中信建投期货
CHINA FUTURES

CHINA FUTURES RESEARCH

股指期货策略：2021 年 2 月 26 日

股指套利系列(二)-基于基本面逻辑评分的价差预测效果

股指期货策略报告

摘要：

2016 年以来，IC 与 IH 当月合约的价差呈现下行趋势，上证 50 中市场龙头企业强者恒强带领指数稳健上行，5 年时间上涨 59%，而中证 500 内的中小企业受经济和政策影响较大，呈现高波动特征，在 2018 年去杠杆政策下大幅下跌，又在 2019 与 2020 年低利率环境下快速上行，中证 500 指数 5 年下跌 9%，因此两指数特征迥异，是通过风格研究进行套利的理想研究标的。

在上一篇系列报告中，我们从经济增长、外部因素、市场风格和资金面 4 个角度出发分析，用特征筛选方法(一元线性回归、逐步递回归、LASSO 和全子集回归)、非线性特征筛选方法(决策树、随机森林和 Adaboost)从宏观经济数据中筛选出显著有效性的指标。值得注意的是，这其中也包含了平时关注度不高或认为可能存在较强自相关性的指标。我们在这里将上一篇报告中筛选出的这些待定指标纳入待定因子池，同时为了使模型产生的信号便于实际策略运用，我们聚焦于分类算法领域，尝试嵌套采用 PCA 主成分分析、Logistic 回归，KNN（最邻近结点算法），K 均值聚类算法进行建模。

