金融科技赋能投研系列之一： 人工智能策略在商品市场的应用

在我们的投研体系中，基本面数据和行情面数据就像我们信息来源的阴阳两极，如何能够在两者之间搭建模型的桥梁，使彼此之间既能有研究结果的互相印证，又能衍生到投资研判的互为指引，是我们量化投研框架建设的核心问题。

特别地，在新世纪第二个十年接近尾声之时，中国对全球贸易的影响力与日俱增，各类行业统计数据越发细化，数据采集更趋数字化、高频化，甚至另类数据也开始逐步纳入大型金融机构的量化模型之中。如何将繁杂而多样的基本面信息进行量化，并最终在有效投研框架下提炼出最具参考意义的投资研判是我们急需解决的具体课题。我们认为黄金在这里具有非常好的样本性质：标的物内在属性复杂；交易规模庞大；交易主体和投资目的层次多样；历史行情数据经历多个经济发展周期。而黄金与基本面因素之间的（周期性）互动关系的研究历史悠久，可以量化印证的内容非常丰富。所以我们选择以黄金作为研究标的物。

本文将围绕黄金的行情数据和基本面量化指标的数据处理方法，各类因子对标的物周期性影响力的量化分析，并最终导向投资策略开发的顺序来展开论述。

文章第二部较系统性地介绍了我们引入的跨学科数据处理方法--突破了过往的研究方法，既能更加有效处理海量数据，又能够从不同的侧面重新挖掘我们最需要的投研信息—基本面数据中的周期性、多尺度波动性，因子与标的物之间的相关性等等。第三部分简要介绍了目前传统线性时序方法和 AI 算法在金融模型开发中的应用。

第四部分，则将前述方法论投入到实证当中，对黄金及各类型因子进行分析研究。我们重点对比了线性模型与 AI 模型在因子解释力和因子预测性能判断之间的差异性。

本文最后探讨在前文研究的框架下--多种类型因子，复杂周期特征的条件下，如何融合不同因子信息做策略导向的研发，并且从典型因子影响力周期的角度来分析策略的收益和风险来源。AI 模型的角度，我们最后选择了随机森林模型建立量化投资策略，并给出历史回测结果。

投资咨询业务资格：
证监许可【2011】1289 号
研究院 量化组

罗剑
量化研究员
☎ 0755-23887993
✉ luojian@htfc.com
从业资格号：F3029622
投资咨询号：Z0012563

陈维嘉
量化研究员
☎ 0755-23887993
✉ chenweijia@htfc.com
从业资格号：T236848
投资咨询号：TZ012046

杨子江
量化研究员
☎ 0755-23887993
✉ yangzijiang@htfc.com
从业资格号：F3034819
投资咨询号：Z0014576

陈辰
☎ 0755-23887993
✉ chenchen@htfc.com
从业资格号：F3024056
投资咨询号：Z0014257

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

目录

图表目录

一、 基本面量化及因子选取

1.1 基本面量化的复杂性

在我们的投研体系中，基本面数据和行情面数据就像我们信息来源的阴阳两极，如何能够在两者之间搭建模型的桥梁，使彼此之间既能有研究结果的互相印证，又能衍生到投资研判的互为指引，是我们量化投研框架建设的核心问题。

特别地，在新世纪第二个十年接近尾声之时，各类行业统计数据越发细化，数据采集更趋数字化、高频化，甚至另类数据也开始逐步纳入大型金融机构的量化模型之中。如何将繁杂而多样的基本面信息进行量化，并最终在有效投研框架下提炼出最具参考意义的投资研判是我们急需解决的具体课题。我们认为黄金在这里具有非常好的样本性质：标的物内在属性复杂；交易规模庞大；交易主体和投资目的层次多样；历史行情数据经历多个经济发展周期。而黄金与基本面因素之间的（周期性）互动关系的研究历史悠久，可以量化印证的内容非常丰富。所以我们选择以黄金作为研究标的物。

黄金是最早进入流通市场的投资品种之一。不同于普通商品，黄金的开采以及货币化流通，在推动人类社会发展历史中发挥了极其重要的作用。上个世纪，黄金经历了战后形成布雷顿森林体系，到 1971 年之后该体系的瓦解，黄金与美元脱钩。黄金作为金融投资产品，则继续发挥着抗通胀功能和避险功能，其本身多重内在属性最终在交易层面获得了更全面、更具象化的体现。

商品、货币和投资避险是黄金的三大基本属性，分析黄金价格的影响因子，也需要从这三大基本属性出发：

- 商品属性，金价受到黄金商品供求关系以及其他大宗商品的影响；
- 货币属性，金价受到全球货币体系变化、全球货币政策变动影响；
- 投资避险属性，金价易受国际政治和经济发展周期的影响，进而表现为与全球权益类市场、债券、房地产等大类资产投资的风险程度变换。

黄金的不同投资功能会在不同的经济发展周期，体现出不同投资者的投资意愿，从而表现出其价格波动主导因素的不断切换。而这些复杂的外部影响因素--从量价影响因子到经济循环周期等不同维度上，都体现了黄金定价的复杂性。

另一个角度来看，还需要考虑到目前全球政治风险可能带来的短时剧烈冲击。回望整个 2019 年，是世界政治格局相当动荡的一年。大国政治角力，区域政治变幻莫测，地区武装冲突依然充斥着新闻媒体的主要版面。对我国而言，经济周期带来的 GDP 增速减缓，叠加中美贸易摩擦已经越来越引起国内政商界的重视。需要特别指出的是，中国（GDP）目前已是

全球第二大经济体，且与跟随其后国家之间的差距越拉越大，在防范全球金融风险方面，也开始实施自己的应对方案：中国央行今年以来，1-10 月已经连续 10 个月购入黄金，其黄金储备总量也位列全球**第 7**。

表格 1：主要经济体央行黄金储备

排名	国家	官方黄金储备（吨）	黄金储备对外汇储备总额占比（%）
1	美国	8133.5	76.9
2	德国	3366.8	73.0
4	意大利	2451.8	68.4
5	法国	2436.1	62.9
7	中国	1942.4	2.9
9	日本	765.2	2.8
19	英国	310.3	9.3

数据截止 2019-11

资料来源：International Financial Statistics（IMF），华泰期货研究院

不难看出，在世界主要经济体、SDR 货币发行国中，目前中国的黄金储备对外汇储备总额的占比依然很低。我国在应对全球金融风险能力的提升上还有相当大的空间。尽管目前，中美贸易争端有所缓和，但是，至少中期来看（2020 年 11 月美国大选之前），中美贸易摩擦依然具有高度不确定性，政治风险依然是影响金价的重要因素之一。

1.2 因子选取

黄金定价是个非常复杂的系统，其价格形成体现了不同层次交易者对黄金不同金融属性投资的叠加。**本文以 LME 的黄金现货作为标的物**，并对其交易价格进行预测。

从历史数据时长和交易时间来看，LME 的黄金现货发展历史远远久于内盘的黄金期货，而且 LME 的黄金现货周一至周五每天 1:00-20:00（伦敦时间）可以在 LME 交易，而且 24 小时都可以通过电话交易，可交易时间完全覆盖内盘，具有更好的价格连续性，历史数据时间

较长。从交易规则角度上看，LME 的黄金现货没有交割期限，可以一直持有，没有涨跌停的限制，在黄金的内在价值发生巨大变化时，市场能够较快的调整到合理价位，是内盘黄金期货价格的重要参考因素。从影响力来看，LME 作为全球最大的黄金交易市场，向全球开放，透明度更高，每日的成交额将近几万亿美元，交易量极其庞大，其价格趋势更能体现黄金的内在价值变化。相比之下，内盘的黄金期货影响力较小，由于交易时段的连续性不足还会时而发生跳空的现象，虽然能够一定程度上反应国内的黄金市场需求，但使用伦敦金作为分析的依据更优。综合各方面，LME 黄金现货更加符合各模型的假设，其可分析性更强，更适用于技术分析。所以我们选择 LME 的黄金现货作为标的物进行分析。

本文仔细梳理了过去学者研究黄金的文献，从黄金的大宗商品、货币以及投资避险三大属性入手，筛选不同类型的（指标）因子，从多个侧面反映黄金价格背后的影响因素。由于不同因子在不同的历史阶段、经济发展周期，对金价的变动可能表现出迥然不同的影响强度。因此在因子选择上，充分考虑了市场上对金价有显著影响力的因子作为测试对象。其中，宏观类型因子包括：美国通胀；美元货币供应量；高盛商品指数；实际无风险利率；信贷风险等。交易类型的指标我们考虑：美国国债收益率；美元指数；美元人民币汇率；S&P 500 指数等。更多交易类型数据包括：金矿商&主要黄金相关公司股票价格等。

表格 2：黄金定价因子选取

序号	因子类型	因子名称	本文简称
1	宏观	美国消费者价格指数	CPI
2	宏观	美元狭义货币供应量	M1
3	宏观	美元广义货币供应量	M2
4	宏观	信贷风险	Credit_Risk
5	宏观	实际无风险利率	Free_risk_rate
6	宏观	高盛商品指数	Comdty
7	宏观	CRB 金属指数	CRB_metal_m
8	宏观	美国通胀率	Inflation
9	交易型	美国十年期国债利率	yield10y_m
10	交易型	美元兑人民币汇率	USDCNY
11	交易型	美元指数	USD_Index
12	交易型	S&P 500 指数	Equity_Market
13	交易型	巴里克黄金公司	ABX
14	交易型	纽蒙特黄金公司	NEM
15	交易型	金罗斯黄金公司	KGC
16	交易型	伊格尔矿业公司	AEM
17	交易型	美国国债跨期利差	Yield_10y_2y

数据来源：华泰期货研究院

1.3 模型基本框架

本文将采用以下方法对黄金的周期性、因子影响力以及模型预测效果进行剖析：

- 数据处理：

1) 将数据按频率周期分解为长、中、短三个时间序列，分别建模分析；

2) 用递归图分析市场定价的结构性变化，选取数据起始时间；

3) 观测价格波动周期性，宏观因子（交易或非交易型数据）的周期性：

- 标的物价格的波动性特征、因子影响力特征：

1) 常规自相关（ACF），协相关模型（CCF）

2) 决策树，随机森林模型和其他森林模型（Decision Tree、Random Forest、Xgboost...）

- 标的物价格预测模型及回测结果：

1) 常规 ARMAX 模型

2) 决策树，随机森林模型和其他树模型（Decision Tree、Random Forest、Xgboost...）

本文中，我们将主要对比业内已经比较流行的随机时间序列模型，与目前发展迅速的 AI 类型模型之间的不同表现。

二、 数据预处理

2.1 按照周期分解数据

金融数据是一个低信噪比的系统。一般来说，交易型数据都很难在统计学意义上显著拒绝随机游走假设。实际上，目前很多流行的量化方法还是在随机游走的框架内进行。但是，这为我们的研究带来了很大的难度：

1) 原始交易数据包含了丰富但是杂乱的信息，数据降噪是策略研究的关键；

2) 随机游走的鞅性质（Martingale）导致我们对价格的量化预判一般情况下没有统计意义

3) 由于噪音的严重干扰，因子虽有较高的解释性，但预测价值较低；

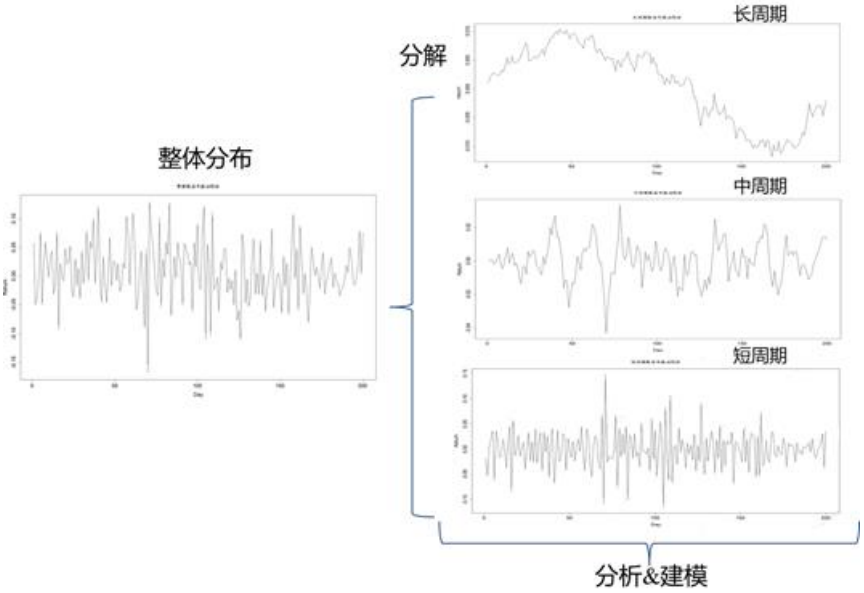
4) 当标的物的市场定价方式发生了结构性的变化，原本有效的交易策略有可能失效，但使用未处理的金融数据分析很难捕捉到该信息。

针对以上问题，常用方法为降低样本数据提取频率（如日度数据转化为月度数据），均值方法和趋势&周期分离等方法为主，有效去除日间价格波动的市场微观特征，在一定意义上有利于突出经济学因素的影响，但也失去了精确分析市场博弈，情绪化分析的机会。

