



## 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

报告日期：2021 年 11 月 25 日

### ★基本面数据整理：

基于螺纹钢上下游产业链关系，共选取 4 大类，共 121 个基本面数据，包括供给类，需求类，库存类以及宏观类。对原始数据进行去极值，标准化，差分，数据可得性调整，频率调整等处理，作为模型的输入因子。

### ★降维与模型选择：

**PCA 降维：**对于数据进行主成分降维，以可解释性方差为依据截取 95%的信息度，将 121 个因子降维至 30 个因子；

**OLS 多元回归：**利用普通最小二乘法进行多元回归，最优拟合曲线应该使各点到直线的距离的平方和（残差平方和 RSS）最小；

**XGBoost 模型：**基于传统 GBDT 模型，将目标函数泰勒展开到了二阶，并且加入了叶子权重的 L2 正则化项；

**RNN (LSTM) 模型：**具有记忆功能的循环神经网络，当前隐藏层的状态受到之前隐藏层状态的影响。

### ★模型结果：

综合模型回测结果达到 2017 年至今累计收益率 233.53%，年化收益率 29.74%，年化波动率 13.16%，夏普值 2.08，最大回撤-10.07%，胜率 56.05%，平均持仓时间 14.22 天。

### ★致谢：

感谢东方证券金融工程首席分析师朱剑涛老师的指导与帮助。

### ★风险提示：

市场风格的变换会造成特征有效性变化，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](mailto: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](mailto:yilun.xie@orientfutures.com)

### 相关报告

《衍生品量化择时系列专题之二：螺纹钢指标筛选与大类因子研究》

## 目录

1、 基本面因子选取 .....	5
1.1、 商品基本面量化原理 .....	5
1.2、 螺纹钢基本面因子选取 .....	5
2、 降维及模型选择 .....	7
3、 数据处理 .....	8
4、 单因子预测能力分析 & 大类因子合成 .....	9
5、 OLS 多元回归 .....	10
6、 XGBoost 模型预测 .....	12
7、 RNN 模型预测 .....	14
8、 模型合成 .....	15
9、 总结及展望 .....	18
10、 风险提示 .....	18
11、 附录：部分因子回测结果展示 .....	18

## 图表目录

图表 1 螺纹钢上下游产业链.....	6
图表 2 螺纹钢基本面因子（库存类和宏观类） .....	6
图表 3 螺纹钢基本面因子（供给类和需求类） .....	7
图表 4 单因子回测结果展示.....	9
图表 5 各大类因子结果均值.....	9
图表 6 各大类因子等权合成回测结果.....	10
图表 7 降维后因子可解释性方差.....	11
图表 8 OLS 模型回测结果.....	11
图表 9 XGBoost 模型回测结果.....	13
图表 10 各阶段因子重要性排序.....	13
图表 11 动态特征选择 XGBoost 模型回测结果 .....	14
图表 12 RNN 模型原理.....	14
图表 13 RNN 模型回测结果.....	15
图表 14 合成模型 1：OLS+XGBoost 合成结果 .....	16
图表 15 合成模型 2：OLS+特征选择 XGBoost 合成结果.....	16
图表 16 合成模型 3：OLS+XGBoost+RNN 合成结果.....	17
图表 17 合成模型 4：OLS+XGBOOST+RNN 少数服从多数模型.....	17
图表 18 铁矿石进口数量单因子回测结果 .....	19
图表 19 铁精粉矿山开工率单因子回测结果.....	19
图表 20 高炉产能利用率单因子回测结果 .....	20
图表 21 高炉检修限产量（年粗钢产量≤200 万吨）单因子回测结果.....	20
图表 22 高炉检修限产量（年粗钢产量≥600 万吨）单因子回测结果.....	21
图表 23 螺纹钢产能利用率单因子回测结果.....	21
图表 24 中厚板产能利用率单因子回测结果.....	22
图表 25 建筑钢材成交量单因子回测结果 .....	22
图表 26 房地产业固定资产投资完成额单因子回测结果.....	23
图表 27 汽车制造业固定资产投资完成额单因子回测结果.....	23
图表 28 汽车产量单因子回测结果.....	24
图表 29 挖掘机销量单因子回测结果 .....	24
图表 30 铁精粉总库存单因子回测结果.....	25
图表 31 铁精粉库存（70 家矿山企业）单因子回测结果.....	25
图表 32 铁精粉库存（70 家矿山企业：大型）单因子回测结果.....	26
图表 33 宏观经济景气指数单因子回测结果.....	26

图表 34 金属制品、机械和设备修理业企业景气指数单因子回测结果 .....	27
图表 35 建筑业企业景气指数单因子回测结果.....	27
图表 36 房地产业企业景气指数单因子回测结果 .....	28
图表 37 钢铁 PMI 单因子回测结果.....	28
图表 38 社会融资规模单因子回测结果.....	29
图表 39 银行间同业拆借单因子回测结果 .....	29

## 1、基本面因子选取

### 1.1、商品基本面量化原理

基本面量化试图有效地结合“基本面研究”和“量化投资”两套独立而成熟的交易理论——以大量的基本面数据为基础对未来价格做出预测，并自动生成有效的交易信号，不仅遵从基本面研究的严谨逻辑，又兼具量化交易的高效分析。具体而言，在本研究报告中，以螺纹钢的基本面数据作为输入，通过模型对未来价格进行预测，接着根据预测价格生成相应的多空信号进行交易。

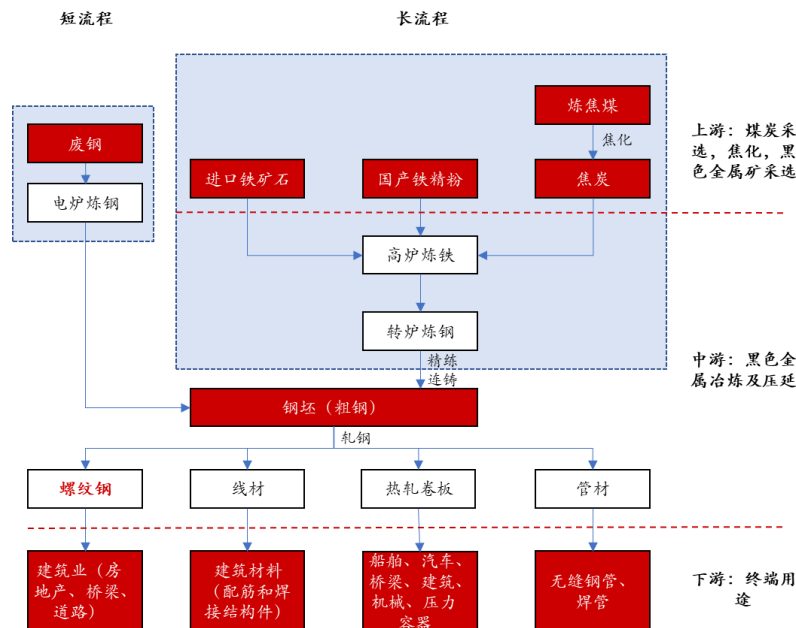
### 1.2、螺纹钢基本面因子选取

根据螺纹钢产业链的上下游关系，选取相应的基本面数据作为模型的输入因子。螺纹钢是整个钢铁产业最主要的制成品之一。钢铁行业的上游主要涵盖采选（焦煤）、焦化（焦炭）以及黑色金属矿采选（铁矿石）三个行业，中游是黑色金属冶炼及压延，下游则是各种轧钢制品的应用。本报告根据螺纹钢的上下游产业链关系，选取三大类因子，分别是：供给类因子（进口铁矿石数量，高炉开工率，产能利用率，按产能开工率等），需求类因子（房屋新开工面积，房地产投资国内贷款，汽车产量，汽车销售量等），以及库存类因子（铁矿石库存，铁精粉库存，烧结粉库存等）。

除此之外，以螺纹钢为首的黑色系商品与国内宏观经济关系密切。尤其是近年来，钢铁产业链日趋金融化，与其他大类资产的相关性增强，基于此基础上，本报告还整理了宏观数据作为宏观类因子（活期存款利率，宏观经济景气指数，PMI，社会融资规模等）。

以上四大类因子共计 121 个指标，其中产业相关数据均经过筛选取自于专业钢铁行业数据供应商，宏观相关数据取自万得。

图表1 螺纹钢上下游产业链



资料来源：东证衍生品研究院

图表2 螺纹钢基本面因子（库存类和宏观类）

	库存类因子		宏观类因子
库存1	库存:铁矿石:港口合计	宏观1	活期存款利率(月)
库存2	块矿: 总库存: 45个港口 (周度)	宏观2	宏观经济景气指数:一致指数
库存3	球团矿: 总库存: 45个港口 (周度)	宏观3	宏观经济景气指数:先行指数
库存4	进口贸易矿: 总库存: 45个港口 (周度)	宏观4	宏观经济景气指数:滞后指数
库存5	铁精粉: 总库存: 45个港口 (周度)	宏观5	中金CMI指数:工业生产:初值
库存6	库存:铁矿石:澳洲	宏观6	企业景气指数:金属制品、机械和设备修理业
库存7	库存:铁矿石:巴西	宏观7	企业景气指数:汽车制造业
库存8	港口外矿:铁矿石全国港口库存总量	宏观8	企业景气指数:开采专业及辅助性活动
库存9	澳洲矿-港口库存	宏观9	企业景气指数:橡胶和塑料制品业
库存10	巴西矿-港口库存	宏观10	企业景气指数:建筑业
库存11	铁精粉: 库存: 中国: 70家矿山企业 (136座矿山) (周度)	宏观11	企业景气指数:交通运输、仓储和邮政业
库存12	铁精粉: 库存: 中国: 70家矿山企业 (136座矿山) : 小型 (周度)	宏观12	企业景气指数:批发和零售业
库存13	铁精粉: 库存: 中国: 70家矿山企业 (136座矿山) : 中型 (周度)	宏观13	企业景气指数:房地产业
库存14	铁精粉: 库存: 中国: 70家矿山企业 (136座矿山) : 大型 (周度)	宏观14	PMI
库存15	国产烧结粉: 库存: 64家钢铁企业 (周度)	宏观15	PMI:生产
库存16	进口烧结粉: 库存: 64家钢铁企业 (周度)	宏观16	非制造业PMI:商务活动
库存17	进口铁矿石: 平均天数: 中国 (周度)	宏观17	中国综合PMI:产出指数
库存18	全国钢铁企业进口矿库存	宏观18	钢铁PMI
		宏观19	钢铁PMI:生产
		宏观20	M0
		宏观21	M1
		宏观22	M2
		宏观23	金融机构:各项贷款余额
		宏观24	社会融资规模:当月值
		宏观25	银行间同业拆借:加权平均利率:当月值

资料来源：东证衍生品研究院

图 表 3 螺纹钢基本面因子（供给类和需求类）

	供给类因子		需求类因子
供给1	进口数量:铁矿石:合计:当月值	需求1	产量:粗钢:当月值
供给2	矿山产量:铁精粉:全国:当期值	需求2	高炉开工率:全国
供给3	矿山开工率:铁精粉:全国:当期值	需求3	国产烧结粉:日均消耗量:64家钢铁企业(周度)
供给4	全国国产矿开工率	需求4	进口烧结粉:日均消耗量:64家钢铁企业(周度)
供给5	国产矿样本统计:原矿日产能	需求5	进口铁矿石:日均疏港量合计:45个港口(周度)
供给6	国产矿样本统计:精粉日产能	需求6	表观消费量:粗钢:当月值
供给7	国产矿样本统计:样本企业数	需求7	表观消费量:钢材:当月值
供给8	高炉:产能利用率:中国:钢铁企业(周度)	需求8	建筑钢材:成交量合计:中华人民共和国:主流贸易商(日度)
供给9	高炉:产能利用率(不含淘汰产能):中国:钢铁企业(周度)	需求9	线螺采购量:上海
供给10	高炉:开工率:中国:钢铁企业(周度)	需求10	固定资产投资完成额:房地产业:累计同比
供给11	高炉:开工率:中国:钢铁企业:年粗钢产量≤200万吨(周度)	需求11	房屋施工面积:累计值
供给12	高炉:开工率:中国:钢铁企业:年粗钢产量200-600万吨(周度)	需求12	房屋新开工面积:累计值
供给13	高炉:开工率:中国:钢铁企业:年粗钢产量≥600万吨(周度)	需求13	商品房销售额:累计值
供给14	高炉:检修限产量:中国:钢铁企业(周度)	需求14	商品房销售面积:累计值
供给15	高炉:检修限产量:中国:钢铁企业:年粗钢产量≤200万吨(周度)	需求15	商品房现房销售面积_累计值
供给16	高炉:检修限产量:中国:钢铁企业:年粗钢产量200-600万吨(周度)	需求16	商品房现房销售面积_累计值
供给17	高炉:检修限产量:中国:钢铁企业:年粗钢产量≥600万吨(周度)	需求17	房地产土地购置费_累计值
供给18	高炉:检修数量:中国:钢铁企业:年粗钢产量≤200万吨(周度)	需求18	房地产业土地购置面积_累计值
供给19	高炉:检修数量:中国:钢铁企业:年粗钢产量200-600万吨(周度)	需求19	本年实际到位资金合计_累计值
供给20	高炉:检修数量:中国:钢铁企业:年粗钢产量≥600万吨(周度)	需求20	房地产投资国内贷款_累计值
供给21	高炉:盈利钢厂:中国:钢铁企业(周度)	需求21	房地产投资利用外资_累计值
供给22	高炉:盈利钢厂:中国:钢铁企业:年粗钢产量≤200万吨(周度)	需求22	房地产投资各项应付款_累计值
供给23	高炉:盈利钢厂:中国:钢铁企业:年粗钢产量200-600万吨(周度)	需求23	房地产投资工程款_累计值
供给24	高炉:盈利钢厂:中国:钢铁企业:年粗钢产量≥600万吨(周度)	需求24	固定资产投资完成额:基础设施建设投资:累计同比
供给25	高炉样本统计:样本企业数	需求25	固定资产投资完成额:交通运输、仓储和邮政业:累计同比
供给26	高炉样本统计:高炉个数	需求26	固定资产投资完成额:电力、热力、燃气及水的生产和供应业:累计同比
供给27	高炉样本统计:日产能	需求27	固定资产投资完成额:水利、环境和公共设施管理业:累计同比
供给28	高炉样本统计:总容积	需求28	固定资产投资完成额:制造业:汽车制造业:累计同比
供给29	全国按产能高炉开工率	需求29	产量:汽车:累计同比
供给30	螺纹钢:产能:中国:建筑钢材钢铁企业(周度)	需求30	销量:汽车:当月同比
供给31	螺纹钢:产能利用率:中国:建筑钢材钢铁企业(周度)	需求31	销量:汽车:累计同比
供给32	螺纹钢:短流程:产能利用率:中国:建筑钢材钢铁企业(周度)	需求32	产量:乘用车:当月值
供给33	螺纹钢:长流程:产能利用率:中国:建筑钢材钢铁企业(周度)	需求33	产量:商用车:客车:当月值
供给34	中厚板:产能利用率:中国:钢铁企业(周度)	需求34	销量:汽车:当月值
供给35	全国按产能螺纹钢开工率	需求35	销量:商用车:客车:当月值
供给36	全国按产能热轧卷板开工率	需求36	产量:挖掘机:当月值
供给37	全国按产能带钢开工率	需求37	销量:挖掘机:工程机械行业:当月值
供给38	全国按产能焊管开工率	需求38	造船完工量:中国:累计值
		需求39	新接船舶订单量:中国:累计值
		需求40	手持船舶订单量:中国:累计值

