文章编号:1003－207(2018)06－0026－13
DOI:10.16381/j.cnki.issn1003－207x.2018.06.004

FEPA－金融时间序列自适应组合预测模型

潘和平[1,4],张承钊[2,3]

(1.重庆金融学院,重庆 400067; 2.电子科技大学经济与管理学院,四川 成都 611731;
3. 成都职业技术学院,四川 成都 610041; 4. Swingtum Prediction,澳大利亚)

摘 要:本文报告一种金融时间序列预测的信号分析、信息融合与智能计算组合模型,简称 FEPA,由针对金融时间序列(FTS)信号分析的经验模态分解(EMD)、用于数据降维的主成分分析(PCA)和用于非线性建模的人工神经网络(ANN)三部分组成。该模型首先应用滑动窗口截取原始金融时间序列最近期数据集,应用 EMD 分解算法把数据集分解成不同尺度的本征模态函数(IMF),然后通过主成分分析将分解后的数据降维,提取最有信息量的特征;然后将这些特征输入到神经网络进行组合预测。本文提出的组合预测模型 FEPA 是基于分解－提优－合成的信息融合思想,有效提高了预测可靠性。其创新点在于:1)首次给出了 EMD 算法的结构化表达,提供了今后融合更多信息的算法接口;2)通过多步长预测输出深入研究 EMD 分解的有效信息结构;3)通过切换到更细时间框架来处理 EMD 的端点效应,并探索了两级时间框架下的预测效果;4)给出了金融时间序列组合预测模型的一般性架构,具有可升级性和可扩展性。并且通过滑动窗口 EMD 使得实证更能切近实际。通过在沪深 300 股指和澳大利亚股指上的实证,结果表明 FEPA 预测模型在沪深 300 股指日线和 15 分钟线上的预测命中率高达 78%和 82%,在澳大利亚股指日线上也达到了 74%的命中率,经比较,明显高于文献中常见的 5 种模型。

关键词:金融时间序列预测;经验模态分解;主成分分析;人工神经网络;本征模态函数;信息融合;预测可靠性
中图分类号:F832.5 **文献标识码**:A

1 引言

本文研究金融市场价格的非线性预测模型。原则上,这里所论及的原理、方法、模型和实施应该适用于股市(股指、股指期货、个股价格等)、期货(以商品期货为主)与外汇等三大类主要金融市场。但作为实证研究的对象,本文只限于股票市场指数,简称股指。在具体选择实证对象上,这里选取了中国的沪深 300 股指(HS300)和澳大利亚的大盘指数(AXJO)。在时间框架上,日线预测在中国股市 T＋1 的制度下具有典型的意义。因而,本文研究的主要目标就是在日线框架下股指预测的非线性建模与实证。

股指作为一国股票市场代表性股票群的价格加权平均,比个股价格变化具有更为复杂的行为模式,具有非线性、混沌、长记忆等特点。就因果关系而言,股指水平的波动受到政治经济、金融市场、社会情绪、自然环境等诸多因素的影响。这些影响因素之间有着复杂的交互过程,对市场波动的影响有着不同的时滞和强度。股票市场是现代市场经济的主体,是一个风险和利益共存的复杂系统。鉴于股市周期一般领先于经济周期几个月的时间,股市总是扮演着经济晴雨表的角色。股票市场的分析与预测建模研究有着多重重要意义:在宏观层面,若能有效地预测股市的股指走势和波动,对于国家层面制定宏观经济政策和国民经济发展计划具有重要的信息支持作用;在中观层面,股指或股价的走势与波动风险预测是企业管理资产风险和上下游供应链风险所不可或缺的;在微观层面,股市投资者在决策投资股票时自然首先要考虑股市总体的走势方向与涨跌空间、以及所关注行业和个股的走势方向和波动风险。股票投资上最大的风险首先在于投资方向与市场方向的不一致。

然而,无论是股市大盘指数、还是标准行业或自定义板块指数、还是个股价格,就数据形式而言都具有一般金融时间序列(FTS – Financial Time Se-

收稿日期:2015－10－23; 修订日期:2017－10－15
基金项目:国家社科基金资助项目(17BGL231);中国智能金融研究院(香港)资助项目
通讯作者简介:潘和平(1961—),男(汉族),陕西西安人,重庆金融学院教授,博导,研究方向:智能金融,E-mail:178372311@qq.com.

ries)的标准数据结构。因而,股指的预测本身也是金融时间序列预测的重要类型。关于金融时间序列预测问题,计量金融经济学文献中已经有大量的文章和资料。Box 和 Jenkins 提出了自回归移动平均模型(ARIMA)用于处理非平稳信号之间的线性相关性。Engle 提出了自回归条件异方差模型(ARCH)已经被许多金融分析师采用。Bollerslev 提出了广义自回归条件异方差模型(GARCH)进一步预测误差方差,该模型的分析对投资者的交易决策发挥了非常重要的作用。但应该说明,这些自回归类型的金融时间序列模型主要是针对金融时间序列的波动率的。而金融学中长期占主导地位的"有效市场理论"主要表现为金融市场价格的几何布朗运动模型,即认为市场价格是一种随机游走,因而不具有可预测性。

自 1980 年代以来,由多学科交叉而形成的智能计算金融实证研究表明,金融市场具有一定的概率可预测性。其中包括有混沌理论模型、金融物理模型与计算智能金融模型等,Pan 等[1－2]给出了较全面的文献综述,已经回答了金融市场的概率可预测性问题,这里不再赘述。

尽管有人工智能所产生的多种智能计算模型,如 ANN,SVM, 以及决策树与随机森林等都是具有极强的非线性建模潜力的机器学习乃至深度学习模型,但建立真正有效的金融市场价格预测模型的关键点首先在于如何从金融时间序列数据中提取有信息负荷的特征。这一个步骤一般称为特征提取。过去主要依赖对问题本身的经验观察和技术性理解来人工设计特征空间。现在有了深度学习的思路,我们当然也可以直接将原始数据作为输入然后构建多隐蔽层的 ANN 来让系统自学习出一种完全黑箱模型。但我们仍然不满足于这种黑箱模型,而是试图发展出一种基于信号分析的信息融合预测模型,因为这样的模型有助于构建理性的预测模型。

综合文献中出现的各种时间序列预测模型,我们认识到,一般的金融市场预测模型的一般架构应该有三部分组成:第一部分是针对原始金融时间序列数据的信号分析,第二部分是在信号分析的基础上进行特征提取,第三部分是以这些特征为输入并以未来走势为输出建立预测计算模型。