本文将科研领域（物理学，海洋学、生态学等）中使用较为成熟的数据处理方法引入到金融数据处理中，通过按照周期分解数据，将数据按照周期长度分解为多条不同时间序列，就是将不同频率，甚至不同类型的数据与标的物价格波动做不同周期的匹配，进而测试其相关性（或影响力）的强弱，挖掘多层次的数据特征；相对而言，传统数据处理方法可能较难做直接的观测和定量信息提取。:

- 中长周期数据较好地保留了宏观面层次上影响因子、标的物价格波动特征
- 短周期数据便于捕捉价格序列微观上的变化，加强对市场行情发展的把握。

图 1：按照周期分解数据



数据来源：华泰期货研究院

在分析对贵金属影响力较强的宏观因子数据时，需要判断出宏观因子的类周期性表现。因此，选择数据处理（特别是较长历史数据处理时）的方法时，不能忽略市场发生结构性变化的可能性。因为，在市场发生相变的前后，任何模型都不能保证一定有相同甚至接近的参数。

所以，我们首先使用 RQA 的方法分析市场的历史相变点，寻找可以稳定估算周期的时间段。然后，针对可能出现的波动周期，对数据进行分解，并且保持分解后数据的稳定性（stationarity）。最后，再使用不同的线性或非线性的方法去测试数据中是否存在可观测的类周期现象。

本文将在不同周期尺度上，结合提取的信息，探索不同类型因子在特定周期尺度上对标的物的影响力，从而挖掘出对应该周期尺度上最合适的因子组合，进而建立因子模型。后文中实验结果也说明各个周期尺度上的不同因子影响力差异十分显著。而预测的结果将是各个周期尺度上预测结果的叠加。

2.2 市场结构性相变研究

在研究时间序列数据时，数据往往会受到外界的多种因素影响，从而表现出看似无规律的变动。但实际上，这些数据内在高度结构化，携带着丰富的信息—比如周期性，相变点等。为了能发掘数据的一些本质上的统计特征，我们需要根据这些特性去寻找最有力的分析工具。其中一个强有力的工具是 1987 年，Eckmann 提出的递归图模型[1]，它在非线性时间序列的研究上起着极其重要的作用。

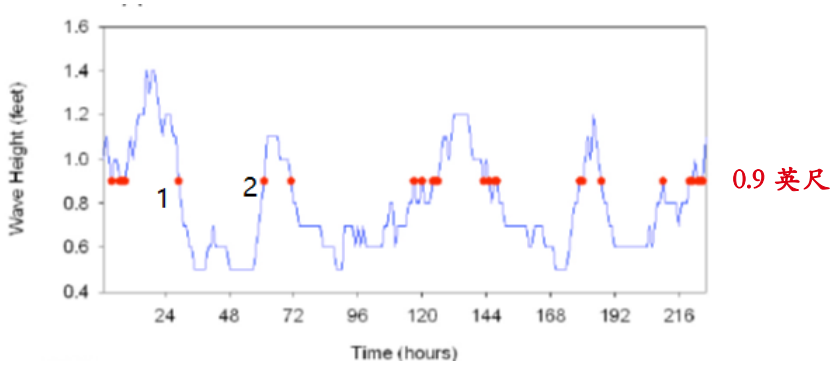
该模型主要研究的是重现（Recurrence）。即在真实世界中，存在一些场景或相似场景的一次又一次的重现，虽然并非精确周期性。在数据处理中，这种重现的现象可以使用递归图来描述。递归图矩阵的定义如下，在空间中有两个状态$\vec{x_i}$和$\vec{x_j}$分别发生在 i 时刻和 j 时刻，那么递归图矩阵可以显示为：

$$R_{i,j} = \begin{cases} 1 & if \left\| \vec{x_i} - \vec{x_j} \right\| < \epsilon \\ 0 & otherwise \end{cases}$$

其中ϵ是阈值。

接下来，我们以研究海浪的数据为例来展示递归图的结构形态。图 2 是海浪的高度随着时间变化图[2]，数据来自 Islip(纽约州，美国)海岸线以南 33 海里的观测站。这些数据初看，似乎没有太多规律可循。

图 2：海浪随时间变化图

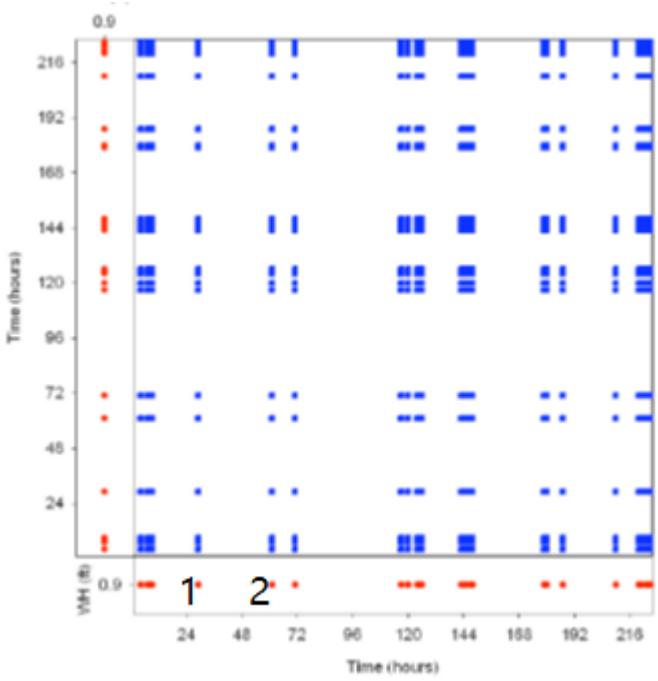


数据来源：参考文献[2] ，华泰期货研究院

我们截取海浪在高度等于 0.9 英尺的点（即图中的红点），然后以横纵坐标为时间轴，按照红点在海浪图中的时间排列顺序，在图中描出对应的点。图 3 则为对应递归图。为更精确的描述图 3，将两个状态x_i和x_j分别发生在 i 时刻和 j 时刻，那么递归图矩阵可以表示为：

$$R_{i,j} = \begin{cases} 1 & if \ \ x_i = x_j = 0.9 \\ 0 & otherwise \end{cases}$$

图 3：在海浪高度等于 0.9 英尺时的递归图



数据来源：参考文献[2]，华泰期货研究院

通俗地讲，递归图中每个点代表不同时间节点处在空间上观测值相近的点，空白处代表横、纵坐标对应的两个时间点在空间上的观测值相差较大（超过阈值ϵ）。递归图的定义决定其沿着对角线对称的性质，同时也能捕捉到任何时刻海浪高度是过去哪些时间点的复现，能较好的捕捉到不规则的周期特性。换言之，通过递归图从原本冗杂的数据中提炼出一些周期性的信息

海浪图 2 中点 1 和点 2 对应的区域即为递归图 3 中点 1 与点 2 对应的区域，且红点代表所有海浪在高度等于 0.9 英尺的点。沿着递归图的对角线看，可以很清晰地看到递归图上的点与空白地区以一定的周期性交错排列着。通过这个简单的例子，可以看到递归图在处理复杂数据尤其是时间序列数据上的有效性。

在对递归图做定量分析时，通常有如下几种指标可供参考：第一个指标是递归率（recurrence rate，RR），它计算的是在递归图除了对角线以外的部分，递归点在所有点中的占比。它的计算公式如下：

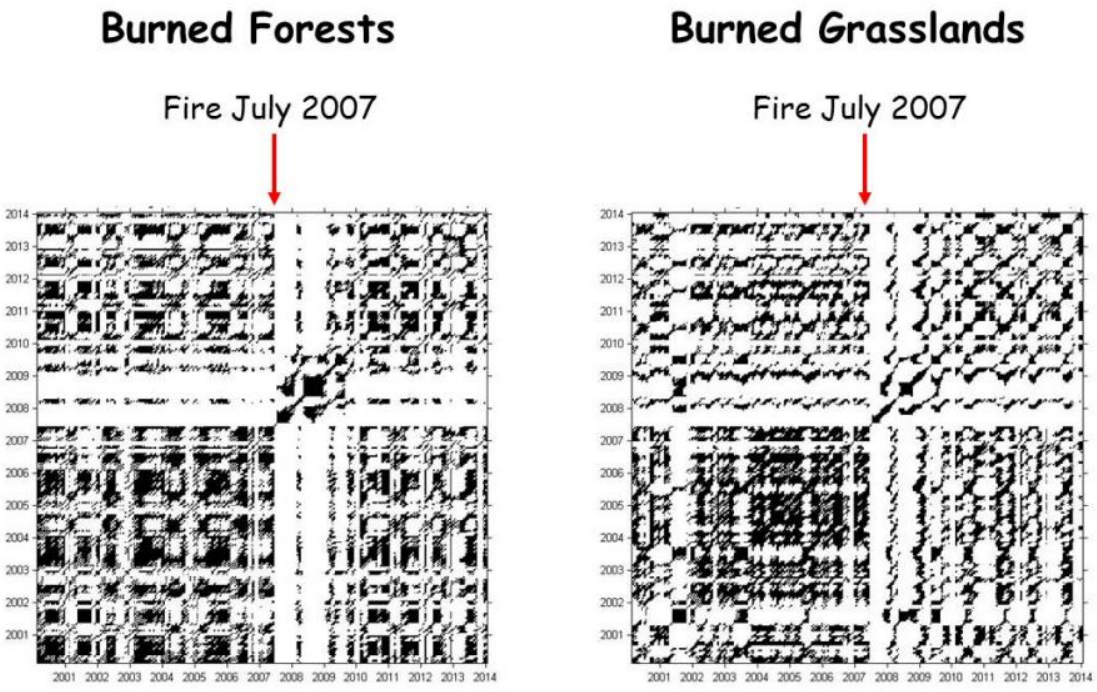
$$RR(\varepsilon,N)=\frac{1}{N^2-N}\sum_{i\neq j=1}^{N}R_{i,J}^{m,\varepsilon}$$

其中 N 为递归图边长的点的个数，ϵ是阈值，i,j 分别为递归图的横纵坐标，m 为递归图的维度，在海浪高度的递归图中 m=2。

其次，还会有诸如确定性（DET，递归点在递归图的对角线中的占比），层流率（LAM,递归点在递归图中垂直线的占比）之类的递归图指标，因篇幅有限，在此不做详细叙述。通过对这些指标的运用，我们就可以从定量的角度来评估递归图以及发现背后的规律。其中 RR 指标是我们用到金融数据中的关键指标（参考下文）。

接下来我们再看一个例子，这个例子更加深刻的揭示了系统在相空间运动中，递归图发挥的重要作用。（相空间是一个用以表示系统所有可能状态的空间；而目标系统所处状态随时间的变化轨迹则为其在相空间中的一组时间序列。）人们通过美国宇航局的空间遥感仪 MODIS，收集地中海区域植被的覆盖情况。并且通过植被指数（EVI）来分析森林，草原等植被覆盖，在不时发生野火的状况下，出现的植被延续、过火后恢复和植被结构性变化等特征。植被指数的递归图显示了如图 4[3]的场景。在大火烧过的前后（2007 年 7 月之前及 2009 年 1 月之后），森林和草地的递归图都表现出了比较显著的类周期性特征—系统稳定，具有一定可预测性。而 2007 年 7 月至 2009 年 1 月之间的白色的断层带则清晰表明了该时段，植被系统正在经历重要的相变（野火发生时及后续生态恢复期），森林和草原植被指数在其他时段都没有重现。此例子中的相变特征正是我们寻找金融数据相变点的关键依据之一。（黄金历史相变点，参考 4.2 节）。

图 4：发生火灾前后森林和草地的 EVI 递归图



数据来源：参考文献[3]，华泰期货研究院

递归图在处理时间序列数据上还有其他优势，如在处理数据方面，递归图不仅可以处理噪音数据，还可以测试时序数据的稳态性质。

三、传统算法与人工智能算法

3.1 传统线性模型

在线性模型方面，我们主要使用自相关（ACF）、协相关（CCF）、ARMAX 模型。所以，这里我们简要介绍一下此类型模型，以及在我们目前研究问题范围内，该模型的优势和不足。

对于任意一个时间序列来说，如果按一定时间顺序取得的观测值之间存在相关性，那么我们称该时间序列“自相关”。金融数据的时间序列常常符合自相关的特征，如 GDP、价格指数、失业率等常见指标都会呈现出一定的惯性作用。由于宏观经济的内在规律驱动，这些时间序列当前数据的取值往往与上一期或者上几期取值有关，一些过往的信号通过类似反射、折射等方式延时后，仍有着对后续时间序列的影响力，从而使得序列数据表现出较强

的延续性质。

对于金融时间序列的自相关性分析，能够让我们在给定的置信区间内把握其内在的规律性质，同时对于不同维度自相关性的分析，也能够让我们对影响时间序列背后因素的内在结构做出较为可靠的猜测。但值得一提的是，统计意义上显著的自相关性也不一定意味着完全的线性自相关性，仅仅依赖自相关性的分析，我们只能得到对时间序列较为粗浅的、非结构化的理解。在此基础上，通过对内在结构的猜测，选择适当的分析方法是自相关分析能够提供的更为有效的逻辑链条。

如果说自相关性是探寻时间序列本身的周期性结构，那么协相关性更多的则是寻找两个不同时间序列背后相同的周期分量。同样回到对于金融数据的分析，GDP、价格指数、失业率这些指标往往只是经济周期在高维维度上的投影。通过对两个时间序列相关系数的分析，我们能够对两个序列间的联动方式做出一定猜测。但与自相关性一样，协相关性对于时间序列的理解也较为片面，一方面，相关系数低只能说明可能不存在线性相关性，但不能保证数据间不存在更为复杂的相关关系；另一方面，一对时间序列组合的相关系数并不能直接简单的横向对比，因此在把握多时间序列间关系时协相关性也存在一定问题。

