DOI：10.13718/j.cnki.xdzk.2022.01.014

基于机器学习和经验模态分解的跨期套利研究

周亮[1]，　陈辰[2]，　李宁[1]

1. 湖南财政经济学院 财政金融学院，长沙 410205；2. 西南财经大学 金融学院，成都 611130

摘要：采用滚动经验模态分解(EMD)方法对沪深 300 股指期货当月和下月合约的价差波动进行分解，分别利用 Elman 网络、随机森林(RF)、支持向量回归(SVM)3 种机器学习模型及自回归移动平均模型(ARIMA)对不同频率信号进行分析，合成最终的预测结果，并根据预测结果设计跨期套利策略. 研究结果表明：SVM，RF 和 ARIMA 模型的预测精确度相对 Elman 网络较高，所有模型均能取得较高的套利收益，将非线性模型和线性模型融合使用能够改善模型的风险控制能力；将机器学习预测与 EMD 分解技术相融合可以在不提高风险的同时大幅度提高模型的收益率，从而使得模型的夏普比率和索提诺比率均有较大幅度上涨；分样本检验、全 IMF 信号预测以及基于商品期货市场的套利分析，均证明融合 EMD 的机器学习模型可以获得比纯机器学习模型更优异的套利效果. 研究结论有助于促进人工智能与金融学的交叉融合研究，同时也为期货投资提供了理论和现实参考.

关　键　词：机器学习；经验模态分解；跨期套利；期货投资；人工智能

中图分类号：F830.9　　**文献标志码**：A

文 章 编 号：1673－9868(2022)01－0148－12

开放科学(资源服务)标识码(OSID)：

Research on Intertemporal Arbitrage Based on Machine Learning and Empirical Mode Decomposition

ZHOU Liang[1]，　CHEN Chen[2]，　LI Ning[1]

1. *School of Finance，Hunan University of Finance and Economics，Changsha 410205，China*；
2. *School of Finance，Southwestern University of Finance and Economic，Chengdu 611130，China*

Abstract：This paper used rolling EMD(Empirical Mode Decomposition) method to decompose the price gap of the CSI 300 stock index futures contract of the current month and the next month，and used three machine learning models (Elman network，RF，SVM) and ARIMA model to analyze and synthesize signals of different frequencies，and designed intertemporal arbitrage strategies based on the forecast results. The research results show that：the prediction accuracy of SVM，RF and ARIMA models is higher than that of Elman network. All models can achieve higher arbitrage returns，and the use of model fusion which combines liner and nonliner models can improve the risk control ability of the model. The combination of machine learning prediction and EMD decomposition technology can greatly increase the profitability of the

收稿日期：2021－01－04

基金项目：国家社会科学基金项目(20BJL061)；湖南省教育厅科学研究项目(21B0839).

作者简介：周亮，博士研究生，讲师，主要从事金融工程研究.

model without increasing the risk, so that the Sharpe ratio and the Sotino ratio of the model are both larger. Sub-sample test, full IMF signal prediction and arbitrage analysis based on the commodity futures market have all proved that the machine learning model integrated with EMD can achieve better arbitrage effects than pure machine learning models. The research conclusions help to promote the cross-integration research of artificial intelligence and finance, and also provide theoretical and practical references for futures investment.

Key words: machine learning; empirical mode decomposition; intertemporal arbitrage; futures investment; artificial intelligence

跨期套利是利用同一种期货品种、不同到期时间合约间价差的不寻常变动，进而实施反向交易，在两个合约间价差回归常态时进行平仓获利的投资方式. 相对于股票等金融工具的买入并持有策略而言，跨期套利由于交易的是同一种期货品种不同合约之间的价差，相对风险更低. 相对于跨品种或者跨市场套利，跨期套利的合约价差更为稳定，因此投资的稳定性更高，风险也相对较低. 跨期套利在价差超过正常值较远的时候进行反向交易，单笔利润相对于买入持有的趋势投资策略往往更低，由于期货市场具有较高的杠杆属性，且 T+0 的交易模式使得交易频率可以更高，致使套利交易的风险调整后收益往往更高[1-4]，致使越来越多的基金公司在实践中引入套利交易. 同时，套利交易与买入持有策略间的相关性极低甚至为负，因此是分散投资风险及规避尾部风险的重要手段，如 2020 年年初新冠肺炎疫情导致全球股票市场、债券市场、商品市场均发生了大幅回撤，如果在投资组合中加入套利交易，则可以对尾部风险进行极为有效的控制.

对价差的准确预测是跨期套利成功实施的关键所在，现有绝大部分文献及实际投资者均是利用价差均值回复原理的标准距离法设计策略，即当价差超过合理范围(常见的为均值±1 倍或多倍标准差)的时候进行反向交易，待价差回到均值附近时进行平仓[5-7]. 随着机器学习模型在金融预测领域应用得越来越广泛、且预测精度高，众多学者和投资者利用机器学习模型对价差进行预测，并在预测价差超过一定阈值后进行交易，从而获得套利收益. 常用来进行套利交易的机器学习模型包括人工神经网络[8-12]、支持向量机[13-14]和随机森林[15]等.

但是，直接对价差进行预测无疑丧失了许多细节信息，如熊志斌[16]和周亮[17]对人民币汇率的研究均发现，用 ARIMA 模型预测线性部分、用机器学习模型预测非线性部分或残差部分能够实现对离岸人民币汇率更精准的预测. Huang 等[18]提出的经验模态分解(EMD)模型在工程信号领域有着广泛的应用，该模型可以将信号分解为多个本征模函数(IMF)及残差项，每个本征模函数及残差项均有自身的特征益于分析及预测. 自 EMD 模型提出后，众多学者将该模型应用于经济问题分析，包括原油价格分析[19-20]、环境问题分析[21-23]等，相对于对原始数据的直接分析，利用分解信号进行分析的研究结果更为准确和稳健.