资料来源：东证衍生品研究院

## 2、降维及模型选择

在利用基本面数据进行预测之前，首先对数据进行 PCA 降维，去除冗余的数据信息，接下来分别采用 OLS，XGBoost 以及 RNN 模型对数据进行回归处理，获取价格预测数据。此外，对于 XGBoost 模型，尝试利用动态特征选择的方式对模型进行改进。最



后，通过等权合成以及“少数服从多数”的方式结果多个模型的预测结果建立综合预测模型。

报告采用主成分分析法（PCA）进行降维处理，PCA 通过线性变换将原始数据变换为一组各维度线性无关的表示，可用于提取数据的主要特征分量，常用于高维数据的降维。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。

### 3、数据处理

**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；

**异常值处理：**对所有偏离均值 3 个标准差的数据进行处理；

**移仓换月处理：**为避免期货展期导致的价格影响，本报告以螺纹钢复权价格进行回测；

**周期性影响：**为剔除数据的周期性影响，对数据进行差分处理；

**可得性处理：**按指标具体可得性进行相应滞后处理；

**回测参数：**本报告基于日频数据生成周频级别的多空信号，每周更新仓位，手续费设置为双边万三。



#### 4、单因子预测能力及大类因子合成

针对所有单因子进行 OLS 回归，预测未来一周的螺纹钢价格，根据该价格进行多空交易，由此生成回测结果。下图为单因子回测结果展示，由于篇幅限制，只展示收益率前 10 和后 10 的因子，在本文附录部分列出了部分单因子的回测曲线。结果显示，大部分单因子的年化收益在 20% 以下，夏普值低于 1，而由于基本面数据多为低频数据（周频或月频），故单因子的平均持仓时间较长，普遍达到 30 天以上。

图表 4 单因子回测结果展示

	总收益	年化收益	年化波动率	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
宏观23	154.60%	21.37%	15.01%	1.26	-13.43%	1.59	0.53	1.17	2.45	39.23
宏观17	124.67%	18.26%	13.27%	1.20	-25.22%	0.72	0.52	1.17	2.42	55.27
宏观22	100.12%	15.46%	16.82%	0.78	-16.64%	0.93	0.52	1.10	1.45	36.85
宏观25	96.78%	15.06%	16.53%	0.77	-23.79%	0.63	0.52	1.11	1.32	40.53
宏观21	83.05%	13.35%	24.00%	0.46	-35.54%	0.38	0.54	0.98	0.71	39.23
供给1	78.06%	12.70%	17.52%	0.59	-30.11%	0.42	0.52	1.06	1.01	60.80
库存7	77.74%	12.66%	17.49%	0.59	-27.65%	0.46	0.52	1.06	1.01	43.43
库存10	77.74%	12.66%	17.49%	0.59	-27.65%	0.46	0.52	1.06	1.01	43.43
宏观18	77.43%	12.62%	17.12%	0.60	-26.48%	0.48	0.53	1.04	1.01	55.27
库存18	77.17%	12.58%	17.45%	0.58	-30.49%	0.41	0.52	1.06	1.00	60.80
.....	.....									
供给11	30.25%	5.63%	18.68%	0.17	-38.91%	0.14	0.51	1.05	0.30	67.56
供给9	29.29%	5.47%	18.72%	0.16	-39.10%	0.14	0.51	1.04	0.28	76.00
供给13	27.88%	5.23%	18.74%	0.15	-39.92%	0.13	0.51	1.03	0.26	60.80
供给30	26.82%	5.05%	22.41%	0.12	-38.27%	0.13	0.52	1.00	0.19	60.80
宏观4	18.73%	3.62%	22.24%	0.05	-38.37%	0.09	0.51	1.00	0.09	93.54
宏观10	16.06%	3.14%	19.19%	0.04	-45.05%	0.07	0.51	1.00	0.06	152.00
宏观2	12.29%	2.43%	19.42%	0.00	-44.61%	0.05	0.50	1.02	0.00	152.00
宏观8	11.38%	2.26%	19.42%	-0.01	-46.44%	0.05	0.51	1.01	-0.01	152.00
库存3	8.78%	1.76%	19.71%	-0.03	-45.25%	0.04	0.51	0.99	-0.05	60.80
宏观13	1.82%	0.38%	20.96%	-0.10	-48.91%	0.01	0.51	0.99	-0.16	152.00
宏观6	-14.40%	-3.17%	24.08%	-0.23	-53.78%	-0.06	0.50	0.99	-0.39	135.11

资料来源：东证衍生品研究院

在上表中，可以观察到排名前列的因子多为宏观因子。通过计算四大类因子的平均回测指标发现，显然宏观类因子的平均表现好于其他三类因子，这一现象也体现了螺纹钢与某些宏观经济因素的紧密连接。另一方面，在表现较差的 10 个单因子中，也存在较多的宏观因子，说明只有部分宏观因子对螺纹钢期货价格具有较好的预测能力。此外，供给 1（铁矿石进口数量），库存 7（巴西铁矿石库存），库存 10（巴西矿港口库存），库存 18（全国钢铁企业进口库存）等因子也具有较好的表现。

图表 5 各大类因子结果均值

下图统计了各大类因子的平均表现，总体而言，库存类因子的表现相对较好。

	总收益	年化收益	年化波动率	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
供给类平均	52.87%	9.12%	18.13%	0.37	-34.39%	0.27	0.52	1.05	0.64	66.97
需求类平均	55.35%	9.51%	17.97%	0.40	-33.21%	0.29	0.51	1.05	0.68	78.23
库存类平均	57.72%	9.80%	17.96%	0.42	-32.89%	0.31	0.52	1.05	0.71	67.12
宏观类平均	53.71%	8.78%	18.73%	0.38	-34.17%	0.34	0.52	1.05	0.67	81.65

资料来源：东证衍生品研究院

接下来，对各个大类内部因子的预测结果进行等权合成，观察其回测结果。

图表 6 各大类因子等权合成回测结果

	总收益	年化收益	年化波动率	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
供给类等权	46.63%	8.25%	18.16%	0.32	-35.92%	0.23	0.51	1.05	0.55	67.56
需求类等权	53.24%	9.25%	18.00%	0.38	-33.11%	0.28	0.51	1.05	0.65	86.86
库存类等权	87.80%	13.95%	17.35%	0.67	-28.93%	0.48	0.53	1.05	1.14	67.56
宏观类等权	39.51%	7.14%	18.33%	0.26	-39.43%	0.18	0.51	1.04	0.44	67.56
全部因子等权	66.80%	11.18%	17.70%	0.50	-31.59%	0.35	0.52	1.05	0.85	60.80

资料来源：东证衍生品研究院

通过上表发现，对预测结果进行等权合成并不会显著提升回测结果。在四大类因子等权合成的结果中，只有库存类因子等权合成的回测结果好于其大类平均值。

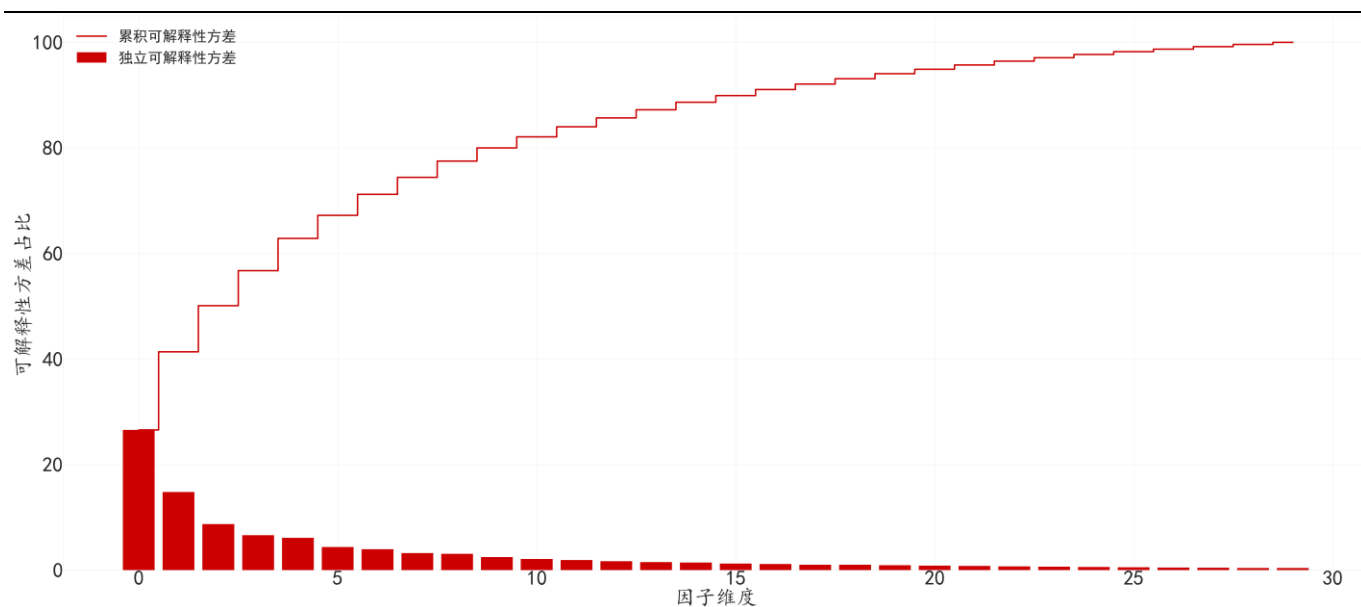
## 5、OLS 多元回归

在该部分，针对各大类因子进行多元回归，观测其回测结果。OLS（普通最小二乘法）多元回归的原理为，最优拟合曲线应该使各点到直线的距离的平方和（残差平方和 RSS）最小：

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

在进行多元回归之前，需要对数据进行降维，通过计算因子的可解释性变量，截取 95% 的信息度。在信号处理过程中，有效信号具有较大方差，而噪声具有较小方差，信噪比就是信号和噪声的方差比，该值越大越好。为了确定 PCA 降维的维度数量，需要计算可解释性方差，即将各个特征值大小除以求和的总特征值得到各个特征重要性占比。通过累计可解释性方差，截取 95% 的信息度，截取得到的维度数量为 30 个，表明前 30 个主成分包含了所有特征 95% 的信息。

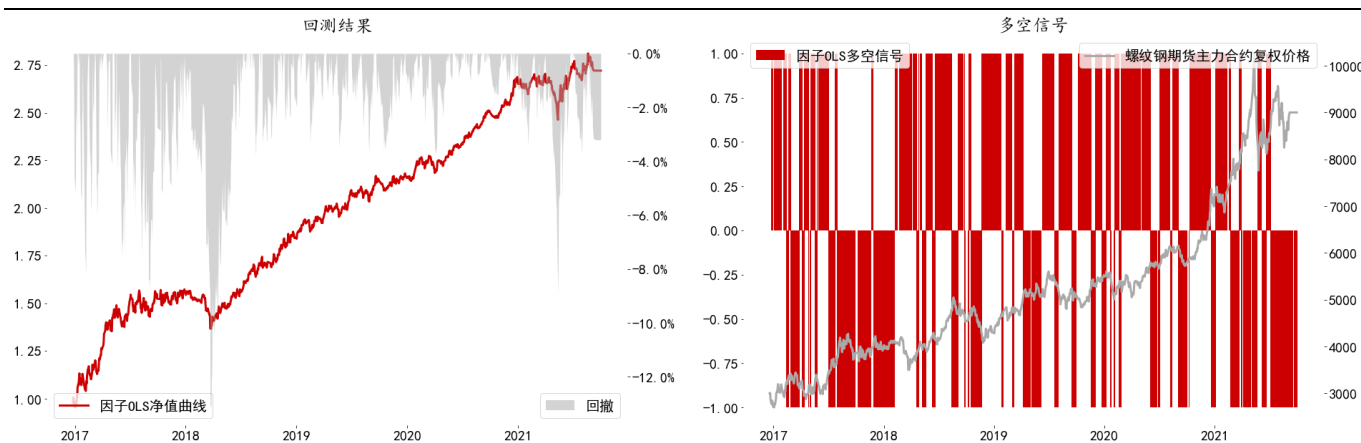
图表7 降维后因子可解释性方差



资料来源：东证衍生品研究院

利用经过降维后的 30 个特征对螺纹钢价格作多元回归，最后得到的模型表现为：  
**2017 年至今累计收益率 127.21%，年化收益率 24.15%，年化波动率 15.23%，夏普值 1.43，最大回撤-13.13%，胜率 54.68%，平均持仓时间 17.67 天。**通过 OLS 多元回归之后的策略表现明显好于之前的单因子策略，收益率以及策略稳定性都得到了较大的提升，另一方面，平均持仓时间也相应缩短了。

图表8 OLS 模型回测结果



资料来源：东证衍生品研究院

相对于单因子预测，OLS 多元回归包含了更多的基本面因子的信息，因此其回测结果更加稳健。其中两次比较大的回撤出现在 2018 年第一季度和 2021 年第二季度，比对期货合约主力价格可以发现，这两个时间节点螺纹钢期货价格出现比较大的波动。因此，可以认为 OLS 固然相对于单因子模型表现更为优秀，但面对大波动的行情时，依

然无法做到有效预测。接下来，尝试用非线性模型进行预测，看看效果如何。

## 6、XGBoost 模型预测

XGBoost 是基于 boosting 集成方法的一种集成学习算法，相较于传统的梯度提升树 (GBDT) 模型，XGBoost 在性能和效果上均有明显的提升。作为一个加法模型，XGBoost 的基模型一般选择树模型，但也包括逻辑回归等类型。XGBoost 模型的两大改进在于目标函数泰勒展开到了二阶，并且加入了叶子权重的 L2 正则化项，XGBoost 模型的目标函数为：