针对一般时间序列的信号分析问题,1998 年由来自美国 NASA 的信号处理专家 Huang 等[3]提出了 Hilbert－Huang 变换(简称 HHT)。这是针对非线性非平稳数据的一种经验性的、非参数化的、而又完备的、兼顾时域和频域的分析方法。它包含经验模态分解(EMD)算法和 Hilbert 变换两个过程。通过 EMD,任何复杂的信号都可以被分解成数目有限的且常常不太多的几个本征模态函数(Intrinsic Mode Function, IMF),而这些 IMF 序列具有性能良好的 Hilbert 变换,能够很好地刻画原始数据在每一个局部的振荡结构,以此为基础得到的 Hilbert 谱具有很好的时频特性。自 EMD 方法提出以来已被初步用于金融时间序列的预测,如原油[4－5]、外汇[6]、铜价[7]、房地产[8]和股票市场[9－12]等。然而,几乎所有这些文献都没有注意到 EMD 分解本身具有一个先天性的缺陷,那就是 EMD 分解的过程具有端点效应,也就是在时间序列的最右端所对应的分解结果是不稳定的,随着新数据的加入,分解的结果要不断更新。因而,绝大多数已经发表的基于 EMD 的金融预测模型在实证时都隐含了这种先天缺陷而变得并不可靠。另外,这些现有文献对于 EMD 的算法表述也是完全基于原创作者 Huang 等的表述,其算法的表达仍然是流水式的顺序算法,而没有清晰的结构化,并且若干关键计算过程并未清楚表达。这两点也是本文的重点关注问题。

对于预测建模中的特征提取问题,在大数据时代,这个问题也可以表述为数据降维问题。主成分分析(Principal Component Analysis,PCA)又称主分量分析,是由皮尔逊(Pearson)于 1901 年首先引入,后来由霍特林(Hotelling)于 1933 年进行了发展。主成分分析是一种通过降维技术把多个变量转化为少数几个主成分(即综合变量)的多元统计方法,这些主成分能够反映原始变量的大部分信息,通常表示为原始变量的线性组合,为使得这些主成分所包含的信息互不重叠,要求各主成分之间互不相关。一般来说,当研究的问题涉及很多变量,并且变量间相关性明显,即包含的信息有所重叠时,可以考虑用主成分分析,这样更容易抓住事物的主要矛盾,使得问题得到简化。对于基于 EMD 的预测模型而言,一个金融时间序列经过 EMD 分解后形成多级 IMF 序列, 这些 IMF 序列数据本身包含着冗余信息, 因而需要降维处理。使用 PCA 对 EMD 变换产生的 IMF 序列进行降维处理后再来构建非线性预测模型,是一条必由之路,目前文献中还很少提及。

最近,在基于 EMD 的预测模型研究上有一些新的发展。Dragomiretskiy 和 Zosso[13]提出了一种变分模式分解 VMD (Variational Mode Decomposition), 用来实现多分辨率的模式分解。Samiri[14]

将 VMD 同回归神经网络相结合。Wei Liangying[14]将 EMD 与模糊逻辑相结合。这些新发展同本文对于 EMD 的深化研究不再一个方向上。

从以上研究文献可以看出，虽然各种非线性智能计算模型在不同的条件下具有较高的预测精度，但是目前的主流趋势是利用各单一智能算法模型的优点构建组合预测模型来实现更好的预测可靠性。Krogh 和 Vedelsby[16]证明如下思想：当构成组合预测模型的单一模型足够精确且足够多样化时，组合预测模型一定能获得比单一模型更好的预测效果。当前智能计算模型的发展趋势正是发展各种算法的组合方法，因为任何一种典型方法不能在所有场合都优于其他方法。虽然国内外学者已经意识到应该采用 EMD 方法研究股票市场、外汇市场和原油市场，但是他们大多采用长达几年的窗口采集数据进行 EMD 分解。然而，股市预测的实时性要求很高，长达几年的窗口采集数据会造成时延，而且在实证中大都忽视了 EMD 的端点效应，从而影响市场预测的实效性。本文提出了一种前向滑动的滑动窗口 EMD 技术，分解后得到很多具有重叠信息的 IMF 序列，必须对冗余的数据降维。本文提出用 PCA 算法降维，将降维后的几个主成分输入神经网络以实现组合预测。本文提出了一种将 EMD、PCA、ANN 的组合预测模型，基于分解－择优－综合的信息融合思想构建了一个新的多尺度组合预测模型。在该模型的基本框架下分析沪深 300 指数和澳大利亚股指的波动特征及其走势。

本文提出一类具有一般性的金融市场预测模型，称作 FEPA，由三部分组成：第一部分是基于 EMD 的信号分析，第二部分是通过 PCA 的特征提取，第三部分是基于 ANN 的信息融合预测建模。本文所报告的工作有四个创新点：1）在学术文献上首次给出 EMD 算法的结构化表达，这种表达一方面明确了 EMD 算法的具体计算流程，另一方面提供了融合更多信息的算法接口；2）通过深入到更细时间框架来克服 EMD 的最大缺陷，即端点效应，并且探索两个时间框架下预测效能；3）通过多步长预测来深入研究 EMD 的信息结构；4）FEPA 模型展示了金融时间序列预测模型的一般性架构，具有可升级性与可扩展性。另外，采用滑动 EMD 分解使得实证更切近实际。为了便于阐明这些创新的具体内容和含义，我们需要综述一下基于 EMD 的金融预测研究的文献，并给出 EMD 结构化算法和整个预测模型的理论结构。然后我们再给出实证与分析，并得出结论。

2 金融时间序列经验模式分解的结构化算法

在 Huang 等[3]首创了一般时间序列数据的经验模式分解（EMD）算法之后，后来至今的所有的扩展和应用性研究文献上在介绍 EMD 算法时都沿用了 Huang 等人的流水瀑布式（waterfall）的算法表述，我们这里将其称为 HuangEMD 算法（用类 Matlab 语言表达）具体如下：

HuangEMD 算法：

{对于任一给定的实数时间序列数据 x(t)，首先找到其上全部的局部极值点（extrema），包括极大值点（maxima）与极小值点（minima）；

（1）分别用三次样条插值法将极大值点和极小值点连接成上包络线 $u(t)$ 与下包络线 $l(t)$；

（2）找到上下包络线的均线 $m(t)=(u(t)+l(t))/2$；

（3）取其数据与均线的差 $h(t)=x(t)-m(t)$；这个从（1）～（4）的过程称为一个筛滤过程（sifting）。

（4）检查看 $h(t)$ 是否满足本征模态函数（IMF）的定义并且是否满足筛滤停止标准；

（5）如果（4）的两个条件有一个不满足，就令 $x(t)=h(t)$；然后重复筛滤过程（1）～（4）；

（6）如果（4）的两个条件都满足，这是 $h(t)$ 就是一个 IMF 成分，并保存为 $c(t)=h(t)$，并且得到余差 $r(t)=x(t)-c(t)$。

（7）令 $x(t)=r(t)$，重复（1）～（6）步整个过程。

（8）当余差 $r(t)$ 中所包含的极值点个数小于 2 时，整个运算终止。

}

注意到 Huang 等[3]提出的 EMD 作为一种信号分解方法，不是一种参数化的解析方法，而是一种经验性的、非参数的分解算法，因而也只能通过一个算法来表达。这个算法中涉及到两个关键准则，一个是 IMF 的定义，另一个是筛滤过程的停止准则。关于这两点，我们下面再详述。

我们看到，文献中大家都直接引用的这个 HuangEMD 算法是典型的流水瀑布式的表达，看起来简单好算，实际上许多细节并不清楚，也不便于扩展。下面我们将这个算法转换成一种结构化的形式算法，由三个函数组成，PanEMD 作为主算法调用单级分解过程 PanEMD1，而后者又调用单一筛滤

