金融科技赋能投研系列之十一：

智 AI 科技 慧投未来（上）

投资咨询业务资格：
证监许可【2011】1289 号
研究院 量化组
陈辰
☎ 0755-23887993
✉ chenchen@htfc.com
从业资格号：F3024056
投资咨询号：Z0014257

何绪纲
☎ 0755-23887993
✉ hexugang@htfc.com
从业资格号：F3069194

高天越
量化研究员
☎ 0755-23887993
✉ gaotianyue@htfc.com
从业资格号：F3055799

这是最好的时代，也是最坏的时代。对于金融市场来说，本世纪第一个庚子年是个极为动荡的一年。这一年里，全球经济遭受了全球流行病影响。权益、债券、衍生品等各类市场均出现了大幅的波动，量化投资模式也因此受到了前所未有的挑战。由于传统量化策略所依赖时间段的数据越来越无法反映当今环境的快速变化，因此市场史无前例的动荡使得传统量化投资模型无所适从，众多量化对冲基金在多变的行情中也亏损惨重。

然而，投研的脚步并未减缓。相反，正是在这种情况下，投资领域对于新技术的研发和应用，表现出了更大的热情。其中，尤以能够快速学习市场新模式的人工智能技术，为最受瞩目的热点课题。实际上，我们观察到海内外业界明显提升了对 AI 技术开发和应用落地的投入--应用场景更为务实，创新技术逐渐渗透到投研领域的各个角落，研究目标也更趋向于多样化且多有不俗表现。

人工智能最显著的特征就是其属于前沿技术。对于前沿技术而言，从萌芽阶段到成为共识主导技术，并非完全以排他为主要特征。在具备一定成熟度的领域，我们观察到更多的情况是，新技术与原有方法的长期共存，互为补充，互为促进，反复博弈共进。最终的技术形态可能离最初预设相差甚远，甚至是新旧技术充分融合之后的全新技术形态。

基于上述的观察和思考，我们认为，正确的研究方向应该是充分理解当前 AI 技术发展的核心驱动因素，类比其成功领域的应用场景，找到 AI 技术在金融投研的最佳突破点；并以此建立对应的数据框架，特征分析模式，并将其融入到我们已经较成熟的投研框架中来。

本文将延续上一年我们开发的多周期数据分析技术，并且将更为深化的应用场景逐一展现。同时，我们将会第一次推出商品多因子模型（时序构建类型、截面回归类型），同时与同类股票因子模型做横向比较，并观察期股联动现象。最后，我们将简要分析不同类型的 AI 模型的应用场景，基于我们的特征分析方法，结合浅度和深度等不同算法优势，构建投研框架。

本文第一部分，将会结合数据处理方法和模型背景知识介绍为读者提供一个较完整的商品投研框架。

第二部分，延续我们上一年开始研发的多周期数据分析框架，做进一步的完善，并对其在大宗商品领域的应用做系统性阐述。同时，这也是我们后续商品因子分析，期股联动分析的关键性工具。实际上，到了在深度学习领域，这一套数据分析工具将和模型进一步深层次融合，为提取商品因子特征信息，量身打造深度学习模型。

第三部分，我们将推出国内期货市场的大宗商品多因子模型，包括时序基础因子和截面因子两大类。据我们所知，这是目前国内第一个推出该类型商品多因子模型。在参考了大量海外发达商品市场的研究结果的基础上，我们将利用时序回归模型针对不同商品，计算其因子（市场中性化）暴露程度，进而利用截面回归方法获得截面多因子(不可投资)收益率。这一套商品多因子模型包含了国家因子，宏观风格因子，市场风格因子以及商品板块因子。

这套多因子组的构造离不开（可投资）基础时序因子的计算。事实上，在这一基础上，模型还实现了宏观类因子的截面回归分析方法，这为我们利用商品全市场日度表现追踪低频宏观指数提供了方法论基础。我们将详细介绍这两类因子的定义、联系和区别，并且简要探讨他们各自的应用场景。

第四部分，我们将对期股因子联动的内在驱动因素进行探索。通常意义下，不同类型的市场的主导影响因素并不相同（相关性测试也将验证这一点）。然而，当我们放开测试条件，则可以进一步考察期股之间的提前/滞后等联动关系，为深入理解权益类市场和商品市场之间的内在联系提供数据基础。

