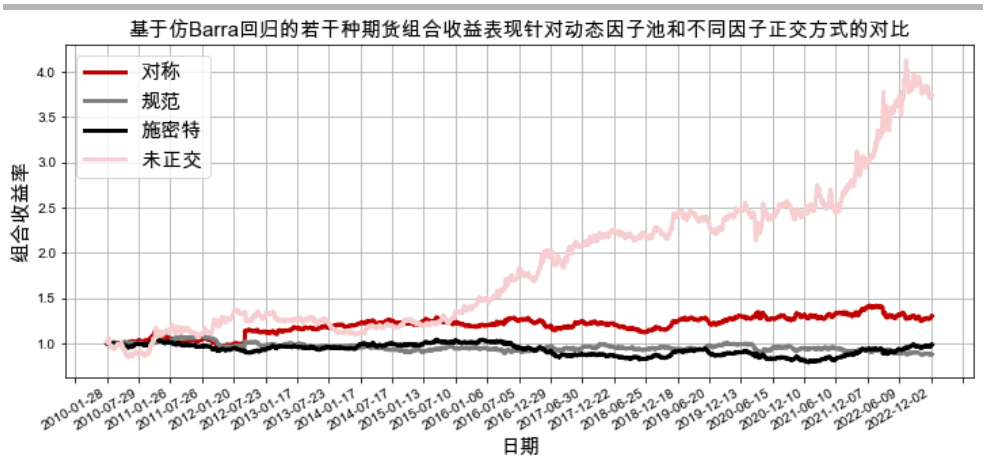
图表9：针对非正交化/正交化因子的使用 SMA 优化方式的回测净值统计（不跳过、但人为截取“理论个数”个正交向量的回测，$n=5$）

正交+优化	年化收益%	年化波动%	最大回撤	夏普	卡玛
未正交+SMA	2.97	10.64	0.31	0.28	0.1
对称+SMA	0.98	4.9	0.14	0.2	0.07
规范+SMA	3.09	5.86	0.09	0.53	0.33
Schmidt+SMA	-0.89	4.77	0.24	-0.19	-0.04

资料来源：同花顺 iFind、中信期货研究所

其次是 $n=20$ 的简单移动平均 SMA：

图表10：针对非正交化/正交化因子的使用 SMA 优化方式的回测净值图（不跳过、但人为截取“理论个数”个正交向量的回测，$n=20$）



资料来源：同花顺 iFind、中信期货研究所

以下是净值统计结果：

图表11：针对非正交化/正交化因子的使用 SMA 优化方式的回测净值统计（不跳过、但人为截取“理论个数”个正交向量的回测，$n=20$）

正交+优化	年化收益%	年化波动%	最大回撤	夏普	卡玛
未正交+SMA	10.95	12.93	0.2	0.85	0.55
对称+SMA	2.11	7.38	0.15	0.29	0.14
规范+SMA	-1.01	6.11	0.2	-0.17	-0.05
Schmidt+SMA	-0.11	6.16	0.26	-0.02	0

资料来源：同花顺 iFind、中信期货研究所

从上面表格中我们发现：

1. 对因子收益率估计采取简单移动平均 SMA（$n=5$）优化后，从（年化收益，最大回撤，夏普）的三元组视角来看，“规范正交+SMA”的回测净值以（3.09,0.09,0.53）优于“未正交+SMA”的净值表现（2.97,0.31,0.28），此外这两者明显优于“对称正交+SMA”和“Schmidt 正交+SMA”，见图表 9；对因子收益率估计采取简单移动平均 SMA（$n=20$）优化后，“未正交+SMA”的组合方式则明显优于任意一种“正交+SMA”的方式，见图表 11；

2. 针对未正交的因子收益率估计采取简单移动平均优化后，其结果显著优于优化前（即未优化）的结果，具体体现在：前者年化收益 10.95 优于后者的 3.49，夏普前者 0.85 优于后者 0.28，最大回撤前者 0.2 低于 0.28；“对称正交+SMA”也优于仅仅对称正交的净值表现；而“规范正交+SMA”和“Schmidt 正交+SMA”这两种则出现相反的情形。

六、总结

本篇报告中，我们以综述的形式讨论了市面上常见的 4 种正交化方法、阐明了正交实践中遇到的问题与解决办法、设计了 1 种商品期货组合的配置模式、配比了 1 种因子收益率估计的优化方式。