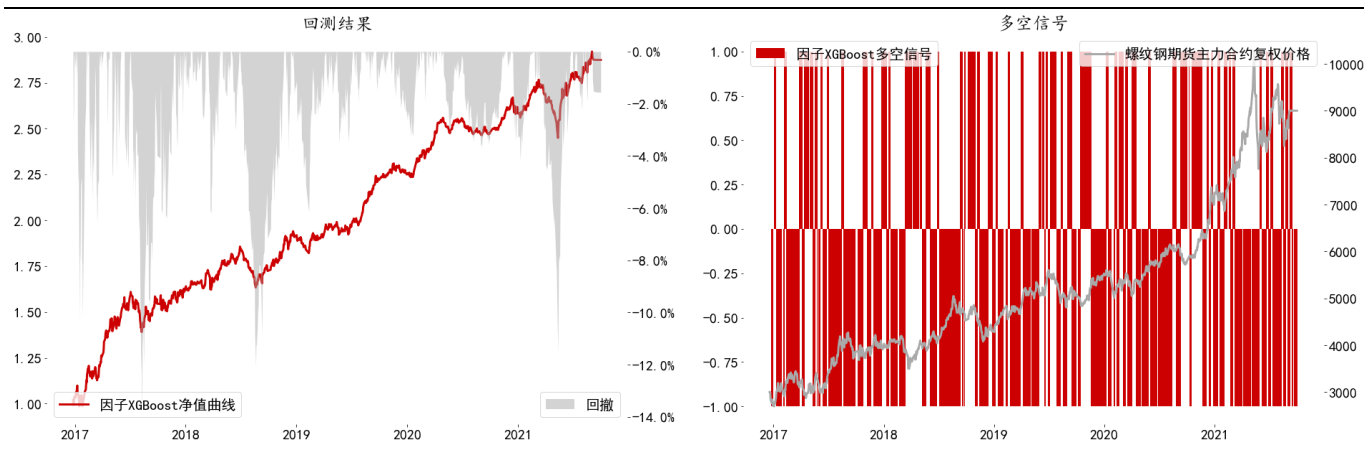
$$Obj(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$
$$where \Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$$

其中  $i$  表示第  $i$  个样本， $l(\hat{y}_i, y_i)$  表示第  $i$  个样本的预测误差，该值越小越好。后面的

$\sum_k \Omega(f_k)$  表示树的复杂度的函数，越小复杂度越低，表示其泛化能力越强， $T$  代表叶子节点数， $w$  代表叶子节点的分数。加入正则项的好处是防止过拟合，这个好处是由两方面体现的：一是预剪枝，因为正则项中有限定叶子节点数；二是正则项里 leaf score 的 L2 模平方的系数，对 leaf score 做了平滑。

同样利用降维后的特征进行 XGBoost 模型训练，得到的结果显示：**2017 年至今累计收益率 187.57%，年化收益率 25.64%，年化波动率 14.92%，夏普值 1.56，最大回撤 -13.59%，胜率 55.10%，平均持仓时间 11.54 天。**可以发现 XGBoost 的训练结果相对于 OLS 有一些提升，但是提升幅度较小，并未达到预期，尤其是在今年上半年出现了较大的回撤。考虑到随着市场行情的变化，各个特征的有效性也在不断变化，为了进一步考虑到特征的动态变化，接下来尝试使用特征动态选择的方式作 XGBoost 模型训练。

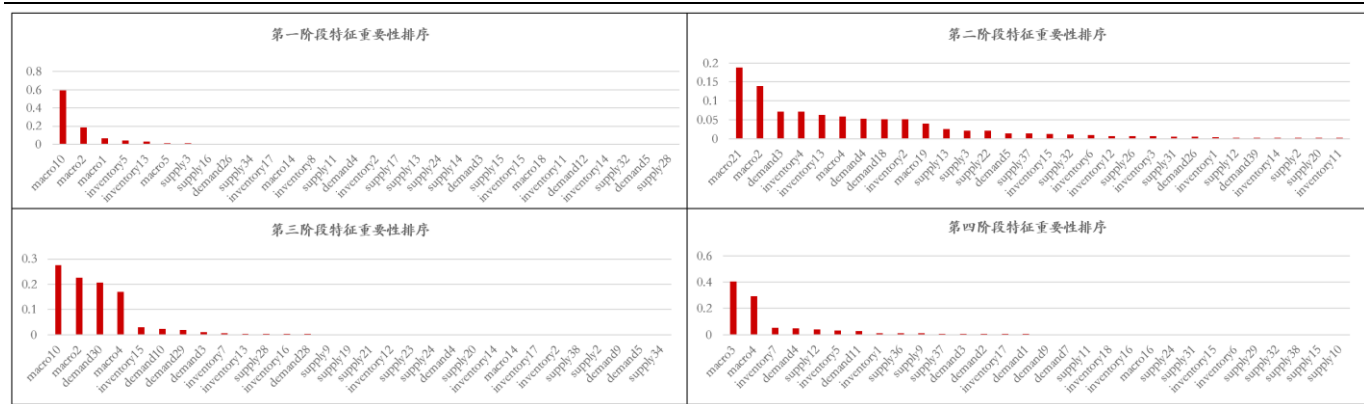
图表 9 XGBoost 模型回测结果



资料来源：东证衍生品研究院

针对未被降维的基本面信息，将总回测时段（共 1215 个交易日）均分为 5 个时段，每个时段包含约 243 个交易日（大致对应一个自然年的交易日）。在每个时段结束之后，统计该阶段重要性最大的 30 个特征，作为下个阶段的输入特征。此处选择 30 个特征是因为之前使用 PCA 降维后得到的维度数量为 30，因此采用同样的特征数量，方便进行对比。下图显示了前四个阶段的特征重要性排序，将第一阶段的重要特征作为第二阶段的输入特征，以此类推，而第一阶段则采用 PCA 降维后的特征，这一操作也是为了和前述模型进行对比。

图表 10 各阶段因子重要性排序

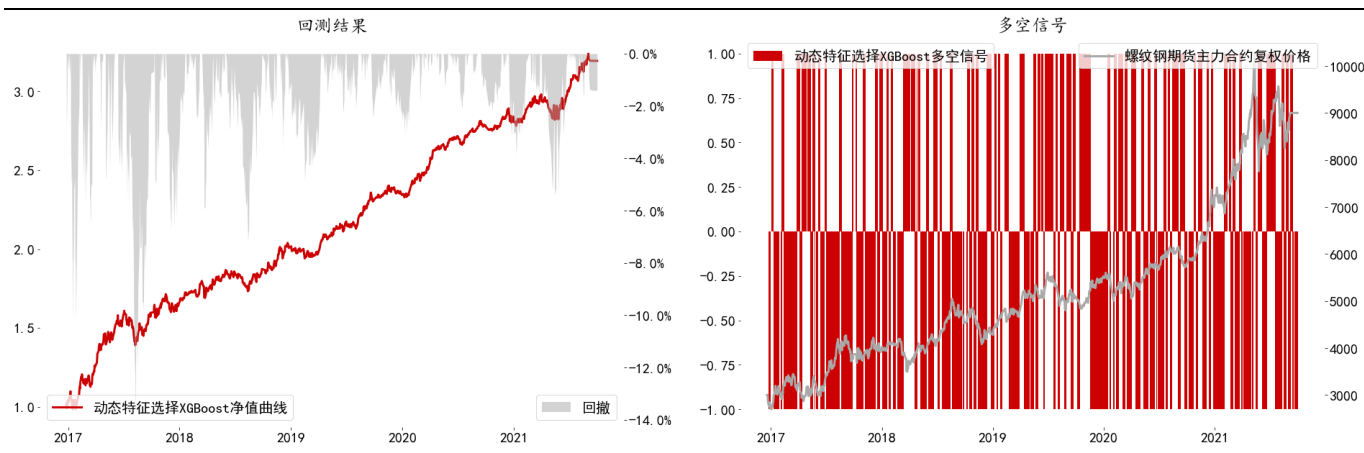


资料来源：东证衍生品研究院

经过模型训练，得到的结果显示：**2017 年至今累计收益率 219.21%，年化收益率 28.51%，年化波动率 14.54%，夏普值 1.80，最大回撤-13.59%，胜率 55.96%，平均持仓时间 9.18 天。**对比可得，采用动态特征选择之后的 XGBoost 模型有了更好的回测效果，收益率和夏普值都得到了明显的提升，相应的，平均持仓时间也进一步缩短

了。此外，对比两者的回撤图，可以发现，未采用动态特征选择的 XGBoost 模型在今年上半年的大级别行情下有较大的回撤，而采用动态特征选择的 XGBoost 模型则有效减少了该时段的下撤。因此可以初步认为，采用动态特征选择的 XGBoost 模型能够更加有效地应对较大级别的价格波动。

图表 11 动态特征选择 XGBoost 模型回测结果

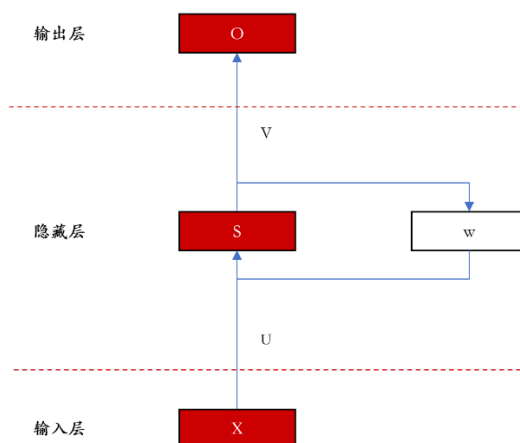


资料来源：东证衍生品研究院

## 7、RNN 模型预测

RNN (Recurrent Neural Network)，又称循环神经网络，是一类用于处理序列数据的神经网络。基础的神经网络包括输入层，隐藏层，输出层，并通过激活函数控制输出，层与层之间通过权值连接，而 RNN 在此基础上，对层之间的神经元也建立了权连接，基于此，RNN 便具有了记忆功能。

图表 12 RNN 模型原理



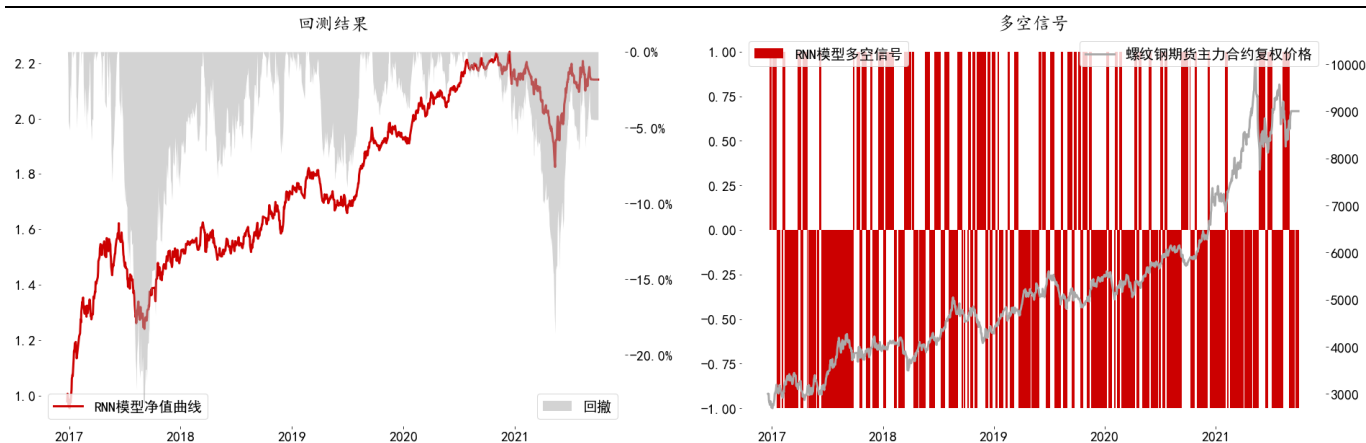
资料来源：东证衍生品研究院



上图显示， $U$  是输入层到隐藏层的权重矩阵， $V$  是隐藏层到输出层的权重矩阵。循环神经网络中隐藏层的值  $S$  不仅仅取决于当前这次的输入  $X$ ，还取决于上一次隐藏层的值  $S$ ，权重矩阵  $W$  就是隐藏层上一次的值作为这一次的输入的权重，因此 RNN 网络对上一时刻的隐藏层具有记忆功能。

然而传统的 RNN 会遇到的一个问题是后面时间的节点对于前面节点节点的感知力下降。于是在 RNN 记忆模型中再加入遗忘门，输入门和输出门，使得可以记忆更加长期的信息，这就是 LSTM（长短期记忆）模型，LSTM 是一种特殊的 RNN 模型。遗忘门（forget gate）在 LSTM 中一定的概率控制是否遗忘上一层的隐藏细胞状态；输入门（input gate）负责处理当前序列位置的输入，确定什么样的新信息被存放在细胞状态中；输出门（output gate）则对更新后的细胞状态进行输出。本报告使用 LSTM 模型进行训练，得到的结果为：**2017 年至今累计收益率 114.06%，年化收益率 17.87%，年化波动率 15.80%，夏普值 0.98，最大回撤-23.51%，胜率 53.82%，平均持仓时间 12.40 天。**可以发现，RNN 的训练结果并没有 OLS 和 XGBoost 有效，一部分原因在于 RNN 属于深度学习模型，需要较大的数据量，本报告基于低频的基本面数据生成周频级别的信号，即使已经将数据填充至日频数据，仍然只用千量级的数据，导致训练结果并不有效。

