



衍生品量化择时系列专题（八）： 基于聚类算法的商品基本面大类研究

报告日期：2022 年 9 月 1 日

★研究背景：

在商品基本面量化领域，模型构建和数据处理是两个主要的研究方向。本系列之前的报告着重讨论了基本面量化的模型构建，而本报告则以数据处理为出发点，力求从数据层面给予模型更好的特征输入。

★模型构建：

数据收集：本报告数据均来自于繁微数据平台，数据类别包含库存、情绪、原料、终端等数据；指标来源涵盖了Bloomberg、SHFE、海关总署等数据。总数据量达到 392 个。

数据处理：对原始数据的处理包括频率调整、标准化、可得性处理、数据扩充、移仓换月处理；

聚类模型：聚类指通过某个特定标准将数据划分为不同的类或簇，使得同一个类别内数据的相似性尽可能大。本报告尝试采用DTW+KMeans的方式对期货的基本面数据进行聚类；

数据降维：本报告降维方式以PCA和KPCA为主。

回测框架：采用OLS多元线性拟合，并进行滚动回归。

★模型结果：

根据报告最后构建的多品种横截面多空模型，策略整体的年化收益达到 26.85%，年化波动 11.54%，夏普值 2.12，最大回撤-13.62，胜率 0.61。由于回测阶段未考虑到交易滑点，真实年化收益率应当略低于该值。

★风险提示：

市场风格的切换会造成特征有效性发生变化，导致模型效果下降。

王冬黎 高级分析师(金融工程)
从业资格号: F3032817
投资咨询号: Z0014348
Tel: 8621-63325888-3975
Email: dongli.wang@orientfutures.com

联系人： 谢怡伦（分析师）
从业资格号: F03091687
Tel: 8621-63325888-1585
Email: yilun.xie@orientfutures.com

相关报告

《衍生品量化择时系列专题（二）——螺纹钢指标筛选与大类因子研究》

《衍生品量化择时系列专题（三）——PTA 指标筛选与大类因子合成研究》

《衍生品量化择时系列专题（五）——基于机器学习的螺纹钢价格周度预测》

《衍生品量化择时系列专题（七）——基于遗传规划的期货因子挖掘》

目录

1、研究背景	5
2、模型构建	5
2.1、数据收集与初步筛选	5
2.2、数据处理	7
2.3、聚类模型	7
2.3.1、DTW	7
2.3.2、KMeans	9
2.4、降维	11
2.4.1、PCA	11
2.4.2、KPCA	12
2.5、回测框架	12
3、聚类降维实证	13
3.1、不聚类直接降维	13
3.2、聚类降维	16
4、多品种策略	23
4.1、豆粕	23
4.2、PP	24
4.3、PTA	25
4.4、镍	26
4.5、锌	26
4.6、螺纹钢	27
4.7、铁矿石	28
4.8、焦炭	29
5、多品种横截面多空模型	30
5.1、各品种等权构建	30
5.2、基于波动率对信号强度调整	31
5.3、基于波动率对权重进行调整	32
6、结论	33

图表目录

图表 1：基本面数据分类	6
图表 2：基本面数据来源	6
图表 3：基本面数据举例	6
图表 4：锁步度量	7
图表 5：弹性度量	8
图表 6：“锁步度量”矩阵展开	8
图表 7：“弹性度量”矩阵展开	8
图表 8：KMeans 聚类过程	10
图表 9：KPCA 核函数	12
图表 10：滚动回归	13
图表 11：全量降维回测指标（铜）	13
图表 12：全量降维回测曲线（铜）	14
图表 13：预测正确率 VS 收益率	14
图表 14：全量降维，阈值为 1% 回测指标（铜）	15
图表 15：全量降维，阈值为 1% 回测曲线（铜）	15
图表 16：全量降维不同阈值回测指标（铜）	16
图表 17：聚类 Cluster0	16
图表 18：聚类 Cluster1	17
图表 19：聚类 Cluster2	17
图表 20：聚类 Cluster3	18
图表 21：聚类 Cluster4	18
图表 22：聚类降维，阈值为 0% 回测指标	19
图表 23：聚类降维，阈值为 1% 回测指标	19
图表 24：聚类 cluster0，阈值为 0% 回测曲线	19
图表 25：聚类 cluster1，阈值为 0% 回测曲线	20
图表 26：聚类 cluster2，阈值为 0% 回测曲线	20
图表 27：聚类 cluster3，阈值为 0% 回测曲线	21
图表 28：聚类 cluster4，阈值为 0% 回测曲线	21
图表 29：聚类降维不同阈值下回测指标	22
图表 30：聚类降维，阈值为 0% 回测曲线	22
图表 31：等权合成信号回测指标	22
图表 32：等权合成信号回测曲线	23
图表 33：聚类降维各阈值下回测指标（豆粕）	23
图表 34：聚类降维，阈值为 0% 回测曲线（豆粕）	24

图表 35：聚类降维各阈值下回测指标（PP）	24
图表 36：聚类降维，阈值为 0% 回测曲线（PP）	25
图表 37：聚类降维各阈值下回测指标（PTA）	25
图表 38：聚类降维，阈值为 0% 回测曲线（PTA）	25
图表 39：聚类降维各阈值下回测指标（镍）	26
图表 40：聚类降维，阈值为 0% 回测曲线（镍）	26
图表 41：聚类降维各阈值下回测指标（锌）	27
图表 42：聚类降维，阈值为 0% 回测曲线（锌）	27
图表 43：聚类降维各阈值下回测指标（螺纹钢）	27
图表 44：聚类降维，阈值为 0% 回测曲线（螺纹钢）	28
图表 45：聚类降维各阈值下回测指标（铁矿石）	28
图表 46：聚类降维，阈值为 0% 回测曲线（铁矿石）	28
图表 47：聚类降维各阈值下回测指标（焦炭）	29
图表 48：聚类降维，阈值为 0% 回测曲线（焦炭）	29
图表 49：多品种横截面策略回测曲线	31
图表 50：多品种横截面策略回测指标	30
图表 51：多品种横截面策略各品种收益	31
图表 52：不同品种空仓次数	31
图表 53：信号强度调整后各品种空仓比例	32
图表 54：信号强度调整后策略表现	32
图表 55：权重调整后策略表现	32

1、研究背景

在商品基本面量化领域，模型构建和数据处理是两个主要的研究方向。本系列之前的报告着重讨论了基本面量化的模型构建，而本报告则以数据处理为出发点，力求从数据层面给予模型更好的特征输入。对输入特征的考量取决于模型的回测结果，回测过程采用滚动回归的方式，这样不仅可以避免模型过拟合的风险出现，同时也尽可能还原该策略用于实盘交易的收益情况。

由于商品基本面数据量较多，在作为特征输入给模型之前，需要对数据进行降维处理，这样可以提升模型的稳健性，同时也降低模型计算的复杂程度。传统的降维方式是采用主成分分析法（PCA 降维），在每次滚动更新模型之前，对数据进行全量降维。本文的重点在于考察商品基本面数据不同类别的预测能力，基本面数据的分类通常按照商品产业链上下游关系或者分析师的主观判断进行分类，然而在数据的时序结构层面同一类的数据可能存在较大的差异，如何将在时序上具有较强相关性的数据归到一类是一个值得思考的问题。一个容易想到的方式是根据时序数据之间的相关性进行分类，而该方式无法准确识别数据在时序上的特征的相似性，故本报告通过计算两个数据之间的 DIW 距离来衡量其相似性，并且通过对大类内部的数据进行降维处理来考察各类数据的预测能力。

总之，在商品基本面量化的研究过程中，数据处理的目的是从有限的数量之中挖掘出尽可能多的有效信息，同时也应当考虑模型的运行效率。无论是引入数据的差分、同比、环比等指标，还是对数据进行标准化、降维等处理，都是基于此为目的的。

2、模型构建

2.1、数据收集与初步筛选

本报告数据均来自于繁微数据平台，繁微是东证期货自行研发的投研一体式智能投研平台，该平台集成了目前市面上绝大多数大宗商品数据来源，本报告收集了“铜”目录下所有的基本面数据。数据类别包含库存、情绪、原料、终端等数据；指标来源以 Wind 和上海钢联为主，也涵盖了 Bloomberg、SHFE、海关总署等数据。总数据量达到 392 个。

由于基本面数据经常出现缺失值以及数据停更的情况，首先对数据进行初步筛选。本阶段数据选取时段为 2015 年 1 月 1 日-2022 年 7 月 26 日，为保证数据在时间上的连续性，将 2016 年 1 月 1 日之前未发布的指标删除，同时将 2022 年之后停更的数据指标删除；此外，为剔除超低频数据的影响，将数据点个数不超过 50 个的数据剔除（此步骤主要剔除半年度和年度数据）。经过初步筛选后，数据量从 392 下降至 198 个。

图表 1：基本面数据分类

分类（繁微三级目录）	个数
供需平衡表	6
铜加工	30
铜价格数据	47
铜交易情绪指标	15
铜库存	60
铜冶炼	96
铜原料	91
铜终端	47
总计	392

资料来源：东证衍生品研究院

图表 2：基本面数据来源

指标来源	个数
Bloomberg	9
Cochilco	2
ICSG	3
SHFE	6
Wind	180
东证衍生品研究院	9
国家统计局	1
海关总署	4
上海钢联	171
新闻整理	5
NA	2
总计	392

资料来源：东证衍生品研究院

图表 3：基本面数据举例

指标名称	指标来源	指标单位	指标频度
中国精铜产量（月度）	东证衍生品研究院	万吨（金属）	月度
中国精铜净进口量（月度）	东证衍生品研究院	万吨（金属）	月度
ICSG：全球精铜产量（年度）	ICSG	千吨（金属）	年度
ICSG：全球铜矿产量（年度）	ICSG	千吨（金属）	年度
ICSG：全球精铜需求（年度）	ICSG	千吨（金属）	年度
中国精铜表观需求量（月度）	东证衍生品研究院	万吨（金属）	月度
全球铜矿利润（C1现金成本模型）	东证衍生品研究院	美元/磅	年度
智利铜矿平均C1现金成本	Cochilco	美元/磅	年度
全球铜矿平均C1现金成本	Cochilco	美元/磅	年度
智利铜矿利润（C1现金成本模型）	东证衍生品研究院	美元/磅	年度
产量:铜选矿产品含铜量:当月值	Wind	万吨	月度
产量:铜选矿产品含铜量:累计值	Wind	万吨	月度
产量:铜选矿产品含铜量:累计同比	Wind	%	月度
铜精矿产量：中国（月度）	上海钢联	金属吨	月度
进口数量:铜矿石及精矿:当月值	Wind	万吨	月度
进口数量:铜矿石及精矿:累计值	Wind	万吨	月度
智利铜矿产量（月度值）	Bloomberg	吨（金属）	月度
秘鲁铜矿产量（月度值）	Bloomberg	吨（金属）	月度
墨西哥铜矿产量（月度值）	Bloomberg	吨（金属）	月度

资料来源：东证衍生品研究院

2.2、数据处理

在经过数据的初步筛选之后，认为保留下的数据具有预测价值，接下来需要进一步对这些数据进行处理，操作如下：

频率调整：基本面的原始数据多为低频数据（月频或周频），为便于处理，将所有数据前值填充为日频数据；

标准化：对所有填充后的数据进行z-score标准化处理，提高数据之间的可比性；

可得性处理：基本面数据的获取存在一定的滞后性，即某一期的数据需要在一定时段之后才能够得到，若未作调整直接进行回测，相当于拿当时无法获取的当期数据进行预测，存在调用“未来数据”的风险，故根据不同指标的滞后期数进行相应调整；

数据扩充：为补充时序数据周期性的变化信息，针对每个指标分别进行月度、季度、年度的差分和比值计算以扩充指标数量。

移仓换月处理：为避免期货展期导致的价格影响，本报告以期货复权价格进行回测；

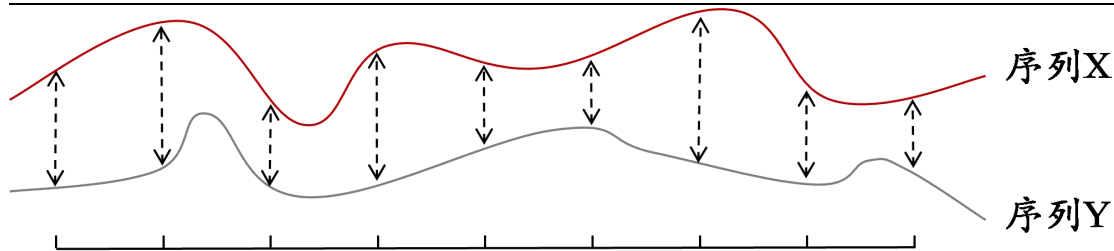
2.3、聚类模型

聚类指通过某个特定标准将数据划分为不同的类或簇，使得同一个类别内数据的相似性尽可能大。本报告尝试采用DTW+KMeans的方式对期货的基本面数据进行聚类。

2.3.1、DTW

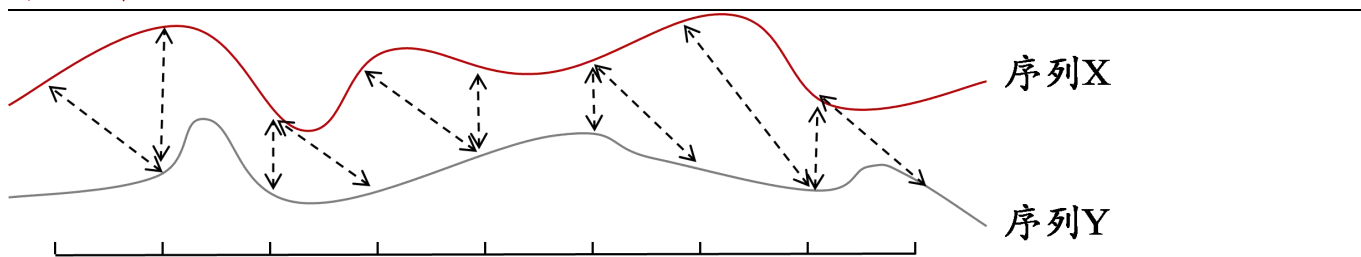
由于基本面数据为时间序列数据，无法通过计算点与点之间的欧式距离或曼哈顿距离来衡量数据之间的相似性。对于时序数据，可以通过“锁步度量”来计算序列的相似性，即依次计算两个序列各点之间的欧式距离再求和，然而这种计算方式存在一系列问题：该方法无法处理序列不同步、步长不一致、长短不一等问题，最重要的是，该方法无法识别数据的时序特征。故本报告采用动态时间规整算法（Dynamic Time Warping, DTW）来计算时序数据之间的“距离”，该算法是一种“弹性度量”算法，DTW通过把时间序列进行延伸和缩短，来计算两个时间序列性之间的相似性。