过程 PanEMD0。具体如下:

函数 PanEMD0, 输入 x(t),输出筛选细节 h(t) 及极值点 maxs 与 mins,即

[h(t),maxs, mins, zeros]=PanEMD0(x(t))

{

(1)检测 $x(t)$的全部极值点与过零点

maxs=Maxima(x(t));

mins=Minima(x(t));

zeros=ZeroCrossing(x(t));

这里 Maxima()与 Minima()是两个具体的函数,检测 $x(t)$的极大值点与极小值点;ZeroCrossing( ) 检测 x(t)的过零点,细节不表。

(2)用这些极值点形成上下包络线

u(t)=Interpolation(maxs, 'spline');

l(t)=Interpolation(mins, 'spline');;

注意到 Interpolation( )作为一个插值函数,可以采用多种插值方式,'spline'表示三次样条。

(1)计算上下包络线的均线

$m(t)=(u(t)+l(t))/2$;

(2)计算细节时间序列(details)

$h(t)=x(t)-m(t)$;

}

这个函数 PanEMD0( ) 输出的 h(t) 仅仅是一种包络中的细节序列,至于是否满足 IMF 的定义还未可知。

函数 PanEMD1, 输入 x(t),输出 c(t),即 [c(t), nextrema]=PanEMD1(x(t))(其中 c(t)是一个标准的 IMF, nextrema 表示极值点的个数)

{令 k 表示筛选迭代的次数。

(1)运行筛选过程一次,计算细节函数

k=1;

$[h(t),maxs, mins, zeros]_k$=PanEMD0(x(t));

(2)如果$[h(t),maxs, mins, zeros]_k$不满足 IMF 的定义或者不满足筛选停止的条件, 就重复筛选:

while (～(IsIMF(h(t),maxs, mins,zeros) ||～StopSifting(h(t),maxs, mins, zeros, k))

repeat

[h(t),maxs, mins, zeros]=PanEMD0(h(t)) ;

(3)c(t) = h(t);

nextrema = #maxs + #mins;

}

这个函数 PanEMD1( ) , 对于输入 x(t),输出 c(t) 已经是一个标准的 IMF 了,这里我们之所以命名为 PanEMD1 是因为这个函数产生一个单一级别的 IMF,当然,PanEMD1( ) 中需要调用两个标准检查的函数如下。

函数 IsIMF(h(t),maxs, mins, zeros)输出 true 或 false

{if |#maxs + #mins － #zeros|< 2

return true;

elsereturn false;

}

这个函数判断如果极值点的个数与过零点的个数相等或不超过 1,细节序列 h(t) 就符合 IMF 的定义了。

然而,单一级别的 IMF 一般还不能满足我们对于金融时间序列进行信号分解和特征提取的需要,所以,一般地,我们还进行多次迭代,产生多级 IMF 序列。另外,停止筛滤的判断标准如下:

函数 StopSifting(h(t),maxs, mins,zeros, k)输出 true 或 false

{令 envmean 与 envamp 表示包络的均值和幅度,

envmean(t)=|u(t)+l(t))|/2;

envamp(t) = |u(t)－ l(t)|/2;

if (any(envmean > T1) || mean(envmean/envamp > T2) >T3)

&& all (#maxs+ #mins >2) && k < T4)

return false;

else

return true;

}

这里,T1, T2, T3 与 T4 是四个提前给定的阈值:首先 T4 是最多迭代的次数,一般地 T4<10; T1 表示包络均值的最小值,T2 表示包络均值与包络幅度的比例的最小值,T3 表示包络均值与包络幅度比例小于 T2 的百分比的最小值,一般地取 T1 = 0.05, T2=0.5, T3 = 0.05。关于筛选停止的条件,在 EMD 的实施中还是有不同的方案。这里我们采用的是 Rilling 等[17]所提出的方案。同时,这里 any( ) 是指只要有其中一个点满足设定的条件就是 true, 而 all( ) 则是要求所有的点都满足设定的条件才是 true. 另外,在 mean( ) 中,若条件满足就是 true =1,否则 false = 0,所以所有点上的条件判断可以求均值。

在明确给出了筛选函数 PanEMD0( ) 和单级 IMF 函数 PanEMD1( ) 之后,我们就可以给出

EMD 分解的主函数 PanEMD( )。令 K 表示 IMF 级数的上限：

函数 PanEMD，输入 x(t)，输出 C(t)，即

C(t)=PanEMD(x(t),K)

{令 k 表示分解迭代的次数。

(1)运行筛选过程一次，计算细节函数

k=1;

$[c_1(t), nextrema]$=PanEMD1(x(t));

$r_1(t) = x(t)-c_1(t)$;

(2)while (k <K && nextrema >T)

repeat

$\{[c_{k+1}(t), nextrema]$=PanEMD1$(r_k(t))$;

$r_{k+1}(t)=r_k(t)-c_k+1(t)$;

k=k+1;}

(3)$C(t) =\{c_1(t), c_2(t), \cdots, c_k(t)\}$;

}

这里 K 和 T 是两个阈值，一般取 K=10，T=2。

上面给出的 EMD 分级的结构化算法 PanEMD( ) 由单级 IMF 分解的 PanEMD1( ) 的多次迭代而成，而 PanEMD1( ) 又由单次包络算法 PanEMD0( ) 的多次迭代而成。这种两层嵌套的多次迭代过程揭示了 EMD 的完整实施并不是一个简单直观的过程。因而这种结构化算法为 EMD 的进一步具体化或升级变化提供了可扩展的算法架构。目前已经有的扩展包括面向更高信噪比与可靠性的 EMD 集成 EEMD 与更完备的 CEEMD。另外，对于金融时间序列(FTS)的特殊结构而言，在定义 Maxima( ) 与 Minima( ) 两个函数时应该分别使用当期的最高价与最低价。这些进一步扩展的内容目前我们超出本文的范围。本文所报告的模型计算都是基于收盘价的，这样便于公平地同其它基于收盘价的参考模型比较(见后面第 5.5 小节)。

3 FEPA 预测模型的总体结构与计算流程

有了 EMD 的结构化算法，我们就可以在此基础上构建一个完整的金融时间序列预测模型，这里我们简称 FEPA 模型，表示 FTS－EMD＋PCA＋ANN 模型，即针对金融时间序列(FTS, Financial Time Series)经过经验模态分解(EMD)，再经过主成分分析(PCA)降维，提取特征作为输入经过一个神经网络(ANN)预测出下一个时间点的价格变化。所以说，这个 FEPA 模型由三部分组成：EMD、PCA、ANN。