本文拟采用 EMD 模型对沪深 300 股指期货当月合约与下月合约的价差进行分解，并利用神经网络、支持向量机、随机森林以及 ARIMA 模型分别对高频和低频信号进行预测，再从预测准确性及套利绩效两个方面来评估模型的优劣. 相较于已有期货跨期套利的文献，本文的主要创新之处在于：① 通过 EMD 模型对原始价差变动序列进行滚动分解，再利用各机器学习模型对分信号进行预测，相对于纯机器学习预测模型，对序列信号考虑得更加周全和完整，也大幅提高了模型的预测精度及套利绩效；② 通过将多个机器学习模型及线性的时间序列模型进行比较及综合，既挑选出了更适用于跨期套利的模型，同时也将线性模型和非线性模型整合，在增加模型套利绩效的同时，也增加了机器学习模型的经济解释能力.

1　研究设计

1.1　机器学习模型

1.1.1　Elman 网络

Elman 神经网络是一种简单的循环神经网络，在众多学者的研究中均表现出超过普通反馈神经网络

(如BP网络)的特征[12,24]. Elman神经网络除了常见的输入层、隐藏层和输出层之外，在隐藏层的输入和输出之间增加了一个承接层，该模块存储了隐藏层的输入信号，再作为输入变量影响隐藏层的下期输入，具体结构如图1所示.

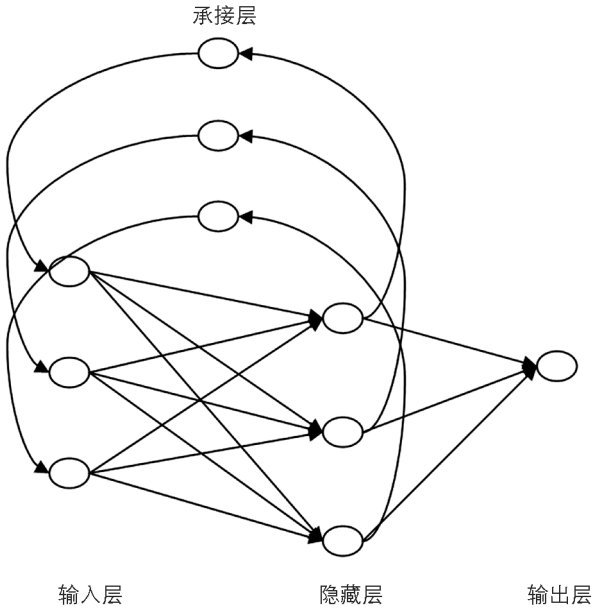

图1 Elman神经网络结构图

Elman神经网络的传导公式为：

$$\begin{aligned} \boldsymbol{x}_c(t) &= \boldsymbol{x}(t-1) \\ \boldsymbol{x}(t) &= f_1(w^1\boldsymbol{x}_c(t) + w^2(\boldsymbol{u}(t-1))) \\ y(t) &= f_2(w^3\boldsymbol{x}(t)) \end{aligned} \tag{1}$$

式(1)中，w^1, w^2, w^3 分别表示承接层到隐藏层、输入层到隐藏层及隐藏层到输出层之间的连接权重；$\boldsymbol{u}$ 为输入向量，$\boldsymbol{x}$ 和 $\boldsymbol{x}_c$ 分别为隐藏层和承接层的输出向量；$f_1(\cdot)$为隐藏层的激励函数，$f_2(\cdot)$为输出层的激励函数，本文采用常见的sigmod激励函数，由于实证中输入层包括20个节点，因此我们将隐藏层设置为40个节点.

1.1.2 随机森林

随机森林(RF)是一种集成学习方法，它的基本单元是决策树，每棵决策树都是一个分类器. 随机森林只关注树的集成学习，在树的集成(森林)产生之后，该模型使用投票的方法来组合预测结果，将投票次数最多的类别指定为最终的输出. 随机森林可以处理大量的数据，而大数据中所谓的“维数灾难”常常会让其他模型失败，同时随机森林对于大多数学习任务的误差率几乎和其他方法处于同等水平，并具有更少的过度拟合倾向. 本文中随机森林采用500颗决策树进行分析.

1.1.3 支持向量回归

SVM模型的目标是最大化支持向量与超平面之间的距离. SVM基于预测函数 $f(\cdot)$设置了一个通道 ε. 如果数据点在通道之内，则损失函数为零；如果数据点在通道之外，则损失函数设为$|y^i - f(w, x^i)| - \varepsilon$. 二次规划问题可以设置为：

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}\xi^i + \xi^i_*\right) \\ &st: \\ &y^i - f(w^T x^i) \leqslant \varepsilon + \xi^i \\ &f(w^T x^i) - y^i \leqslant \varepsilon + \xi^i_* \\ &\xi^i, \xi^i_* \geq 0, \ \forall i \in \{1, 2, \cdots, n\} \end{aligned} \tag{2}$$

式(2)中，ξ^i 与 ξ^i_* 为松弛因子，当数据点在通道内时为零. 通过引入拉格朗日乘子(α_i^1 和 α_i^2)，根据KKT(Karush-Kuhn-Tucker)条件，得到预测模型为：

$$y_* = \sum_{i=1}^{n}(\alpha_i^1 - \alpha_i^2)k(x^i, x_*) + b \tag{3}$$

式(3)中，$k(x^i, x_*)$为核函数，将原始数据的非线性特征映射到高维空间，从而能够采用线性关系来对数据进行预测，本文采用实证中最常见的RBF(Radial Basis Function)径向基核函数.

1.2 经验模态分解(EMD)

EMD是一种非线性、非平稳数据处理方法，它假定数据根据其复杂性可能同时存在多种振荡模式. EMD可以基于数据本身的局部特征，从原始时间序列提取出本征模函数(IMF)，它满足以下两个条件：① 函数的极值和零交叉数相同，或最多相差1；② 函数关于局部零均值是对称的. 这两个条件确保IMF近

似周期性的函数，并且均值为零. IMF 是一种类似谐波的函数，但在不同时间具有可变的幅度和频率.

EMD 具体计算步骤如下：① 确定时间序列 $x(t)$的所有极大值和极小值. ② 用 3 次样条插值生成其上下包络 $e_{\min}(t)$和 $e_{\max}(t)$. ③ 计算上下包络的逐点平均值 $m(t)=(e_{\min}(t)+e_{\max}(t))/2$. ④ 将 $x(t)$和 $m(t)$之差定义为 $d(t)=x(t)-m(t)$. ⑤ 如果 $d(t)$是 IMF，则将 $d(t)$表示为第 i 个 IMF，并用残差 $r(t)=x(t)-d(t)$替换 $x(t)$，第 i 个 IMF 通常表示为 $c_i(t)$；如果 $d(t)$不是 IMF，则用 $d(t)$替换 $x(t)$. ⑥ 重复步骤①至步骤⑤，直到残差项满足某种停止标准为止.