图表 4：锁步度量



资料来源：东证衍生品研究院

图表 5：弹性度量

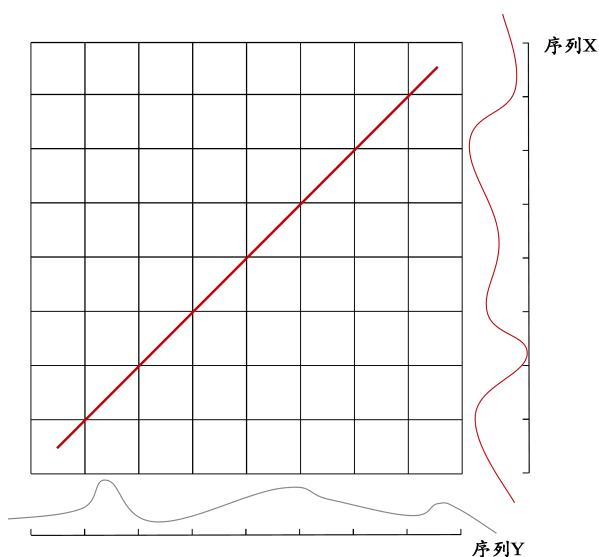


资料来源：东证衍生品研究院

如上图所示，上下两条实线代表两个时间序列，时间序列之间的虚线代表两个时间序列之间的相似的点。DTW 使用所有这些相似点之间的距离的和，称之为归整路径距离 (Warp Path Distance) 来衡量两个时间序列之间的相似性。

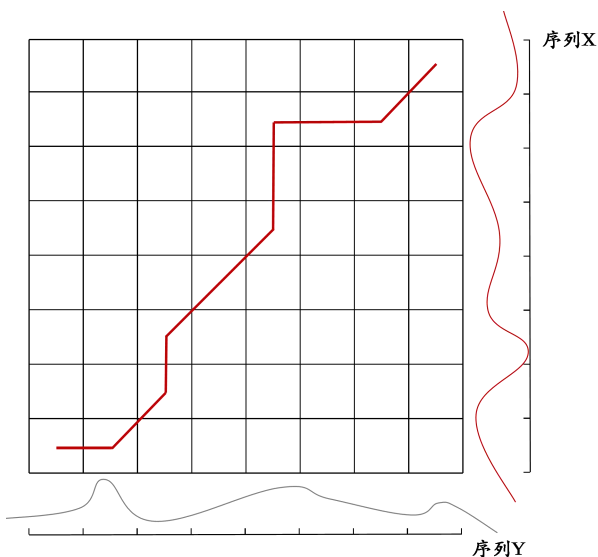
为更加直观地表述，将两个序列通过 $n \times n$ 的矩阵进行展开，每一个单元格表示序列 X 上的某一点到序列 Y 上某一点的距离。左图展示了“锁步度量”的计算方式，该度量路径是从左下直接连接到右上方，表明两条序列之间一一对应的关系；而右图则展示了“弹性度量”的方式，该度量路径通过选择最小化距离之和的方式计算两序列之间的相似性。

图表 6：“锁步度量”矩阵展开



资料来源：东证衍生品研究院

图表 7：“弹性度量”矩阵展开



资料来源：东证衍生品研究院

DTW 算法在寻找最短路径的过程中需要满足三个约束条件：

- 1) **边界条件**：表示两条序列首尾必须匹配，各部分的先后次序匹配；
- 2) **连续性**：这条约束表示在匹配过程中多对一和一对多的情况只能匹配周围一个时间步的情况，也就是不可能跨过某个点去匹配，只能和自己相邻的点对齐；
- 3) **单调性**：路径一定是随时间单调递增的。

2.3.2、KMeans

在定义了时间序列数据的相似度之后，就可以采用 KMeans 算法对时序数据进行聚类操作。

它的基本思想是，通过迭代寻找 K 个簇（Cluster）的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属簇中心点的误差平方和：

$$J(c, \mu) = \sum_{i=1}^M \|x_i - \mu_{c_i}\|^2$$

其中 x_i 代表第 i 个样本， c_i 是 x_i 所属的簇， μ_{c_i} 代表簇对应的中心点， M 是样本总数。

基于时序数据，公式改为：

$$J(c, \mu) = \sum_{i=1}^M DTW(x_i - \mu_{c_i})$$

其中 $DTW()$ 表示时序数据距离计算。

其具体步骤如下：

- 1) 数据预处理；
- 2) 随机选取 k 个中心，记为 $\mu_1^{(0)}, \mu_2^{(0)}, \dots, \mu_k^{(0)}$ ；
- 3) 定义损失函数： $J(c, \mu) = \sum_{i=1}^M DTW(x_i - \mu_{c_i})$ ；

4) 令 $t = 0, 1, 2, \dots$ 为迭代步数，重复下列操作直到 J 收敛；

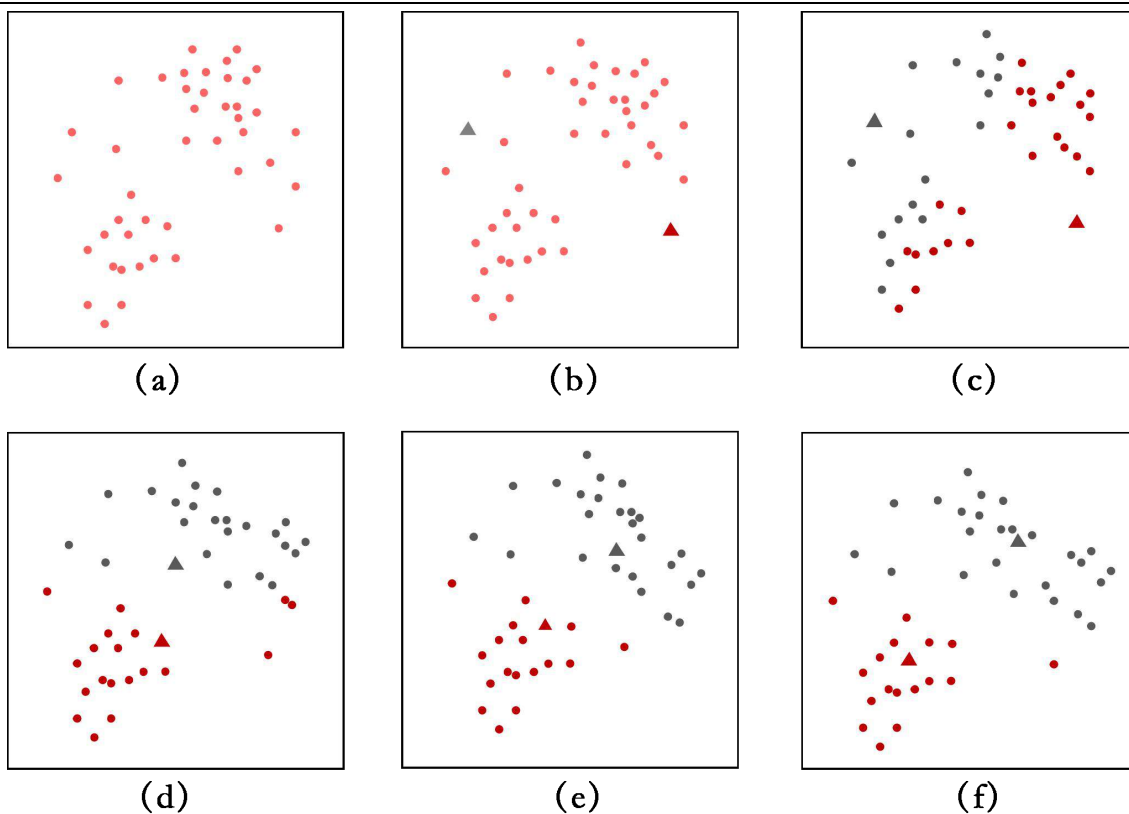
4.1) 对于每一个样本 x_i ，将其分配到距离最近的中心：

$$c_i^t \leftarrow \arg \min_k (DTW(x_i - \mu_k^t)) ;$$

4.2) 对于每一个类的中心 k ，重新计算该类的中心：

$$\mu_k^{t+1} \leftarrow \arg \min_{\mu} \sum_{i: c_i^t = k}^b DTW(x_i - \mu_k^t)$$

图表 8: KMeans 聚类过程



资料来源：东证衍生品研究院

2.4、降维

针对聚类完成后期货基本面数据，本报告对数据进行降维的操作。由于基本面数据彼此之间的相关性较高，且单条数据对期货价格的解释强度低于价量数据，基于以上原因，考虑首先对基本面数据进行降维操作，一方面可以提升单个因子的解释力度，另一方面可以降低模型的复杂程度，以加快模型运行效率。

本报告中涉及两种降维方式，以下对两种降维方式作简要介绍：

2.4.1、PCA

主成分分析算法（PCA）是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中，并期望在所投影的维度上数据的信息量最大（方差最大），以此使用较少的数据维度，同时保留住较多的原数据点的特性。

PCA 降维的目的，就是为了在尽量保证“信息量不丢失”的情况下，对原始特征进行降维，也就是尽可能将原始特征往具有最大投影信息量的维度上进行投影。将原特征投影到这些维度上，使降维后信息量损失最小。其算法步骤如下：

设有 m 条 n 维数据：

- 将原始数据按列组成 n 行 m 列矩阵 X ；
- 将 X 的每一行进行零均值化，即减去这一行的均值；
- 求出协方差矩阵 $C = \frac{1}{m}XX^T$ ；
- 求出协方差矩阵的特征值及对应的特征向量；
- 将特征向量按对应特征值大小从上到下按行排列成矩阵，取前 k 行组成矩阵 P ；
- $Y = PX$ 即为降维到 k 维后的数据。

在进行 PCA 降维之后，可以缓解维度灾难，并对数据进行降噪，同时将数据压缩到低维之后，使得降维之后的数据各特征相互独立。但是在另一方面，由于 PCA 保留了主要信息，舍弃了一些看似无用的信息，但这些“无用信息”只是在训练集上没有有效表现，因此产生了过拟合的可能性，这一问题在模型训练时需要注意。

2.4.2、KPCA

针对非线性的数据，KPCA（核函数主成分分析法）将非线性可分的数据转换到一个适合对其进行线性分类的新的低维子空间上。利用核 PCA 可以通过非线性映射将数据转换到一个高维空间中，在高维空间中使用 PCA 将其映射到另一个低维空间中，并通过线性分类器对样本对其进行划分。

$$\phi: R^d \rightarrow R^k (k \gg d)$$

常见的核函数包括以下几类：

图表 9：KPCA 核函数

核函数	计算公式
线性核	$K(x_i, x_j) = x_i^T x_j$
多项式核	$K(x_i, x_j) = (\gamma x_i^T x_j + b)^d$
径向基函数核/高斯核	$K(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2)$
sigmoid 核	$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + b)$

资料来源：东证衍生品研究院

2.5、回测框架

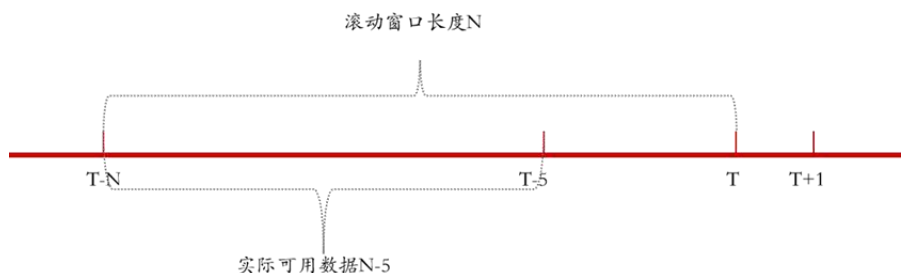
本报告依据 OLS 进行滚动回归预测。OLS（普通最小二乘法）多元回归的原理为，最优拟合曲线使得各点到直线的距离的平方和（残差平方和 RSS）最小：

$$RSS = \sum_{t=1}^T (y_t - \hat{y}_t)^2 = \sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta} x_t)^2$$

本报告采用滚动回归的方式进行回测，以避免使用未来数据。以周度预测为例，首先设置相应的滚动回归窗口长度 N，对每一天 T 都截取 T-N 到 T 时间段的基本面数据，由于为周度预测，需要对基本面数据进行 5 天的移动处理，所以实际可用数据点为 N-5 天内的数据，随后根据训练模型得到一系列预测值，再根据预测值与真实值的比较去构造回测模型。

回测过程基于历史窗口长度为 240 个交易日的基本面周度数据对未来一周的收益率进行滚动回测，回测为周频级别。

图表 10: 滚动回归



资料来源：东证衍生品研究院

3、聚类降维实证

基于繁微数据平台，本报告共整理铜相关基本面因子共 392 个，这些因子涵盖了铜的上下游产业链相关数据，包括库存、原料端、进出口等相关数据，数据来源以 Wind 和上海钢联为主，经过初步筛选过后，剩余 198 个数据。