第五部分，前述的数据分析方法和因子间联动关系，为我们最后应用 AI 模型提取预测信息提供了坚实的基础。为此量身定制决策树和随机森林等模型，为不同标的物锁定有效的因子组合；深度学习方面，则针对多周期分解后因子数据特征，设计合理的模型拓扑结构，利用场景学习概念建模。我们选取的主要标的物是(可投资）华泰商品板块指数。

本文是年报的上篇，着重论述了商品因子体系的构建逻辑和必要的测试方法论。在下篇中，我们会说明更多技术细节，输出完整的测试结果，并将不同模型结果做相应对比。

目录

图表目录

一、 商品投研框架

1.1 基本逻辑

量化投研方法的核心目标是通过模型化方法提取各类金融（甚至密切相关的非金融类）数据背后蕴含的对标的物未来价格判断的信息。

有两个重要假设与本文密切相关，需要深入探讨：

1) 数据集是否蕴含了标的物定价信息

2) 量化模型是否能够提取数据中的有效信息

第一个问题的复杂度较高，我们将其分解成几个层次来考虑：

首先，在经济学层面，经济运行周期对大宗商品定价至关重要。无论是从库存周期（Kitchin inventory cycle~40 个月），还是固定资产投资周期（Juglar cycle~7-11 年），其综合作用的效果将投射到各类生产要素价格的相对强弱上，并体现出价格上下波动。同时，这一类型的影响因素不仅有更加坚实的理论依据和内在规律性，也是定性判断市场宏观特征的主要参考依据。但是，另一个方面，参考这类数据的市场参与者（或规则制定者），也越发娴熟利用这些经济周期规律，甚至为了抑制系统性风险，主动参与市场的逆周期操作。这为提取中长周期宏观数据的有效特征带来了越来越大的难度。

其次，从驱动力角度来看。金融工具一般来说，都有多个驱动因素，而在不同时段不同因素的重要性也有可能发生变化，极端情况下还有可能某个因素成为绝对的主导因素，而难以体现其他因素的影响效果。所以，一般意义上，对于驱动因素的判断主要基于历史数据（针对基本面、宏观指标数据等）的统计分析；同时，结合国内政经环境和更大范围的全球经济态势做出阶段性判断。

再次，从市场博弈角度来看。任何标的物价格的形成都是交易者与其对手方在一次次的交易中形成。虽然，交易者的交易目的，持仓周期和风险偏好各不相同，但是一般都是基于明确的主观目标，并根据自身掌握信息来进行交易。一段时间内的价格形态和技术指标分析，都有助于对市场博弈情绪的判断，从而更敏感把握市场动态。

综上所述，市场的复杂度造成了我们提取市场有效信息的难度。究其原因，上面所提到的各个层面的信息最终都将叠合到市场交易行为中来形成价格，并且一般而言难以确定单笔交易的关键属性。所以，统计分析工具目前依然是金融数据分析的基础性工具；而新引入的数据分析方法是否能够更助于分解出数据中有用的信息就是我们投研方法论的一个重点研究方向。

进一步，金融市场背后的价格影响因素无外乎上面我们分析的几个主要类型，那么其中若干关键因素，我们依然可以用简化且量化的逻辑来理解—多因子模型。本文下一章将会详细介绍商品因子的制作和测试结果。就我们所知，目前国内业界针对大宗商品体系并未有此类多因子模型体系，特别是针对结构化投资模式或全市场 beta 类型风险敞口而设计的市值中性化多因子体系。所以，我们会详细介绍挑选因子的主要考虑依据、制作方法、单因子测试效果、多因子组回归结果，以及和同类型股票因子的对比分析结果。

这里我们的因子组分为四种类型：

1) 国家因子

2) 宏观因子

3) 风格因子

4) 商品板块因子

这些类型因子的挑选和制作，正是基于我们上述对市场复杂性的认识，把对国内期货全市场（从截面和时序两个角度）都较有效的指标量化为影响因子。在和同类型股票因子的对比中，可以看到这一套因子组对全品种期货市场具有很好的整体解释力，同时每个因子都代表了关键的独立风险敞口，若干类型的风格因子还表现出了很好的投资潜力（如商品价值因子）。 有趣的是，这些商品因子与股票因子之间表现出了十分紧密的联动特性（领先/滞后相关性）。更进一步，我们将利用线性模型和 AI 模型，测试商品、股票因子间的价格（波动）传导规律，以及因子对标的物的影响力强弱，从更坚实的预测性角度，为我们研究跨品种金融资产投资奠定数据基础。