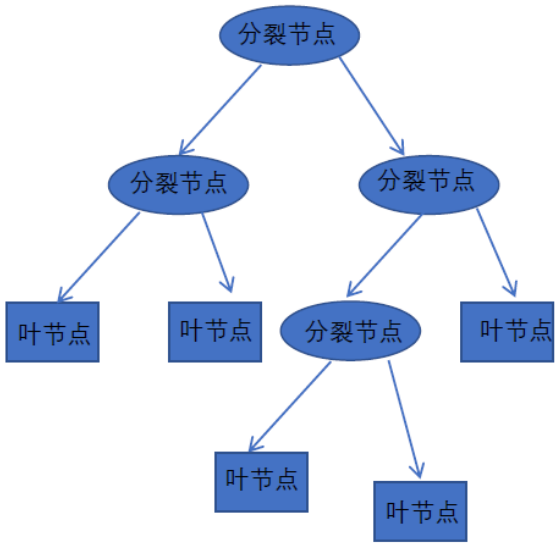
3.2 决策树、森林模型介绍

在非线性模型方面，我们主要使用决策树，随机森林及其衍生模型来作为比较对象。所以，本节将简要介绍此类型模型，以及在目前研究问题范围内该模型的优势和不足。

决策树是一种呈树形结构的分类回归方法，它可以看作是一种以实例为基础的归纳学习算法。决策树通常用来处理一些无次序无规律的事例，并通过自身的分类规则，形成分类器和预测模型。

决策树的本质其实是对于一个给定特征空间上的划分，树上的每一条从根节点到叶节点的路径将特征空间划分成互不相交的区域。决策树的结构如下图，树中一共有两种节点，内部节点（圆形）代表一个属性的分类，叶节点（方形）代表一个类。从根节点开始，对输入例子的某一特征进行测试，根据测试结果将其归到某一个子树中，递归进行直到抵达树的叶节点为止，其叶节点的类别即为输入例子的类别。

图 5：决策树原理示意图



数据来源：华泰期货研究院

在决策树的学习过程中，需要让训练数据按照设置与标准归纳出一组与其吻合的规则，同时又具有较好的泛化能力。这种规则有很多种，因此我们需要采取一些策略去解决这个问题。在决策树中，通常会建立一个正则化后的损失函数，采取最小化损失函数的方法来解决这一问题，比如 ID3、C4.5、CART 等算法。ID3 算法引入了信息熵的概念，通过比较信息熵的大小来决定分裂的节点。信息熵是消息发生后所包含的信息量的数学期望，其计算公式如下：

$$E(x) = P(u_1)I(u_1) + P(u_2)I(u_2) + \cdots + P(u_n)I(u_n)$$

其中，u_n为每个信息，n 为信息的个数

C4.5 算法是在 ID3 算法上的改进。与 ID3 不同的是，C4.5 算法引入分裂信息指标，通过对信息增益率的比较来确定应该分裂的属性。在 CART 算法中，则使用的是 Gini 指标来度量划分数据。这些算法的应用使得决策树具有效率高，处理的信息量大等优点。但同时，决策树本身依然有着不可避免的局限性，如对缺失值不敏感，数据容易过拟合等。

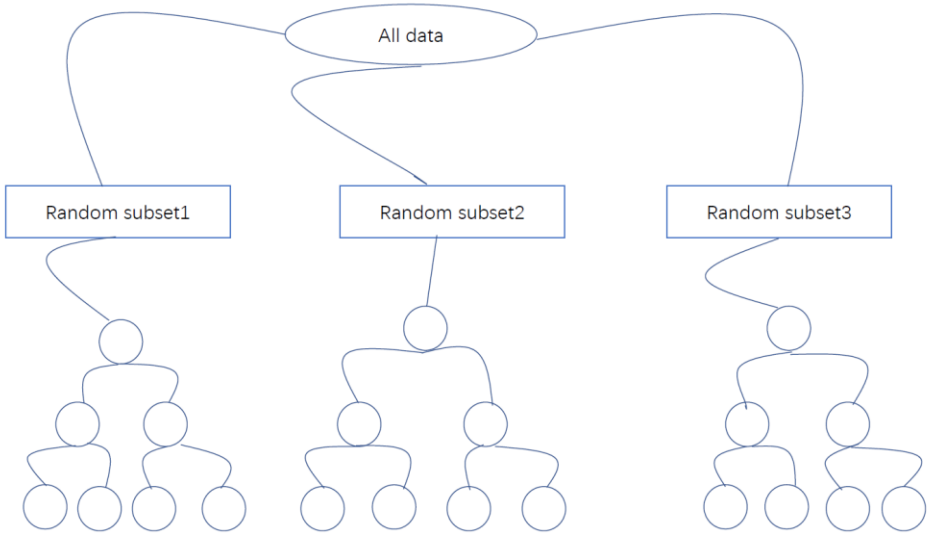
为了解决决策树容易过拟合等缺点，Breiman 在 2001 年组合 Bagging 集成学习和决策树的研究思想，提出一种效果更好的算法——随机森林算法。随机森林算法主要用于分类和回归问题，与此同时，随机森林算法还可以用于降维，处理具有异常值，缺失值以及噪音的数据。

随机森林算法的基本组成元素是决策树。采用对多颗决策树 (可能是二叉树，也可能是多叉) 的随机排列组合来提高分类的准确性。随机森林的主要原理如下图所示。首先，我们对所有

数据做 bootstrap 处理，然后采用 bagging 抽样。

训练集相比原始数据而言，只有 63%的数据被重复抽取，而有 37%的数据从未出现。使用这样的方法可以替代数据集交叉验证法，同时也避免了过高的时间空间复杂度。Bagging 抽样是有放回的抽样，即每棵树的数据集是由原始数据集随机构成，可能重复包含某些数据，也可能不包含某些数据。接着，随机森林会随机选择特征子集。在树的节点分裂时，会随机无放回的选择总属性的子集，这个子集的大小会远远小于总属性特征的数量。每次分裂时，都会根据之前在决策树中提到的指标如信息熵，信息增益零率以及基尼指数来选择分裂的节点。随机森林的结束条件有以下几种方式：决策树达到最大深度、终节点不纯度达到阈值、终节点的样本数达到设定值、待分裂属性用完等。最后，我们会根据对每棵树的评分来对特征的重要程度进行划分。

图 6：随机森林原理示意图



数据来源：华泰期货研究院

通过上述随机森林原理的介绍，我们可以看到随机森林通过种多颗树的方式降低了过拟合的危险。同时也解决了决策树容易受到极值影响等问题。下面会介绍一些衡量随机森林的性能指标。一般来说，这类指标可以分为三种：泛化误差，分类效果指标，运行效率。泛化能力指的是经过训练过的模型对于没有在训练集中出现的样本做出正确反映的能力。在随机森林算法中，可以使用 OOB 估计去估计泛化误差。如前所述，随机森林是使用 bagging 方法进行训练集的生成的，在产生这些数据集的时候，有部分数据是未被抽取的，这类数据就是 OOB（out of bag）。使用这类数据去验证，既减少了数据的复杂度，又保证了验证样本

的一致性。分类效果指标主要是用来考量分类回归以及预测效果的指标。为了描述这些指标我们首先需要考虑一个如下矩阵。

	Classified positive	Classified negative
Positive	TP	FN
Negative	FP	TN

上表是两分类数据集的混淆矩阵。假设数据集中有两个分类，分别叫正类（positive）和负类（negative），TP 和 TN 分别代表正确分类的正类以及负类的样本数量；FN 和 FP 分别是错误分类的正类和负类的样本数量。我们可以通过定义如精确度，负类检验值等一系列指标来评判分类效果。精确度的定义如下。这个指标是用来衡量随机森林算法的总体分类精度，一般而言总体的分类精度越高，算法的分类效果越好。

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

负类检验值（F-value）的定义如下所示，是从负类的角度出发综合评价随机森林性能的一个综合评价指标。它是由负类的查全率（Recall）和查准率（Precision）的组成的。其主要公式如下：

$$F - value = (recall * precision * (1 + \beta^2)/(recall^{*}\beta^2 + precision)$$

其中，$recall = TP/(TP + FN)$ $precision = TP/(TP + FP)$.

运行效率指标一般使用算法的复杂度来衡量。算法的复杂度体现在空间以及时间上。时间的复杂度体现在算法的循环次数上，一般来说，循环次数越多，计算机耗时越多。空间的复杂度体现在执行算法所需要占用的内存空间。但就现阶段而言，随着硬件的迅速发展，随机森林算法的空间复杂度一般不做太多的考虑，而时间复杂度是随机森林算法中需要着重考虑的问题。

所以，将随机森林模型运用在金融数据上，我们可以几乎以遍历的方式去比较各类因子，从而筛选出比较重要的一类因子。在对因子重要性的描述中，我们利用最浅分裂节点分布（distribute of minimum depth），来考虑因子对标的物的影响力。最浅分裂节点分布是指在树生长过程中，最早出现在分裂节点处的因子的分布。也就是说如果一个因子越接近树根，这个因子就越重要。

与随机森林类似的，XGBoost 算法也是一种基于决策树的可扩展的机器学习算法。XGBoost 是一种极限梯度提升算法，通过对损失函数的二阶泰勒近似来提升训练速度。在 Kaggle 大赛的希格斯子信号识别竞赛中，XGBoost 凭借着出众的运算速度和高预测精准率得到了 AI

行业的广泛关注，目前，XGBoost 已经成为了各类 AI 竞赛的常客。

与随机森林类似，XGBoost 算法是树的集成模型，它的基分类器上文所提到的决策树算法 CART。XGBoost 的目标函数为：

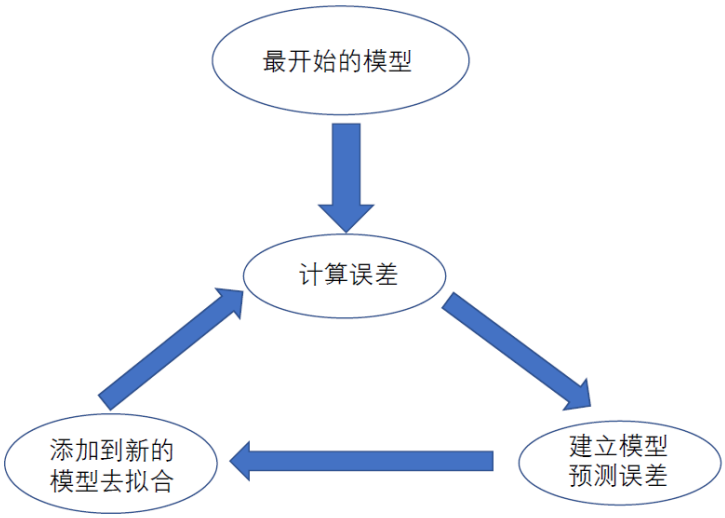
$$\mathrm{Obj} = \mathrm{L} + \ \Omega$$

其中 L 为误差项，常见的误差有平方误差，logistic 误差。Ω是正则项，该项的作用是控制模型的复杂度。误差项鼓励训练复杂的模型，这样样本数据的拟合性会更好。正则项则是鼓励训练更简单的模型，这样可以降低模型中噪声的影响，但这是一个需要权衡的问题。XGBoost 会通过迭代的方式让数据进行学习，从而优化模型。如果迭代 t 次，那么其目标函数为：

$$\mathrm{Obj}^{\mathrm{t}} = \sum_{i=1}^{n} L\left(y_i y_i^{t-1} + f(x_i)\right) + \Omega(\mathrm{f})$$

在此处，用泰勒公式展开后，接着移除常数项，我们可以得到一个比较统一的目标函数。这一目标函数的特点是它只依赖于每个数据点的在误差函数上的一阶导数和二阶导数。接着，XGBoost 对各种枚举出来的结构进行评分，从而选择结构最优的树。当然，枚举出所有的决策树是一项困难耗时的工作，所以此处会使用 CART 算法，在节点分裂时，XGBoost 算法会利用目标函数计算加入分割点后与加入分割点前的目标函数值。进行比较后，再决定是否添加节点。至此 XGBoost 的节点分裂原理已经实现，随着模型的迭代，模型会根据残差不断优化自身。同时，因为节点分裂的目标函数在误差项和正则化项中考虑得比较全面周到，因此该模型拥有很高的精度，同时也拥有很好的对抗过拟合的性能。XGBoost 的运作流程可以如图所示，通过对误差项的不断迭代，来使得模型越来越接近真实值，这样，即使一开始我们的模型与真实模型相差甚远，也可以通过这种反复迭代的方式来让模型变得精确。

图 7：XGBoost 原理示意图



数据来源：华泰期货研究院

值得一提的是 XGBoost 虽然是一个串行的算法，也就是说只有在前面一棵树计算完成以后才可以计算后面一棵树，而其中最耗时的部分就是寻找树的最优分裂节点。在 XGBoost 算法中，通过在遍历特征层面设计并行化的算法来加快计算速率。XGBoost 会对每一个特征进行排序，并保存其结构。在后续的算法计算中，不断的重复使用这一结构，这样就从减少计算量的角度减少了运算时间。

