金融科技赋能投研系列之十三：

Autoencoder 与金融数据应用

摘要：

AI 技术近年来发展迅速，并且已经深入扩散到各个真实世界的应用领域，比如自然语言分析、图像识别等。这些成功的应用为我们将新方法移植到金融数据分析领域提供了学习的案例。

同时，我们也看到 AI 技术依然属于前沿领域，其推进方向并不仅限于成功实例。它发展的巨大潜力蕴含在大量的专业研究领域，特别是在新应用场景,面对更加具体和复杂的问题，提出进一步的模型方法和应用模式。这为我们研究低信噪比的金融数据,尝试更多新方法提供了内容丰富的土壤。

本文，我们将把自编码器方法(autoencoder)应用到金融数据分析的层面上，寻找金融时间序列数据中的多个活跃影响维度的表现形式，并且与已知的一些模型方法进行对比,从而确定方法的有效性,为无监督学习模型的深入应用提供依据。

投资咨询业务资格：

证监许可【2011】1289 号

研究院 量化组

研究员

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

镇谌博

☎ 0755-23887993

✉ zhenchenbo@htfc.com

从业资格号：F3080231

一、背景介绍

机器学习是近年来算法领域的重要发展，目前已经成为最核心的科技研发领域。其研究的范围迅速扩大，从实验性质的算法研发到应用层级的技术开发，都获得了大量的资金和人才投入。现在，我们有大量的技术性文献和案例分析，但是在金融领域鲜有系统性的应用指导。而前沿技术也有其自身的发展模式--技术迭代频率高：大量实验性技术短时间内迅速被更先进的技术取代。所以，我们的研究将会着眼于更为深入的理论探讨，结合创新技术的应用，最终落到金融投研领域，而非时髦技术的简单测试。

从系统性研究的角度来说，金融投研有其独特的复杂性.本质上金融市场是一个开放性系统，持续受到系统内外因素的影响。而内外影响的因素的强弱又具有不连贯的特点，一定的市场条件下，投资工具的内在价值是主要定价逻辑，另一些时段，又可能出现外在政策甚至国际政治风险等成为强烈的干预因素。即使是在市场内部，也会因为市场整体情绪或是交易者投资逻辑改变，出现一定时间范围内的风格转换以及随之而来的标的物价格波动。

因此，对于数据的理解，比如不同周期尺度上的主导因素（及干扰因素）；或是不同因素影响强弱的交替,都是量化方法透过数据理解市场的重要角度。转换成金融研究的语言，市场风格因子的切换，一段时间内的价格驱动力是金融数据最为核心的研究内容，也是金融投研的前置性工作。

然而，传统的数据分析方法具有相当严重的局限性。首先，数据的增量应用模式很难扩展。举例来说，传统投研广泛使用的多因子模型，会因为因子个数的迅速增加，而面临维数诅咒（Curse of Dimensionality）。其次，金融领域存在大量的非线性现象，而这却是传统数据分析工具的弱项。实际上，传统的金工模型需要在较严苛的约束条件下,才能较有效提取数据信息，从而具备预测能力。遗憾的是在真实世界中，这些约束条件几乎都难以满足，甚至有时候直接违反建模条件。这也是业内经常看到的经验模型严重依赖近期历史数据，预测则容易失效的现象。

与此对应，AI 技术对上述类型的问题，则具备更优异的处理能力。首先，目前流行的 AI 模型都具有较高的复杂度，模型本身就具备了处理复杂问题的能力。大部分 AI 模型同时还具备处理大量数据的能力（在硬件环境支持条件下），并且可以通过技术性手段更好的规避因子过多和多重共线性等金融数据里面常见的建模问题。再者，AI 模型目前发展出了相当多样的形态。比如，一般归为浅度学习的决策树、随机森林和 XGBoost 等模型；或归为深度学习的神经网络、卷积网络和 LSTM 等模型。这些模型都能够在各自适用的范围内做到充分解决非线性问题，并且都在近期的应用中有突破性进展。我们在之前的研究中已经涉及到了浅度学习和深度学习方法的应用. 有趣的是,它们都有各自适用的应用场景，在投研中互补性很强，其整合应用将成为我们未来的主要研究方向[1].