这里提到的 4 种正交化方法为“回归取残差、Schmidt 正交化、规范（Canonical）正交化、对称（Symmetric）正交化”；而 1 种商品期货组合配置模式是“作用于动态商品期货池、以截面回归方式来构建组合权重”；1 种因子收益率方式为——简单移动平均。

回测效果来看，可作出以下结论：

（1）本报告提出的配置方式——“作用于动态商品期货池、以截面回归方式来构建组合权重”的配置效果较好，其在本文入选的 6 个量价因子以—“未正交搭配简单移动平均 SMA（$n = 20$）对因子收益率估计”的优化方式回测时，取得了 11%左右的年化收益，夏普为 0.85，最大回撤为 0.2（见图表 11）；

（2）3 种正交化方式是把双刃剑，其实质来源于共线性的双刃性质。在本文入选的 6 个量价类因子的回测中，使用未正交因子的回测净值——无论是对因子收益率未优化时回测得到的年化收益 3.49%（见图表 7）、还是使用简单移动平均 SMA（$n = 20$）优化后回测得到的年化收益 10.95%（见图表 11），均优于相应的基于正交化后因子的回测净值表现。这是由于本文入选的 6 个量价类因子的共线性能够“众人拾柴火焰高”般地有效预测并总体呈“抬升净值”的趋势，但我们在上述回测曲线中也不乏看到在特定的短窗口，基于未正交因子的回测净值线其上升、下跌均较基于正交化后的因子的趋势猛烈。也就是说，我们需要客观的看待正交化这一工具：基于正交化后因子的回测将削弱攀升趋势；而上述配置下的对未正交因子回测的净值走跌时，基于正交化后因子的回测将一扫颓势、抬升净值曲线；但也应该充分关注均线优化中参数对于行情的解释度的重要影响。

（3）对因子收益率估计的优化方式（简单移动平均）无论是针对非正交还是正交化后的因子均能提升策略有效性。

免责声明

除非另有说明，中信期货有限公司拥有本报告的版权和/或其他相关知识产权。未经中信期货有限公司事先书面许可，任何单位或个人不得以任何方式复制、转载、引用、刊登、发表、发行、修改、翻译此报告的全部或部分材料、内容。除非另有说明， 本报告中使用的所有商标、服务标记及标记均为中信期货有限公司所有或经合法授权被许可使用的商标、服务标记及标记。未经中信期货有限公司或商标所有权人的书面许可，任何单位或个人不得使用该商标、服务标记及标记。

如果在任何国家或地区管辖范围内，本报告内容或其适用与任何政府机构、监管机构、自律组织或者清算机构的法律、规则或规定内容相抵触，或者中信期货有限公司未被授权在当地提供这种信息或服务，那么本报告的内容并不意图提供给这些地区的个人或组织，任何个人或组织也不得在当地查看或使用本报告。本报告所载的内容并非适用于所有国家或地区或者适用于所有人。

此报告所载的全部内容仅作参考之用。此报告的内容不构成对任何人的投资建议，且中信期货有限公司不会因接收人收到此报告而视其为客户。

尽管本报告中所包含的信息是我们于发布之时从我们认为可靠的渠道获得，但中信期货有限公司对于本报告所载的信息、观点以及数据的准确性、可靠性、时效性以及完整性不作任何明确或隐含的保证。因此任何人不得对本报告所载的信息、观点以及数据的准确性、可靠性、时效性及完整性产生任何依赖，且中信期货有限公司不对因使用此报告及所载材料而造成的损失承担任 何责任。本报告不应取代个人的独立判断。本报告仅反映编写人的不同设想、见解及分析方法。本报告所载的观点并不代表中信期货有限公司或任何其附属或联营公司的立场。

此报告中所指的投资及服务可能不适合阁下。我们建议阁下如有任何疑问应咨询独立投资顾问。此报告不构成任何投资、法律、会计或税务建议，且不担保任何投资及策略适合阁下。此报告并不构成中信期货有限公司给予阁下的任何私人咨询建议。

深圳总部

地址：深圳市福田区中心三路 8 号卓越时代广场（二期）北座 13 层 1301-1305、14 层

邮编：518048

电话：400-990-8826

传真：(0755)83241191

网址：http://www.citicsf.com