对比上述算法，决策树算法提供一个很好的处理不同特征的大数据的方式，但基于过拟合，对缺失值敏感等局限性（这些局限性在处理金融数据时显得尤为重要），我们更倾向于使用多棵决策树的组合去处理问题，如随机森林模型以及 XGBoost 算法。随机森林模型在对数据进行分类或者回归的精度上发挥着很好的作用，但因为其运算量较大的问题会导致在运算速度上会稍有逊色。而 XGBoost 算法则具有处理模型速度快，且预测准确率较高的优势。

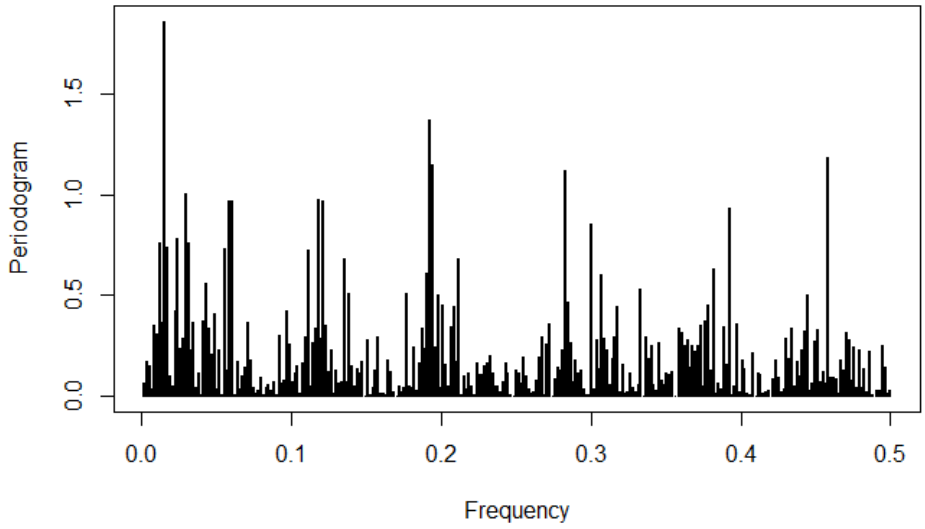
四、主要研究结果

4.1 周期性特征

通过对因子数据进行分析，捕捉宏观因子的类周期性表现，有助于筛选出对贵金属影响力较强的因子。在 2.1 节中介绍了按照周期长度分解的数据处理方法。本节将对分解得到不同周期长度的时间序列转化为频域序列进行分析，观测各因子的类周期特性。

以信贷风险为例，下图为原始数据转化为频域后的图像。很明显，在没有经过数据处理之前，市场数据因为包含了多重周期、若干历史时刻市场结构性变化、噪音干扰等多种因素。所以，即使我们在杂乱的分布中似乎感受到了丰富的信息，也比较难提取并得到确定的量化结论。

图 8：信贷风险指标——原始数据



数据来源：Bloomberg，华泰期货研究院

我们将信贷风险原始数据按照不同的周期（傅里叶变换意义上）范围进行分解，大致分解为三个层次：

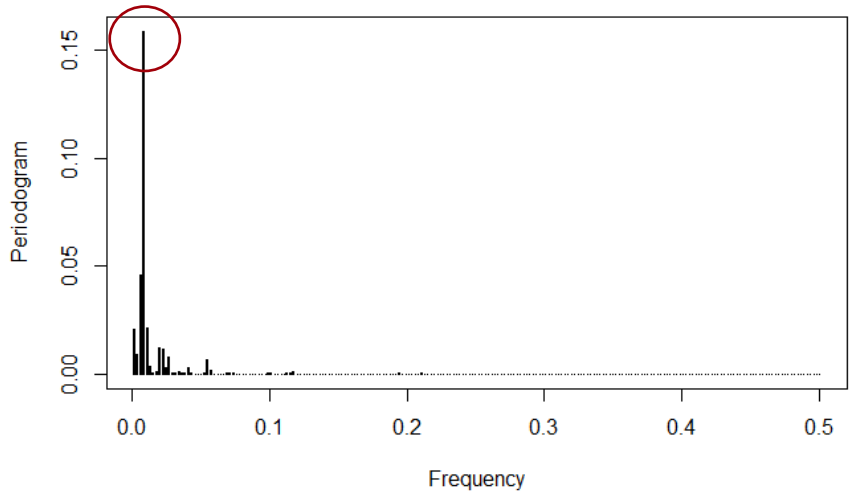
1) 短周期：1 个季度 - 1 年左右；

2) 中周期：1-5 年左右；

3) 长周期：5 年以上。

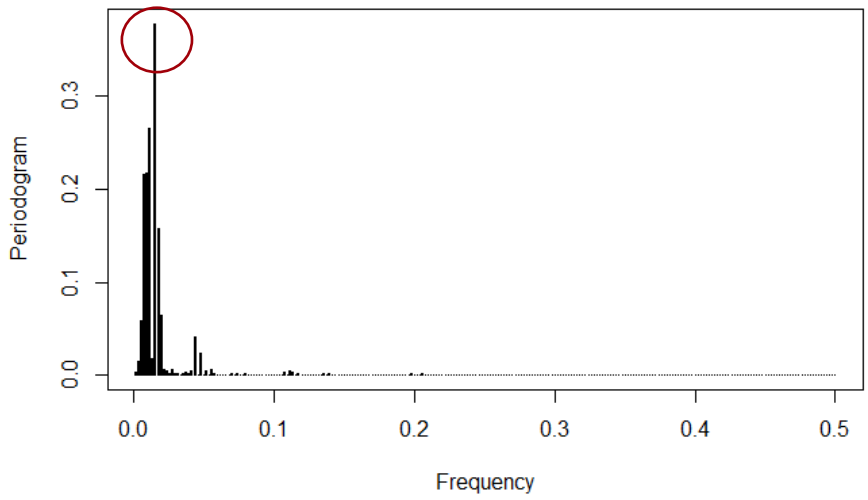
在经过数据分解之后，将三个层次转化为频域后的图像，可以看到长周期与中周期频率数据中，存在单频率的频谱明显高于其他频率的特征，可以说明中长周期的周期特性非常明显，表现出明确的**主导周期**。

图 9：信贷风险指标——长周期



数据来源：Bloomberg，华泰期货研究院

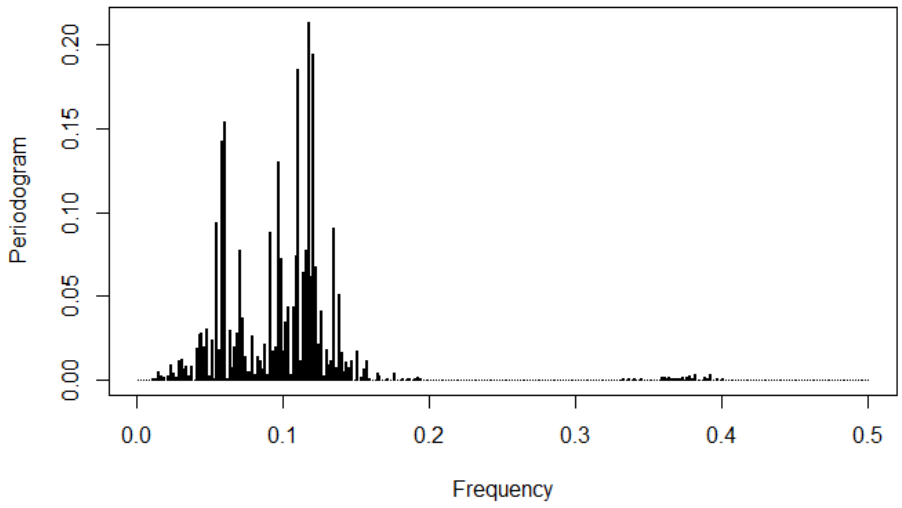
图 10： 信贷风险指标——中周期



数据来源：Bloomberg，华泰期货研究院

短周期频率数据中，难以找到单频率的频谱明显高于其他频率的特征，因此短周期尺度上并不存在主导周期。

图 11： 信贷风险指标——短周期



数据来源：Bloomberg，华泰期货研究院

我们将宏观因子做系统性梳理得到下表。需要特别指出的是，虽然在三个尺度上都能计算出周期，但并不代表因子只具有三种周期特征，实际还需要根据因子的主要特性选取适合的分解尺度寻找相应的周期：

表格 3：主要参考宏观数据周期

	大宗商品	信贷风险	美元 M1	美元 M2	S&P 500	美国 CPI	无风险利率	美元指数
短周期	0.9	0.7	1.0	1.0	0.9	0.7	0.7	0.9
中周期	3.4	5.3	6.0	4.4	3.4	4.5	5.3	4.0
长周期	8.0	9.0	8.0	12.0	6.9	16.7	5.3	8.0

单位：年

资料来源：Bloomberg，华泰期货研究院

4.2 市场结构性相变特征

我们使用 RQA 方法来研究短期波动层次上的复现（Recurrence）统计特征。并且通过递归图来对数据做可视化研究。如下图所示，我们观测到金价在上表现出非常**丰富**而**复杂**的特征。

1) 金价变动的短周期均值为 **6.3 个月。**

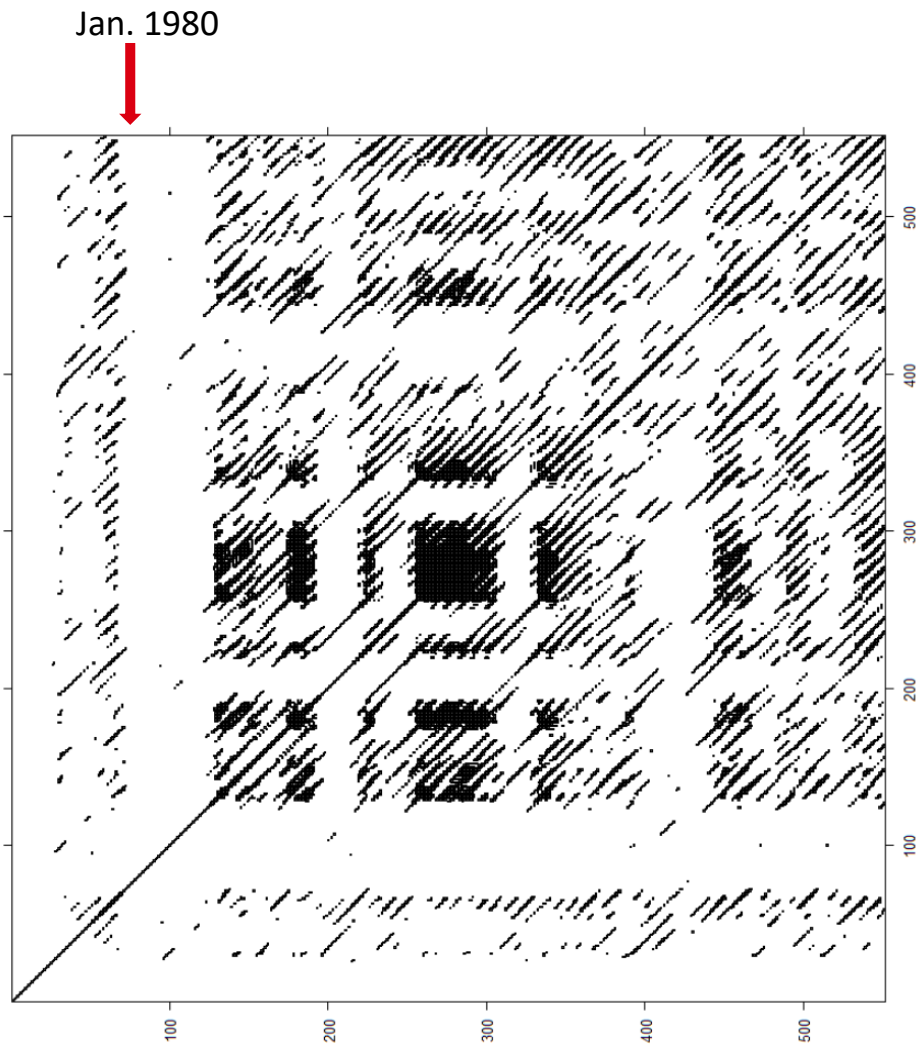
2) 金价不仅有平行与对角线的复现，也有纵向延伸，表现出一定时段内价格波动的周期性。