另外,AI 模型还可以通过学习模式分类为**监督学习**和**无监督学习**。这里的区别在于，监督学习是从带有标记(label)的训练数据中推导出预测函数。也就是说,训练实例数据中已经包含了输入数据以及期望的输出数据，模型训练的目的是模式识别，描述从输入数据到输出数据的映射关系。无监督学习则是从无标记的训练数据中推断出结论. 也就是说，训练实例数据中包含的隐藏信息/数据结构,需要模型自己去寻找，验证其合理性。

可以看出来，监督学习和无监督学习在金融领域都有各自的应用场景. 比如,上文提到的随机森林和深度神经网络都可以构建成监督学习模式. 通过预设标记(比如未来收益率或波动率等)，利用大量数据训练模型. 而无监督学习则具有更为深远的应用,简言之,它可以发现金融数据中我们还不知道的规律。从市场的交易数据出发，并没有一个统一的模型能够精确预测行情走势，这与金融市场是一个开放性系统，且受很多外界(突发)影响分不开的。但是,即使只是在金融市场内部,也会因为不同的交易者结构,不同的交易目的，进而出现不同的交易行为,导致我们观察到的貌似随机的交易现象。而这些不断演化的价格驱动力一般来说是较难及时发现的,大多数情况下,市场结论都是事后分析的结果。

那么是否可以利用无监督学习方法,即在不预设标记的前提下分析数据的特征,让模型自己找到隐含的价格变化模式呢？这是一个较难的问题。但我们可以将它和我们之前建立的投研模式联系起来[2]。在多周期数据分析系列，我们已经将金融数据分解到了不同的周期尺度；同时利用随机森林模型，我们发现在不同周期上，价格的驱动因素可以很不相同，但是重要影响因子的数量并不多（统计显著的意义上）。所以，确定不同周期尺度上的驱动因素，或者说活跃的影响维度是一个重要的模型参数。

本文将结合近期的自编码器（autoencoder）研究成果[3]，分析如何利用无监督学习方法，寻找影响标的物价格的主导因素维度，从而帮助我们从市场动力学的角度确定有效因子数据，帮助我们确定有效模型的数据维度（如多因子模型建模，深度学习中的数据切片维度等）。进一步，我们会和其他非 AI 类型模型结果进行对比以验证方法的有效性。

二、方法论

1. 研究背景

首先，我们在前期的报告中提出了多周期数据分析的概念—把金融数据（如投资标的物收益率）在不同周期层次上进行分解，并寻找在不同周期上该金融数据的变化规律。进一步，分析相关因素的价格影响力大小，或影响力随时间的变化规律，从而为投研模型提供定量支撑。

而在具体某个周期尺度之上，如何理解标的物收益率的市场运动规律呢？我们假设在一定周期范围内，标的物有 N 个核心影响维度，它们对应与经济学层面上，有 N 个具备定价

影响力的因子。那么价格外在表现的波动性等关键特征就是由其背后的若干定价因素共同决定。当然，在不同的经济运行阶段，或市场行情阶段，定价因素的相对强弱也会发生烈度不等的变化，表现出来的影响因素维度可能会不尽相同。而最突出的特征表现为，某标的物价格波动规律在结构性相变时段与市场平稳阶段完全不同，导致其价格变化的类周期规律缺失（无历史复现）。所以，我们将会尽量选取数据量较长的数据进行测试。

通常的研究方法，多为来自于自然科学领域，比如生态学，大气学，物理学等。正如从已经研究过的场景中，我们认为这些方法具有很强的一般性，并且完全可以移植到金融领域，对我们认识金融数据中的非线性规律提供了更先进的数据处理工具。

这里的关键问题在于如何通过挖掘时序数据中的特征确定其运动规律的空间维度，而这一维度与我们通常在定价层面考虑的多因子定价因素存在数量上的对应关系。我们看到研究领域已经有了一些研究方法；同时 AI 技术的发展给我们提供了另一个新颖的角度：利用自编码器（autoencoder；AI 模型的一种），通过合理设置 loss 函数，自动学习最可能的动力学维度，以描述我们观察到的真实数据。这将为我们利用动力学模型研究金融数据打下了坚实的基础。我们将比较两种方法的优缺点，考察 AI 模型的优势，及未来应用的可能性。