Huang 等[25]指出提取 IMF 的停止标准为：残差项满足零交叉数和极值相差不超过一个，并且可以满足下列预定标准：成分 $c_i(t)$或残差项 $r(t)$小于实际结果的预定值，或者残差项 $r(t)$变成单调函数，无法再提取 IMF. IMF 的总数一般限制为 $\log_2 N$，其中 N 是数据序列的长度. 原始时间序列可以表示为所有 IMF 和残差项的总和.

$$x(t)=\sum_{i=1}^{N}c_i(t)+r(t) \tag{4}$$

式(4)中 N 是 IMF 的数量，$r(t)$是最终的残差项.

EMD 往往分解出来的 IMF 信号比较多，如果对每个信号进行建模，无疑会加大计算机的运算难度，从而导致计算时间过长，因此我们借鉴 Zhang 等[22]的方法，将所有的 IMF 合成高频和低频两个部分，实现信号重构. 具体计算步骤为：① 计算每个成分(残差项除外)的 $c_1(t)$到 $c_i(t)$之和的平均值；② 使用 T 检验确定均值显著偏离零的 i；③ 在均值发生突变的变化点，使用 IMF 从该位置进行部分重建，分别合成低频部分和高频部分，即用 $c_1(t)$到 $c_i(t)$合成高频部分，用 $c_{i+1}(t)$到 $c_N(t)$合成低频部分.

1.3　模型绩效评估

为了对模型预测绩效进行评估，本文除选择常见的 *RMSE*，*MAE* 及 *Theil-U* 进行评估外，还选择了方向预测准确度 *DAR* 及样本外预测 R^2_{OS}，计算公式分别如式(5)至式(9)所示.

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{5}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{6}$$

$$Theil\text{-}U=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^{\,2}}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^{\,2}}} \tag{7}$$

$$DAR=\frac{1}{n}\sum_{i=1}^{n}a_i,\ a_i=\begin{cases}1 & \text{if } (y_{i+1}-y_i)(\hat{y}_{t+1}-y_i)>0\\ 0 & \text{其他}\end{cases} \tag{8}$$

$$R^2_{OS}=1-\frac{\sum_{i=1}^{T-1}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{T-1}(\bar{y}_i-y_i)^2} \tag{9}$$

式(5)－式(9)中，y_i，$\hat{y}_i$，$\bar{y}_i$ 分别表示实际值、模型预测值及样本内滚动均值. *RMSE*，*MAE* 及 *Theil-U* 越小，模型预测效果越好；*DAR* 介于[0，1]之间，数值越大模型预测效果越好；R^2_{OS} 是实证研究中常用来衡量样本外绩效的指标，其值介于(－∞，1]之间，数值越大模型预测效果越好.

1.4　套利模型设计

本文利用机器学习的预测结果来构造跨期套利策略，当模型预测下期价差与当期价差的差值大于 α 时，则买入当月合约，卖出下月合约；当模型预测下期价差与当期价差小于$-\alpha$ 时，则卖出当月合约，买入下月合约；当持有套利组合且模型预测值的绝对值小于 α 时平仓. 股指期货的杠杆是 10 倍，交易手续费为

0.23%%，样本区间内两个合约的均价在3 150附近，因此我们假定每单位交易手续费为0.15元. 考虑到期货市场杠杆率较高、风险较大，当出现套利机会时，我们只采用75%的资金进行滚动套利.

2 实证检验

2.1 样本描述

为了检验机器学习融合经验模态分解的跨期套利策略的可行性，本文选择沪深300股指期货的当月连续合约和下月连续合约进行分析，由于沪深300股指期货(以下简称IF合约)2010年4月16日才上市，因此最终选择了IF当月连续和下月连续合约2010年4月16日－2020年7月31日的所有日数据进行分析，共2 503个交易日. 图1报告了两个合约在样本区间的走势，左轴为IF当月连续合约价格曲线，右轴为IF下月连续合约价格曲线. 由图1可以看到，两者走势几乎一致，计算发现两者相关系数高达0.999，两者的价差在－130～70之间波动(99%置信区间)，存在着跨期套利的可行性.

图1 IF合约价格走势

2.2 基于机器学习的预测和套利

2.2.1 预测效果

采用3种机器学习方法(Elman网络、RF、SVM)及ARIMA模型对价差变动进行预测. 对于机器学习模型，采用前20期的数据($t-20$至$t-1$期)作为输入变量来预测第t期的价差；对于ARIMA则根据自相关系数(ACF)和偏自相关系数(PACF)确定模型的参数，并向前一步预测第t期的价差. 所有模型均采用1 000个滚动样本进行建模，即第1 001个价差变动数据是利用1～1 000个价差变动数据进行建模；第1 002个价差变动数据是利用2～1 001个价差变动数据进行建模，依次类推. 表1报告了4个模型的预测效果，可以看到SVM模型的*RMSE*、*Theil-U*指数和R^2_{OS}表现最佳；ARIMA模型的*MAE*和*DAR*表现最优；Elman模型表现相对较差，其R^2_{OS}甚至为负，说明用Elman进行预测略逊于用样本内均值进行预测的效果；RF模型虽然整体误差较SVM模型略高，但是其*DAR*却略优于SVM模型，这与其他很多研究相似，由于RF模型集成了多个决策树，表现出的结果更为稳健.