图表 13 RNN 模型回测结果



资料来源：东证衍生品研究院

## 8、模型合成

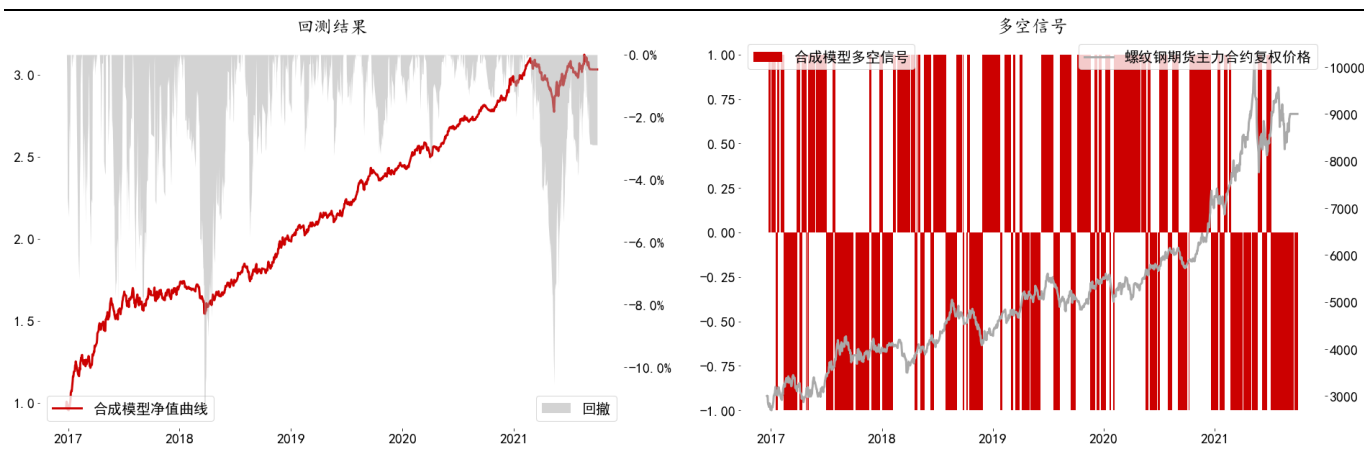
行文至此，我们已经有了四个模型的预测结果了，分别是 OLS，XGBoost，特征选择 XGBoost，以及 RNN 模型。不妨将这四个模型结果进行合成，得到一个综合模型，于是按照一定的权重分别对各个模型赋值，观测回测结果。篇幅所限，本报告并未讨论所有可能的合成组合，但下文所列出的合成组合已经能够充分展示综合模型的预测能



力。

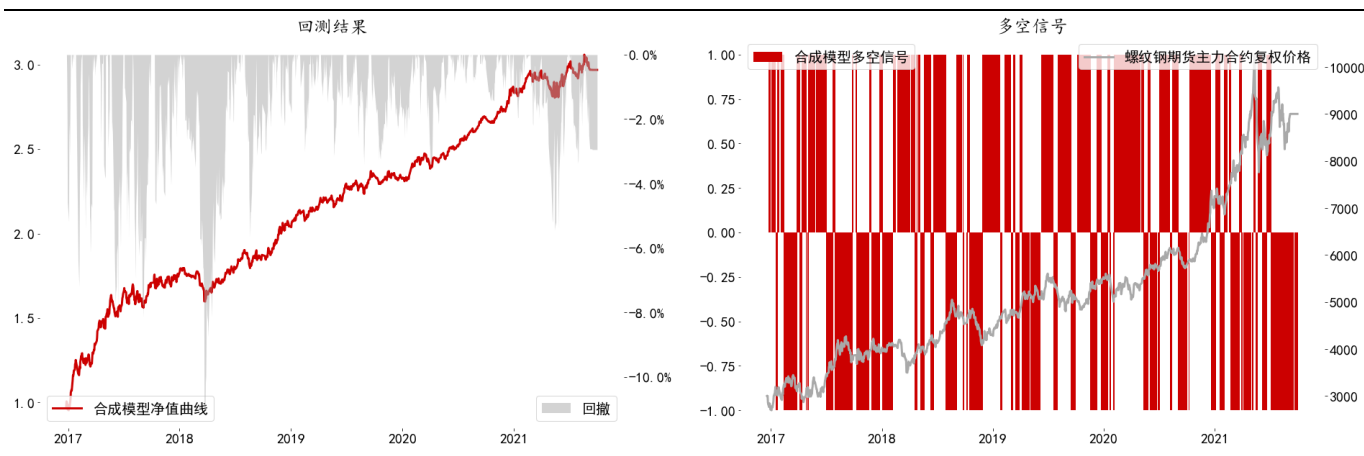
下面两张图依次展示了 OLS 分别与 XGBoost 和特征选择 XGBoost 的合成结果。

图表 14 合成模型 1: OLS+XGBoost 合成结果



资料来源：东证衍生品研究院

图表 15 合成模型 2: OLS+特征选择 XGBoost 合成结果



资料来源：东证衍生品研究院

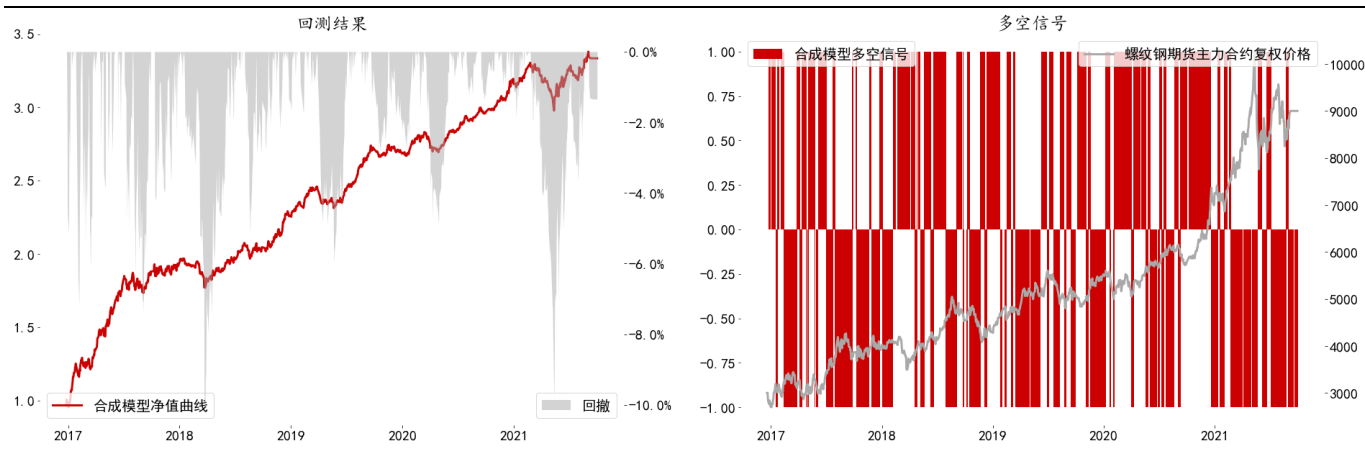
结果显示，合成模型 1 的回测表现为：2017 年至今累计收益率 203.20%，年化收益率 27.09%，年化波动率 14.00%，夏普值 1.76，最大回撤-11.38%，胜率 55.45%，平均持仓时间 16.66 天。

合成模型 2 的回测表现为：2017 年至今累计收益率 196.93%，年化收益率 26.52%，年化波动率 13.95%，夏普值 1.73，最大回撤-11.03%，胜率 55.02%，平均持仓时间 15.76 天。

合成模型 1 和合成模型 2 的回测结果并没有太大差异，但是从回撤图来看，合成模型 2 在 2021 年上半年的行情中回撤更小，这与特征合成 XGBoost 的特点是一致的，接

下来加入 RNN 模型观察结果。

图表 16 合成模型 3: OLS+XGBoost+RNN 合成结果

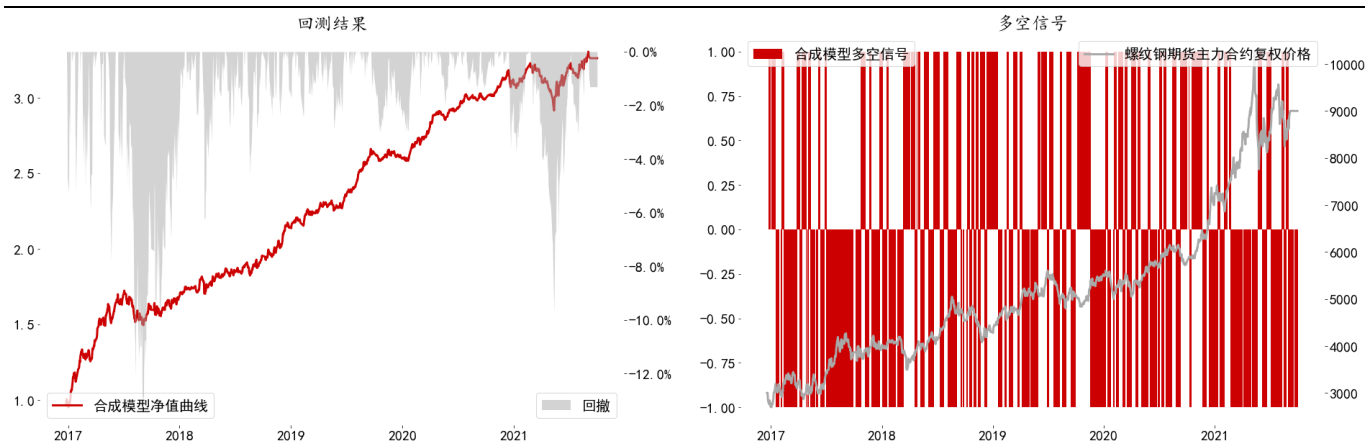


资料来源：东证衍生品研究院

结果显示，合成模型 3 的回测表现为：**2017 年至今累计收益率 233.53%，年化收益率 29.74%，年化波动率 13.16%，夏普值 2.08，最大回撤-10.07%，胜率 56.05%，平均持仓时间 14.22 天。**在加入了 RNN 模型之后，合成模型的回测结果得到进一步提升。

除了等权合成的方式，还可以利用“少数服从多数”的信号输出机制。根据 OLS，XGBoost，RNN 三个模型的多空信号输出，根据少数服从多数的原理进行合成，

图表 17 合成模型 4: OLS+XGBOOST+RNN 少数服从多数模型



资料来源：东证衍生品研究院

结果显示，合成模型 4 的回测表现为：**2017 年至今累计收益率 226.24%，年化收益率 29.11%，年化波动率 13.62%，夏普值 1.96，最大回撤-13.25%，胜率 56.05%，平均持仓时间 11.43 天。**总体而言，“少数服从多数”合成模型与等权模型的差异并不大。

## 9、总结及展望

本文利用螺纹钢基本面数据进行多模型价格预测，比较了不同模型之间的效果，最后尝试合成不同模型的预测结果生成综合预测模型。研究发现，XGBoost 相对于其他模型，具有较好的预测效果，如果采用动态特征选择的方法，可以更加有效地应对大级别地行情。在合成模型方面，三模型的合成模型好于两模型的合成模型，以**合成模型 3**为例（OLS+XGBoost+RNN 等权合成）可以达到超过 2 的夏普值，同时“少数服从多数”的合成方法相对于等权合成并没有明显提升。

本研究方法仍然存在值得改进之处，除了模型本身的优化之外，还可以尝试与商品分析师合作，对于基本面因子进行人工筛选以及特征构造，这也是我下一步预备的研究方向。

在模型的泛化方面，期待在模型得到更多验证之后，构建全品种的预测模型，在横截面和时间序列都能生成交易信号。

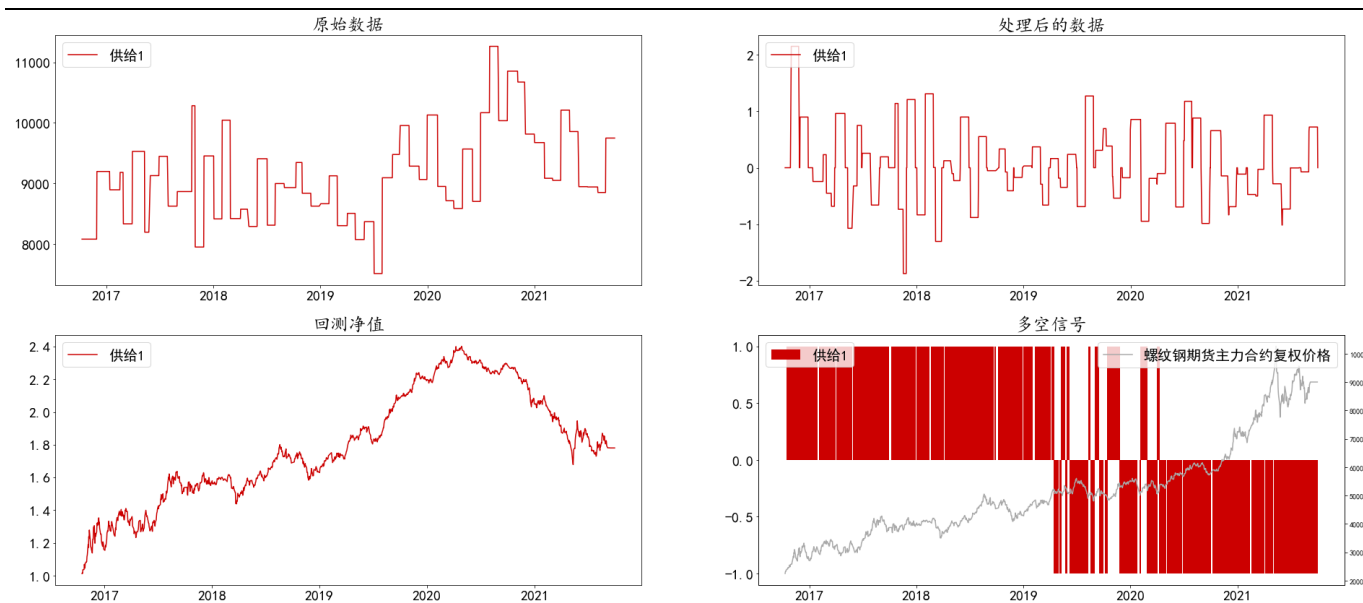
## 10、风险提示

市场风格的变换会造成特征有效性变化，导致模型效果下降。

## 11、附录：部分因子回测结果展示

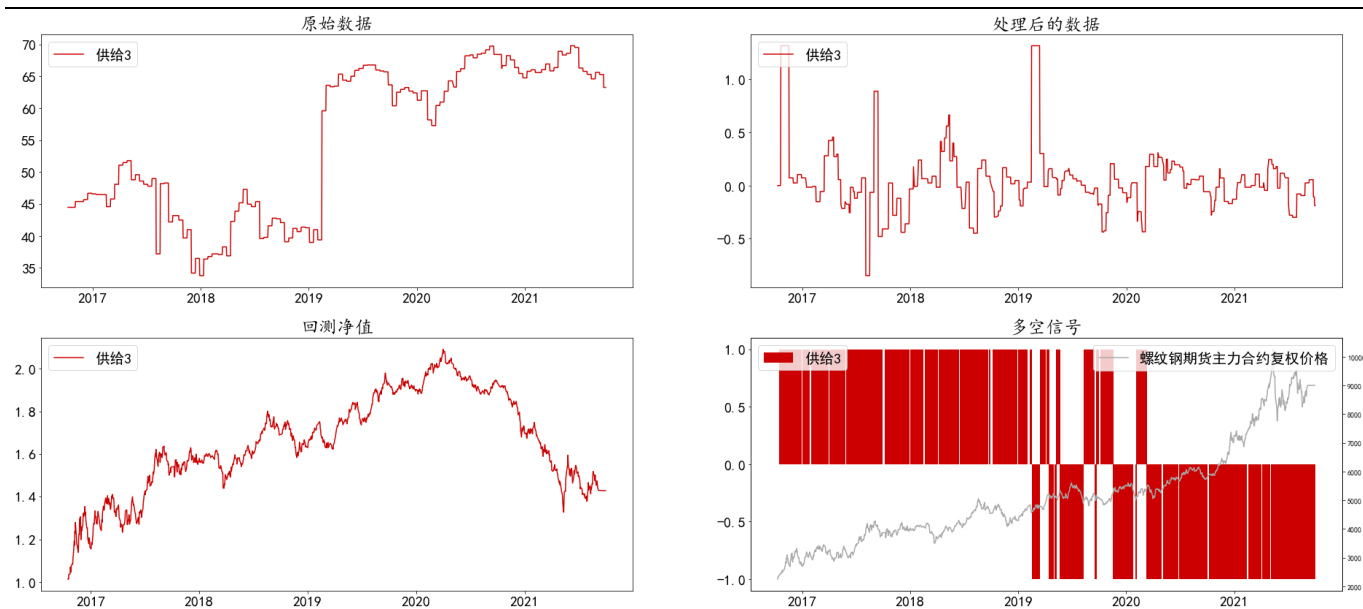
在附录中给出部分因子的回测结果，包括原始数据（左上图），处理后的数据（右上图），回测净值（左下图），以及多空信号（右下图）。