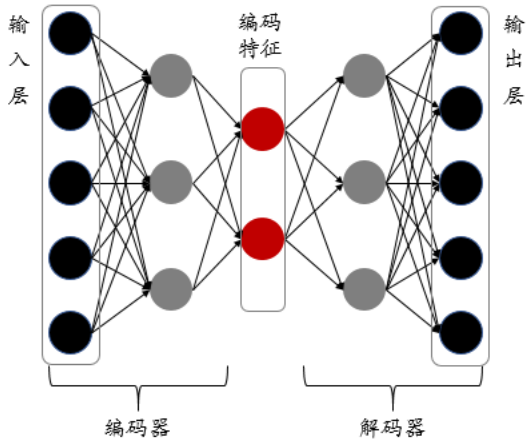
2. 方法介绍

嵌入定理是高维相空间重构理论基础，特别是 Tokens 定理保证了观测量做高维嵌入之后能重构动力学系统[4]。这里我们采用最新出自于生态学的数值研究方法：经验动力学模型（Empirical Dynamic Modeling, EDM）[5,6]。该方法是一种非参数的模型框架，它基于 Tokens 定理，利用收敛交叉映射方法（convergent cross mapping），通过对比不同嵌入维度的预测性能优劣最终确定相空间维度。该方法并不使用 AI 技术，但是可作为我们验证新技术的参照系，便于比较。

自编码器是一类基于神经网络的无监督学习算法，属于较前沿的 AI 技术。它是通过学习（或信息压缩），将数量较庞大的一组数据转换成具有**代表性**的为数不多的几列数据，从而做到数据降维，所以目前其应用范围包括无监督条件下的数据压缩，数据降噪和模式识别等。

自编码器可以分为两个过程：从输入层到中间层的数据处理过程叫做数据编码（encode）过程，从中间层到输出层则为解码(decode)过程，最后希望保证输出尽量等于输入，来减小重构误差，这样就使得中间的每一层都最大程度的保留了原有的数据特征。我们将编码层输出的编码特征视为输入数据的关键表征，实现数据降维。

图 1： 自编码结构示意图



数据来源：华泰期货研究院

而对于重构相空间来说，最关键的一环正是从单变量的时间序列数据中找到最具代表性的维度数目，从而帮助我们理解市场的动力学特征，把握其运动规律。这里我们借鉴机器学习方面的最新进展，通过识别高维空间中“假近邻”（false neighbors），并且改进自编码器的默认 loss 函数，寻找动力学模型意义下的最优维度数目。因为是处理时间序列信息，我们使用 LSTM 层设计编码器和解码器，尽管其他神经网络模型也同样可用--比如 MLP。并且使用两个 loss 项：1. 方差均值（mean squared error）；2.“最近假邻居”(FNN) 正则项。在实际应用中，因为，这两个 loss 项的变动范围不同，我们会根据数据特征对 loss 项做加权处理。

在自编码器学习过程结束后，中间编码层的隐藏单元将按照输出值的方差排序（节点重要性）。理论上判断，对于含有噪音的动力学系统数据，隐藏单元的方差将预期在某个节点出现骤降（活跃维度过度到非活跃维度），模型最后选取方差骤降前的最后一个维度作为相空间维度。我们将使用模型训练结束后，编码节点的方差值大小作为主要参考依据。

三、模型结果对比

1. 测试数据及环境

因为，自编码器算法对数据量有较高要求，我们将只选取一些历史较长的国内金融数据进行测试。数据采样频率为日度。我们采用的合约换月方式充分考虑了持仓量的稳定性和合约流动性，但该方法的主力和次主力合约可能与市面上一些数据服务商提供的连续合约数据存在些许差异。