主要结论：

1. 使用日频基本面因子在机器学习模型中很难预测出价差走势，但是若增加因子可以微弱提升模型预测效果，更多的因子对预测效果会有提升；

2. 使用前 20 大机构持仓净多空单量可以为预测价差提供有效信息，这或是机构研究更专业和资金推动共同令期货价差出现变化的结果。

作者姓名：严晗

yanhan@csc.com.cn
电话：023-86769759
投资咨询从业证书号：Z0014172

研究助理：张仕康

zhangshikang@csc.com.cn
电话：021-58304077
期货从业证书号：F3076198

研究助理：王锴

wangkaiqh@csc.com.cn
电话：021-86769759

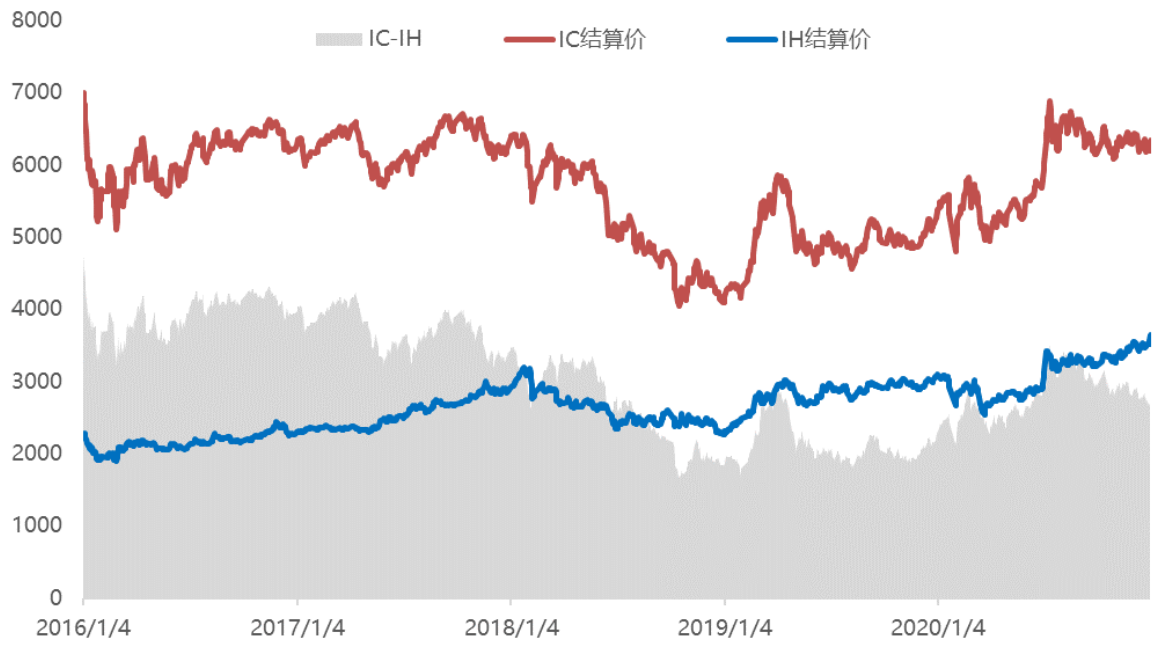
发布日期：2021 年 2 月 26 日

一、研究准备

2016 年以来，IC 与 IH 当月合约的价差呈现下行趋势，上证 50 中市场龙头企业强者恒强带领指数稳健上行，5 年时间上涨 59%，而中证 500 内的中小企业受经济和政策影响较大，呈现高波动特征，在 2018 年去杠杆政策下大幅下跌，又在 2019 与 2020 年低利率环境下快速上行，中证 500 指数 5 年下跌 9%，因此两指数特征迥异，是通过风格研究进行套利的理想研究标的。

在上一篇系列报告中，我们从经济增长、外部因素、市场风格和资金面 4 个角度出发分析，用特征筛选方法(一元线性回归、逐步递回归、LASSO 和全子集回归)、非线性特征筛选方法(决策树、随机森林和 Adaboost)从宏观经济数据中筛选出显著有效性的指标。值得注意的是，这其中也包含了平时关注度不高或认为可能存在较强自相关性的指标。我们在这里将上一篇报告中筛选出的这些待定指标纳入待定因子池，同时为了使模型产生的信号便于实际策略运用，我们聚焦于分类算法领域，尝试嵌套采用 PCA 主成分分析、Logistic 回归，KNN（最邻近结点算法），K 均值聚类算法进行建模。

图 1：IC-IH 价差走势



数据来源：WIND、中信建投期货

二、数据处理

考虑到建模过程中，数据的时间跨度和公布的时间都是影响模型构建的因素，因此为了保证策略实际跟踪中不存在时间滞后差异的问题，上一篇系列报告中我们筛选所得的指标全部为交易日当天收盘后能获得的日频数据，预测的标的为第二交易日与第一交易日的变化方向。

我们对数据进行了如下处理：

1. 取变化率：取每日因子与 IC-IH 价差的变化率，使用当日相对上日的变化率对下一个交易日是否开仓做出判断，数据范围为 2016 年 1 月 1 日-2020 年 12 月 31 日；
2. 3Sigma 筛选：将因子的最大值最小值控制在样本集历史均值的 3Sigma 范围内；
3. Na 填充：线性插值法对日频变化率中 Na 值填充，以减少部分自变量数据缺失对整体模型信息量的影响；
4. 标准化处理：对因子标准化处理；
5. 信号处理：若样本内价差变化率大于零记为 1，样本内价差变化率小于零记为-1，另外我们在 KNN 和 Logistic 模型中对信号进行混合，即用训练集准确度作为权重，加权各大类因子给出的信号，若小于样本中价差变化率中位数的 30%，记为-1，若大于 70%，记为+1，其他记为弱信号，并不给出信号记为 0。