图表 18 铁矿石进口数量单因子回测结果



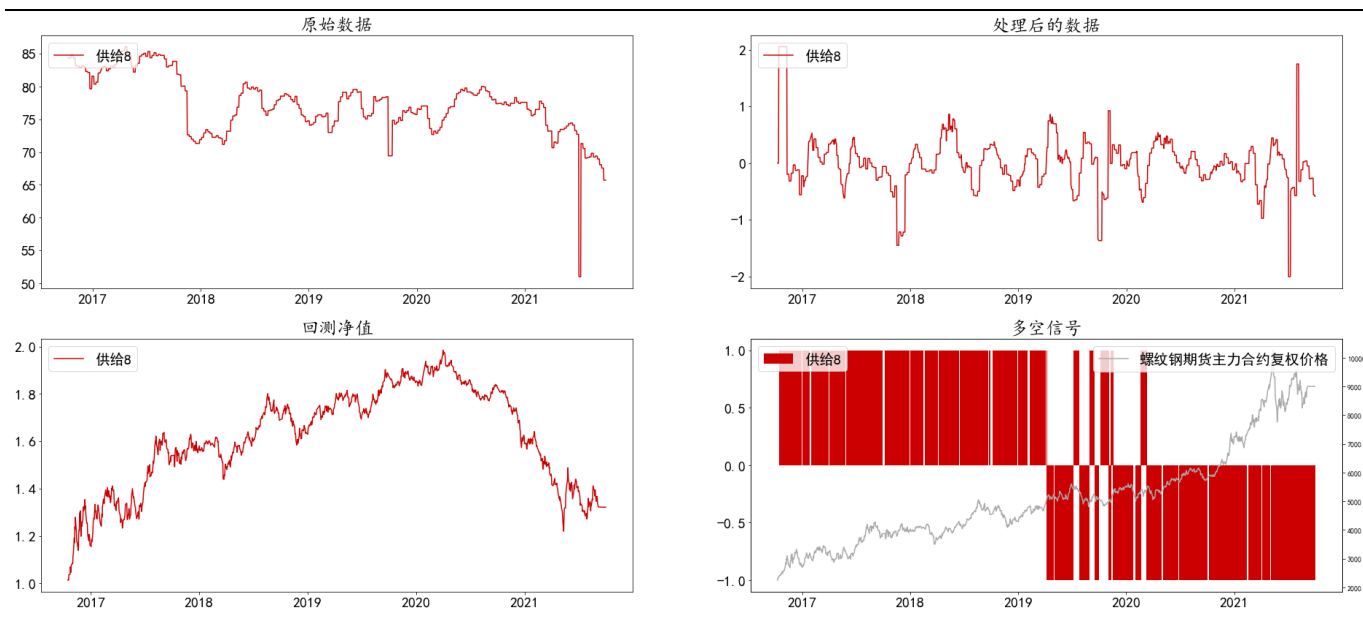
资料来源：东证衍生品研究院

图表 19 铁精粉矿山开工率单因子回测结果



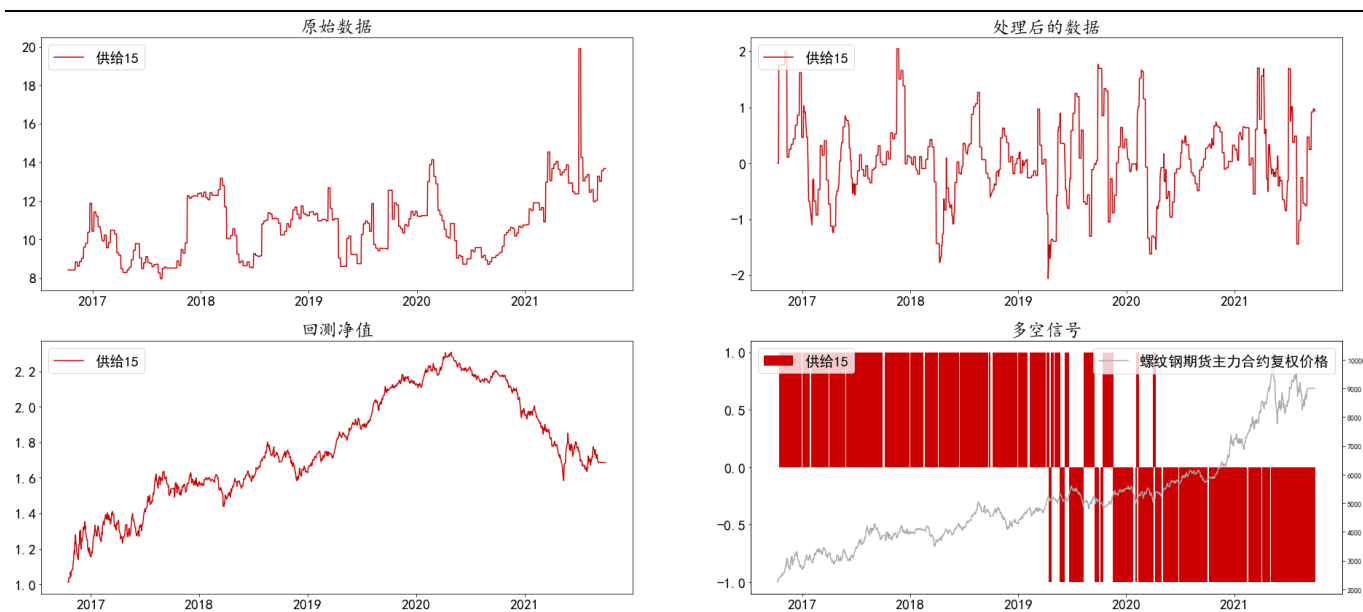
资料来源：东证衍生品研究院

图表 20 高炉产能利用率单因子回测结果



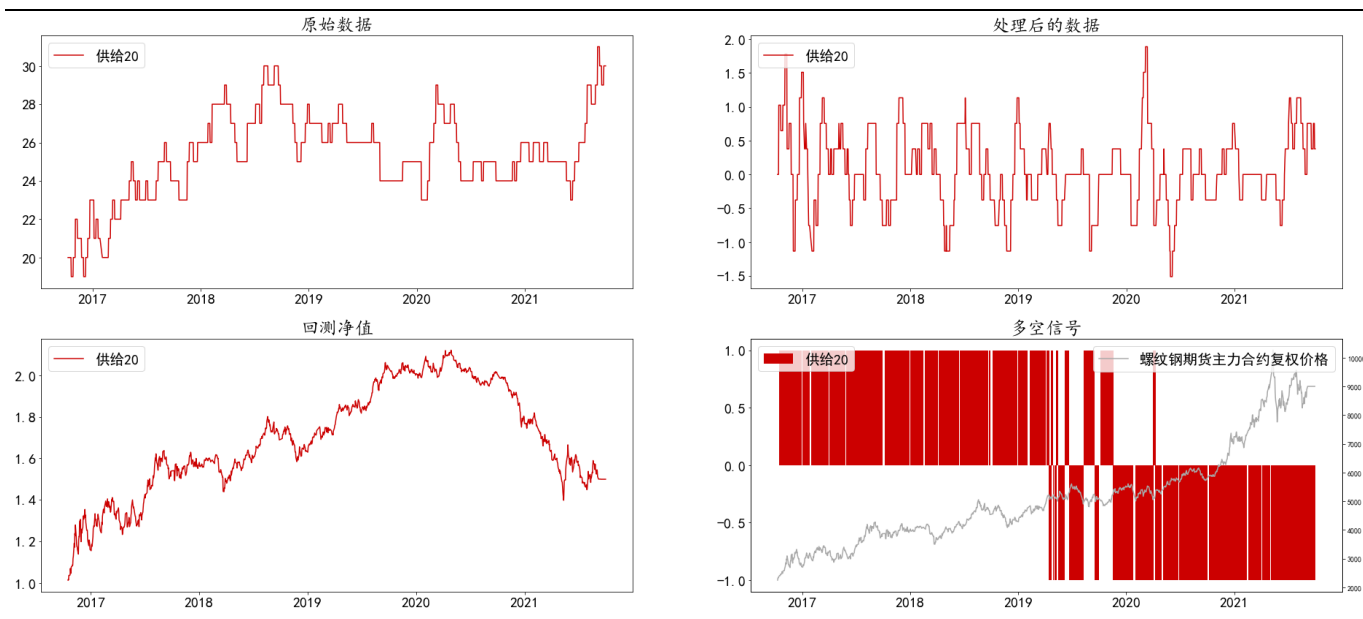
资料来源：东证衍生品研究院

图表 21 高炉检修限产量（年粗钢产量 $\leq 200$ 万吨）单因子回测结果



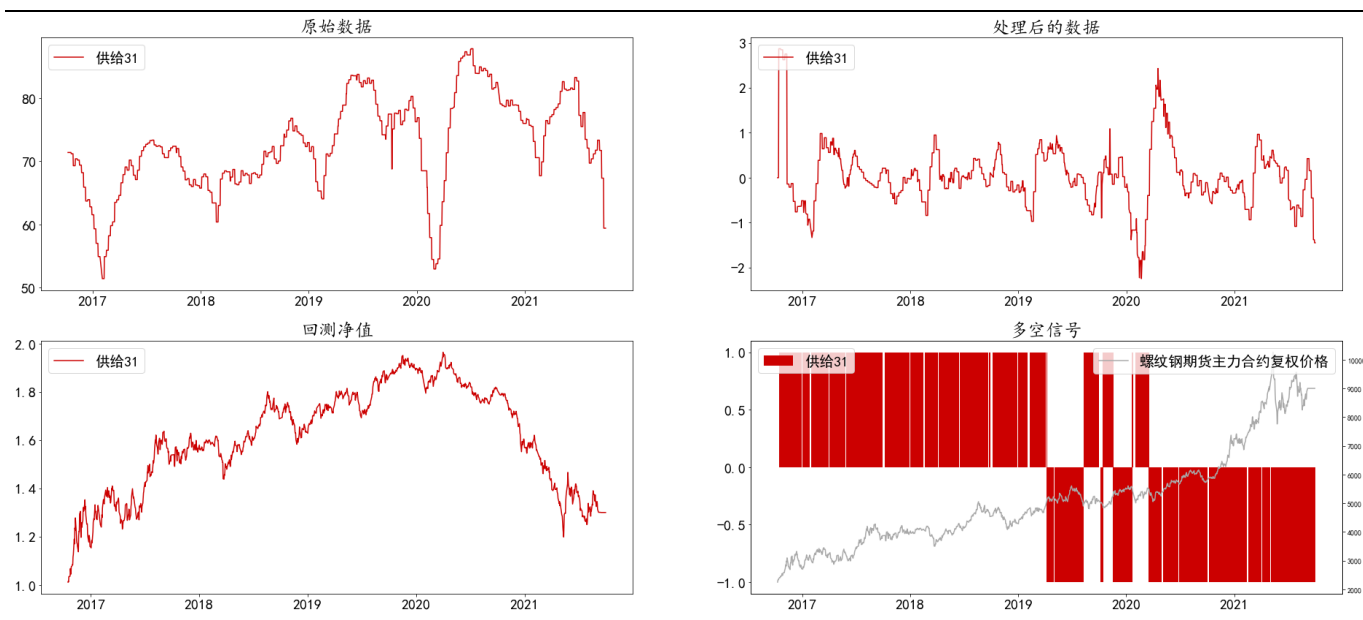
资料来源：东证衍生品研究院

图表 22 高炉检修限产量（年粗钢产量 $\geq 600$ 万吨）单因子回测结果



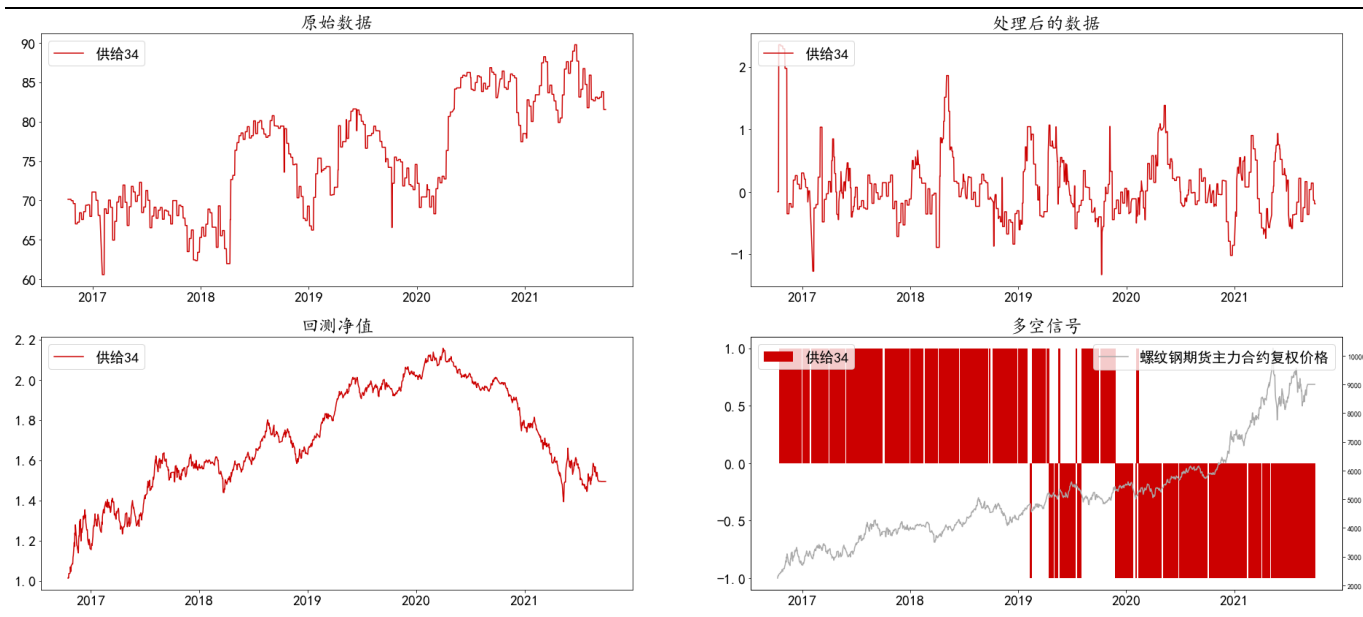
资料来源：东证衍生品研究院

图表 23 螺纹钢产能利用率单因子回测结果



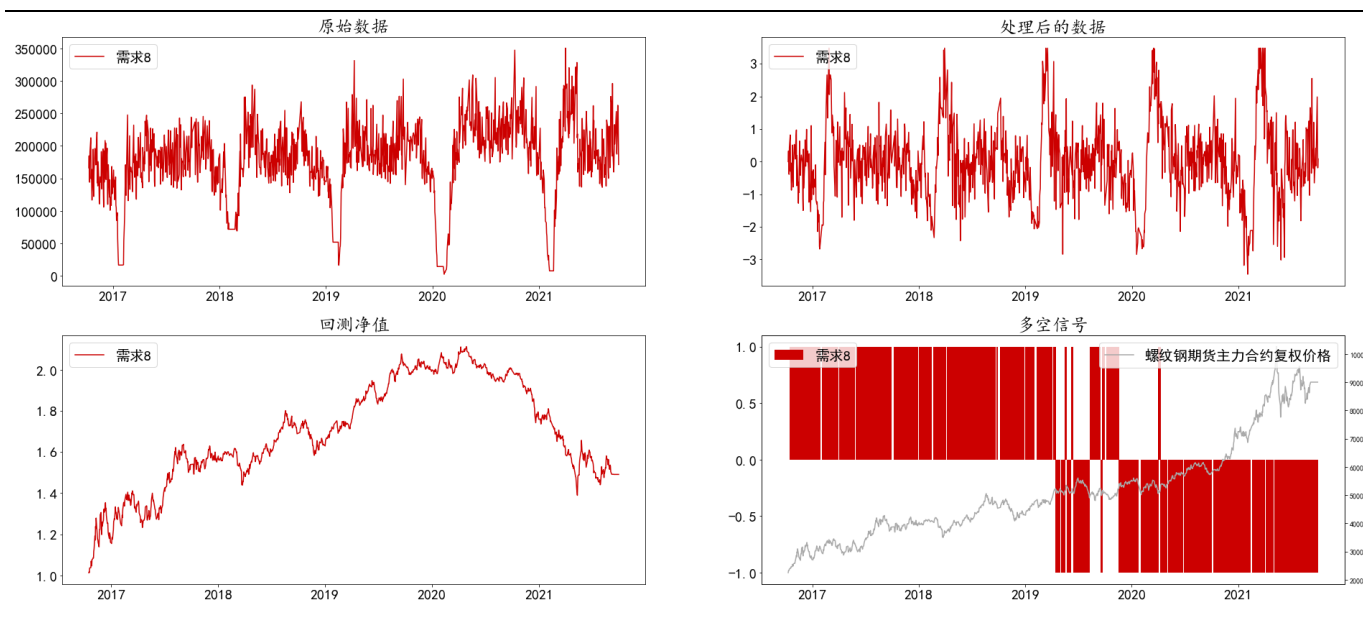