1) 沪深 300 股指期货连续主力/次主力合约（2010-04：2021-02）

2) 沪铜连续主力/次主力合约（2002-01：2021-02）

3) 大商所豆粕连续主力/次主力合约（2000-07：2021-02）

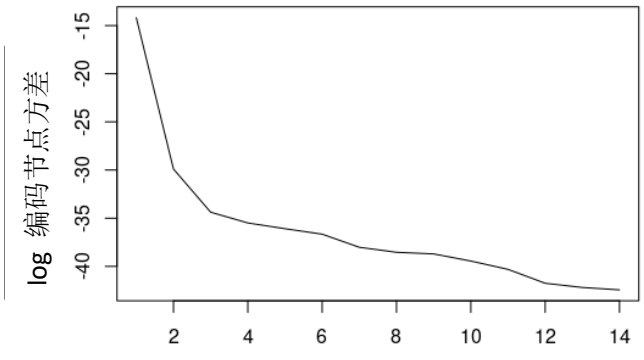
4) 郑商所棉花连续主力/次主力合约（2004-06：2021-02）

本文测试主要使用 AI 工具包有为 Tensorflow 4.2 和 Keras 2.3.0.0。

2. 维度判定方法

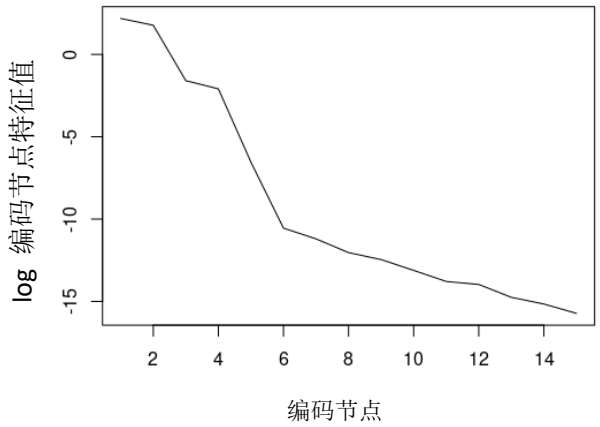
因为，主导的活跃维度体现出最关键的市场动态特征，所以，我们可以计算中间编码节点值的协方差矩阵，通过选取最大特征值的方法来判定活跃维度。尽管我们考虑的是非线性模型，同时编码节点之间可能依然存在相关性，但是作为辅助判断方法，我们也比较不同编码节点的方差值，计算方差最大的节点的方差占比。举个例子，对于 IF（长周期数据）：

图1： 自编码器编码节点对数方差值



数据来源：天软 华泰期货研究院

图2： 自编码器编码节点对数特征值



数据来源：天软 华泰期货研究院

从图上可以看到，活跃维度的特征变化明显，我们需要更加量化，且在不同品种间可横向比较的判定方法。对于特征值比较，可以使用特征值占比的方法度量，我们要求最活跃的若干维度的特征值之和超过总特征值的 99%；辅助以方差值比较，我们要求最活跃的若干维度的方差值之和超过总方差值的 90%。而经验动力学模型（EDM）的判定标准则较为确定，我们将使用模型的预测值与真值之间的相关性作为判断标准。

另外，从上图中，我们观察到自编码器的编码节点值的方差在最开始阶段，就经历了明显的活跃性骤降。所以我们认为，虽然市场波动看似随机变动，而且噪音程度很高，但是对其趋势性（或明确反转）的价格影响因素并不多，甚至接近低维动力学系统。换句话说，特定时段的价格波动主导因素很可能是比较单一（或为数不多）的。

进一步，数据在更高的维度上依然有不可忽视的影响。从模型角度来看，我们也认为，非活跃维度依然提供了有趣的信息。特别地，噪音将会在所有维度出现，这为我们采用同一套模型框架处理噪音提供了思路。另外，寻找编码节点值的内在市场驱动因素则是一个更为有趣的方向。因为，驱动性的活跃维度并不多，所以，对某一周期尺度上的影响因子做筛选（如随机森林方法[1]），并最终寻找匹配的价格主导因素将会让我们的对该周期市场影响因素有更为直观的认识和甚至前瞻性判断。

3. 测试结果

1) 主力连续合约

表格 1：主力连续长/短周期

	方差占比	方差判定维度	特征值占比	特征值判定维度	EDM 维度
短周期					
IF	99.6%	3	99.5%	2	3
CU	96.4%	3	99.9%	4	3
M	99.1%	3	99.1%	2	3
CF	99.5%	3	99.9%	3	3
长周期					
IF	92.9%	3	99.9%	3	6
CU	92.3%	4	99.5%	3	6
M	98.4%	4	99.8%	3	7
CF	91.5%	4	99.9%	3	5

资料来源：天软 华泰期货研究院

2) 次主力连续合约

表格 2：次主力连续长/短周期

	方差占比	方差加权维度	特征值占比	特征值加权维度	EDM 维度
短周期					
IF	97.5%	3	99.74%	3	3
CU	93.5%	3	99.60%	3	3
M	99.3%	4	99.99%	3	3
CF	99.7%	2	99.75%	2	3
长周期					
IF	96.2%	4	99.88%	3	6
CU	97.3%	2	99.31%	5	7

M	90.2%	3	99.99%	3	6
CF	91.8%	4	99.97%	3	5

资料来源：天软 华泰期货研究院

从测试结果可以清晰看出，合约价格变动在长/短周期上显示的特征并不相同，而实际上这样的差异并不能通过波动率观察到，必须从更高阶的统计量来进行分析。表格 3 中，以股指期货 IF 为例，不同周期波动率特征几乎没有差异。我们猜测，维度的信息更有可能与统计分布偏离度和肥尾特征相关。