3.1、不聚类直接降维

首先不进行聚类，直接滚动降维，选取参数滚动窗口为 240 天，信息保留度为 90%，经观察，滚动降维后的数据维度保持在 11 至 14 维。接下来根据拟合结果进行回测，模型会在每一次换仓日滚动建模，根据每一次更新后的模型对最新一期的数据进行收益率预测，根据预测收益率的正负进行相应多空操作。

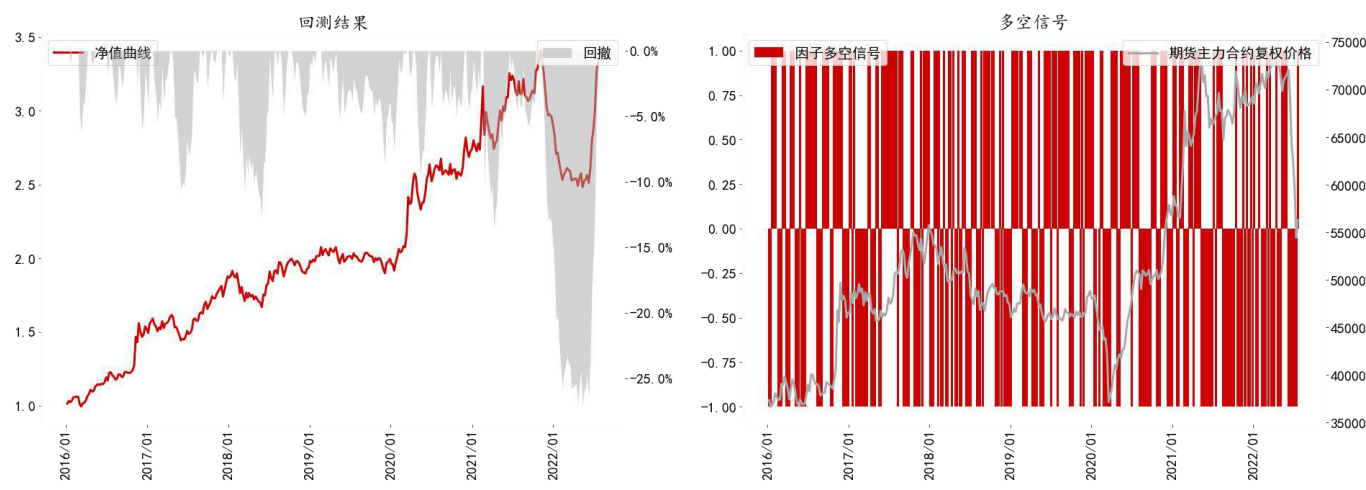
回测结果显示，长期来看年化收益较为显著，达到 19.79%，然而由于在 2022 年 6-7 月份铜期货价格发生暴跌，暴跌前期信号并未及时调整导致在这波行情中策略产生较大回撤。胜率略高于 0.5，达到 0.54，夏普值为 0.43。此外预测值对于真实收益率的 R-square 达到 23.59%。

图表 11: 全量降维回测指标（铜）

总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
233.27%	19.79%	40.64%	0.43	-27.15%	0.73	0.54	1.29	0.9	14.1

资料来源：东证衍生品研究院

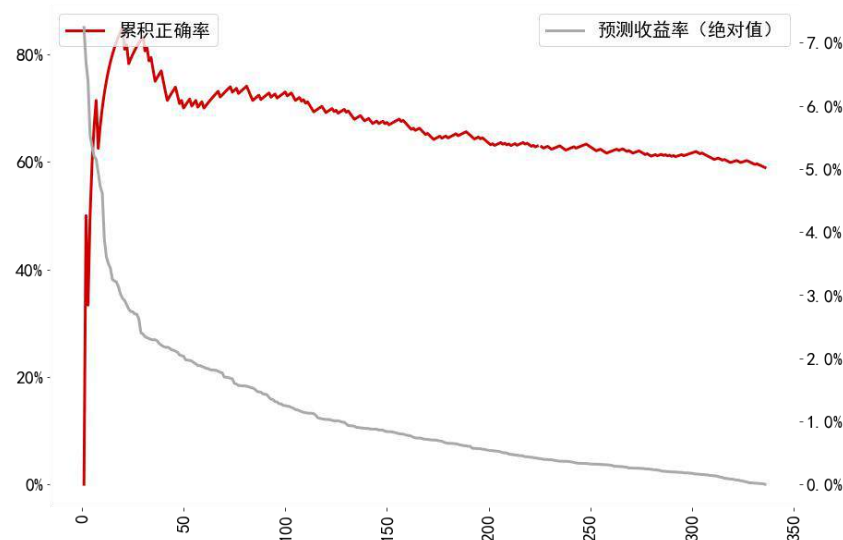
图表 12: 全量降维回测曲线 (铜)



资料来源: 东证衍生品研究院

考虑到预测收益率会给出一些很小的值, 这些值的信号强度并不显著, 不应直接生成多空信号, 为了改进回测结果, 设置多空阈值, 只有当预测收益率的绝对值大于该阈值时才进行相应的多空操作。为了验证这样的猜想, 统计模型的累积预测正确率随阈值逐渐下降的改变。通过下图可以观察到, 除了前期由于预测点较少导致曲线波动外, 随着阈值逐渐减少到 0, 模型预测的累积正确率向下收敛到 55% 左右, 该数据验证了之前的猜想, 即预测收益率绝对值越大, 信号强度越明显, 相应的预测正确率也越高。

图表 13: 预测正确率 VS 收益率



资料来源: 东证衍生品研究院

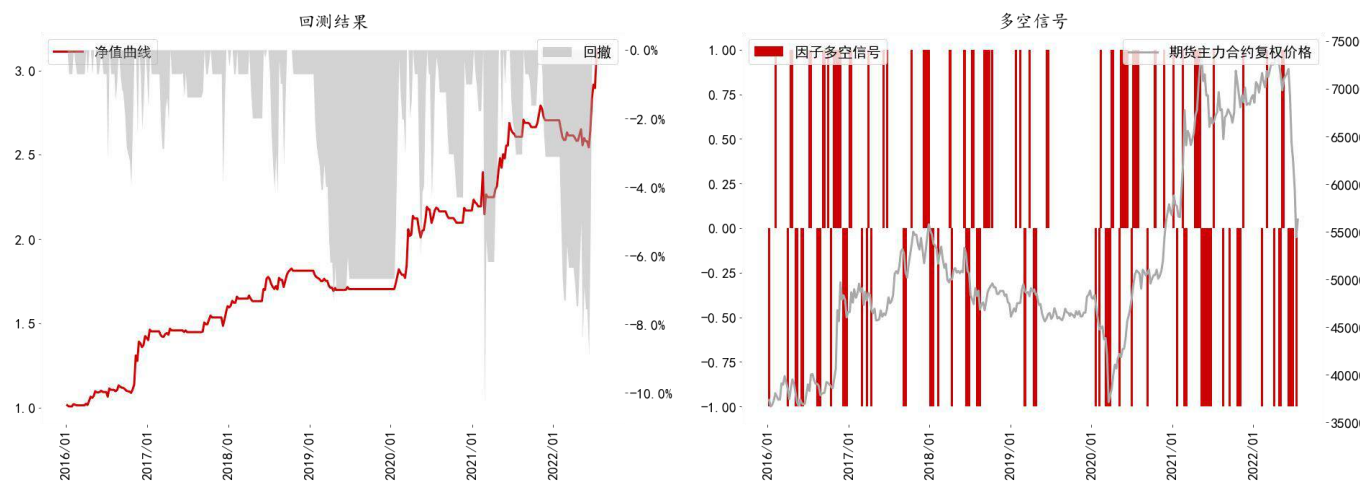
尝试将阈值设置为1%，即只有预测收益率大于1%或小于-1%时才进行相应多空操作，结果如下。从回测指标来看，年化收益并无太大改变，而年化波动显著降低，夏普值也相应提升了，最大回撤控制在-10.39%，胜率达到了0.71。此外，收益曲线图像表明，策略整体收益更为稳健，整体开仓时间显著减少。

图表 14: 全量降维，阈值为1%回测指标（铜）

总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
210.06%	18.50%	32.27%	0.5	-10.39%	1.78	0.71	0.88	1.24	13.9

资料来源：东证衍生品研究院

图表 15: 全量降维，阈值为1%回测曲线（铜）



资料来源：东证衍生品研究院

事实表明，设置合理的开仓阈值有助于策略整体表现的提升。然而，阈值是否“合理”取决于个人的投资偏好。数据表明，开仓阈值越高，策略整体胜率越高，但收益率会降低，另一方面，随着开仓阈值提升，策略开仓次数显著减少，意味着策略的容错率降低，故阈值的选择显得十分重要，下表展示了各个阈值下策略的各项回测指标，读者可以作为参考。

图表 16: 全量降维不同阈值回测指标 (铜)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	233.27%	19.79%	40.64%	0.43	-27.15%	0.73	0.54	1.29	0.90	14.10
0.20%	221.09%	19.12%	39.36%	0.42	-21.58%	0.89	0.52	1.41	0.92	11.15
0.40%	206.82%	18.31%	37.30%	0.43	-17.51%	1.05	0.55	1.31	0.96	10.25
0.60%	202.32%	18.05%	35.81%	0.44	-13.95%	1.29	0.58	1.25	1.01	10.35
0.80%	241.90%	20.25%	33.71%	0.53	-10.39%	1.95	0.64	1.13	1.30	11.85
1.00%	210.06%	18.50%	32.27%	0.50	-10.39%	1.78	0.71	0.88	1.24	13.90
1.20%	192.18%	17.45%	30.24%	0.50	-10.32%	1.69	0.78	0.66	1.24	17.30
1.50%	142.06%	14.18%	28.55%	0.41	-10.91%	1.30	0.81	0.52	0.99	19.75
2.00%	68.10%	8.10%	25.08%	0.23	-15.56%	0.52	0.88	0.29	0.52	28.95
3.00%	41.61%	5.36%	19.64%	0.15	-6.64%	0.81	0.94	0.16	0.57	62.20
4.00%	54.13%	6.71%	18.27%	0.24	-1.21%	5.52	0.98	0.22	2.71	120.00
5.00%	39.23%	5.09%	17.12%	0.16	-1.21%	4.19	0.99	0.17	2.00	168.00

资料来源: 东证衍生品研究院

上图显示, 当阈值小于1%时, 年化收益并无明显区别, 而当阈值大于1%时, 随着阈值增大, 年化收益逐渐降低; 年化波动率和最大回撤均随着阈值增加逐渐降低; 夏普值在阈值为0.8%时为最高值; 胜率随着阈值增加而变高。

3.2、聚类降维

在本节, 根据上文提到的DTW+KMeans算法对筛选过后的数据进行降维。设置聚类数量为5, 由于数据集数量较少, 且各大类数据存在明显差异, 故KMeans算法收敛速度较快, 大约5次之内就收敛, 故将迭代次数设置为5次。下列表格为聚类结果, 共五类, 每类因子个数为11到23个不等, 表格中标红数据为每类的“中心”。

图表 17: 聚类 Cluster0

名称	频率	单位	来源
进口数量:阳极铜:美国:当月值	月度	吨	Wind
出口数量:精炼铜:当月值	月度	吨	Wind
出口数量:精炼铜:韩国:当月值	月度	吨	Wind
出口数量:精炼铜:中国台湾:当月值	月度	吨	Wind
LME铜:注销仓单:合计:全球	日度	吨	Wind
LME铜:库存:韩国:釜山	日度	吨	Wind
LME铜:库存:韩国:光阳	日度	吨	Wind
LME铜:库存:中国台湾:高雄	日度	吨	Wind
进口数量:铜精矿:分国别:土耳其-中国:终值	月度	吨	上海钢联
进口数量:废铜:分国别:泰国-中国	月度	吨	上海钢联
中国地区铜现货升贴水:天津:平水铜	日度	元/吨	上海钢联

资料来源: 东证衍生品研究院

图表 18: 聚类 Cluster1

名称	频率	单位	来源
进口数量:精炼铜:智利:当月值	月度	吨	Wind
进口数量:精炼铜:日本:当月值	月度	吨	Wind
进口数量:精炼铜:澳大利亚:当月值	月度	吨	Wind
进口数量:精炼铜:波兰:当月值	月度	吨	Wind
进口数量:精炼铜:比利时:当月值	月度	吨	Wind
进口数量:精炼铜:德国:当月值	月度	吨	Wind
进口数量:精炼铜:秘鲁:当月值	月度	吨	Wind
进口数量:精炼铜:巴西:当月值	月度	吨	Wind
LME铜:库存:美国:芝加哥	日度	吨	Wind
LME铜:库存:美国:新奥尔良	日度	吨	Wind
进口数量:铜精矿:分国别:蒙古-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:加拿大-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:缅甸-中国:终值	月度	吨	上海钢联
墨西哥铜产量(月度值)	月度	吨(金属)	Bloomberg
ICSG全球铜矿产能利用率(月度值)	月度	%	Wind
进口数量:废铜:分国别:日本-中国	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:保税区仓储转口货物:终值	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:边境小额贸易:终值	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:进料加工贸易:终值	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:一般贸易:终值	月度	吨	上海钢联
ICSG澳大利亚精炼铜产量(月度值)	月度	千吨(金属)	Wind
ICSG智利精炼铜产量(月度值)	月度	千吨(金属)	Wind
中汽协:汽车经销商库存预警指数	月度	NaN	Wind