3) 即使时间跨度很大也会出现复现（并非只有紧邻对角线的平行线），说明价格波动确实包含了中长周期特征。

4) 但是，沿对角线的重复一般不长，也说明了，价格波动特征一般延续性不强。

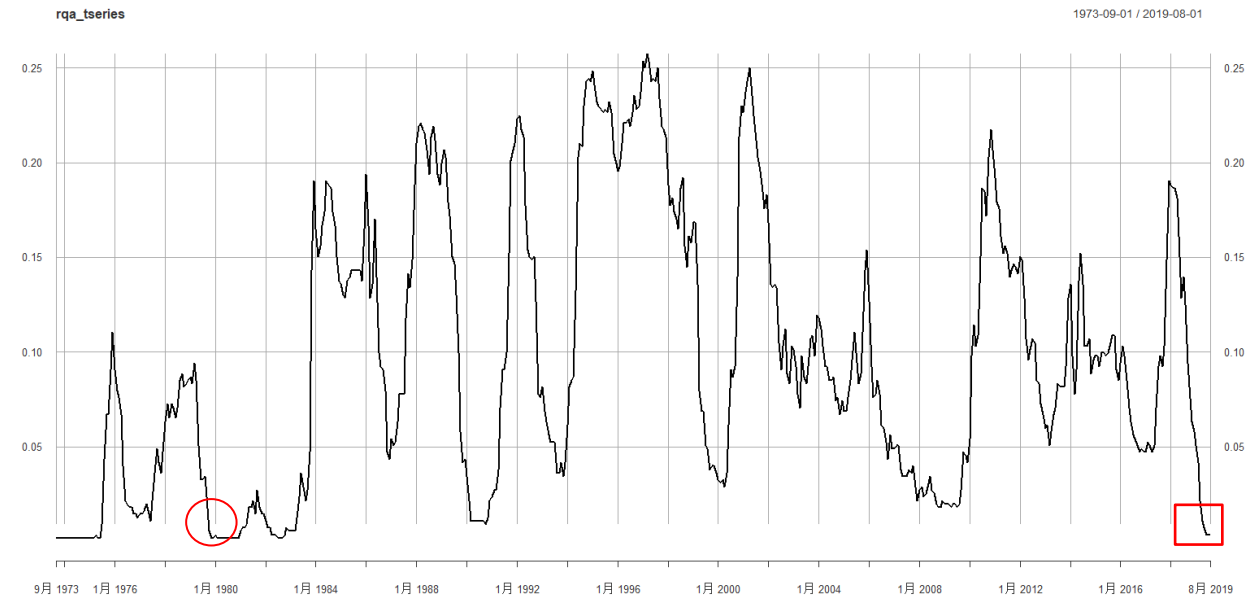
5) 金价在一些重要时间点出现了断层（空白部分），表现出相变特征（参考方法介绍 2.2）。这和重要历史事件相吻合。比如 1980 年 1 月，是金价与美元脱钩以后达到的第一个极值点，随后的 20 多年间都无法再触及，直到 2008 年次贷海啸才明确突破。

图 12： 黄金递归图



数据来源：Bloomberg，华泰期货研究院

图 13： 黄金递归图复现比率分析（RR 按时序展开）



数据来源：Bloomberg，华泰期货研究院

图 14： 黄金历史价格

数据来源：Bloomberg，华泰期货研究院

上图清晰显示出金价复现的类周期性特征，并且很好得捕捉到了的市场相变特征。另外，我们警觉地注意到（红色方框），金价似乎在走向另一个历史性相变点。在下文中会结合其他宏观因子的相关性表现作进一步论述。

按照完全一样的方法，我们使用较长历史数据（自 1971 年），来分析宏观数据的短周期复

现规律，计算出主要参考宏观数据短周期均值：

表格 4：主要参考宏观数据短周期均值

	大宗商品	信贷风险	美元 M1	美元 M2	S&P 500	美国 CPI	无风险利率	美元指数
短周期均值	6.2	5.9	6.6	5.1	5.5	6.7	6.5	5.0

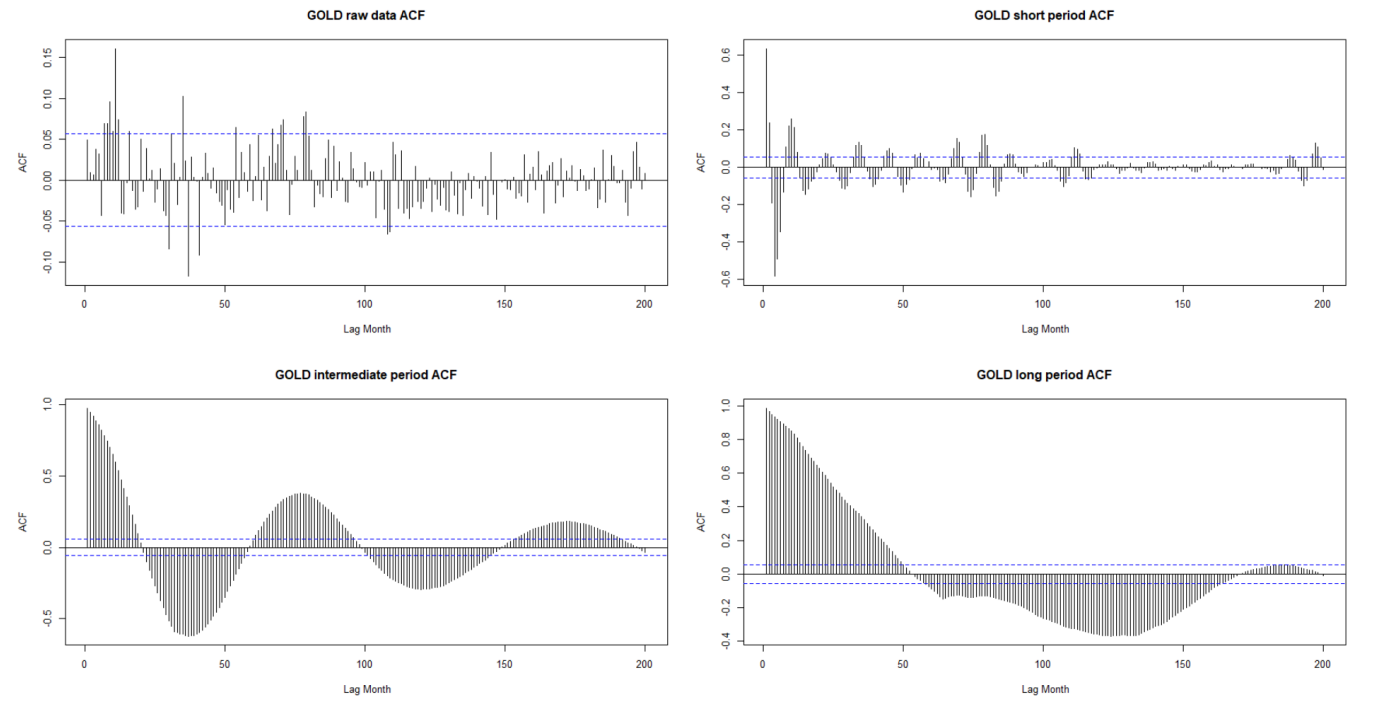
单位：月

资料来源：Bloomberg 华泰期货研究院

4.3 自相关性特征

多周期尺度数据处理方法为我们理解金融时间序列中的自相关性，标的物与相关因子之间的相关性提供了较好的分析工具。以本文的主要预测标的物黄金为例：

图 15： 金价月度收益率自相关性



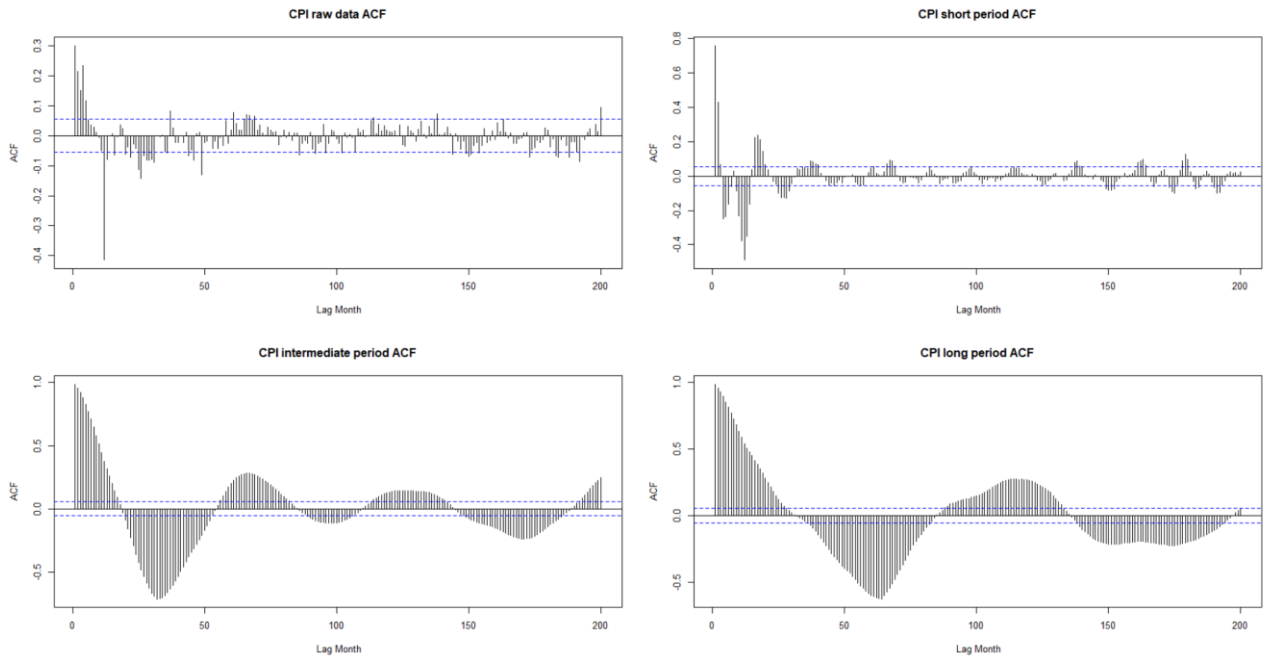
数据来源：Bloomberg，华泰期货研究院

在原始数据中（左上角图），金价月度收益率的自相关性并不明显，而超过95%置信区间的异常值也并不多见，这符合我们一般常见的交易型数据特征。实际上，如果只有这样的数据结构，我们并不能做出任何有效的判断，也比较难决定选用什么样的模型或者分析工具做更深入的数据特征分析。

通过数据分解可以清楚地看到，数据结构呈现出来越来越明显的规律性。这在一定程度上符合我们分解的原则，处理后的数据展现出来不同的周期性特征。同时在3个周期尺度上，数据的自相关特征在统计意义上非常显著，远超过无自相关性的95%置信区间，且衰减速度很慢。注意，随时间延迟没有看到ACF趋于饱和，我们不认为这里是long-term memory特征。下文中将使用ARMA模型做定量拟合。但同时，我们也持一定谨慎的态度，认为上图中显示的自相关性*不一定*是线性自相关性。事实上，我们认为数据具有相当有趣的内在结构。

上图的结果具有很强的一般性，实际上，至今没有看到其他任何数据表现出和上图严重不符的反例。从侧面验证了该方法的普适性。但限于篇幅，我们不再一一展示。不过，非常值得一提的是，不仅仅是交易型数据，对于机构统计型数据，也存在不同波动周期尺度上的自相关特征。我们这里举一个例子—美国劳工部统计局记录的CPI同比变化率（始于1913-01-01）。

图16：　CPI月度波动自相关性

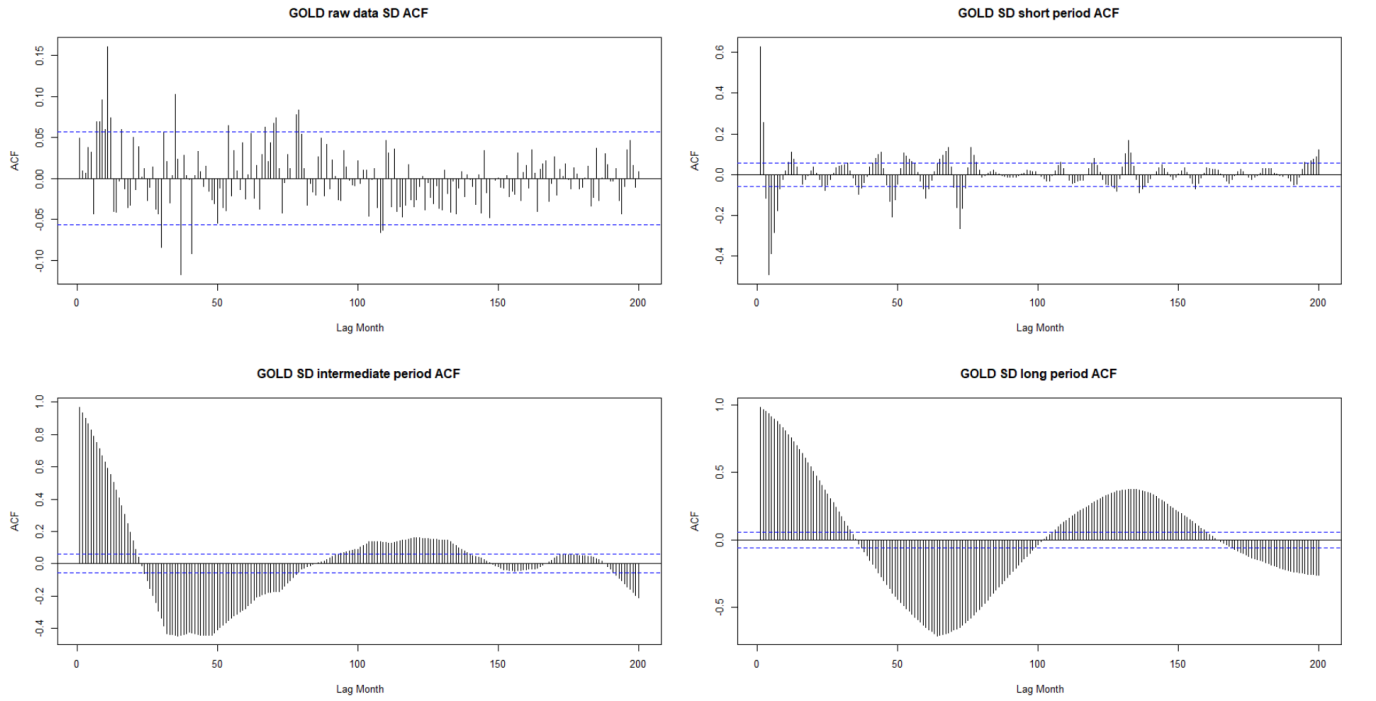


数据来源：BLS，华泰期货研究院

CPI 原始数据似乎已经表现出一定的结构特征。在数据分解处理之后，展现了和交易型数据非常类似的显著特征。更加有趣的是，测试发现这样的多周期尺度数据分解后的自相关结构并非仅限于收益率（或者一阶差分）数据。实际上，更高阶的统计量也具有类似特征。