表1 对价格变动序列的预测效果

模型	*RMSE*	*MAE*	*Theil-U*	*DAR*	R^2_{OS}
Elman	0.145 8	0.095 0	0.147 7	0.636 0	－0.001 0
RF	0.140 8	0.091 5	0.142 5	0.625 8	0.057 7
SVM	0.140 4	0.091 5	0.142 3	0.605 1	0.064 0
ARIMA	0.142 3	0.090 9	0.143 8	0.723 2	0.034 4

2.2.2　套利分析

采用不同的 α 阈值进行套利，表 2 报告了 4 个模型的套利结果，其中 Panel A 是 $\alpha=1$ 时的套利效果，Panel B 是 $\alpha=4$ 时的套利效果，Panel C 是 $\alpha=8$ 时的套利效果. 第 2 至第 5 列分别报告了基于 Elman,RF,SVM 及 ARIMA 模型预测结果的套利效果，为了避免单一模型进行预测时的弊端，第 6 列和第 7 列综合了 RF 模型和 ARIMA 模型预测结果进行套利. 本文采用 RF 模型是因为其表现较为稳健，预测效果介于 Elman 和 SVM 之间；综合一个非线性的机器学习模型(RF)和一个线性的时序预测模型(ARIMA)，预期会增加套利模型的稳健性；第 6 列是将两个模型预测值进行平均，第 7 列是只有两个模型预测值都超过阈值时才进行套利.

表 2　套利结果分析

指标	Elman	RF	SVM	ARIMA	平均	综合
Panel A：$\alpha=1$						
年化收益率/%	36.04	44.82	30.39	46.29	41.14	43.77
波动率//%	29.62	32.36	14.74	30.44	32.65	27.10
下行波动率/%	15.74	15.03	6.09	13.57	15.84	11.10
最大回撤/%	25.04	22.91	7.43	19.60	31.77	13.05
夏普比率	1.115 4	1.292 4	1.857 9	1.422 2	1.168 3	1.504 4
索提诺比率	2.099 1	2.782 5	4.498 6	3.190 7	2.408 3	3.673 1
胜率/%	57.55	59.91	60.03	61.71	63.22	66.84
持仓时间占比/%	59.96	42.94	40.38	44.97	39.84	25.86
Panel B：$\alpha=4$						
年化收益率/%	22.59	22.24	13.46	29.37	29.75	22.97
波动率/%	26.03	23.28	10.17	25.63	24.50	20.46
下行波动率/%	13.79	12.48	3.29	10.23	9.04	8.04
最大回撤/%	21.40	22.91	5.66	15.85	11.07	11.07
夏普比率	0.752 3	0.826 4	1.028 5	1.028 8	1.091 6	0.975 8
索提诺比率	1.419 9	1.541 2	3.182 6	2.576 5	2.958 1	2.483 6
胜率/%	58.47	67.11	70.15	65.66	68.12	68.75
持仓时间占比/%	15.94	10.26	4.52	11.21	9.32	6.48
Panel C：$\alpha=8$						
年化收益率/%	13.70	13.16	6.93	15.43	14.21	13.75
波动率/%	21.41	18.28	7.33	21.56	17.81	17.24
下行波动率/%	8.84	8.60	0.01	8.38	7.66	7.32
最大回撤/%	13.90	13.13	0.02	11.07	11.07	11.07
夏普比率	0.499 8	0.555 7	0.536 9	0.576 4	0.629 1	0.623 3
索提诺比率	1.210 8	1.180 3	383.060 0	1.482 4	1.463 4	1.468 1
胜率/%	67.69	72.09	91.67	68.52	75.00	81.82
持仓时间占比/%	4.39	2.90	0.81	3.65	2.97	2.23

注：计算夏普比率时采用银行一年期定期存款利率作为无风险利率；索提诺比率是用超额收益除以下行标准差，相对夏普比率仅分母不同；下同.

由表 2 可知，所有模型在任何阈值下均能取得较高的套利收益，收益率最高的是 ARIMA 模型在 $\alpha=1$ 时，年化收益率高达 46.29%；胜率最高的是 SVM 模型在 $\alpha=8$ 时，胜率高达 91.67%，但是其交易时间很短，仅交易了 0.81%的样本时间，即仅交易了 12 次，其他所有模型的胜率均在 57%以上，说明机器学习在

进行股指期货跨期套利时，总体胜率均不错；最大回撤最低值是SVM模型在$\alpha=8$时，仅回撤了0.02%，同时可以看到，绝大部分模型最大回撤均能控制在20%以内，说明套利模型风险控制较好；所有模型的波动率均低于33%，下行波动率均低于16%，因此模型夏普比率和索提诺比率均较好. 夏普比率最高的是SVM模型在$\alpha=1$时，高达1.857 9；索提诺比率最高的是SVM模型在$\alpha=8$时，高达383.06，但是由于此时交易量过低导致下行波动率极低，索提诺比率次高的是SVM模型在$\alpha=1$时，达到了4.498 6. 从第6列和第7列可以看出，相对于仅采用RF或ARIMA进行预测，混合模型的风险控制更好，表现为更低的波动率、下行波动率以及最大回撤，尤其是第7列，只有当两个模型预测值均大于阈值时才进行套利，风险控制更为出色，说明将非线性模型和线性模型融合使用能够改善模型的风险控制能力，在实践中可能应用价值更高. 实际上，采用SVM与ARIMA相结合的模型风险控制更佳，限于篇幅，结果未列出.

2.3 EMD分解及机器学习预测

为了更好地了解跨期价差的微观结构，提高跨期套利的绩效表现，本文采用EMD模型对原始价差变动数据进行信号分解(图2). 由图2可知，EMD模型将原始信号分成了10个IMF信号及1个残差信号，从IMF1-IMF10分别表示从高频到低频的本征模函数. 图2中越低频的信号越平稳，IMF10及残差信号已经变成了一条非常平滑的曲线. 由于对所有序列进行建模会加大计算机的工作量，本文后面的分析将借鉴Zhang等[22]的方法，将所有IMF合成一个高频信号和一个低频信号，其中高频信号波动剧烈，与原始信号相似性较强，而低频信号及残差信号则表现出较强的线性特征.