资料来源：东证衍生品研究院

图表 24 中厚板产能利用率单因子回测结果



资料来源：东证衍生品研究院

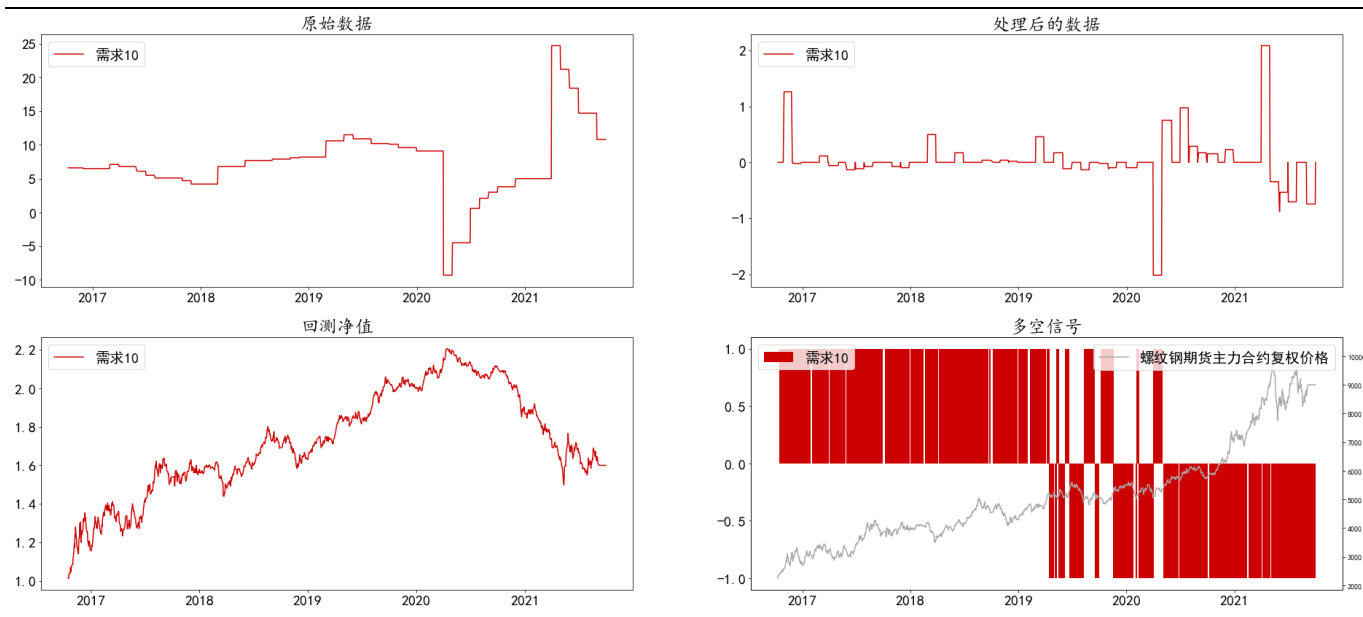
图表 25 建筑钢材成交量单因子回测结果



资料来源：东证衍生品研究院

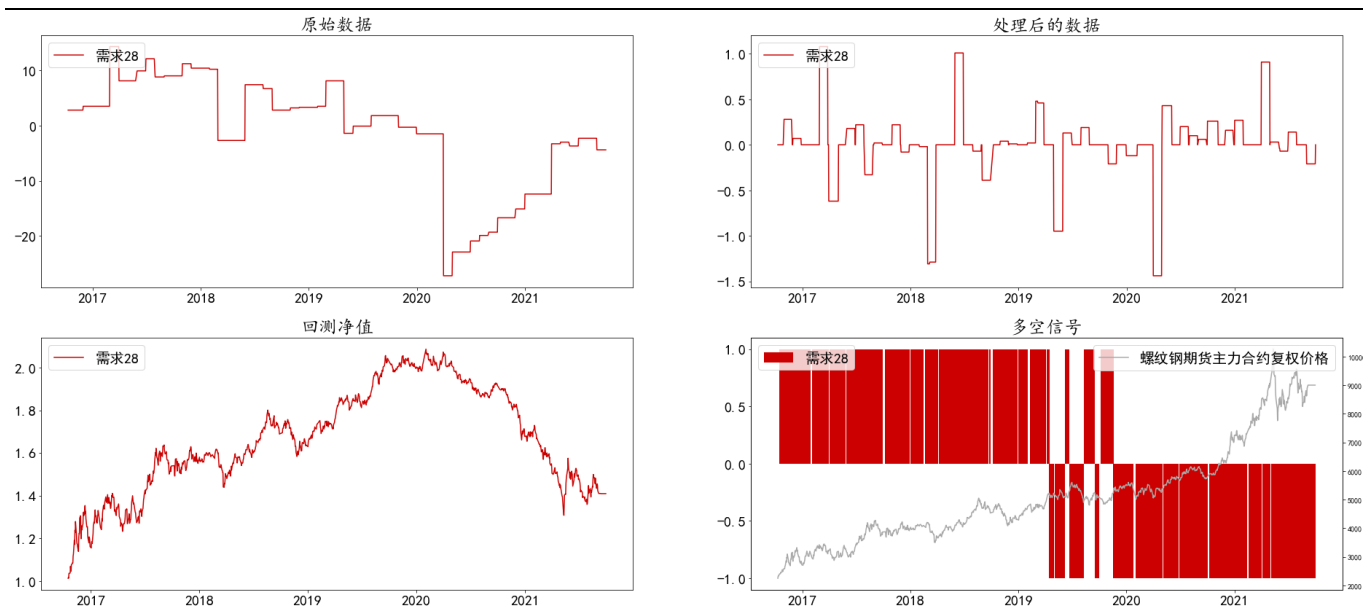


图表 26 房地产业固定资产投资完成额单因子回测结果



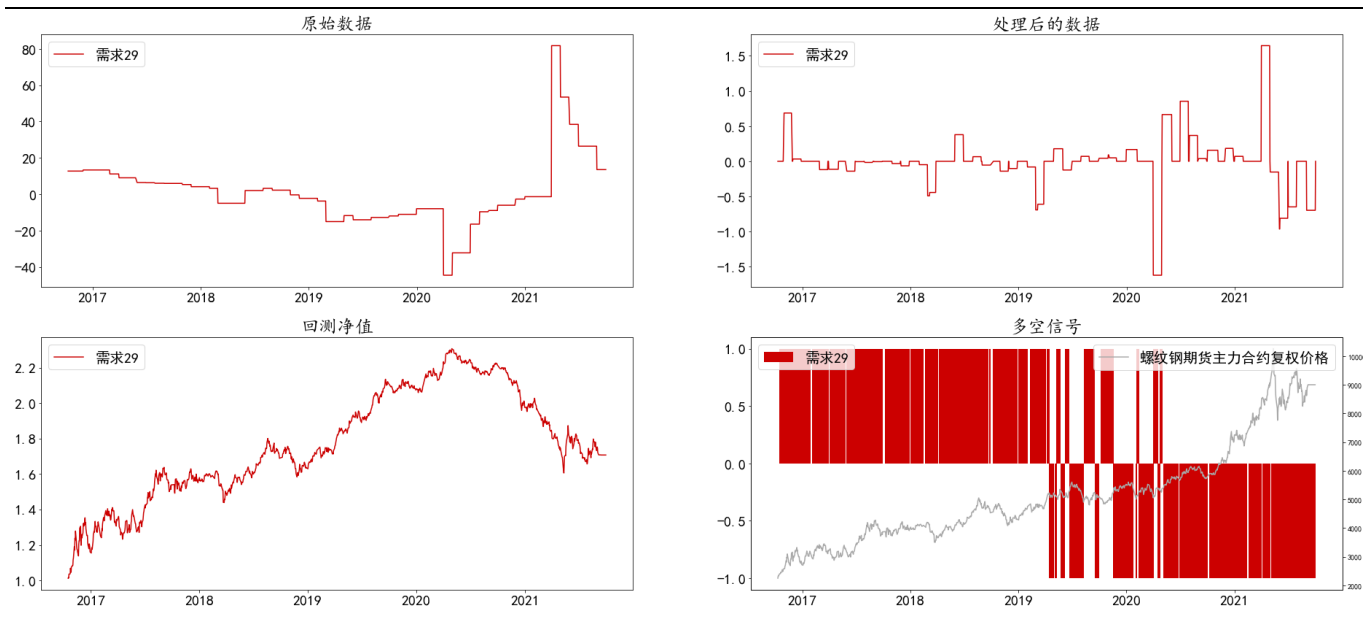
资料来源：东证衍生品研究院

图表 27 汽车制造业固定资产投资完成额单因子回测结果



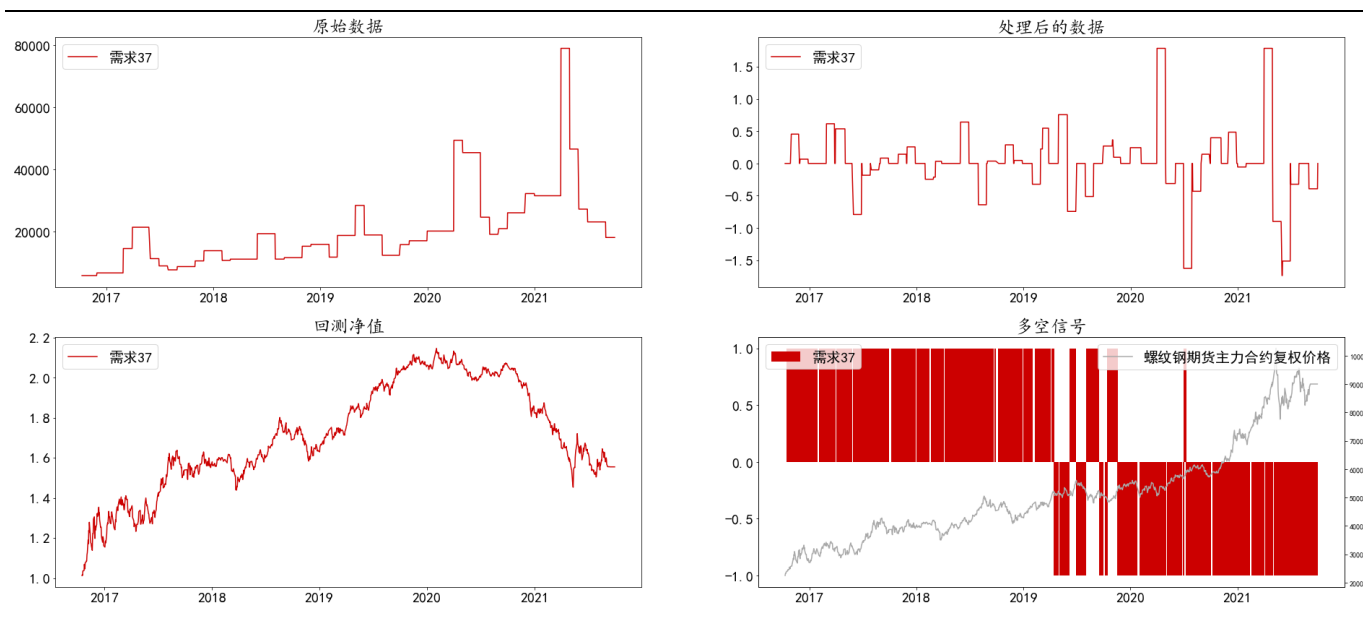
资料来源：东证衍生品研究院

图表 28 汽车产量单因子回测结果



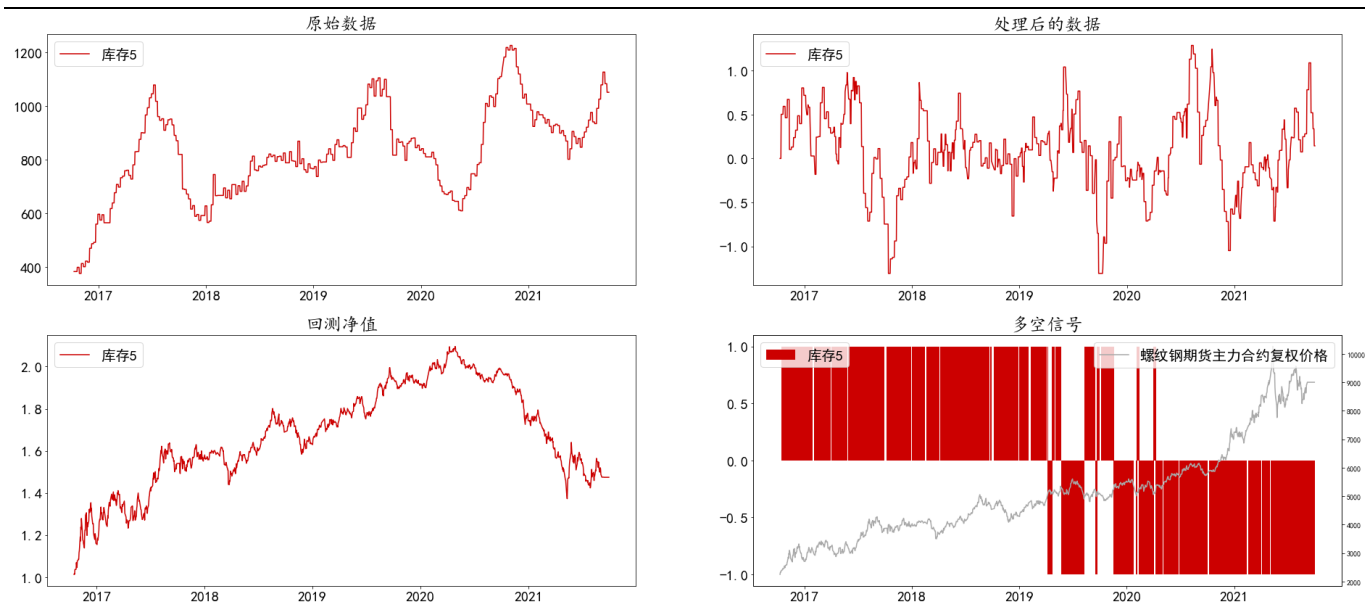
资料来源：东证衍生品研究院

图表 29 挖掘机销量单因子回测结果



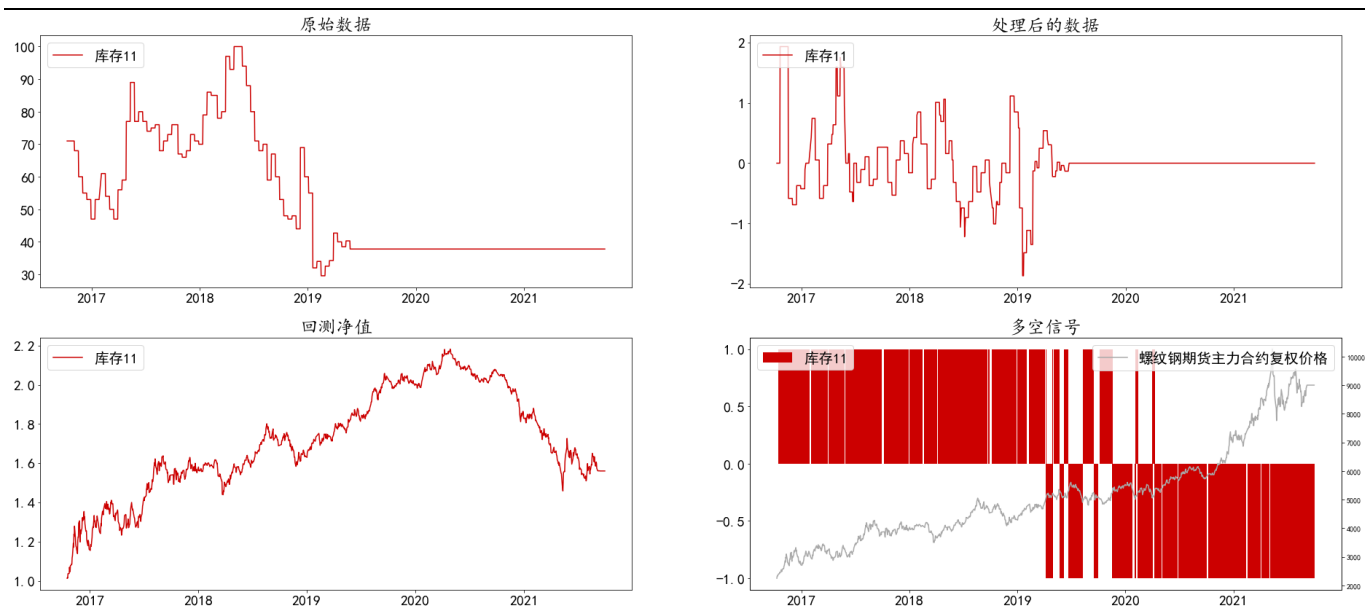
资料来源：东证衍生品研究院

图表 30 铁精粉总库存单因子回测结果