接下来，我们看一下黄金历史收益率的标准差：

图 17： 金价月度 Volatility 自相关性



数据来源：Bloomberg，华泰期货研究院

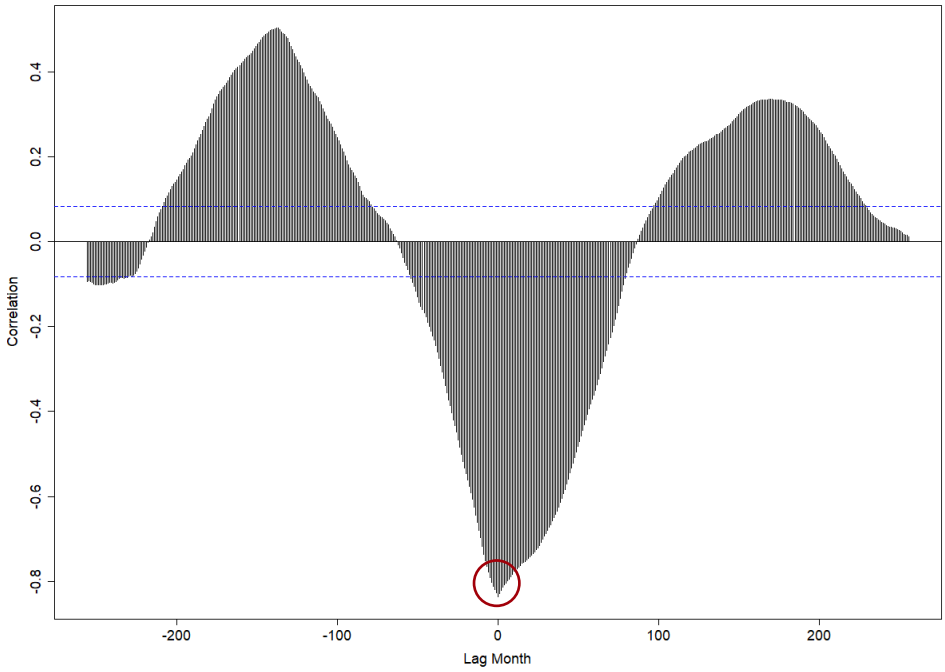
从而，我们相信本文所面对的金融数据的确蕴含了非常丰富的信息，于是问题的关键变成了使用什么样的方法去提取解决问题所需要的最关键信息，比如本文考虑的自相关性以及自相关性表现出来的周期波动特征。而对于其他统计量的多周期尺度分解，有助于我们对数据进行更全面的理解。从策略研发的角度来看，也为提高收益，管控风险提供了数据基础。

4.4 协方差相关性及因子影响力

本节将从线性和非线性两个角度来考虑因子对于标的物的影响力。线性方法来自线性回归分析；非线性方法则主要采用随机森林方法。因为，内容较多，本节将筛选若干长周期的典

型例子来做对比，并且简要叙述如何利用相关性考虑投资策略的设计。从大类资产配置的角度来看，我们分析美国 S&P 500 指数收益率与黄金的相关性。

图 18： 金价与 S&P 500 指数线性相关性



数据来源：Bloomberg，华泰期货研究院

表格 5：主要因子对黄金价格的解释力（基于线性相关性）

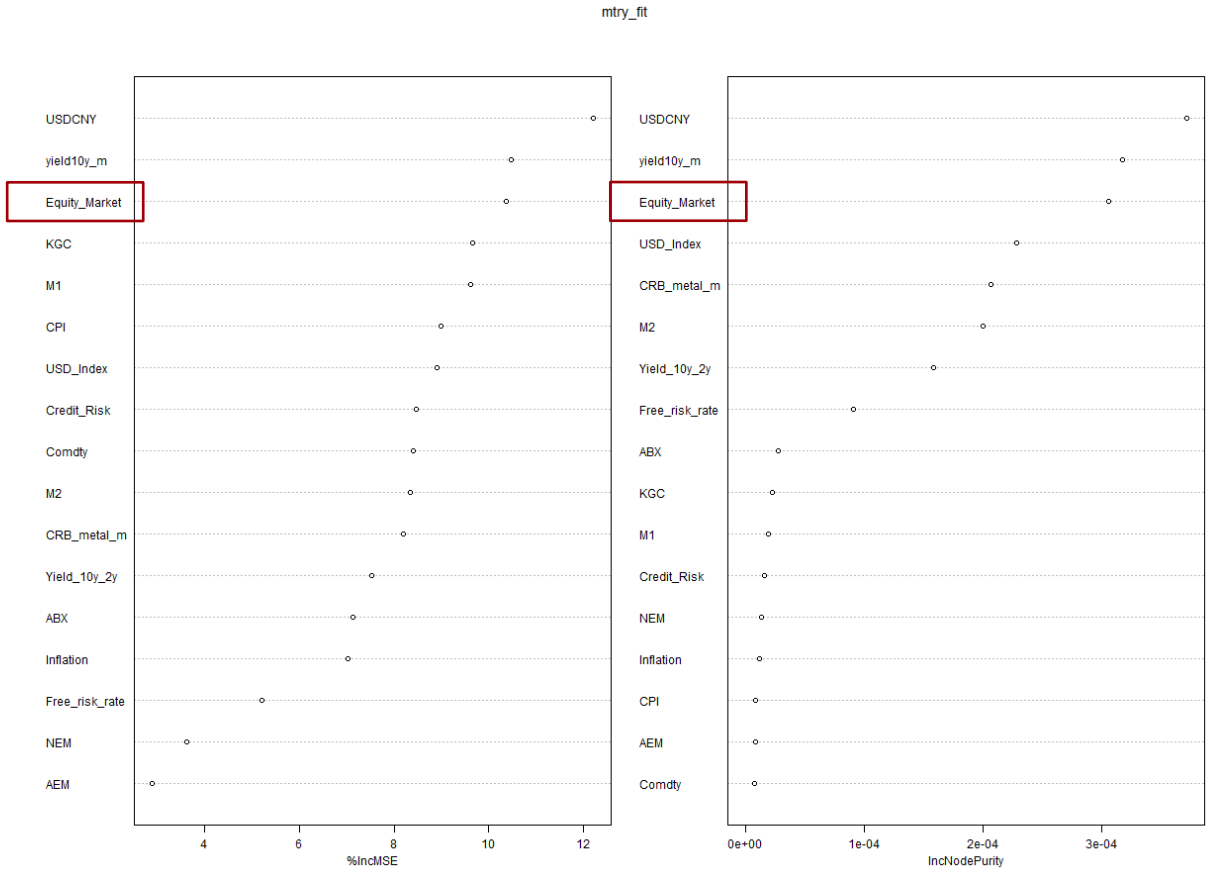
	R-squared	Adjusted R-squared	T-Value
商品指数	0.1492	0.1475	9.3092
信贷风险	0.4640	0.4629	20.6784
S&P 500 指数	0.2506	0.2489	-11.9924
美国 CPI	0.5144	0.5134	22.8742
美国实际利率	0.3625	0.3613	-17.2960
美元指数	0.1831	0.1815	-10.8571

资料来源：Bloomberg，华泰期货研究院

尽管，我们预期黄金和权益类市场之间具有负相关性，但是长周期的相关性绝对值如此之高令人略感意外。同时，我们也看到因为相关性衰减速度很慢（实际上，是延续若干年），那么这样的负相关性并不适合做交易策略的（择时）战术设计，更适合在大类配置的过程中长期持有，利用这两类资产的对冲特征来分散资产配置风险。

随机森林模型也印证了这个长周期的相关性。首先，我们给出因子重要性判断图。该图是根据树生长过程中，不同变量作为节点分裂参数，带来的 loss 函数的优化程度，来确定其重要性。由下图可知 S&P500 指数依然居于影响力前三的位置。