3.1 FEPA 预测模型的总体结构

FEPA 预测模型本身是一种计算智能模型，模型需要不断地根据市场演化而不断学习进化。一般地，我们取一段足够长的历史数据。首先要确定数据的时间框架，在本文中我们重点研究日数据。任一个时间 t 在日数据上对应于一天。价格时间序列数据 $X(t)$ 包括四个价格分量和一个交易量：

$$X(t) = (X.O(t), X.H(t), X.L(t), X.C(t), X.V(t)) \quad (1)$$

其中 $X.O(t)$、$X.H(t)$、$X.L(t)$、$X.C(t)$ 分别代表市场在这个时间框架下在 t 时间(区间)的开盘价、最高价、最低价、收盘价；$X.V(t)$ 代表交易量。对于日线级别，这四个价格就代表了当日的开盘价、最高价、最低价、收盘价。一般而言，我们假设有足够的历史数据

$$DX(t,N) = (X(t-N+1), X(t-N+2), \cdots, X(t)) \quad (2)$$

在日线级别，t 代表我们得到最新数据的那一天，N 代表我们得到的全部数据的天数。在本文中，我们只考虑使用收盘价 $X.C(t)$ 时间序列的预测问题，而且无论是预测模型的输入和输出只与收盘价有关。因而，在后面的论述中，我们用 $X(t)$ 只代表 $X.C(t)$。本文的后续论文中会发表更全面的预测模型。

对于历史数据 $DX(t,N)$ 中的每一个数据点 $X(t)$，我们都可以计算其当日的相对收益率

$$R(t,\tau) = \frac{X(t)-X(t-\tau)}{X(t-\tau)} \quad (3)$$

其中 τ 是要预测的步长，最基本的预测对应于 $\tau = 1$。若无其它说明，一般我们用 $R(t) = R(t,\tau)$。因而针对所有历史数据 $DX(t,N)$，也有对应的历史收益率数据

$$DR(t,N) = (R(t-N+1), R(t-N+2), \cdots, R(t)) \quad (4)$$

所以，一般的 FEPA 预测模型的输入输出可以表达为：

$$FEPA: \quad F(t) \Rightarrow ANN(S,\theta) \Rightarrow R(t+\tau) \quad (5)$$

也可以用数学上函数的形式

$$FEPA: \quad R(t+\tau) = ANN(F(t), S, \theta) \quad (6)$$

其中 $F(t)$ 是从 t 时间起前面的价格时间序列中提取出来的特征信息集(一个一维数组)，作为预测模型输入；(S, θ) 是预测模型的结构和参数(神经网络的神经元之间的连接结构与连接权重参数等)；$R(t+\tau)$ 是预测模型的输出。注意到 FEPA 模型使用 EMD 和 PCA 来生成输入特征 $F(t)$，所以，FEPA 预测模型可以进一步具体化为：

$$FEPA:\ DX(t,m)\Rightarrow EMD\Rightarrow PCA\Rightarrow F(t)\Rightarrow ANN(S,\theta)\Rightarrow R(t+\tau) \tag{7}$$

或者用数学函数的形式表示为:

$$R(t+\tau)=ANN(FE(PCA(EMD(DX(t,m)))),S,\theta) \tag{8}$$

其中 $FE(\ )$ 表示在 PCA 之后还有一个特征提取的过程(FE, feature extraction)。这里 $m<<N$ 表示截取 t 时间前面原始时间序列数据的窗口宽度。也就是说,对于任意时间点 t 而言,我们截取前面长度为 m 的时间序列数据段 $DX(t,m)$ 作为 EMD 算法的输入。

下面我们逐一说明 FEPA 模型的三个组成部分及计算流程。

3.2　前向滑动 EMD 分解

在 FEPA 模型中,我们要首先准备选定的全部历史训练数据集。在应用 PCA 之前,我们要针对全部历史数据 $DX(i-1,N)$ 形成一个 EMD 分解产生 IMF 作为输入与 $R(i)$,$i=t,t-1,\cdots,t-N+1$,作为输出,因而,预测模式训练输入输出数据集 $DT(t,N-m)$:

$$DT(t,N-m)=\begin{Bmatrix} EMD(DX(t-1,m) & \rightarrow & R(t) \\ \vdots & \vdots & \vdots \\ EMD(DX(t-N-m,m) & \rightarrow & R(t-N+1-m) \end{Bmatrix} \tag{9}$$

注意到这个训练数据集中针对每一个时间点 $i=t,t-1,\cdots,t-N+1$,都是截取了该时间点起前面长度为 $m<<N$ 的一个数据段,这里形象地称为'前向滑动'时间窗口。

经过审慎观察我们得知,EMD 分解的 IMF 会随着时间序列新数据的加入而不断变化,这就是 EMD 的内在缺陷,称作"端点效应"(end effect)。需要指出的是,在迄今发表的基于 EMD 的预测模型文献中,绝大多数在实证时都没有考虑 EMD 的这种端点效应,而直接对选取的全部历史数据 $DX(t,N)$ 进行一次性 EMD 分解,然后从分解的 IMF 中提取特征作为输入经过 ANN 进行预测;而这种实证方式是不切实际的。

在本文所提出的 FEPA 模型在实证时使用公式(9)所表达的前向滑动 EMD 分解,这样实证结果更切近实际。同时,我们认识到,只有在靠近端点时深入到下面更细一级的时间框架才能克服 EMD 的端点效应。对于当前时间框架为日线而言,更细一级的时间框架可以取 H1(小时线)或 M15(15 分钟线)。对于外汇和国际期货市场而言,由于市场是每天 24 小时连续交易,可以取 H1;而对于股指而言,每天股市只开 4 个小时,可以取 M15,这样一天就有 16 个 M15 数据点。

给定一个时间框架,如日线,注意到在 EMD 分解之后产生的是多层的 IMF:

$$DX(t-1,m)\Rightarrow EMD\Rightarrow\begin{Bmatrix} IMF_1(t-1,m) \\ \vdots \\ IMF_k(t-1,m) \end{Bmatrix} \tag{10}$$

一般地,我们令 $k\leqslant 5$。就将这些 IMF 串联起来与 $R(t)$ 构成一个输入－输出模式:

$$(IMF_1(t-1,m),IMF_2(t-1,m),\cdots,IMF_k(t-1,m))\rightarrow R(t) \tag{11}$$

这样训练数据集(9)中的每一数据点(每一行)都取(11)的形式,这样(9)式具体化为:

$$DT(t,N-m)=\{\boldsymbol{D}\rightarrow\boldsymbol{R}\} \tag{12}$$

$$\boldsymbol{D}=\begin{Bmatrix} IMF_1(t-1,m) & IMF_2(t-1,m) & \cdots & IMF_k(t-1,m) \\ IMF_1(t-2,m) & IMF_2(t-2,m) & \cdots & IMF_k(t-2,m) \\ IMF_1(t-N+m,m) & IMF_2(t-N+m,m) & \cdots & IMF_k(t-N+m,m) \end{Bmatrix} \tag{13}$$

$$\boldsymbol{R}=(R(t)\quad R(t-1)\quad\cdots\quad R(t-N+m+1))' \tag{14}$$

注意到式(13)中的数据矩阵 D 中的每一项 IMF 都是一个 IMF 时间序列,因而在列的方向(每一行)都是一个高维数组,其中必然包含冗余信息。为了对 IMF 数据矩阵 D 进行降维处理,我们使用主成分分析(PCA)。

3.3　主成分分析(PCA)

在式(12)中,由 $N-m$ 个时间点上,每个点上 k 层长度为 m 的 IMF 时间序列所构成的数据集 D 是一个 $n\times p$ 矩阵,其中 $n=N-m$,$p=k\times m$。为了便于下面的讨论,令

$$\boldsymbol{D}=\begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{np} \end{pmatrix} \tag{15}$$