资料来源:东证衍生品研究院

图表 19: 聚类 Cluster2

名称	频率	单位	来源
进口数量:阳极铜:智利:当月值	月度	吨	Wind
进口数量:阳极铜:墨西哥:当月值	月度	吨	Wind
进口数量:精炼铜:印度:当月值	月度	吨	Wind
出口数量:精炼铜:马来西亚:当月值	月度	吨	Wind
LME铜:库存:英国:赫尔	日度	吨	Wind
进口数量:废铜:分国别:智利-中国	月度	吨	上海钢联
进口数量:废铜:分国别:意大利-中国	月度	吨	上海钢联
进口数量:废铜:分国别:西班牙-中国	月度	吨	上海钢联
进口数量:废铜:分国别:沙特-中国	月度	吨	上海钢联
进口数量:废铜:分国别:韩国-中国	月度	吨	上海钢联
进口数量:废铜:分国别:菲律宾-中国	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:境外设备进口:终值	月度	吨	上海钢联
ICSG全球精炼铜产能利用率(月度值)	月度	%	Wind
ICSG赞比亚精炼铜产量(月度值)	月度	千吨(金属)	Wind
手机产量	月度	台	Wind
期货成交量:阴极铜	日度	手	Wind

资料来源:东证衍生品研究院

图表 20: 聚类 Cluster3

名称	频率	单位	来源
进口数量:阳极铜:中国台湾:当月值	月度	吨	Wind
进口数量:阳极铜:巴基斯坦:当月值	月度	吨	Wind
进口数量:精炼铜:赞比亚:当月值	月度	吨	Wind
进口数量:精炼铜:韩国:当月值	月度	吨	Wind
LME铜:库存:西班牙:毕尔巴鄂	日度	吨	Wind
LME铜:库存:荷兰:鹿特丹	日度	吨	Wind
进口数量:铜精矿:分国别:刚果(金)-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:印尼-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:赞比亚-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:美国-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:南非-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:墨西哥-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分贸易方式:保税区仓库进出境货物:终值	月度	吨	上海钢联
进口数量:阳极铜:刚果(金):当月值	月度	吨	Wind
进口数量:废铜:分国别:越南-中国	月度	吨	上海钢联
进口数量:废铜:分国别:葡萄牙-中国	月度	吨	上海钢联
进口数量:废铜:分国别:马来西亚-中国	月度	吨	上海钢联
进口数量:精炼铜:分贸易方式:其他贸易方式:终值	月度	吨	上海钢联
ICSG波兰精炼铜产量(月度值)	月度	千吨(金属)	Wind
国内铜库存:广东	周度	万吨	上海钢联
LME铜升贴水(0-3)	日度	美元/吨	Wind
LME铜升贴水(3-15)	日度	美元/吨	Wind

资料来源:东证衍生品研究院

图表 21: 聚类 Cluster4

名称	频率	单位	来源
进口数量:精炼铜:哈萨克斯坦:当月值	月度	吨	Wind
出口数量:精炼铜:越南:当月值	月度	吨	Wind
出口数量:精炼铜:印度尼西亚:当月值	月度	吨	Wind
出口数量:精炼铜:泰国:当月值	月度	吨	Wind
LME铜:库存:意大利:的里雅斯特	日度	吨	Wind
进口数量:铜精矿:分国别:秘鲁-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:菲律宾-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:哈萨克斯坦-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分国别:巴西-中国:终值	月度	吨	上海钢联
进口数量:铜精矿:分贸易方式:保税区仓储转口货物:终值	月度	吨	上海钢联
进口数量:铜精矿:分贸易方式:进料加工贸易:终值	月度	吨	上海钢联
智利铜矿产量(月度值)	月度	吨(金属)	Bloomberg
秘鲁铜矿产量(月度值)	月度	吨(金属)	Bloomberg
进口数量:废铜:分国别:中国台湾-中国	月度	吨	上海钢联
进口数量:废铜:分国别:印尼-中国	月度	吨	上海钢联
进口数量:废铜:分国别:俄罗斯-中国	月度	吨	上海钢联
进口数量:废铜:分国别:阿联酋-中国	月度	吨	上海钢联
ICSG日本精炼铜产量(月度值)	月度	千吨(金属)	Wind
ICSG俄罗斯精炼铜产量(月度值)	月度	千吨(金属)	Wind
中国电线电缆企业开工率	月度	%	SMM
中汽协:汽车经销商库存系数	月度	NaN	Wind
进口铜:仓单溢价均价(上海)	日度	美元/吨	上海钢联

资料来源:东证衍生品研究院

通过聚类表可以发现，聚类得到的结果与我们的直觉存在一定出入，一些看似具有高相关性的数据被分到了不同的类别当中，比如 Cluster1 中的“进口数量:阳极铜:美国:当月值”与“LME 铜:库存:韩国:釜山”距离较近，而同样是阳极铜进口数据的 Cluster3 中的“进口数量:阳极铜:智利:当月值”与“进口数量:精炼铜:印度:当月值”距离较近，由于此处的度量为 DTW 距离度量，能够有效提取时间序列数据的时序特征，故聚类结果与人工分类结果存在差异。

下列图表分别展示了阈值为 0% 和 1% 时，各类别的回测指标，同时也展示了在阈值为 0% 时，各类别的回测曲线。数据表明，除 cluster0 表现不尽如人意之外，其余类别的表现均展示了较好的预测能力。不同阈值的设置对回测结果的影响与上文的讨论一致，随着阈值的升高胜率提升，盈亏比下降。

图表 22：聚类降维，阈值为 0% 回测指标

	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
cluster0	22.46%	3.09%	40.21%	0.02	-35.82%	0.09	0.49	1.14	0.03	27.55
cluster1	172.89%	16.25%	38.70%	0.36	-23.29%	0.70	0.51	1.41	0.81	31.10
cluster2	211.30%	18.57%	38.95%	0.42	-27.18%	0.68	0.51	1.44	0.95	44.20
cluster3	148.50%	14.63%	39.07%	0.31	-21.89%	0.67	0.50	1.43	0.68	29.45
cluster4	199.19%	17.87%	39.61%	0.39	-29.64%	0.60	0.50	1.51	0.89	30.00

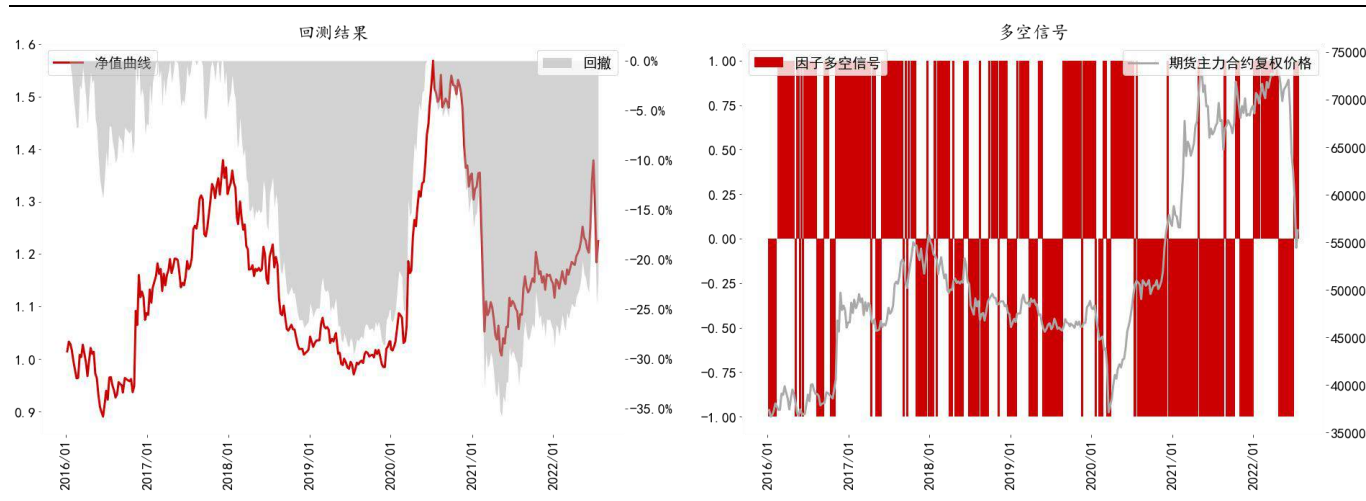
资料来源：东证衍生品研究院

图表 23：聚类降维，阈值为 1% 回测指标

	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
cluster0	11.52%	1.65%	14.83%	-0.05	-9.17%	0.18	0.87	0.20	-0.10	38.20
cluster1	183.31%	16.91%	25.62%	0.57	-4.23%	3.99	0.88	0.53	2.25	31.70
cluster2	174.79%	16.37%	23.89%	0.58	-5.82%	2.81	0.86	0.56	2.15	35.75
cluster3	194.32%	17.58%	23.78%	0.64	-6.28%	2.80	0.85	0.55	2.13	33.60
cluster4	153.78%	14.99%	22.92%	0.55	-12.64%	1.19	0.83	0.54	1.62	29.45

