股债量化基本面迭代更新--IM 的增量信息扩充

国投安信期货研究院
王锴 期货投资咨询号 Z0016943

在上一期关于股债 CTA 策略的报告中，我们以量化基本面为核心，结合持仓量模型，形成了多周期结构的策略。从截面维度考虑，相较于大宗商品，股债涉及的品种池偏少，因子在横截面上排序多空信号所带来的的对冲和套利机制不太稳定，风险分散能力有限。随着今年 7 月中证 1000 指数期货合约上市，其所代表的中小市值板块与已有的 IH、IF 和 IC 合约在衍生品风险管理方面形成了有效的互补。对于以量化基本面为核心的 CTA 策略而言，由于中证 1000 指数中新能源、电力设备、TMT、医药生物等行业的新兴成长性科技企业的占比更大，这些行业受中高频经济数据变化的影响相比其他指数具有一定差异，为截面多空套利带来了更多可能性。

在之前的报告中，我们分别使用中高频的金融数据和低频的宏观月度数据对于股债轮动的相关因素进行分析，结合样本外的持续跟踪，我们发现基于市场预期的持仓量指标在更多时间上能反映出短期内风险偏好所带来的跷跷板的现象，这一周度级别的现象在今年三季度以来资金面较为宽松的格局下尤为明显。从更长的时间轴来看，长周期对于股债轮动的反应更具有领先意义，而在板块风格轮动的问题上，短周期模型对于 6 月以来风格频繁高低切换的状况也有一定捕捉和反映，主要体现在短周期在周度行情内与持仓量模型形成的联动与反转关系。

短周期模型聚焦于市场风格，外部因素，资金面三大高频金融数据板块。在原有的样本外跟踪过程中，我们发现 IC 在股指合约中收益表现一直较为稳定，尤其是在 IC/IH 风格切换中，预测胜率较高。加入 IM 合约后，同样延续了短周期模型信号对于中小市值板块的反应较为敏锐的特点。通过用中证 1000 指数代替历史期货数据的模型训练，我们发现 IM 在样本内与样本外测试集中差异不大，区间内的平均收益率和回撤均优于除 IC 以外的其他合约。结合 IC 此前样本外的表现，间接反映出目前采用的中高频金融经济数据库对于成长风格的映射更为紧密，也为之后扩充预测变量提供了一定参考依据。

在对于低频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则难以提供稳定收益，这也符合指标在不同行情周期下局部有效的规律。我们更新了长周期模型的类型，经过样本内的训练和样本外的比较，我们发现目前来看新模型仅对于 IM 合约较为有效，因此在其它合约上仍旧采用之前的 Logistic 模型进行预测。综合训练和测试集的表现，IM 的平均收益率 7.5%，略高于 CTA 策略在近 1 年的样本外跟踪年化，最大回撤在 10.83%，高于综合策略的回，但是在单品种中这一数值相对较小，对比 Logistic 模型也有一定优势。考虑到训练集中采用现货指数代替因变量拟合的信息缺失，可以认为采用 Lasso 拟合 IM 合约长周期信号是值得跟踪的方法之一。

收缩算法的延伸与运用

之前的长周期模型中，我们主要采用 stepwise 逐步递回归算法和 BestSubset 最优子集法来实现数据特征筛选，从样本外的跟踪测试来看基本没有呈现出样本内过大的差异性，但是考虑到最优子集法筛选出的靠前最有因子组合差异仍然偏大，有理由认为结构具有偶然性，并且通过逐步递回归方法得出的结果也可能存在先后顺序跳空方面的问题。因此，这次我们采用收缩算法中的 Lasso 特征筛选方法来进行尝试。

图：各合约 Lasso 模型样本内外表现

	IC_train	IC_test	IC_combine	IF_train	IF_test	IF_combine
累积收益	30.69	-8.09	20.12	57.75	-12.85	37.48
年化收益	6.35	-8.09	3.45	11.95	-12.85	6.42
最大回撤	16.07	14.20	18.55	40.75	22.43	48.14
夏普比率	0.70	-1.17	0.38	0.69	-1.27	0.37
卡玛比率	0.40	-0.59	0.19	0.29	-0.59	0.13
年开仓次数	7.45	12.00	8.23	13.86	18.00	14.57
	IH_train	IH_test	IH_combine	IM_train	IM_test	IM_combine
累积收益	44.00	-8.54	31.70	27.72	13.02	44.34
年化收益	9.10	-8.54	5.43	5.73	13.02	7.60
最大回撤	19.57	12.26	32.44	6.58	8.48	10.83
夏普比率	0.60	-1.03	0.37	0.98	1.31	1.01
卡玛比率	0.47	-0.72	0.17	0.88	1.58	0.71
年开仓次数	11.79	10.00	11.49	2.90	8.00	3.77
	T_train	T_test	T_combine	TF_train	TF_test	TF_combine
累积收益	2.02	-0.17	1.84	4.97	0.03	5.00
年化收益	0.42	-0.17	0.32	1.03	0.03	0.86
最大回撤	7.63	1.10	7.63	3.84	0.72	3.84
夏普比率	0.16	-0.14	0.13	0.59	0.03	0.53
卡玛比率	0.06	-0.16	0.04	0.27	0.04	0.22
年开仓次数	12.41	8.00	11.66	9.10	6.00	8.57