第二个问题，直接关系到我们的投研框架是否具备一定效率而非仅仅历史信息的解读。由于真实的金融市场并没有一个第一驱动力的“真实”模型，所以我们观察到的各种金融数据并不能直接对“真实”模型进行拟合，从而进一步判断数据拟合程度（如参数精度），甚至估算数据噪音水平等。

相反，我们必须不断尝试不同模型去解读数据，对比它们的模型效能，利用最优模型总结出有用的市场规律，指导我们的投资行为。随着算法技术的高速发展，模型优化迭代的竞争态势越发激烈。本文我们将看到多种跨学科技术的整合应用，汇聚来自传统金工领域的统计类型时序模型、信号分析方法和受到高度关注的 AI 技术等，挖掘多种新技术在投入到金融领域以后所发挥的各自效能。

这里我们关注的重点将集中在不同方法对比的结果。首先，对于跨领域的技术移植而言，技术的适用性是我们最为关心的，那么是否能发挥出比原有技术更好的效率就是一个比较客观的评判标准。其次，技术的跨领域应用往往还涉及到算法本身的逻辑深化，信息提取方式的优化，以及特征工程优化等方面。这将是大量新旧技术融合的地方，也是本文将会深入讨论的部分。

1.2 因子选取

商品因子一直是一个讨论热度很高，但却体现出较大分化的领域。这方面，海内外都出现了大量的参考文献，在投资领域更是热度难减。这其中最主要的出发点是基于大宗商品的经济学特征和市场交易特点出发，选取合适的多因子组，既能最大程度描述市场的系统性收益/风险敞口，又能作为数据基础为新型模型开发铺平道路。下文将详细介绍我们推出的国内全商品市场多因子模型构建方案，并在年报下篇展示更多（技术性）测试结果。

1.3 模型基本方法

本文将对比多种模型方法进行数据处理，因子效率判断，期股联动观察，因子解释力&预测能力的研究结果。

- 数据处理：

  沿用我们之前开发的多周期数据分解方法[1-4]，将相关金融数据按周期长度分解为长/短周期，分别建模分析。

- 相关性特征：

1) 使用传统金工方法分析标的物与因子之间，期股因子之间相关性；

2) 并利用协相关性的方法，观察上述因子间领先/滞后相关性

- 因子对标的物的重要性判断及回测结果：

1) 随机森林模型（Random Forest）；

2) 基于进一步特征分析的深度学习模型。

二、 数据预处理

2.1 多周期数据分解

从上一年开始，我们就系统性的引入了多周期数据分解的方法，已累计了相当数量的研究报告，包括：高频策略研究；商品因子研究；CTA 策略开发；套期保值研究等。该方法持续深化、推广的主要原因是从数据预处理的层面，多周期数据分解方法就能最大程度帮助我们拆分数据中不同周期范围上的主导驱动因素。现在该方法已经全面融入了我们的研究体系，实际上，其深刻地改变了我们看待数据的角度和提取信息的方式，比如深度神经网络模型就是根据这样的数据分析方式而量身定制了模型拓扑结构。

金融数据是一个低信噪比的系统。数据之间的关联性几乎难以通过统计方法在原始数据中挖掘，即使一段时间内出现的高度相关性，也难以保持其稳定的联系而外推到预测场景中去。所以，为了降低数据噪音，并提取不同周期上的数据特征，我们在对基础因子数据，截面类型商品因子数据分析研究时，都将对数据做适当的多周期分解，并在不同周期尺度上分别观察标的物的时序特征、挑选影响力较强的影响因子、分析因子之间的相关性等。

我们使用测试数据的历史长度是 2010 年以来的国内商品期货数据，数据量较少，难以使用中长周期数据进行测试（如月度数据）；而日度数据则又频率过高，从更广的实际投资角度来说有较高的门槛和技术性限制，所以本文主要给出周度的测试结果。我们的数据分解将分出短周期和长周期两类。

预览已结束，完整报告链接和二维码如下：
https://www.yunbaogao.cn/report/index/report?reportId=1_1065