表 1：四大类因子展示

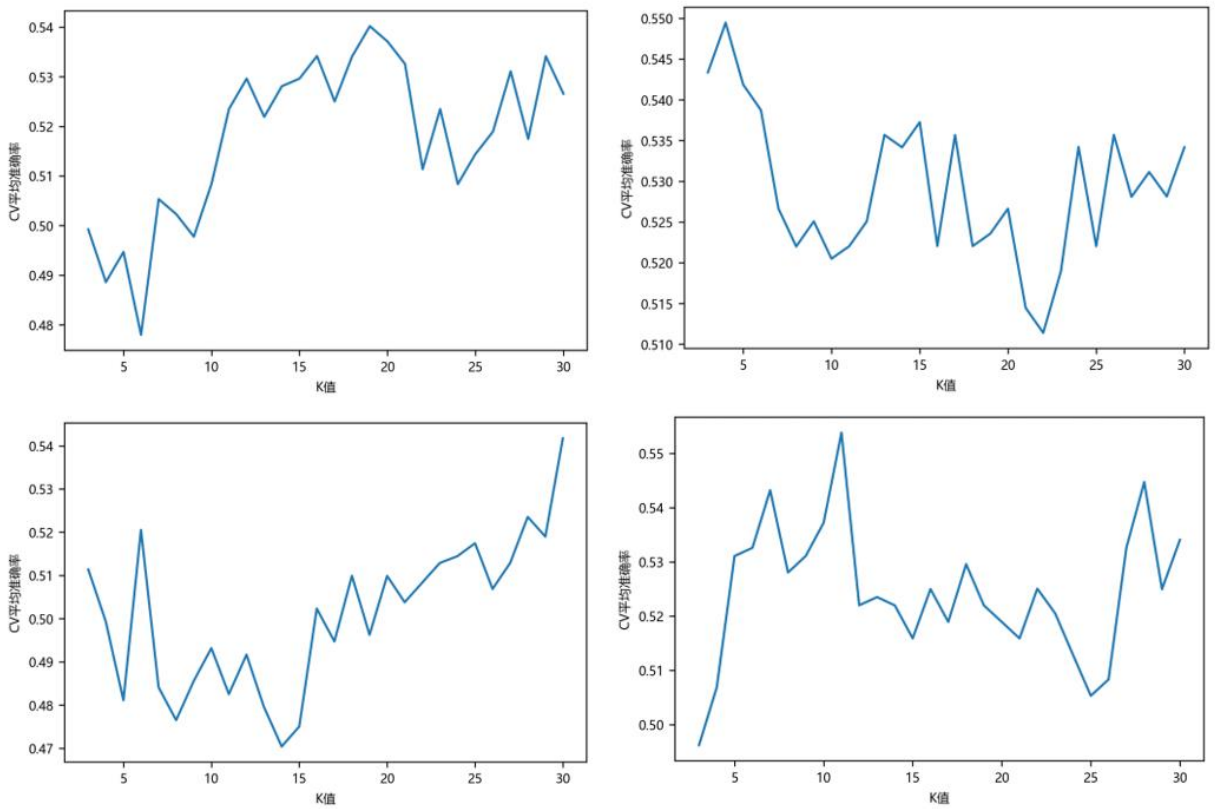
因子大类	因子
海外因素	纳斯达克综合指数
市场估值	行业 PE-TTM（食品饮料、农林牧渔、国防军工、银行、非银金融、传媒、家用电气）、指数 PE-TTM（上证 50、中证 500）
流动性	GC007、陆股通当日买入卖出成交金额、沪市融资融券余额
市场情绪	500ETF 与 50ETF 当日净流入资金的差值