资料来源：Wind，国投安信期货

Lasso 模型是广义线性模型，在经过特征值筛选、标签设定、特征值预处理以及遍历参数之后产生日频信号。通过交叉检验的方式筛选最优参数。由于长周期宏观数据的样本较少，为避免模型过度拟合，保持样本外的泛化能力，主要从线性模型入手分析，同时为了提高预测精度，以方向判断来替代连续性数值的预测，即样本内价格到达某一区间时，定义多空信号。根据结果统计，正则化模型在宏观因子上对于模型的预测精度在测试集上表现欠佳，合约在测试集上出现回撤，因为模型训练样本容量小，以及样本的特征已经证实其有效性，降低了正则化的效果，从而减少了模型的相对鲁棒性，并在测试集上出现亏损，在信号分化较为明显的 IF 合约上出现近 48%的最大回撤。相对而言 Logistic Lasso 模型在 IM 以及 TF 合约的测试集表现相对良好。

图：品种回测净值曲线

资料来源：Wind，国投安信期货

图：品种回测净值曲线

资料来源：Wind，国投安信期货

KNN 算法的迭代更新

随着数据容量的增大，在日频的宏观数据上，使用无估计参数的 KNN 模型进行因子筛选，同时也可以规避数据量纲的影响。长周期模型关注市场预期，聚焦于宏观经济数据等低频指标。主要逻辑是从高维数据中选出有效且相对独立的因子，如果公布的实际金融数据与市场预期存在一定偏差，则部分投资者可能会根据自身投资风格在数据的公布的时间段内进行短期博弈。持仓量模型主要考虑机构多空单持仓量并进行合成，使用机构会员持仓可以为价格提供有效信息，这或是更专业的机构行为和资金推动的结果。在 2016-2021 期间，策略胜率为 57%。

从训练集和测试集可以看出，KNN 模型在股指期货的方向预测上的胜率较高，除去 TF 合约的测试集表现一般，其余合约均正收益。从信号强度可以看出多空信号相对分布

均匀，最大回撤一般出现在训练集，而在测试集的择时效果更优，表征模型具有较优秀的预测效果。在股指期货 IH 上产生 19%的年化收益，而新加入的 IM 股指期货则产生 21%的年化收益。同时我们注意到尽管中证 1000 指数作为 IM 合约的标的，价格关联性很高，但是由于样本内采用非合约标的替代的缘故，IM 是期指中唯一在测试集中收益表现比训练集差的合约。因此，未来随着 IM 合约价格数据样本的延伸，将有更多空间去解决模型在样本内的潜在过拟合问题。

图：各合约 KNN 模型样本内外表现

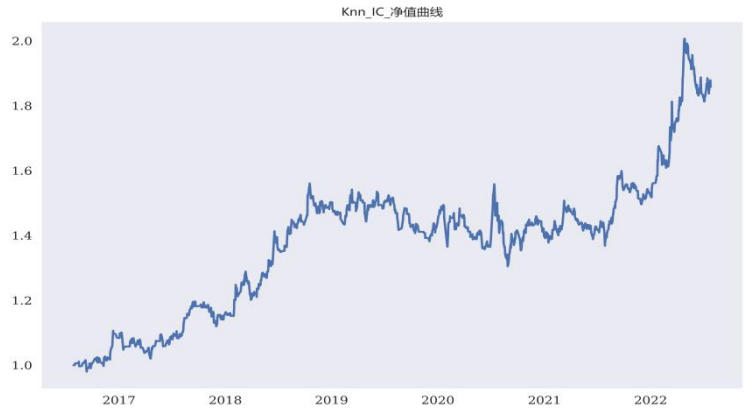
	IC_train	IC_test	IC_combine	IF_train	IF_test	IF_combine
累积收益	43.93	29.08	85.79	41.68	11.83	58.44
年化收益	9.11	29.08	14.73	8.64	11.83	10.03
最大回撤	25.57	13.49	25.57	37.84	9.10	37.84
夏普比率	0.53	1.61	0.78	0.41	0.92	0.49
卡玛比率	0.36	2.16	0.58	0.23	1.31	0.27
年开仓次数	98.88	110.00	100.79	102.61	91.00	100.62
	IH_train	IH_test	IH_combine	IM_train	IM_test	IM_combine
累积收益	94.54	9.33	112.69	99.77	12.20	124.14
年化收益	19.60	9.33	19.35	20.68	12.20	21.31
最大回撤	32.63	10.58	32.63	16.29	22.75	45.44
夏普比率	0.92	0.54	0.81	1.40	0.53	0.91
卡玛比率	0.60	0.89	0.59	1.27	0.54	0.47
年开仓次数	92.87	116.00	96.84	103.86	127.00	107.83
	T_train	T_test	T_combine	TF_train	TF_test	TF_combine
累积收益	26.42	3.92	31.38	14.38	-0.70	13.58
年化收益	5.48	3.92	5.39	2.98	-0.70	2.33
最大回撤	2.14	1.45	2.14	4.48	3.24	4.71
夏普比率	1.69	1.53	1.66	1.12	-0.36	0.90
卡玛比率	2.56	2.71	2.52	0.67	-0.22	0.50
年开仓次数	98.47	102.00	99.07	115.05	119.00	115.73