矩阵 $\boldsymbol{D}$ 的列代表了预测模型输入信息，维度为 p；行代表这些输入信息的观测取样。注意到这个维数 p 可以很大。比如若 EMD 变换的窗口宽度为 300，若仅仅只取 3 层 IMF，就是 $p=k\times m=3\times 300=900$，相当于 900 个变量的动态复杂系统。如此之多个自由度的复杂系统，若在日线数据上，至少需要 10 倍的数据点，即 9000 个日数据，每年只有 250 个交易日，就相当于需要 36 年的日数据。而中国股市一共只有约 25 年的历史，且股市体制从 2005 年才开始正常化，因而只有约 12 年的可用日线数据。显然，我们认为这些输入信息（列）必然包含有冗余信息，因而，我们希望能有一种变换将这 p 维输入信息约简到一个更小的的维度 $q\ll p$。在经典统计学和线性代数中有一个这样的变换存在，即主成分分析（PCA，Principal Component Analysis），也叫主分量变换。

PCA 是一种在数据挖掘领域被广泛使用的降维技术。PCA 试图减少数据集的维数，将数据映射到低维空间，同时保持数据集中的对方差贡献最大的特征，尽可能保留隐藏在数据中的固有信息。PCA 作为一个强有力的技术用于提取高维度数据集的结构特征，在许多领域获得广泛的应用。

PCA 计算的原理是建立在数据矩阵 $\boldsymbol{D}$ 的归一化 $norm(\boldsymbol{D})$ 之上的：

$$\boldsymbol{Z}=\begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ z_{p1} & z_{p2} & \cdots & z_{pn} \end{pmatrix}=(norm(\boldsymbol{D}))^T \tag{16}$$

其中

$$z_{it}=\frac{(d_{ti}-\bar{d}_i)}{\sigma_i},\quad \bar{d}_i=\frac{1}{n}\sum_{t=1}^{n}d_{ti},\quad \sigma_i=\sqrt{\frac{\sum_{t=1}^{n}(d_{ti}-\bar{d}_i)^2}{n}} \tag{17}$$

然后就可得到矩阵 Z 的奇异值分解为：

$$\boldsymbol{Z}=\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^T \tag{18}$$

其中 $p\times p$ $\boldsymbol{W}$ 是 $\boldsymbol{Z}\boldsymbol{Z}^T$ 的特征向量矩阵，是正交矩阵；$p\times n$ $\boldsymbol{\Sigma}$ 是非负矩形对角矩阵，其内部左边 $p\times p$ 子矩阵就是由 $\boldsymbol{Z}\boldsymbol{Z}^T$ 的 p 个特征值作为对角元素组成的对角矩阵；$n\times n$ $\boldsymbol{V}$ 是 $\boldsymbol{Z}^T\boldsymbol{Z}$ 的特征向量矩阵。

据此，有：

$$\boldsymbol{Y}^T=\boldsymbol{Z}^T\boldsymbol{W}=(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T\boldsymbol{W}=\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{V}\boldsymbol{\Sigma}^T \tag{19}$$

当 $p<n-1$ 时，$\boldsymbol{V}$ 通常是不唯一的，而 $\boldsymbol{Y}$ 是唯一定义的。$\boldsymbol{Y}^T$ 的第一列由第一主成分组成，第二列由第二主成分组成，以此类推。

在实际应用中，一般地，数据矩阵 Z 的信息量主要集中在前面若干个特征维上。这样我们就可以取 p 个特征值中的前 $q\ll p$ 个，组成新的非负矩形对角矩阵：

$$q\times n\ \boldsymbol{\Sigma}^*\ \boldsymbol{\Sigma}^*=\boldsymbol{I}_{q\times p}\boldsymbol{\Sigma} \tag{20}$$

其中 $\boldsymbol{I}_{q\times p}$ 是 $q\times p$ 单位矩阵。

与 $\boldsymbol{\Sigma}^*$ 对应地，我们将特征向量矩阵 W 约简为只包含前 q 个特征矢量组成的低维矩阵 W^*，这样我们就可以对数据矩阵 Z 进行降维处理得到新的数据矩阵：

$$\boldsymbol{Y}^*=\boldsymbol{W}^{*T}\boldsymbol{Z}=\boldsymbol{\Sigma}^*\ \boldsymbol{V}^T \tag{21}$$

注意到，实际计算中，$\boldsymbol{Z}\boldsymbol{Z}^T$ 就是原始数据矩阵 $\boldsymbol{D}$ 的相关系数矩阵。关于 PCA 的具体计算过程，一般统计学软件如 SPSS 或 Matlab 中都有，就不赘述。这里需要特别说明的是如何确定新的维度 q。这里我们要求前 q 个主成分的累积贡献率 CCR（Cumulative Contribution Rate）必须大于某个预设的阈值，如 85%：

$$CCR_q=\left(\sum_{i=1}^{q}\lambda_i\right)/\left(\sum_{i=1}^{p}\lambda_i\right)>85\% \tag{22}$$

在确定了 q 后，我们就可以得到降维后的数据矩阵 $\boldsymbol{Y}^{*T}$，取代原始数据矩阵 $\boldsymbol{D}$，作为训练数据的输入部分。

3.4 多层前馈神经网络及后向传播训练算法

上述 FEPA 预测模型最终归结到输入－输出的一般映射问题。对于一个从 n 维实空间（输入变量）$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$ 到 m 维实空间（输出变量）$\boldsymbol{z}=(z_1,z_2,\cdots,z_m)$ 的一般映射，可以看成一个一般的非线性函数：

$$\boldsymbol{z}=(z_1,z_2,\cdots z_m)=F(\boldsymbol{x})=F(x_1,x_2,\cdots x_n) \tag{23}$$

这个映射若通过一个含有 h 层隐含神经元的多层前馈神经网络（MFNN，Multilayer Feedforward Neural Network）来实现，就可以表达成下列数学计算结构：

$$\begin{cases} z_k = f(y_{h,1}, y_{h,2}, \cdots y_{h,n_h}) = \Psi(\sum_{i=1}^{n_h} w_{h+1,k,j} y_{h,i} + b_{h+1,k}), & k = 1,2,\cdots,m \\ y_{l,j} = g(y_{l-1,1}, y_{l-1,2}, \cdots y_{l-1,n_{h-1}}) = \Phi(\sum_{i=1}^{n_{l-1}} w_{l,j,i} y_{l-1,i} + b_{l,j}), & \begin{matrix} l = 1,2,\cdots,h \\ j = 1,2,\cdots,n_l \end{matrix} \\ y_{1,j} = g(x_1, x_2, \cdots x_n) = \Phi(\sum_{i=1}^{n} w_{1,j,i} x_i + b_{1,j}), & j = 1,2,\cdots,n_1 \end{cases} \tag{24}$$