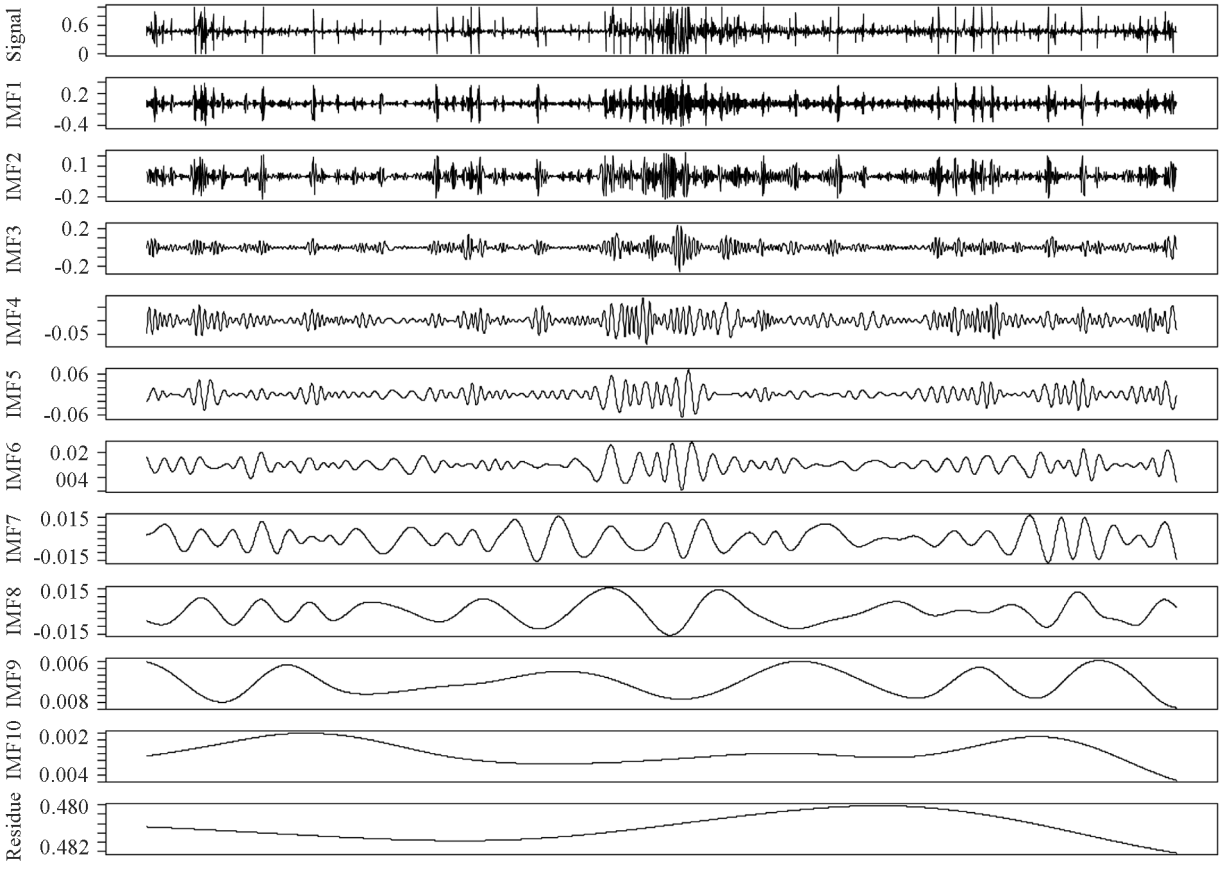

图2 EMD分解

图2是将整个样本区间的价差变动进行分解，如果直接对其进行分析，则在套利过程中用到了未来信息，这显然不符合实际情况，因此本文采用滚动窗口来进行信号分解，即在训练机器学习模型的滚动窗口区间(1 000个数据)进行EMD分解，借鉴Li等[22]的方法，将多个IMF合成一个高频信号和一个低频信号，然后分别利用Elman,RF,SVM和ARIMA这4种模型来对高频信号、低频信号及残差信号进行预测，再将3个信号的预测值汇总得到最终的价差变动预测值. 表3报告了EMD分解信号的预测结果，其中Panel A报告了模型的预测效果，Panel B报告了$\alpha=1$时的套利结果(限于篇幅，这里仅报告了$\alpha=1$时的

结果). 从 Panel A 可知，Elman 网络的预测效果较差，其 R^2_{OS} 远小于 0，说明效果不及利用均值的预测效果，RF,SVM 和 ARIMA 的预测效果相对较好，因此在 Panel B 中本文仅采用了 RF,SVM 和 ARIMA 这 3 种模型进行套利. 同时，本文还报告了 RF 与 ARIMA 平均值及两者同时超过阈值的综合模型的套利结果. 将 Panel B 的结果与表 2 进行比较可以发现，除了 SVM 模型套利的风险有所提高外，其他模型的风险均与表 2 大体相当，但是所有模型的收益率都有了大幅度上升，尤其是 ARIMA、平均模型及综合模型，从而使得模型的夏普比率和索提诺比率均有较大幅度上涨. 表现最好的是 EMD-ARIMA 模型，其年化收益率高达 96.52%，夏普比率和索提诺比率分别高达 2.854 9 和 8.271 1.

表 3　EMD 滚动套利结果

Panel A：预测结果					
模型	*RMSE*	*MAE*	*Theil-U*	*DAR*	R^2_{OS}
Elman	0.192 8	0.094 6	0.187 9	0.639 6	−1.349 4
RF	0.099 5	0.061 3	0.100 9	0.685 3	0.096 9
SVM	0.098 0	0.060 5	0.099 7	0.664 4	0.119 7
ARIMA	0.095 8	0.059 0	0.097 0	0.757 7	0.152 8
Panel B：套利结果($\alpha=1$)					
指标	RF	SVM	ARIMA	平均	综合
年化收益率/%	49.17	55.64	96.52	81.09	82.63
波动率/%	31.84	25.95	32.76	31.52	29.19
下行波动率/%	14.79	13.76	11.31	11.30	10.22
最大回撤/%	30.09	15.14	16.52	17.91	21.03
夏普比率	1.449 9	2.028 6	2.854 9	2.477 6	2.727 7
索提诺比率	3.121 2	3.826 2	8.271 1	6.909 6	7.790 3
胜率/%	59.73	63.14	63.70	63.73	65.02
持仓时间占比/%	49.97	48.55	57.66	50.64	32.82

图 3 展示了套利模型的净值走势图. 图 3 总体来看，各曲线均能保持较平稳的上升趋势，尤其是 EMD-ARIMA 模型，最终净值接近 60，平均模型和综合模型也获得了较高的回报，最终净值均在 30 以上. 综合来看，跨期套利相对于买入持有策略，风险更低(绝大部分股票指数的年化波动率均在 30%以上，最大回撤一般在 50%以上，个股的波动率和最大回撤更高)，如果能够选择到合适的套利模型，同样能够获得非常高的投资收益，进而大幅提高投资的风险调整后收益.