数据来源：WIND、中信建投期货

三、Knn 模型

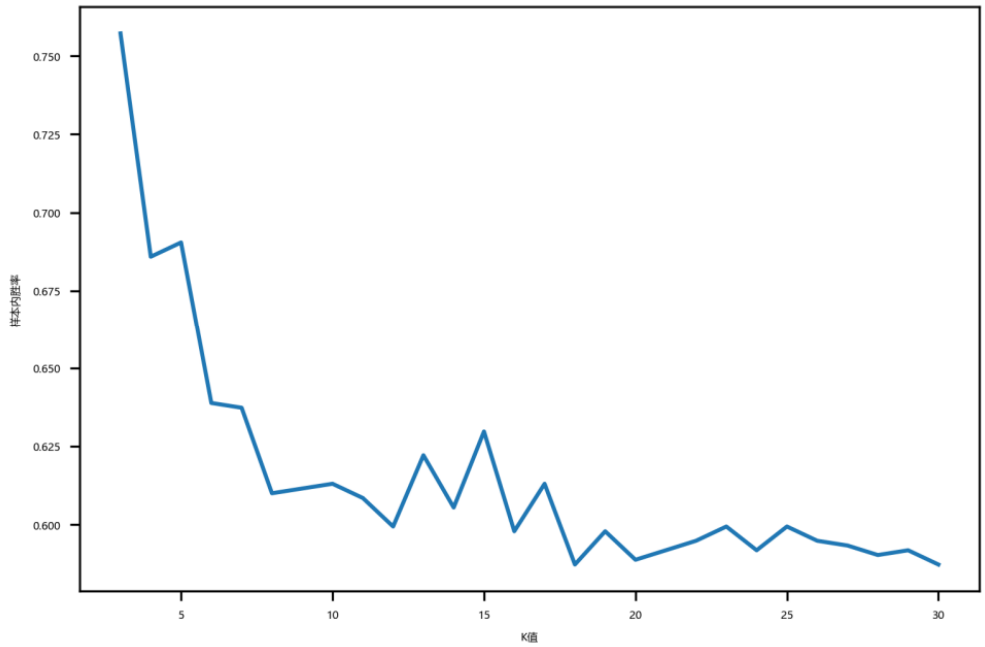
Knn 算法属于惰性算法，其特点是不必事先建立全局的判别公式或规则人，当新样本需要分类时，根据每个新样本和原样本之间的距离，取最近的 K 个样本点的众数或均值作为新样本的预测值。由于 KNN 算法对于解释变量的类型没有限制，最主要的超参数就是 K，我们运用交差验证的方法，在样本内确定最优参数 K，力求在拟合和泛化能力上取得平衡。

图 2：四类因子交叉验证确定最优参数 K



数据来源：WIND、中信建投期货

图 3：所有因子在不同 K 值下的样本内胜率



数据来源：WIND、中信建投期货

从结果来看，我们发现使用更多因子的效果虽然可以部分提升最终组合净值，但是仍然不够稳定，我们通过使用混合信号（在数据准备中提到了构建方法，即根据训练集内的胜率，加权平均各类因子给出的信号），最终稳定性仍然难以提升，胜率无法稳定达到 50%以上，我们考虑了主观经验上比较符合逻辑的策略，也就是 IC 净多单量与 IH 净多单量的差值去预测下一日的 IC-IH 价差变化率，即前 20 机构投资者更加专业，更能代表市场的看法，并且期货端也可以影响合约走势，若模型给出的信号为 1，而代表 IC 净多单减去 IH 净多单量的 Delta 小于 0，我们就不去买入，若模型信号为 1 且 Delta 大于 0，则我们买入 IC-IH 价差，可以发现该策略筛选后对最大回撤和胜率有明显的提升。