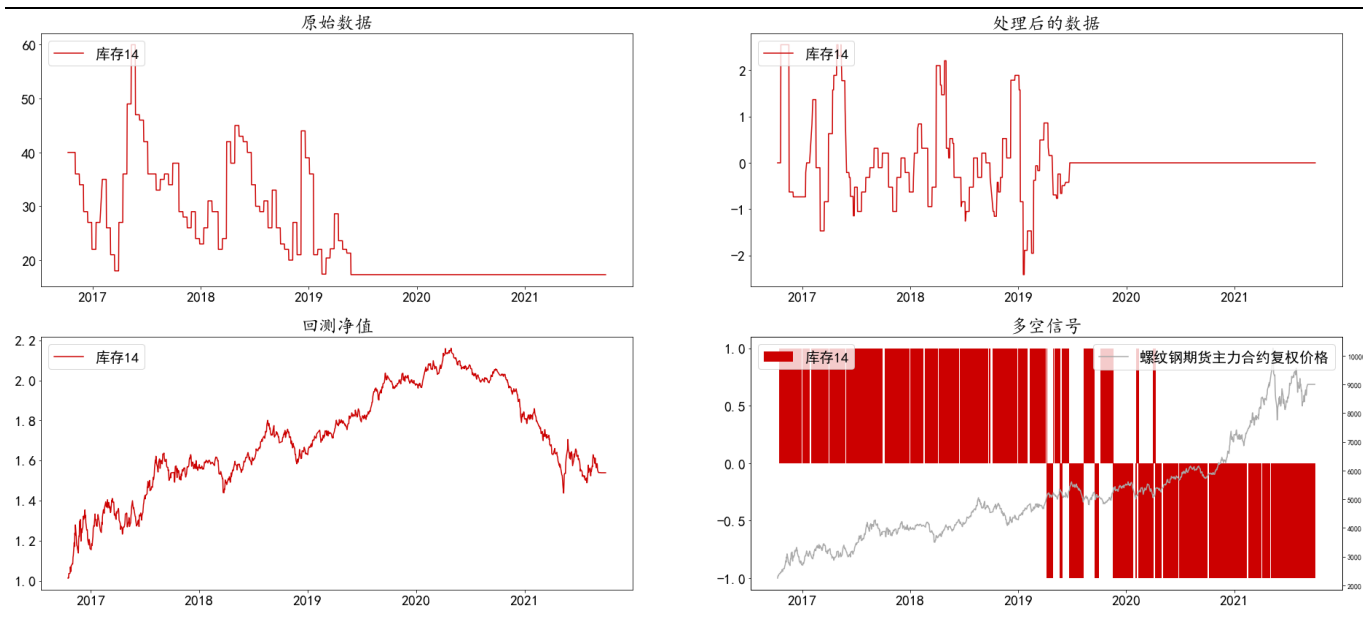
资料来源：东证衍生品研究院

图表 31 铁精粉库存（70 家矿山企业）单因子回测结果



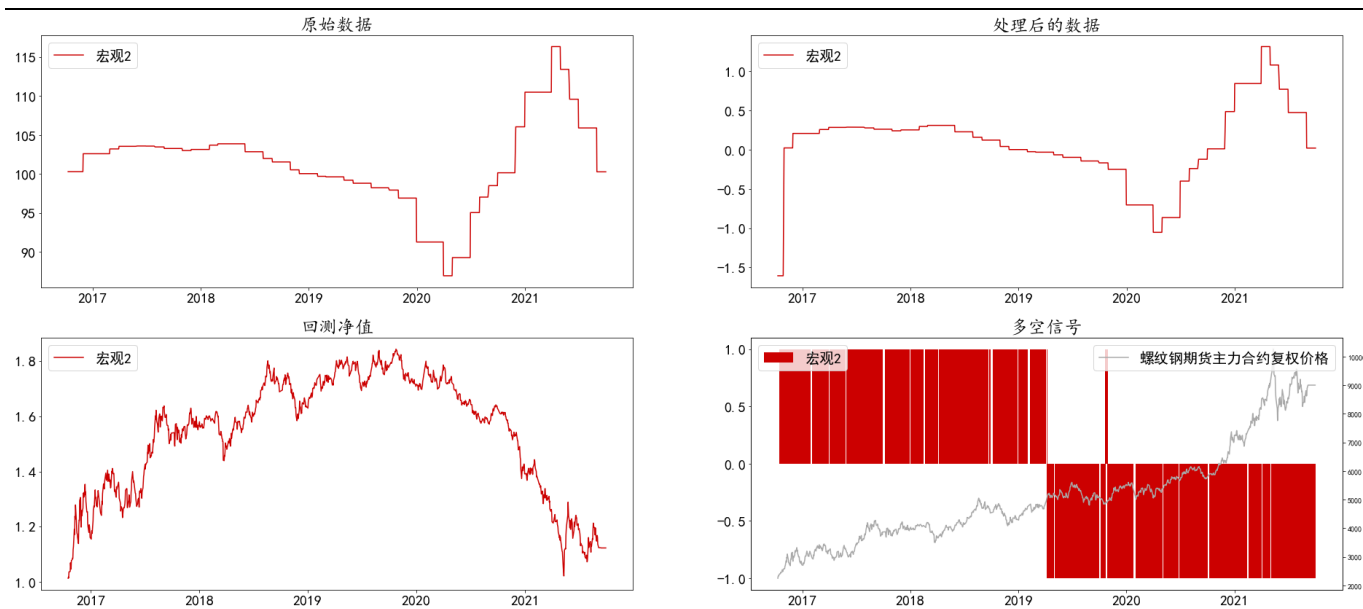
资料来源：东证衍生品研究院

图表 32 铁精粉库存（70 家矿山企业：大型）单因子回测结果



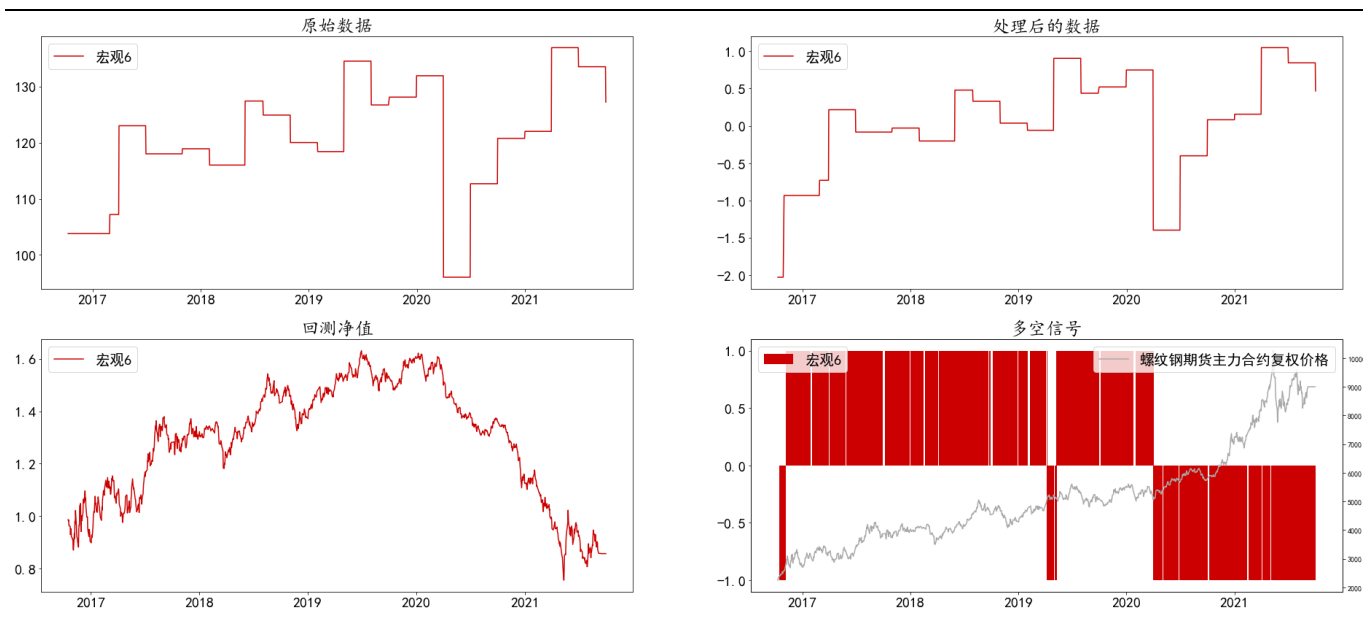
资料来源：东证衍生品研究院

图表 33 宏观经济景气指数单因子回测结果



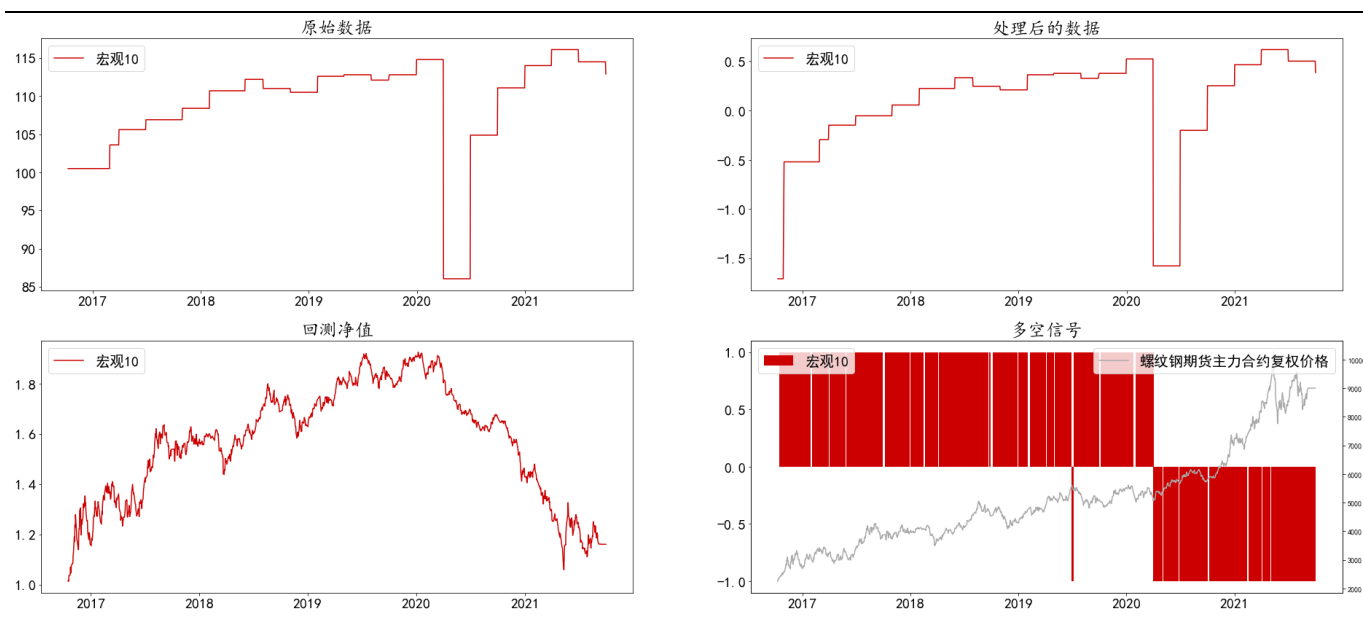
资料来源：东证衍生品研究院

图表 34 金属制品、机械和设备修理业企业景气指数单因子回测结果

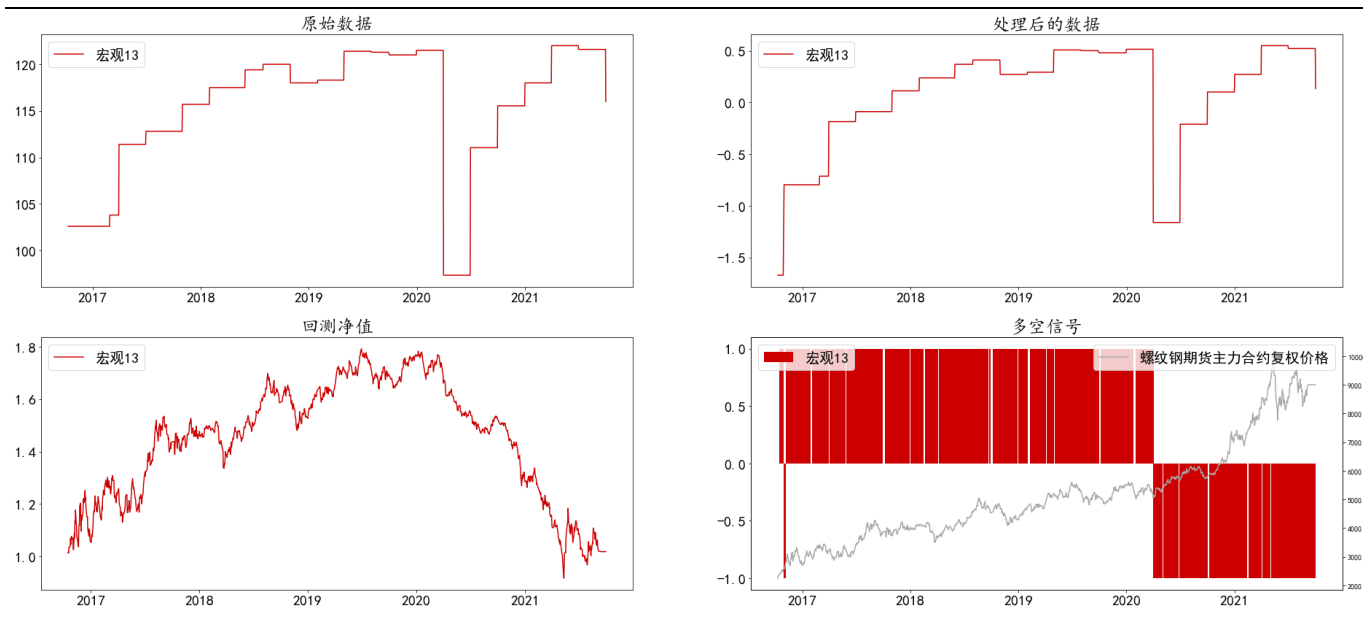


资料来源：东证衍生品研究院

图表 35 建筑业企业景气指数单因子回测结果

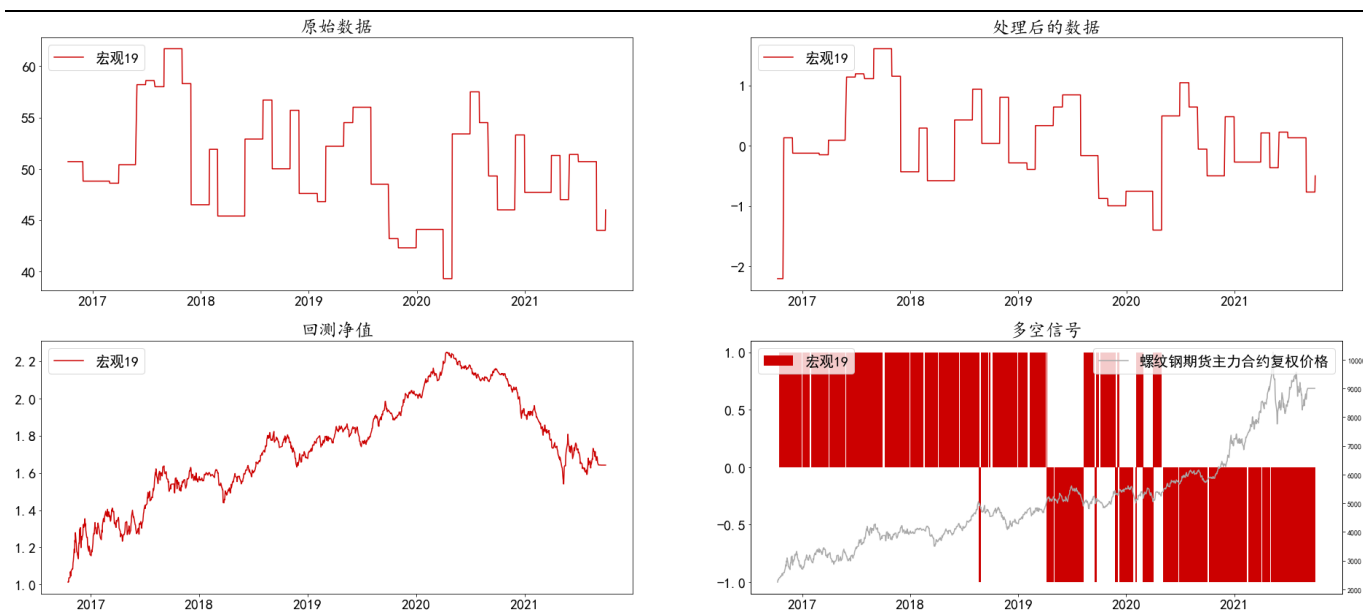


图表 36 房地产业企业景气指数单因子回测结果