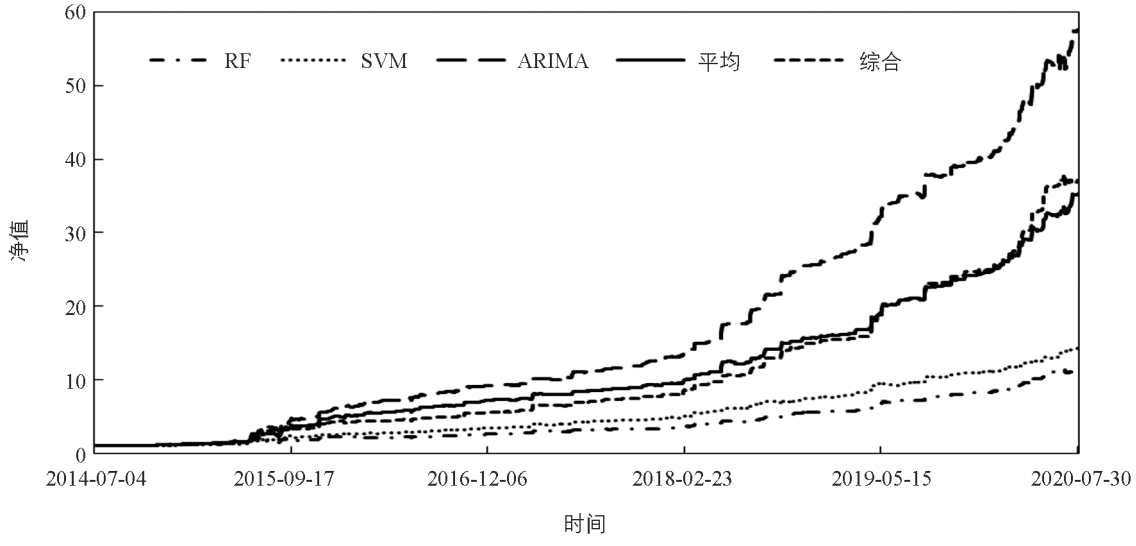

图 3　套利模型净值走势图

2.4 分样本稳健性检验

为了检验研究结论的稳健性，本文将整个套利区间划分为两个时间相等的分样本，各包括3年时间，分别是2014年7月－2017年7月、2017年8月－2020年7月. 表4报告了分样本检验结果，其中Panel A和Panel B是2014年7月－2017年7月的套利结果，Panel C和Panel D是2017年8月－2020年7月的套利结果；Panel A和Panel C仅采用了机器学习模型，Panel B和Panel D采用了机器学习与EMD相结合的套利模型(限于篇幅，同样仅报告了$\alpha=1$时的套利结果). 由表4可以看到与全样本相似的结果，无论是2017年7月以前还是以后，机器学习加EMD模型的套利风险虽然与纯机器学习模型相当，但其套利收益却要显著高于纯机器学习模型(除第一阶段RF＋EMD的投资收益相对RF模型略有降低外)，从而使得机器学习加EMD模型的夏普比率和索提诺比率均显著高于纯机器学习模型，本文的研究结论稳健. 从表4还可知，相对于2017年7月之前，2017年7月之后的套利收益有所下降，套利风险也有所降低，这也间接说明随着期货市场的不断发展，市场有效性在逐步提高，从而使得套利空间有所收窄.

表4 分样本稳健性检验

指标	Panel A：机器学习(2014.7－2017.7)					Panel B：机器学习＋EMD(2014.7－2017.7)				
	RF	SVM	ARIMA	平均	综合	RF	SVM	ARIMA	平均	综合
年化收益率/%	57.41	23.88	62.76	51.71	55.71	46.52	62.64	124.24	104.59	92.15
波动率/%	42.99	13.04	39.56	43.12	35.41	42.09	32.88	42.83	41.40	38.47
下行波动率/%	20.16	4.65	17.06	20.79	14.66	20.25	18.38	15.18	15.34	14.07
最大回撤/%	22.91	4.12	19.60	31.77	13.05	30.09	15.14	16.52	17.91	21.03
夏普比率	1.265 6	1.601 0	1.510 6	1.129 6	1.488 6	1.033 9	1.813 9	2.830 9	2.453 7	2.317 4
索提诺比率	2.699 6	4.493 4	3.502 4	2.342 7	3.596 3	2.149 0	3.243 9	7.985 7	6.623 8	6.335 6
胜率/%	60.32	62.02	61.02	62.69	66.96	58.02	63.16	61.04	62.41	60.77
持仓时间占比/%	49.73	27.73	49.60	44.67	29.87	54.00	40.53	61.60	53.20	34.67
指标	Panel C：机器学习(2017.8－2020.7)					Panel D：机器学习＋EMD(2017.8－2020.7)				
	RF	SVM	ARIMA	平均	综合	RF	SVM	ARIMA	平均	综合
年化收益率/%	33.23	37.23	31.49	31.31	32.75	51.88	48.95	72.23	60.28	73.58
波动率/%	14.90	16.28	16.38	15.79	14.12	15.35	15.95	16.97	15.80	14.37
下行波动率/%	5.79	7.26	8.39	7.65	5.20	3.80	5.98	4.26	3.65	2.42
最大回撤/%	6.22	7.43	7.30	6.11	6.15	4.24	6.43	3.73	3.17	3.78
夏普比率	2.029 4	2.103 2	1.739 8	1.793 0	2.106 4	3.184 6	2.881 7	4.079 5	3.625 3	4.910 5
索提诺比率	5.225 5	4.716 5	3.396 8	3.699 9	5.721 4	12.846 3	7.683 0	16.252 0	15.676 7	29.164 0
胜率/%	59.32	58.97	62.59	63.92	66.67	61.79	63.13	66.84	65.24	69.91
持仓时间占比/%	35.98	53.28	40.16	34.84	21.72	45.83	56.69	53.55	48.02	30.87