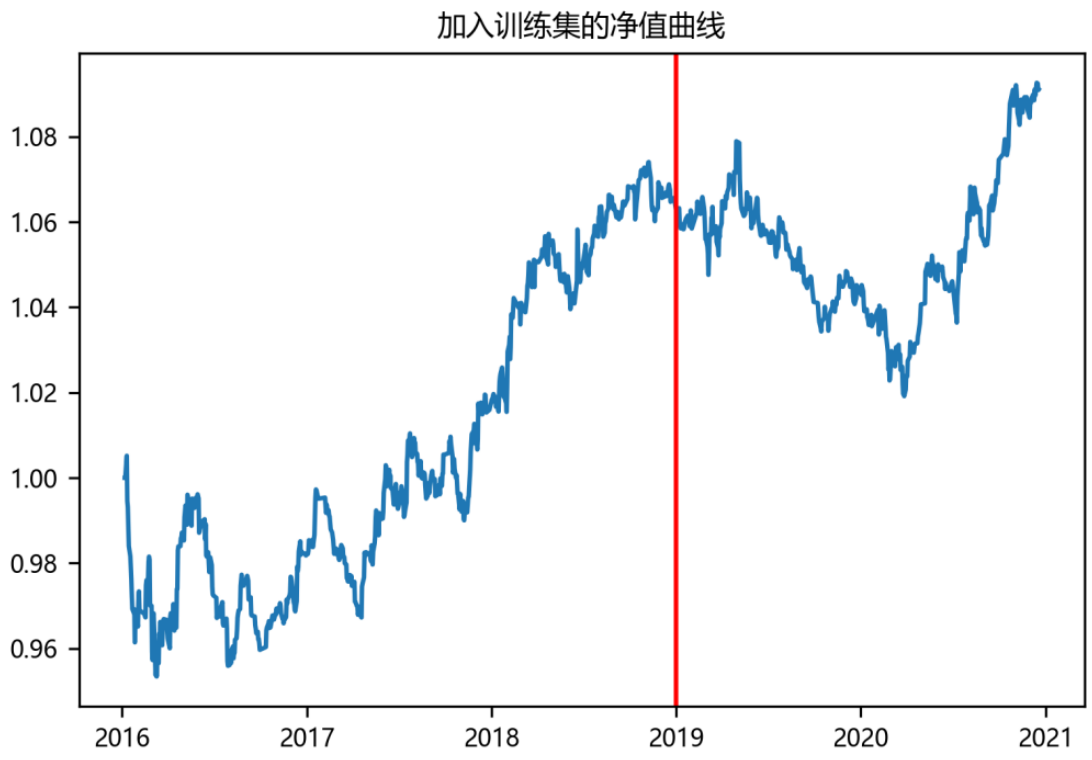
另外值得注意的是，在单一类别因子的基本面回测中，我们发现不同类型因子不同时间段表现存在差异。如果采用单因子模型难以提供稳定收益，这也符合指标在不同行情周期下的有效性规律。若将所有因子放入模型中后，我们发现其效果基本要优于单一类别，收益和风险控制上表现都有所提升。

表 2：使用最优 K 值下 KNN 模型效果

	样本内胜率	样本外胜率	最终净值	最大回撤
海外市场	0.5	0.5	0.81	0.24
行业估值	0.54	0.51	0.96	0.16
流动性指标	0.51	0.51	0.75	0.51
情绪指标	0.5	0.52	0.94	0.22
所有指标	0.5	0.49	1.14	0.26
混合信号		0.51	0.84	0.21
Delta 筛选过后的混合信号		0.53	0.94	0.12

数据来源：WIND、中信建投期货

图 3：所有指标-KNN 模型净值曲线



数据来源：WIND、中信建投期货

四、Logistic 模型

跨品种套利聚焦价差的相对变化位置，离散型变量相对于连续型变量在模型预测方面提供了更灵活的

空间，输出结果与套利的方向操作有较大的兼容性。因此我们我们对因变量做二分类处理，扩大记为‘1’，缩小记为‘0’。Logistic 回归不同于线性回归，它不要求模型变量间具有线性的相关关系，不要求服从协方差矩阵相等和残差项服从正态分布等，使得模型较为简洁高效。通常来讲，logistic 回归基于极大似然估计方法逐步选择重要的解释变量，无法避免多重共线性和对原始数据依附性的问题。由于在该案例中，原始数据已经由上述步骤所得的三个独立的主成分代替，规避了类似问题。

$$\ln\left(\frac{p_1}{p_2}\right) = \mu_1 + \omega_1\beta_0 + \gamma_1\beta_1 + \lambda_1\beta_2$$

$$p = probability(y = 1|x) \ ， \ 1 - p = probability(y = 0|x)$$

我们首先使用 Logistic 对四大类因子做预测，后将因子全部纳入模型，可以发现加入更多的因子后，不管是样本外还是样本内的胜率都是有所提升的，并且在回撤、夏普比率上都有提升，这与 KNN 模型预测结论相符，即使用更多因子可以增加模型效果。此外，我们继续使用 Delta 筛选法则可以有效提升混合信号的效果，不仅对净值有所增益，也减少了最大回撤。