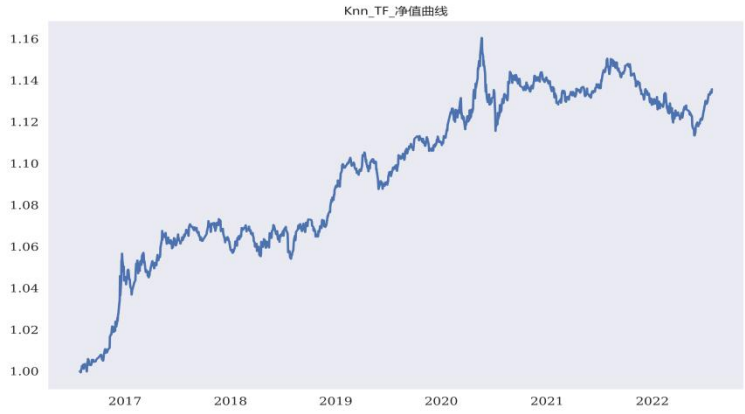
资料来源：Wind，国投安信期货

图：品种回测净值曲线

资料来源：Wind，国投安信期货

图：品种回测净值曲线





资料来源：Wind，国投安信期货

样本外跟踪概况

从今年的风格结构上来看， 1 月到 4 月下旬的下跌趋势中，美联储加息预期升温，地缘政治因素叠加疫情影响经济复苏对成长股造成较大压力，价值股占优，量化信号在这段时间的表现主要以持续做空 IC 为主，一直持续到 4 月初，期指长周期逐渐高于短周期，并且趋向接近期债，从磨底转向修复的迹象初现。

而 5 月来的反弹修复行情中，由于成长股前期估值受到明显压缩，再加上 5 月和 6 月社融放水，地产销售还未改善，因此在后期的修复性行情中，成长股表现明显优于价值股。在这段时间内，期指的长周期信号开始反超期债，开始出现间歇性的做多 IC 的信号。从最近 1 个多月的跟来看，自 IM 合约上市以来基本以做多信号为主，这也体现 7 月以来成长强于价值的风格情况，临近 8 月底，大盘整固回调之后 IH 为代表的价

值开始出现回归，IM 持仓量和短周期开始走低，以中性震荡方向。随着 IM 加入成长板块标的的补充，使原来在 IH/IC 套利波动得到放大和增强。

综合来看，IM 在加入策略品种池后，对于周期模型提供了历史平均跟踪表现以上的增益作用。短周期方面，周内的板块风格间轮动加快，均值回归与倾斜交替，以 IM 代替 IC 参与品种间的套利可以获得相对更低的相关性和更完全的对冲作用。长周期方面，新品种上市后机构资金流入量较大，市场参与度快速提升，IM 在测试区间内显著体现出股债间的资金轮动和跷跷板效应。

免责声明

本研究报告由国投安信期货有限公司撰写,研究报告中所提供的信息仅供参考。报告根据国际和行业通行的准则，以合法渠道获得这些信息，尽可能保证可靠、准确和完整，但并不保证报告所述信息的准确性和完整性。本报告不能作为投资研究决策的依据，不能作为道义的、责任的和法律的依据或者凭证，无论是否已经明示或者暗示。国投安信期货有限公司将随时补充、更正和修订有关信息，但不保证及时发布。对于本报告所提供信息所导致的任何直接的或者间接的投资盈亏后果不承担任何责任。