2.5 EMD全分解滚动套利效果

表3和表4的分析均是基于EMD分解后再将多个本征模函数合成一个高频信号和一个低频信号，这样的操作方式可以极大地提高计算机的运算速度，但是也会丧失较多的信号信息，因此本文利用RF,SVM和ARIMA分别对每个本征模函数及残差信号进行预测，再综合为最终的预测值. 相对于合成两个信号，这种方法利用到了更多的信息，但是运行速度慢了约5倍. 表5报告了对每个分解信号单独进行预测的套利结果，其中Panel A是模型的预测偏差，Panel B是基于预测值的套利结果，同样仅报告了$\alpha=1$时的套利绩效. 与表3相比较可知，基于EMD所有信号的套利模型，RF模型和SVM模型的预测精度有所提高，ARIMA略有下降. 所有模型的投资收益均有一定幅度的上升，波动率也略有上升，而下行波动率反而有所下降(除SVM模型略有上升)，因此模型的夏普比率和索提诺比率均大幅上升，同时模型的胜率也显著提高. 总体来看，基于EMD所有信号预测值的套利模型相对于将信号合成高频和低频的模型，投资绩效又有了一定程度的上升，只是损失了计算机的运行速度，在实际投资过程中可能会因价格变动过快而导致

实际投资收益与回测收益有一定的偏差，比较适合于较低频率及较稳定市场的套利投资.

表 5　基于 EMD 所有信号的套利结果

模型	Panel A：预测效果				
	RMSE	*MAE*	*Theil-U*	*DAR*	R^2_{OS}
RF	0.095 6	0.056 3	0.095 5	0.760 2	0.155 2
SVM	0.090 6	0.053 1	0.091 4	0.758 9	0.226 9
ARIMA	0.098 1	0.057 2	0.098 0	0.795 0	0.118 8
指标	Panel B：套利结果($\alpha=1$)				
	RF	SVM	ARIMA	平均	综合
年化收益率/%	94.20	70.20	102.47	104.87	109.80
波动率/%	32.78	30.42	32.60	32.66	32.04
下行波动率/%	8.19	16.42	6.55	6.56	6.09
最大回撤/%	9.41	22.31	9.41	9.41	12.53
夏普比率	2.782 2	2.208 9	3.051 4	3.119 3	3.333 8
索提诺比率	11.129 3	4.093 6	15.191 5	15.531 0	17.543 1
胜率/%	73.46	73.48	74.15	75.02	77.25
持仓时间占比/%	82.17	81.23	81.77	82.17	68.26

2.6　商品期货跨期套利研究

为了进一步检验研究结论的稳健性，本文还对商品期货进行了检验. 螺纹钢是商品期货中交易量最大的品种，因此选择螺纹钢期货 2020 年 8 月和 2020 年 9 月到期合约的 30 min 数据进行分析，为了避免合约刚上市及快要交割时价格波动幅度过大的弊端，本文选择了这两个合约 2019 年 10 月 15 日－2020 年 7 月 15 日的所有 30 min 数据，共 2 190 组. 同样滚动采用 1 000 组数据来进行建模，通过 EMD 进行分解后将信号合成高频和低频两部分，并分别利用 RF，SVM 和 ARIMA 进行预测并整合. 表 6 报告了机器学习＋EMD 套利结果，其中 Panel A 是模型的预测结果，Panel B 是模型的套利效果.

表 6　螺纹钢期货机器学习＋EMD 套利

模型	Panel A：预测结果				
	RMSE	*MAE*	*Theil-U*	*DAR*	R^2_{OS}
RF	0.173 7	0.126 1	0.173 5	0.723 6	0.099 9
SVM	0.171 3	0.126 8	0.170 8	0.698 4	0.125 1
ARIMA	0.166 3	0.124 3	0.165 4	0.784 7	0.174 9
指标	Panel B：套利效果				
	RF	SVM	ARIMA	平均	综合
年化收益率/%	95.82	96.65	125.27	118.95	98.00
波动率/%	27.08	26.68	27.46	27.19	25.20
下行波动率/%	11.15	10.54	11.04	11.10	8.88
最大回撤/%	6.95	11.14	11.33	9.32	5.77
夏普比率	3.427 9	3.510 0	4.452 6	4.263 7	3.769 9
索提诺比率	8.327 7	8.886 1	11.071 3	10.444 3	10.700 2
胜率/%	56.76	54.76	57.78	58.06	59.12
持仓时间占比/%	76.63	76.46	80.91	79.62	74.93

由表 6 可知，3 个模型均能对螺纹钢期货的价差变动进行较好的预测，且基于预测值的套利模型能够取得非常不错的套利绩效. ARIMA 模型的套利绩效最优，其夏普比率和索提诺比率分别高达 4.45 和 11.07；而综合模型的套利风险最低，下行波动率和最大回撤分别为 8.88%和 5.77%. 总体来看，EMD 分

解能够改善机器学习模型的套利绩效，而将线性ARIMA模型和非线性机器学习模型结合使用的综合模型，能够更好地控制投资风险，是更为稳健的投资方式.

3 结论与讨论

选择IF当月连续和下月连续合约2010年4月16日—2020年7月31日的所有日数据，利用3种机器学习方法(Elman,RF,SVM)及ARIMA模型对两个合约的价差变动序列进行预测并构建套利模型. 研究结果发现：① SVM和ARIMA模型的预测精确度相对较高，Elman模型表现较差，而RF模型由于集成了多个弱分类器，表现出的结果较为稳健. ② 所有模型在任何阈值下均能取得较高的套利收益，同时绝大部分模型最大回撤均能控制在20%以内，波动率均低于33%，下行波动率均低于16%，说明套利模型风险控制较好；相对于仅采用RF或ARIMA进行预测，混合模型(将预测值进行平均或作为并列条件)的风险控制更好，表现为更低的波动率、下行波动率及最大回撤，说明将非线性模型和线性模型融合使用能够改善模型的风险控制能力. ③ 将机器学习预测与EMD分解技术相融合可以在不提高风险的同时大幅提高模型的收益率，从而使得模型的夏普比率和索提诺比率均有较大幅度上升，表现最好的是EMD-ARIMA模型，其年化收益率高达96.52%，夏普比率和索提诺比率分别高达2.854 9和8.271 1. ④ 分样本检验、全IMF信号预测及基于商品期货市场的套利分析，均证明融合EMD的机器学习模型可以获得比纯机器学习模型更优异的套利效果.