表 3：Logistic 模型测试效果

	样本内胜率	样本外胜率	最终净值	最大回撤
海外市场	0.52	0.47	0.66	0.38
行业估值	0.54	0.53	1.02	0.13
流动性指标	0.55	0.52	0.85	0.26
情绪指标	0.51	0.51	0.97	0.14
所有指标	0.56	0.47	1.07	0.11
混合信号		0.51	0.89	0.2
Delta 筛选过后的混合信号		0.49	0.98	0.11

数据来源：WIND、中信建投期货

五、PCA 聚类

Pca 和 k-means 作为非监督学习的两大经典方法，能够较为客观地从高维数据中提取具有代表性的特

征，在减少白噪声干扰的同时平滑了建立模型过程中的输入条件。

为了避免多重共线性，实现数据降维的同时最大程度减少原始数据信息的丢失，我们首先引入主成分分析法，基本思路是：从 p 个相关的解释变量中提起出 k 个不相关的主成分，每一个主成分都是原始变量的线性拟合，第一个主成分最大程度地解释了原始变量数据的方差，具有最大的特征值。第二主成分与第一主成分之间不存在线性关系，它最大程度解释了剩余方差，以此类推。因此，参照（1）中的六个解释变量，我们从原始数据中提取出前六个主成分，分别可以解释样本中 38.3%、11.4%、8.0%、7.6%、6.0%和 5.0%的方差，共计可以解释 76%的方差。

我们将以上使用的所有因子先使用 PCA 处理，得到前六个能够解释因变量的因子类，再放入同样的 Logistic 模型和 KNN 模型中，最终虽然可以提升测试集胜率，但是盈亏比较低，仍然未能有效减少回撤和提升收益。

表 4：PCA 聚类处理后回测效果

	训练集胜率	测试集胜率	最终净值	最大回撤
PCA 加入 KNN	0.59	0.52	0.8	0.35
PCA 加入 Logistic	0.56	0.53	1.08	0.16

数据来源：WIND、中信建投期货

六、K-Means 聚类

K-Means 算法：随机选择 K 个聚类的初始中心；对任意一个样本点，求其到 K 个聚类中心的距离，将样本点归类到距离最小的中心的聚类，如此迭代 n 次；每次迭代过程中，利用均值等方法更新各个聚类的中心点(质心)；迭代更新后，如果位置点变化很小(可以设置阈值)，则认为达到稳定状态，迭代结束，对不同的聚类块和聚类中心可选择不同的标注。

将 KMeans 分别加入到 KNN 和 Logistic 都并没有能够显著提升模型预测效率，但是如果我们按照前面加入 Delta 筛选的话，可以提升模型预测的准确度和减少回撤。

表 5：KMeans 聚类处理后回测效果

	训练集胜率	测试集胜率	最终净值	最大回撤
KMeans 加入 KNN	0.52	0.47	0.66	0.38
KMeans 加入 Logistic	0.54	0.53	1.02	0.13
混合信号（加入 KMeans）		0.51	0.84	0.21
Delta 筛选过后的混合信号（加入 KMeans）		0.53	1.01	0.08

数据来源：WIND、中信建投期货

七、结论与展望

综上，我们可以得到如下结论：

1. 使用日频基本面因子在机器学习模型中很难预测出价差走势，但是若增加因子可以微弱提升模型预测效果，更多的因子对预测效果会有提升；

2. 使用前 20 大机构持仓净多空单量可以为预测价差提供有效信息，这或是机构研究更专业和资金推动共同令期货价差出现变化的结果。

接下来，我们将使用量价因子，并且借鉴本篇报告中前 20 机构持仓的有效信息，从资金博弈的角度去挖掘影响 IC-IH 价差的因子，并且在之后会使用到我们在第一篇报告中特征筛选出来的月度因子，结合相对短周期的模型，探究是否可以通过“看大做小”挖掘 IC-IH 价差的投资机会。