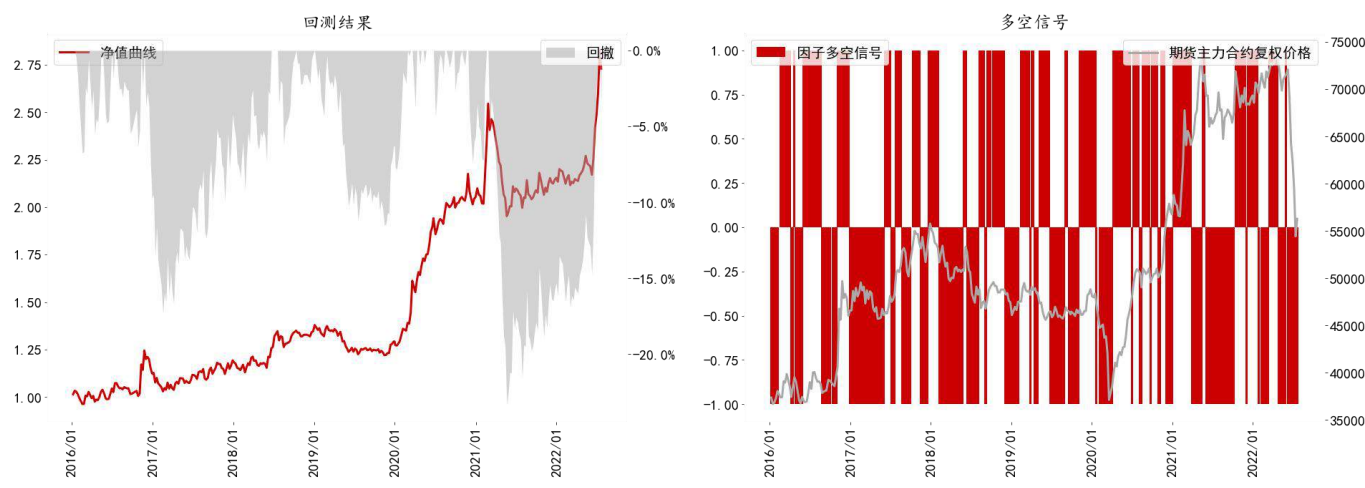
资料来源：东证衍生品研究院

图表 24：聚类 cluster0，阈值为 0% 回测曲线



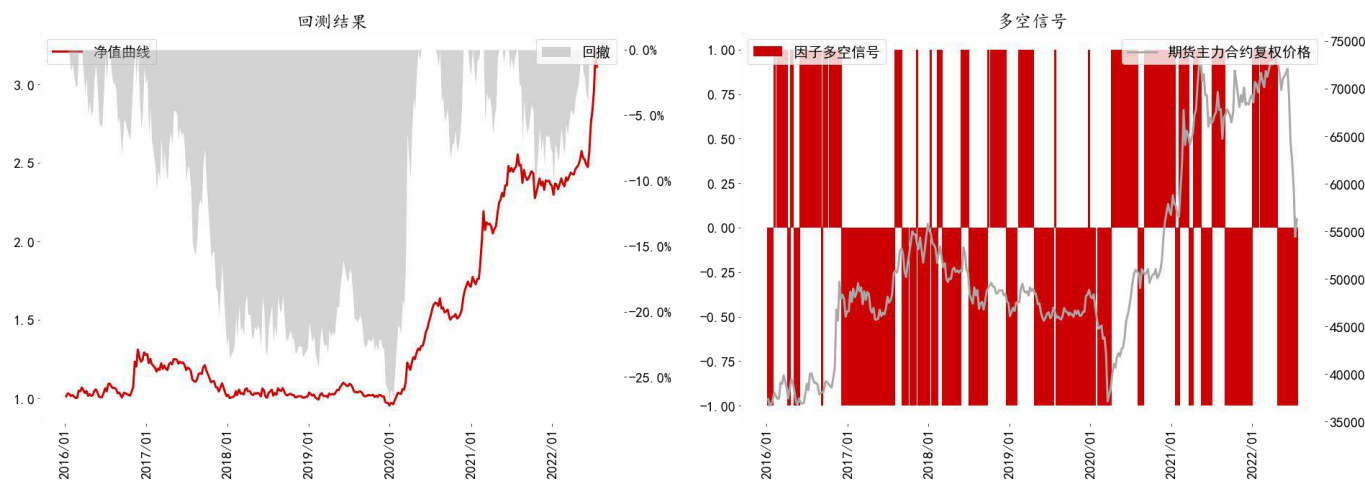
资料来源：东证衍生品研究院

图表 25: 聚类 cluster1, 阈值为 0% 回撤曲线



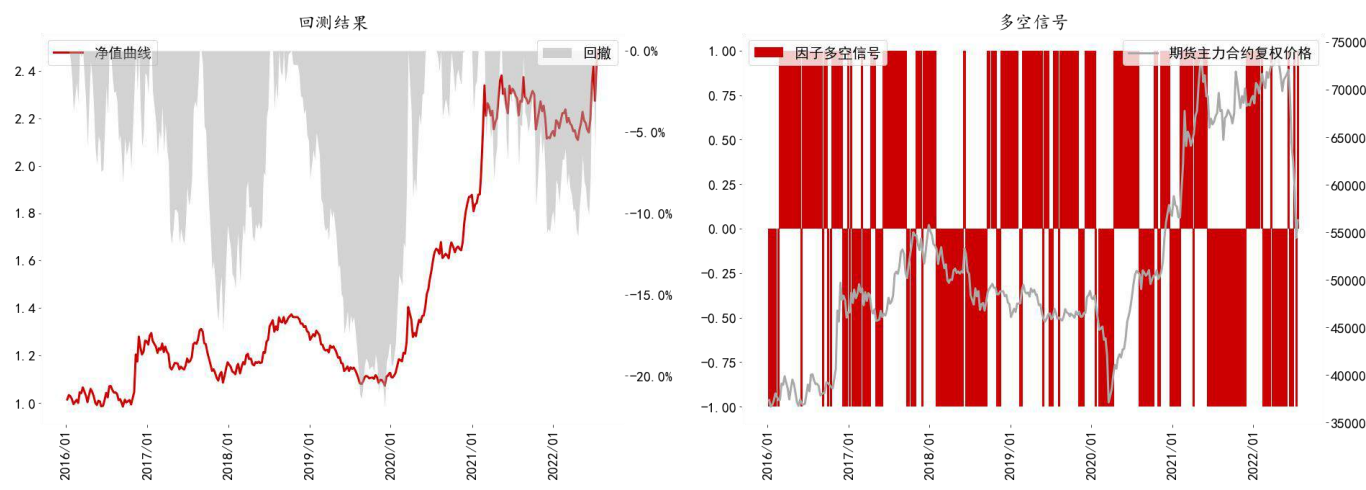
资料来源: 东证衍生品研究院

图表 26: 聚类 cluster2, 阈值为 0% 回撤曲线



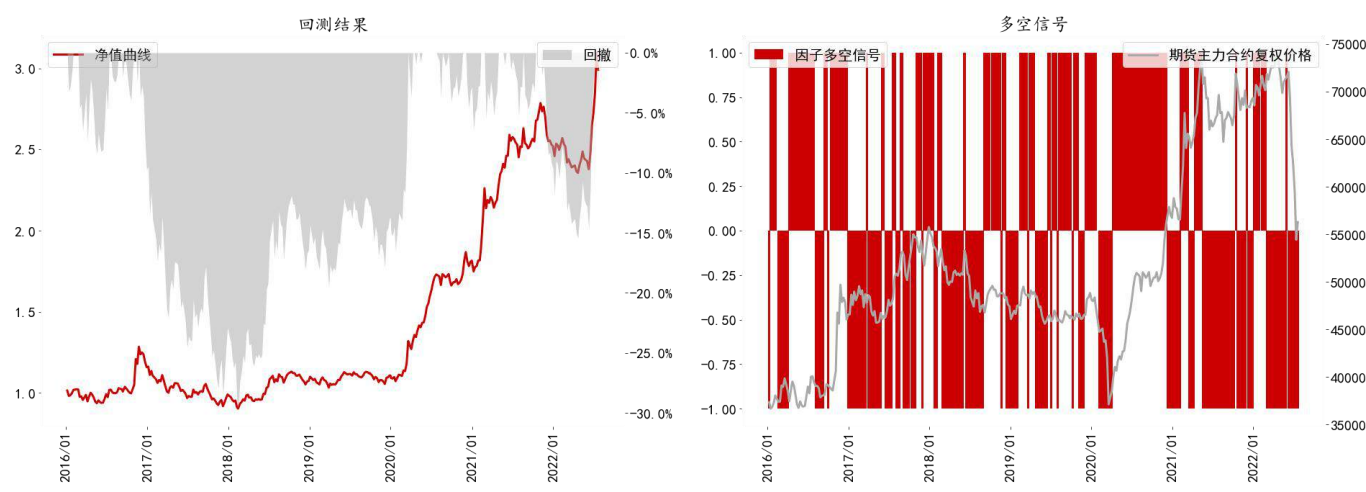
资料来源: 东证衍生品研究院

图表 27: 聚类 cluster3, 阈值为 0% 回撤曲线



资料来源: 东证衍生品研究院

图表 28: 聚类 cluster4, 阈值为 0% 回撤曲线



资料来源: 东证衍生品研究院

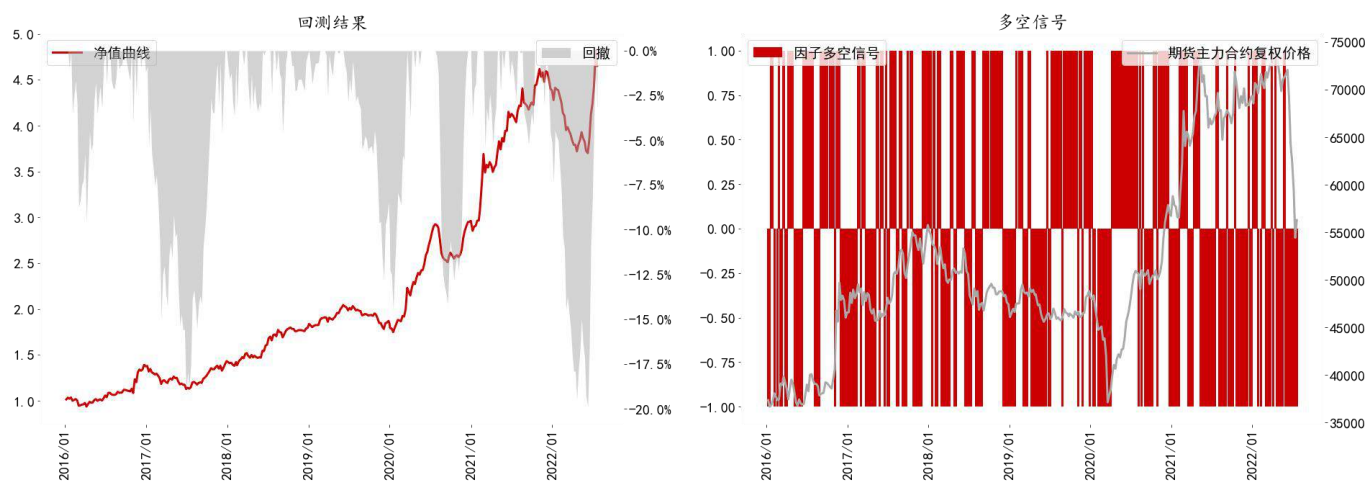
同样尝试在不同阈值下, 全部大类聚类降维后的回测结果。数据显示, 聚类降维的方式相较于直接降维表现出明显的提升, 在个别阈值下 (0.04%), 策略的年化收益达到了 30%, 相应地, 夏普值也达到了 0.75 以上。

图表 29：聚类降维不同阈值下回测指标

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	365.44%	25.94%	39.58%	0.59	-19.88%	1.31	0.54	1.48	1.35	18.25
0.20%	401.00%	27.34%	38.53%	0.65	-20.15%	1.36	0.52	1.67	1.57	14.75
0.40%	479.36%	30.15%	36.65%	0.76	-13.24%	2.28	0.52	1.85	2.02	12.45
0.60%	457.14%	29.39%	35.26%	0.77	-13.41%	2.19	0.57	1.65	2.16	12.35
0.80%	297.61%	23.00%	32.28%	0.64	-13.41%	1.71	0.62	1.28	1.73	12.75
1.00%	238.09%	20.05%	31.80%	0.55	-13.41%	1.49	0.65	1.10	1.48	13.25
1.20%	215.66%	18.82%	31.18%	0.53	-14.88%	1.26	0.70	0.89	1.39	15.25
1.50%	165.86%	15.80%	29.46%	0.45	-11.06%	1.43	0.74	0.74	1.30	16.95
2.00%	129.39%	13.26%	26.58%	0.41	-10.26%	1.29	0.80	0.59	1.33	18.90
3.00%	50.78%	6.35%	20.47%	0.19	-10.40%	0.61	0.89	0.29	0.71	34.30
4.00%	24.70%	3.37%	15.78%	0.06	-5.52%	0.61	0.93	0.16	0.21	48.00
5.00%	7.48%	1.09%	12.97%	-0.10	-7.17%	0.15	0.96	0.08	-0.49	93.35

资料来源：东证衍生品研究院

图表 30：聚类降维，阈值为 0%回测曲线



资料来源：东证衍生品研究院

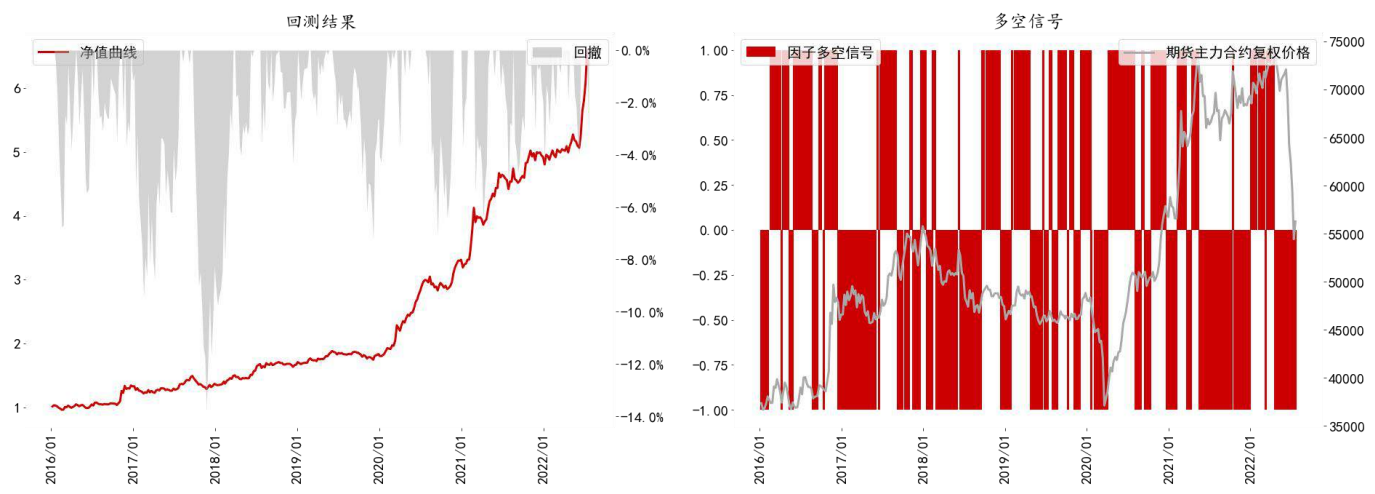
此外，尝试另外一种结合各大类信号的方式，即采用等权重的方式进行结合。将五大类信号单独导出，再根据等权重的方式合成综合信号，从结果来看，对于等权合成信号，阈值越小信号表现越好。整体来看，等权重合成的方式相较于全量降维，各回测指标均有小幅提升。

图表 31：等权合成信号回测指标

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	537.00%	32.01%	37.76%	0.78	-13.75%	2.33	0.55	1.67	2.01	32.30
0.40%	373.82%	26.28%	32.41%	0.74	-8.10%	3.25	0.70	1.10	2.35	21.00
1.00%	128.18%	13.17%	19.53%	0.55	-4.23%	3.11	0.94	0.34	2.44	67.20

资料来源：东证衍生品研究院

图表 32：等权合成信号回测曲线



资料来源：东证衍生品研究院

4、多品种策略

在上文中，铜期货的回测结果表明聚类降维的方式相较于直接降维能够起到一定的提升，然而模型有效性的验证显然在一个品种上是远远不够的。在本章节，基于期货市场关注度较高的几大品种运用上述方式进行进一步验证。期货品种的选择不仅考虑到市场关注度，同时也覆盖了各大期货分类，具体为：有色金属（镍、锌）；黑色系（螺纹钢、铁矿石、焦炭）；能源化工（PTA、PP）；农产品（豆粕）。

此外，在本章节最后，基于各个期货品种的预测数据，构建横截面上的多空策略，目的是提升策略整体的稳健性。

4.1、豆粕

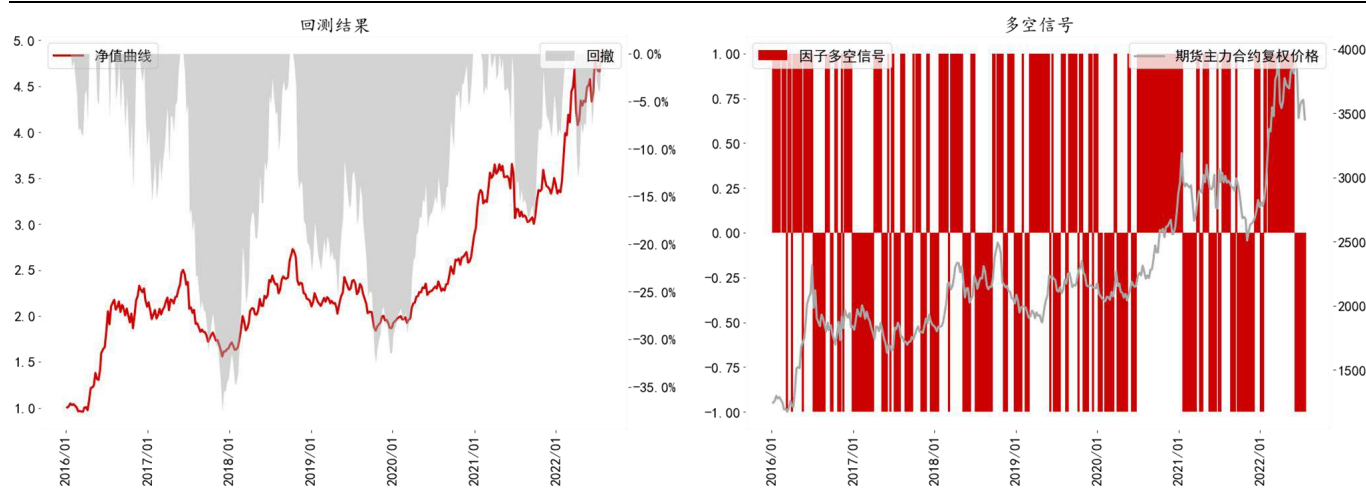
在阈值为 0 时，豆粕的回测结果显示策略表现为：年化收益 26.74%，年化波动 59.90%，夏普值 0.41，最大回撤 -37.63%。整体来看，收益率较为可观，然而策略波动率较高，相应的最大回撤也较大。