其中 Ψ 和 Φ 是两个激发函数，一般应该是非线性的，如 S 型函数；$y_l = (y_{l,1}, y_{l,2}, \cdots, y_{l,n_l})$ 表示第 l 隐含层，共有 n_l 个隐含神经元；w 表示权重，b 代表偏差，w 和 b 的总体 $\{(\{w_{l,j,i}, i = 1,2,\cdots n_l\}, b_{l,j})\}$ 就是整个 MFNN 的参数集。这些参数的最优取值就是 MFNN 学习或训练的目标。当 $h = 1$ 时，MFNN 就约简为普通的只带一个隐含层的前馈神经网络（FNN），即：

$$FNN = MFNN(h = 1) \tag{25}$$

有必要说明的是，在深度（学习）神经网络（如智能围棋 AlphaGo）出现之前，这种多层前馈神经网络的结构本身就已经存在了，然而有三个原因导致了这种神经网络进一步深入研究的停滞：其一是由于 Hornik 等从数学上证明了只需要一个隐含层但包含足够多个神经元的前馈神经网络就能够逼近任何连续函数，从而人们觉得不一定非要使用更多的隐含层就能够解决问题；其二是若使用多层网络，系统复杂性自然就会增大，训练网络就需要更多的数据，若数据不够或训练算法不强就会导致模型系统的不稳定或落在局部最优；其三，由 Rumelhart 等所提出了反向误差传播（BP）算法针对浅层前馈网络的训练上还是足够有效，但对于多层前馈网络的训练上，若直接用，就会出现梯度消失或梯度爆炸问题。普通多层神经网络模型是高度非凸的，若直接用 BP 算法来训练，由于存在大量的局部最优点而且收敛性差，很难获得好的学习结果。实际应用时会出现严重的过度拟合现象。

在本文所报告的实证中，我们仍然使用的只包含一个隐含层的前馈神经网络（FNN），并且仍然使用反向误差传播（BP）算法来训练，由于目前 FEPA 模型仍然使用常规金融时间序列数据，还未用到大数据。今后，在扩展到多市场联立预测模型时将会用到包含多层深度神经网络，并且将会使用深度学习算法来训练这类深度神经网络模型。公式（24）给出了今后对接深度学习神经网络的一般形式。

4　历史数据及 EMD 分解图示

4.1　数据来源及说明

FEPA 预测模型的实证测试使用两个股票市场指数：中国沪深 300 指数 HS300 和澳大利亚股指（SP200）AXJO。沪深 300 指数是中国的基准股票指数。在世界股票市场生态系统中，美国股票市场具有最大影响力，其走势不大受其他国家股市的影响。除美国之外的其他发达国家（如 G7 国家和澳大利亚）的股票市场属于第二梯队，受美国股票市场的影响较大，具有更好的可预测性。正因如此，澳大利亚股指具有更好的风险可控性和预测的经济价值。本文的实证数据集样本是 2011 年 2 月 18 日至 2016 年 4 月 8 日的沪深 300 指数 HS300（图 1）和 2011 年 4 月 20 日至 2016 年 3 月 24 日的澳大利亚股指 AXJO，剔除节假日等因素影响，共有 1250 组数据。数据集分为两个子集：前 4 年的数据为样本内训练数据集，共 1000 个数据；最近一年的数据为样本外测试数据集，共 250 个数据。样本的时间跨度覆盖了许多重要的经济事件，因此我们认为这一时间跨度对训练模型是足够有代表性的。下面我们主要展示对于 HS300 预测模型的实证过程，并在最后表 5 中给出了 AXJO 的预测实证结果。

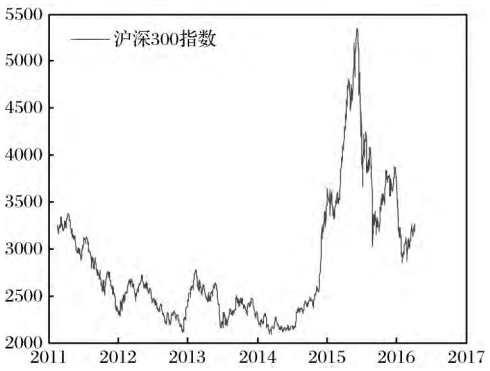

图 1　股指 HS300 收盘价

4.2 沪深 300 股指的 IMF 分量

根据 EMD 算法的定义，原始信号经 EMD 算法分解后得到的各 IMF 序列相互正交，即线性无关。然而，正如 Huang 等[3]所述，正交性取决于分解方法，EMD 算法是一个非线性的方法，这样保证了经 EMD 分解后的各 IMF 序列相互正交。按照前述的研究思路，首先对沪深 300 指数和澳大利亚股指进行 EMD 分解。图 2 为滑动窗口取 300 天时两个股票指数的 IMF 分量和残差图。EMD 分解的 IMF 序列层数越多，细节信号和近似信号平稳性就越好，预测值也越精确，但是同时分解过程本身存在计算误差，误差随着分解层数增加而增大，这会降低预测精度。所以选择适当的分解层数是非常重要的。从图中可以看出，所有 IMF 的频率和振幅都是随时间变化的，IMF 的振幅由高频至低频依次减小。IMF1 捕获了投资者情绪的高频波动特征，IMF5 反映了投资者情绪最低频率的波动特征，残差项 R 反映了投资者情绪的平均趋势部分。EMD 分解就是把一个非线性非平稳信号分解成若干个平稳的不同频率的分量以及一个趋势项的过程。

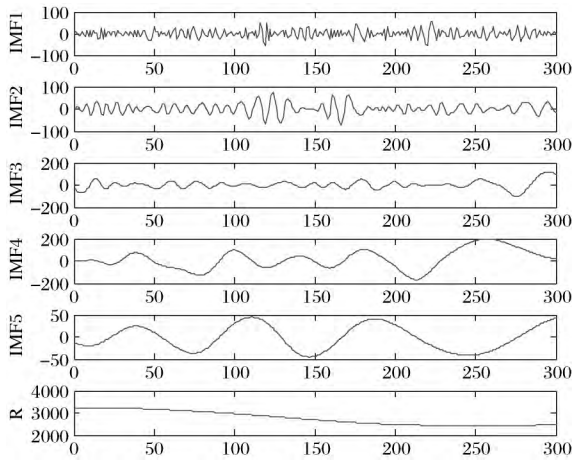

图 2 窗口长度为 300 天 HS300 收盘价的 IMF 分量

4.3 主成分分析后的特征值和累积贡献率

HS300 的 IMF 序列经主成分分析后的前 4 个特征值对应的主成分的累计贡献率就超过了 85%，而 AXJO 的前 5 个特征值对应的主成分的累计贡献率超过 85%。为了取得更好的预测效果，本文将提取两个市场的前 22 个主成分作为训练数据输入神经网络。

5 具体 FEPA 预测模型实证与讨论

对于上述 FEPA 预测模型的一般理论，我们实现了三个具体模型并进行了历史数据实证。下面我们首先给出具体 FEPA 预测模型的简明表达及预测效能测度，然后给出三个具体模型及实证结果。