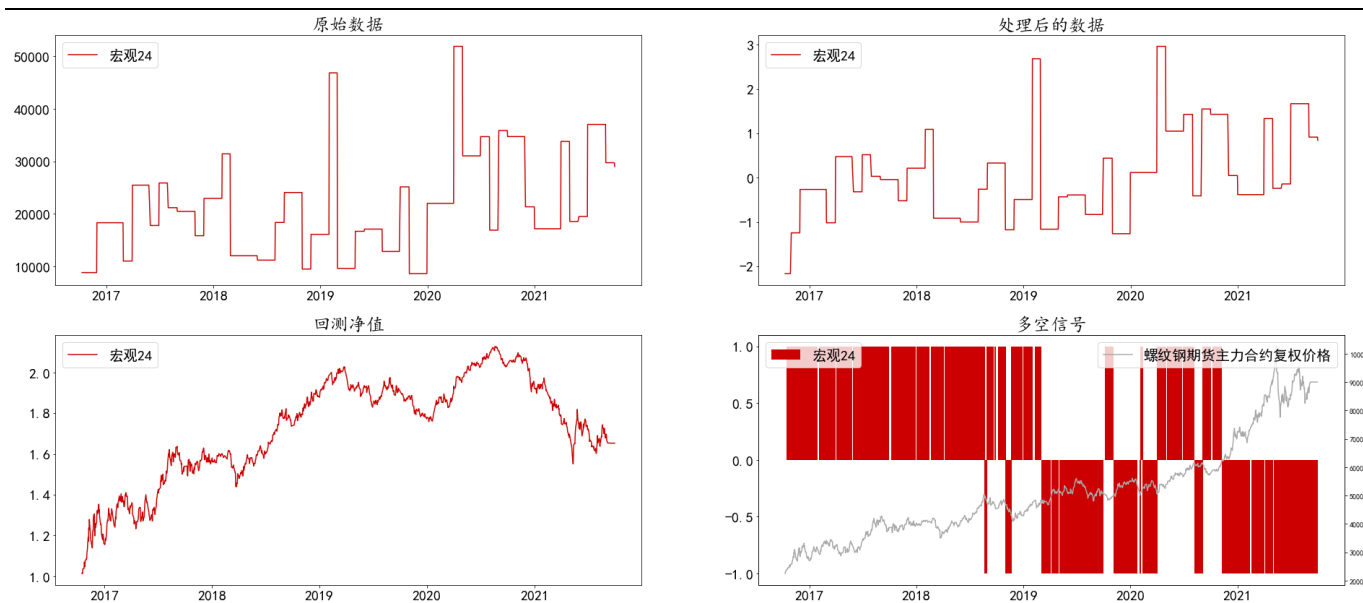
资料来源：东证衍生品研究院

图表 37 钢铁 PMI 单因子回测结果



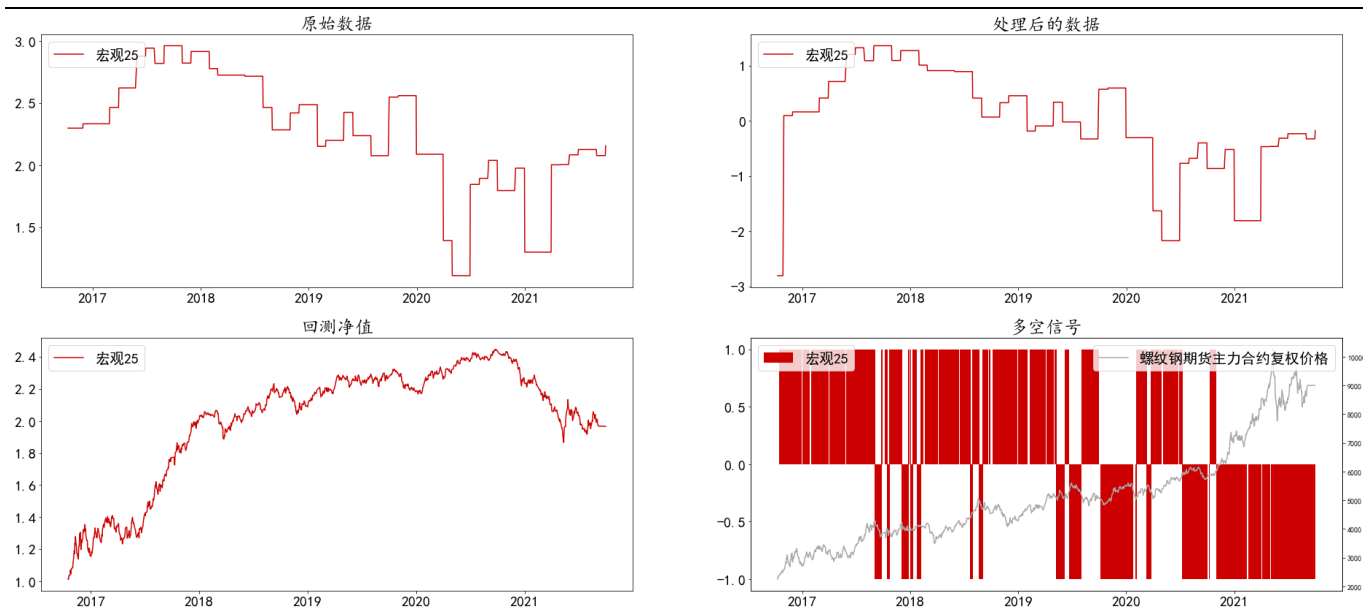
资料来源：东证衍生品研究院

图表 38 社会融资规模单因子回测结果



资料来源：东证衍生品研究院

图表 39 银行间同业拆借单因子回测结果



资料来源：东证衍生品研究院



### 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，需注明出处为东证期货研究所，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22楼

联系人：梁爽

电话：8621-63325888-1592

传真：8621-33315862

网址：[www.orientfutures.com](http://www.orientfutures.com)

Email：[research@orientfutures.com](mailto:research@orientfutures.com)