图表 33：聚类降维各阈值下回测指标（豆粕）

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
0.00%	385.34%	26.74%	59.90%	0.41	-37.63%	0.71	0.57	1.12	0.78	23.00
0.20%	475.39%	30.02%	57.51%	0.48	-33.62%	0.89	0.56	1.26	0.96	16.95
0.60%	378.08%	26.45%	53.64%	0.45	-32.38%	0.82	0.56	1.23	0.93	13.00
1.00%	426.58%	28.30%	49.94%	0.52	-19.86%	1.42	0.64	1.01	1.08	13.90
1.20%	536.21%	31.99%	48.07%	0.62	-17.55%	1.82	0.69	0.93	1.35	15.00
1.50%	522.91%	31.57%	45.20%	0.65	-17.46%	1.81	0.74	0.81	1.56	16.45
2.00%	508.16%	31.10%	39.96%	0.72	-10.22%	3.04	0.83	0.60	2.18	24.35
5.00%	45.84%	5.82%	20.90%	0.16	-13.11%	0.44	0.97	0.09	0.54	120.00

资料来源：东证衍生品研究院

图表 34: 聚类降维, 阈值为 0% 回测曲线 (豆粕)



资料来源: 东证衍生品研究院

4.2、PP

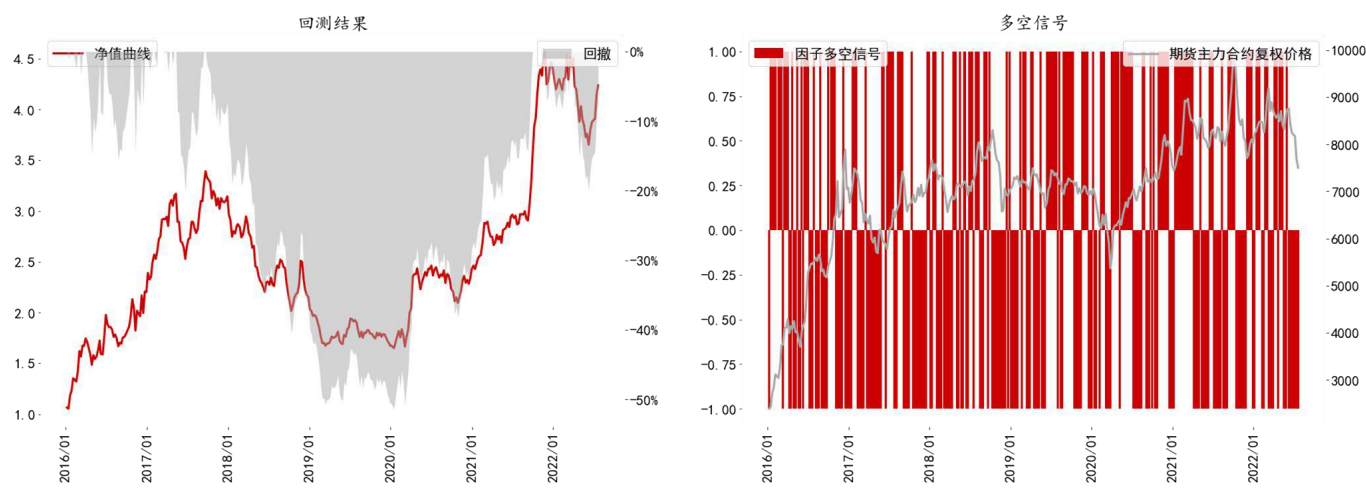
在阈值为 0 时, PP 的回测结果显示策略表现为: 年化收益 24.19%, 年化波动 61.85%, 夏普值 0.35, 最大回撤-51.36%。PP 策略在 2018 年至 2019 年这两年时间里面出现了较长周期的回撤, 且回撤幅度较大, 达到了-51.36%。

图表 35: 聚类降维各阈值下回测指标 (PP)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
0.00%	323.82%	24.19%	61.85%	0.35	-51.36%	0.47	0.51	1.35	0.72	15.25
0.20%	260.34%	21.20%	61.01%	0.31	-49.90%	0.42	0.50	1.39	0.64	13.00
0.60%	445.72%	28.99%	56.76%	0.47	-46.28%	0.63	0.54	1.34	1.07	10.90
1.00%	345.10%	25.10%	54.60%	0.42	-45.14%	0.56	0.57	1.18	0.96	11.00
1.20%	389.59%	26.90%	53.06%	0.46	-37.59%	0.72	0.60	1.12	1.09	11.65
1.50%	480.40%	30.18%	50.25%	0.55	-31.10%	0.97	0.66	0.99	1.37	13.10
2.00%	331.18%	24.51%	47.09%	0.47	-25.83%	0.95	0.71	0.78	1.17	14.10
5.00%	53.02%	6.59%	32.21%	0.13	-14.41%	0.46	0.90	0.19	0.30	42.00

资料来源: 东证衍生品研究院

图表 36: 聚类降维, 阈值为 0% 回测曲线 (PP)



资料来源: 东证衍生品研究院

4.3、PTA

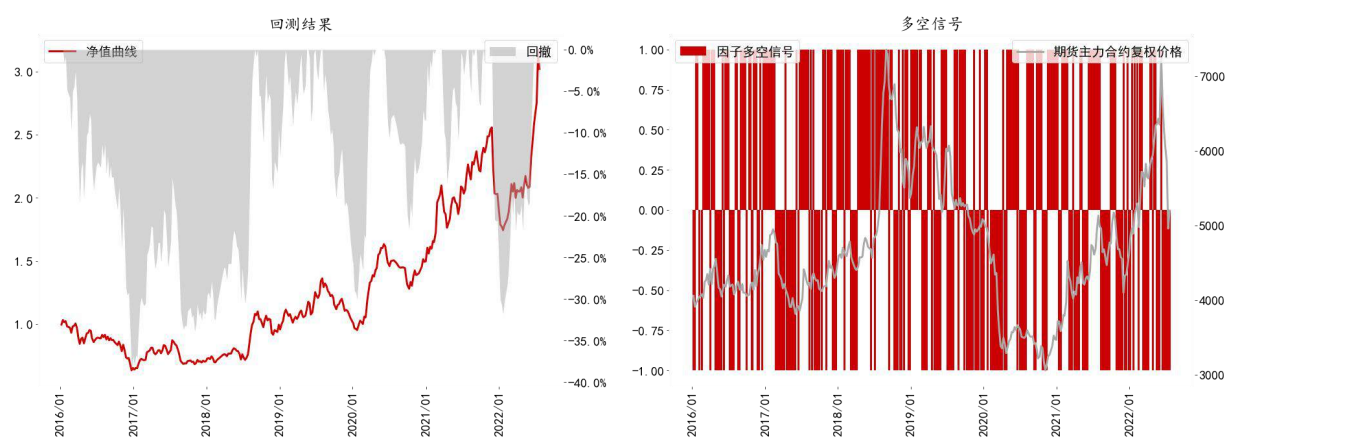
在阈值为 0 时, PTA 的回测结果显示策略表现为: 年化收益 18.04%, 年化波动 60.84%, 夏普值 0.26, 最大回撤-38.59%。PTA 策略整体收益情况不及 PP, 但在风险控制层面表现较好, 最大回撤为-38.59%。

图表 37: 聚类降维各阈值下回测指标 (PTA)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
0.00%	202.07%	18.04%	60.84%	0.26	-38.59%	0.47	0.52	1.23	0.49	15.00
0.20%	252.48%	20.80%	59.14%	0.31	-36.04%	0.58	0.52	1.28	0.61	12.65
0.60%	356.40%	25.57%	54.88%	0.42	-23.21%	1.10	0.53	1.36	0.94	11.75
1.00%	390.13%	26.92%	51.34%	0.48	-20.64%	1.30	0.57	1.20	1.11	11.00
1.20%	343.65%	25.04%	50.51%	0.45	-19.73%	1.27	0.59	1.11	1.04	11.30
1.50%	255.30%	20.95%	45.58%	0.41	-19.58%	1.07	0.65	0.88	0.95	12.10
2.00%	123.44%	12.82%	40.80%	0.26	-23.85%	0.54	0.72	0.59	0.56	14.35
5.00%	91.73%	10.26%	24.91%	0.32	-5.43%	1.89	0.96	0.17	1.23	67.20

资料来源: 东证衍生品研究院

图表 38: 聚类降维, 阈值为 0% 回测曲线 (PTA)



资料来源: 东证衍生品研究院

4.4、镍

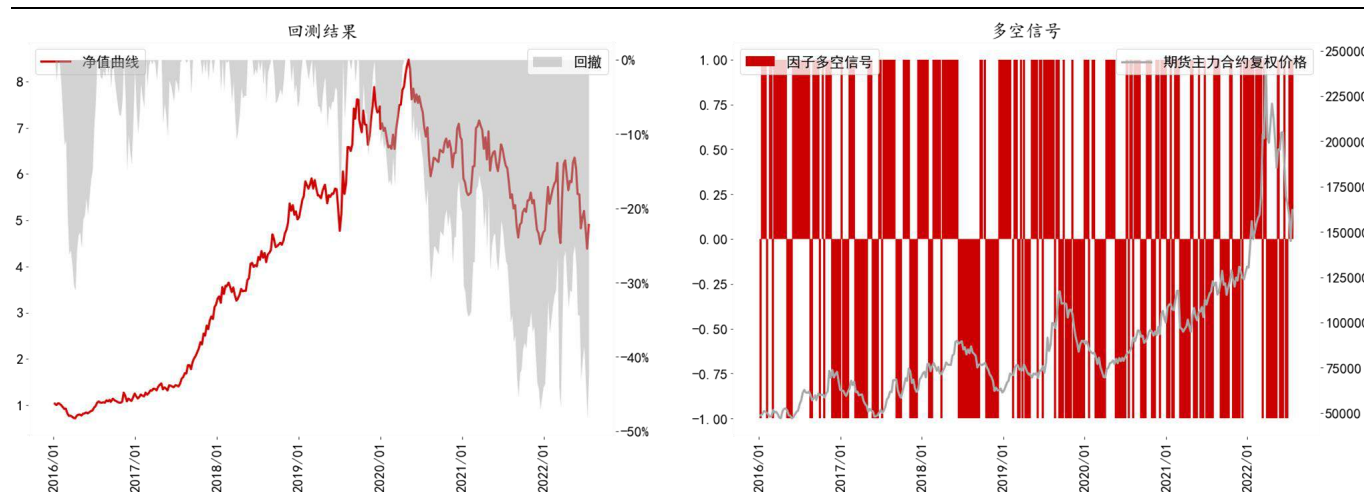
在阈值为0时，镍的回测结果显示策略表现为：年化收益26.92%，年化波动81.54%，夏普值0.3，最大回撤-48.26%。在2020年之前镍的回测结果较为理想，然而在2020年之后策略净值出现较为明显下滑，策略整体的年化波动率达到了81.54%，表明该策略在镍上的风险较高。

图表 39：聚类降维各阈值下回测指标（镍）

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	389.95%	26.92%	81.54%	0.3	-48.26%	0.56	0.54	1.14	0.55	17.30
0.20%	446.07%	29.00%	79.16%	0.34	-37.79%	0.77	0.54	1.19	0.62	14.10
0.60%	326.60%	24.31%	76.98%	0.28	-40.00%	0.61	0.53	1.21	0.54	12.15
1.00%	255.22%	20.94%	73.21%	0.25	-36.43%	0.57	0.54	1.14	0.48	11.20
1.20%	365.65%	25.95%	69.70%	0.34	-38.12%	0.68	0.55	1.15	0.71	11.20
1.50%	298.15%	23.03%	67.37%	0.31	-36.01%	0.64	0.56	1.12	0.65	10.65
2.00%	263.71%	21.37%	60.18%	0.32	-31.05%	0.69	0.64	0.83	0.64	11.85
5.00%	11.14%	1.60%	34.49%	-0.02	-22.98%	0.07	0.88	0.17	-0.05	29.45

资料来源：东证衍生品研究院

图表 40：聚类降维，阈值为0%回测曲线（镍）



资料来源：东证衍生品研究院

4.5、锌

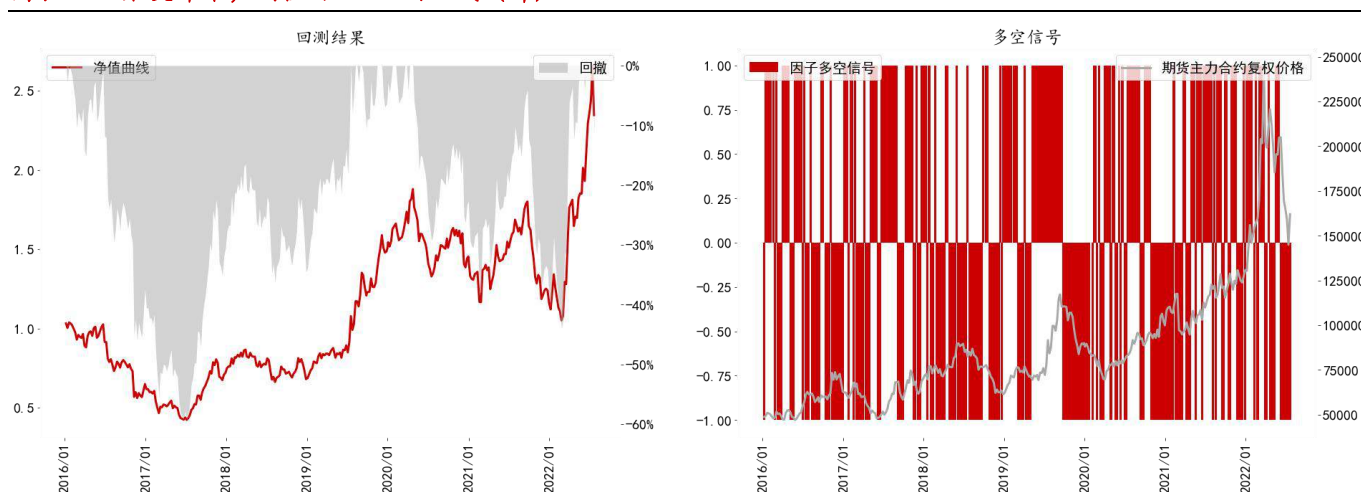
在阈值为0时，锌的回测结果显示策略表现为：年化收益13.66%，年化波动82.64%，夏普值0.14，最大回撤-59.31%。锌的策略虽然取得了13.66%的正收益，但是波动较为剧烈，最大回撤接近60%，导致策略的夏普值只有0.14。