联系我们

中信建投期货总部

地址：重庆市渝中区中山三路131号希尔顿商务中心27楼、30楼

电话：023-86769605

中信建投期货有限公司上海分公司

地址：中国（上海）自由贸易试验区浦电路 490 号，世纪大道 1589 号 8 楼 10-11 单元

电话：021-68765927

中信建投期货有限公司湖南分公司

地址：长沙市芙蓉区五一大道 800 号中隆国际大厦 903

电话：0731-82681681

南昌营业部

地址：南昌市红谷滩新区红谷中大道 998 号绿地中央广场 A1#办公楼-3404 室

电话：0791-82082702

中信建投期货有限公司河北分公司

地址：廊坊市广阳区金光道 66 号圣泰财富中心 1 号楼 4 层西侧 4010、4012、4013、4015、4017

电话：0316-2326908

漳州营业部

地址：漳州市龙文区九龙大道以东漳州碧湖万达广场 A2 地块 9 幢 1203 号

电话：0596-6161588

西安营业部

地址：西安市高新区高新路 56 号电信广场裙楼 6 层北侧 6G

电话：029-89384301

北京朝阳门北大街营业部

地址：北京市东城区朝阳门北大街 6 号首创大厦 207 室

电话：010-85282866

北京北三环西路营业部

地址：北京市海淀区中关村南大街 6 号 9 层 912

电话：010-82129971

武汉营业部

地址：武汉市武昌区中北路 108 号兴业银行大厦 3 楼

电话：027-59909521

中信建投期货有限公司杭州分公司

地址：杭州市上城区庆春路 137 号华都大厦 811、812 室

电话：0571-28056983

太原营业部

地址：太原市小店区长治路 103 号阳光国际商务中心 A 座 902 室

电话：0351-8366898

中信建投期货有限公司济南分公司

地址：济南市历下区泺源大街 150 号中信广场 A 座十层 1016、1018、1020 室

电话：0531-85180636

中信建投期货有限公司大连分公司

地址：大连市沙河口区会展路 129 号大连国际金融中心 A 座大连期货大厦 2901、2904、2905、2906 室

电话：0411-84806316

中信建投期货有限公司河南分公司

地址：郑州市未来大道 69 号未来大厦 2205、2211、1910 房

电话：0371-65612397

广州东风中路营业部

地址：广州市越秀区东风中路 410 号时代地产中心 20 层自编 2004-05 房

电话：020-28325286

重庆龙山一路营业部

地址：重庆市渝北区龙山街道龙山一路 5 号扬子江商务小区 4 幢 24-1

电话：023-88502020

成都营业部

地址：成都市武侯区科华北路 62 号（力宝大厦）1 栋 2 单元 18 层 2、3 号

电话：028-62818701

中信建投期货有限公司深圳分公司

地址：深圳市福田区深南大道和泰然大道交汇处绿景纪元大厦 11I

电话：0755-33378759

上海徐汇营业部

地址：上海市徐汇区斜土路 2899 甲号 1 幢 1601 室

电话：021-64040178

南京营业部

地址：南京市黄埔路 2 号黄埔大厦 11 层 D1、D2 座

电话：025-86951881

中信建投期货有限公司宁波分公司

地址：浙江省宁波市鄞州区和济街 180 号国际金融中心 F 座 1809 室

电话：0574-89071681

合肥营业部

地址：合肥市包河区马鞍山路 130 号万达广场 C 区 6 幢 1903、1904、1905

电话：0551-2889767

广州黄埔大道营业部

地址：广州市天河区黄埔大道西 100 号富力盈泰大厦 B 座 1406

电话：020-22922102

上海浦东营业部

地址：上海自由贸易试验区世纪大道 1777 号 3 楼 F1 室

电话：021-68597013

重要声明

本报告中的信息均来源于公开可获得资料，中信建投期货力求准确可靠，但对这些信息的准确性及完整性不做任何保证，据此投资，责任自负。本报告不构成个人投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需要。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。

全国统一客服电话：400-8877-780

网址：www.cfc108.com