5.1 具体 FEPA 预测模型的简明表达

一个具体的 FEPA 预测模型有三个关键结构参数：1)EMD 分解的窗口长度(后视长度)，即公式(8)中的 m；2)要预测的步长，即公式(3)中的 τ；3) PCA 变换后降维后所选取的主分量的个数，即公式((21)中的 q。这样，一个具体的 FEPA 预测模型可以由一般公式(8)简化为：

$$FEPA: R(t+\tau) = FNN\{PCA^*[EMD(DX(t, m)), q]\} \tag{26}$$

其中 $PCA^* = FE(PCA)$ 对应于 PCA 变换后特征提取(降维)，FNN 是只有一个隐含层的前馈神经网络。

5.2 具体 FEPA 预测模型的效能测度

为了测度预测模型的有效性，我们自 2004 年以来就一直着重使用预测方向正确性的命中率(HR: Hit Rate)作为股指预测的主要测度，考虑到命中率对于市场风险管理和投资交易策略具有第一重要性。命中率 HR 的定义如下：

$$HR = \frac{100}{n}\sum_{i=1}^{n} d_i(X_i, Y_i),\ d_i(X_i, Y_i) = \begin{cases} 1 & X_i Y_i \geqslant 0 \\ 0 & else \end{cases} \tag{27}$$

其中 X 和 Y 分别表示样本点上相对收益率的真实值和预测值；n 是样本点总数目。

当然我们也知道一般时间序列预测的其它效能指标，如平均绝对百分比误差(MAPE)，均方根误差(RMSE)，平均绝对误差(MAD)，用来测度实际值与预测值偏差。在本项研究中心，我们只使用命中率 HR，因为对于量化投资而言，预测模型的命中率 HR 是最重要的效能指标。只有在预测方向正确率有了显著性之后，其它效能指标才有意义。

5.3 具体模型 FEPA_HS300D1 预测沪深 300 股指日线收益率

这个预测模型针对沪深 300 股指日线收盘价数据，用日线数据的 EMD 分解然后经过 PCA 变换后降维处理作为输入，预测日线收益率，该模型可从公式(26)具体化为：

$$FEPA\_HS300D1:\ R(t+\tau) = FNN\{PCA^{*}[EMD(HS300D1\_DX(t,\ m)),\ q]\} \tag{28}$$

该模型在用历史数据训练后在样本外检验数据上的命中率在后视窗口长度 m 取不同值时的命中率见表 1,其中最好的命中率 78.57% 出现在 $m=300$。在 $m=300$ 条件下,预测未来不同步长 τ 上的命中率见表 2,其中最好的命中率出现在 $\tau=1$,即 $t+1$;而 5 步之后基本上就不具有可预测性了。

表 1　FEPA_HS300D1 预测 t+1 在不同后视窗口长度 m 上的命中率($q=22$)

后视长度 m	命中率 HR(%)
50	63.67
100	62.28
150	64.61
200	65.63
250	76.92
300	78.57
350	73.08
400	73.08

表 2　FEPA_HS300D1 预测未来不同步长 τ 上的命中率($q=22$)

未来步长 t+τ	命中率 HR(%)
t+1	78.57
t+2	67.86
t+3	60.16
t+4	56.12
t+5	51.03

5.4　具体模型 FEPA_HS300M15 预测沪深 300 股指 15 分钟线收益率

为了进一步消减 EMD 端点效应,除了使用滑动窗口 EMD 变换外,这个预测模型深入到更细一级的时间框架 M15,即 15 分钟线,考虑到中国股市每天开场时间只有 4 个小时,股指期货也仅开场 4 个半小时。(若针对 G8 股指期货,可以考虑使用 H1 时间框架(一小时线))。这个预测模型针对沪深 300 股指 M15 收盘价数据,用 M15 数据的 EMD 分解然后经过 PCA 变换后降维处理作为输入,预测 M15 收益率,该模型可从公式(26)具体化为:

$$FEPA\_HS300M15:\ R(t+\tau) = FNN\{PCA^{*}[EMD(HS300M15\_DX(t,\ m)),\ q]\} \tag{29}$$

该模型在用历史数据训练后在样本外检验数据上的命中率在后视窗口长度 m 取不同值时的命中率见表 3,其中最好的命中率 82.14% 出现在 $m=300$。在 $m=300$ 条件下,预测未来不同步长 τ 上的命中率见表 4,其中最好的命中率出现在 $\tau=1$,即 $t+1$;而 5 步之后仍然具有可预测性。这里针对 HS300 在 M15(15 分钟线上)预测命中率和多步长上的表现与 HS300 在 D1(日线)比较,可以看出,M15 可预测性更高,也可能解释为中国特有的 T+1 交易制度,使得日间市场比日内市场具有更好的效率;因而内日股票市场具有更好的可预测性。

表 3　FEPA_HS300M15 预测 t+1 在不同后视窗口长度 m 上的命中率($q=22$)

后视长度 m	命中率 HR(%)
50	64.75
100	62.72
150	66.29
200	67.19
250	73.08
300	82.14
350	71.79
400	71.79

表 4　FEPA_HS300M15 预测未来不同步长 τ 上的命中率($q=22$)

未来步长 t+τ	命中率 HR(%)
t+1	82.14
t+2	75.00
t+3	64.29
t+4	58.76
t+5	53.02

关于三种 FEPA 具体模型的参数取值:

(1)EMD 窗口长度:从 150 到 400,每 10 递增一次。三个模型经过实证都是 300 最优;

(2)主成分特征值个数:主成分特征值个数从 3 到 22,后来三个模型都用 22 个。

(3)神经网络隐含层神经元个数从 2 至 10 个遍历实证,最后由程序自动选择最优个数。

5.5　与其它现有模型的预测效能比较

为了比较本文所提出的 FEPA 预测模型同文

献中现有的代表性预测模型，我们也选择了五种常见的预测模型，做了软件实现，并用同一数据集（样本内和样本外）进行了历史数据实证，包括：

1）ARIMA – 自回归移动平均模型，

2）GARCH – 广义自回归条件异方差模型

3）BPNN – 经典单隐含层神经网络（用 BP 算法训练），

4）EMD－BPNN – 用 EMD 分解作为输入的 BPNN，

5）WD－BPNN – 用小波分解作为输入的 BPNN。

用这些模型得到的样本外预测命中率和 FEPA 的比较见表 5。这五种参考模型的参数取值都是在同样的训练数据集上各自优化确定，然后应用在后面的样本外检验数据集上。

这五种从这个表中可以看出，FEPA 模型对于无论中国股指 HS300 还是澳大利亚股指 AXJO 在日线上的预测命中率都是远远高于这些现有模型。当然不排除文献中还有其它更复杂的模型会有优异的表现，但我们无法掌握其软件实现。

表 5 FEPA 与其它预测模型的命中率实证比较

预测模型	HS300	AXJO
ARIMA	54.10	52.85
GARCH	53.65	53.58
BPNN	59.75	58.95
EMD－BPNN	65.78	66.80
WD－BPNN	63.49	64.13
FEPA	78.57	74.79

6 结语

本文给出了 FEPA 预测模型的理论基础、模型架构、与计算过程，并在中国股指 HS300 的日线（D1）和 15 分钟线（M15）和澳大利亚股指 AXJO 日线上作了全面的软件实现和历史数据检验，并同 5 种文献中现有的预测模型进行了比较。实证结果表明 FEPA 模型在股指日线上实现了 74％－78％的命中率，在 15 分钟线上达到 82％，这些结果本身就是明显远离了有效市场假设的命中率（50％上下），因而可以说中国股指和澳大利亚股指在日线上具有概率可预测性。当然这种概率可预测性具体如何用于股市投资中的风险管理和量化投资，还需要进入到主动投资组合理论和智能交易策略技术的研究之中。FEPA 模型同 5 种现有模型的实证比较结果也表明 FEPA 模型目前具有更好的预测能力。