图表 41: 聚类降维各阈值下回测指标 (锌)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	134.75%	13.66%	82.64%	0.14	-59.31%	0.23	0.53	1.08	0.24	16.80
0.20%	276.28%	21.99%	78.90%	0.25	-53.97%	0.41	0.53	1.19	0.46	14.10
0.60%	577.38%	33.24%	72.56%	0.42	-36.30%	0.92	0.53	1.31	0.90	12.65
1.00%	726.16%	37.26%	70.02%	0.5	-21.69%	1.72	0.55	1.29	1.12	12.00
1.20%	588.06%	33.55%	67.90%	0.46	-30.16%	1.11	0.57	1.19	1.06	11.60
1.50%	634.51%	34.87%	65.70%	0.49	-27.34%	1.28	0.58	1.18	1.21	11.05
2.00%	335.57%	24.70%	60.80%	0.37	-29.10%	0.85	0.62	0.94	0.89	11.15
5.00%	67.65%	8.06%	38.38%	0.15	-15.56%	0.52	0.88	0.23	0.39	28.00

资料来源: 东证衍生品研究院

图表 42: 聚类降维, 阈值为 0%回测曲线 (锌)



资料来源: 东证衍生品研究院

4.6、螺纹钢

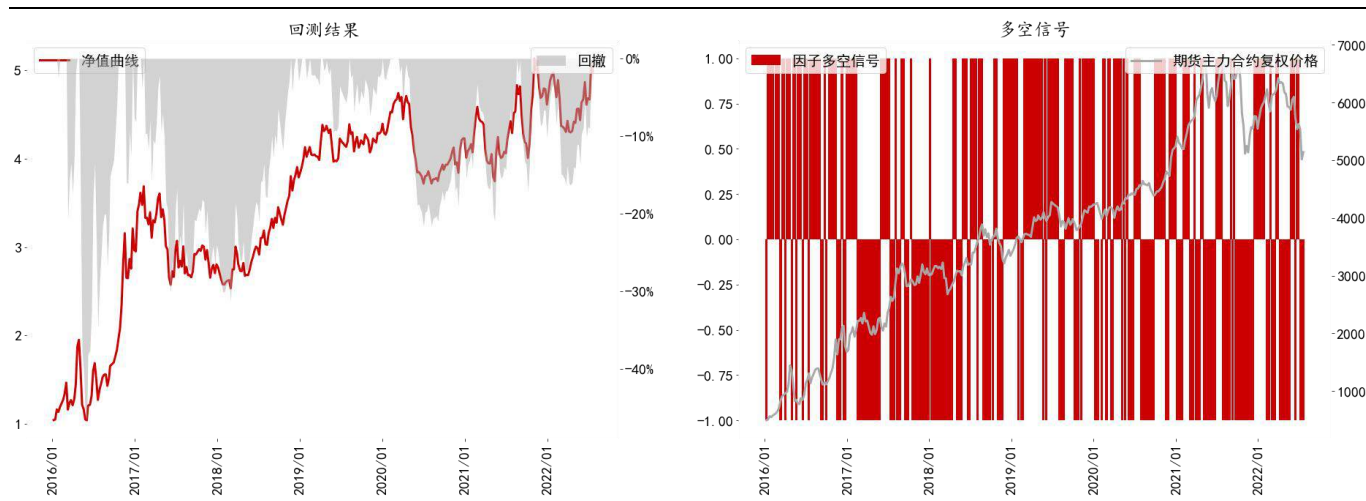
在阈值为 0 时, 螺纹钢的回测结果显示策略表现为: 年化收益 27.24%, 年化波动 85.31%, 夏普值 0.29, 最大回撤 -46.62%。螺纹钢的大幅回撤主要发生在 2016 年初的价格剧烈波动, 改策略在近些年表现较为稳定。

图表 43: 聚类降维各阈值下回测指标 (螺纹钢)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	398.24%	27.24%	85.31%	0.29	-46.62%	0.58	0.56	1.12	0.51	17.85
0.20%	447.70%	29.06%	83.70%	0.32	-46.62%	0.62	0.54	1.23	0.57	14.60
0.60%	328.08%	24.37%	80.88%	0.27	-46.62%	0.52	0.51	1.37	0.50	11.20
1.00%	483.55%	30.29%	77.43%	0.36	-46.62%	0.65	0.55	1.27	0.69	10.90
1.20%	407.10%	27.57%	76.70%	0.33	-46.62%	0.59	0.57	1.14	0.63	11.15
1.50%	570.61%	33.04%	75.50%	0.41	-46.62%	0.71	0.63	1.00	0.78	12.00
2.00%	688.31%	36.30%	70.65%	0.48	-28.76%	1.26	0.67	0.91	1.05	13.00
5.00%	181.56%	16.80%	46.33%	0.31	-19.25%	0.87	0.86	0.34	0.62	26.25

资料来源: 东证衍生品研究院

图表 44: 聚类降维, 阈值为 0% 回测曲线 (螺纹钢)



资料来源: 东证衍生品研究院

4.7、铁矿石

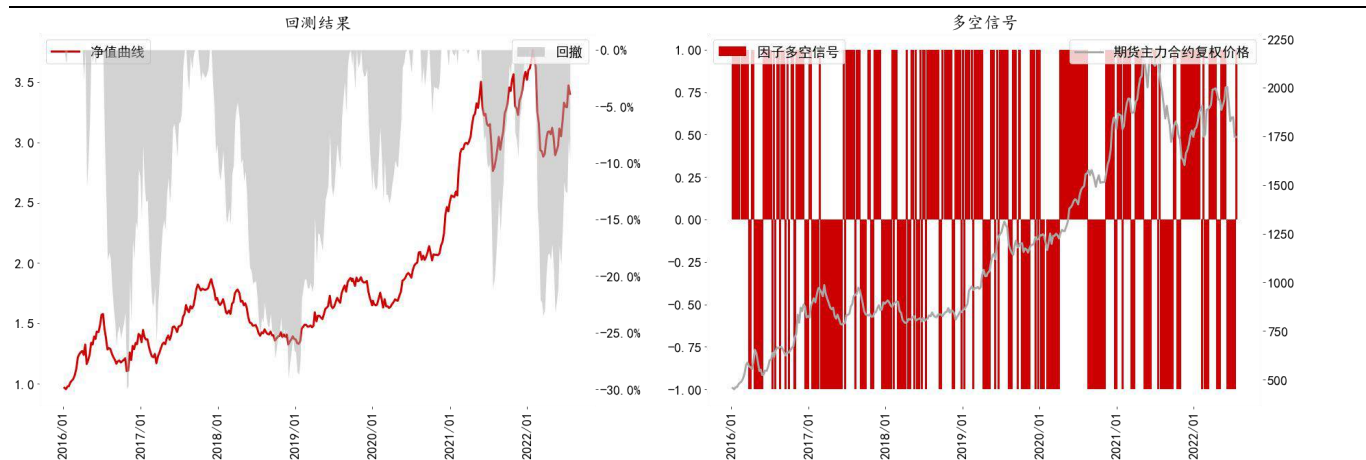
在阈值为 0 时, 铁矿石的回测结果显示策略表现为: 年化收益 20.16%, 年化波动 50.00%, 夏普值 0.36, 最大回撤 -30.00%。同为黑色系的铁矿石相较于螺纹钢表现更好, 在阈值为 0 的情况下, 胜率达到了 0.58。

图表 45: 聚类降维各阈值下回测指标 (铁矿石)

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率	平均持仓时间
0.00%	240.15%	20.16%	50.00%	0.36	-30.00%	0.67	0.58	1.02	0.62	17.15
0.20%	248.67%	20.60%	47.98%	0.38	-28.04%	0.73	0.55	1.20	0.69	13.25
0.60%	127.72%	13.14%	43.97%	0.24	-29.35%	0.45	0.54	1.12	0.45	10.75
1.00%	78.11%	9.04%	38.31%	0.17	-26.29%	0.34	0.58	0.92	0.31	10.55
1.20%	86.87%	9.83%	36.89%	0.2	-21.70%	0.45	0.63	0.80	0.36	11.30
1.50%	43.15%	5.53%	31.46%	0.1	-23.39%	0.24	0.65	0.65	0.19	11.50
2.00%	39.30%	5.10%	25.09%	0.11	-19.00%	0.27	0.74	0.47	0.19	14.85
5.00%	16.42%	2.31%	11.22%	-0.01	-5.63%	0.41	0.98	0.06	-0.02	168.00

资料来源: 东证衍生品研究院

图表 46: 聚类降维, 阈值为 0% 回测曲线 (铁矿石)



资料来源: 东证衍生品研究院

4.8、焦炭

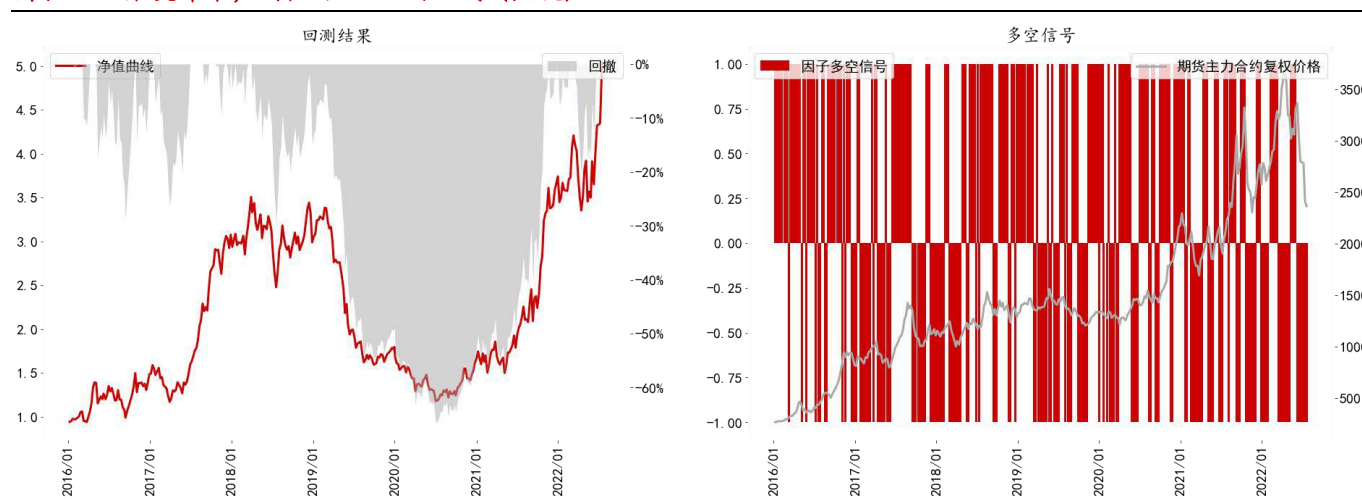
在阈值为0时，焦炭的回测结果显示策略表现为：年化收益27.40%，年化波动84.74%，夏普值0.330，最大回撤-66.40%。焦炭策略在2020年中出现幅度较大的回撤，最大回撤达到了-66.40%，风险程度较高，于2021年开始逐渐修复该回撤。

图表 47：聚类降维各阈值下回测指标（焦炭）

开仓阈值	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率	平均持仓时间
0.00%	402.46%	27.40%	84.74%	0.3	-66.40%	0.41	0.58	0.97	0.52	18.05
0.20%	370.66%	26.16%	83.26%	0.29	-66.92%	0.39	0.56	1.06	0.51	14.60
0.60%	327.29%	24.34%	78.70%	0.28	-67.90%	0.36	0.56	1.04	0.51	12.80
1.00%	208.43%	18.41%	75.53%	0.21	-63.28%	0.29	0.59	0.90	0.38	12.25
1.20%	391.86%	26.99%	73.04%	0.34	-51.39%	0.53	0.62	0.87	0.63	12.75
1.50%	193.63%	17.54%	69.09%	0.22	-59.24%	0.30	0.65	0.72	0.40	12.00
2.00%	177.71%	16.56%	63.96%	0.22	-55.47%	0.30	0.68	0.62	0.41	12.45
5.00%	173.44%	16.29%	36.02%	0.38	-24.69%	0.66	0.91	0.23	1.04	38.20

资料来源：东证衍生品研究院

图表 48：聚类降维，阈值为0%回测曲线（焦炭）



资料来源：东证衍生品研究院

在回测了几大主流品种之后，其回测结果显示出一定的共性：策略可以提供显著的正向收益，然而整体波动较大，最大回撤也较高；胜率均高于50%，范围在52%到56%不等，夏普值0.14到0.59不等。

数据结果表明，该策略对于单品种而言能够提供具有预测能力的信号，只需将策略的波动率进行控制即可，一个显而易见可以提高策略稳定性的方式就是纳入多个品种去构建横截面的多空策略。