表格 3：IF 收益率统计值

IF	Mean	SD	Skewness	Kurtosis
主力；短周期	1.1E-18	16.5%	-0.02	4.5
主力；长周期	0.11	16.7%	-0.42	9.5
次主力；短周期	-9E-18	17.5%	0.06	4.6
次主力；长周期	0.09	17.7%	-0.38	9.6

资料来源：天软 华泰期货研究院

同时我们也需要强调，目前主要测试维度数目而非寻找影响因素，所以即使活跃影响维度数目一样，也不代表影响因素相同。而影响因素即使相同，也不代表影响力大小一致。举例来说，当期限结构代表主导因素时，因为主力和次主力合约所处的位置存在天然差异，那么其影响方式依然可能不同。

最后需要指出，因为金融数据是一个低信噪比系统，我们不能忽视噪音带来的影响。实际上在对“最近假邻居”(FNN)估算时，我们并不能排除因为噪音干扰而无法区别真假邻居的可能性。 一般的处理方法是提高维度数目，这样一定程度上可以尽量减小噪音干扰，虽然这样明显会增加计算资源消耗。成熟的 EDM 方法针对我们测试品种得到相同或者更高的活跃维度数；而 autoencoder 算法则看起来大致保持不变，我们认为有更好的抗噪效果。

四、总结

本文我们延续了 AI 算法的应用研究，并且尝试使用**无监督学习**的模型来自动寻找金融数据中的隐含特征，并且与来自其他领域的方法进行比较。因为，无监督学习的训练数据集不会预先给出标记（人工分类），所以模型就有机会找到数据中尚未被人发现的数据特征，这是无监督学习持续吸引研究资源投入的主要原因。而在金融领域，试图发现别人尚未发现的数据规律自然是交易策略研究的主要目的之一。但是，金融数据并没有一个通用

的模型作为研究出发点，同时噪音程度又普遍较高，于是我们尝试一步步推进，将新的算法与已有算法进行对比测试，探索新 AI 算法的应用场景。

本文利用 autoencoder 模型尝试解码金融时间序列在不同分解周期上的变动规律；特别地，我们尝试挖掘金融数据的多维度运动特征。我们根据最新理论研究结果，搭建模型，并对不同品种的期货数据进行了测试比较。

我们的测试结果显示，不同品种在不同周期上似乎都体现了低维度特征，即使是 EDM 算法给出的长周期活跃维度最多也只达到 7 维。这与我们经常看到的业内多因子模型（alpha 或 beta 因子），动辄十几个因子的多因子模型并不相符。我们猜测对于一个金融工具来说，通常一段时间内的主导因素并不多。而 autoencoder 这一类偏向于数据压缩的算法则很好的抓住了这一类特征，在训练模型的中间层编码节点上体现了较为明确的信息量递减规律。

无监督学习算法也可以作为我们挖掘新的影响因子，或者因子合成的模型框架。我们将会在随后的研究过程中持续深化这类模型的研究。

五、参考文献

[1] 华泰期货量化策略年报 20201207：金融科技赋能投研系列之十一、十二：智 AI 科技慧投未（上，下）

[2] 华泰期货金融时序专题：金融科技赋能投研系列之（二至八）

[3] William Gilpin, “*Deep reconstruction of strange attractors from time series*”, arXiv: 2002.05909, (2020)

[4] Floris Takens, “*Detecting strange attractors in turbulence*”, Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898. Springer-Verlag. pp. 366 – 381, (1981)

[5] G. Sugihara, and R.M. May, “*Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series*” Nature, 344:734-741 (1990)

[6] G. Sugihara, “Nonlinear forecasting for the classification of natural time series”, Philosophical Transactions of the Royal Society of London: Mathematical, Physical and Engineering Sciences 348: 477-495 (1994)

免责声明

本报告基于本公司认为可靠的、已公开的信息编制，但本公司对该等信息的准确性及完整性不作任何保证。本报告所载的意见、结论及预测仅反映报告发布当日的观点和判断。在不同时期，本公司可能会发出与本报告所载意见、评估及预测不一致的研究报告。本公司不保证本报告所含信息保持在最新状态。本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的观点、结论和建议仅供参考，投资者并不能依靠本报告以取代行使独立判断。对投资者依据或者使用本报告所造成的一切后果，本公司及作者均不承担任何法律责任。

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