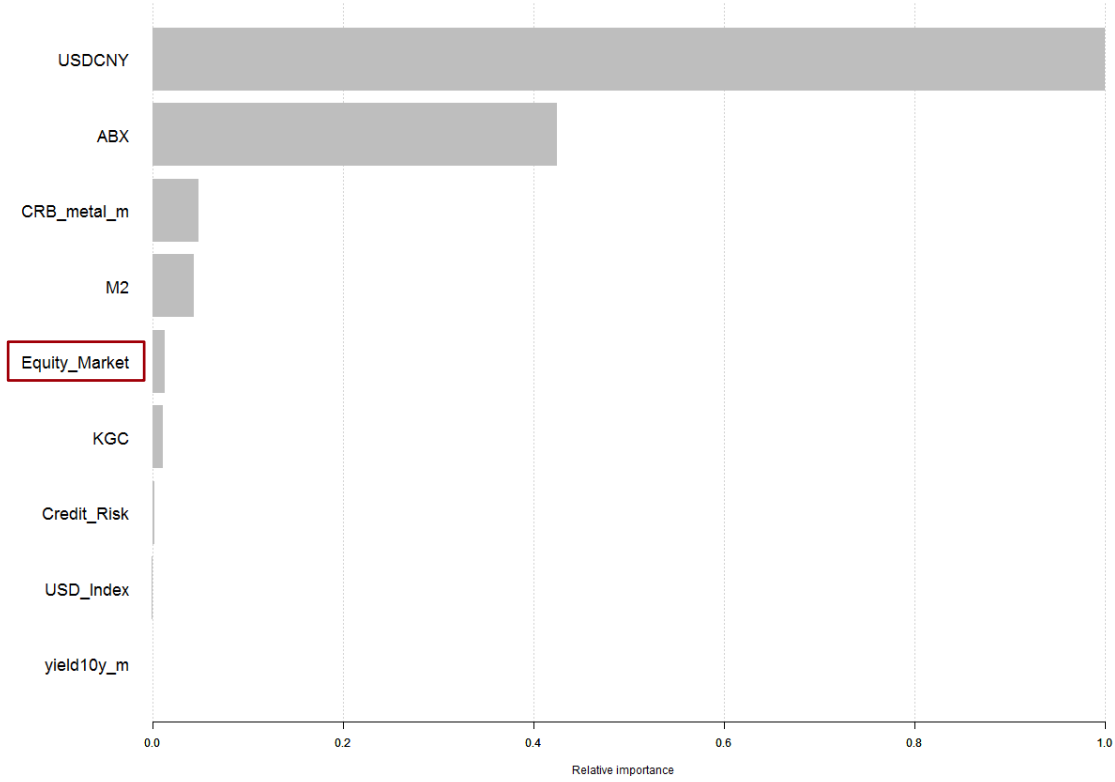
图 19： 因子对金价影响力（基于随机森林模型）



数据来源：Bloomberg，Wind，华泰期货研究院

这个结果与 XGBoost 给出的结果有一定差异，尽管 S&P500 指数的排名依然位于前五。

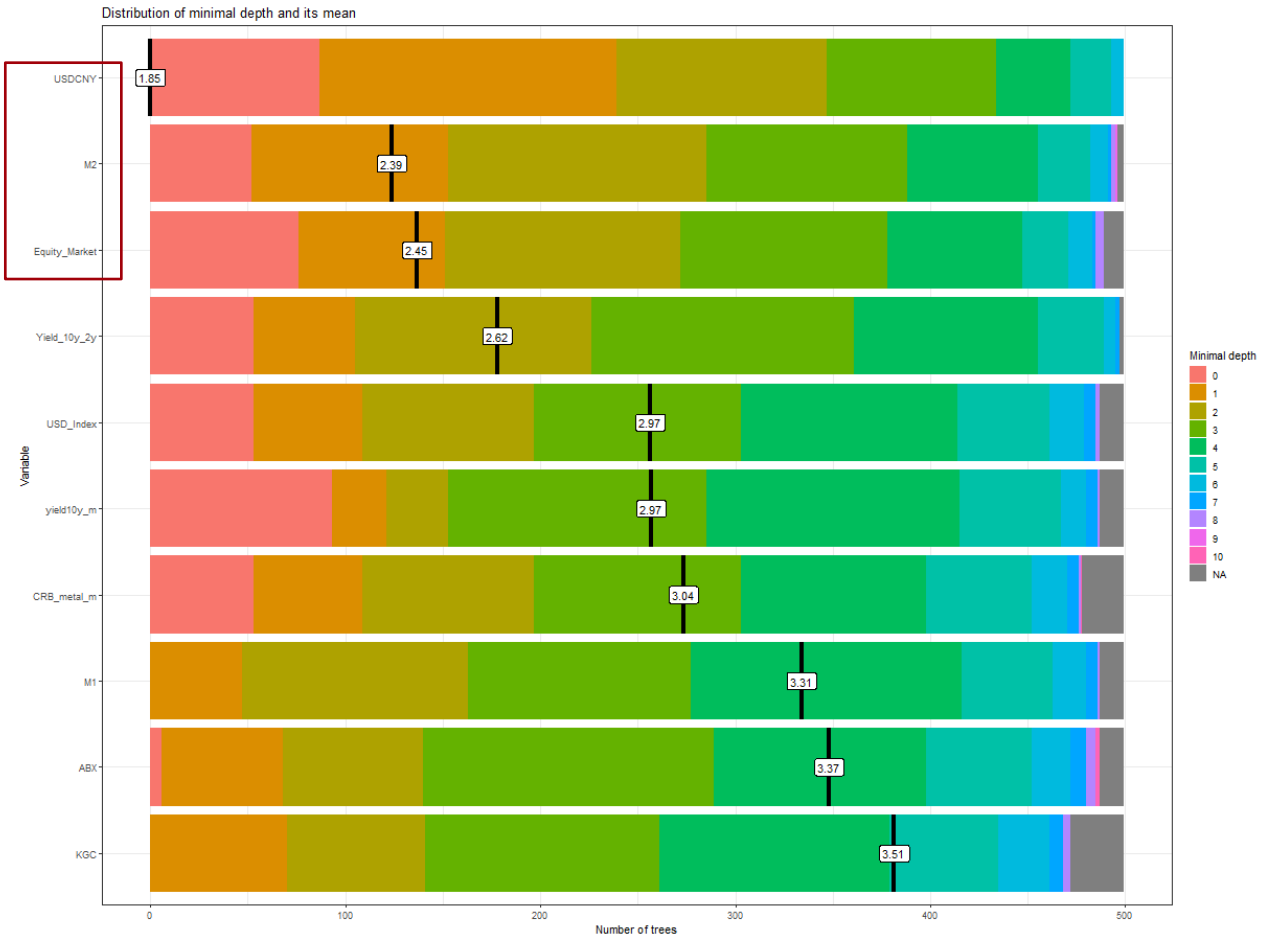
图 20： 因子对金价影响力（基于 XGBoost）



数据来源：Bloomberg，Wind，华泰期货研究院

然后，利用最浅分裂节点分布（distribute of minimum depth）和分布均值测量因子对标的物的影响力。需要指出，这个方法与上述因子重要性的计算方法并不相同，最浅分裂节点并不直接依赖于误差率的计算，而是依赖于树形的**拓扑结构**以及在森林中的**分布**[4]。长周期来看（下图），股票市场对金价的总体影响仅次于中美汇率（近 20 年）和美元供应量 M2。

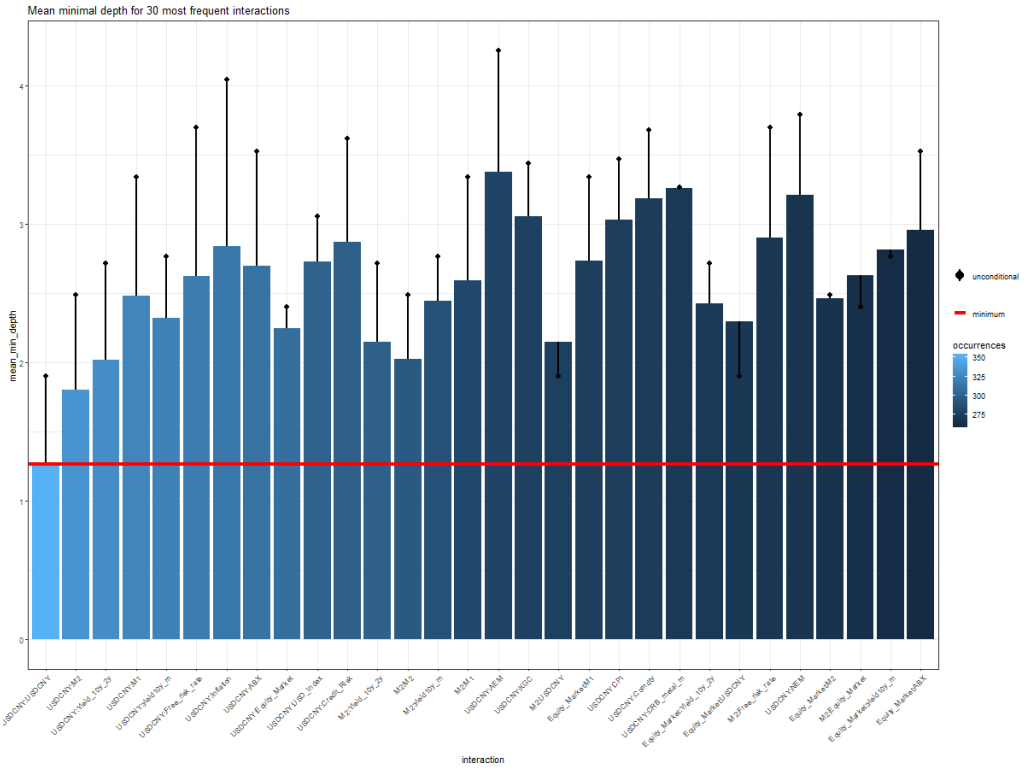
图 21： S&P 500 指数对金价影响力（随机森林）



数据来源：Bloomberg，Wind，华泰期货研究院

在随机森林模型中，我们并没有假设任何线性相关性；换句话说，可以提取因子与标的物之间的非线性相关性（如果存在的话）。下图就提供了一种直观的方法去寻找相互作用因子--一对因子组合的最浅分裂节点均值，我们看到中美汇率在因子对中也是出现频率最高的单因子。同时，S&P 500 指数与其他因子作用的频率也很高（出现 9 次，仅次于中美汇率和美元供应量 M2）。

图 22： 频率最高因子对（平均最浅分裂节点）



数据来源：Bloomberg，Wind，华泰期货研究院

使用相同的方法，再分别考虑中周期和短周期条件下各种因子的影响力。在树形结构下，有如下结论：

- **长周期**因子影响力前 5 名为：**中美汇率，美元货币供应量 M2，S&P500 指数，国债收益率利差，美元指数**。
- **中周期**影响力最强前 5 名为：**实际无风险利率，金矿公司股价，美元指数，10 年期国债收益率，美元货币供应量 M1**
- **短周期**影响力最强前 5 名为：**金矿公司股价，CRB 金属指数，美元指数，实际无风险利率，信贷风险**

上述结果也是对我们过往很多行业研究结论的重要量化佐证。比如，短线来看，对金价互动比较密切的包含了美元指数的波动（黄金的全球定价主导依然是美元），金属商品指数，金矿公司个股的股价变动，以及市场对无风险利率的判断，信贷风险的与预估等等。充分体现了，黄金的货币属性，商品属性和金融属性。并且，我们也注意到这些不同属性其实是在不同的周期尺度上对金价产生不同程度的影响。

同时，我们也发现了新的内容，比如黄金与**股市**整体存在较高的长期负相关性。我们猜测这

可能是在西方金融市场对全球金融市场主导的几十年来，其金融资产多元化配置，分散化投资，风险管理走向成熟的重要表现。这对我们在大类资产配置的理解和方法创新等方面提供了重要的启示。

令人颇感意外的是**中美汇率**（近 20 年数据）对金价长期走势的影响如此显著。但是，这个结果本身符合近 20 年来，中国作为全球最重要的经济发展引擎，而美国作为最重要的技术创新国和产品终端消费国，两者互动带来的全球经济发展的新格局。我们觉得，这几十年来的中美合作确实为全球经济的发展提供了良好的条件，带来了全球经济发展较为平稳的 20 年。可以说在这一过程中，中美间互惠互利已成事实，全球金融避险并非长周期主导因素。但当我们结合 4.2 节中黄金递归图复现比率分析，也不无担心，近年来中美贸易摩擦的升级，以至于现在暂无贸易战彻底解决方案的时候，金价作为一个重要的避险资产和抗通胀工具，也许会迎来历史性的相变阶段。中美经济互补的发展模式甚至有可能迎来一个重要转变阶段，需要引起我们的高度重视。

我们注意到，CPI 指数在三个周期尺度中对金价的影响均不明显，可参看报告《宏观策略看资产（一）——黄金（修订版）》中的详细解释。

五、金价走势预判初探

5.1 预测方法介绍

上文中对数据的分解，线性和非线性模型的测试已经帮助我们更深入理解了金融数据的特征。这同时也为我们寻找合适的金价预测模型提供了线索。

从线性时序模型的角度出发，因为，数据分解后在短周期，中周期和长周期尺度上分别表现了其独特的自相关特征，那么，我们就可以利用 ARMAX 模型来建立基于回归分析的多因子模型，并且通过蒙特卡洛的方法获得预测性结果。这里的模型不同于过往我们所使用的单一时间序列模型，也就是一个标的物对应一组 arma 阶数和一套因子组合。我们将首先尝试在每一个周期尺度上，利用该周期尺度上挑选出来的影响力最强因子，建立 ARMAX 模型。然后，分别做模型预测，最后，再将各尺度的结果叠加得到最后的预测结果。

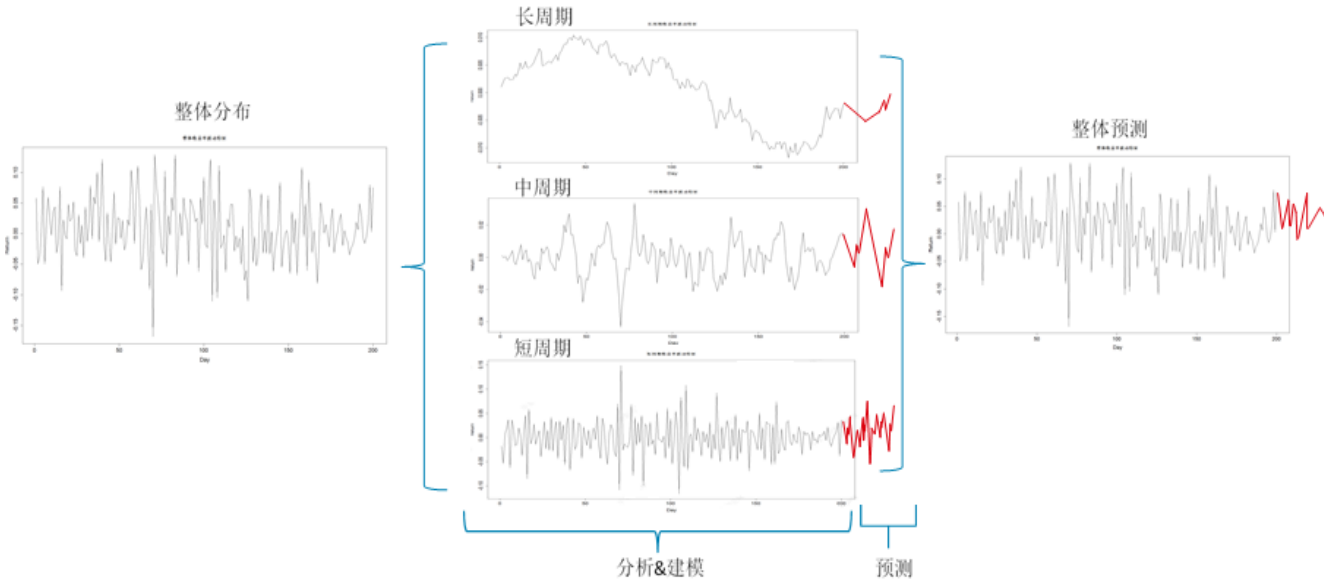
AI 模型测试具有一定的复杂度，为此，我们考虑从最容易获得确定性结论的角度出发设计测试模型。首先，我们的历史回测采用分类的方法，在做样本外预测时只判断涨跌两种情况。如果判断为涨就持有黄金，如果判断为跌就空仓。这样，最直观的考察指标就是：持有期间是否获得一定的超额收益；在市场发生回撤时，能否更好地管控风险。

从我们目前模型的测试来看，随机森林模型给出了较好的结果。我们发现，在我们选取的因子（指标）组合里面，因为若干因子互相之间表现了非常强的相关性，在做进一步特征

工程处理之前，Boosting 类型的模型比较难发挥应有的效率。

请注意，本文的主要目的是展示我们目前解读数据，研究大数据的方法论，而能够处理复杂类型的数据，充分提取各个周期尺度上的信息是模型首要考虑的功能。所以，策略的描述将以实现前文数据分析及研究结果转换为主。目前，实战投资策略的开发亦是按照相似思路展开。