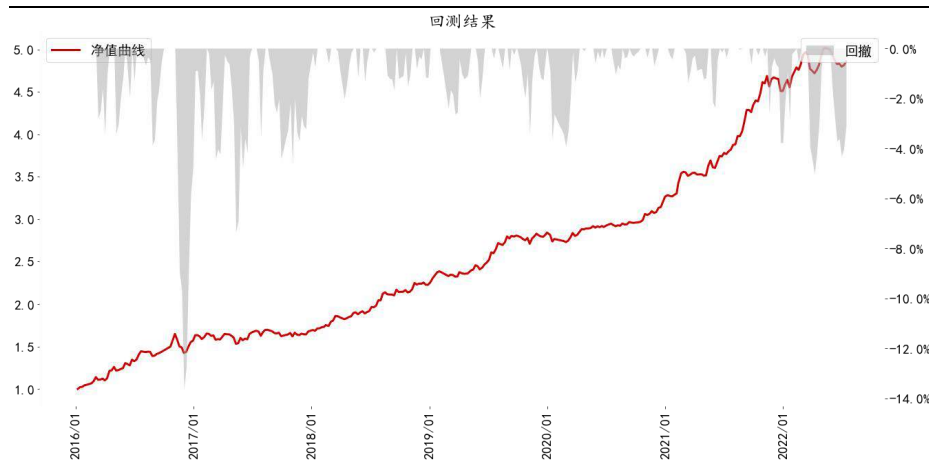
5、多品种横截面多空模型

5.1、各品种等权构建

基于上述期货品种的回测结果，尝试构建横截面多空策略，策略共包括九大期货品种（铜、镍、锌、螺纹钢、铁矿石、焦炭、PTA、PP、豆粕），构建逻辑为在每个换仓周期，做多3个预测收益率最高的品种，做空3个预测收益率最低品种，排名中间的3个进行空仓。本报告不重点研究仓位的权重分配策略，为简单起见，将做多与做空品种进行等权处理。此外，与前述对单品种的研究不同的是，由于多品种的多空机制的问题，不设置阈值参数。

下面展示了多品种策略的净值曲线以及回测指标。从净值曲线上来看，整体收益较为平滑，并未出现明显的回撤阶段；具体到回测指标，策略整体的年化收益达到26.85%，年化波动25.82%，夏普值0.95，最大回撤-13.62%，胜率0.61，由于回测阶段未考虑到交易滑点，真实年化收益率应当略低于该值。

图表 49：多品种横截面策略回测曲线



资料来源：东证衍生品研究院

图表 50：多品种横截面策略回测指标

总收益	385.86%
年化收益	26.85%
年化波动	11.54%
夏普值	2.12
最大回撤	-13.62%
收益风险比	1.97
胜率	0.61
盈亏比	1.46
sortino 比率	1.95

资料来源：东证衍生品研究院

多品种策略的净值曲线显著好于单品种策略，由于该策略对于在品种做多上的仓位等于做空的仓位，故该策略为中性化策略，规避了期货市场的beta风险，这一点从收益曲线可以看出，该策略与整个商品市场的整体涨跌关系较低。由于策略考虑了九大品种的交易信号，故在单品种上出现的较大回撤某种程度上被策略平滑掉了。

图表 51：多品种横截面策略各品种收益

品种	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino比率
铜	-58.45%	-12.38%	26.77%	-0.55	-59.81%	-0.21	0.51	0.53	-0.70
镍	286.98%	22.58%	70.22%	0.29	-50.90%	0.44	0.49	1.46	0.54
锌	-55.56%	-11.49%	49.15%	-0.28	-69.57%	-0.17	0.44	1.04	-0.43
螺纹钢	1254.98%	48.01%	80.72%	0.57	-35.18%	1.36	0.53	1.65	1.12
铁矿石	159.39%	15.42%	36.58%	0.36	-14.57%	1.06	0.53	1.48	0.68
焦炭	752.61%	38.05%	71.69%	0.50	-33.68%	1.13	0.57	1.21	0.88
PP	369.58%	26.20%	53.36%	0.45	-35.57%	0.74	0.53	1.48	0.89
PTA	393.97%	27.16%	51.81%	0.48	-37.94%	0.72	0.50	1.70	0.95
豆粕	369.36%	26.19%	52.66%	0.45	-27.75%	0.94	0.53	1.52	0.87

资料来源：东证衍生品研究院

5.2、基于波动率对信号强度调整

根据收益率排序进行品种多空选择可能存在问题：对于波动率比较高的品种，被选择的概率更高；而波动率较低品种，被选择的概率更低，导致空仓概率过高。下图统计各品种的历史长期波动率和在策略组合中的空仓比例（空仓次数/总开仓次数）。结果表明，波动率较低品种（如铜、铁矿石）空仓比例较高，以铜为例，其空仓比例达到了54.17%，而具有最高波动率的螺纹钢，其空仓比例只有27.38%。

图表 52：不同品种空仓次数

	cu	ni	zn	rb	i	j	ta	pp	m
做空次数	76	130	142	127	80	121	104	128	100
做多次数	78	129	121	117	107	120	124	95	117
空仓次数	182	77	73	92	149	95	108	113	119
波动率	15.83%	33.12%	24.86%	33.88%	18.77%	28.01%	21.89%	28.12%	22.75%
空仓比例	54.17%	22.92%	21.73%	27.38%	44.35%	28.27%	32.14%	33.63%	35.42%

资料来源：东证衍生品研究院

为了尽可能均衡各品种的空仓比例，尝试基于各品种一段时间内的价格波动率对预测信号的强度进行调整，即将每一期的预测收益率除以当期之前一段时间内的价格波动率，再基于调整后的收益率进行回测。下图分别展示了基于10天、20天、30天、60天、90天内价格波动率调整后各品种的空仓比例（NA表示未调整），数据表明在调整后，各品种空仓比例仍然存在较大差异，原因在于该调整方式是利用实际历史波动率对预测收益率进行调整，而在该模型当中预测值的波动与真实值的波动有较大差异，导致调整后仍然有较多品种空仓时期过多。因此，在最后一列加入了基于预测收益率序列的波动率进行调整，以此作为参照可以发现，该种调整方式可以有效均衡各品种的空仓时间。

图表 53: 信号强度调整后各品种空仓比例

波动率计算周期(天)	cu	ni	zn	rb	i	j	ta	pp	m
NA	54.17%	22.92%	21.73%	27.38%	44.35%	28.27%	32.14%	33.63%	35.42%
10	43.15%	52.08%	46.13%	23.51%	11.61%	23.81%	33.63%	44.35%	21.73%
20	42.56%	52.08%	45.54%	21.73%	11.61%	27.08%	34.23%	46.43%	18.75%
30	42.56%	52.08%	45.54%	21.73%	11.61%	27.08%	34.23%	46.43%	18.75%
60	43.45%	53.27%	48.81%	21.73%	11.61%	23.81%	33.33%	43.75%	20.24%
90	40.18%	52.98%	48.81%	22.62%	12.50%	25.00%	31.85%	45.83%	20.24%
预测收益率序列	30.95%	27.08%	37.50%	44.64%	31.55%	33.33%	33.63%	35.71%	25.60%

资料来源: 东证衍生品研究院

调整后, 策略的表现如下图: 基于预测收益率序列的波动率进行调整后策略整体表现反而更差, 然而年化波动率是所有策略中最低的。分析原因, 在于模型对于波动率较高的品种预测能力偏强, 波动率较高的品种其预测信号的强度往往更加显著, 模型在这些品种上的表现也更高。故均衡空仓时间之后, 策略整体收益降低, 但整体波动率也进一步降低。

图表 54: 信号强度调整后策略表现

波动率计算周期(天)	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率
NA	387.26%	26.90%	11.60%	2.10	-13.61%	1.98	0.61	1.51	1.94
10	499.60%	30.93%	11.11%	2.57	-13.12%	2.36	0.68	1.29	5.48
20	609.01%	34.27%	10.76%	2.96	-11.97%	2.86	0.67	1.48	7.03
30	540.57%	32.23%	10.69%	2.79	-12.06%	2.67	0.67	1.44	6.36
60	501.59%	30.99%	10.15%	2.82	-10.48%	2.96	0.68	1.37	6.45
90	455.48%	29.43%	10.39%	2.60	-11.60%	2.54	0.65	1.47	5.74
预测收益率序列	250.21%	20.75%	9.70%	1.89	-12.52%	1.66	0.61	1.44	4.09

资料来源: 东证衍生品研究院

5.3、基于波动率对权重进行调整

另外一种基于波动率调整的方式是摒弃横截面等权的思路, 直接基于波动率构建各品种的权重矩阵, 根据该矩阵进行回测, 这样一来也就不存在某些空仓时间过多的问题, 任何时刻每个品种都有一定的仓位。回测结果如下图, 同样展示了基于 10 天、20 天、30 天、60 天、90 天内价格波动率调整后的策略表现。该调整对参数敏感性较强, 且波动率计算周期越短, 策略整体表现较好, 但相应的年化波动率也越高。

图表 55: 权重调整后策略表现

	总收益	年化收益	年化波动	夏普值	最大回撤	收益风险比	胜率	盈亏比	sortino 比率
NA	387.26%	26.90%	11.60%	2.10	-13.61%	1.98	0.61	1.51	1.94
10	858.74%	40.36%	17.23%	2.20	-12.55%	3.22	0.62	1.40	4.69
20	508.27%	31.10%	15.89%	1.81	-10.78%	2.89	0.60	1.37	3.84
30	397.51%	27.21%	15.16%	1.64	-9.49%	2.87	0.59	1.33	3.49
60	402.13%	27.39%	14.70%	1.70	-10.17%	2.69	0.58	1.39	3.64
90	335.44%	24.69%	14.32%	1.56	-10.95%	2.26	0.58	1.33	3.33

资料来源: 东证衍生品研究院

6、结论

本文提出了一种新的期货收益率预测模型中基本面数据处理的方法：通过DTW+KMeans的方式对数据进行分类，再基于分类数据对期货收益率进行预测。相较于全量降维的方式，回测结果显示该模型对于策略整体的预测能力有一定提升。

此外，为提升策略的稳健性，可以通过设置相应的开仓阈值，阈值的提高能够提升策略的胜率，然而同时也会降低其收益率和盈亏比，实证表明，阈值设定为1%是一个较为有效的参数。

在选择基本面数据的过程中，我们建议对数据进行一定的筛选。有效的筛选不仅可以降低模型的复杂性，还可以避免过多的无效、冗余数据，为后续的建模提供坚实的基础。

模型的有效性在多个品种的回测上均得到验证，结果表明，对于多数品种，该策略均能带来显著正收益，唯一需要担忧的是在一些市场波动剧烈时段，单品种的回撤风险较大。故可以构建多品种横截面策略以降低策略回撤，与预期的一样，多策略模型相较于单品种模型具有更高的稳健性。在策略中，尽可能地纳入更多的期货品种可以带来更为平滑的收益曲线。

在报告最后考虑了基于波动率对仓位进行调整：若对信号强度进行调整，策略整体并无太大提升，原因在于策略对波动率较高品种的预测能力更强，而调整之后波动率较高的品种持仓次数降低；若对权重进行调整，策略表现对参数敏感性较大，波动率计算周期越短，策略收益越高，年化波动也相应增高。

期货走势评级体系（以收盘价的变动幅度为判断标准）

走势评级	短期（1-3个月）	中期（3-6个月）	长期（6-12个月）
强烈看涨	上涨 15%以上	上涨 15%以上	上涨 15%以上
看涨	上涨 5-15%	上涨 5-15%	上涨 5-15%
震荡	振幅-5%-+5%	振幅-5%-+5%	振幅-5%-+5%
看跌	下跌 5-15%	下跌 5-15%	下跌 5-15%
强烈看跌	下跌 15%以上	下跌 15%以上	下跌 15%以上

上海东证期货有限公司

上海东证期货有限公司成立于2008年，是一家经中国证券监督管理委员会批准的经营期货业务的综合性公司。东证期货是东方证券股份有限公司全资子公司，注册资本金23亿元人民币，员工近600人。公司主要从事商品期货经纪、金融期货经纪、期货投资咨询、资产管理、基金销售等业务，拥有上海期货交易所、大连商品交易所、郑州商品交易所和上海国际能源交易中心会员资格，是中国金融期货交易所全面结算会员。公司拥有东证润和资本管理有限公司，上海东祺投资管理有限公司和东证期货国际（新加坡）私人有限公司三家全资子公司。

东证期货以上海为总部所在地，在大连、长沙、北京、上海、郑州、太原、常州、广州、青岛、宁波、深圳、杭州、西安、厦门、成都、东营、天津、哈尔滨、南宁、重庆、苏州、南通、泉州、汕头、沈阳、无锡、济南等地共设有33家营业部，并在北京、上海、广州、深圳多个经济发达地区拥有134个证券IB分支机构，未来东证期货将形成立足上海、辐射全国的经营网络。

自2008年成立以来，东证期货秉承稳健经营、创新发展的宗旨，坚持市场化、国际化、集团化的发展道路，打造以衍生品风险管理为核心，具有研究和技术两大核心竞争力，为客户提供综合财富管理平台的一流衍生品服务商。

分析师承诺

王冬黎、谢怡伦

本人具有中国期货业协会授予的期货执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

免责声明

本报告由上海东证期货有限公司（以下简称“本公司”）制作及发布。

本研究报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本研究报告是基于本公司认为可靠的且目前已公开的信息撰写，本公司力求但不保证该信息的准确性和完整性，客户也不应该认为该信息是准确和完整的。同时，本公司不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司会适时更新我们的研究，但可能会因某些规定而无法做到。除了一些定期出版的报告之外，绝大多数研究报告是在分析师认为适当的时候不定期地发布。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买投资标的的邀请或向人作出邀请。

在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

本报告主要以电子版形式分发，间或也会辅以印刷品形式分发，所有报告版权均归本公司所有。未经本公司事先书面授权，任何机构或个人不得以任何形式复制、转发或公开传播本报告的全部或部分内容，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。

如需引用、刊发或转载本报告，需注明出处为东证衍生品研究院，且不得对本报告进行任何有悖原意的引用、删节和修改。

东证衍生品研究院

地址：上海市中山南路318号东方国际金融广场2号楼21楼

联系人：梁爽

电话：8621-63325888-1592

传真：8621-33315862

网址：www.orientfutures.com

Email：research@orientfutures.com