本文的研究结论不仅是对期货投资理论及人工智能方法在金融领域中应用的补充，同时也具有较强的实践价值：① 跨期套利是一种有效的投资策略，相对于买入持有等基于价格预测的投资策略，套利策略的风险更低，如果方法得当，收益却反而可能获得提高. 同时，大量理论研究及实践均证明，商品期货策略(尤其是套利策略)与股市等投资策略的相关性极低甚至为负，因此在股票投资策略中增加跨期套利策略，可以有效降低整体投资组合的风险，从而提高投资收益率，并且可以在极端的市场风险下保护资产的安全性. ② 机器学习模型在对非线性金融时间序列数据进行预测时具有较好的效果，但是机器学习模型完全由数据驱动，其经济基础较为薄弱，因此将其与经济基础更为稳健的线性预测模型相结合，可以在提升模型预测能力的同时，增加模型的经济解释能力. ③ 金融时间序列具有较高的复杂性及噪声比率，采用单一模型进行预测无疑会丧失很多信息，通过EMD等信号分解模型将金融时间序列进行分解，通过趋势成分或波动成分的提取分别进行预测，可以实现对金融时间序列更为准确的预测，并进而提升跨期套利成功的几率.

参考文献：

[1] 杨云飞，鲍玉昆，胡忠义，等. 基于EMD和SVMs的原油价格预测方法[J]. 管理学报，2010，7(12)：1884-1889.

[2] JACOBS H，WEBER M. On the Determinants of Pairs Trading Profitability[J]. Journal of Financial Markets，2015，23：75-97.

[3] 张波，刘晓倩. 基于EGARCH-M模型的沪深300股指期货跨期套利研究——一种修正的协整关系[J]. 统计与信息论坛，2017，32(4)：34-40.

[4] 刘海飞，李伟，李冬昕，等. 股指期货跨期套利自适应机制理论与实证——基于沪深300股指期货高频数据的证据[J]. 华东经济管理，2018，32(11)：102-111.

[5] KRAUSS C，DO X A，HUCK N. Deep Neural Networks，Gradient-Boosted Trees，Random Forests：Statistical Arbitrage on the S&P 500[J]. European Journal of Operational Research，2017，259(2)：689-702.

[6] HAIN M，HESS J，UHRIG-HOMBURG M. Relative Value Arbitrage in European Commodity Markets[J]. Energy Economics，2018，69：140-154.

[7] 邢亚丹，劳兰珺，孙谦. 跨期套利收益与风险来源探究——基于沪深300股指期货高频跨期套利策略[J]. 投资研究，2015，34(10)：98-109.

[8] DUNIS C L，LAWS J，EVANS B. Modelling and Trading the Soybean-Oil Crush Spread with Recurrent and Higher Or-

der Networks：a Comparative Analysis [J]. Neural Network World，2006，16(3)：193-213.

[9] HUCK N. Pairs Selection and Outranking：an Application to the S&P 100 Index [J]. European Journal of Operational Research，2009，196(2)：819-825.

[10] WILES P S，ENKE D. Nonlinear Modeling Using Neural Networks for Trading the Soybean Complex [J]. Procedia Computer Science，2014，36：234-239.

[11] 王文波，费浦生，羿旭明. 基于 EMD 与神经网络的中国股票市场预测 [J]. 系统工程理论与实践，2010，30(6)：1027-1033.

[12] 刘建和，梁仁方，王玉斌，等. 大豆期货合约均值回归套利策略和 Elman 神经网络套利策略对比研究 [J]. 湖南财政经济学院学报，2016(3)：8-15.

[13] 邓亚东，王波. 基于高斯核支持向量机的商品期货市场套利研究 [J]. 经济数学，2018，35(1)：27-30.

[14] 周亮. 基于价差预测的商品期货跨期套利研究 [J]. 金融理论与实践，2019(7)：84-92.

[15] HUCK N. Large Data Sets and Machine Learning：Applications to Statistical Arbitrage [J]. European Journal of Operational Research，2019，278(1)：330-342.

[16] 熊志斌. ARIMA 融合神经网络的人民币汇率预测模型研究 [J]. 数量经济技术经济研究，2011，28(6)：64-76.

[17] 周亮. 机器学习融合 ARIMA 模型的离岸人民币汇率预测 [J]. 统计学报，2020，1(2)：48-56.

[18] HUANG N E，SHEN Z，LONG S R，et al. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis [J]. Proceedings of the Royal Society of London Series A：Mathematical，Physical and Engineering Sciences，1998，454(1971)：903-995.

[19] ZHANG X，LAI K K，WANG S Y. A New Approach for Crude Oil Price Analysis Based on Empirical Mode Decomposition [J]. Energy Economics，2008，30(3)：905-918.

[20] 杨云飞，鲍玉昆，胡忠义，等. 基于 EMD 和 SVMs 的原油价格预测方法 [J]. 管理学报，2010，7(12)：1884-1889.

[21] 米子川，姜天英. 煤炭大数据指数编制及经验模态分解模型研究 [J]. 统计与信息论坛，2016，31(8)：71-77.

[22] LI H T，BAI J C，CUI X，et al. A New Secondary Decomposition-Ensemble Approach with Cuckoo Search Optimization for Air Cargo Forecasting [J]. Applied Soft Computing，2020，90(1)：1-19.

[23] SUN S L，WANG S Y，WEI Y J. A New Multiscale Decomposition Ensemble Approach for Forecasting Exchange Rates [J]. Economic Modelling，2019，81：49-58.

[24] 吴曼曼，徐建新. 基于 EMD 改进的 Elman 神经网络对股票的短期预测模型 [J]. 计算机工程与科学，2019，41(6)：1119-1127.

[25] HUANG N E，WU M L C，LONG S R，et al. A Confidence Limit for the Empirical Mode Decomposition and Hilbert Spectral Analysis [J]. Proceedings of the Royal Society of London Series A：Mathematical，Physical and Engineering Sciences，2003，459(2037)：2317-2345.

责任编辑　夏娟