图 23： 预测模型原理



数据来源：华泰期货研究院

模型开发的极简化流程为：

数据分解>>数据分析>>建模>>预测>>叠加

我们的策略开发思路同时加入了如下的考虑：

1) 融合多重维度的数据信息，分尺度判断其后市波动特征；

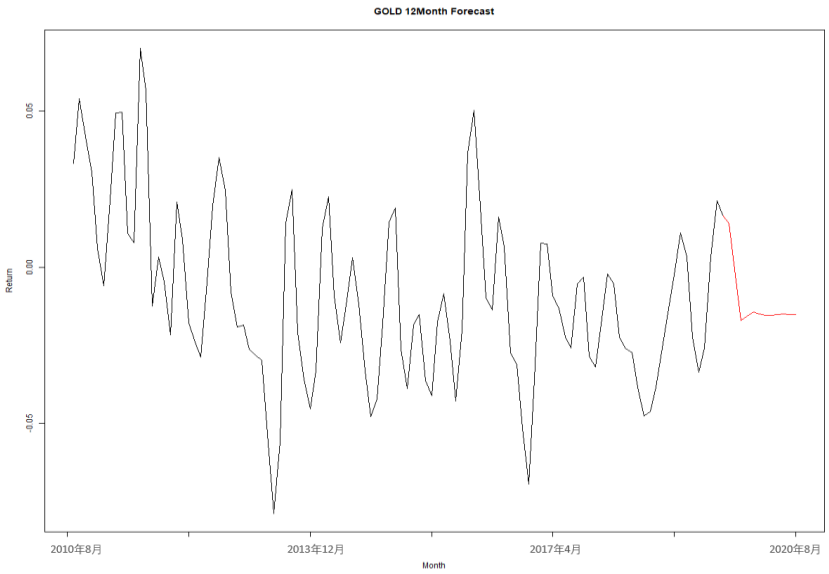
2) 因子组是否具有一定的预测能力（如领先性）；

3) 相关性衰减速度等。

5.2 ARMAX 模型预判结果

直接给出日收益率的预测结果：

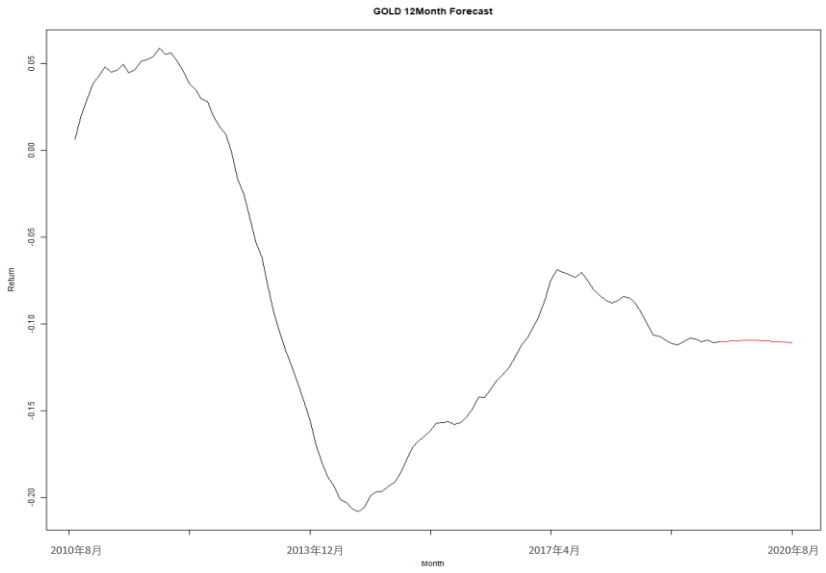
图 24： ARMAX 模型短周期预判



数据来源：华泰期货研究院

短周期尺度来看，将出现短暂反转，然后趋稳。美元指数和美国实际利率的变动牵扯了整个市场的神经，全球投资人都在密切关注，近期全球经济加速下行，而美国基本面独好的情景能维持多长时间。可能出现的场景是，美元一段时间内依然保持优势，黄金价格承压，但会逐渐走稳。而商品在短周期波动中反向推动金价，有助于后半段金价变动趋稳。

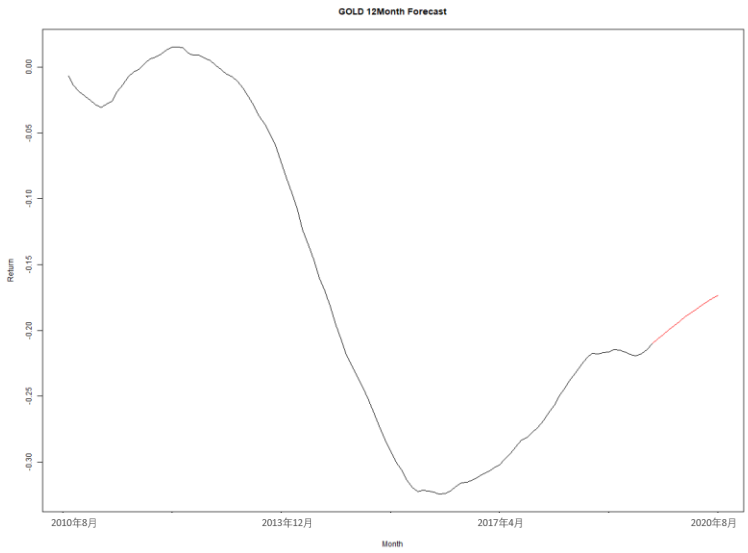
图 25： ARMAX 模型中周期预判



数据来源：华泰期货研究院

中周期尺度来看，合力效果是几乎中性，但其实影响因素较为复杂。其中美国实际利率，通货膨胀和信贷风险的影响最为显著，互动关系有可能进一步上升。商品价格的变动和可能的货币宽松在做正向推动，美元指数依然是主要的压制因素。

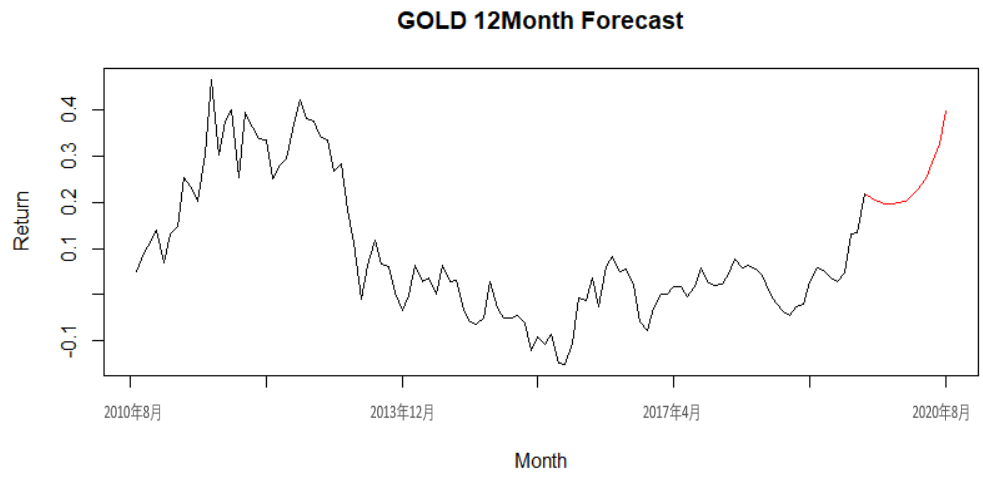
图 26： ARMAX 模型长周期预判



数据来源：华泰期货研究院

长周期尺度来看，金价推升效果比较明确。主要得益于美国实际利率的进一步下滑和货币政策宽松化，美元长期趋势将寻求走贬，通货膨胀的修复，信贷风险的进一步提升，都将成为主要的长线推动因素。同时注意到，股票市场的方向的最终转变极可能成为黄金走高的最后推手。

图 27： ARMAX 模型黄金综合预判



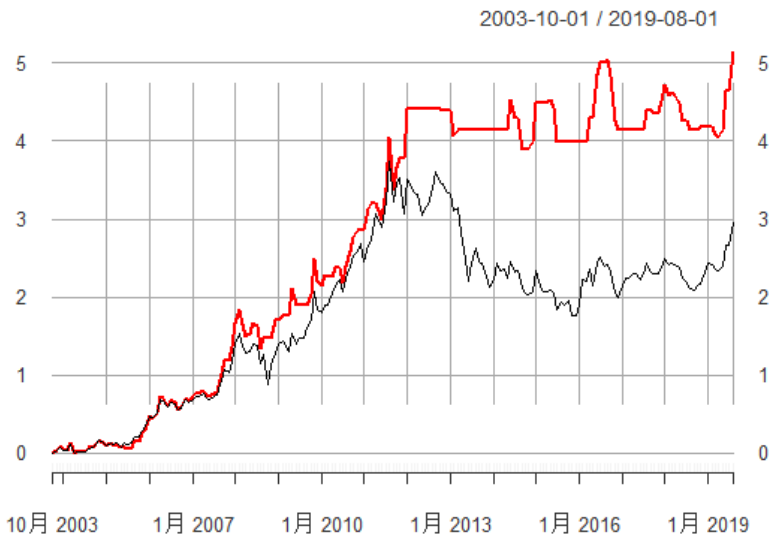
数据来源：华泰期货研究院

综合来看，我们认为未来大约一年内，黄金将在经历一段时间的修整之后，体现充足的后发动能。

5.3 AI 模型回测结果

我们使用大约十年的数据点建模，然后月度滚动对黄金收益率做预判，在此基础上简单选择持有（预测涨）或者空仓黄金（预测跌）。我们使用了 1.2 节提到的全部因子，在叠加了不同周期尺度上的预测结果后进行回测。

图 28： 随机森林模型累计收益率回测结果



数据来源：华泰期货研究院

表格 6：随机森林模型回测结果主要指标

指标	黄金	黄金策略
Alpha	0.000	0.004
Beta	1.000	0.741
Avg. Return	10.132	12.745
Volatility	17.263	16.264
Sharpe	0.169	0.226
Sortino	0.188	0.222
Skewness	-0.059	0.531
CVaR	0.099	0.090
Max DrawDown (%)	41.864	17.518
Max DrawDown Duration	2011-08:2015-12	2008-02:2008-08
预测胜率 (%)	-	57.92

数据来源：华泰期货研究院

回测中我们看到几个重要的策略表现：

1) 策略明确抓到标的物的**长周期主升段**，并且获得一定的超额收益。

2) 策略在若干黄金市场极端条件下表现出来较优异的风险控制能力。

3) 策略明确捕捉到了黄金在 2012 之后的**中长周期熊市**特征，与 2011 年及之前只有零星时段空仓情况完全不同。

4) 今年以来的黄金牛市，策略累积收益率已经突破新高。

结合，我们之前对重要宏观因子影响力的研究，不难发现，黄金市场作为重要的金融资产，具有长期持有配仓，分散风险的重要意义。特别是，我们前文提到的黄金和美股在中长周期上的显著负相关性（请参看 4.3 节），现在演变成了，黄金策略在 2009 年之后，多次出现的长时间空仓配置。所以，我们策略的设计正是抓住了中长周期相关性研究的主要结论，并且使之变成了策略的主要获利来源和管控风险的主要依据。

注意，这样的结论非常关键。因为 AI 模型在提供更高效输出和更小预测误差时，通常会损失更多的特征解释力（相比较线性模型），但是我们的策略设计则有针对性得克服了这一难题。

六、总结

投研之路如同逆水行舟，不进则退。面对大数据时代的挑战，我们不仅需要重新审视已有投研方法的合理性和有效性，更要采用跨学科的方法，利用各门类量化科学的研究成果，重新解构基本面信息，并结合市场行情，用定量的方法挖掘经济运行的重要信息，最终形成投资研判。

本文虽然只是从黄金这个单一的投资资产类型出发，简要展示了我们近期的研究结果。但是，其中定量研究方法，合理投研框架的搭建才是我们关注的焦点。归纳起来，我们有如下几个核心投研观点：

1) 数据的深度解构、挖掘、分析是后续模型提取有效信息的关键。

这一结论对基本面数据，行情数据和另类数据都适用，而跨学科研发是实践中的重点。实际上，我们目前认为，多个周期尺度，不同时间阶段，各种基本面信息会对不同资产定价产生差异显著的影响。而金融资产价格则是这些影响因素的全部叠加，并通过不同层次交易者的市场参与，得到最终具象化体现。

2) 大数据时代的投研框架模型必须具备数据综合处理能力强，运算效率高等重要特征。

尽管，传统金工模型依然是我们工具库的重要组成部分。但是，我们认为 AI 类型的模型，在大数据处理方面具有巨大优势，也是目前我们重点研发方向。同时，我们也要看

到很多 AI 模型也有解释性模糊，计算结果不稳定，特征工程更复杂等挑战。如何佐证基本面研究结论，并最终将数据分析结论转化为有效的投资策略和风控管控方法则是研究的主要难点。而本文正是基于此类问题，展示我们的一些突破性进展。

3) 投研框架需要具备可扩展性。

虽然，本文的案例分析是围绕黄金价格展开，但是，我们落脚点则是一般性方法的讨论。面对其他类型的资产，我们同样需要分析其价格变化现象背后的基本面因素，经济发展的周期性影响，或其他相关类型资产的行情走势影响等等。所以本文的一些重要结论其实是更广阔投研课题的起始点。目前，我们也在沿着这个指导性方向，将各种数据分析方法和 AI 模型推广到其他类型的金融资产研究中去，在进一步丰富研究工具库的同时，用更具一致性的方法论来探索更有效的投资逻辑。

七、参考文献

[1] Eckmann, J. P , S. O. Kamphorst , and D. Ruelle . "*Recurrence Plots of Dynamical Systems.*" Europhysics Letters (EPL) 4.9:973-977 (1987)

[2] Michael A. Rley,Guy C. Van Orden," *Tutorials in Contemporary Nonlinear Mehtods for the Behavioral Sciences",* Arlington, VA: National Science Foundation, Internet resource (2005)

[3] Teodoro Semeraro, Nobert Marwan, Bruce K.Jones, " *Recurrence Analysis of Vegetation Time Series and Phase Transitions in Mediterranean Rangelands*", arXiv:1705.04813 (2017)

[4] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn and S. Lauer, "*High-Dimensional Variable Selection for Survival Data*", Journal of the American Statistical Association, 105:489, 205-217 (2012)

免责声明

公司总部

地址：广东省广州市越秀区东风东路761号丽丰大厦20层

电话：400-6280-888

网址：www.htfc.com