从本文所提出的 FEPA 预测模型的理论创新上讲，这里首次给出了 EMD 分解的结构化算法，这一方面有利于精确理解现有的 EMD 的算法计算过程，也有利于今后进一步扩展到金融时间序列分解的更特殊或更全面的信息融合可能性，比如将最高价和最低价纳入到 EMD 分解当中以确定更有代表性的金融时间序列上下包络，将交易量和资金流纳入到 EMD 分解当中得出更可靠的信息，将相关市场和财经指标时间序列数据融合到目标市场价格时间序列 EMD 分解当中，实现多市场多数据流的 EMD 分解与信息融合。同时对于 EEMD（加载噪音的 EMD 分解），也可以扩展到加载随机游走噪音而不仅仅是白噪音。总而言之，EMD 分解的结构化为金融时间序列的更可靠的信号分解与更多源信息融合提供了一个算法架构。

考虑到 EMD 分解的端点效应，FEPA 模型在实证时使用了滑动窗口，这样更切近实际。同时我们在日线预测的同时还测试了更细时间框架（15 分钟线）的 EMD 分解。这事实上就是应对 EMD 端点效应的主要措施。当然，如何将多个时间框架的 EMD 分解融合起来建立统一的预测模型，这里也仅仅是一个开端，今后还有更大的研究空间。

在 EMD 分解后，FEPA 模型用 PCA 来降维，已经表现出明显的效果。但是，我们也知道 PCA 是一种线形变换，对于非线性非稳态金融时间序列而言，我们仍然需要探索非线性的降维方法。目前一种可能有效的非线性降维方法就是深度学习中的‘自编码器’，属于深度学习多层前馈神经网络。另外一种可能会很有效的非线性预测模型是从决策树到模糊决策树发展而来的‘随机森林’。所以，狭义的 FEPA 模型中的‘A’代表‘ANN’，而广义的 FEPA 模型中的‘A’代表‘AI’。这个 FEPA 模型构成了深度智能投资理论的重要组成部分[18]。

参考文献：

[1] Pan Heping, Sornette D, Kortanek K. Intelligent finance－An emerging direction [J]. Quantitative Finance, 2006,6(4):273－277.

[2] Pan Heping. A basic theory of intelligent finance [J]. New Mathematics and Natural Computation, 2011, 7(2):197－227.

[3] Huang N E, Shen Zheng, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for

nonlinear and nonstationary time series analysis [J]. Proceedings: Mathematical Physical and Engineering Sciences, 1998,454(1971): 903－995.

[4] Yu Lean, Wang Shouyang, Lai K K. Forecasting crude oil price with an EMD－based neural network ensemble learning paradigm [J]. Energy Economics,2008,30(5): 2623－2635.

[5] 周德群,鞠可一,周鹏,等. 石油价格波动预警分级机制研究[J]. 系统工程理论与实践,2013,33(3):585－592.

[6] Lin C S, Chiu S H, Lin T Y. Empirical mode decomposition－based least squares support vector regression for foreign exchange rate forecasting [J]. Economic Modelling, 2012,29(6):2583－2590.

[7] 王书平,胡爱梅,吴振信. 基于多尺度组合模型的铜价预测研究[J]. 中国管理科学,2014,22(8):21－28.

[8] 阮连法,包洪洁. 基于经验模态分解的房价周期波动实证分析[J]. 中国管理科学,2012,20(3):41－46.

[9] 张秀艳,徐立本. 基于神经网络集成系统的股市预测模型[J]. 系统工程理论与实践,2003, 9(9):67－70.

[10] 秦宇. 应用经验模态分解的上海股票市场价格趋势分解及周期性分析[J]. 中国管理科学,2008,16(S1):220－225.

[11] 王文波,费浦升,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] 张承钊,潘和平. 基于前向滚动 EMD 技术的预测模型[J]. 技术经济, 2015,34(5):70－76.

[13] Dragomiretskiy K, Zosso D. Variational mode decomposition[J]. IEEE Transactions on Signal Processing, 2014,62, 531－544 .

[14] Wei Liangying. A hybrid ANFIS model based on empirical mode decomposition for stock time series forecasting [J]. Applied Soft Computing, 2016,42:368－376.

[15] Lahmiri S. A variational mode decomposition approach for analysis and forecasting of economic and financial time series [J]. Expert Systems with Applications, 2016,55:268－273.

[16] Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning [C]//Proceedings of the 7th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, 1994.

[17] Rilling G, Flandrin P, Goncalves P. On empirical mode decomposition and its applications [C]//Proceedings of IEEE－EURASIP Workshop on Nonlinear Signal and Image Processing. 2003,NSIP－03, Grado.

[18] 潘和平. 深度智能投资－智能投资组合理论、强势投资法与深度学习智能交易策略的统一[C]//第四届全国金融大数据战略与应用研讨会特邀报告,中科院大学主办, 2016 年 10 月 9－11 日,北戴河.

FEPA:An Adaptive Integrated Prediction Model of Financial Time Series

PAN He-ping[1,4], ZHANG Cheng-zhao[2,3]

(1. Chongqing Institute of Finance,Chongqing 400067, China; 2. School of Economics and Management, University of Electronic Science and Technology,Chengdu 611731, China; 3. Chengdu Polytechnic,Chengdu 610041, China; 4. Swingtum Prediction,Australia)

Abstract: In this paper, an adaptive model is documented for predicting financial time series integrating signal processing, information fusion and computational intelligence. The model consists of financial time series (FTS)－specific Empirical Mode Decomposition (EMD) for signal processing, Principal Component Analysis (PCA) for dimension reduction, and Artificial Neural Networks (ANN) for nonlinear prediction. The model uses a sliding window to capture the most recent time series data, applies EMD to transform the data into multilevel Intrinsic Mode Functions (IMF's). PCA is then used to reduce the dimension of IMF's and to generate a set of information－rich features which are input into an ANN to generate the output as prediction. This novel model of prediction implements an information fusion process consisting of signal decomposition, dimension reduction and nonlinear synthesis. This model lifts the prediction capability to a new level. The originality of this work exhibits in fouraspects: 1) a structural reformulation of EMD algorithm, providing an interface to more information fusion, 2)deepening into finer time frames for tackling the end effect of EMD andimplementation and testing on two levels of time frameimplementation,3) investigation of multi－step prediction, 4) a generic framework of prediction models for financial time series with upgradability and extensibility. The use of sliding window for EMD also gets the test closer to the re-

ality. The new model is tested on the historical data of two stock indices – Chinese HS300 and Australian AORD, the performance, achieving a hit rate of 78% and 82% on HS300 D1 and M15, and 74% on AXJO D1 respectively, significantly higher than 5 existing models after comparison.

Key words: financial time series prediction; Empirical Mode Decomposition (EMD); Intrinsic Mode Function (IMF); Principal Component Analysis (PCA); Artificial Neural Networks (ANN)